Bernoulli **17**(2), 2011, 687–713
DOI: 10.3150/10-BEJ288

Margin-adaptive model selection in statistical learning

SYLVAIN ARLOT[1] and PETER L. BARTLETT[2]

[1]*CNRS, Willow Project-Team, Laboratoire d'Informatique de l'Ecole Normale Superieure (CNRS/ENS/INRIA UMR 8548), 23, avenue d'Italie, CS 81321, 75214 Paris Cedex 13, France. E-mail: sylvain.arlot@ens.fr*
[2]*Computer Science Division and Department of Statistics, University of California, Berkeley, 367 Evans Hall #3860, Berkeley, CA 94720-3860, USA. E-mail: bartlett@cs.berkeley.edu*

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 one to take into account that learning within a small model can be much easier than within a large one. Requiring this "strong margin adaptivity" makes the model selection problem more challenging. We first prove, in a 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.

Keywords: adaptivity; empirical minimization; empirical risk minimization; local Rademacher complexity; margin condition; model selection; oracle inequalities; statistical learning

1. Introduction

We consider in this paper the model selection problem in a general framework. Since our main motivation comes from the supervised binary classification setting, we focus on this framework in this introduction. Section 2 introduces the natural generalization to empirical (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, where $\mathcal{Y} = \{0, 1\}$. The 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. In other words, the 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 s of the prediction error, when it exists, is called the Bayes predictor. Define the regression function $\eta(X) = \mathbb{P}_{(X,Y)\sim P}(Y = 1|X)$. 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 from the data some predictor t minimizing the prediction error, or equivalently the excess loss $\ell(s, t) := P\gamma(t) - P\gamma(s)$.

A classical approach to the prediction 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

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

We expect that the risk of $\hat{s}_m$ is close to that of

$$s_m \in \arg\min_{t\in S_m} P\gamma(t),$$

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 $\hat{s}_m$ is more or less distant from that of s_m. For instance, when S_m has a finite Vapnik–Chervonenkis dimension V_m [26,27] and $s \in S_m$, it has been proven (see, e.g., [19]) that

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

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 uniformly over all distributions P such that $s \in S_m$, up to the numerical factor C [14].

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 [21]. If, for some $\varepsilon_0, C_0 > 0$ and $\alpha \geq 1$,

$$\forall \varepsilon \in (0,\varepsilon_0] \qquad \mathbb{P}\big(|2\eta(X)-1| \leq \varepsilon\big) \leq C_0\varepsilon^{\alpha}, \tag{1}$$

if the Bayes predictor s belongs to S_m, and if S_m is a VC-class of dimension V_m, then the prediction error of $\hat{s}_m$ is smaller than $L(C_0,\varepsilon_0,\alpha)\ln(n)(V_m/n)^{\kappa/(2\kappa-1)}$ in expectation, where $\kappa = (1+\alpha)/\alpha$ and $L(C_0,\varepsilon_0,\alpha) > 0$ only depends on C_0, ε_0 and α. Corresponding minimax lower bounds [23] and other upper bounds can be obtained under other complexity assumptions (e.g., Assumption (A2) of Tsybakov [24], involving bracketing entropy). In the extreme situation where $\alpha = +\infty$, that is, for some $h > 0$,

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

then the same result holds with $\kappa = 1$ and $L(h) \propto h^{-1}$. More precisely, as proved in [23]

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

Following the approach of Koltchinskii [16], we will consider the following generalization of the margin condition:

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

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 $\hat{s}_m$ use only that (1) implies (3) with $\varphi(x)=L(C_0,\varepsilon_0,\alpha)x^{2\kappa}$ and $\kappa=(1+\alpha)/\alpha$, and that (2) implies (3) with $\varphi(x)=hx^2$. (See, e.g., Proposition 1 in [24].)

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 strong assumption.

A less restrictive framework is the following. First, we do not assume that $s \in S_m$. 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:

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

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 global margin condition is loose. As proven in Section 5.2 (Proposition 1), such situations certainly exist.

Note that when $\varphi_m(x)=h_m x^2$, (3) and (4) can be traced back to mean–variance conditions on γ that were used in several papers for deriving convergence rates of some minimum contrast estimators on some given model S_m (see, e.g., [11] and references therein).

1.2. Adaptive model selection

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

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

where the leading constant $C \geq 1$ should be close to one and the remainder term $R_{m,n}$ should be close to $P\gamma(\hat{s}_m)-P\gamma(s_m)$. 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 of [23], one cannot hope in general

to prove an oracle inequality (5) with a remainder $R_{m,n}$ smaller than

$$\min\left\{\frac{\ln(n)V_m}{nh_m}, \sqrt{\frac{V_m}{n}}\right\},$$

where the $\ln(n)$ term may only be necessary for some VC classes S_m (see [23]).

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 (e.g., its VC-dimension, or the ρ appearing in Tsybakov's assumption [24]). Then, define $R_n(C_m, \varphi_m)$ as the minimax prediction error over the set of distributions P such that $s \in S_m$ and the local margin condition (4) is satisfied in S_m with φ_m, where S_m has a complexity at most C_m. Massart and Nédélec [23] have proven tight upper and lower bounds on $R_n(C_m, \varphi_m)$ with several complexity measures; their results are stated with the margin condition (3), but they actually use its local version (4) only.

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

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

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

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

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* [8,10,16,20], 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 wealth of literature on margin adaptivity in the context of model selection as well as model aggregation. Most of the papers consider the uniform margin condition,

that is, when $\varphi_m \equiv \varphi$. Barron, Birgé and Massart [7] have proven oracle inequalities for deterministic penalties under some mean–variance condition on γ close to (3) with $\varphi(x) = hx^2$. Following a similar approach, margin adaptive oracle inequalities (with more general φ) have been proven with localized random penalties [8,10,16,20] and with other penalties in a particular framework [25].

Adaptivity to the margin has also been considered with a regularized boosting method [12], the hold-out [13] and in a PAC-Bayes framework [5]. Aggregation methods have been studied in [24] and [17]. 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; 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 [16]. 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 1 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 that 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 at understanding 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 [16]), 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 [16]), 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. A family of models exists such that, for every sample size n and every (model) selection procedure $\widehat{m}$, a distribution P exists for which $\widehat{m}$ fails to be strongly margin adaptive with a positive probability (Theorem 2). Hence, the previous positive results (Theorem 1 and Corollary 1) cannot 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 Section 1.4. We sketch in Section 5 some situations where strong margin adaptivity is possible. More precisely,

we state a general oracle inequality (Theorem 3), 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 minimization framework

Although our main motivation comes from the classification problem, it turns out that all our results can be proven in the general setting of empirical 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 its expectation $P(f) := \mathbb{E}_{\xi \sim P}[f(\xi)]$ 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 $\Xi = \mathcal{X} \times \mathcal{Y}$, $\xi_i = (X_i, Y_i)$,

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

where $\gamma : S \times \Xi \mapsto [0,1]$ is any contrast function. Then, $f^\star$ is equal to $\gamma(s;\cdot)$, where s is the Bayes predictor. In the binary classification framework, $\mathcal{Y} = \{0,1\}$ and we can take the 0–1 contrast $\gamma(t;(x,y)) = \mathbb{1}_{t(x) \neq y}$, for instance. We then recover the setting described in Section 1. In the bounded regression framework, assuming that $\mathcal{Y} = [0,1]$, we can take the least-squares contrast,

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

Many other contrast functions γ can be considered, provided that they take their values in $[0,1]$. Notice the one-to-one correspondence between predictors t and functions $\overline{f}_t := \gamma(t;\cdot)$ in the prediction framework.

