Optimal model selection in heteroscedastic regression using piecewise polynomial functions

Adrien Saumard*

Department of Statistics, University of Washington, Seattle, WA 98195, USA
INRIA Saclay Île-de-France, France
e-mail: `asaumard@uw.edu`

Abstract: We consider the estimation of a regression function with random design and heteroscedastic noise in a nonparametric 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. The optimal penalty is shown to be twice the minimal one and to satisfy a non-asymptotic pathwise oracle inequality with leading constant almost one. Finally, the ideal penalty being unknown in general, we propose a hold-out penalization procedure and show that the latter is asymptotically optimal.

Contents

1. Introduction

Given a collection of models and associated estimators, two different model selection 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 or a loss (efficiency problem). We focus on the efficiency problem, where the leading idea of penalization, that goes back to early works of Akaike [2, 3] and Mallows [33], is to perform an unbiased - or uniformly biased - estimation of the risk of the estimators. FPE and AIC procedures proposed by Akaike respectively in [2] and [3], as well as Mallows' C_p or C_L [33], aim to do so by adding to the empirical risk a penalty which depends on the dimension of the models.

The first analysis of such procedures had the drawback of being fundamentally asymptotic, considering in particular that the number of models as well as their dimensions are fixed while the sample size tends to infinity. As explained for instance in Massart [34], in various statistical settings it is natural to let these quantities depend on the amount of data. Thus, pointing out the importance of Talagrand's type concentration inequalities in the nonasymptotic approach, Birgé and Massart [16, 18] and Barron, Birgé and Massart [11] have been able to build nonasymptotic oracle inequalities for penalization procedures. Their framework takes into account the complexity of the collection of models as a parameter depending on the sample size.

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 less accurate global penalties, see Koltchinskii [27], Bartlett et al. [12], to the refined local Rademacher complexities in the case where some favorable noise conditions hold (see for instance Bartlett, Bousquet and Mendelson [13], Koltchinskii [28]). But as a price to pay for generality, the above penalties suffer from their dependence on unknown constants. These penalized procedures are very difficult to implement and calibrate in practice. Moreover, the existing risk bounds for these procedures contain very large leading constants. Other general-purpose penalties have been proposed, such as the bootstrap penalties of Efron [26] and the resampling and V-fold penalties of Arlot [5, 6]. These penalties are essentially resampling estimates of the difference between the empirical risk and the risk. Arlot [5, 6] proved 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 conjectured that the restriction to histograms is mainly technical and that his results can be extended to more general situations.

Model selection *via* penalization is not the only method which provides sharp oracle inequalities for the estimation of a nonparametric regression function. Indeed, aggregation techniques and PAC-Bayesian bounds also allow to obtain nearly optimal constants in the oracle inequalities. Bunea et al. [21] derived some sharp oracle inequalities for different aggregation tasks by means of a single unifying procedure. However, the authors asked for a fixed design and

homoscedastic Gaussian noise. By using aggregation with exponential weights, Dalalyan and Tsybakov obtained in [25] oracle inequalities of a PAC-Bayesian flavor with leading constant one and optimal rate of the remainder term for the estimation of a regression function with deterministic design and homoscedastic errors. Furthermore, these authors allowed error distributions which are symmetric or n-divisible. PAC-Bayesian methods are systematically investigated in Catoni, [23]. The work of Lecué and Mendelson [29] concerning the aggregation by empirical risk minimization of a finite family of functions seems to handle the case of a random design and heteroscedastic noise, even if this example is not explicitly developed. The oracle inequalities obtained by Lecué and Mendelson are sharp and valid with probability close to one. In particular, they are related to oracle inequalities obtained, in expectation, by Catoni in [23].

A difference between aggregation and model selection studies, is that in most aggregation results, the estimators at hand are considered as deterministic functions. However, notable exceptions are the following. Leung and Barron [32] proved sharp oracle inequalities for the aggregation of projection estimators in the Gaussian sequence model. Rigollet and Tsybakov [35] recently showed sharp bounds for the aggregation of some linear estimators, including projection estimators, in a regression setting, with fixed design and homoscedastic Gaussian noise. More general PAC-Bayesian type inequalities were also recently obtained by Dalalyan and Salmon [24], considering the aggregation of affine estimators in heteroscedastic regression, with Gaussian noise and fixed design.

