The Slope Heuristics in Heteroscedastic Regression

Adrien Saumard

The Slope Heuristics in Heteroscedastic Regression

A. Saumard
University Rennes 1, IRMAR
adrien.saumard@univ-rennes1.fr

August 29, 2010

Abstract

We consider the estimation of a regression function with random design and heteroscedastic noise in a non-parametric setting. More precisely, we address the problem of characterizing the optimal penalty when the regression function is estimated by using a penalized least-squares model selection method. In this context, we show the existence of a minimal penalty, defined to be the maximum level of penalization under which the model selection procedure totally misbehaves. Moreover, the optimal penalty is shown to be twice the minimal one and to satisfy a nonasymptotic pathwise oracle inequality with leading constant almost one. When the shape of the optimal penalty is known, this allows to apply the so-called slope heuristics initially proposed by Birgé and Massart [14], which further provides with a data-driven calibration of penalty procedure. Finally, the use of the results obtained in [30], considering the least-squares estimation of a regression function on a fixed finite-dimensional linear model, allows us to go beyond the case of histogram models, which is already treated by Arlot and Massart in [6].

Keywords: Optimal model selection, Slope heuristics, Heteroscedastic regression, data-driven penalty.

1 Introduction

Model selection by penalization has been the object of intensive research in the last decades. Given a collection of models and associated estimators, two different tasks can be tackled: find out the smallest true model (consistency problem), or select an estimator achieving the best performance according to some criterion, called a *risk* (efficiency problem). We only focus on the efficiency problem, where the leading idea of penalization, that goes back to early works of Akaike [1], [2] and Mallows [27], is to perform an unbiased estimation of the risk of the estimators. FPE and AIC procedures proposed by Akaike respectively in [1] and [2], as well as Mallows' C_p or C_L [27], aim to do so by adding to the empirical risk a penalty which depends on the dimension of the models. But the first analysis of such procedures had the drawback to be fundamentally asymptotic, considering in particular that the number of models as well as their dimensions are fixed while the number of data tends to infinity. As explained for example in Massart [28], various statistical situations require to let these quantities depend on the number of data. Pointing out the importance of Talagrand's type concentration inequalities in this nonasymptotic approach, Birgé and Massart [13], [15] and Barron, Birgé and Massart [8] have thus been able to build nonasymptotic oracle inequalities for penalization procedures that take into account the complexity of the collection of models. In an abstract risk minimization framework, which includes statistical learning problems such as classification or regression, many distribution-dependent and data-dependent penalties have been proposed, from the more general and thus less accurate global penalties, see Koltchinskii [22], Bartlett & *al.* [9], to the refined local Rademacher complexities in the case where some margin relations hold (see for instance Bartlett, Bousquet and Mendelson [10], Koltchinskii [23]). But as a prize to pay for generality, the above penalties suffer from their dependence on unknown or unrealistic constants. They are very difficult to implement and calibrate in practice and satisfy oracle inequalities with possibly huge leading constants. In the general purpose, there are other penalties such as the bootstrap penalties of Efron [19] and the resampling and V-fold penalties of Arlot [4] and [3]. These penalties are essentially resampling estimates of the difference between the empirical risk and the risk and can be used in practice since, in particular, they avoid the practical drawbacks of the local Rademacher complexities. Arlot [4], [3]

also proves sharp pathwise oracle inequalities for the resampling and V-fold penalties in the case of regression with random design and heteroscedastic noise on histograms models, and conjectures that the restriction on histograms is mainly technical and that his results can be extended to more general situations.

We address in this article the problem of optimal model selection, in a bounded heteroscedastic with random design regression setting. A penalty will be said to be optimal if it achieves a nonasymptotic oracle inequality with leading constant almost one, i.e. converging to one when the number of data tends to infinity. In the following we restrict ourselves to "small" collections of models, where the number of models is not more than polynomial in the number of data, a case where such an optimal penalty can exist. In more general settings, where the collection of models is large, one should gather the models of equal or equivalent complexity and derive an oracle inequality with respect to the infimum of the risk on the union of models with the same complexities, as explained in Birgé and Massart [14]. This would allow to consider optimal penalties for large collections of models, but this problem is anyway beyond the scope of this article. Birgé and Massart [14] have discovered in a generalized linear Gaussian model setting, that the optimal penalty is closely related to the minimal one, defined to be the maximal penalty under which the procedure totally misbehaves. They prove sharp upper and lower bounds for the minimal penalty and show that the optimal penalty is two times the minimal one, both for small and large collections of models. These facts are called by the authors *the slope heuristics*. The authors also exhibit a jump in the dimension of the selected model occurring around the value of the minimal penalty, and use it to estimate the minimal penalty from the data. Taking a penalty equal to two times the previous estimate then gives a nonasymptotic quasi-optimal data-driven model selection procedure. The algorithm proposed by Birgé and Massart [14] to estimate the minimal penalty relies on the previous knowledge of the shape of the latter, which is a known function of the dimension of the models in their setting, and thus their procedure gives a data-driven *calibration* of the minimal penalty. Considering the case of Gaussian least-squares regression with unknown variance, Baraud, Giraud and Huet [7] have also derived lower bounds for the penalty terms for small and large collection of models, as well as Castellan [18] in the case of maximum likelihood estimation of density on histograms where a lower bound on the penalty term is given only for small collections of models. Then the slope phenomenon has been extended by Arlot and Massart [6] in a bounded heteroscedastic with random design regression framework. They consider least-squares estimators on a "small" collection of histograms models. Heteroscedasticity of the noise allows them to validate the slope heuristics without assuming a particular shape of the penalty, and in particular to consider situations where the shape of the penalty is not a function of the dimension of the models. In such general cases, the authors propose to estimate the shape of the penalty by using Arlot's resampling or V-fold penalties, proved to be efficient in their regression framework by Arlot [3] and [4], in order to derive an accurate data-driven calibration of the optimal penalty. Moreover, their approach is more general than the histogram case, except for some identified technical parts of their proofs, thus providing with some quite general algebra that can be applied in other frameworks to derive sharp model selection results. The authors have also identified the minimal penalty as the mean of the empirical excess risk on each model, and the ideal penalty to be estimated as the sum of the empirical excess risk and true excess risk on each model. The slope heuristics then heavily relies on the fact that the empirical excess risk is equivalent to the true excess risk for models of reasonable dimensions. Arlot and Massart [6] conjecture that this equivalence between the empirical and true excess risk is a quite general fact in M-estimation, as well as, by rather direct consequence, the slope phenomenon for models not too badly chosen in terms of approximation properties. A general result supporting this conjecture is the high dimensional Wilks' phenomenon discovered by Boucheron and Massart [16] in the setting of bounded contrast minimization under margin conditions, where the authors derive concentrations inequalities for the true and empirical excess risk when the considered model satisfies some general condition on the moment of first order of the surpremum of the empirical process on localized slices of variance in the loss class. This assumption can be explicated under suitable covering entropy conditions on the model. Lerasle [25] proved the validity of the slope heuristics in a least-squares density estimation setting, under rather mild conditions on the considered linear models. The approach developed by Lerasle in this framework allows sharp computations and the empirical excess risk is shown by the author to be exactly equal to the true excess risk. Moreover, some improvements comparing to the technology of proofs given by Arlot and Massart [6] can be found in [25], where Lerasle considers comparison between all pairs of models, allowing in particular a more refined use of the bias of the models. Lerasle also proves in the least-squares density estimation setting the efficiency of Arlot's resampling penalties, and generalizes these results for weakly dependent data, see [26]. Arlot and Bach [5] recently consider the problem of selecting among linear estimators in non-parametric regression. Their

framework includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. In such cases, the minimal penalty is not necessarily half the optimal one, but the authors propose to estimate the unknown variance by the minimal penalty and to use it in a plug-in version of Mallows' C_L. The latter penalty is proved to be optimal by establishing a nonasymptotic oracle inequality with constant almost one.

In this article, we prove the validity of the slope heuristics in a bounded heteroscedastic with random design regression framework, by considering a "small" collection of finite-dimensional linear models, a setting that extends the case of histograms already treated by Arlot and Massart [6]. Two main assumptions must be satisfied. First, we require that the models have a uniform localized orthonormal basis structure in $L_2\left(P^X\right)$, where P^X is the law of the explicative variable X. This kind of analytical property describing the L_∞-structure of the models has already been used in a model selection framework by Birgé and Massart [13] and Barron, Birgé and Massart [8] (see also Massart [28]). Considering for example the unit cube of $\mathbb{R}^q$ and taking $P^X = \text{Leb}$ the Lebesgue measure on it, it is shown in Birgé and Massart [13] that the assumption of localized orthonormal basis are satisfied for some wavelet expansions and piecewise polynomials uniformly bounded in their degrees. It is also known, Massart [28], that in the case of histograms the property of localized basis in $L_2\left(P^X\right)$ is equivalent to the lower regularity of the considered partition with respect to P^X, an assumption required by Arlot and Massart in [6]. Moreover, we show in [30] that if P^X has a density with respect to the Lebesgue measure on the unit interval that is uniformly bounded away from zero then, assuming the lower regularity of the partition defining piecewise polynomials of uniformly bounded degrees ensures that the assumption of localized basis is satisfied for such a model. The second property that must be satisfied in our setting is that the least-squares estimators are uniformly consistent over the collection of models and converge to the orthogonal projections of the unknown regression function. Again, such a property is shown in [30] to be satisfied for suitable histograms and more general piecewise polynomial models. This allows us to recover the results of Arlot and Massart [6] with the same set of assumptions when the noise is uniformly bounded by upper and by below, and to extend it to models of piecewise polynomials uniformly bounded in their degrees. Taking advantage of the sharp estimates of the empirical and true excess risks for a fixed model given in [30], our proofs then rely on the same algebra of proofs as those given in Arlot and Massart [6].

The article is organized as follows. We describe in Section 2 the statistical framework, the slope heuristics and the subsequent data-driven algorithm of calibration of penalties. We state in Section 3 our main results and derive their proofs in the remainder of the paper.