The *empirical 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 expect that its expectation $P(\widehat{f}_m)$ 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 framework, defining $\mathcal{F}_m := \{\overline{f}_t \text{ s.t. } t \in S_m\}$, we have $\widehat{f}_m = \overline{f}_{\widehat{s}_m}$ and $f_m = \overline{f}_{s_m}$.

We can now write the global margin condition as follows:

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

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

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

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, e.g., [23] and [16]). Assume that $\mathcal{F}_m$ is a VC-type class of dimension V_m. If $\varphi_m(x) = h_m x^2$,

$$\mathbb{E}[P(\widehat{f}_m - f^\star)] \leq 2P(f_m - f^\star) + C \min\left\{\left(\frac{\ln(n) V_m}{n h_m}\right), \sqrt{\frac{V_m}{n}}\right\}$$

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

$$\mathbb{E}[P(\widehat{f}_m - f^\star)] \leq 2P(f_m - f^\star) + C \min\left\{\left[L(h_m, \kappa_m) \ln(n) \left(\frac{V_m}{n h_m}\right)^{\kappa_m/(2\kappa_m - 1)}\right], \sqrt{\frac{V_m}{n}}\right\}$$

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

Given a collection $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$ 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

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

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 numerical constant C, and $R_{m,n}$ of the order of the minimax risk $R_n(C_m, \varphi_m)$, where C_m is some complexity measure of $\mathcal{F}_m$.

Defining penalization methods as

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

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

3. Margin-adaptive model selection for nested models

3.1. General result

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

Theorem 1. *Fix $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$ and $(\varphi_m)_{m\in\mathcal{M}_n}$ such that the local margin conditions (9) hold. Let $(t_m)_{m\in\mathcal{M}_n}$ be a sequence of positive reals that is non-decreasing (with respect to the inclusion ordering on $\mathcal{F}_m$). Assume that some constants $c, \eta \in (0,1)$ and $C_1, C_2 \geq 0$ exist such that the following holds*:

- *The models $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$ are nested.*
- *Lower bounds on the penalty*: *with probability at least $1-\eta$, for every $m, m' \in \mathcal{M}_n$,*

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

$$\begin{aligned} \mathcal{F}_{m'} \subset \mathcal{F}_m \Rightarrow c\operatorname{pen}(m) &\geq v(m) - C_1 v(m') - C_2 P(f_{m'} - f^\star), \\ \text{where } v(m) &:= \sqrt{\frac{2t_m}{n}\operatorname{var}_P(f_m - f^\star)}. \end{aligned} \tag{13}$$

Then, if $\widehat{m}$ is defined by (11), with probability at least $1-\eta-2\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-t_m}$, we have for every $\varepsilon \in (0,1)$

$$\begin{aligned} P(\widehat{f}_{\widehat{m}} - f^\star) \leq \frac{1}{1-\varepsilon} \inf_{m\in\mathcal{M}_n} \Bigg\{ & (1+\varepsilon+C_2+\varepsilon C_1) P(f_m - f^\star) + \operatorname{pen}(m) \\ & + (1+\max\{1, C_1\}) \min\left\{ \varphi_m^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right), \sqrt{\frac{2t_m}{n}} \right\} + \frac{t_m}{3n} \Bigg\}, \end{aligned} \tag{14}$$

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

Theorem 1 is proved in Section 7.1.

Remark 1.

1. If $\operatorname{pen}(m)$ is of the right order, that is, not much larger than $\mathbb{E}[\operatorname{pen}_{\mathrm{id}}(m)]$, then Theorem 1 is a strong margin adaptivity result. Indeed, assuming that $\varphi_m(x) = (h_m x^2)^{\kappa_m}$, the remainder term is not too large, since

$$\varphi_m^\star(x) = L(h_m, \kappa_m) x^{2\kappa_m/(2\kappa_m - 1)}$$

for some positive constant $L(h_m, \kappa_m)$. Hence, choosing $\varepsilon = 1/2$, for instance, we can rewrite (14) as

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

for some positive constants $L(C_1, C_2)$ 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. Nevertheless, 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 [16]). On the contrary, the φ_m are not assumed to be known in Theorem 1, and conditions (12) and (13) are satisfied by completely data-dependent penalties, as shown in Section 3.2. Also, Theorem 7 of [16] shows that adaptivity is possible using a comparison method, provided that $f^\star$ belongs to one of the models. However, it is not clear whether this comparison method achieves the optimal bias–variance trade-off in the general case, as in Theorem 1.

3.2. Local Rademacher complexities

Although Theorem 1 applies to any penalization procedure satisfying assumptions (12) and (13), we now focus on methods based on local Rademacher complexities. Let us define precisely these complexities. We mainly use the notation of [16]:

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

$$\mathcal{F}_{m,P}(\delta) := \Big\{ f \in \mathcal{F}_m \text{ s.t. } P(f) - \inf_{g\in\mathcal{F}_m} P(g) \le \delta \Big\},$$

- the $L^2(P)$ diameter of the δ minimal set of $\mathcal{F}_m$ is

$$D_P^2(\mathcal{F}_m;\delta) = \sup_{f,g\in\mathcal{F}_{m,P}(\delta)} P\big((f-g)^2\big),$$

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

$$\phi_n(\mathcal{F}_m; P; \delta) = \mathbb{E} \sup_{f,g\in\mathcal{F}_{m,P}(\delta)} |(P_n - P)(f-g)|.$$

We then define

$$U_n(\mathcal{F}_m;\delta;t) := \overline{K}\left(\phi_n(\mathcal{F}_m; P; \delta) + D_P(\mathcal{F}_m;\delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\right),$$

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,

$$\overline{\delta}_n(\mathcal{F}_m;t) := \inf\left\{ \delta > 0 \text{ s.t. } \sup_{\sigma\ge\delta}\left\{ \frac{U_n(\mathcal{F}_m;\sigma;t)}{\sigma} \right\} \le \frac{1}{2q} \right\}, \tag{15}$$

where $q > 1$ is a numerical constant.