Birgé and Massart [19] discovered, in a generalized linear Gaussian model setting, that the optimal penalty is closely related to the minimal one. An optimal penalty is a penalty which gives an oracle inequality with leading constant converging to one when the sample size tends to infinity. The minimal penalty is defined to be the maximal penalty under which the procedure totally misbehaves (in a sense to be specified below). Birgé and Massart [19] proved sharp upper and lower bounds for the minimal penalty. These authors also showed that the optimal penalty is twice the minimal one, both for small and large collections of models. These facts are called the *slope heuristics*. The authors also exhibited a jump in the dimension of the selected model occurring around the value of the minimal penalty, and used it to estimate the minimal penalty from the data. Taking a penalty equal to twice the previous estimate then gives a nonasymptotic quasi-optimal data-driven model selection procedure. The algorithm proposed by Birgé and Massart [19] 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. Thus, their procedure gives a data-driven *calibration* of the minimal penalty.

Considering the case of Gaussian least-squares regression with unknown variance, Baraud et al. [10] have also derived lower bounds on the penalty terms for small and large collections of models. In the setting of maximum likelihood estimation of density on histograms, Castellan [22] obtained a lower bound on the penalty term, in the case of small collections of models.

The slope heuristics has been then extended by Arlot and Massart [9] in a bounded regression framework, with heteroscedastic noise and random design.

The authors considered least-squares estimators on a "small" collection of histograms models. Their analysis differs from the one of Birgé and Massart [19] in an important way. Indeed, Arlot and Massart [9] did not assume a particular *shape* of the penalty term. As a matter of fact, the penalties considered by Birgé and Massart [19] were known functions of the dimension of the models, whereas heteroscedasticity of the noise allowed Arlot and Massart to consider situations where the shape of the penalty is not even a function of the dimension of the models. In such general cases, the authors proposed 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 [5, 6].

The approach developed in [9] is more general than the histogram case, except for some identified technical parts of the proofs, thus providing a general framework that can be applied to other problems. The authors have also identified, in the case of histograms, the minimal penalty as the mean of the empirical excess loss on each model, and the ideal penalty to be estimated as the sum of the empirical excess loss and true excess loss on each model. The slope heuristics then heavily relies on the fact that the empirical excess loss is equivalent to the true excess loss for models of reasonable dimensions.

Arlot and Massart [9] conjectured that this equivalence between the empirical and true excess loss is a quite general fact in M-estimation. A general result supporting this conjecture is the high dimensional Wilks' phenomenon investigated by Boucheron and Massart [20] in the setting of bounded contrast minimization. The authors derive in [20] concentration inequalities for the empirical excess loss, under some margin conditions (called "noise conditions" by the authors) and when the considered model satisfies some general "complexity condition" on the first moment of the supremum of the empirical process on localized slices of variance in the loss class. The latter assumption can be explicated under suitable covering entropy conditions on the model.

Lerasle [31] 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 the author in this framework allows sharp computations and the empirical excess loss is shown to be exactly equal to the true excess loss. Lerasle [31] also proved in the least-squares density estimation setting the efficiency of Arlot's resampling penalties. Moreover, Lerasle [30] generalized the previous results to weakly dependent data. Arlot and Bach [8] recently considered the problem of selecting among linear estimators in nonparametric 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 close to one, converging to one when the sample size tends to infinity.

In this paper, we prove the validity of the slope heuristics in the framework of bounded regression with random design and heteroscedastic noise. This is done

by considering a "small" collection of finite-dimensional linear models of piece-wise polynomial functions. This setting extends the case of histograms already treated by Arlot and Massart [9]. An interesting consequence is that piecewise polynomial functions are known to have good approximation properties in Besov spaces and can lead to minimax rates of convergence, see for instance [11, 37]. As a matter of fact, histograms allow minimax procedures only on Hölder spaces.

Our validation of the slope heuristics is of asymptotic nature. However, the complexity of the collection of models as well as their dimensions are not constant terms in our analysis. These quantities are indeed allowed to depend on the sample size n.

