Submitted to the Annals of Statistics

MARGIN-ADAPTIVE MODEL SELECTION IN STATISTICAL LEARNING

By Sylvain Arlot, and Peter L. Bartlett

Universite Paris-Sud and University of California, Berkeley

A classical condition for fast learning rates is the margin condition, first introduced by Mammen and Tsybakov. We tackle in this paper the problem of adaptivity to this condition in the context of model selection, in a general learning framework. Actually, we consider a weaker version of this condition that allows us to take into account that learning within a small model can be much easier than in a large one. Requiring this "strong margin adaptivity" makes the model selection problem more challenging. We first prove, in a very general framework, that some penalization procedures (including local Rademacher complexities) exhibit this adaptivity when the models are nested. Contrary to previous results, this holds with penalties that only depend on the data. Our second main result is that strong margin adaptivity is not always possible when the models are not nested: for every model selection procedure (even a randomized one), there is a problem for which it does not demonstrate strong margin adaptivity.

1. Introduction. We consider in this paper the model selection problem in a quite general framework. Since our main motivation comes from the supervised binary classification setting,

*The authors gratefully acknowledge the support of the NSF under award DMS-0434383.

AMS 2000 subject classifications: Primary 68T05, 62H30; secondary 68Q32, 62G08

Keywords and phrases: statistical learning, classification, empirical risk minimization, margin condition, model selection, oracle inequalities, adaptivity, local Rademacher complexity

we focus on this framework in this introduction. Section 2 introduces the natural generalization to risk minimization problems, which we consider in the remainder of the paper.

We observe independent realizations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \dots, n$ of a random variable with distribution P, with $\mathcal{Y} = \{0, 1\}$. Our goal is to build a (data-dependent) predictor t (*i.e.* a measurable function $\mathcal{X} \mapsto \mathcal{Y}$) such that $t(X)$ is as often as possible equal to Y, where $(X, Y) \sim P$ is independent from the data. This is the *prediction* problem, in the setting of supervised binary classification. Our goal is to find t minimizing the prediction error $P\gamma(t; \cdot) := \mathbb{P}_{(X,Y)\sim P}(t(X) \neq Y)$, where γ is the 0-1 loss.

The minimizer of the prediction error, when it exists, is called the Bayes predictor s. Define the regression function $\eta(X) = \mathbb{P}_{(X,Y)\sim P}(Y = 1 \mid X)$. Then, a classical argument shows that $s(X) = \mathbb{1}_{\eta(X)\geq 1/2}$. However, s is unknown, since it depends on the unknown distribution P. Our goal is to build some predictor t from the data minimizing the prediction error, or equivalently the excess loss $\ell(s, t) := P\gamma(t) - P\gamma(s)$.

A classical approach to this problem is *empirical risk minimization*. Let $P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i,Y_i)}$ be the empirical measure and S_m be any set of predictors, which is called a *model*. The *empirical risk minimizer* over S_m is then defined as

$$\widehat{s}_m \in \arg\min_{t\in S_m} P_n\gamma(t) = \arg\min_{t\in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{t(X_i)\neq Y_i} \right\} \quad .$$

We hope that the risk of $\widehat{s}_m$ is close to that of

$$s_m \in \arg\min_{t\in \mathcal{F}_m} P\gamma(t) \quad ,$$

assuming that such a minimizer exists.

1.1. *Margin condition.* Depending on some properties of P and the complexity of S_m, the prediction error of $\widehat{s}_m$ is more or less distant from that of s_m. For instance, when S_m has a finite

Vapnik-Chervonenkis dimension V_m [24, 23] and $s \in S_m$, it has been proven (*e.g.* [16]) that

$$\mathbb{E}\left[\ell(s, \widehat{s}_m)\right] \leq C\sqrt{\frac{V_m}{n}}$$

for some numerical constant $C > 0$. This is optimal without any assumption on P, in the minimax sense: no estimator can have a smaller prediction risk (up to a numerical factor in front of C) uniformly over all distributions P [12].

However, there exist favorable situations where much smaller prediction errors ("fast rates," up to n^{-1} instead of $n^{-1/2}$) can be obtained. A sufficient condition, the so-called "margin condition," has been introduced by Mammen and Tsybakov [18]. If, for some $\epsilon_0, C_0 > 0$ and $\alpha \geq 1$,

$$\forall \epsilon \in (0, \epsilon_0], \qquad \mathbb{P}\left(|2\eta(X) - 1| \leq \epsilon\right) \leq C_0 \epsilon^{\alpha}, \tag{1}$$

if the Bayes predictor s is in S_m, and if S_m is a VC-class of dimension V_m, then the prediction error of $\widehat{s}_m$ is smaller than $L(C_0, \epsilon_0, \alpha) \ln(n) \left(V_m/n\right)^{\frac{\kappa}{2\kappa-1}}$ in expectation, where $\kappa = (1+\alpha)/\alpha$ and $L(C_0, \epsilon_0, \alpha) > 0$ only depends on C_0, ϵ_0 and α. Minimax lower bounds [20] and other upper bounds can be obtained under other complexity assumptions (*e.g.* assumption (A2) of Tsybakov [21], involving bracketing entropy). In the extreme situation where $\alpha = +\infty$, *i.e.*, for some $h > 0$,

$$\mathbb{P}\left(|2\eta(X) - 1| \leq h\right) = 0, \tag{2}$$

then the same result holds with $\kappa = 1$ and $L(h) \propto h^{-1}$. More precisely,

$$\mathbb{E}\left[\ell(s, \widehat{s}_m)\right] \leq C\left(\frac{V_m\left(1 + \ln\left(nh^2 V_m^{-1}\right)\right)}{nh}\right) \wedge \sqrt{\frac{V}{n}} \ .$$

Following the approach of Koltchinskii [14] (as well as Massart and Nédélec [20] for instance), we will consider the following generalization of the margin condition:

$$\forall t \in S, \qquad \ell(s,t) \geq \varphi\left(\sqrt{\mathrm{var}_P\left(\gamma\left(t;\cdot\right) - \gamma\left(s;\cdot\right)\right)}\right) \ , \tag{3}$$

where S is the set of predictors, and φ is a convex non-decreasing function on $[0, \infty)$ with $\varphi(0) = 0$. Indeed, the proofs of the above upper bounds on the prediction error of $\widehat{s}_m$ use only

that (1) implies (3) with $\varphi(x) = L(C_0, \epsilon_0, \alpha)x^{2\kappa}$ and $\kappa = (1+\alpha)/\alpha$, and that (2) implies (3) with $\varphi(x) = hx^2$. (See, for instance, Proposition 1 in [21].)

All these results show that the empirical risk minimizer is *adaptive to the margin condition*, since it leads to an optimal excess risk under various assumptions on the complexity of S_m. However, obtaining such rates of estimation requires knowledge of some S_m to which the Bayes predictor belongs, which is a very strong assumption.

A less restrictive framework is the following. First, we do not assume that $s \in S_m$, but only that s is "not too far" from it, in the sense that $\ell(s, s_m)$ is small. Second, we do not assume that the margin condition (3) is satisfied for all $t \in S$, but only for $t \in S_m$, which can be seen as a "local" margin condition:

$$\forall t \in S_m, \qquad \ell(s,t) \geq \varphi_m\left(\sqrt{\operatorname{var}_P\left(\gamma\left(t;\cdot\right) - \gamma\left(s;\cdot\right)\right)}\right) \quad , \tag{4}$$

where φ_m is a convex non-decreasing function on $[0,\infty)$ with $\varphi_m(0) = 0$. The fact that φ_m can depend on m allows situations where we are lucky to have a strong margin condition for some small models but the only global margin condition is loose. As proven in Section 5.2 (Proposition 2), such situations certainly exist.

1.2. *Adaptive model selection.* Assume now that we are not given a single model but a whole family $\left(S_m\right)_{m\in\mathcal{M}_n}$. By empirical risk minimization, we obtain a family $\left(\widehat{s}_m\right)_{m\in\mathcal{M}_n}$ of predictors, from which we would like to select some $\widehat{s}_{\widehat{m}}$ with a prediction error $P\gamma\left(\widehat{s}_{\widehat{m}}\right)$ as small as possible. The aim of such a *model selection procedure* $((X_1,Y_1),\ldots,(X_n,Y_n)) \mapsto \widehat{m} \in \mathcal{M}_n$ is to satisfy an *oracle inequality* of the form

$$\ell(s, \widehat{s}_m) \leq C \inf_{m\in\mathcal{M}_n}\left\{\ell(s,s_m) + R_{m,n}\right\} \quad , \tag{5}$$

where the leading constant $C \geq 1$ should be close to one and the remainder term $R_{m,n}$ should be close to the value $P\gamma\left(\widehat{s}_m\right) - P\gamma\left(s_m\right)$. Typically, one proves that (5) holds either in expectation, or with high probability.

Assume for instance that $\varphi_m(x) = h_m x^2$ for some $h_m > 0$ and S_m has a finite VC-dimension $V_m \geq 1$. In view of the aforementioned minimax lower bounds, one cannot hope to prove an oracle inequality (5) with a remainder $R_{m,n}$ smaller than

$$\frac{V_m}{nh_m} \wedge \sqrt{\frac{V_m}{n}} \ .$$

Then, *adaptive model selection* occurs when $\widehat{m}$ satisfies an oracle inequality (5) with $R_{m,n}$ of the order of this minimax lower bound. More generally, let C_m be some complexity measure of S_m (such as its VC-dimension, the ρ appearing in Tsybakov's assumption [21], *etc.*) and define $R_n(C_m, \varphi_m)$ as the minimax prediction error over the set of distributions P such that $s \in S_m$, S_m has a complexity at most C_m and the local margin condition (4) is satisfied with φ_m. A margin adaptive model selection procedure should satisfy an oracle inequality of the form

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m \in \mathcal{M}_n} \{ \ell(s, s_m) + R_n(C_m, \varphi_m) \} \tag{6}$$

without using the knowledge of C_m and φ_m. We call this property "strong margin adaptivity", to emphasize the fact that this is more challenging than adaptivity to a margin condition that holds uniformly over the models.

1.3. *Penalization.* We focus in particular in this article on *penalization* procedures, which are defined as follows. Let $\mathrm{pen} : \mathcal{M}_n \mapsto [0, \infty)$ be a (data-dependent) function, and define

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ P_n \gamma (\widehat{s}_m) + \mathrm{pen}(m) \} \ .$$

Since our goal is to minimize the prediction error of $\widehat{s}_m$, the *ideal penalty* would be

$$\mathrm{pen}_{\mathrm{id}}(m) := P\gamma (\widehat{s}_m) - P_n \gamma (\widehat{s}_m) \ , \tag{7}$$

but it is unknown because it depends on the distribution P. A classical way of designing a penalty is to estimate $\mathrm{pen}_{\mathrm{id}}(m)$, or at least a tight upper bound on it.

We consider in particular *local complexity measures* [17, 9, 7, 14], because they estimate $\mathrm{pen}_{\mathrm{id}}$ tightly enough to achieve fast estimation rates when the margin condition holds true. See Section 3.2 for a detailed definition of these penalties.