Two important points, which follow from Theorems 1 and 3 of Koltchinskii [16], are that:

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

2. There is a completely data-dependent $\hat{\delta}_n(\mathcal{F}_m; t)$ such that

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

This data-dependent $\hat{\delta}_n(\mathcal{F}_m; t)$ 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 $\hat{\delta}_n(\mathcal{F}_m; t)$, as in [16]. We need the following additional notation:

- for every $\delta > 0$, the empirical δ minimal set of $\mathcal{F}_m$ is

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

- the empirical $L^2(P)$ diameter of the empirical δ minimal set of $\mathcal{F}_m$ is

$$\widehat{D}_n(\mathcal{F}_m; \delta) = \sup_{f,g \in \widehat{\mathcal{F}}_{n,m}(\delta)} P_n\big((f-g)^2\big),$$

- the modulus of continuity of the Rademacher process $f \mapsto n^{-1}\sum_{i=1}^n \varepsilon_i f(\xi_i)$ over $\mathcal{F}_m$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables (i.e., ε_i takes the values $+1$ and -1 with probability $1/2$ each):

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

Defining

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

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

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

where $q > 1$ is a numerical constant.

Corollary 1 (Strong margin adaptivity for local Rademacher complexities). *There exist numerical constants $\overline{K} > 0$ and $q > 1$ such that the following holds. Let $t > 0$. Assume that a numerical constant $L > 0$ exists and an event of probability at least $1 - L\log_q(n/t)\,\mathrm{Card}(\mathcal{M}_n)\mathrm{e}^{-t}$ exists on which*

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

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*

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

Then, an event of probability at least $1 - [2 + (L+1)\log_q(\frac{n}{t})]\operatorname{Card}(\mathcal{M}_n)\mathrm{e}^{-t}$ *exists on which, for every* $\varepsilon \in (0,1)$,

$$P(\widehat{f}_{\widehat{m}} - f^\star) \le \frac{1}{1-\varepsilon}\inf_{m\in\mathcal{M}_n}\left\{\left(1 + \frac{2}{\overline{K}q} + \varepsilon\big(1+\sqrt{2}\big)\right)P(f_m - f^\star) + \operatorname{pen}(m) + \big(1+\sqrt{2}\big)\min\left\{\varphi_m^\star\left(\sqrt{\frac{2t}{\varepsilon^2 n}}\right), \sqrt{\frac{2t}{n}}\right\} + \frac{t}{3n}\right\}. \tag{18}$$

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

Corollary 1 is proved in Section 7.1.

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 $\operatorname{pen}(m)$ or $\hat{\delta}_n(\mathcal{F}_m;t)$). 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 (18) 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) that 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 and Corollary 1.

The reason why Corollary 1 implies strong 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 $K_1 U_n(\mathcal{F}_m;K_2\delta;t)$ for some numerical constants $K_1, K_2 > 0$, related to $\widehat{K}$ and $\hat{c}$) such that

$$\forall m \in \mathcal{M}_n \qquad \mathbb{P}\big(\widetilde{\delta}_n(\mathcal{F}_m;t) \ge \hat{\delta}_n(\mathcal{F}_m;t) \ge \overline{\delta}_n(\mathcal{F}_m;t)\big) \ge 1 - 5\log_q\left(\frac{n}{t}\right)\mathrm{e}^{-t}.$$

(See Theorem 3 of [16].) This leads to several upper bounds on $\hat{\delta}_n(\mathcal{F}_m;t)$ under the local margin condition (9), by combining Lemma 5 of [16] 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 $\varphi_m(x) = h_m x^2$, we have for every $t > 0$ and $\varepsilon \in (0, 1]$,

$$\overline{\delta}_n(\mathcal{F}_m; t) \leq \varepsilon P(f_m - f^\star) + \frac{K_3}{nh_m}\left[\varepsilon^{-1} t + \varepsilon^{-2} V_m \ln\left(\frac{n\varepsilon^2 h_m}{K_4 V_m}\right)\right], \tag{19}$$

where K_3 and K_4 depend only on $\overline{K}$. (Similar upper bounds hold under several other complexity assumptions on the models $\mathcal{F}_m$, see [16].) In particular, when each model S_m is a VC-class of dimension V_m, $\varphi_m(x) = h_m x^2$, $\mathrm{pen}(m) = \frac{7}{2}\hat{\delta}_n(\mathcal{F}_m; t)$ and $t = \ln(\mathrm{Card}(\mathcal{M}_n)) + 3\ln(n)$, (18) implies that

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

with probability at least $1 - Kn^{-2}$, for some numerical constants $C, K > 0$. Up to some $\ln(n)$ factor, this is a strong margin-adaptive model selection result, provided that $\mathrm{Card}(\mathcal{M}_n)$ is smaller than some power of n. Notice that the $\ln(n)$ factor is sometimes necessary (as shown by [23]), meaning that this upper bound is then optimal.

4. Lower bound for some non-nested models

In this section, we investigate the assumption in Theorem 1 that the models $\mathcal{F}_m$ are nested. To this aim, let us consider the case where models are singletons $\mathcal{F}_m = \{f_m\}$. Then, any estimator $\widehat{f}_m \in \mathcal{F}_m$ is deterministic and equal to f_m, so that model selection amounts to selecting among a family $\{f_m \text{ s.t. } m \in \mathcal{M}_n\}$ of functions. Theorem 2 below shows that no selection procedure can be strongly margin-adaptive in general.

Theorem 2. *Let γ be the* 0–1 *loss and $\mathcal{F}^{0-1} := \{\gamma(u;\cdot)$ s.t. $u : \mathcal{X} \mapsto \{0,1\}$ is measurable$\}$ be the associated loss function class. If* $\mathrm{Card}(\mathcal{X}) \geq 2$, *two functions $f_0, f_1 \in \mathcal{F}^{0-1}$ and absolute constants $C_3, C_4 > 0$ exist such that the following holds. For every integer $n \geq 2$ and $\widehat{m}$ a selection procedure* (*that is, a function $(\mathcal{X} \times \mathcal{Y})^n \mapsto \mathcal{M} = \{0,1\}$*), *a distribution P exists such that*

$$\mathbb{P}\left(P(f_{\widehat{m}} - f^\star) \geq \frac{C_4\sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}}\left\{P(f_m - f^\star) + \overline{v}(m) + \frac{\ln(n)}{nh_m}\right\}\right) \geq C_3 \tag{20}$$

and

$$\mathbb{E}[P(f_{\widehat{m}} - f^\star)] \geq \frac{C_3 C_4\sqrt{n}}{\ln(n)} \min_{m \in \{0,1\}}\left\{P(f_m - f^\star) + \overline{v}(m) + \frac{\ln(n)}{nh_m}\right\}, \tag{21}$$

where $\forall m \in \{0,1\}$

$$\overline{v}(m) := \sqrt{\frac{2\ln(n)}{n}\mathrm{var}_P(f_m - f^\star)} \quad \text{and} \quad h_m := \frac{P(f_m - f^\star)}{\mathrm{var}_P(f_m - f^\star)}.$$

Theorem 2 is proved in Section 7.2. A straightforward corollary of Theorem 2 is that in the classification setting with the 0–1 loss, *strong margin-adaptive model selection is not always possible when the models are not nested*. Indeed, when $\mathcal{F}_m = \{f_m\}$ for every $m \in \mathcal{M}_n = \{0, 1\}$, (20) shows that for any model selection procedure $\widehat{m}$, some distribution P exists such that results like Theorem 1 or Corollary 1 do not hold if $t_m = \ln(n)$ for every m.