If the noise is homoscedastic, then the shape of the ideal penalty is known, and is linear in the dimension of the models as in the case of Mallows' C_p. However, if the noise is heteroscedastic, then Arlot [7] showed that the ideal penalty is not even a function of the linear dimensions of the models. So, it is necessary to give a suitable estimator of this shape. As emphasized by Arlot [5, 6], V-fold and resampling penalties are good, natural candidates for this task. In this paper, we show that a hold-out penalty - which is closely related to a special case of resampling penalty - is indeed asymptotically optimal under very mild conditions on the data split. As a matter of fact, a half-and-half split leads to an optimal penalization. It is worth noticing that hold-out type procedures have also been exploited in Chapter 8 of Massart [34] as simple tools to overcome the margin adaptivity issue in classification.

The paper is organized as follows. In Section 2, we describe the statistical framework. The slope heuristics is presented in Section 3, and the hold-out penalization is considered in Section 4. The proofs are collected in Section 5.

2. Statistical framework

2.1. Penalized least-squares model selection

Let us take n independent observations $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ with common distribution P. In Sections 2.2, 3.2-4 the feature space $\mathcal{X} = [0, 1]$. The marginal distribution of X_i is denoted by P^X. We assume that the data satisfy the following relation

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

where $s_* \in L_2 (P^X)$. Conditionally to X_i, the residual ε_i is assumed to have zero mean and variance equal to one. The function $\sigma : \mathcal{X} \to \mathbb{R}_+$ is the unknown heteroscedastic noise level. A generic random variable with distribution P, independent of the sample $(\xi_1, \ldots, \xi_n)$, is denoted by $\xi = (X, Y)$.

It follows from (1) that s_* is the unknown regression function of Y with respect to X. Our aim is to estimate s_* from the sample. To do so, we are given a finite collection of models $\mathcal{M}_n$, with cardinality depending on the sample size n. Each model $M \in \mathcal{M}_n$ is assumed to be a finite-dimensional vector space. We denote by D_M the linear dimension of M. In the main part of this paper,

we focus on models of piecewise polynomial functions, that are introduced in Section 2.2 below.

We denote by $\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X \right)^{1/2}$ the usual norm in $L_2 \left(P^X \right)$ and by s_M the linear projection of s_* onto M in the Hilbert space $\left(L^2 \left(P^X \right), \|\cdot\|_2 \right)$. For a function $f \in L_1 (P)$, we write $P(f) = Pf = \mathbb{E} \left[f \left(\xi \right) \right]$. By setting $K : L_2 \left(P^X \right) \to L_1 (P)$ the least-squares contrast, defined by

$$K (s) : (x, y) \mapsto (y - s(x))^2 \ , \qquad s \in L_2 \left(P^X \right) \ , \tag{2}$$

the regression function s_* satisfies

$$s_* = \arg \min_{s \in L_2(P^X)} P(K(s)) \ . \tag{3}$$

For the linear projections s_M we get

$$s_M = \arg \min_{s \in M} P(K(s)) \ . \tag{4}$$

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

$$s_n (M) \in \arg \min_{s \in M} \left\{ P_n (K(s)) \right\}$$

$$= \arg \min_{s \in M} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(X_i))^2 \right\} \ ,$$

where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{\xi_i}$ is the empirical measure built from the data.

In order to avoid cumbersome notations, we will often write Ks in place of $K(s)$ for the image of a suitable function s by the contrast K. We measure the performance of the least-squares estimators by their excess loss,

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

We have the following decomposition,

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

where

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

and

$$\ell (s_M, s_n (M)) := P(Ks_n (M) - Ks_M) \geq 0 \ .$$

The quantity $\ell (s_*, s_M)$ is called the bias of the model M and $\ell (s_M, s_n (M))$ is the excess loss of the least-squares estimator $s_n (M)$ on the model M. By the Pythagorean identity, we have

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

Given the collection of models $\mathcal{M}_n$, an oracle model M_* is defined as a minimizer of the losses - or equivalently excess losses - of the estimators at hand,

$$M_* \in \arg \min_{M \in \mathcal{M}_n} \left\{ \ell \left(s_*, s_n \left(M \right) \right) \right\} . \tag{5}$$

The associated oracle estimator $s_n \left(M_* \right)$ thus achieves the best performance in terms of excess loss among the collection $\left\{ s_n \left(M \right) ; M \in \mathcal{M}_n \right\}$. The oracle model is a random quantity because it depends on the data and it is also unknown as it depends on the distribution P of the data. We propose to estimate the oracle model by a penalization procedure.