2 Statistical framework and the slope heuristics

2.1 Penalized least-squares model selection

We assume that we have n independent observations $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ with common distribution P. The marginal law of X_i is denoted by P^X. We assume that the data satisfy the following relation

$$Y_i = s_*\left(X_i\right) + \sigma\left(X_i\right)\varepsilon_i \ , \tag{1}$$

where $s_* \in L_2\left(P^X\right)$, ε_i are i.i.d. random variables with mean 0 and variance 1 conditionally to X_i and $\sigma : \mathcal{X} \longrightarrow \mathbb{R}$ is an heteroscedastic noise level. A generic random variable of law P, independent of the sample $(\xi_1, ..., \xi_n)$, is denoted by $\xi = (X, Y)$.

Hence, s_* is the regression function of Y with respect to X, that we want to estimate. We are given a finite collection of models $\mathcal{M}_n$, with cardinality depending on the number of data n. Each model $M \in \mathcal{M}_n$ is assumed to be a finite-dimensional vector space, and we denote by D_M its linear dimension and s_M the linear projection of s_* onto M in $L^2\left(P^X\right)$. Furthermore, by setting $K : L_2\left(P^X\right) \longrightarrow L_1\left(P\right)$ the least-squares contrast, defined by

$$K\left(s\right) = (x, y) \longmapsto \left(y - s\left(x\right)\right)^2, \quad s \in L_2\left(P^X\right) \ ,$$

the regression function s_* satisfy

$$s_* = \arg \min_{s \in L_2(P^X)} PK\left(s\right)$$

and for the linear projections s_M we have

$$s_M = \arg \min_{s \in M} PK\left(s\right) \ .$$

For each model $M \in \mathcal{M}_n$, we consider a least-squares estimator $s_n(M)$, satisfying

$$\begin{aligned} s_n(M) &\in \arg\min_{s \in M} \{P_n(K(s))\} \\ &= \arg\min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(X_i))^2 \right\} \end{aligned}$$

where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ is the empirical measure built from the data. We measure the performance of the least-squares estimators by their excess risk,

$$l(s_*, s_n(M)) := P(Ks_n(M) - Ks_*) = \|s_n(M) - s_*\|_2^2$$

where $\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X\right)^{1/2}$ is the quadratic norm in $L_2\left(P^X\right)$. Moreover, we have

$$l(s_*, s_n(M)) = l(s_*, s_M) + l(s_M, s_n(M)) \ ,$$

where the quantity

$$l(s_*, s_M) := P(Ks_M - Ks_*) = \|s_M - s_*\|_2^2$$

is called the bias of the model M and $l(s_M, s_n(M)) := P(Ks_n(M) - Ks_M) \geq 0$ is the excess risk of the least-squares estimator $s_n(M)$ on M. By the Pythagorean identity, we have

$$l(s_M, s_n(M)) = \|s_n(M) - s_M\|_2^2$$

and we prove sharp bounds for the latter quantity in [30], based on the expansion of the least-squares contrast to the sum of a linear part and a quadratic part.
Given the collection of models $\mathcal{M}_n$, an oracle model M_* is defined to be

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{l(s_*, s_n(M))\} \tag{2}$$

and the associated oracle estimator $s_n(M_*)$ thus achieves the best performance in terms of excess risk among the collection $\{s_n(M); M \in \mathcal{M}_n\}$. Unfortunately, the oracle model is unknown as it depends on the unknown law P of the data, and we propose to estimate it by a model selection procedure via penalization. Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}_+$, we thus consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}(M)\} \ . \tag{3}$$

Our goal is then to find a good penalty, such that the selected model $\widehat{M}$ satisfies an oracle inequality of the form

$$l\left(s_*, s_n\left(\widehat{M}\right)\right) \leq C \times \ell(s_*, s_n(M_*)) \ ,$$

with some positive constant C as close to one as possible and with high probability, typically more than $1 - Ln^{-2}$ for some positive constant L.

2.2 The slope heuristics

Let us rewrite the definition of the oracle model M_* given in (2). As for any $M \in \mathcal{M}_n$, the excess risk $l(s_*, s_n(M)) = P(Ks_n(M)) - P(Ks_*)$ is the difference between the risk of the estimator $s_n(M)$ and the risk of the target s_*, and as $P(Ks_*)$ is a constant of the problem, it holds

$$\begin{aligned} M_* &\in \arg\min_{M \in \mathcal{M}_n} \{P(Ks_n(M))\} \\ &= \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M)) + \text{pen}_{\text{id}}(M)\} \end{aligned}$$

where for all $M \in \mathcal{M}_n$,

$$\text{pen}_{\text{id}}(M) := P(Ks_n(M)) - P_n(Ks_n(M)) \ .$$

The penalty function $\mathrm{pen}_{\mathrm{id}}$ is called the *ideal penalty*, as it allows to select the oracle, but it is unknown because it depends on the distribution of the data. As pointed out by Arlot and Massart [6], the leading idea of penalization in the efficiency problem is thus to give some sharp estimate of the ideal penalty, in order to perform an unbiased or asymptotically unbiased estimation of the risk over the collection of models, leading to a sharp oracle inequality for the selected model. A penalty term $\mathrm{pen}_{\mathrm{opt}}$ is said to be optimal if it achieves an oracle inequality with constant almost one, tending to one when the number n of data tends to infinity.
Concerning the estimation of the optimal penalty, Arlot and Massart [6] conjecture that the mean of the empirical excess risk $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ satisfies the following slope heuristics in a quite general framework:

(i) If a penalty $\mathrm{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ is such that, for all model $M \in \mathcal{M}_n$,

$$\mathrm{pen}\left(M\right) \leq \left(1-\delta\right)\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

with $\delta > 0$, then the dimension of the selected model $\widehat{M}$ is "very large" and the excess risk of the selected estimator $s_n\left(\widehat{M}\right)$ is "much larger" than the excess risk of the oracle.

(ii) If $\mathrm{pen} \approx \left(1+\delta\right)\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ with $\delta > 0$, then the corresponding model selection procedure satisfies an oracle inequality with a leading constant $C\left(\delta\right) < +\infty$ and the dimension of the selected model is "not too large". Moreover,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$$

is an optimal penalty.

The mean of the empirical excess risk on M, when M varies in $\mathcal{M}_n$, is thus conjectured to be the maximal value of penalty under which the model selection procedure totally misbehaves. It is called the *minimal penalty,* denoted by $\mathrm{pen}_{\mathrm{min}}$:

$$\text{for all } M \in \mathcal{M}_n, \qquad \mathrm{pen}_{\mathrm{min}}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \ .$$

The optimal penalty is then close to two times the minimal one,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}} \ .$$

Let us now briefly explain why points (i) and (ii) below are natural. We give in Section 3 precise results which validate the slope heuristics for models such as histograms or piecewise polynomials uniformly bounded in their degrees. If the penalty is the minimal one, then for all $M \in \mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}_{\mathrm{min}}\left(M\right) \\
&= P_n\left(Ks_n\left(M\right)\right) + \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \\
&= P\left(Ks_M\right) + \left(P_n - P\right)\left(Ks_M\right) + \left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\
&\approx P\left(Ks_M\right) \ .
\end{aligned}$$

In the above lines, we neglect $\left(P_n - P\right)\left(Ks_M\right)$ as it is a centered quantity and if the empirical excess risk $P_n\left(Ks_n\left(M\right) - Ks_M\right)$ is close enough to its expectation, then the selected model almost minimizes its bias, and so its dimension is among the largest of the models and the excess risk of the selected estimator blows up. As shown by Boucheron and Massart [16], the empirical excess risk satisfies a concentration inequality in a general framework, which allows to neglect the difference with its mean, at least for models that are not too small.
Now, if the chosen penalty is less than the minimal one, $\mathrm{pen} \approx \left(1-\delta\right)\mathrm{pen}_{\mathrm{min}}$ with $\delta \in \left(0,1\right)$, the algorithm minimizes over $\mathcal{M}_n$,

$$\begin{aligned}
&P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}\left(M\right) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M\right) \\
&\quad + \left(1-\delta\right)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\
&\approx P\left(Ks_M\right) - \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) \ ,
\end{aligned}$$

where in the last identity we neglect the deviations of the empirical excess risk and the difference between the empirical and true risk of the projections s_M. As the empirical excess risk is increasing and the risk of the projection s_M is decreasing with respect to the complexity of the models, the penalized criterion is decreasing with respect to the complexity of the models, and the selected model is again among the largest of the collection.
If on the contrary, the chosen penalty is more than the minimal one, $\text{pen} \approx (1+\delta)\,\text{pen}_{\min}$ with $\delta > 0$, then the selected model minimizes the following criterion, for all $M \in \mathcal{M}_n$,

$$\begin{aligned} & P_n\left(Ks_n\left(M\right)\right) + \text{pen}\left(M\right) - P_n\left(Ks_*\right) \\ & \approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M - Ks_*\right) \\ & \quad + (1+\delta)\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right) \\ & \approx \ell\left(s_*, s_M\right) + \delta P_n\left(Ks_M - Ks_n\left(M\right)\right) \ , \end{aligned} \tag{4}$$

So the selected model achieves a trade-off between the bias of the models which decreases with the complexity and the empirical excess risk which increases with the complexity of the models. The selected dimension will be then reasonable, and the trade-off between the bias and the complexity of the models is likely to give some oracle inequality.
Finally, if we take $\delta = 1$ in the above case, that is $\text{pen} \approx 2 \times \text{pen}_{\min}$ and if we assume that the empirical excess risk is equivalent to the excess risk,

$$P_n\left(Ks_M - Ks_n\left(M\right)\right) \sim P\left(Ks_n\left(M\right) - Ks_M\right) \ , \tag{5}$$

then according to (4) the selected model almost minimizes

$$P\left(Ks_M - Ks_*\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) \approx \ell\left(s_*, s_M\right) + P\left(Ks_n\left(M\right) - Ks_M\right) \approx \ell\left(s_*, s_n\left(M\right)\right) \ .$$

Hence,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \approx \ell\left(s_*, s_n\left(M_*\right)\right)$$

and the procedure is nearly optimal. We give in [30] some results showing that (5) is a quite general fact in least-squares regression.