1.4. *Related results.* There is a considerable literature on margin adaptivity, in the context of model selection as well as model aggregation. First, notice that the results of Massart and Nédélec [20] show that with the knowledge of C_m and φ_m, one can choose a model realizing the bias-complexity trade-off, so that (6) is satisfied (maybe up to some $\ln(n)$ factor in front of the remainder term; these results are stated with the margin condition (3) but actually use its local version (4) only). Hence, the challenge of strong margin adaptivity (6) is mainly to find a completely data-driven procedure having a similar performance.

Most of the papers consider the uniform margin condition, *i.e.* with $\varphi_m \equiv \varphi$. Penalization methods have been studied in [17, 9, 7, 14] (with localized random penalties) and [22] (with other penalties, in a particular framework). Adaptivity to the margin has also been considered with a regularized boosting method [10], the hold-out [11] and in a PAC-Bayes framework [5]. Aggregation methods have been studied in [21, 15]. Notice also that a completely different approach is possible: estimate first the regression function η (possibly through model selection), then use a plug-in classifier, and this works provided η is smooth enough [6].

It is quite unclear whether any of these results can be extended to strong margin adaptivity (actually, we will prove that this needs additional restrictions in general). To our knowledge, the only results allowing φ_m to depend on m can be found in [14]. First, when the models are nested, a comparison method based on local Rademacher complexities attains strong margin adaptivity, assuming that $s \in \bigcup_{m \in \mathcal{M}_n} S_m$ (Theorem 7; and it is quite unclear whether this still holds without the latter assumption). Second, a penalization method based on local Rademacher complexities has the same property in the general case, but it uses the knowledge of $(\varphi_m)_{m \in \mathcal{M}_n}$ (Theorems 6 and 11).

Our claim is that when φ_m does strongly depend on m, it is crucial to take it into account to choose the best model in $\mathcal{M}_n$. And such situations occur, as proven by our Proposition 2 in Section 5.2. But assuming either $s \in \bigcup_{m\in\mathcal{M}_n} S_m$ or that φ_m is known is not realistic. Our goal is to investigate the kind of results which can be obtained with *completely data-driven* procedures, in particular when $s \notin \bigcup_{m\in\mathcal{M}_n} S_m$.

1.5. *Our results.* In this paper, we aim to understand when strong margin adaptivity can be obtained for data-dependent model selection procedures. Notice that we do not restrict ourselves to the classification setting. We consider a much more general framework (as in [14] for instance), which is described in Section 2. We prove two kinds of results. First, when models are nested, we show that some penalization methods are strongly margin adaptive (Theorem 1). In particular, this result holds for the local Rademacher complexities (Corollary 1). Compared to previous results (in particular the ones of [14]), our main advance is that our penalties do not require the knowledge of $(\varphi_m)_{m\in\mathcal{M}_n}$, and we do not assume that the Bayes predictor belongs to any of the models.

Our second result probes the limits of strong margin adaptivity, without the nested assumption. For every sample size n, there is a distribution P and a family of models such that any "reasonable" penalization procedure fails to be strongly margin adaptive with a positive probability (Theorem 2). In addition, the same negative result holds for any model selection procedure — even a randomized one — provided that the distribution P is allowed to depend on the model selection rule (Theorem 3). Hence, the previous positive results can not be extended outside of the nested case for a general distribution P.

Where is the boundary between these two extremes? Obviously, the nested assumption is not necessary. For instance, when the global margin assumption is indeed tight ($\varphi = \varphi_m$ for every $m \in \mathcal{M}_n$), margin adaptivity can be obtained in several ways, as mentioned in the previous section. We try in Section 5 to sketch some situations where strong margin adaptivity is possible.

More precisely, we state a general oracle inequality (Theorem 4), valid for any family of models and any distribution P. We then discuss assumptions under which its remainder term is small enough to imply strong margin adaptivity.

This paper is organized as follows. We describe the general setting in Section 2. We consider in Section 3 the nested case, in which strong margin adaptivity holds. Negative results (*i.e.* lower bounds on the prediction error of a general model selection procedure) are stated in Section 4. The line between these two situations is sketched in Section 5. We discuss our results in Section 6. All the proofs are given in Section 7.

2. The general empirical risk minimization framework. Although our main motivation comes from the classification problem, it turns out that all our results can be proven in the much more general setting of empirical risk minimization. As explained below, this setting includes binary classification with the 0-1 loss, bounded regression and several other frameworks. In the rest of the paper, we will use the following general notation, in order to emphasize the generality of our results.

We observe independent realizations $\xi_1, \dots, \xi_n \in \Xi$ of a random variable with distribution P, and we are given a set $\mathcal{F}$ of measurable functions $\Xi \mapsto [0,1]$. Our goal is to build some (data-dependent) f such that $P(f) := \mathbb{E}_{\xi \sim P}\left[f(\xi) \right]$ is as small as possible. For the sake of simplicity, we assume that there is a minimizer $f^\star$ of $P(f)$ over $\mathcal{F}$.

This includes the prediction framework, in which $\Xi = \mathcal{X} \times \mathcal{Y}$, $\xi_i = (X_i, Y_i)$,

$$\mathcal{F} := \left\{ \xi \mapsto \gamma\left(t; \xi\right) \text{ s.t. } t \in S \right\} \quad ,$$

where $\gamma : S \times \Xi \mapsto [0,1]$ is any contrast function. Then, $f^\star$ is equal to $\gamma\left(s; \cdot\right)$, where s is the Bayes predictor. In the binary classification framework,$\mathcal{Y} = \left\{0, 1\right\}$ and we can take the 0-1 contrast $\gamma(t; (x,y)) = \mathbb{1}_{t(x) \neq y}$ for instance. This is the case that we have considered in introduction. In the

bounded regression framework, assuming that $\mathcal{Y} = [0,1]$, we can take the least-squares contrast,

$$\gamma(t;(x,y)) = (t(x)-y)^2 .$$

Notice that many other contrast functions γ can be considered, provided that they take their values in $[0,1]$. Because of the one-to-one correspondence between predictors t and functions $f_t := \gamma(t;\cdot)$ in the prediction framework, in the following we will call functions $f \in \mathcal{F}$ *predictors*, even if we do not restrict ourselves to the prediction problem.

The *empirical risk minimizer* over $\mathcal{F}_m \subset \mathcal{F}$ (called a model) can then be defined as

$$\widehat{f}_m \in \arg\min_{f\in\mathcal{F}_m} P_n(f) \ .$$

We hope that its risk is close to that of $f_m \in \arg\min_{f\in\mathcal{F}_m} P(f)$, assuming that such a minimizer exists. In the prediction case, defining $\mathcal{F}_m := \{ f_t \text{ s.t. } t \in S_m \}$, we have $\widehat{f}_m = f_{\widehat{s}_m}$ and $f_m = f_{s_m}$.

We can now write the global margin condition as follows:

$$\forall f \in \mathcal{F}, \qquad P(f-f^\star) \geq \varphi\left(\sqrt{\mathrm{var}_P(f-f^\star)}\right) \ , \tag{8}$$

where φ is a convex non-decreasing function on $[0,\infty)$ with $\varphi(0)=0$. Similarly, the local margin condition is

$$\forall f \in \mathcal{F}_m, \qquad P(f-f^\star) \geq \varphi_m\left(\sqrt{\mathrm{var}_P(f-f^\star)}\right) \ . \tag{9}$$

Notice that most of the upper and lower bounds on the risk under the margin condition given in the introduction stay valid in the general empirical minimization framework, at least when $\varphi_m(x) = (h_m x^2)^{\kappa_m}$ for some $h_m > 0$ and $\kappa_m \geq 1$ (see for instance [20, 14]). Assume that S_m is a VC-class of dimension V_m. If $\varphi_m(x) = h_m x^2$,

$$\mathbb{E}\left[\ell(s,\widehat{s}_m)\right] \leq 2\ell(s,s_m) + C\left(\frac{\ln(n)V_m}{nh_m}\right) \wedge \sqrt{\frac{V_m}{n}}$$

for some numerical constant $C > 0$. If $\varphi_m(x) = \left(h_m x^2\right)^{\kappa_m}$ for some $h_m > 0$ and $\kappa_m \geq 1$,

$$\mathbb{E}\left[\ell(s, \widehat{s}_m)\right] \leq 2\ell(s, s_m) + C\left[L(h_m, \kappa_m)\ln(n)\left(\frac{V_m}{nh_m}\right)^{\frac{2\kappa_m}{\kappa_m - 1}}\right] \wedge \sqrt{\frac{V_m}{n}}$$

for some constant $L(h_m, \kappa_m) > 0$.

Given a collection $\left(\mathcal{F}_m\right)_{m \in \mathcal{M}_n}$ of models, we are looking for a model selection procedure $(\xi_1, \ldots, \xi_n) \mapsto \widehat{m} \in \mathcal{M}_n$ satisfying an *oracle inequality* of the form

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \leq C \inf_{m \in \mathcal{M}_n}\left\{P\left(\widehat{f}_m - f^\star\right) + R_{m,n}\right\} \quad , \tag{10}$$

with a leading constant C close to 1 and a remainder term $R_{m,n}$ as small as possible. Similarly to (6), we define a strongly margin adaptive procedure as any $\widehat{m}$ such that (10) holds with some constant C and $R_{m,n}$ of the order of the minimax risk $R_n(C_m, \varphi_m)$.

Defining penalization methods as

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n}\left\{P_n\left(\widehat{f}_m\right) + \mathrm{pen}(m)\right\} \tag{11}$$

for some data-dependent $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}$, the ideal penalty is $\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\left(\widehat{f}_m\right)$.

3. Margin adaptive model selection for nested models.

3.1. *General result.* Our first result is a sufficient condition for penalization procedures to attain margin adaptivity when the models are nested (Theorem 1). Since this condition is satisfied by local Rademacher complexities, this leads to a completely margin adaptive penalization procedure (Corollary 1).