Remark 3.

1. Theorem 2 and its corollary for model selection also hold 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 f_1). Hence, aggregating models instead of selecting one does not modify the conclusion of Theorem 2.
2. The most reasonable selection procedure among two functions f_0 and f_1 (or two models $\{f_0\}$ and $\{f_1\}$) clearly is empirical minimization. The proof of Theorem 2 yields explicitly some distribution P, called P_1, such that (20) and (21) hold for empirical minimization. Note that when models are singletons, most penalization procedures coincide with empirical minimization, for instance, when $\mathrm{pen}(m)$ is proportional to the local Rademacher complexity $\hat{\delta}_n(\mathcal{F}_m; t)$, or to the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)(\widehat{f}_m - f_m)$, its expectation or some quantile of $\mathrm{pen}_{\mathrm{id}}(m)$.
3. Theorem 2 focuses on 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 could hope for learning rates of order $1/(nh_m)$ if strong margin adaptivity were possible. The meaning of Theorem 2 is then mainly that one cannot guarantee to learn at a rate better than $1/\sqrt{n}$, whereas for some model, the excess loss and $1/(nh_m)$ both are of order $1/n$.
4. The counterexample given in the proof of Theorem 2 is highly non-asymptotic, since the distribution P strongly depends on n. If P and f_0, f_1 were fixed, it is well known that empirical minimization leads to asymptotic optimality, because $(f_m)_{m \in \{0,1\}}$ is finite and fixed when n grows. This illustrates a significant difference between the asymptotic and non-asymptotic frameworks. Another example of such a difference occurs when the number of candidate functions (or models) is infinite, or grows to infinity with the sample size, see (iv) in Proposition 1 in Section 5.2.

With Theorem 1, we have proven a strong margin adaptivity result for nested models, which holds true when the penalty is built upon local Rademacher complexities. Therefore, 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 model selection procedure can satisfy an oracle inequality (10) with a leading constant smaller than $C_4\sqrt{n}/(\ln(n))$.

Looking carefully at the selection problems considered in the proof of Theorem 2, 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. When looking for a more general strong margin adaptivity result, we then must keep in mind that this is a hopeless task in such situations.

Let us finally mention a related result in a close but slightly different framework. In the classification framework, under a *global* margin condition with $\varphi(x) \propto x^{2\kappa}$ with $\kappa \geq 1$, Theorem 3 in [18] shows that for any $M_n \geq 2$, a family $(u_m)_{m \in \mathcal{M}_n}$ of M_n classifiers exists for which, for any selection procedure $\widehat{m}$, some distribution P exists such that

$$\mathbb{E}[P(f_{\widehat{m}} - f^\star)] \geq \inf_{m \in \mathcal{M}_n} \{P(f_m - f^\star)\} + C\left(\frac{\ln(M_n)}{n}\right)^{\kappa/(2\kappa-1)},$$

where $f_m = \gamma(u_m; \cdot)$ for some loss function γ. When $\widehat{m}$ is (penalized) empirical minimization, the remainder term is shown to be as large as $C\sqrt{\ln(M_n)/n}$ when the margin condition holds with $\kappa > 1$.

This result and Theorem 2 focus on different problems. In [18], the margin condition is only assumed to hold *globally*, and the focus is on the dependence of the remainder term on the cardinality M_n of $\mathcal{M}_n$. Therefore, the counterexample given in [18] implies nothing about local margin conditions for $(f_m)_{m \in \mathcal{M}_n}$. Note that using these arguments, we could probably generalize Theorem 2 to a family of $M_n \geq 2$ functions and obtain a lower bound depending on M_n as in [18].

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 3. *Let $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$ be any countable family of models, and $(t_m)_{m \in \mathcal{M}_n}$ be any sequence of positive numbers. Let $\widehat{m}$ be defined by (11) and assume that some $c \in (0,1)$ exists such that*

$$\forall m \in \mathcal{M}_n \qquad (1-c)\operatorname{pen}(m) \geq (P - P_n)(\widehat{f}_m - f_m) + \frac{t_m}{n} \geq 0 \tag{22}$$

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

Then, there exists an event of probability at least $1 - \eta - 2\sum_{m \in \mathcal{M}_n} \mathrm{e}^{-t_m}$ on which the following holds: for every $\varepsilon \in (0,1)$,

$$P(\widehat{f}_{\widehat{m}} - f^\star) \leq \frac{1}{1-\varepsilon} \inf_{m \in \mathcal{M}_n} \left\{ P(f_m - f^\star) + \operatorname{pen}(m) + v(m) + \frac{t_m}{3n} \right\} + V_n, \tag{23}$$

where

$$V_n := \frac{1}{1-\varepsilon} \sup_{m \in \mathcal{M}_n} \{v(m) - \varepsilon P(f_m - f^\star) - c\operatorname{pen}(m)\}$$

and

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

Theorem 3 is proved in Section 7.1. Let us make a few comments.

First, without V_n, (23) 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 $\varphi_m(x) = h_m x^2$. Then,

$$v(m) \le \sqrt{\frac{2t_m P(f_m - f^\star)}{h_m n}} \le \varepsilon P(f_m - f^\star) + \frac{t_m}{2\varepsilon h_m n}$$

for any $\varepsilon \in (0,1)$. Hence, the first term of the right-hand side of (23) is smaller than

$$\frac{1+\varepsilon}{1-\varepsilon} \inf_{m \in \mathcal{M}_n} \left\{ P(f_m - f^\star) + \operatorname{pen}(m) + \frac{t_m}{2\varepsilon h_m n} + \frac{t_m}{3n} \right\},$$

which is the right-hand side of a margin-adaptive oracle inequality like (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 (22) (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 difference between the ideal penalty and the right-hand side of the lower bound (22), that is $(P - P_n)(f_m)$. 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 $P(f_m - f^\star) + \operatorname{pen}(m)$, m is selected with a positive probability, whatever the value of $P(\widehat{f}_m - f^\star)$. In that case, the expectation of $\widehat{f}_{\widehat{m}}$ is worse than the oracle by at least $v(m)$ (for any of these "bad" models). Hence, V_n certainly is unavoidable in (23).