Given some known penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}$, we consider the following data-dependent model, also called selected model,

$$\widehat{M} \in \arg \min_{M \in \mathcal{M}_n} \left\{ P_n \left(K s_n \left(M \right) \right) + \text{pen} \left(M \right) \right\} . \tag{6}$$

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

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

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

2.2. Piecewise polynomial functions

Let us take $\mathcal{X} = [0, 1]$ the unit interval and $\mathcal{P}$ a finite partition of $\mathcal{X}$. For a positive integer r and any $(I, j) \in \mathcal{P} \times \{0, \ldots, r\}$, we set

$$p_{I,j} : x \in \mathcal{X} \mapsto x^j \mathbf{1}_I \left(x \right) .$$

Definition 1. A finite dimensional vector space M is said to be a model of piecewise polynomial functions, with respect to the finite partition $\mathcal{P}$ of $\mathcal{X} = [0, 1]$ and of degrees not larger than $r \in \mathbb{N}$, if

$$M = \text{Span} \left\{ p_{I,j} ; \ (I, j) \in \mathcal{P} \times \{0, \ldots, r\} \right\} .$$

The linear dimension of M is then equal to $(r + 1) |\mathcal{P}|$.

Notice that models of histograms on the unit interval are exactly models of piecewise polynomial functions with degrees not larger than 0. In [36], it is shown that models of piecewise polynomial functions have nice analytical and statistical properties. Let us recall two of them.

In Lemma 8 of [36], it is proved that if the distribution P^X has a density with respect to the Lebesgue measure Leb on $\mathcal{X} = [0, 1]$ which is uniformly bounded away from zero and if the considered partition $\mathcal{P}$ is lower regular with respect to Leb - that is there exists a positive constant c such that $|\mathcal{P}| \inf_{I \in \mathcal{P}} \text{Leb} \left(I \right) \geq c > 0$ - then the associated model of piecewise polynomial functions is equipped

with a localized orthonormal basis in $L_2\left(P^X\right)$. For a formal definition of a localized basis, see Section 5 below. Since the pioneering work of Birgé and Massart [15, 17, 34], the property of localized basis is known to play a key role in M-estimation and model selection using vector spaces or more general sieves.

Considering models of piecewise polynomial functions on the unit interval, where the density of P^X with respect to Leb is both uniformly bounded and bounded away from 0 and where the underlying partition is lower regular with respect to Leb, it is shown in Lemma 9 of [36] that the least-squares estimator $s_n\left(M\right)$ converges in sup-norm to the linear projection s_M of the regression function s_*.

Assumptions of lower regularity of the considered partitions as well as the existence of a uniformly bounded density of P^X with respect to the Lebesgue measure on $\mathcal{X}$, will thus naturally arise when dealing with least-squares model selection using piecewise polynomial functions - see Section 3.2 below. Furthermore, the interested reader will find in Section 5 a more general version of our results, available for linear models equipped with a localized basis and where least-squares estimators converge in sup-norm to the linear projections of the regression function onto the models.

3. The slope heuristics

3.1. Underlying concepts

In order to clarify our approach and to highlight the connection of the present paper with the results previously established in [36], we first give a brief heuristic explanation of the major mathematical facts underlying the slope phenomenon.

We rewrite the definition of the oracle model M_* given in (5). For any $M \in \mathcal{M}_n$, the excess loss $\ell\left(s_*, s_n\left(M\right)\right) = P\left(Ks_n\left(M\right)\right) - P\left(Ks_*\right)$ is the difference between the loss of the estimator $s_n\left(M\right)$ and the loss of the target s_*. As $P\left(Ks_*\right)$ is independent of M varying in $\mathcal{M}_n$, it holds

$$M_* \in \arg\min_{M\in\mathcal{M}_n}\left\{P\left(Ks_n\left(M\right)\right)\right\}$$
$$= \arg\min_{M\in\mathcal{M}_n}\left\{P_n\left(Ks_n\left(M\right)\right) + \mathrm{pen}_{\mathrm{id}}\left(M\right)\right\} ,$$

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

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

The penalty function $\mathrm{pen}_{\mathrm{id}}$ is called the *ideal penalty* - as it allows to select the oracle - and is unknown because it depends on the distribution of the data. As pointed out by Arlot and Massart [9], the main idea of penalization in the efficiency problem is to give some sharp estimate, up to a constant, of the ideal penalty. This would yield an (asymptotically) unbiased - or uniformly biased over the collection of models $\mathcal{M}_n$ - estimation of the loss. Such a penalization would lead 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 leading constant converging to one when the sample size n tends to infinity.