THEOREM 1. *Fix $\left(\varphi_m\right)_{m \in \mathcal{M}_n}$ such that the local margin conditions* (9) *hold. Let $t > 0$ and assume that there are constants $c, \eta \in (0,1)$ and $C_1, C_2 \geq 0$ such that the following holds:*

- *the models $\mathcal{F}_m$ are nested and $\mathcal{M}_n$ is finite.*

- *lower bounds on the penalty: with a probability at least* $1-\eta$*, for every* $m, m' \in \mathcal{M}_n$*,*

$$(1-c)\,\mathrm{pen}(m) \geq (P-P_n)\left(\widehat{f}_m - f_m\right) + \frac{t}{n} \geq 0 \tag{12}$$

$$\mathcal{F}_{m'} \subset \mathcal{F}_m \;\Rightarrow\; c\,\mathrm{pen}(m) \geq v(m) - C_1 v(m') - C_2 P(f_{m'} - f^\star)\ , \tag{13}$$

$$\text{where} \quad v(m) := \sqrt{\frac{2t}{n}\,\mathrm{var}_P(f_m - f^\star)}\ .$$

Then, if $\widehat{m}$ *is defined by* (11)*, with probability at least* $1-\eta-\mathrm{Card}(\mathcal{M}_n)e^{-t}$*, we have for every* $\epsilon \in (0,1)$

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \leq \frac{1}{1-\epsilon}\inf_{m\in\mathcal{M}_n}\left\{(1+\epsilon+C_2)P\left(f_m - f^\star\right) + (2-c)\,\mathrm{pen}(m) \right. \\ \left. +(1+C_1)\varphi_m^\star\left(\sqrt{\frac{2t}{\epsilon^2 n}}\right) \wedge \frac{1}{\sqrt{n}} + \frac{2t}{3n}\right\}\ , \tag{14}$$

where $\varphi_m^\star(x) := \sup_{y\geq 0}\{xy - \varphi_m(y)\}$ *is the convex conjuguate of* φ_m*.*

REMARK 1. 1. If $\mathrm{pen}(m)$ is of the right order, *i.e.* not much larger than $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$, then Theorem 1 is a strong margin adaptivity result. Indeed, assuming that $\varphi_m(x) = \left(h_m x^2\right)^{\kappa_m}$, the remainder term is not too large, since $\varphi_m^\star(n^{-1}) = L(h_m,\kappa_m)x^{2\kappa_m/(2\kappa_m-1)}$ for some positive constant $L(h_m,\kappa_m)$. Hence, choosing $\epsilon = 1/2$ for instance, we can rewrite (14) as

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \leq L(C_2,C_1)\inf_{m\in\mathcal{M}_n}\left\{P\left(f_m - f^\star\right) + \mathrm{pen}(m) + L(h_m,\kappa_m)\left(\frac{t}{n}\right)^{\frac{\kappa_m}{2\kappa_m-1}}\right\}\ ,$$

for some positive constants $L(C_2,C_1)$ and $L(h_m,\kappa_m)$. When φ_m is a general convex function, minimax estimation rates are no longer available, so that we do not know whether the remainder term in (14) is of the right order. However, no better risk bound is known, even for a single model to which s belongs.

2. In the case that the φ_m are known, methods involving local Rademacher complexities and $(\varphi_m)_{m\in\mathcal{M}_n}$ satisfy oracle inequalities similar to (14) (see Theorems 6 and 11 in [14]). Also, Theorem 7 of [14] shows that adaptivity is possible using a comparison method, provided

that $f^\star$ belongs to one of the models. It is not clear whether this method achieves the optimal bias-variance trade-off in the general case, as in our result.

3. In Theorem 1, we cannot make t depend on m. If we want the probability bound to be good enough, we have to choose $t \geq \ln \operatorname{Card}(\mathcal{M}_n) + 2\ln(n)$, so that the result is only meaningful when $\mathcal{M}_n$ does not grow exponentially with n. We do not know whether Theorem 1 can be extended to very large collections $\mathcal{M}_n$, assuming for instance that $t = t_m$ is a nondecreasing function of m.

3.2. *Local Rademacher complexities.* Although Theorem 1 applies to any penalization procedure that satisfies assumptions (12) and (13), we focus on methods based on local Rademacher complexities. Let us define precisely these complexities. We mainly use the notation of [14]:

- for every $\delta > 0$, the δ minimal set of $\mathcal{F}_m$ w.r.t the distribution P is
$$\mathcal{F}_{m,P}(\delta) := \left\{ f \in \mathcal{F}_m \text{ s.t. } P(f) - \inf_{g \in \mathcal{F}_m} P(g) \leq \delta \right\}$$
- the $L^2(P)$ diameter of the δ minimal set:
$$D_P(\mathcal{F}_m; \delta) = \sup_{f,g \in \mathcal{F}_{m,P}(\delta)} P\left((f-g)^2 \right)$$
- the expected modulus of continuity of $(P - P_n)$ over $\mathcal{F}_m$:
$$\phi_n(\mathcal{F}_m; P; \delta) = \mathbb{E} \sup_{f,g \in \mathcal{F}_{m,P}(\delta)} |(P_n - P)(f-g)| \quad .$$

We then define
$$U_n(\mathcal{F}_m; \delta; t) := \overline{K} \left(\phi_n(\mathcal{F}_m; P; \delta) + D_P(\mathcal{F}_m; \delta) \sqrt{\frac{t}{n}} + \frac{t}{n} \right) \quad ,$$
where $\overline{K} > 0$ is a numerical constant (to be chosen later). The (ideal) local complexity $\overline{\delta}_n(\mathcal{F}_m; t)$ is (roughly) the smallest positive fixed-point of $r \mapsto U_n(\mathcal{F}_m; r; t)$. More precisely,
$$\overline{\delta}_n(\mathcal{F}_m; t) := \inf \left\{ \delta > 0 \text{ s.t. } \sup_{\sigma \geq \delta} \left\{ \frac{U_n(\mathcal{F}_m; \sigma; t)}{\sigma} \right\} \leq \frac{1}{2q} \right\} \tag{15}$$

where $q > 1$ is a numerical constant.

Two important points, which follow from Theorem 1 and 3 of Koltchinskii [14], are that:

1. $\overline{\delta}_n(\mathcal{F}_m; t)$ is large enough to satisfy assumption (12) with a probability at least $1 - \log_q(n/t)\, e^{-t}$ for each model $m \in \mathcal{M}_n$.
2. there is a completely data-dependent $\widehat{\delta}_n(\mathcal{F}_m; t)$ such that

$$\forall m \in \mathcal{M}_n, \quad \mathbb{P}\left(\widehat{\delta}_n(\mathcal{F}_m; t) \geq \overline{\delta}_n(\mathcal{F}_m; t)\right) \geq 1 - 5 \ln_q\left(\frac{n}{t}\right) e^{-t} \ .$$

   This data-dependent $\widehat{\delta}_n(\mathcal{F}_m; t)$ is a resampling estimate of $\overline{\delta}_n(\mathcal{F}_m; t)$, called the "local Rademacher complexity".

Before stating the main result of this section, let us recall the definition of $\widehat{\delta}_n(\mathcal{F}_m; t)$, as in [14]. We need the following additional notation:

- for every $\delta > 0$, the empirical δ minimal set of $\mathcal{F}_m$ w.r.t the distribution P is

$$\widehat{\mathcal{F}}_{n,m,P}(\delta) := \left\{ f \in \mathcal{F}_m \text{ s.t. } P_n(f) - \inf_{g \in \mathcal{F}_m} P_n(g) \leq \delta \right\}$$

- the empirical $L^2(P)$ diameter of the empirical δ minimal set:

$$D_{n,P}(\mathcal{F}_m; \delta) = \sup_{f,g \in \widehat{\mathcal{F}}_{n,m,P}(\delta)} P_n\left((f-g)^2\right)$$

- the modulus of continuity of $(P - P_n)$ over $\mathcal{F}_m$:

$$\widehat{\phi}_n(\mathcal{F}_m; P; \delta) = \sup_{f,g \in \widehat{\mathcal{F}}_{n,m,P}(\delta)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(\xi_i) - g(\xi_i)) \right| \ ,$$

  where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables (*i.e.*, ϵ_i takes the values $+1$ and -1 with probability $1/2$ each).

Defining

$$\widehat{U}_n(\mathcal{F}_m; \delta; t) := \widehat{K}\left(\widehat{\phi}_n(\mathcal{F}_m; P; \widehat{c}\delta) + \widehat{D}_{n,P}(\mathcal{F}_m; \widehat{c}\delta)\sqrt{\frac{t}{n}} + \frac{t}{n} \right)$$

(where $\widehat{K}, \widehat{c} > 0$ are numerical constants, to be chosen later), the *local Rademacher complexity* $\widehat{\delta}_n(\mathcal{F}_m; t)$ is (roughly) the smallest positive fixed-point of $r \mapsto \widehat{U}_n(\mathcal{F}_m; r; t)$. More precisely,

$$\widehat{\delta}_n(\mathcal{F}_m; t) := \inf \left\{ \delta > 0 \text{ s.t. } \sup_{\sigma \geq \delta} \left\{ \frac{\widehat{U}_n(\mathcal{F}_m; \sigma; t)}{\sigma} \right\} \leq \frac{1}{2q} \right\} \tag{16}$$

where $q > 1$ is a numerical constant.

COROLLARY 1 (Margin adaptivity for local Rademacher complexities). *There exist numerical constants $L > 0$, $\overline{K} > 0$ and $q > 1$ such that the following holds. Assume that*

$$\forall m \in \mathcal{M}_n, \quad \mathrm{pen}(m) \geq \frac{7}{2} \overline{\delta}_n(\mathcal{F}_m; t) \ , \tag{17}$$

where $\overline{\delta}_n(\mathcal{F}_m; t)$ is defined by (15) *(and depends on both $\overline{K}$ and q). Assume moreover that the models $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$ are nested and*

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \left(\widehat{f}_m \right) + \mathrm{pen}(m) \right\} \ .$$

Then, with probability at least $1 - L \ln_q \left(\frac{n}{t} \right) \mathrm{Card}(\mathcal{M}_n) e^{-t}$, for every $\epsilon \in (0,1)$,

$$\begin{aligned} P \left(\widehat{f}_{\widehat{m}} - f^\star \right) \leq \frac{1}{1-\epsilon} \inf_{m \in \mathcal{M}_n} \left\{ \left(1 + \epsilon + \frac{2}{\overline{K} q} \right) P \left(f_m - f^\star \right) + \frac{9}{7} \mathrm{pen}(m) \right. \\ \left. + (1+\sqrt{2}) \varphi_m^\star \left(\sqrt{\frac{2t}{\epsilon^2 n}} \right) + \frac{2t}{3n} \right\} \ . \end{aligned} \tag{18}$$

In particular, this holds when $\mathrm{pen}(m) = \frac{7}{2} \widehat{\delta}_n(\mathcal{F}_m; t)$, provided that $\widehat{K}, \widehat{c} > 0$ are larger than some constants depending only on $\overline{K}, q$.

REMARK 2. One can always enlarge the constants $\overline{K}$ and q, making the leading constant of the oracle inequality (18) closer to one, at the price of enlarging $\overline{\delta}_n(\mathcal{F}_m; t)$ (hence $\mathrm{pen}(m)$ or $\widehat{\delta}_n(\mathcal{F}_m; t)$). We do not know whether it is possible to make the leading constant closer to one without changing the penalization procedure itself.

As we show in Section 5.2, there are distributions P and collections of models $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$ such that this is a strong improvement over the "uniform margin" case, in terms of prediction error. It seems reasonable to expect that this happens in a significant number of practical situations.

In Section 5, we state a more general result (from which Theorem 1 is a corollary) which suggests why it is more difficult to prove Corollary 1 when φ_m really depends on m. This general result is also useful to understand how the nestedness assumption might be relaxed in Theorem 1.