2.3 A data-driven calibration of penalty algorithm

The slope heuristics stated in points (i) and (ii) in Section 2.2, include that a jump in the dimensions of the selected models should occur around the minimal penalty, which can be used to estimate the minimal penalty and by consequence, the optimal one. Let us denote by $\text{pen}_{\text{shape}}$ the shape of the minimal penalty which is, according to the slope heuristics, equal to the shape of the optimal penalty. Thus, for two unknown positive constants $A_{\min}$ and A_* depending on the unknown distribution of the data, we have

$$\text{pen}_{\min} = A_{\min}\,\text{pen}_{\text{shape}} \quad \text{and} \quad \text{pen}_{\text{opt}} = A_*\,\text{pen}_{\text{shape}} \ ,$$

where

$$A_* = 2 \times A_{\min}$$

whenever the optimal penalty is twice the minimal one. We assume now that the shape of the minimal penalty is known, from some prior knowledge or because it has been estimated from the data, for example by using Arlot's resampling and V-fold penalties as suggested in Arlot and Massart [6]. Then, Arlot and Massart [6] propose to *calibrate* the optimal penalty by the following procedure and by doing so, they extend to general penalty shapes a previous algorithm proposed by Birgé and Massart [14].

Algorithm of data-driven calibration of penalties :

1. Compute the selected model $\widehat{M}\left(A\right)$ as a function of $A > 0$,

$$\widehat{M}\left(A\right) \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n K\left(s_n\left(M\right)\right) + A\,\text{pen}_{\text{shape}}\left(M\right)\right\} \ .$$

2. Find $\hat{A}_{\min} > 0$ such that the dimension $D_{\widehat{M}(A)}$ is "very large" for $A < \hat{A}_{\min}$ and "reasonably small" for $A > \hat{A}_{\min}$.

3. Select the model $\widehat{M} = \widehat{M}\left(2\hat{A}_{\min}\right)$.

In this paper, since our aim is not to apply the above algorithm in practice, we refer to Arlot and Massart [6] for a detailed presentation of the algorithm and to Baudry, Maugis and Michel [12] for an overview on the slope heuristics and further discussions on implementation issues. Data-driven calibration of penalties algorithms have already been applied successively in many statistical frameworks such as mixture models [29], clustering [11], spatial statistics [31], estimation of oil reserves [24] and genomics [32], to name but a few. These applications tend to support the conjecture of Arlot and Massart [6] that the slope heuristics is valid in a quite general framework.

3 Main Results

We state here our results that theoretically validate the slope heuristics in our bounded heteroscedastic regression setting. In particular, we recover the results stated in Theorems 2 and 3 of Arlot and Massart [6] for histogram models and extend them to models of piecewise polynomials uniformly bounded in their degrees. The proofs are postponed to the end of the paper, and heavily rely on results obtained in [30] where we consider a fixed model, and on the general algebra of proofs developed by Arlot and Massart [6]. We state now the assumptions required to derive our results.

3.1 Main assumptions

Let us begin with the set of assumptions needed in the general case of models that are provided with localized basis in $L_2\left(P^X\right)$.

General set of assumptions : (GSA)

(P1) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}\left(\mathcal{M}_n\right) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $M \in \mathcal{M}_n$, $1 \leq D_M \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2} \leq n$.

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$.

(Ab) A positive constant A exists, that bounds the data and the projections s_M of the target s_* over the models M of the collection $\mathcal{M}_n$: $|Y_i| \leq A < \infty$, $\|s_M\|_\infty \leq A < \infty$ for all $M \in \mathcal{M}_n$.

(An) Uniform lower-bound on the noise level: $\sigma\left(X_i\right) \geq \sigma_{\min} > 0$ *a.s.*

(Ap$_u$) The bias decreases as a power of D_M: there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} .$$

(Alb) Each model is provided with a localized basis: there exists a constant $r_{\mathcal{M}}$ such that for each $M \in \mathcal{M}_n$ one can find an orthonormal basis $(\varphi_k)_{k=1}^{D_M}$ satisfying that, for all $(\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M}$,

$$\left\| \sum_{k=1}^{D_M} \beta_k \varphi_k \right\|_\infty \leq r_{\mathcal{M}} \sqrt{D_M} \left|\beta\right|_\infty ,$$

where $\left|\beta\right|_\infty = \max\left\{|\beta_k| \ ; \ k \in \{1, ..., D_M\}\right\}$.

($\mathbf{Ac}_\infty$) Consistency in sup-norm of least-squares estimators: an event Ω_∞ of probability at least $1-n^{-2-\alpha_{\mathcal{M}}}$, a positive constant A_{cons}, a positive integer n_1 and a collection of positive numbers $(R_{n,D_M})_{M\in\mathcal{M}_n}$ exist, such that

$$\sup_{M\in\mathcal{M}_n} R_{n,D_M} \leq \frac{A_{cons}}{\sqrt{\ln n}} \tag{6}$$

and for all $M \in \mathcal{M}_n$ it holds on Ω_∞, for all $n \geq n_1$,

$$\|s_n(M) - s_M\|_\infty \leq R_{n,D_M} \ . \tag{7}$$

We turn now to the set of assumptions needed for histogram models and models by piecewise polynomials, respectively.

Set of assumptions for histogram models :

Given some linear histogram model $M \in \mathcal{M}_n$, we denote by $\mathcal{P}_M$ the associated partition of $\mathcal{X}$.
Take assumptions (**P1**), (**P2**), (**P3**), (**An**) and ($\mathbf{Ap}_u$) from the general set of assumptions. Assume moreover that the following conditions hold true:

(Ab') A positive constant A exists, that bounds the data: $|Y_i| \leq A < \infty$.

(Alrh) Lower regularity of the partitions: there exists a positive constant $c^h_{\mathcal{M},P}$ such that,

$$\text{for all } M \in \mathcal{M}_n, \quad \sqrt{|\mathcal{P}_M| \inf_{I\in\mathcal{P}_M} P^X(I)} \geq c^h_{\mathcal{M},P} > 0 \ .$$

Set of assumptions for piecewise polynomials models :

In this case we take $\mathcal{X} = [0,1]$, Leb is the Lebesgue measure on $\mathcal{X}$, and given a linear model $M \in \mathcal{M}_n$ of piecewise polynomials, we denote by $\mathcal{P}_M$ the associated partition of $\mathcal{X}$.
Take assumptions (**P1**), (**P2**), (**P3**), (**An**) and ($\mathbf{Ap}_u$) from the general set of assumptions. Assume moreover that the following additional conditions hold.

(Ab') A positive constant A exists, that bounds the data: $|Y_i| \leq A < \infty$.

(Aud) Uniformly bounded degrees: there exists $r \in \mathbb{N}^*$ such that, for all $M \in \mathcal{M}_n$, all $I \in \mathcal{P}_M$ and all $p \in M$,

$$\deg\left(p_{|I}\right) \leq r \ .$$

($\mathbf{Ad}_{\text{Leb}}$) Density bounded from upper and from below: P^X has a density f with respect to Leb satisfying for some constants $c_{\min}$ and $c_{\max}$, that

$$0 < c_{\min} \leq f(x) \leq c_{\max} < \infty, \quad \forall x \in [0,1] \ .$$

(Alrpp) Lower regularity of the partition: a positive constant $c^{pp}_{\mathcal{M},P}$ exists such that, for all $M \in \mathcal{M}_n$,

$$0 < c^{pp}_{\mathcal{M},\text{Leb}} \leq \sqrt{|\mathcal{P}_M| \inf_{I\in\mathcal{P}_M} \text{Leb}(I)} < +\infty \ .$$

The sets of assumptions will be discussed in Section 3.3.

3.2 Statement of the theorems

Theorem 1 *Under the general set of assumptions (**GSA**) of Section 3.1, for $A_{\text{pen}} \in [0,1)$ and $A_p > 0$, we assume that with probability at least $1 - A_p n^{-2}$ we have*

$$0 \leq \text{pen}(M_1) \leq A_{\text{pen}} \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M_1)\right)\right] \ , \tag{8}$$

*where the model M_1 is defined in assumption (**P3**) of (**GSA**). Then there exist two positive constants A_1, A_2 independent of n such that, with probability at least $1 - A_1 n^{-2}$, we have, for all $n \geq n_0((\textbf{GSA}), A_{\text{pen}})$,*

$$D_{\widehat{M}} \geq A_2 n \ln(n)^{-2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \ln(n) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n(M)\right)\right\} \ . \tag{9}$$

Moreover, in the case of histograms and piecewise polynomials models, taking their respective set of assumptions defined in Section 3.1 yields the same results.

Thus, Theorem 1 justifies the first part (**i**) of the slope heuristics exposed in Section 2.2. As a matter of fact, it shows that there exists a level such that if the penalty is smaller than this level for one of the largest models, then the dimension of the output is among the largest dimensions of the collection and the excess risk of the selected estimator is much bigger than the excess risk of the oracle. Moreover, this level is given by the mean of the empirical excess risk of the least-squares estimator on each model.
The following theorem validates the second part of the slope heuristics.