As shown by Theorem 2, V_n can be much larger than the expectation of a strong margin-adaptive estimator. Nevertheless, V_n is not always the main term on the right-hand side of (23). 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 is non-decreasing (with respect to the inclusion ordering on $\mathcal{F}_m$), and pen satisfies the additional condition (13); see Section 3.
2. Models are nested, t_m is non-decreasing and $v(m)$ is decreasing (or at least not increasing too much) when $\mathcal{F}_m$ increases. Indeed, let us fix $m, m^\star \in \mathcal{M}_n$ (think of $m^\star$ as a minimizer of the infimum on the right-hand side of (23)). When models are nested, either $\mathcal{F}_{m^\star} \subset \mathcal{F}_m$ so that $v(m) \le \sup_{\mathcal{F}_{m^\star} \subset \mathcal{F}_{m'}} \{v(m')\}$, or $\mathcal{F}_m \subset \mathcal{F}_{m^\star}$ so that $\varphi_{m^\star} \le \varphi_m$ hence $\varphi_m^\star \le \varphi_{m^\star}^\star$. In the second case,

$$v(m) - \varepsilon P(f_m - f^\star) \le \varphi_m^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right) \le \varphi_{m^\star}^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right) \le \varphi_{m^\star}^\star\left(\sqrt{\frac{2t_{m^\star}}{\varepsilon^2 n}}\right)$$

since $t_m \le t_{m^\star}$ and $\varphi^\star_{m^\star}$ is non-decreasing. As a consequence, for any $m^\star \in \mathcal{M}_n$,

$$V_n \le \frac{1}{1-\varepsilon} \max\left\{ \sup_{\mathcal{F}_{m^\star} \subset \mathcal{F}_{m'}} \{v(m')\}; \varphi^\star_{m^\star}\left(\sqrt{\frac{2t_{m^\star}}{\varepsilon^2 n}}\right) \right\},$$

which is not too large provided that $v(m)$ never increases too much. Notice that we can understand 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 $t_m \le t$. Indeed, the straightforward upper bound $v(m) \le \sqrt{2t_m/n}$ shows that $V_n \le (1-\varepsilon)^{-1}\sqrt{2t/n}$.
4. The margin condition does not depend on m and $t_m \le t$. Indeed, when $\varphi_m \equiv \varphi$ (or $\inf_m \varphi_m \ge \varphi$), we have

$$V_n \le \frac{1}{1-\varepsilon} \sup_{m \in \mathcal{M}_n} \left\{ \varphi^\star_m\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right) \right\} \le \frac{1}{1-\varepsilon} \varphi^\star\left(\sqrt{\frac{2t}{\varepsilon^2 n}}\right).$$

5. The penalty satisfies $c\,\mathrm{pen}(m) \ge v(m)$ 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 (22). An example of this method is the one proposed by Koltchinskii [16] (Section 5.2), and in that case (23) 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 their straightforward consequences.

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 2 is that $v(m)$ cannot be estimated accurately enough uniformly over the set of all distributions P. Consider the proposal, in Section 5.1 of [16], to add

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

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 small 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 1. *Let $\kappa \in (1; +\infty)$ and assume that $\mathcal{X}$ is infinite. Let γ be the* 0–1 *loss and $\mathcal{F}^{0-1} := \{\gamma(u;\cdot) \text{ s.t. } u : \mathcal{X} \mapsto \{0,1\}$ is measurable$\}$ be the associated loss function class. Then there exist a probability distribution P on $\mathcal{X} \times \{0,1\}$, a sequence $(f_j)_{j\in\mathbb{N}}$ of elements of $\mathcal{F}^{0-1}$ and positive constants $(C_i)_{5\le i\le 7}$ (depending on κ only) such that*:

(i) $\forall k \in \mathbb{N}$, $P(f_{2k+1} - f^\star) = P(f_{2k} - f^\star) = b(k)$ *and* $2^{-k\kappa-2} \le b(k) \le 2^{-k\kappa-1}$.
(ii) *The global margin condition* (8) *is satisfied over $\mathcal{F} = \mathcal{F}^{0-1}$ with $\varphi(x) = C_5 x^{2\kappa}$, and it is tight*: $\forall k \in \mathbb{N}, \varphi(\sqrt{\operatorname{var}(f_{2k+1} - f^\star)}) \ge C_6 P(f_{2k+1} - f^\star)$.
(iii) *A tighter local margin condition* (9) *holds over $\{f_{2k} \text{ s.t. } k \in \mathbb{N}\} : \forall k \in \mathbb{N}$, $P(f_{2k} - f^\star) \le \operatorname{var}_P(f_{2k} - f^\star)$.*
(iv) *For every $m \in \mathbb{N}$, define $\mathcal{F}_m = \{f_m\}$ and consider the model selection problem among $(\mathcal{F}_m)_{0\le m\le M_n}$ with $M_n \ge 2\ln_2(n)$. Then, the right-hand side of a strong margin-adaptive oracle inequality of the form* (10) *is at most proportional to*

$$\inf_{0\le 2k\le M_n} \left\{ P(f_{2k} - f^\star) + \frac{\ln(n)}{n} \right\} \le \frac{2\ln(n)}{n},$$

whereas the right-hand side of a global margin-adaptive oracle inequality is larger than $C_7 n^{-\kappa/(2\kappa-1)} \gg (\ln(n))/n$.

Proposition 1 is proved in Section 7.3. It gives an example of a model selection problem where *strong margin adaptivity implies a faster rate of convergence than adaptivity to the global margin condition*. Note that the same argument works with many other model selection problems, such as selecting among $(\{f_{2k+1} \text{ s.t. } 0 \le k \le m\})_{m\in\{1,\dots,(\ln(n))^2\}}$.

6. Discussion

6.1. Other penalization procedures

We have focused in Section 3.2 on penalties defined in terms of local Rademacher complexities in order to prove that strong margin adaptivity is attainable for some data-driven penalties. An interesting question is whether such a result can be extended to penalties that can be computed faster.

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 in [2] and [3] and called "resampling penalties" (RP), generalizing the bootstrap penalty suggested by Efron [15]. Resampling penalties can be computed faster than local Rademacher complexities, because they are not defined as fixed points of the resampling estimate of a function. In particular, the V-fold penalties defined in [2] have the same computational cost as V-fold cross-validation.

In addition, RP are easy to calibrate, since they depend on a single tuning parameter – the multiplicative factor in front of it – which can, for instance, be estimated from the data by using the "slope heuristics" (see [4]). On the contrary, local Rademacher complexities depend on two more constants, whose theoretical values are certainly too large for practical application.

Extending Corollary 1 to RP would require to prove that RP satisfy both assumptions (12) and (13). On the one hand, (12) means essentially that the penalty is larger than the expectation of the ideal penalty with large probability. Hence, one can conjecture that (12) holds for RP; a partial proof of (12) for RP in our general setting can be found in Chapter 7 of [1], together with an agenda for a complete proof, which seems to be a difficult theoretical problem. On the other hand, (13) seems less likely to hold for RP, and we may have to modify RP so that (13) can be satisfied in general.

Proving such results would be quite interesting, since it would provide a strong margin-adaptive penalization procedure with a reasonably small computational cost.

6.2. 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 counterexample of Theorem 2 and look at what would happen if we make the models nested.

Assume that $P = P_1$ is the distribution defined in the proof of Theorem 2. On the one hand, comparing $\{f_0\}$ and $\{f_0, f_1\}$, the model selection problem would be easy because the margin parameter h_m is 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 $\{f_1\}$ and $\{f_0, f_1\}$ is more challenging because f_1 is really better than 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.