The reason why Corollary 1 implies margin adaptivity is that the local Rademacher complexities are not too large when the local margin condition is satisfied, together with a complexity assumption on $\mathcal{F}_m$. Indeed, there exists a distribution-dependent $\widetilde{\delta}_n(\mathcal{F}_m;t)$ (defined as $\overline{\delta}_n(\mathcal{F}_m;t)$ with $U_n(\mathcal{F}_m;\delta;t)$ replaced by $K_1U_n(\mathcal{F}_m;K_2\delta;t)$ for some numerical constants $K_1,K_2>0$, related to $\widehat{K}$ and $\widehat{c}$) such that

$$\forall m\in\mathcal{M}_n,\quad \mathbb{P}\left(\widetilde{\delta}_n(\mathcal{F}_m;t)\geq\widehat{\delta}_n(\mathcal{F}_m;t)\geq\overline{\delta}_n(\mathcal{F}_m;t)\right)\geq 1-5\ln_q\left(\frac{n}{t}\right)e^{-t}\ .$$

(See Theorem 3 of [14].) This leads to several upper bounds on $\widehat{\delta}_n(\mathcal{F}_m;t)$ under the local margin condition (9) (combining Lemma 5 of [14] with the examples of its Section 2.5). For instance, in the binary classification case, when $\mathcal{F}_m$ is the class of 0-1 loss functions associated with a VC-class S_m of dimension V_m, such that the margin condition (9) holds with $\varphi_m(x)=h_mx^2$, we have for every $t>0$ and $\epsilon\in(0,1]$,

$$\overline{\delta}_n(\mathcal{F}_m;t)\leq\epsilon P(f_m-f^\star)+\frac{K_3}{nh_m}\left[\epsilon^{-1}t+\epsilon^{-2}V_m\ln\left(\frac{n\epsilon^2h_m}{K_4V_m}\right)\right]\ , \tag{19}$$

where K_3 and K_4 depend only on $\overline{K}$. (Similar upper bounds hold under several other complexity assumptions on the models $\mathcal{F}_m$, *cf.* [14].) In particular, when each model S_m is a VC-class of dimension V_m, $\varphi_m(x)=h_mx^2$, $\mathrm{pen}(m)=\frac{7}{2}\widehat{\delta}_n(\mathcal{F}_m;t)$ and $t=\ln(\mathrm{Card}(\mathcal{M}_n))+3\ln(n)$, (18) implies that

$$P\left(\widehat{f}_{\widehat{m}}-f^\star\right)\leq C\inf_{m\in\mathcal{M}_n}\left\{P(f_m-f^\star)+\frac{\ln(\mathrm{Card}(\mathcal{M}_n))+\ln(n)+V_m\ln(enh_m/V_m)}{nh_m}\right\}$$

with probability at least $1 - Kn^{-2}$, for some numerical constants $C, K > 0$. Up to some $\ln(n)$ factor, this is a margin adaptive model selection result, provided that $\mathrm{Card}(\mathcal{M}_n)$ is not larger than some power of n. Notice that the $\ln(n)$ factor is sometimes necessary (as shown by [20]), meaning that this upper bound is optimal.

4. Lower bounds for some non-nested models. In this section, we investigate the assumption in Theorem 1 that the models $\mathcal{F}_m$ are nested. We show that (strong) margin adaptive model selection is not always possible, even for randomized model selection procedures, if we relax this assumption.

4.1. *"Reasonable" penalization procedures.* First assume that $\widehat{m}$ is obtained from a penalization procedure that assigns a null penalty to singletons. This is, for instance, the case when $\mathrm{pen}(m)$ is proportional to the expectation of $\mathrm{pen}_{\mathrm{id}}$, or when $\mathrm{pen}(m)$ is any quantile of the distribution of $(P - P_n)(\widehat{f}_m - f_m)$. Then, the following result shows that there exists a model selection problem for which such a method fails with positive probability.

THEOREM 2. *If* $\mathrm{Card}(\mathcal{X}) \geq 2$, *there are two classes* $\mathcal{F}_0$ *and* $\mathcal{F}_1$ *of functions* $\mathcal{X} \times \{0,1\} \mapsto [0,1]$ *and a numerical constant* $\kappa > 0$ *such that the following holds. For every* $\gamma > 0$, *there is a constant* L_γ *such that for every* $n \geq 1$ *and* $t_0, t_1 \in [0, \gamma \ln(n)]$, *there is a distribution* P *such that for any penalization procedure satisfying*

$$\mathrm{Card}(\mathcal{F}_m) = 1 \Rightarrow \mathrm{pen}(m) = 0 \tag{20}$$

and any

$$\widehat{m} \in \arg\min_{m \in \{0,1\}} \left\{ P_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \right\} \quad ,$$

we have

$$\mathbb{P}\left(P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \geq \frac{L_\gamma \sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}} \left\{ P\left(\widehat{f}_m - f^\star\right) + v(m) + \frac{t_m}{n h_m} \right\} \right) \geq \kappa > 0 \ , \tag{21}$$

$$\text{where} \quad \forall m \in \{0,1\}, \quad h_m := \inf_{f \in \mathcal{F}_m} \left\{ \frac{P(f - f^\star)}{\mathrm{var}_P(f - f^\star)} \right\} \ .$$

In addition, (21) *implies*

$$\mathbb{E}\left[P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \right] \geq \frac{\kappa L_\gamma \sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}} \left\{ \mathbb{E}\left[P\left(\widehat{f}_m - f^\star\right) \right] + v(m) + \frac{t_m}{n h_m} \right\} \ . \tag{22}$$

In other words, Theorem 1 cannot be generalized to non-nested models without some other assumption on P or on the models.

REMARK 3.

1. The counterexample in the proof of Theorem 2 holds in the classification case with the 0-1 loss.
2. Assumption (20) holds for $\mathrm{pen}(m) = (P - P_n)(\widehat{f}_m - f_m)$ as well as any quantile of its distribution and its expectation

$$\mathbb{E}\left[(P - P_n)(\widehat{f}_m - f_m) \right] = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) \right] \ .$$

   It also holds for the local Rademacher complexities $\overline{\delta}_n(\mathcal{F}_m; t_m)$ and $\widehat{\delta}_n(\mathcal{F}_m; t_m)$ (up to the t_m/n term, but t_m should not depend on m when there are only two models).
3. Assumption (20) is reasonable if we do not have more information on P: when a model $\mathcal{F}_m$ is a singleton, there is no hope to estimate its "complexity." Without additional information about P, comparing the prediction abilities of two singletons can only be made by comparing their empirical risks.
4. We consider margin adaptivity with $\varphi_m(x) = h_m x^2$, whereas the margin condition is also satisfied with other functions φ_m. This is both for simplicity reasons, and because this choice emphasizes that one can hope for learning rates of order $1/(nh_m)$. The meaning of Theorem 2 is then mainly that one can not guarantee to learn at a rate better than $1/\sqrt{n}$,

whereas the best model has an excess loss of order $1/n$, even with the additional term $1/(nh_m)$.

5. The counterexample given in the proof of Theorem 2 is highly nonasymptotic, in the sense that the distribution P strongly depends on n. If P and $\mathcal{F}_0, \mathcal{F}_1$ were fixed, it is well known that empirical risk minimization leads to asymptotic optimality. This illustrates a significant difference between the asymptotic and non-asymptotic frameworks.

4.2. *Lower bound for any model selection procedure.* More precisely, we can modify the proof of Theorem 2 in order to make it valid for *any model selection procedure*, at the price of making the distribution P depend on it (otherwise, for any P, one among the deterministic choices $\widehat{m} \equiv 0$ and $\widehat{m} \equiv 1$ is optimal). In other words, we have the following "minimax model selection lower bound":

THEOREM 3. *If* $\operatorname{Card}(\mathcal{X}) \geq 2$, *there are two classes* $\mathcal{F}_0$ *and* $\mathcal{F}_1$ *of functions* $\mathcal{X} \times \{0,1\} \mapsto [0,1]$ *and a numerical constant* $\kappa' > 0$ *such that the following holds. For any* $\gamma > 0$, *there is a constant* $L_\gamma > 0$ *such that for every* $n \in \mathbb{N}$, $t_0, t_1 \in [0, \gamma \ln(n)]$ *and* $\widehat{m}$ *a model selection procedure (*i.e. *a function* $(\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{M} = \{0,1\}$*), there is a distribution* P *such that*

$$\mathbb{P}\left(P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \geq \frac{L_\gamma \sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}} \left\{ P\left(\widehat{f}_m - f^\star\right) + v(m) + \frac{t_m}{n h_m} \right\} \right) \geq \kappa' \tag{23}$$

$$\mathbb{E}\left[P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \right] \geq \frac{\kappa' L_\gamma \sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}} \left\{ \mathbb{E}\left[P\left(\widehat{f}_m - f^\star\right) + v(m) + \frac{t_m}{n h_m} \right] \right\} \ , \tag{24}$$

where

$$h_m := \inf_{f \in \mathcal{F}_m} \left\{ \frac{P(f - f^\star)}{\operatorname{var}_P(f - f^\star)} \right\} \ .$$

This generalizes the conclusions of the previous subsection, since we have proven that Theorem 1 cannot be generalized to non-nested models without further assumptions on P or on the models.

REMARK 4. The same result holds for randomized rules $\widehat{m} : (\mathcal{X} \times \mathcal{Y})^n \mapsto [0,1]$ (where the value of $\widehat{m}((X_i, Y_i)_{1 \le i \le n})$ is the probability assigned to the choice of the model $\mathcal{F}_1$). Hence, aggregating models instead of selecting one does not modify the conclusion of Theorem 3.

4.3. *Comments.* With Theorem 1, we have proven a margin adaptivity result for nested models, which holds true when the penalty is built upon local Rademacher complexities. This means that adaptive model selection is attainable for nested models, whatever the distribution of the data. On the other hand, Theorem 2 gives a simple example where no "reasonable" penalty can satisfy an oracle inequality (10) with a leading constant smaller than n^β, where β can be made arbitrarily close to $1/2$. By "reasonable", we mean that we do not penalize two singletons in a different way, which would imply choosing a model with larger empirical risk. Even if we admit "unreasonable" model selection rules, Theorem 3 shows that we cannot attain our adaptive model selection goals uniformly over all distributions.

Looking carefully at the examples given in the proofs of Theorem 2 and 3, it appears that the main reason why they are particularly tough is that we are quite "lucky" with one of the models: it has simultaneously a very small bias, a very small size and a large margin parameter, while other models with very similar appearance are much worse. See also Remark 5 in Section 7.2. When looking for more general margin adaptivity result, we then must keep in mind that this is a hopeless task in such situations.

5. General collections of models. As proven in Section 4, we cannot hope to obtain margin adaptivity without any assumption on either P or the models. The purpose of this section is to explain what can still be proven in the general case, and why this is weaker than our Theorem 1.

5.1. *A general oracle inequality.* We start with a general result for penalties satisfying the lower bound (12).

THEOREM 4. *Let $(t_m)_{m\in\mathcal{M}_n}$ be any sequence of positive numbers. Let $\widehat{m}$ be defined by* (11) *and assume that there is some $c \in (0,1)$ such that*

$$\forall m \in \mathcal{M}_n, \qquad (1-c)\operatorname{pen}(m) \geq (P-P_n)\left(\widehat{f}_m - f_m\right) + \frac{t_m}{n} \geq 0 \tag{25}$$

on an event of probability at least $1-\eta$.

Then, there is an event of probability at least $1-\eta-\sum_{m\in\mathcal{M}_n} e^{-t_m}$ on which the following holds: for every $\epsilon \in (0,1)$,

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \leq \frac{1}{1-\epsilon}\inf_{m\in\mathcal{M}_n}\left\{P\left(\widehat{f}_m - f^\star\right) + \operatorname{pen}(m) + v(m) + \frac{2t_m}{3n}\right\} + V_n \tag{26}$$

$$\text{where} \quad V_n := \frac{1}{1-\epsilon}\sup_{m\in\mathcal{M}_n}\left\{v(m) - \epsilon P\left(f_m - f^\star\right) - c\operatorname{pen}(m)\right\}$$

$$\text{and} \quad v(m) := \sqrt{\frac{2t_m}{n}\operatorname{var}_P(f_m - f^\star)} \ .$$

Let us make a few comments.