Concerning the estimation of the optimal penalty, Arlot and Massart [9] conjectured that the mean of the empirical excess loss $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ satisfies the following slope heuristics in a quite general M-estimation framework:

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

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

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

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

is an optimal penalty.

The mean of the empirical excess loss 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 or, equivalently, the minimum value of penalty above which the procedure achieves an oracle inequality. 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 twice the minimal one,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}} \ . \tag{7}$$

Let us now briefly explain the points (i) and (ii) above. We give in Section 3.3 precise results which validate the slope heuristics for models of piecewise polynomial functions.

If the chosen penalty is less than the minimal one, $\mathrm{pen} = \left(1 - \delta\right)\mathrm{pen}_{\mathrm{min}}$ with $\delta \in \left[0, 1\right]$, the algorithm minimizes over $\mathcal{M}_n$,

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

In the latter identity, we neglect the difference between the empirical and true loss of the projections s_M and the deviations of the empirical excess loss

$P_n \left(K s_M - K s_n \left(M \right) \right)$. Indeed, as shown by Boucheron and Massart [20], the empirical excess loss 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.

As the empirical excess loss is increasing and the excess loss of the projection s_M is decreasing with respect to the complexity of the models, the penalized criterion is (almost) decreasing with respect to the complexity of the models, and the selected model is among the largest of the collection.

On the contrary, if the chosen penalty is greater than the minimal one, pen $=$ $\left(1 + \delta \right) \text{pen}_{\min}$ with $\delta > 0$, then by the same kind of manipulations, the selected model minimizes the following criterion, for all $M \in \mathcal{M}_n$,

$$P_n \left(K s_n \left(M \right) \right) + \text{pen} \left(M \right) - P_n \left(K s_* \right) \approx \ell \left(s_*, s_M \right) + \delta P_n \left(K s_M - K s_n \left(M \right) \right) . \tag{8}$$

The selected model thus achieves a trade-off between the bias of the models which decreases with the complexity and the empirical excess loss which increases with the complexity of the models. The selected dimension would then be 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 latter case, pen $= 2 \times \text{pen}_{\min}$, and if we assume that the empirical excess loss is equivalent to the excess loss,

$$P_n \left(K s_M - K s_n \left(M \right) \right) \sim P \left(K s_n \left(M \right) - K s_M \right) , \tag{9}$$

then according to (8) the selected model almost minimizes

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

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.

One can find in [36] some results showing that (9) is a quite general fact in least-squares regression and is in particular satisfied when considering models of piecewise polynomial functions. Thus, these results represent a preliminary material for the present study, and we shall base our arguments on the results exposed in [36].

3.2. Assumptions and comments

We take $\mathcal{X} = [0, 1]$, Leb is the Lebesgue measure on $\mathcal{X}$, and linear models $M \in \mathcal{M}_n$ are models of piecewise polynomial functions. We denote by $\mathcal{P}_M$ the partition of $\mathcal{X}$ underlying the model M.

Set of assumptions for piecewise polynomial functions: (SAPP)

(P1) there exist two positive constants $c_{\mathcal{M}}, \alpha_{\mathcal{M}}$ such that $\mathrm{Card}\,(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(P2) 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) there exist $c_{rich} > 0$, $A_{rich} > 0$ and $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in \left[n^{1/(1+\beta_+)}, c_{rich} n^{1/(1+\beta_+)} \right]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$, where β_+ is defined in $(\mathbf{Ap}_u)$.

($\mathbf{Ap}_u$) there exist $\beta_+ > 0$ and $C_+ > 0$ such that

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

(An) There exists a constant $\sigma_{\min}$ such that $\sigma\left(X_i\right) \geq \sigma_{\min} > 0$ a.s.