To conclude, improving significantly the prediction performance of the final estimator by making the models nested requires some prior knowledge, such as a natural ordering between the (non-nested) models. Otherwise, Theorem 2 shows that choosing how to make the models nested, either from data or randomly, is not successful with probability at least $C_3 > 0$, whatever the sample size.

7. Proofs

7.1. Oracle inequalities

We give the proofs in a logical order, that is, first Theorem 3, then Theorem 1 (which is a corollary of it), and finally Corollary 1.

Proof of Theorem 3. First, by definition of $\widehat{m}$, for every $m \in \mathcal{M}_n$ we have

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

which can be rewritten as

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

In the event that (22) holds, we then have

$$\begin{aligned} &P(\widehat{f}_{\widehat{m}} - f^\star) + (P_n - P)(f_{\widehat{m}} - f^\star) + c\operatorname{pen}(\widehat{m}) + \frac{t_{\widehat{m}}}{n} \\ &\quad \le \inf_{m\in\mathcal{M}_n} \{P(f_m - f^\star) + (P_n - P)(f_m - f^\star) + \operatorname{pen}(m)\}. \end{aligned} \tag{24}$$

By Bernstein's inequality (see, e.g., Proposition 2.9 in [22]), for every $m \in \mathcal{M}_n$, there is an event of probability $1 - 2\mathrm{e}^{-t_m}$ on which

$$|(P_n - P)(f_m - f^\star)| \le v(m) + \frac{t_m}{3n}.$$

On the intersection of these events with the one in which (22) holds, we derive from (24) that

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

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

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

The result follows. □

Proof of Theorem 1. We consider the event in which (23) holds. By Theorem 3, we know that it has probability at least $1 - \eta - 2\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-t_m}$. We first bound the first term on the right-hand side of (23). From (9), we have

$$\forall m \in \mathcal{M}_n \qquad v(m) \le \sqrt{\frac{2t_m}{n}}\varphi_m^{-1}\big(P(f_m - f^\star)\big).$$

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

$$\forall m \in \mathcal{M}_n \qquad v(m) \le \varphi_m^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right) + \varphi_m\big(\varepsilon\varphi_m^{-1}\big(P(f_m - f^\star)\big)\big).$$

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, using also that $\operatorname{var}_P(f_m - f^\star) \leq 1$,

$$\forall m \in \mathcal{M}_n \qquad v(m) \leq \min\left\{ \sqrt{\frac{2t_m}{n}}, \varphi_m^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right) + \varepsilon P(f_m - f^\star) \right\}, \tag{25}$$

and the right-hand side of (23) is smaller than

$$\frac{1}{1-\varepsilon} \inf_{m \in \mathcal{M}_n} \left\{ (1+\varepsilon) P(f_m - f^\star) + \operatorname{pen}(m) + \min\left\{ \varphi_m^\star\left(\sqrt{\frac{2t_m}{\varepsilon^2 n}}\right), \sqrt{\frac{2t_m}{n}} \right\} + \frac{t_m}{3n} \right\} + V_n. \tag{26}$$

It now remains to upperbound V_n.

Let $m, m' \in \mathcal{M}_n$. Since models $\mathcal{F}_m$ are nested, two cases can occur:

1. $\mathcal{F}_m \subset \mathcal{F}_{m'}$, which implies $t_m \leq t_{m'}$ and $\varphi_m \geq \varphi_{m'}$, hence $\varphi_m^\star \leq \varphi_{m'}^\star$. Using, in addition, (25) and that $\varphi_{m'}^\star$ is non-decreasing , we have

$$v(m) \leq \min\left\{ \sqrt{\frac{2t_{m'}}{n}}, \varphi_{m'}^\star\left(\sqrt{\frac{2t_{m'}}{\varepsilon^2 n}}\right) \right\} + \varepsilon P(f_m - f^\star).$$

2. $\mathcal{F}_{m'} \subset \mathcal{F}_m$. Using (13) and (25),

$$\begin{aligned} v(m) &\leq C_1 v(m') + C_2 P(f_{m'} - f^\star) + c \operatorname{pen}(m) \\ &\leq C_1 \min\left\{ \sqrt{\frac{2t_{m'}}{n}}, \varphi_{m'}^\star\left(\sqrt{\frac{2t_{m'}}{\varepsilon^2 n}}\right) \right\} + (C_2 + C_1\varepsilon) P(f_{m'} - f^\star) + c \operatorname{pen}(m). \end{aligned}$$

Therefore,

$$V_n \leq \frac{1}{1-\varepsilon} \inf_{m' \in \mathcal{M}_n} \left\{ \max\{1, C_1\} \min\left\{ \sqrt{\frac{2t_{m'}}{n}}, \varphi_{m'}^\star\left(\sqrt{\frac{2t_{m'}}{\varepsilon^2 n}}\right) \right\} + (C_2 + C_1\varepsilon) P(f_{m'} - f^\star) \right\}$$

and the result follows. □

Proof of Corollary 1. From [16] (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 $t_m = t$, $c = 5/7$ and $\eta = (L+1)\ln_q(\frac{n}{t}) \operatorname{Card}(\mathcal{M}_n) \mathrm{e}^{-t}$.

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 with $t_m = t$. □

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

$$v(m) \leq 2\overline{\delta}_n(\mathcal{F}_m; t) + \sqrt{2} v(m') + \frac{2P(f_{m'} - f^\star)}{q\overline{K}}. \tag{27}$$

Proof. Since $\mathcal{F}_{m'} \subset \mathcal{F}_m$, $f_{m'} \in \mathcal{F}_m$ (as well as f_m), so that

$$\begin{aligned} D_P\big(\mathcal{F}_m; P(f_{m'} - f_m)\big) &\geq \sqrt{P(f_m - f_{m'})^2} \geq \sqrt{\operatorname{var}_P(f_m - f_{m'})} \\ &\geq \sqrt{\frac{\operatorname{var}_P(f_m - f^\star)}{2}} - \sqrt{\operatorname{var}_P(f_{m'} - f^\star)}. \end{aligned} \tag{28}$$

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 $x, y \geq 0$.

First, assume that the lower bound in (28) is non-positive. This implies

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

so that (27) holds.