Theorem 2 *Assume that the general set of assumptions (**GSA**) of Section 3.1 hold.*
Moreover, for some $\delta \in [0,1)$ and $A_p, A_r > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}(\ln n)^3$, it holds

$$(2-\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \leq \text{pen}(M) \leq (2+\delta)\,\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \tag{10}$$

together with

$$\text{pen}(M) \leq A_r \frac{(\ln n)^3}{n} \tag{11}$$

*for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}(\ln n)^3$. Then, for $\frac{1}{2} > \eta > \left(1 - \beta_+\right)_+ / 2$, there exist a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**GSA**) and on A_p, a positive constant A_4 only depending on constants in the set of assumptions (**GSA**), a positive constant A_5 only depending on constants in the set of assumptions (**GSA**) and on A_r and a sequence*

$$\theta_n = A_4 \sup_{M \in \mathcal{M}_n} \left\{\varepsilon_n(M) \, , \ A_{\mathcal{M},+}(\ln n)^3 \leq D_M \leq n^{\eta+1/2}\right\} \leq \frac{A_4\left(1 \vee \sqrt{A_{cons}}\right)}{(\ln n)^{1/4}} \tag{12}$$

such that with probability at least $1 - A_3 n^{-2}$, it holds for all $n \geq n_0((\textbf{GSA}), \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\left((\ln n)^{-2} + \theta_n\right)}{(1-\delta)^2}\right) \ell\left(s_*, s_n(M_*)\right) + A_5 \frac{(\ln n)^3}{n} \ . \tag{13}$$

Assume that in addition, the following assumption holds,

(Ap) *The bias decreases like a power of D_M : there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_M^{-\beta_-} \leq \ell\left(s_*, s_M\right) \leq C_+ D_M^{-\beta_+} \ .$$

Then it holds with probability at least $1 - A_3 n^{-2}$*, for all* $n \geq n_0 \left((\boldsymbol{GSA}), C_-, \beta_-, \eta, \delta \right)$,

$$A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{\eta+1/2}$$

and

$$\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2} \right) \ell \left(s_*, s_n \left(M_* \right) \right) \ . \tag{14}$$

Likewise, in the case of models of histograms and piecewise polynomials, taking their respective set of assumptions defined in Section 3.1, together with assumption (10) and, for the second part of the theorem, assumption (***Ap***)*, yields the same results.*

The quantity $\varepsilon_n (M)$ used in (12) controls the deviations of the true and empirical excess risks on the model M and is more precisely defined in Remark 3 above. From Theorems 1 and 2, we identify the minimal penalty with the mean of the empirical excess risk on each model,

$$\mathrm{pen}_{\min} (M) = \mathbb{E} \left[P_n \left(K s_M - K s_n (M) \right) \right] \ .$$

Moreover, Theorem 2 states in particular that if the penalty is close to two times the minimal procedure, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal.

3.3 Comments on the sets of assumptions

Let us now explain the sets of assumptions given in Section 3.1. Assumption (**P1**) states that the collection of models has a small complexity, more precisely a polynomially increasing one with respect to the amount of data. For this kind of complexities, if one wants to perform a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model. Indeed, as Talagrand's type inequalities for the empirical process are pre-Gaussian, they allow to neglect the deviations of the quantities of interest from their mean, uniformly over the collection of models. This is not the case for too large collections of models, where one has to put an extra-log factor depending on the complexity of the collection of models inside the penalty (see for example [13] and [8]). In assumption (**P2**) we restrict the dimensions of the models by upper, in a way that is not too restrictive since we allow the dimension to be of the order of the amount of data within a power of a logarithmic factor. We assume in (**P3**) that the collection of models contains a model M_0 of reasonably large dimension and a model M_1 of high dimension, which is necessary since we prove the existence of a jump between high and reasonably large dimensions. We demand in ($\mathbf{Ap}_u$) that the quality of approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of linear projections of the regression function onto the models. Assumptions (**Ab**), (**An**), (**Alb**) and ($\mathbf{Ac}_\infty$) essentially allow us to apply results of Section **??**, as further explained in Remark 3 below. The assumption (**Ab**) is also necessary to control in the proofs the empirical bias term centered by the true bias by using Bernstein's inequality (see Lemma 5).
Assumption (**Ab'**) implies in the histogram case assumption (**Ab**), see Section 4 of [30]. Moreover, assumption (**Alrh**) allows us in this case to deduce assumptions (**Alb**) and ($\mathbf{Ac}_\infty$) of the general set of assumptions (see Lemma 5 and 6 of [30]). Moreover, using Lemma 6, it is straightforward to see that in the histogram case we have

$$R_{n,D_M} \leq A_{cons} \sqrt{\frac{D_M \ln n}{n}} \ ,$$

where A_{cons} is a uniform positive constant over the models of $\mathcal{M}_n$. We obtain in the case of histograms the same set of assumptions as given in Arlot and Massart [6]. Arlot and Massart [6] also notice that they can weaken assumptions (**Ab'**) and (**An**), for example by assuming conditions on the moment of the noise instead of considering that this quantity is bounded in sup-norm. This latter improvement seems to be beyond the reach of our method, due to the use of Talagrand's type inequalities that require conditions in sup-norm. Arlot and Massart [6] also show that the condition ($\mathbf{Ap}_u$) is satisfied when $\mathcal{X} \subset \mathbb{R}^k$ and the regression function s_* is α-Hölderian. Moreover, they show that (**Ap**) is satisfied when in addition, s_* is non-constant with respect to the sup-norm.

As in the case of histogram models, assumption (**Ab'**) implies in the piecewise polynomial case assumption (**Ab**), see Section 5 of [30]. Assumptions (**Aud**), ($\mathbf{Ad}_{\mathrm{Leb}}$) and (**Arpp**) allow us to guaranty the statements (**Alb**) and ($\mathbf{Ac}_\infty$) of the general set of assumptions in this case (see Lemmas 8 and 9 of [30]). Moreover, we still have

$$R_{n,D_M} \propto \sqrt{\frac{D_M \ln n}{n}} \ ,$$

within a uniform constant over the models of $\mathcal{M}_n$. It is well-known that piecewise polynomials uniformly bounded in their degrees have good approximation properties in Besov spaces. More precisely, as stated in Lemma 12 of Barron, Birgé and Massart [8], if $\mathcal{X} = [0,1]$ and the regression function s_* belongs to the Besov space $B_{\alpha,p,\infty}(\mathcal{X})$ (see the definition in [8]), then taking models of piecewise polynomials of degree bounded by $r > \alpha - 1$ on regular partitions with respect to the Lebesgue measure Leb on $\mathcal{X}$, and assuming that P^X has a density with respect to Leb which is bounded in sup-norm, assumption ($\mathbf{Ap}_u$) is satisfied. It remains to find conditions in this context such that the lower bound on the bias in (**Ap**) is also satisfied.

Remark 3 *Since constants in the general set of assumptions (**GSA**) made above are uniform over the collection $\mathcal{M}_n$, we deduce from Theorem 3 of [30] applied with $\alpha = 2 + \alpha_{\mathcal{M}}$ and $A_- = A_+ = A_{\mathcal{M},+}$ that if assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and ($\mathbf{Ac}_\infty$) hold, then a positive constant A_0 exists, depending on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$ and on the constants A, $\sigma_{\min}$ and $r_{\mathcal{M}}$ defined in the general set of assumptions, such that for all $M \in \mathcal{M}_n$ satisfying*

$$0 < A_{\mathcal{M},+} (\ln n)^2 \leq D_M \ ,$$

by setting

$$\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \ \left(\frac{D_M \ln n}{n}\right)^{1/4} ; \ \sqrt{R_{n,D_M}} \right\} \tag{15}$$

we have, for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$,

$$\mathbb{P}\left[(1 - \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P(Ks_n(M) - Ks_M) \leq (1 + \varepsilon_n(M)) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 10 n^{-2-\alpha_{\mathcal{M}}} \tag{16}$$

and

$$\mathbb{P}\left[\left(1 - \varepsilon_n^2(M)\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq \left(1 + \varepsilon_n^2(M)\right) \frac{1}{4} \frac{D_M}{n} \mathcal{K}_{1,M}^2 \right] \geq 1 - 5 n^{-2-\alpha_{\mathcal{M}}} \ . \tag{17}$$

Moreover, for all $M \in \mathcal{M}_n$, we have by Theorem 4 of [30], for a positive constant A_u depending on $A, A_{cons}, r_{\mathcal{M}}$ and $\alpha_{\mathcal{M}}$ and for all $n \geq n_0(A_{cons}, n_1)$,

$$\mathbb{P}\left[P(Ks_n(M) - Ks_M) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-2-\alpha_{\mathcal{M}}} \tag{18}$$

and

$$\mathbb{P}\left[P_n(Ks_M - Ks_n(M)) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3 n^{-2-\alpha_{\mathcal{M}}} \ . \tag{19}$$

The remainder of this paper is devoted to the proofs.

4 Proofs

Before stating the proofs of Theorems 2 and 1, we need two technical lemmas. In the first lemma, we intend to evaluate the minimal penalty $\mathbb{E}[P_n(Ks_M - Ks_n(M))]$ for models of dimension not too large and not too small.

Lemma 4 *Assume (**P2**), (**Ab**), (**An**), (**Alb**) and ($\mathbf{Ac}_\infty$) of the general set of assumptions defined in Section 3.1. Then, for every model $M \in \mathcal{M}_n$ of dimension D_M such that*

$$0 < A_{\mathcal{M},+} (\ln n)^2 \leq D_M \ ,$$

we have for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$,

$$\left(1 - L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{20}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ , \tag{21}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}; \left(\frac{D_M \ln n}{n}\right)^{1/4}; \sqrt{R_{n,D_M}}\right\}$ *is defined in Remark 3.*

Proof. As explained in Remark 3, for all $n \geq n_0\left(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}}\right)$, we thus have on an event $\Omega_1\left(M\right)$ of probability at least $1 - 5n^{-2-\alpha_{\mathcal{M}}}$,

$$\left(1 - \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \ , \tag{22}$$

where $\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4};\ \left(\frac{D_M \ln n}{n}\right)^{1/4};\ \sqrt{R_{n,D_M}}\right\}$. Moreover, as $|Y_i| \leq A$ *a.s.* and $\|s_M\|_\infty \leq A$ by (**Ab**), it holds

$$0 \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq P_n Ks_M = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - s_M\left(X_I\right)\right)^2 \leq 4A^2 \tag{23}$$

and as $D_M \geq 1$, we have

$$\varepsilon_n\left(M\right) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4};\ \left(\frac{D_M \ln n}{n}\right)^{1/4};\ \sqrt{R_{n,D_M}}\right\} \geq A_0 n^{-1/8} \ . \tag{24}$$

We also have

$$\begin{aligned}&\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]\\ &= \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \ .\end{aligned} \tag{25}$$

Now notice that by (**An**) we have $\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0$. Hence, as $D_M \geq 1$, it comes from (23) and (24) that

$$0 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \leq 20A^2 n^{-2-\alpha_{\mathcal{M}}} \leq \frac{80A^2}{A_0^2\sigma_{\min}^2}\varepsilon_n^2\left(M\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{26}$$

Moreover, we have $\varepsilon_n\left(M\right) < 1$ for all $n \geq n_0\left(A_0, A_{\mathcal{M},+}, A_{cons}\right)$, so by (22),

$$0 < \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 - \varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\mathbf{1}_{\Omega_1(M)}\right] \tag{27}$$

$$\leq \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 + \varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{28}$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2}\varepsilon_n^2\left(M\right)$ by (24), we use (26), (27) and (28) in (25) to conclude by straightforward computations that

$$L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}} = \frac{80A^2}{A_0^2\sigma_{\min}^2} + 5A_0^{-2} + 1$$

is convenient in (20) and (21), as A_0 only depends on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$, A, $\sigma_{\min}$ and $r_{\mathcal{M}}$. ■

Lemma 5 *Let $\alpha > 0$. Assume that (**Ab**) of Section 3.1 is satisfied. Then a positive constant A_d exists, depending only in A, $A_{\mathcal{M},+}$, $\sigma_{\min}$ and α such that, by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq A_d\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right)\right) \leq 2n^{-\alpha} \ . \tag{29}$$

*If moreover, assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac**$_\infty$) of the general set of assumptions defined in Section 3.1 hold, then for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ and for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha)$, we have*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[p_2(M)\right]\right) \leq 2n^{-\alpha} \ , \tag{30}$$

where $p_2(M) := P_n(Ks_M - Ks_n(M)) \geq 0$.