First, without V_n, (26) is the kind of oracle inequality we are looking for, since the leading constant is close to 1 (provided ϵ is small enough). For the sake of simplicity, assume that a margin condition (9) holds for every model $m \in \mathcal{M}_n$, with $\varphi_m(x) = h_m x^2$. We then have

$$v(m) \leq \sqrt{\frac{2t_m P\left(f_m - f^\star\right)}{h_m n}} \leq \epsilon P\left(f_m - f^\star\right) + \frac{t_m}{2\epsilon h_m n} \ ,$$

for any $\epsilon \in (0,1)$. Hence, applying (25), the first term of the right-hand side of (26) is smaller than

$$\frac{1+\epsilon}{1-\epsilon}\inf_{m\in\mathcal{M}_n}\left\{P\left(f_m - f^\star\right) + (2-c)\operatorname{pen}(m) + \frac{t_m}{2\epsilon h_m n}\right\} \ ,$$

which is the right-hand side of a margin adaptive oracle inequality (6) (at least when the penalty is itself of the right order). A similar result holds for a more general φ_m; see the proof of Theorem 1.

Once we have a penalty satisfying (25) (for instance, a local Rademacher penalty), the main difficulty for proving a strong margin adaptivity result then lies in V_n. It arises from the presence

of $(P - P_n)(f_m)$ in the ideal penalty but not in the right-hand side of the lower bound (25). This random quantity is centered, and (up to a quantity independent of m) has deviations of order $v(m)$, Bernstein's inequality being unimprovable. Then, if $v(m)$ happens to be much larger than $P(f_m - f^\star) + \mathrm{pen}(m)$, m is selected with a positive probability, whatever the quality of m for prediction. In that case, our risk is worse than the oracle by at least $v(m)$ (for any of these "bad" models). In that sense, V_n is unavoidable in (26).

As shown by Theorem 2 and 3, V_n can be much larger than the prediction risk of a margin adaptive procedure. However, V_n is not always the main term in the right-hand side of (26). Let us now describe a set of favorable situations, in which it is possible to prove that V_n is small enough.

1. Models are nested, $t_m \equiv t$, and pen satisfies the additional condition (13): see Section 3.
2. Models are nested, $t_m \equiv t$ and $v(m)$ is decreasing (or at least not increasing too much) when m increases. Indeed, when models are nested, either $\mathcal{F}_{m^\star} \subset \mathcal{F}_m$ so that $v(m) \leq \sup_{m' \geq m^\star} \{ v(m') \}$, or $\mathcal{F}_m \subset \mathcal{F}_{m^\star}$ so that $\varphi_{m^\star} \leq \varphi_m$ hence $\varphi^\star_m \leq \varphi^\star_{m^\star}$. In the second case,
$$v(m) - \epsilon P(f_m - f^\star) \leq \varphi^\star_m \left(\sqrt{\frac{2t}{\epsilon^2 n}} \right) \leq \varphi^\star_{m^\star} \left(\sqrt{\frac{2t}{\epsilon^2 n}} \right) \ .$$
As a consequence,
$$V_n \leq \max \left\{ \sup_{m' \geq m^\star} \{ v(m') \} ; \varphi^\star_{m^\star} \left(\sqrt{\frac{2t}{\epsilon^2 n}} \right) \right\} \ ,$$
which is not too large provided that $v(m)$ never increases too much. Notice that we can view assumption (13) as ensuring that the penalty compensates a possible increase of $v(m)$.
3. The oracle model prediction error does not decrease to zero faster than $n^{-1/2}$ and $t_m \equiv t$. Indeed, the straightforward upper bound $v(m) \leq \sqrt{2t_m/n}$ shows that $V_n \leq \sqrt{2t/n}$.
4. The margin condition does not depend on n and $t_m \equiv t$. Indeed, when $\varphi_m \equiv \varphi$ (or

$\inf_m \varphi_m \geq \varphi$), we have

$$V_n \leq \sup_{m \in \mathcal{M}_n} \left\{ \varphi_m^\star \left(\sqrt{\frac{2t_m}{\epsilon^2 n}} \right) \right\} \leq \varphi^\star \left(\sqrt{\frac{2t}{\epsilon^2 n}} \right) \ .$$

5. The penalty satisfies $c\,\mathrm{pen}(m) \geq v(m)$ for every $m \in \mathcal{M}_n$, which can be ensured for instance by adding $c^{-1}v(m)$ (or an estimate of it) to a penalty satisfying (25). This method is for instance the one proposed by Koltchinskii [14] (Section 5.2), and in that case (26) coincides with his Theorem 6.

Points 3 and 4 above show that the challenging situations are the ones where the margin condition indeed depends on the model, and fast rates of estimation are attainable. We prove in Section 5.2 that such situations can occur, enlightening how our Theorem 1 is an improvement on existing results and straightforward consequences of them.

On the other hand, point 5 may seem contradictory with the negative results of Section 4. The explanation is that using $v(m)$ in the penalty means that $\widehat{m}$ is not only a function of the data, but also of the unknown distribution P. Then, it cannot be considered adaptive. A more surprising consequence of this remark combined with Theorem 3 is that it is not possible to estimate $v(m)$ accurately enough uniformly over the set of all distributions P. Consider the proposal, in Section 5.1 of [14], to add

$$L\sqrt{\frac{t_m P_n(\widehat{f}_m)}{n}}$$

to the penalty, which is sufficient to give a result like (14). The point is that such a penalty is generally much too large (at least for "not too large" models), which often results in an upper bound of order $n^{-1/2}$. In the examples we have in mind (as well as in the counterexamples of Section 4), the excess risk of the oracle is much smaller, typically of order $n^{-\beta}$ for some $\beta \in (1/2; 1]$.

5.2. *The local margin conditions can be significantly tighter than the global one.* In this section, we show that there exist challenging situations, in which the margin condition holds for functions φ_m strongly depending on m.

PROPOSITION 2. *Let $\kappa \in (2; +\infty)$ and assume that $\mathcal{X}$ is infinite. Then, there is a probability distribution P on $\mathcal{X} \times \{0,1\}$ and two constants $C, L > 0$ (depending on κ only) such that the global margin condition* (8) *is satisfied with $\varphi(x) = Cx^\kappa$. In addition, for every $\epsilon > 0$, there exist $f_{0,\epsilon}, f_{1,\epsilon} \in \mathcal{F}$ such that*

$$0 < P\left(f_{1,\epsilon} - f^\star\right) = P\left(f_{0,\epsilon} - f^\star\right) \le \epsilon \tag{27}$$

$$\frac{\operatorname{var}_P\left(f_{1,\epsilon} - f^\star\right)}{\operatorname{var}_P\left(f_{0,\epsilon} - f^\star\right)} \le \epsilon \tag{28}$$

$$P\left(f_{0,\epsilon} - f^\star\right) \ge \varphi\left(\sqrt{\operatorname{var}_P\left(f_{0,\epsilon} - f^\star\right)}\right) \ge LP\left(f_{0,\epsilon} - f^\star\right) \ . \tag{29}$$

This result means that the global margin condition (8) is tight at any scale (29), but not uniformly over $\mathcal{F}$ (28). In particular, for every $\epsilon \in (0,1)$, we have

$$\begin{aligned}\varphi\left(\sqrt{\operatorname{var}_P\left(f_{1,\epsilon} - f^\star\right)}\right) \le \varphi\left(\sqrt{\epsilon \operatorname{var}_P\left(f_{0,\epsilon} - f^\star\right)}\right) &\le \sqrt{\epsilon}\varphi\left(\sqrt{\operatorname{var}_P\left(f_{0,\epsilon} - f^\star\right)}\right) \\ &\le \sqrt{\epsilon}P\left(f_{0,\epsilon} - f^\star\right) = \sqrt{\epsilon}P\left(f_{1,\epsilon} - f^\star\right) \ .\end{aligned}$$

Defining $\mathcal{F}_1 := \{f_{1,\epsilon}\}$ and $\mathcal{F}_2 := \{f_{0,\epsilon}, f_{1,\epsilon}\}$, the local margin condition (9) is then satisfied with $\varphi_1 \ge \epsilon^{-1/2}\varphi$ and $\varphi_2 \le L^{-1}\varphi$. With the particular formula for φ given in Proposition 2 (which is not improvable in general), we have

$$\forall x > 0, \quad \varphi^\star(x) = \frac{\kappa - 1}{\kappa^{\kappa/(\kappa-1)} C^{1/(\kappa-1)}} x^{\kappa/(\kappa-1)} \quad \text{and} \quad \varphi_1^\star(x) \le \epsilon^{1/(2(\kappa-1))} L^{-1/(\kappa-1)} \varphi_2^\star(x) \ .$$

Hence, an oracle inequality with remainder terms of order $\varphi_m^\star(t_m/n)$ is much better than an oracle inequality with remainder terms of order $\varphi^\star(t_m/n)$.

6. Discussion.

6.1. *Large collection of models.* As noticed in Remark 3, Theorem 1 may not be meaningful when $\mathcal{M}_n$ is very large, *i.e.* when $\ln(\mathrm{Card}(\mathcal{M}_n))$ is much larger than $\ln(n)$. Can a similar result be obtained in this case by replacing t by some t_m, a nondecreasing function of m? This would be interesting for instance in the ordered variable selection problem when the number of variables is much larger than the sample size n.

6.2. *Other penalization procedures.* Throughout this paper, we have focused on penalties defined in terms of local Rademacher complexities. However, most of our results are stated for general penalization procedures, the assumption (12) meaning essentially that the penalty is larger than the expectation of the ideal one with a large probability.

For instance, it is natural to think of estimating $\mathrm{pen}_{\mathrm{id}}(m)$ itself by resampling, instead of the local complexity $\overline{\delta}_n(\mathcal{F}_m; t)$. Such penalties, with several kinds of resampling schemes, have been proposed by Arlot [3, 2] and called "Resampling Penalties" (RP), generalizing the bootstrap penalty suggested by Efron [13]. Compared to the local Rademacher complexities, RP have two main interests. First, they can be computed much faster, because they are not defined as a fixed point of the resampling estimate of a function. In particular, the V-fold penalties defined in [3] have the same computational cost as V-fold cross-validation. Second, RP mainly depend on a single tuning parameter, which is a multiplicative factor in front of it. Hence, it is much easier to calibrate them than the local Rademacher complexities, which depend on two more constants, whose theoretical values are certainly too large for practical application. See for instance [4] for a completely data-driven calibration procedure for the multiplicative factor in front of a penalty.

It would therefore be interesting to prove that RP satisfy the assumptions of Theorem 1 (and are not too large), in order to obtain a margin adaptive penalization procedure with a much smaller computation time. Unfortunately, this is still a seemingly hard open problem, for which partial results can be found in Chapter 7 of [1], together with an agenda for a complete proof. In particular, controlling the expectation of RP is quite hard compared to local Rademacher

complexities, since we can no longer use a symmetrization argument.