Otherwise, the assumptions of Lemma 4 below hold with

$$D_0 = \sqrt{\frac{\operatorname{var}_P(f_m - f^\star)}{2}} - \sqrt{\operatorname{var}_P(f_{m'} - f^\star)} > 0$$

and

$$\sigma_0 = P(f_{m'} - f_m).$$

We deduce from (29) that

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

and (27) also holds. □

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*:

$$\max\left\{\overline{\delta}_n(\mathcal{F}_m; t); \frac{\sigma_0}{q\overline{K}}\right\} \geq D_0 \sqrt{\frac{t}{n}}. \tag{29}$$

Proof. First, (29) clearly holds when $\frac{\sigma_0}{q\overline{K}} \geq D_0\sqrt{t/n}$. Otherwise, let $\sigma_1 = \max\{q\overline{K}, 1\}$ $D_0\sqrt{t/n} > \sigma_0$. From the definition of U_n, we have

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

Then, according to the definition (15) of $\overline{\delta}_n(\mathcal{F}_m; t)$, $\overline{\delta}_n(\mathcal{F}_m; t) \geq \sigma_1 \geq D_0\sqrt{t/n}$ and the result follows. □

7.2. Lower bounds (proof of Theorem 2)

For every $m \in \{0,1\}$, let $f_m : (x,y) \mapsto \mathbb{1}_{y \neq m}$; $f_m \in \mathcal{F}^{0-1}$, since $f_m = \gamma(u_m; \cdot)$, where for every $x \in \mathcal{X}$, $u_m(x) = m$. Let $\alpha = (2n)^{-1}$ and $h = (2n)^{-1/2}$. Let $a \neq b$ be any two elements of $\mathcal{X}$. We define a probability distribution P_1 on $\mathcal{X} \times \{0,1\}$ as follows: if $(X,Y) \sim P_1$, then $\mathbb{P}(X = a) = \alpha$, $\mathbb{P}(X = b) = 1 - \alpha$, $\mathbb{P}(Y = 1 | X = a) = 0$ and $\mathbb{P}(Y = 1 | X = b) = \frac{1}{2} + h$. We also define P_0 as the distribution of $(X, 1 - Y)$, where $(X,Y) \sim P_1$. In the following, for any distribution Q on $\mathcal{X} \times \{0,1\}$, we use the notation $\mathbb{P}_Q$ as a shortcut for $\mathbb{P}_{(X_i,Y_i)_{1 \leq i \leq n} \sim Q^{\otimes n}}$.

First, under distribution P_1, the Bayes predictor is $s = \mathbb{1}_b$,

$$P_1(f_0 - f^\star) = 2(1-\alpha)h, \qquad P_1(f_1 - f^\star) = \alpha \quad \text{and} \quad \operatorname{var}_{P_1}(f_1 - f^\star) = \alpha - \alpha^2.$$

Hence,

$$\min_{m \in \{0,1\}} \left\{ P_1(f_m - f^\star) + \overline{v}(m) + \frac{\ln(n)}{n h_m} \right\}$$
$$\leq P_1(f_1 - f^\star) + \overline{v}(1) + \frac{\ln(n)}{n h_1} \leq \alpha + \sqrt{\frac{2\alpha \ln(n)}{n}} + \frac{\ln(n)}{n} \leq \frac{2 + 3\ln(n)}{2n}.$$

Therefore, if $\mathbb{P}_{P_1}(\widehat{m} = 0) \geq C_3$, then (20) holds when $P = P_1$, with $C_4 = 1/3$. Similarly, $\mathbb{P}_{P_0}(\widehat{m} = 1) \geq C_3$ implies (20) with $P = P_0$ and $C_4 = 1/3$. So, in order to prove (20), we only need to prove that

$$\max_{j \in \{0,1\}} \{\mathbb{P}_{P_j}(\widehat{m} = 1 - j)\} \geq C_3 > 0. \tag{30}$$

The proof of (30) relies on three main facts. First,

$$\forall j \in \{0,1\} \qquad \mathbb{P}_{P_j}(\forall i, X_i = b) = (1-\alpha)^n = \left(1 - \frac{1}{2n}\right)^n \geq \frac{1}{2}. \tag{31}$$

Second, for every $j \in \{0,1\}$, under P_j, conditionally to $\{\forall i, X_i = b\}$, $\operatorname{Card}\{i \text{ s.t. } Y_i = 1\}$ is a binomial random variable with parameters (n, p_j), where

$$p_j = \mathbb{P}_{(X,Y) \sim P_j}(Y = 1) = \frac{1}{2} + (-1)^{j+1} h.$$

So, Lemma 5 shows that for every $j \in \{0,1\}$ and every $k \in \mathbb{N} \cap [\frac{n}{2} - \sqrt{n}, \frac{n}{2} + \sqrt{n}]$,

$$\mathbb{P}_{P_j}(\operatorname{Card}\{i \text{ s.t. } Y_i = 1\} = k | \forall i, X_i = b) \geq \frac{C}{\sqrt{n}} > 0, \tag{32}$$

where C is an absolute constant.

Third, let us define, for every $k \in \{0, \ldots, n\}$,

$$\pi_k := \mathbb{P}_{P_U}\left(\widehat{m}((X_i,Y_i)_{1 \leq i \leq n}) = 1 | \operatorname{Card}\{i \text{ s.t. } Y_i = 1\} = k \text{ and } \forall i, X_i = b\right),$$

where P_U is the uniform distribution on $\{a, b\} \times \{0, 1\}$. A crucial remark is that P_U can be replaced by either P_0 or P_1 in the definition of π_k, since the conditioning event determines $(X_i, Y_i)_{1 \le i \le n}$ up to the ordering of the observations; in the definition of π_k, the probability only refers to the ordering of the (X_i, Y_i), and any product measure on $\mathcal{X} \times \{0, 1\}$ assigns equal probabilities to the $n!$ permutations of the n observations. Note also that the definition of π_k stays valid when $\widehat{m}$ is a randomized selection rule, which proves the generalization of Theorem 2 pointed out in Remark 3. For any given selection rule $\widehat{m}$,

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

is either larger or smaller than $\sqrt{n}$. If it is larger, (31), (32) and the definition of the π_k (with P_0 instead of P_U) show that

$$\mathbb{P}_{P_0}\big(\widehat{m}((X_i, Y_i)_{1 \le i \le n}) = 1\big) \ge \sqrt{n} \times \frac{C}{\sqrt{n}} \times \frac{1}{2} = \frac{C}{2} = C_3 > 0,$$

so that (30) is satisfied. Otherwise, choosing P_1 instead of P_0 shows that (30) holds true. This proves (20), which clearly implies (21), since $P(f_{\widehat{m}} - f^\star) \ge 0$ a.s.