Proof. We set

$$A_d = \max\left\{4A\sqrt{\alpha};\ \frac{8A^2}{3}\alpha;\ \frac{8A^2\alpha}{\sqrt{A_{\mathcal{M},+}}\sigma^2_{\min}} + \frac{16A^2\alpha}{3A_{\mathcal{M},+}\sigma_{\min}}\right\} \ . \tag{31}$$

Since by (**Ab**) we have $|Y| \leq A$ *a.s.* and $\|s_*\|_\infty \leq A$, it holds $\|s_*\|_\infty = \|\mathbb{E}[Y|X]\|_\infty \leq A$, and so $\|s_M - s_*\|_\infty \leq 2A$. Next, we apply Bernstein's inequality (96) to $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$. Notice that

$$K(s_M)(x,y) - K(s_*)(x,y) = (s_M(x) - s_*(x))(s_M(x) + s_*(x) - 2y) \ ,$$

hence $\|Ks_M - Ks_*\|_\infty \leq 8A^2$. Moreover, as $\mathbb{E}[Y - s_*(X)|X] = 0$ and $\mathbb{E}\left[(Y - s_*(X))^2|X\right] \leq \frac{(2A)^2}{4} = A^2$ we have

$$\begin{aligned}
&\mathbb{E}\left[(Ks_M(X,Y) - Ks_*(X,Y))^2\right] \\
&= \mathbb{E}\left[\left(4(Y - s_*(X))^2 + (s_M(X) - s_*(X))^2\right)(s_M(X) - s_*(X))^2\right] \\
&\leq 8A^2\mathbb{E}\left[(s_M(X) - s_*(X))^2\right] \\
&= 8A^2\ell(s_*, s_M) \, ,
\end{aligned}$$

and therefore, by (96) we have for all $x > 0$,

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{16A^2\ell(s_*, s_M)\,x}{n}} + \frac{8A^2x}{3n}\right) \leq 2\exp(-x) \ .$$

By taking $x = \alpha\ln n$, we then have

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{16A^2\alpha\ell(s_*, s_M)\ln n}{n}} + \frac{8A^2\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \ , \tag{32}$$

which gives the first part of Lemma 5 for A_d given in (31). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and by using it in (32) with $a = \ell(s_*, s_M)$, $b = \frac{4A^2\alpha\ln n}{n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + \left(4\sqrt{D_M} + \frac{8}{3}\right)\frac{A^2\alpha\ln n}{n}\right) \leq 2n^{-\alpha} \ . \tag{33}$$

Then, for a model $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$, we apply Lemma 4 and by (20), it holds for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$,

$$\left(1 - L_{A_{\mathcal{M},-}, A, \sigma_{\min}, r_{\mathcal{M}}, \alpha_{\mathcal{M}}}\varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}^2_{1,M} \leq \mathbb{E}\left[p_2(M)\right] \tag{34}$$

where $\varepsilon_n = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}, \left(\frac{D_M \ln n}{n}\right)^{1/4}, \sqrt{R_{n,D_M,\alpha}}\right\}$. Moreover as $D_M \leq A_{\mathcal{M},+} n (\ln n)^{-2}$ by (**P2**), $R_{n,D_M} \leq A_{cons} (\ln n)^{-1/2}$ by (6) and $A_{\mathcal{M},+} (\ln n)^2 \leq D_M$, we deduce that for all $n \geq n_0 (A_{\mathcal{M},+}, A, A_{cons}, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$,

$$L_{A_{\mathcal{M},-},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}} \varepsilon_n^2 (M) \leq 1/2 \ .$$

Now, since $\mathcal{K}_{1,M} \geq 2\sigma_{\min} > 0$ by (**An**), we have by (34), $\mathbb{E}[p_2(M)] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n}$ for all $n \geq n_0 (A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$. This allows, using (33), to conclude the proof for the value of A_d given in (31) by simple computations. ■

In order to avoid cumbersome notations in the proofs of Theorems 2 and 1, when generic constants L and n_0 depend on constants defined in the general set of assumptions stated in Section 3.1, we will note $L_{(\mathbf{GSA})}$ and $n_0 ((\mathbf{GSA}))$.

Proof of Theorem 2. From the definition of the selected model $\widehat{M}$ given in (3), $\widehat{M}$ minimizes

$$\mathrm{crit}(M) := P_n (Ks_n (M)) + \mathrm{pen}(M) \ , \tag{35}$$

over the models $M \in \mathcal{M}_n$. Hence, $\widehat{M}$ also minimizes

$$\mathrm{crit}'(M) := \mathrm{crit}(M) - P_n (Ks_*) \ . \tag{36}$$

over the collection $\mathcal{M}_n$. Let us write

$$\begin{aligned}
\ell (s_*, s_n (M)) &= P (Ks_n (M) - Ks_*) \\
&= P_n (Ks_n (M)) + P_n (Ks_M - Ks_n (M)) + (P_n - P)(Ks_* - Ks_M) \\
&\quad + P (Ks_n (M) - Ks_M) - P_n (Ks_*) \ .
\end{aligned}$$

By setting

$$\begin{aligned}
p_1 (M) &= P (Ks_n (M) - Ks_M) \ , \\
p_2 (M) &= P_n (Ks_M - Ks_n (M)) \ , \\
\bar{\delta}(M) &= (P_n - P)(Ks_M - Ks_*)
\end{aligned}$$

and

$$\mathrm{pen}'_{\mathrm{id}} (M) = p_1 (M) + p_2 (M) - \bar{\delta}(M) \ ,$$

we have

$$\ell (s_*, s_n (M)) = P_n (Ks_n (M)) + p_1 (M) + p_2 (M) - \bar{\delta}(M) - P_n (Ks_*) \tag{37}$$

and by (36),

$$\mathrm{crit}'(M) = \ell (s_*, s_n (M)) + (\mathrm{pen}(M) - \mathrm{pen}'_{\mathrm{id}} (M)) \ . \tag{38}$$

As $\widehat{M}$ minimizes crit$'$ over $\mathcal{M}_n$, it is therefore sufficient by (38), to control $\mathrm{pen}(M) - \mathrm{pen}'_{\mathrm{id}} (M)$ - or equivalently $\mathrm{crit}'(M)$ - in terms of the excess risk $\ell (s_*, s_n (M))$, for every $M \in \mathcal{M}_n$, in order to derive oracle inequalities. Let Ω_n be the event on which:

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$, (10) hold and

$$|p_1 (M) - \mathbb{E}[p_2 (M)]| \leq L_{(\mathbf{GSA})} \varepsilon_n (M) \mathbb{E}[p_2 (M)] \tag{39}$$

$$|p_2 (M) - \mathbb{E}[p_2 (M)]| \leq L_{(\mathbf{GSA})} \varepsilon_n^2 (M) \mathbb{E}[p_2 (M)] \tag{40}$$

$$\left|\bar{\delta}(M)\right| \leq \frac{\ell (s_*, s_M)}{\sqrt{D_M}} + L_{(\mathbf{GSA})} \frac{\ln n}{\sqrt{D_M}} \mathbb{E}[p_2 (M)] \tag{41}$$

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell (s_*, s_M) \ln n}{n}} + \frac{\ln n}{n}\right) \tag{42}$$

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$, (11) holds together with

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \tag{43}$$

$$p_2(M) \leq L_{(\mathbf{GSA})} \frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} \tag{44}$$

$$p_1(M) \leq L_{(\mathbf{GSA})} \frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} \tag{45}$$

By (16), (17), (18) and (19) in Remark 3, Lemma 4, Lemma 5 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, and since (10) holds with probability at least $1 - A_p n^{-2}$, we get for all $n \geq n_0((\mathbf{GSA}))$,

$$\mathbb{P}(\Omega_n) \geq 1 - A_p n^{-2} - 24 \sum_{M \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{A_p, c_{\mathcal{M}}} n^{-2} \ .$$

Control on the criterion crit' **for models of dimension not too small:**

We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$. Notice that (41) implies by (15) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$, for all $n \geq n_0((\mathbf{GSA}))$,

$$\begin{aligned} \left|\bar{\delta}(M)\right| &\leq L_{(\mathbf{GSA})} \left(\frac{(\ln n)^3}{D_M} \cdot \frac{\ln n}{D_M} \right)^{1/4} \times \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq L_{(\mathbf{GSA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \ , \end{aligned}$$

so that on Ω_n we have, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$,

$$\begin{aligned} &\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \\ &\leq \left|p_1(M) + p_2(M) - \mathrm{pen}(M)\right| + \left|\bar{\delta}(M)\right| \\ &\leq \left|p_1(M) + p_2(M) - 2\mathbb{E}\left[p_2(M)\right]\right| + \delta \mathbb{E}\left[p_2(M)\right] + L_{(\mathbf{GSA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq L_{(\mathbf{GSA})} \varepsilon_n(M) \, \mathbb{E}\left[p_2(M)\right] + \delta \mathbb{E}\left[p_2(M)\right] + L_{(\mathbf{GSA})} \varepsilon_n(M) \, \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \ . \end{aligned} \tag{46}$$