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

($\mathbf{Ad}_{\mathrm{Leb}}$) P^X has a density f with respect to Leb satisfying for some constants $c_{\min}$ and $c_{\max}$, that

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

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

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

(Alr) a positive constant $c_{\mathcal{M},\mathrm{Leb}}$ exists such that, for all $M \in \mathcal{M}_n$,

$$0 < c_{\mathcal{M},\mathrm{Leb}} \leq |\mathcal{P}_M| \inf_{I \in \mathcal{P}_M} \mathrm{Leb}\left(I\right) < +\infty \ .$$

The set of assumptions (**SAPP**) can be divided into three groups. Firstly, assumptions (**P1**), (**P2**), (**P3**) and ($\mathbf{Ap}_u$) are linked to properties of the collection of models $\mathcal{M}_n$. Secondly, assumptions (**An**), (**Ab**) and ($\mathbf{Ad}_{\mathrm{Leb}}$) give some constraints on the general regression relation stated in (1). Thirdly, assumptions (**Aud**) and (**Alr**) specify some quantities related to the choice of the models of piecewise polynomial functions.

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 design a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model, up to a constant. Indeed, as Talagrand's type concentration inequalities for the empirical process are exponential, 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 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 instance [16, 11].

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. One can notice that in practice, the parameter β_+, which depends on the

bias of the model is not known and so the existence of M_0 is not straightforward. However, it suffices for the statistician to take at least one model per dimension lower than the chosen upper bound to ensure the existence of M_0 and M_1.

We require in $(\mathbf{Ap}_u)$ for the quality of approximation of the collection of models to be good enough in terms of the quadratic loss. More precisely, we ask for a polynomial decrease of excess loss of linear projections of the regression function onto the models. It is well-known that piecewise polynomial functions uniformly bounded in their degrees have good approximation properties in Besov spaces. More precisely, as stated in Lemma 12 of Barron, Birgé and Massart [11], 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 [11]), then taking models of piecewise polynomial functions 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.

Assumption $(\mathbf{Ab})$ is rather restrictive, since it excludes Gaussian noise. However, the assumption of bounded noise is somehow classical when dealing with M-estimation and related procedures. Indeed, a central tool in this field is empirical process theory and more especially, concentration inequalities for the supremum of the empirical process. We used the classical inequalities of Bousquet, and Klein and Rio in [36]. As a matter of fact, we do not know yet if an adaptation of our proofs (including results established in [36]) by using extensions of the latter inequalities to some unbounded cases - as for instance in Adamczak's concentration inequalities [1] - would be possible.

The noise restriction stated in $(\mathbf{An})$ is needed to derive our results which are optimal to the first order. More precisely, it allows in [36] to obtain sharp lower bounds for the true and empirical excess losses on a fixed model. This assumption is also needed in the work of Arlot and Massart [9] concerning the case of histogram models. As it is noticed in Section 5.3 of [36], assumption $(\mathbf{An})$ could be replaced by the following assumption, which states that the partitions underlying the models of piecewise polynomial functions are regular from above with respect to the Lebesgue measure on $[0,1]$.

(Aur) a positive constant $c^+_{\mathcal{M},\mathrm{Leb}}$ exists such that, for all $M \in \mathcal{M}_n$,

$$|\mathcal{P}_M| \sup_{I \in \mathcal{P}_M} \mathrm{Leb}(I) \leq c^+_{\mathcal{M},\mathrm{Leb}} \ .$$

Assumptions $(\mathbf{Ad}_{\mathrm{Leb}})$, $(\mathbf{Aud})$ and $(\mathbf{Alr})$ imply several important properties for the models of piecewise polynomial functions, such as the existence of an orthonormal localized basis in each model or the consistency in sup-norm of least-squares estimators toward the projections of the target onto the models. See also Sections 2.2 and 5.1 for further comments about these properties.

3.3. Statement of the theorems

We are now able to state our main results leading to the slope heuristics. They describe the behavior of the penalization procedure defined in (6).