6.3. *Should we make collections of models nested?.* A natural question coming from our results is whether one should make any collection of models nested before performing model selection, in order to improve performance. Let us consider the counterexamples of Section 4 and look at what would happen if we made the models nested.

Assume that $P = P_1$ is the distribution defined in the proof of Theorem 2. Then, comparing $\mathcal{F}_0$ and $\mathcal{F}_0 \cup \mathcal{F}_1$, the model selection problem would be easy because the margin parameter h_m would be the same in both models, making the remainder term of order $n^{-1/2}$ (the remainder term $(nh_m)^{-1}$ can be replaced by $n^{-1/2}$ when $h_m \leq n^{-1/2}$ because of the upper bound $\mathrm{var}_P(f_m - f^\star) \leq 1/4$). And margin adaptivity is not challenging when the margin condition is merely not satisfied. On the other hand, when $P = P_1$, comparing $\mathcal{F}_1$ and $\mathcal{F}_0 \cup \mathcal{F}_1$ is more challenging because $\mathcal{F}_1$ is really better than $\mathcal{F}_0$. Here, contrary to the non-nested case, the large increase of the term $\mathrm{var}_P(f_m - f^\star)$ induces a similar increase in the $L^2(P_1)$ diameter of the class. Hence, local Rademacher complexities can detect it, as shown by Theorem 1.

This shows that the final performance strongly depends on how we make the models nested. Without the knowledge of which model is indeed less complex with respect to the distribution P, we can not make the right choice with a probability going to 1 when n goes to infinity.

7. Proofs.

7.1. *Oracle inequalities.* We give the proofs in a logical order, that is first Theorem 4, then Theorem 1 (which is a corollary of it), and finally Corollary 1.

PROOF OF THEOREM 4. First, by definition of $\widehat{m}$, for every $m \in \mathcal{M}_n$ we have

$$P_n\left(\widehat{f}_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) \leq P_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \ ,$$

which can be rewritten as

$$\begin{aligned} &P\left(\widehat{f}_{\widehat{m}} - f^\star\right) + (P_n - P)\left(\widehat{f}_{\widehat{m}} - f_{\widehat{m}}\right) + (P_n - P)\left(f_{\widehat{m}} - f^\star\right) + \operatorname{pen}(\widehat{m}) \\ &\leq P\left(\widehat{f}_m - f^\star\right) + (P_n - P)\left(\widehat{f}_m - f_m\right) + (P_n - P)\left(f_m - f^\star\right) + \operatorname{pen}(m) \enspace . \end{aligned}$$

On the event where (25) holds, we then have

$$\begin{aligned} &P\left(\widehat{f}_{\widehat{m}} - f^\star\right) + (P_n - P)\left(f_{\widehat{m}} - f^\star\right) + c\operatorname{pen}(\widehat{m}) + \frac{t_{\widehat{m}}}{n} \\ &\leq \inf_{m \in \mathcal{M}_n} \left\{ P\left(\widehat{f}_m - f^\star\right) + (P_n - P)\left(f_m - f^\star\right) + \operatorname{pen}(m) \right\} \enspace . \end{aligned} \tag{30}$$

By Bernstein's inequality (see for instance Proposition 2.9 in [19]), for every $m \in \mathcal{M}_n$, there is an event of probability $1 - e^{-t_m}$ on which

$$|(P_n - P)\left(f_m - f^\star\right)| \leq v(m) + \frac{2t_m}{3n} \enspace .$$

On the intersection of these events with the one on which (25) holds, we derive from (30) that

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) - v(\widehat{m}) + c\operatorname{pen}(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ P\left(\widehat{f}_m - f^\star\right) + \operatorname{pen}(m) + v(m) + \frac{2t_m}{3n} \right\} \enspace .$$

For any $\epsilon > 0$, the left-hand side is larger than

$$\begin{aligned} &(1-\epsilon)P\left(\widehat{f}_{\widehat{m}} - f^\star\right) + \epsilon P\left(f_{\widehat{m}} - f^\star\right) + c\operatorname{pen}(\widehat{m}) - v(\widehat{m}) \\ &\geq (1-\epsilon)P\left(\widehat{f}_{\widehat{m}} - f^\star\right) - \sup_{m \in \mathcal{M}_n} \left\{ v(m) - \epsilon P\left(f_m - f^\star\right) - c\operatorname{pen}(m) \right\} \enspace . \end{aligned}$$

The result follows. □

PROOF OF THEOREM 1. We consider the event on which (26) holds. By Theorem 4, we know that it has probability at least $1 - \eta - \operatorname{Card}(\mathcal{M})e^{-t}$. We first bound the first term in the right-hand side of (26). From (9), we have

$$\forall m \in \mathcal{M}_n, \quad v(m) \leq \sqrt{\frac{2t_m}{n}} \varphi_m^{-1}\left(P\left(f_m - f^\star\right)\right) \enspace .$$

Then, using that $xy \leq \varphi_m(x) + \varphi_m^\star(y)$ for every $x, y \geq 0$,

$$v(m) \leq \varphi_m^\star\left(\sqrt{\frac{2t_m}{\epsilon^2 n}}\right) + \varphi_m\left(\epsilon\varphi_m^{-1}\left(P\left(f_m - f^\star\right)\right)\right) \enspace .$$

Since φ_m is convex with $\varphi_m(0) = 0$, we have $\varphi_m(\lambda x) \leq \lambda \varphi_m(x)$ for every $\lambda \in (0,1)$ and $x \geq 0$. Then,

$$v(m) \leq \varphi_m^\star\left(\sqrt{\frac{2t_m}{\epsilon^2 n}}\right) + \epsilon P\left(f_m - f^\star\right) \quad , \tag{31}$$

and using once more (25), the right-hand side of (26) is smaller than

$$\frac{1}{1-\epsilon} \inf_{m \in \mathcal{M}_n} \left\{ (1+\epsilon) P\left(f_m - f^\star\right) + (2-c)\,\mathrm{pen}(m) + \varphi_m^\star\left(\sqrt{\frac{2t}{\epsilon^2 n}}\right) + \frac{2t}{3n} \right\} + V_n \ . \tag{32}$$

It now remains to upperbound V_n.

Let $m^\star$ be any model which realizes the infimum in (32). For any $m \in \mathcal{M}_n$, there can be two situations:

1. $\mathcal{F}_m \subset \mathcal{F}_{m^\star}$, which implies $\varphi_m \geq \varphi_{m^\star}$, hence their conjugates satisfy $\varphi_m^\star \leq \varphi_{m^\star}^\star$. Using again (31), we have
$$v(m) \leq \varphi_m^\star\left(\sqrt{\frac{2t}{\epsilon^2 n}}\right) + \epsilon P\left(f_m - f^\star\right) \leq \varphi_{m^\star}^\star\left(\sqrt{\frac{2t}{\epsilon^2 n}}\right) + \epsilon P\left(f_m - f^\star\right) \ .$$
2. $\mathcal{F}_{m^\star} \subset \mathcal{F}_m$. Using (13) (with $m' = m^\star$), this implies that
$$v(m) \leq C_1 v(m^\star) + C_2 P(f_{m^\star} - f^\star) + c\,\mathrm{pen}(m) \ .$$

Combining those two upper bounds on $v(m)$, we get that V_n is of the same order as the first term of (32). The result follows. □

PROOF OF COROLLARY 1. From [14] (Theorem 1 and (9.2) in the proof of its Lemma 2), we know that there exist numerical constants $\overline{K} > 0$ and $q > 1$ such that (12) holds with $c = 5/7$ and $\eta = L \ln_q\left(\frac{n}{t}\right) \mathrm{Card}(\mathcal{M}_n) e^{-t}$.

In addition, Lemma 3 below shows that (13) holds with $C_1 = \sqrt{2}$ and $C_2 = 2/(\overline{K}q)$.

The result follows from Theorem 1. □

LEMMA 3. *Let $\mathcal{F}_{m'} \subset \mathcal{F}_m$ and $\overline{\delta}_n$ be defined by* (15)*. Then,*

$$v(m) \leq 2\overline{\delta}_n(\mathcal{F}_m; t) + \sqrt{2}v(m') + \frac{2P(f_{m'} - f^\star)}{q\overline{K}} \quad . \tag{33}$$

PROOF OF LEMMA 3. Since $\mathcal{F}_{m'} \subset \mathcal{F}_m$, $f_{m'} \in \mathcal{F}_m$ (as well as f_m), showing that

$$\begin{aligned} D_P(\mathcal{F}_m; P(f_{m'} - f_m)) &\geq P\left(f_m - f_{m'}\right) \geq \sqrt{\operatorname{var}_P\left(f_m - f_{m'}\right)} \\ &\geq \sqrt{\frac{\operatorname{var}_P\left(f_m - f^\star\right)}{2}} - \sqrt{\operatorname{var}_P\left(f_{m'} - f^\star\right)} \quad . \end{aligned} \tag{34}$$

For the last inequality, we used that $\operatorname{var}(X) \leq 2\operatorname{var}(X+Y) + 2\operatorname{var}(Y)$ for any random variables X, Y, and the inequality $\sqrt{x-y} \leq \sqrt{x} - \sqrt{y}$ for every $x \geq y \geq 0$.

First, assume that the lower bound in (34) is nonpositive. This implies

$$v(m) = \sqrt{\frac{t}{n}\operatorname{var}_P\left(f_m - f^\star\right)} \leq \sqrt{2}v\left(m'\right) \quad ,$$

so that (33) holds.

Otherwise, the assumptions of Lemma 4 hold with

$$D_0 = \sqrt{\frac{\operatorname{var}_P\left(f_m - f^\star\right)}{2}} - \sqrt{\operatorname{var}_P\left(f_{m'} - f^\star\right)} > 0 \quad \text{and} \quad \sigma_0 = P(f_{m'} - f_m) \quad .$$

We deduce from (35) that

$$\frac{v(m)}{2} - \frac{v(m')}{\sqrt{2}} \leq \overline{\delta}_n(\mathcal{F}_m; t) + \frac{P(f_{m'} - f_m)}{q\overline{K}} \leq \overline{\delta}_n(\mathcal{F}_m; t) + \frac{P(f_{m'} - f^\star)}{q\overline{K}} \quad ,$$

and (33) holds also. □

LEMMA 4. *Let $\overline{\delta}_n(\mathcal{F}_m; t)$ be defined by* (15)*. Assume that there is some $D_0, \sigma_0 > 0$ such that $D_P(\mathcal{F}_m; \sigma_0) \geq D_0$. Then, we have the following lower bound:*

$$\max\left\{\overline{\delta}_n(\mathcal{F}_m; t); \frac{\sigma_0}{q\overline{K}}\right\} \geq D_0\sqrt{\frac{t}{n}} \quad . \tag{35}$$

PROOF OF LEMMA 4. First, (35) clearly holds when $\frac{\sigma_0}{q\overline{K}} \geq D_0\sqrt{t/n}$. Otherwise, let $\sigma_1 = q\overline{K}D_0\sqrt{t/n} > \sigma_0$. From the definition of U_n, we have

$$\frac{U_n(\mathcal{F}_m;\sigma_1)}{\sigma_1} \geq \frac{\overline{K}D_P(\mathcal{F}_m;\sigma_1)}{\sigma_1}\sqrt{\frac{t}{n}} \geq \frac{\overline{K}D_0}{q\overline{K}D_0\sqrt{t/n}}\sqrt{\frac{t}{n}} = \frac{1}{q} > \frac{1}{2q} \ .$$

Then, according to the definition (15) of $\overline{\delta}_n(\mathcal{F}_m;t)$, $\overline{\delta}_n(\mathcal{F}_m;t) \geq D_0\sqrt{t/n}$ and the result follows.