Now notice that using (**P2**) and (6) in (15) gives that for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0((\mathbf{GSA}))$, $0 < L_{(\mathbf{GSA})} \varepsilon_n(M) \leq \frac{1}{2}$. As $\ell(s_*, s_n(M)) = \ell(s_*, s_M) + p_1(M)$, we thus have on Ω_n, for all $n \geq n_0((\mathbf{GSA}))$,

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\ell(s_*, s_M) + p_2(M)\right] \\ &\leq \ell(s_*, s_n(M)) + \left|p_1(M) - \mathbb{E}\left[p_2(M)\right]\right| \\ &\leq \ell(s_*, s_n(M)) + \frac{L_{(\mathbf{GSA})} \varepsilon_n(M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n(M)} p_1(M) \qquad \text{by (39)} \\ &\leq \frac{1 + L_{(\mathbf{GSA})} \varepsilon_n(M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n(M)} \ell(s_*, s_n(M)) \\ &\leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \end{aligned} \tag{47}$$

Hence, using (47) in (46), we have on Ω_n for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0((\mathbf{GSA}))$,

$$\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \tag{48}$$

By consequence, for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0((\mathbf{GSA}))$, it holds on Ω_n, using (38) and (48),

$$\left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \leq \mathrm{crit}'(M) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \tag{49}$$

Control on the criterion crit' for models of small dimension:

We consider models $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$. By (11), (43) and (44), it holds on Ω_n, for any $\tau > 0$ and for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\begin{aligned}
&|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)| \\
&\leq p_1(M) + p_2(M) + \mathrm{pen}(M) + \left|\bar{\delta}(M)\right| \\
&\leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} + A_r \frac{(\ln n)^3}{n} + L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \\
&\leq L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} + \tau \ell(s_*, s_M) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \\
&\leq L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} + \tau \ell(s_*, s_n(M)) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \ . \qquad (50)
\end{aligned}$$

Hence, by taking $\tau = (\ln n)^{-2}$ in (50) we get that for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$, it holds on Ω_n,

$$|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)| \leq \frac{\ell(s_*, s_n(M))}{(\ln n)^2} + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \qquad (51)$$

Moreover, by (38) and (51), we have on the event Ω_n, for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\left(1 - (\ln n)^{-2}\right) \ell(s_*, s_n(M)) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}'(M) \qquad (52)$$

$$\leq \left(1 + (\ln n)^{-2}\right) \ell(s_*, s_n(M)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \qquad (53)$$

Oracle inequalities:

Recall that by the definition given in (2), an oracle model satisfies

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{\ell(s_*, s_n(M))\} \ . \qquad (54)$$

By Lemmas 6 and 7 below, we control on Ω_n the dimensions of the selected model $\widehat{M}$ and the oracle model M_*. More precisely, by (66) and (68), we have on Ω_n, for any $\frac{1}{2} > \eta > \left(1 - \beta_+\right)_+ / 2$ and for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{1/2+\eta} \ , \qquad (55)$$

$$D_{M_*} \leq n^{1/2+\eta} \ . \qquad (56)$$

Now, from (55) we distinguish two cases in order to control $\mathrm{crit}'\left(\widehat{M}\right)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta}$, we get by (49), for all $n \geq n_0((\mathbf{GSA}))$,

$$\mathrm{crit}'\left(\widehat{M}\right) \geq \left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n\left(\widehat{M}\right)\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) \ . \qquad (57)$$

Otherwise, if $D_{\widehat{M}} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (52),

$$\left(1 - (\ln n)^{-2}\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}'\left(\widehat{M}\right) \ . \qquad (58)$$

In all cases, we have by (57) and (58), for all $n \geq n_0\left((\mathbf{GSA})\right)$,

$$\mathrm{crit}'\left(\widehat{M}\right) \geq \left(1-\delta-(\ln n)^{-2}-L_{(\mathbf{GSA})} \sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/2+\eta}} \varepsilon_n(M)\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) - L_{(\mathbf{GSA}),A_r}\frac{(\ln n)^3}{n} \ . \tag{59}$$

Similarly, from (56) we distinguish two cases in order to control $\mathrm{crit}'(M_*)$. If $A_{\mathcal{M},+}(\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta}$, we get by (49), for all $n \geq n_0\left((\mathbf{GSA})\right)$,

$$\mathrm{crit}'(M_*) \leq \left(1+\delta+L_{(\mathbf{GSA})}\varepsilon_n(M_*)\right)\ell\left(s_*, s_n(M_*)\right) \ . \tag{60}$$

Otherwise, if $D_{M_*} \leq A_{\mathcal{M},+}(\ln n)^3$, we get by (53),

$$\mathrm{crit}'(M_*) \leq \left(1+(\ln n)^{-2}\right)\ell\left(s_*, s_n(M_*)\right) + L_{(\mathbf{GSA}),A_r}\frac{(\ln n)^3}{n} \ . \tag{61}$$

In all cases, we deduce from (60) and (61) that we have for all $n \geq n_0\left((\mathbf{GSA}),\delta\right)$,

$$\mathrm{crit}'(M_*) \leq \left(1+\delta+(\ln n)^{-2}+L_{(\mathbf{GSA})} \sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/2+\eta}} \varepsilon_n(M)\right)\ell\left(s_*, s_n(M_*)\right) + L_{(\mathbf{GSA}),A_r}\frac{(\ln n)^3}{n} \ . \tag{62}$$

Hence, by setting

$$\theta_n = L_{(\mathbf{GSA})} \times \sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/2+\eta}} \varepsilon_n(M) \ ,$$

we have by (15) and (6), for all $n \geq n_0\left((\mathbf{GSA}),\eta,\delta\right)$,

$$\theta_n \leq \frac{L_{(\mathbf{GSA})}}{(\ln n)^{1/4}} \ , \quad (\ln n)^{-2}+\theta_n+\delta<1 \ , \quad (\ln n)^{-2}+\theta_n < \frac{1-\delta}{2}$$

and we deduce from (59) and (62), since $\frac{1}{1-x} \leq 1+2x$ for all $x \in \left[0,\frac{1}{2}\right)$, that for all $n \geq n_0\left((\mathbf{GSA}),\eta,\delta\right)$, it holds on Ω_n,

$$\begin{aligned}\ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\leq \left(\frac{1+\delta+(\ln n)^{-2}+\theta_n}{1-\delta-(\ln n)^{-2}-\theta_n}\right)\ell\left(s_*, s_n(M_*)\right) + \frac{L_{(\mathbf{GSA}),A_r}}{1-\delta-(\ln n)^{-2}-\theta_n}\frac{(\ln n)^3}{n} \\ &\leq \left(\frac{1+\delta}{1-\delta}+\frac{5\left((\ln n)^{-2}+\theta_n\right)}{(1-\delta)^2}\right)\ell\left(s_*, s_n(M_*)\right) + L_{(\mathbf{GSA}),A_r}\frac{(\ln n)^3}{n} \ . \end{aligned} \tag{63}$$

Inequality (13) is now proved.
It remains to prove the second part of Theorem 2. We assume that assumption (**Ap**) holds. From Lemmas 6 and 7, we have that for any $\frac{1}{2} > \eta > \left(1-\beta_+\right)_+/2$ and for all $n \geq n_0\left((\mathbf{GSA}),C_-,\beta_-,\eta,\delta\right)$, it holds on Ω_n,

$$A_{\mathcal{M},+}(\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta} \ , \tag{64}$$

$$A_{\mathcal{M},+}(\ln n)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{65}$$

Now, using (57) and (60), by the same kind of computations leading to (63), we deduce that it holds on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}),C_-,\beta_-,\eta,\delta\right)$,

$$\begin{aligned}\ell\left(s_*, s_n\left(\widehat{M}\right)\right) &\leq \left(\frac{1+\delta+\theta_n}{1-\delta-\theta_n}\right)\ell\left(s_*, s_n(M_*)\right) \\ &\leq \left(\frac{1+\delta}{1-\delta}+\frac{5\theta_n}{(1-\delta)^2}\right)\ell\left(s_*, s_n(M_*)\right) \ . \end{aligned}$$

Thus inequality (14) is proved and Theorem 2 follows. ■

Lemma 6 (Control on the dimension of the selected model) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta > \left(1-\beta_+\right)_+/2$. If $n \geq n_0\left((\textbf{GSA}),\eta,\delta\right)$ then, on the event Ω_n defined in the proof of Theorem 2, it holds*

$$D_{\widehat{M}} \leq n^{1/2+\eta} \ . \tag{66}$$

*If moreover (**Ap**) holds, then for all $n \geq n_0\left((\textbf{GSA}),C_-,\beta_-,\eta,\delta\right)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta} \ . \tag{67}$$

Lemma 7 (Control on the dimension of oracle models) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta > \left(1-\beta_+\right)_+/2$. If $n \geq n_0\left((\textbf{GSA}),\eta\right)$ then, on the event Ω_n defined in the proof of Theorem 2, it holds*

$$D_{M_*} \leq n^{1/2+\eta} \ . \tag{68}$$

*If moreover (**Ap**) holds, then for all $n \geq n_0\left((\textbf{GSA}),C_-,\beta_-,\eta\right)$, we have on the event Ω_n,*

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{69}$$

Proof of Lemma 6. Recall that $\widehat{M}$ minimizes

$$\mathrm{crit}'\left(M\right) = \mathrm{crit}\left(M\right) - P_n K s_* = \ell\left(s_*,s_M\right) - p_2\left(M\right) + \bar{\delta}\left(M\right) + \mathrm{pen}\left(M\right) \tag{70}$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\mathrm{crit}'\left(M\right)$ for small models in the case where (**Ap**) hold : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}\left(\ln n\right)^3$. We then have on Ω_n,

$$\begin{aligned}
&\ell\left(s_*,s_M\right) \geq C_- A_{\mathcal{M},+}^{-\beta_-}\left(\ln n\right)^{-3\beta_-} \text{ by } (\textbf{Ap}) \\
&\mathrm{pen}\left(M\right) \geq 0 \\
&p_2\left(M\right) \leq L_{(\textbf{GSA})}\frac{\left(\ln n\right)^3}{n} \qquad \text{from (44)} \\
&\bar{\delta}\left(M\right) \geq -L_{(\textbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*,s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \text{ from (43).}
\end{aligned}$$