A key tool in the proof of Theorem 2 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 $c \in (0, 1/2)$, a positive constant $C(a, b, c)$ exists such that for any positive integer n,*

$$\inf_{\substack{k \in \mathbb{N}, |k - n/2| \le \min\{an^{1/2}, n/2\} \\ |p - 1/2| \le \min\{bn^{-1/2}, c\}}} \left\{\sqrt{n}\mathbb{P}_{Z \sim \mathcal{B}(n,p)}(Z = k)\right\} \ge C(a, b, c) > 0. \tag{33}$$

Proof. 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 $|\eta| \le \min\{an^{-1/2}, 1/2\}$ and $|\delta| \le \min\{bn^{-1/2}, c\}$. In addition,

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

We now use Stirling's formula:

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

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

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

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

$$\forall x > -1 \qquad \ln(1+x) = x\big(1+h(x)\big).$$

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

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

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

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

with

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

for some numerical constant $L > 0$, and this lower bound is uniform over $n \geq n_0$ and k, p such that the conditions of the infimum in (33) are satisfied. On the other hand,

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

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

7.3. Proof of Proposition 1

Let $(x_j)_{j\in\mathbb{N}}$ be any infinite sequence of distinct elements of $\mathcal{X}$ and $\lambda > 0$ to be chosen later. We define P as follows, by denoting (X,Y) 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 $p_k = 2^{-k-1}$ and $q_k \in [0, 1]$ is to be chosen later; note that $\sum_{k\in\mathbb{N}} p_k = 1$. For every $k \in \mathbb{N}$, $\mathbb{P}(Y = 1|X = x_{2k}) = 0$ and $\mathbb{P}(Y = 1|X = x_{2k+1}) = (1 + \delta_k)/2$ where $\delta_k = 2^{-k\lambda}$. As a consequence, the Bayes predictor is $s := \mathbb{1}_{\{x_{2k+1} \text{ s.t. } k\in\mathbb{N}\}}$. Let us define for every $j \in \mathbb{N}$,

$$u_j(x) := \begin{cases} s(x), & \text{if } x \neq x_j, \\ 1 - s(x), & \text{if } x = x_j \end{cases} \quad \text{and} \quad f_j = \gamma(u_j; \cdot),$$

where γ is the 0–1 loss. Then, for any $k \in \mathbb{N}$,

$$P(f_{2k+1} - f^\star) = \delta_k p_k(1 - q_k), \qquad P(f_{2k} - f^\star) = p_k q_k, \tag{34}$$

$$\operatorname{var}_P(f_{2k+1} - f^\star) = p_k(1 - q_k) - \big(\delta_k p_k(1 - q_k)\big)^2, \tag{35}$$

$$\operatorname{var}_P(f_{2k} - f^\star) = p_k q_k - (p_k q_k)^2. \tag{36}$$

We can now prove the four statements of Proposition 1.

(i) By (34), choosing $q_k = \delta_k/(1 + \delta_k)$ and $\lambda = \kappa - 1 > 0$ implies (i) with $b(k) = p_k q_k$.

(ii) For every $t \in (0, 1)$,

$$\mathbb{P}\big(|2\eta(X) - 1| \le t\big) = \sum_{k\in\mathbb{N}} \mathbb{P}(X = x_{2k+1}) \mathbb{1}_{\delta_k \le t} \le \sum_{k \text{ s.t. } 2^{-k\lambda} \le t} 2^{-k-1} \le t^{1/\lambda}. \tag{37}$$

By Lemma 9 of [9], (37) implies the global margin condition over $\mathcal{F}^{0-1}$ with function $\varphi(x) = C_5 x^{2(\lambda+1)}$, where C_5 only depends on λ. This implies the first part of (ii) since $\lambda = \kappa - 1 > 0$. For the second part, (35) implies that

$$\operatorname{var}_P(f_{2k+1} - f^\star) \ge p_k(1 - q_k)(1 - p_k) \ge \frac{p_k(1 - q_k)}{2} \ge \frac{p_k}{4} = 2^{-k-3},$$

hence the second part of (ii) holds with $C_6 = C_5 2^{2-3\kappa}$.

(iii) By (36), $\operatorname{var}_P(f_{2k} - f^\star) = p_k q_k(1 - p_k q_k) \le p_k q_k = P(f_{2k} - f^\star)$.

(iv) By (iii), for every $k \in \mathbb{N}$, a local margin condition holds on $\mathcal{F}_{2k}$ with function $\varphi_{2k} : x \mapsto x^2$. So, the right-hand side of a strong margin-adaptive oracle inequality is at most (keeping only even values of m) proportional to

$$\inf_{0\le k\le M_n/2} \left\{ P(f_{2k} - f^\star) + \frac{\ln(n)}{n} \right\} \le 2^{-\ln_2(n)-1} + \frac{\ln(n)}{n} \le \frac{2\ln(n)}{n}.$$

Note that the $\ln(n)$ factor may be replaced by a smaller quantity depending on the framework. The last statement on global margin adaptivity holds according to (ii), since $\varphi^\star(x) = L(\kappa) x^{2\kappa/(2\kappa-1)}$, where $L(\kappa) > 0$ only depends on κ.

Acknowledgements

The authors gratefully acknowledge the support of the NSF under awards DMS-0434383 and DMS-0707060. The first author's research was mostly carried out at Univ Paris-Sud (Laboratoire

de Mathematiques, CNRS – UMR 8628), with the additional support of Inria Saclay (Select Project). The authors would also like to thank an anonymous referee for numerous comments that improved the presentation and some of the results of the paper.

References

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

[2] Arlot, S. (2008). V-fold cross-validation improved: V-fold penalization. Available at arXiv:0802.0566v2.

[3] Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.* **3** 557–624 (electronic). MR2519533

[4] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** 245–279 (electronic).

[5] Audibert, J.-Y. (2004). Classification under polynomial entropy and margin assumptions and randomized estimators. Laboratoire de Probabilites et Modeles Aleatoires. Preprint.

[6] Audibert, J.-Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[7] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[8] Bartlett, P.L., Bousquet, O. and Mendelson, S. (2005). Local rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[9] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[10] Bartlett, P.L., Mendelson, S. and Philips, P. (2004). Local complexities for empirical risk minimization. In *Learning Theory*. *Lecture Notes in Comput. Sci.* **3120** 270–284. Berlin: Springer. MR2177915

[11] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** 329–375. MR1653272

[12] Blanchard, G., Lugosi, G. and Vayatis, N. (2004). On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** 861–894. MR2076000

[13] Blanchard, G. and Massart, P. (2006). Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [*Ann. Statist.* **34** (2006) 2593–2656] by V. Koltchinskii. *Ann. Statist.* **34** 2664–2671. MR2329460

[14] Devroye, L. and Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition* **28** 1011–1018.

[15] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

[16] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[17] Lecué, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721. MR2351102

[18] Lecué, G. (2007). Suboptimality of penalized empirical risk minimization in classification. In *COLT 2007 Lecture Notes in Artificial Intelligence* **4539**. Berlin: Springer. MR2397584

[19] Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of Nonparametric Learning (Udine, 2001). CISM Courses and Lectures* **434** 1–56. Vienna: Springer. MR1987656

[20] Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32** 1679–1697. MR2089138

[21] Mammen, E. and Tsybakov, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

[22] Massart, P. (2003). *Concentration inequalities and model Selection. Lecture Notes in Mathematics* **1896**. Berlin: Springer. MR2319879

[23] Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502

[24] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[25] Tsybakov, A.B. and van de Geer, S.A. (2005). Square root penalty: Adaptation to the margin in classification and in edge estimation. *Ann. Statist.* **33** 1203–1224. MR2195633

[26] Vapnik, V.N. (1998). *Statistical Learning Theory*. New York: Wiley. MR1641250

[27] Vapnik, V.N. and Cervonenkis, A.J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. (Russian. English summary) *Teor. Verojatnost. i Primenen.* **16** 264–279. MR0288823