□

PROOF OF PROPOSITION 2. Let $(x_j)_{j\in\mathbb{N}}$ be any infinite sequence of elements of $\mathcal{X}$, and $p \in (0,1)$ and $\lambda > 0$ to be chosen later. We define P as follows ((X,Y) denotes a pair of random variables with joint distribution P). For every $k \in \mathbb{N}$, $\mathbb{P}(X = x_{2k}) = p_k q_k$ and $\mathbb{P}(X = x_{2k+1}) = p_k(1-q_k)$, where $p_k = p^k(1-p)$ and $q_k \in [0,1]$ are to be chosen later, so that $\sum_{k\in\mathbb{N}} p_k = 1$. For every $k \in \mathbb{N}$, $\mathbb{P}(Y = 1 \,|\, X = x_{2k}) = 0$ and $\mathbb{P}(Y = 1 \,|\, X = x_{2k+1}) = (1+\delta_k)/2$ with $\delta_k = p^{k\lambda}$. As a consequence, the Bayes predictor is $s := \mathbb{1}_{\{x_{2k+1} \text{ s.t. } k\in\mathbb{N}\}}$. Moreover, defining $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$, we have for every $t > 0$

$$\mathbb{P}(|2\eta(X) - 1| \leq t) \leq \sum_{k \text{ s.t. } \delta_k \leq t} p_k(1-q_k) \leq \sum_{k \text{ s.t. } p^{k\lambda} \leq t} p^k(1-p) \leq t^{1/\lambda} \ . \tag{36}$$

According to Lemma 9 of [8], this implies the global margin condition

$$\forall f \in \mathcal{F}, \quad \mathrm{var}_P(f - f^\star) \leq P(f - f^\star)^2 \leq P(f \neq f^\star) \leq C_\lambda [P(f - f^\star)]^{1/(1+\lambda)}$$

where $C_\lambda > 0$ only depends on λ. As a consequence, (8) is satisfied with

$$\varphi(x) = C_\lambda^{-(1+\lambda)} x^{2(\lambda+1)} \ .$$

We now take $\lambda = \kappa/2 - 1$ and $C = C_\lambda^{-(1+\lambda)}$, so that the first statement of Proposition 2 holds true.

We now define, for every $k \in \mathbb{N}$,

$$t_{k,1}(x) := \begin{cases} s(x) & \text{if } x \notin I_k \\ 1 & \text{if } x \in I_k \end{cases} \qquad t_{k,0}(x) := \begin{cases} s(x) & \text{if } x \notin I_k \\ 0 & \text{if } x \in I_k \end{cases}$$

and $f_{k,0} = \gamma\left(t_{k,0};\cdot\right)$, $f_{k,1} = \gamma\left(t_{k,1};\cdot\right)$. It follows that

$$P\left(f_{k,0} - f^\star\right) = \delta_k p_k(1-q_k) \qquad P\left(f_{k,1} - f^\star\right) = p_k q_k$$
$$\operatorname{var}_P\left(f_{k,0} - f^\star\right) = p_k(1-q_k) - \left(\delta_k p_k(1-q_k)\right)^2 \qquad \operatorname{var}_P\left(f_{k,1} - f^\star\right) = p_k q_k - \left(p_k q_k\right)^2 \ .$$

As a consequence, choosing $q_k = \delta_k/(1+\delta_k)$, we have $P\left(f_{k,0} - f^\star\right) = P\left(f_{k,1} - f^\star\right)$, and (27) holds as soon as $p_k q_k \leq p^k(1-p) \leq \epsilon$, which holds when k is large enough.

Moreover,

$$\frac{\operatorname{var}_P\left(f_{k,1} - f^\star\right)}{\operatorname{var}_P\left(f_{k,0} - f^\star\right)} = \frac{q_k}{1-q_k}\frac{1-p_k q_k}{1-\delta_k^2 p_k(1-q_k)} = \delta_k \frac{1-p_k q_k}{1-\delta_k p_k q_k} \leq \delta_k = p^{k\lambda} \ ,$$

so that (28) holds when k is large enough.

We now have to check (29). When k is large enough, $q_k \leq 1/2$ so that

$$\begin{aligned}\frac{\varphi\left(\sqrt{\operatorname{var}_P\left(f_{0,\epsilon} - f^\star\right)}\right)}{P\left(f_{0,\epsilon} - f^\star\right)} &= \frac{C\left(p_k(1-q_k)(1-\delta_k q_k p_k)\right)^{\kappa/2}}{p_k q_k} \\ &= \frac{C(1+\delta_k)(1-q_k)^{\kappa/2}(1-\delta_k q_k p_k)^{\kappa/2}}{2} \geq \frac{C}{2^{\kappa+1}} \ .\end{aligned}$$

□

7.2. *Lower bounds.*

PROOF OF THEOREM 2. Let $t_0 : \mathcal{X} \mapsto \{0,1\}$ and $t_1 : \mathcal{X} \mapsto \{0,1\}$ defined by $t_0(x) \equiv 0$ and $t_1(x) \equiv 1$. Let $\gamma(t;(x,y)) := \mathbb{1}_{t(x)\neq y}$ be the 0-1 loss, and $\forall m \in \{0,1\}$, $\mathcal{F}_m = \{f_m\}$ with $f_m : (x,y) \mapsto \gamma\left(t;(x,y)\right)$. Since these models are singletons, we have $f_m = \widehat{f}_m$ a.s. for every $m \in \{0,1\}$.

Let $\alpha, h > 0$ to be chosen later, and P the probability distribution on $\mathcal{X} \times \{0,1\}$ defined by $P(X=a) = \alpha$, $P(X=b) = 1-\alpha$, $\mathbb{P}\left(Y=1 \,|\, X=a\right) = 0$ and $\mathbb{P}\left(Y=1 \,|\, X=b\right) = \frac{1}{2} + h$. The Bayes predictor (w.r.t. the 0-1 loss) is then defined by $s(a) = 0$ and $s(b) = 1$, and its values

outside $\mathrm{supp}(X) = \{a, b\}$ do not matter. Defining $f^\star : (x, y) \mapsto \gamma(s; (x, y))$, we have

$$P(f^\star) = (1-\alpha)\left(\frac{1}{2} - h\right) \quad P(f_0 - f^\star) = 2(1-\alpha)h \quad P(f_1 - f^\star) = \alpha$$
$$\mathrm{var}_P(f_0 - f^\star) = 1 - \alpha - (2(1-\alpha)h)^2 \quad \text{and} \quad \mathrm{var}_P(f_1 - f^\star) = \alpha - \alpha^2 \ .$$

Then,

$$h_0 \geq h \quad \text{and} \quad h_1 \geq 1 \ ,$$

so that with probability one (since $\widehat{f}_m$ is deterministic),

$$\min_{m \in \{0,1\}} \left\{ P\left(\widehat{f}_m - f^\star\right) + v(m) + \frac{t_m}{n h_m} \right\} \leq 2\alpha + \sqrt{\frac{2\gamma \ln(n)\alpha}{n}} + \frac{\gamma \ln(n)}{n} \leq 3\alpha + \frac{3\gamma \ln(n)}{2n} \ . \tag{37}$$

Since both $\mathcal{F}_0$ and $\mathcal{F}_1$ are singletons, (20) means that $\widehat{m} \in \arg\min_{m \in \{0,1\}} P_n(f_m)$. Now, notice that $P_n(f_0) = 1 - P_n(f_1)$, so that we choose the model $\widehat{m} = 0$ if and only if this random variable is smaller than $n/2$. In addition, $nP_n(f_0)$ is a binomial random variable with parameters (n, p), where

$$p = \mathbb{P}(Y = 0) = \alpha + (1-\alpha)\left(\frac{1}{2} - h\right) = \frac{1}{2} + \frac{\alpha}{2} - h(1-\alpha) \ .$$

From Lemma 5 with $a = 1$, we have $\mathbb{P}(nP_n(f_0) = k) \geq C(1, b, c) n^{-1/2} > 0$ for every $n/2 - \sqrt{n} \leq k < n/2$, as soon as

$$\left| \frac{\alpha}{2} - h(1-\alpha) \right| \leq \min\left\{ \frac{b}{\sqrt{n}}, c \right\} < \frac{1}{2} \ . \tag{38}$$

Summing these probabilities, it follows that for $n \geq 1$, $\mathbb{P}(P_n(f_0) = k) \geq C(1, b, c) 2^{-1/2} =: \kappa > 0$. On the corresponding event, $\widehat{m} = 0$ so that

$$P(\widehat{f}_{\widehat{m}} - f^\star) = P(f_0 - f^\star) = 2(1-\alpha)h \ .$$

Now compare this risk with (37). If $\alpha = (2n)^{-1}$ and $h = (2n)^{-1/2}$, then for $n \geq 1$, (38) is satisfied with $b = 1$, $c = 1/4$ and

$$2(1-\alpha)h \geq \frac{\sqrt{2n}}{3(1 + \gamma \ln(n))}\left(3\alpha + \frac{3\gamma \ln(n)}{2n}\right) \ .$$

Combined with (37), this gives the first result.

In order to prove (22), we just remark that the quantity in the min is deterministic on the right-hand side of the inequality (because both models are singletons), and the quantity on the left-hand side of (21) is a.s. nonnegative. □

REMARK 5 (on the proof of Theorem 2). The example built in the proof of Theorem 2 is a typical situation where model selection is really hard. Indeed, we are comparing two singletons, so that the only information we have on the models is their empirical risks at $X = a$ and $X = b$. Most of the points we observe are for $X = b$, but they are not significantly far from pure noise, because h is too close to 0. As a matter of fact, we have a positive probability of making a mistake at $X = b$, so that we cannot hope to make the right choice for $X = b$. On the other hand, we have very few points such that $X = a$, but we are sure to make no mistake since Y is deterministic at this point. Combining these two facts, it seems impossible to choose t_1 with a probability close to 1, since this predictor fails at the "infrequent but easy" point a, and seems worse than t_0 at point b with a positive probability (even if it is not significant). Hence, any "reasonable" method should choose t_1, *i.e.* be suboptimal within a very large factor.

PROOF OF THEOREM 3. This relies on a similar argument to the one of the proof of Theorem 2. Let $\mathcal{F}_0$ and $\mathcal{F}_1$ be as in the proof of Theorem 2, P_1 be the distribution on $\mathcal{X} \times \{0,1\}$ of the proof of Theorem 2 (*i.e.* $P_1(X = a) = \alpha$, $P_1(X = b) = 1-\alpha$, $\mathbb{P}_{(X,Y)\sim P_1}(Y = 1 \mid X = a) = 0$ and $\mathbb{P}_{(X,Y)\sim P_1}(Y = 1 \mid X = b) = \frac{1}{2} + h$, where α, h are chosen at the end of the proof of Theorem 2), and P_2 be the distribution of $(X, 1 - Y)$ when $(X, Y) \sim P_1$. In other words, we exchange the roles of $\mathcal{F}_0$ and $\mathcal{F}_1$ when switching from P_1 to P_2.