   Since by (**Ab**), we have $0 \leq \ell\left(s_*,s_M\right) \leq 4A^2$, we deduce that for all $n \geq n_0\left((\textbf{GSA}),C_-,\beta_-\right)$,

$$\mathrm{crit}'\left(M\right) \geq \frac{C_- A_{\mathcal{M},+}^{-\beta_-}}{2}\left(\ln n\right)^{-3\beta_-} \ . \tag{71}$$

2. Lower bound for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (10) and (40) we have on Ω_n,

$$\mathrm{pen}\left(M\right) - p_2\left(M\right) \geq \left(1-\delta-L_{(\textbf{GSA})}\varepsilon_n^2\left(M\right)\right)\mathbb{E}\left[p_2\left(M\right)\right] \ .$$

   Using (**P2**), (6) and the fact that $D_M \geq n^{1/2+\eta}$ in (15), we deduce that for all $n \geq n_0\left((\textbf{GSA}),\eta,\delta\right)$, $L_{(\textbf{GSA})}\varepsilon_n^2\left(M\right) \leq \frac{1}{2}\left(1-\delta\right)$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 4 that for all $n \geq n_0\left((\textbf{GSA}),\eta\right)$, $\mathbb{E}\left[p_2\left(M\right)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0\left((\textbf{GSA}),\eta,\delta\right)$,

$$\mathrm{pen}\left(M\right) - p_2\left(M\right) \geq \frac{\sigma_{\min}^2}{4}\left(1-\delta\right)\frac{D_M}{n} \ . \tag{72}$$

   From (42) it holds on Ω_n,

$$\bar{\delta}\left(M\right) \geq -L_{(\textbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*,s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \ . \tag{73}$$

   Hence, as $D_M \geq n^{1/2+\eta}$ and as by (**Ab**), $0 \leq \ell\left(s_*,s_M\right) \leq 4A^2$, we deduce from (70), (72) and (73) that we have on Ω_n, for all $n \geq n_0\left((\textbf{GSA}),\eta,\delta\right)$,

$$\mathrm{crit}'\left(M\right) \geq \left(1-\delta\right)L_{(\textbf{GSA})}n^{-1/2+\eta} \ . \tag{74}$$

3. A better model exists for $\mathrm{crit}'(M)$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Then, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} .$$

Using ($\mathbf{Ap}_u$),

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2} . \tag{75}$$

By (41), we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq \frac{\ell(s_*, s_{M_0})}{\sqrt{D_{M_0}}} + L_{(\mathbf{GSA})}\frac{\ln n}{\sqrt{D_{M_0}}}\mathbb{E}[p_2(M_0)] \tag{76}$$

and by (10),

$$\mathrm{pen}(M_0) \leq 3\mathbb{E}[p_2(M_0)] .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ and $\ell(s_*, s_{M_0}) \leq 4A^2$ by (**Ab**) and as for all $n \geq n_0((\mathbf{GSA}))$ $\varepsilon_n(M) \leq 1$, we deduce from inequalities (75), (76) and Lemma 4 that for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq L_{(\mathbf{GSA})}\left(n^{-\left(\beta_+/2+1/4\right)} + \ln(n)\, n^{-3/4}\right)$$

and

$$\mathrm{pen}(M_0) \leq L_{(\mathbf{GSA})} n^{-1/2} .$$

By consequence, we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\begin{aligned}\mathrm{crit}'(M_0) &\leq \ell(s_*, s_{M_0}) + \left|\bar{\delta}(M_0)\right| + \mathrm{pen}(M_0) \\ &\leq L_{(\mathbf{GSA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) .\end{aligned} \tag{77}$$

To conclude, notice that the upper bound (77) is smaller than the lower bound given in (74) for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$. Hence, points 2 and 3 above yield inequality (66). Moreover, the upper bound (77) is smaller than lower bounds given in (71), derived by using (**Ap**), and (74), for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta, \delta\right)$. This thus gives (67) and Lemma 6 is proved. ■

Proof of Lemma 7. By definition, M_* minimizes

$$\ell(s_*, s_n(M)) = \ell(s_*, s_M) + p_1(M)$$

over the models $M \in \mathcal{M}_n$.

1. Lower bound on $\ell(s_*, s_n(M))$ for small models : let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}(\ln n)^3$. In this case we have

$$\ell(s_*, s_n(M)) \geq \ell(s_*, s_M) \geq C_- A_{\mathcal{M},+}^{-\beta_-}(\ln n)^{-3\beta_-} \text{ by } (\mathbf{Ap}). \tag{78}$$

2. Lower bound of $\ell(s_*, s_n(M))$ for large models : let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/2+\eta}$. From (39) we get on Ω_n,

$$p_1(M) \geq \left(1 - L_{(\mathbf{GSA})}\varepsilon_n(M)\right)\mathbb{E}[p_2(M)] .$$

Using (**P2**), (6) and the fact that $D_M \geq n^{1/2+\eta}$ in (15), we deduce that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $L_{(\mathbf{GSA})}\varepsilon_n(M) \leq \frac{1}{2}$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 4 that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $\mathbb{E}[p_2(M)] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$. By consequence, it holds for all $n \geq n_0((\mathbf{GSA}), \eta)$, on the event Ω_n,

$$\ell(s_*, s_n(M)) \geq p_1(M) \geq \frac{\sigma_{\min}^2}{4}\frac{D_M}{n} \geq \frac{\sigma_{\min}^2}{4} n^{-1/2+\eta} . \tag{79}$$

3. A better model exists for $\ell(s_*, s_n(M))$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n}$. Moreover, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$A_{\mathcal{M},+}(\ln n)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich}\sqrt{n} \leq n^{1/2+\eta} .$$

Using $(\mathbf{Ap}_u)$,

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/2}$$

and by (39)

$$p_1(M_0) \leq \left(1 + L_{(\mathbf{GSA})}\varepsilon_n(M)\right)\mathbb{E}[p_2(M_0)]$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (6) and (15), for all $n \geq n_0((\mathbf{GSA}))$ it holds $\varepsilon_n(M) \leq 1$, we deduce from Lemma 4 that for all $n \geq n_0((\mathbf{GSA}))$, on the event Ω_n,

$$p_1(M_0) \leq L_{(\mathbf{GSA})}\frac{D_M}{n} \leq L_{(\mathbf{GSA})}n^{-1/2} .$$

By consequence, on Ω_n, for all $n \geq n_0((\mathbf{GSA}))$,

$$\begin{aligned}\ell(s_*, s_n(M_0)) &= \ell(s_*, s_{M_0}) + p_1(M_0) \\ &\leq L_{(\mathbf{GSA})}\left(n^{-\beta_+/2} + n^{-1/2}\right) . \end{aligned} \tag{80}$$

The upper bound (80) is smaller than the lower bound (79) for all $n \geq n_0((\mathbf{GSA}), \eta)$, and this gives (68). If (**Ap**) hold, then the upper bound (80) is smaller than the lower bounds (78) and (79) for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta\right)$, which proves (69) and allows to conclude the proof of Lemma 7. ■

Proof of Theorem 1. Similarly to the proof of Theorem 2, we consider the event Ω'_n of probability at least $1 - L_{c_{\mathcal{M}},A_p}n^{-2}$ for all $n \geq n_0((\mathbf{GSA}))$, on which: (8) holds and

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ it holds

$$|p_1(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\mathbb{E}[p_2(M)] , \tag{81}$$

$$|p_2(M) - \mathbb{E}[p_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\mathbb{E}[p_2(M)] . \tag{82}$$

- For all models $M \in \mathcal{M}_n$ with $D_M \leq A_{\mathcal{M},+}(\ln n)^2$ it holds

$$p_2(M) \leq L_{(\mathbf{GSA})}\frac{(\ln n)^2}{n} . \tag{83}$$

- For every $M \in \mathcal{M}_n$,

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) . \tag{84}$$

Let $d \in (0, 1)$ to be chosen later.

Lower bound on $D_{\widehat{M}}$. Remind that $\widehat{M}$ minimizes

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_n K s_* = \ell(s_*, s_M) - p_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) . \tag{85}$$

1. Lower bound on $\mathrm{crit}'(M)$ for "small" models : assume that $M \in \mathcal{M}_n$ and

$$D_M \leq dA_{rich}n(\ln n)^{-2} .$$

We have

$$\ell(s_*, s_M) + \mathrm{pen}(M) \geq 0 \tag{86}$$

and from (84), as $\ell(s_*, s_M) \leq 4A^2$ by (**Ab**), we get on Ω'_n, for all $n \geq n_0((\mathbf{GSA}),d)$,

$$\begin{aligned}\bar{\delta}(M) &\geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \\ &\geq -L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} \\ &\geq -d \times A^2 A_{rich}(\ln n)^{-2} \ . \end{aligned} \tag{87}$$

Then, if $D_M \geq A_{\mathcal{M},+}(\ln n)^2$, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (6) and (15), for all $n \geq n_0((\mathbf{GSA}))$ it holds $L_{(\mathbf{GSA})}\varepsilon_n(M) \leq 1$, we deduce from (82) and Lemma 4 that for all $n \geq n_0((\mathbf{GSA}))$,

$$p_2(M) \leq 2\mathbb{E}[p_2(M)] \leq 36A^2\frac{D_M}{n} \leq d \times 36A^2 A_{rich}(\ln n)^{-2} \ .$$

Whenever $D_M \leq A_{\mathcal{M},+}(\ln n)^2$, (83) gives that, for all $n \geq n_0((\mathbf{GSA}),d)$, on the event Ω'_n,

$$p_2(M) \leq L_{(\mathbf{GSA})}\frac{(\ln n)^2}{n} \leq d \times 36A^2 A_{rich}(\ln n)^{-2} \ .$$