Under P_1, the proof of Theorem 2 shows that for n large enough, $\widehat{m} = 0$ implies that

$$P\left(\widehat{f}_{\widehat{m}} - f^\star\right) \geq n^\beta \min_{m\in\{0,1\}} \left\{ P\left(\widehat{f}_m - f^\star\right) + v(m) + \frac{t_m}{nh_m} \right\} .$$

Similarly, under P_2, this holds as soon as $\widehat{m} = 1$. What we have to prove is that for every model

selection rule $\widehat{m}$,

$$\max\left\{\mathbb{P}_{(X_i,Y_i)_{1\le i\le n}\sim P_1^n}\left(\widehat{m}\left((X_i,Y_i)_{1\le i\le n}\right)=0\right),\mathbb{P}_{(X_i,Y_i)_{1\le i\le n}\sim P_2^n}\left(\widehat{m}\left((X_i,Y_i)_{1\le i\le n}\right)=1\right)\right\} \ge \kappa'>0 \ . \tag{39}$$

Since we have chosen $\alpha=(2n)^{-1}$, under both P_1 and P_2, for every $n\ge 1$,

$$\mathbb{P}\left(\forall i,\, X_i=b\right)=\left(1-\frac{1}{2n}\right)^n\ge e^{-1/4}>0 \ . \tag{40}$$

Conditioned on this event, $\mathrm{Card}\left\{i \text{ s.t. } Y_i=1\right\}$ is a binomial random variable with parameters (n,p_j) under the distribution P_j, $j\in\{1,2\}$, with $|p_j-1/2|\le h$. As a consequence, the choice of h in the proof of Theorem 2 ensures that for every $j\in\{1,2\}$ and $k\in\mathbb{N}\cap\left[\frac{n}{2}-\sqrt{n},\frac{n}{2}+\sqrt{n}\right]$,

$$\mathbb{P}_{(X_i,Y_i)_{1\le i\le n}\sim P_j^n}\left(\mathrm{Card}\left\{i \text{ s.t. } Y_i=1\right\}=k \,\middle|\, \forall i,\, X_i=b\right)\ge\frac{C}{\sqrt{n}}>0, \tag{41}$$

where C is numerical (it comes from Lemma 5).

We now explain which distribution has to be considered, according to the rule $\widehat{m}$, so that (39) holds true. For every $k\in\{0,\dots,n\}$, define

$$p_k:=\mathbb{P}_{(X_i,Y_i)_{1\le i\le n}\sim P^n}\left(\widehat{m}\left((X_i,Y_i)_{1\le i\le n}\right)=1 \,\middle|\, \forall i,\, X_i=b \text{ and } \mathrm{Card}\left\{i \text{ s.t. } Y_i=1\right\}=k\right) \ ,$$

where P is the uniform distribution on $\{a,b\}\times\{0,1\}$. Actually, this quantity would be the same with any distribution assigning positive probabilities to both $(b,0)$ and $(b,1)$ (*e.g.* P_1 and P_2), since the product measure does not give different weights when the ordering of the variables change (although $\widehat{m}$ is allowed to change its outcome according to the order of the variables). In addition, this definition stays valid when $\widehat{m}$ is a randomized selection rule, which proves the generalization of Theorem 3 pointed out in Remark 4. For any given $\widehat{m}$,

$$\mathrm{Card}\left\{k\in\mathbb{N}\cap\left[\frac{n}{2}-\sqrt{n},\frac{n}{2}+\sqrt{n}\right] \text{ s.t. } p_k>\frac{1}{2}\right\}$$

is either larger or smaller than $\sqrt{n}/2$. If it is larger, (40), (41) and the definition of the p_k shows that

$$\mathbb{P}_{(X_i,Y_i)_{1\leq i\leq n}\sim P_2^n}\left(\widehat{m}\left((X_i,Y_i)_{1\leq i\leq n}\right)=1\right)\geq\frac{\sqrt{n}}{2}\frac{C}{\sqrt{n}}e^{-1/4}=\kappa'>0\ ,$$

so that (39) is satisfied. In the second situation, the same holds by choosing P_1 instead of P_2, showing that (39) is satisfied also.

This proves (23), which clearly implies (24), since everything inside the expectation on the right-hand side of the inequality is deterministic, and the quantity on the left-hand side is a.s. nonnegative. □

Finally, a key tool in the proofs of Theorem 2 and 3 is the following uniform lower bound on the density of the binomial distribution w.r.t. the counting measure on $\mathbb{N}$.

LEMMA 5. *For every $n\in\mathbb{N}$ and $p\in[0,1]$, let $\mathcal{B}(n,p)$ denote the binomial distribution with parameters (n,p). For every $a,b>0$ and $c\in(0,1/2)$, there is a positive constant $C(a,b,c)$ such that*

$$\forall n\in\mathbb{N}\setminus\{0\},\quad \inf_{\substack{k\in\mathbb{N},\,\left|k-\frac{n}{2}\right|\leq\min\left\{an^{-1/2},\frac{n}{2}\right\}\\ \left|p-\frac{1}{2}\right|\leq\min\left\{bn^{-1/2},c\right\}}}\left\{\sqrt{n}\mathbb{P}_{Z\sim\mathcal{B}(n,p)}\left(Z=k\right)\right\}\geq C(a,b,c)>0\ . \tag{42}$$

PROOF OF LEMMA 5. Let n,k,p satisfy the above conditions, $Z\sim\mathcal{B}(n,p)$, and define

$$\eta:=\frac{2k}{n}-1\qquad\delta:=p-\frac{1}{2}\ .$$

The assumption on k and p becomes $|\eta|\leq\min\left\{an^{-1/2},1/2\right\}$ and $|\delta|\leq\min\left\{bn^{-1/2},c\right\}$. In addition,

$$\mathbb{P}\left(Z=k\right)=p^k(1-p)^{n-k}\binom{n}{k}=\left(\frac{1}{2}+\delta\right)^k\left(\frac{1}{2}+\delta\right)^{n-k}\frac{n!}{k!(n-k)!}\ .$$

We now use Stirling's formula:

$$\ln(n!)=n\ln(n)-n+\frac{1}{2}\ln(2\pi n)+\epsilon_n$$

for some sequence $\epsilon_n \to 0$ when $n \to +\infty$ (one has $(12n+1)^{-1} \leq \epsilon_n \leq (12n)^{-1}$). Then,

$$\begin{aligned}\ln \mathbb{P}\left(Z=k\right) &= k \ln\left(\frac{1}{2}+\delta\right) + (n-k)\ln\left(\frac{1}{2}-\delta\right) + \ln \frac{n!}{k!(n-k)!} \\ &= \frac{n}{2}\left[(1-\eta)\ln\left(\frac{1-2\delta}{1-\eta}\right) + (1+\eta)\ln\left(\frac{1+2\delta}{1+\eta}\right)\right] \\ &\quad - \frac{1}{2}\ln(n) + \frac{1}{2}\ln\left(\frac{2}{\pi}\right) - \frac{1}{2}\ln\left(1-\eta^2\right) + \epsilon_n - \epsilon_k - \epsilon_{n-k} \ .\end{aligned}$$

Define $h : (-1, +\infty) \mapsto \mathbb{R}$ by $h(x) := x^{-1}\ln(1+x) - 1$, so that

$$\forall x > -1, \quad \ln(1+x) = x\left(1+h(x)\right) \ .$$

Recall that $|h(x)| \leq 2\,|x|$ as soon as $x \geq -1/2$, by the Taylor-Lagrange formula. In particular, $\lim_{x\to 0} h(x) = 0$. We then have

$$\begin{aligned}&\ln \mathbb{P}\left(Z=k\right) \\ &= \frac{n}{2}\left[4\delta\eta - 2\eta^2 - 2\delta(1-\eta)h(-2\delta) + \eta(1-\eta)h(-\eta) + 2\delta(1+\eta)h(2\delta) - \eta(1+\eta)h(\eta)\right] \\ &\quad - \frac{1}{2}\ln(n) + \frac{1}{2}\ln\left(\frac{2}{\pi}\right) + \frac{\eta^2}{2}h(-\eta^2) + \epsilon_n - \epsilon_k - \epsilon_{n-k} \ .\end{aligned}$$

Assuming that $n \geq n_0$ such that $\max\{a,b\}\, n^{-1/2} \leq 1/2$, it follows that

$$\ln \mathbb{P}\left(Z=k\right) = -\frac{1}{2}\ln(n) + R(k,n,p)$$

with

$$R(k,n,p) \geq L\left(1 + a^2 + ab + b^2\right)$$

for some numerical positive $L > 0$, and this lower bound is uniform over $n \geq n_0$ and k, p such that the conditions of the infimum in (42) are satisfied. On the other hand,

$$\inf_{n \leq n_0,\, 1 \leq k \leq n} \left\{ \mathbb{P}_{Z\sim\mathcal{B}(n,p)}\left(Z=k\right)\right\} \geq \kappa(p) > 0$$

as soon as $p \in (0,1)$. Since $\mathbb{P}_{Z\sim\mathcal{B}(n,p)}\left(Z=k\right)$, seen as a function of p, is increasing on $(0,k/n)$ and decreasing on $(k/n,1)$, $\kappa(p)$ is uniformly larger than $\min\{\kappa(1/2-c), \kappa(1/2+c)\}$. The result follows. □

REFERENCES

[1] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[2] Sylvain Arlot. Model selection by resampling penalization, March 2008. hal-00262478.

[3] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566.

[4] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression, March 2008. arXiv:0802.0837.

[5] Jean-Yves Audibert. Classification using Gibbs estimators under complexity and margin assumptions. Preprint, Laboratoire de Probabilites et Modeles Aleatoires, 2004.

[6] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[7] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[8] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[9] Peter L. Bartlett, Shahar Mendelson, and Petra Philips. Local complexities for empirical risk minimization. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 270–284. Springer, Berlin, 2004.

[10] Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, 4(5):861–894, 2004.

[11] Gilles Blanchard and Pascal Massart. Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6):2664–2671, 2006.

[12] Luc Devroye and Gábor Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7):1011–1018, 1995.

[13] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[14] Vladimir Koltchinskii. 2004 IMS Medallion Lecture: Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6), 2006.

[15] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 2007.

[16] Gábor Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 1–56. Springer, Vienna, 2002.

[17] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.

[18] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[19] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

[20] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[21] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[22] Alexandre B. Tsybakov and Sara A. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33(3):1203–1224, 2005.

[23] Vladimir N. Vapnik. *Statistical learning theory*. John Wiley & Sons Inc., New York, 1998.

[24] Vladimir N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Sylvain Arlot
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathematiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select
E-mail: sylvain.arlot@math.u-psud.fr

Peter L. Bartlett
University of California, Berkeley
Computer Science Division and Department of Statistics
367 Evans Hall #3860
Berkeley, CA 94720-3860
USA
E-mail: bartlett@cs.berkeley.edu