Hence, we have checked that for all $n \geq n_0((\mathbf{GSA}),d)$, on the event Ω'_n,

$$-p_2(M) \geq -d \times 36A^2 A_{rich}(\ln n)^{-2} \ , \tag{88}$$

and finally, by using (86), (87) and (88) in (85), we deduce that on Ω'_n, for all $n \geq n_0((\mathbf{GSA}),d)$,

$$\mathrm{crit}'(M) \geq -d \times 37A^2 A_{rich}(\ln n)^{-2} \ . \tag{89}$$

2. There exists a better model for $\mathrm{crit}'(M)$: By (**P3**), for all $n \geq n_0(A_{\mathcal{M},+}, A_{rich})$ a model $M_1 \in \mathcal{M}_n$ exists such that

$$A_{\mathcal{M},+}(\ln n)^2 \leq \frac{A_{rich}n}{(\ln n)^2} \leq D_{M_1} \ .$$

We then have on Ω'_n,

$$\begin{aligned} \ell(s_*, s_{M_1}) &\leq A_{rich}^{-\beta_+}(\ln n)^{2\beta_+} n^{-\beta_+} && \text{by } (\mathbf{Ap}_u) \\ p_2(M_1) &\geq \left(1 - L_{(\mathbf{GSA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] && \text{by (82)} \\ \mathrm{pen}(M_1) &\leq A_{\mathrm{pen}}\mathbb{E}[p_2(M_1)] && \text{by (8)} \\ \left|\bar{\delta}(M_1)\right| &\leq L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} && \text{by (84) and } (\mathbf{Ab}) \end{aligned}$$

and therefore,

$$\mathrm{crit}'(M_1) \leq \left(-1 + A_{\mathrm{pen}} + L_{(\mathbf{GSA})}\varepsilon_n^2(M_1)\right)\mathbb{E}[p_2(M_1)] + L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} + A_{rich}^{-\beta_+}\frac{(\ln n)^{2\beta_+}}{n^{\beta_+}} \ . \tag{90}$$

Hence, as $-1 + A_{\mathrm{pen}} < 0$, and as by (6), (15), (**An**) and Lemma 4 it holds for all $n \geq n_0((\mathbf{GSA}), A_{\mathrm{pen}})$

$$L_{(\mathbf{GSA})}\varepsilon_n^2(M_1) \leq \frac{1 - A_{\mathrm{pen}}}{2} \quad \text{and} \quad \mathbb{E}[p_2(M_1)] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n} \geq \frac{\sigma_{\min}^2 A_{rich}}{2}(\ln n)^{-2} \ ,$$

we deduce from (90) that on Ω'_n, for all $n \geq n_0((\mathbf{GSA}), A_{\mathrm{pen}})$,

$$\mathrm{crit}'(M_1) \leq -\frac{1}{4}(1 - A_{\mathrm{pen}})\sigma_{\min}^2 A_{rich}(\ln n)^{-2} \ . \tag{91}$$

Now, by taking

$$0 < d = \left(\frac{1}{149} \left(1 - A_{\text{pen}} \right) \left(\frac{\sigma_{\min}}{A} \right)^2 \right) \wedge \frac{1}{2} < 1 \tag{92}$$

and by comparing (89) and (91), we deduce that on Ω_n', for all $n \geq n_0 \left((\mathbf{GSA}), A_{\text{pen}} \right)$, for all $M \in \mathcal{M}_n$ such that $D_M \leq dA_{rich} n \left(\ln n \right)^{-2}$,

$$\text{crit}' \left(M_1 \right) < \text{crit}' \left(M \right)$$

and so

$$D_{\widehat{M}} > dA_{rich} n \left(\ln n \right)^{-2} \ . \tag{93}$$

Excess Risk of $s_n\left(\widehat{M}\right)$. We take d with the value given in (92). First notice that for all $n \geq n_0 \left(A_{\mathcal{M},+}, A_{rich}, d \right)$, we have $dA_{rich} n \left(\ln n \right)^{-2} \geq A_{\mathcal{M},+} \left(\ln n \right)^2$. Hence, for all $M \in \mathcal{M}_n$ such that $D_M \geq dA_{rich} n \left(\ln n \right)^{-2}$, by (6), (15), (**P2**), (**An**) and Lemma 4, it holds on Ω_n' for all $n \geq n_0 \left((\mathbf{GSA}), A_{\text{pen}} \right)$, using (81),

$$\ell \left(s_*, s_n \left(M \right) \right) \geq p_1 \left(M \right) \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n} \geq \frac{d\sigma_{\min}^2 A_{rich}}{2} \left(\ln n \right)^{-2} \ .$$

By (93), we thus get that on Ω_n', for all $n \geq n_0 \left((\mathbf{GSA}), A_{\text{pen}} \right)$,

$$\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \geq \frac{d\sigma_{\min}^2 A_{rich}}{2} \left(\ln n \right)^{-2} \ . \tag{94}$$

Moreover, the model M_0 defined in (**P3**) satisfies, for all $n \geq n_0 \left((\mathbf{GSA}) \right)$,

$$A_{\mathcal{M},+} \left(\ln n \right)^3 \leq \sqrt{n} \leq D_{M_0} \leq c_{rich} \sqrt{n}$$

and so using $(\mathbf{Ap}_u)$,

$$\ell \left(s_*, s_{M_0} \right) \leq C_+ n^{-\beta_+/2} \ .$$

In addition, by (39),

$$p_1 \left(M \right) \leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n \left(M \right) \right) \mathbb{E} \left[p_2 \left(M \right) \right] \ .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (6) and (15), for all $n \geq n_0 \left((\mathbf{GSA}) \right)$ it holds $\varepsilon_n \left(M \right) \leq 1$, we deduce from Lemma 4 that for all $n \geq n_0 \left((\mathbf{GSA}) \right)$

$$p_1 \left(M \right) \leq L_{(\mathbf{GSA})} \frac{D_M}{n} \leq L_{(\mathbf{GSA})} n^{-1/2} \ .$$

By consequence, for all $n \geq n_0 \left((\mathbf{GSA}) \right)$,

$$\ell \left(s_*, s_n \left(M_0 \right) \right) \leq L_{(\mathbf{GSA})} \left(n^{-\beta_+/2} + n^{-1/2} \right) \tag{95}$$

and the ratio between the two bounds (94) and (95) is larger than $\ln \left(n \right)$ for all $n \geq n_0 \left(L_{(\mathbf{GSA})}, A_{\text{pen}} \right)$, which yields (9). ■

Probabilistic Tools We recall here the main probabilistic results that are instrumental in our proofs. The following tool is the well known Bernstein's inequality, that can be found for example in [28], Proposition 2.9.

Theorem 8 *(Bernstein's inequality) Let $(X_1, ..., X_n)$ be independent real valued random variables and define*

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mathbb{E} \left[X_i \right] \right).$$

Assuming that

$$v = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[X_i^2 \right] < \infty$$

and

$$X_i \leq b \quad a.s.$$

we have, for every $x > 0$,

$$\mathbb{P}\left[|S| \geq \sqrt{2v\frac{x}{n}} + \frac{bx}{3n}\right] \leq 2\exp\left(-x\right). \tag{96}$$

We now turn to concentration inequalities for the empirical process around its mean. Bousquet's inequality [17] provides optimal constants for the deviations above the mean. Klein-Rio's inequality [20] gives sharp constants for the deviations below the mean, that slightly improves Klein's inequality [21].

Theorem 9 *Let* $(\xi_1, ..., \xi_n)$ *be* n *i.i.d. random variables having common law* P *and taking values in a measurable space* $\mathcal{Z}$*. If* $\mathcal{F}$ *is a class of measurable functions from* $\mathcal{Z}$ *to* $\mathbb{R}$ *satisfying*

$$|f\left(\xi_i\right) - Pf| \leq b \quad a.s., \text{ for all } f \in \mathcal{F},\ i \leq n,$$

then, by setting

$$\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} \left\{P\left(f^2\right) - \left(Pf\right)^2\right\},$$

we have, for all $x \geq 0$,

Bousquet's inequality *:*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{3n}\right] \leq \exp\left(-x\right) \tag{97}$$

and we can deduce that, for all $\varepsilon, x > 0$*, it holds*

$$\mathbb{P}\left[\|P_n - P\|_{\mathcal{F}} - \mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + \frac{1}{3}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{98}$$

Klein-Rio's inequality :

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{99}$$

and again, we can deduce that, for all $\varepsilon, x > 0$*, it holds*

$$\mathbb{P}\left[\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] - \|P_n - P\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + 1\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{100}$$

References

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] Sylvain Arlot. Model selection by resampling penalization, March 2008. oai:hal.archives-ouvertes.fr:hal-00262478_v2.

[4] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[5] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, September 2009. arXiv:0909.1884v1.

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

[7] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[8] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[9] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[10] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[11] Jean-Patrick Baudry. Clustering through model selection criteria, June 2007. Poster session at One Day Statistical Workshop in Lisieux.

[12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope Heuristics : Overview and Implementation. Technical Report 7223, INRIA, 2010.

[13] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[14] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[15] L. Birgé and P. Massart. Gaussian model selection. *J.Eur.Math.Soc.*, 3(3):203–268, 2001.

[16] S. Boucheron and P. Massart. A high dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 2010. To appear.

[17] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[18] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université de Paris-Sud.*, 1999.

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

[20] R. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of Probability*, 1:63–87 (electronic), 2005.

[21] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C.R. Acad. Sci. Paris, Ser I*, 334:500–505, 2002.

[22] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001.

[23] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimisation. *Ann. Statist.*, 34(6):2593–2656, 2006.

[24] Vincent Lepez. *Some estimation problems related to oil reserves*. PhD thesis, University Paris-Sud XI, 2002.

[25] Matthieu Lerasle. Optimal model selection in density estimation, 2009. arXiv:0910.1654.

[26] Matthieu Lerasle. *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes.* PhD thesis, INSA Toulouse, June 2009.

[27] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[28] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007.

[29] Cathy Maugis and Bertrand Michel. A nonasymptotic penalized criterion for Gaussian mixture model selection. a variable selection and clustering problems. Technical Report 6549, INRIA, 2008.

[30] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, August 2010. hal-00512304, v1.

[31] N. Verzelen. High-dimensional Gaussian model selection on a Gaussian design. Technical Report RR-6616, INRIA, 2008.

[32] Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris-Sud XI, December 2007.