Theorem 1. *Take a positive penalty: for all $M \in \mathcal{M}_n$, pen $(M) \geq 0$. Suppose that the assumptions (**SAPP**) of Section 3.2 hold, and furthermore suppose that for $A_{\mathrm{pen}} \in [0,1)$ and $A_p > 0$ the model M_1 of assumption (**P3**) satisfies*

$$0 \leq \mathrm{pen}\,(M_1) \leq A_{\mathrm{pen}}\mathbb{E}\left[P_n\left(Ks_{M_1} - Ks_n\left(M_1\right)\right)\right] \ , \tag{10}$$

*with probability at least $1 - A_p n^{-2}$. Then there exist a constant $A_1 > 0$ only depending on constants in (**SAPP**), as well as an integer n_0 and a positive constant A_2 only depending on A_{pen} and on constants in (**SAPP**) such that, for all $n \geq n_0$, it holds with probability at least $1 - A_1 n^{-2}$,*

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{n^{\beta_+/(1+\beta_+)}}{\left(\ln n\right)^3} \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ , \tag{11}$$

*where $\beta_+ > 0$ is defined in assumption (**Ap$_u$**) of (**SAPP**).*

Theorem 1 justifies the first part (**i**) of the slope heuristics exposed in Section 3. 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 loss of the selected estimator is much larger than the excess loss of the oracle. Moreover, this level is given by the mean of the empirical excess loss of the least-squares estimator on each model. Let us also notice that the lower bound given in (11) gets worse as β_+ increases. This is due to the fact that when β_+ increases, the approximation properties of the models improve and the performances in terms of excess loss for the oracle estimator also improve.

The following theorem validates the second part of the slope heuristics.

Theorem 2. *Suppose that the assumptions (**SAPP**) of Section 3.2 hold, and furthermore suppose that for some $\delta \in [0,1)$ and $A_p, A_r > 0$, there exists an event of probability at least $1 - A_p n^{-2}$ on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^3$, it holds*

$$\left|\mathrm{pen}(M) - 2\mathbb{E}\left[P_n(Ks_M - Ks_n(M))\right]\right| \leq \delta\left(\ell(s_*, s_M) + \mathbb{E}\left[P_n(Ks_M - Ks_n(M))\right]\right) \tag{12}$$

together with

$$\left|\mathrm{pen}\left(M\right)\right| \leq A_r\left(\frac{\ell\left(s_*, s_M\right)}{\left(\ln n\right)^2} + \frac{\left(\ln n\right)^3}{n}\right) \ . \tag{13}$$

*Then, for any $\eta \in (0, \beta_+/\left(1+\beta_+\right))$, there exist an integer n_0 only depending on η, δ and β_+ and on constants in (**SAPP**), a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**SAPP**) and on A_p, two positive constants A_4 and A_5 only depending on constants in (**SAPP**) and on A_r and a sequence*

$$\theta_n \leq \frac{A_4}{\left(\ln n\right)^{1/4}} \tag{14}$$

such that it holds for all $n \geq n_0$, with probability at least $1 - A_3 n^{-2}$,

$$D_{\widehat{M}} \leq n^{\eta + 1/(1+\beta_+)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{\left(1-\delta\right)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) + A_5 \frac{\left(\ln n\right)^3}{n} \ . \quad (15)$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{\eta + 1/(1+\beta_+)} \quad (16)$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{\left(1-\delta\right)^2}\right) \ell\left(s_*, s_n\left(M_*\right)\right) \ . \quad (17)$$

Theorem 2 states that if the penalty is close to twice the minimal one, then the selected estimator satisfies a pathwise oracle inequality with constant almost one, and so the model selection procedure is approximately optimal. Moreover, the dimension of the selected model is of reasonable dimension, bounded by a power less than one of the sample size.

Condition (**Ap**) allows to remove the remainder terms from the oracle inequality (15) by ensuring that the selected model is of dimension not too small, as stated in (16). Assumption (**Ap**) is the conjunction of assumption (**Ap**$_u$) with a polynomial lower bound of the bias of the models. On histogram models, Arlot showed in Section 8.10 of [4] that this lower bound is satisfied for non constant α-Hölder, $\alpha \in (0, 1]$, regression functions and for regular partitions.

Finally, from Theorems 1 and 2, we identify the minimal penalty with the mean of the empirical excess loss on each model,

$$\mathrm{pen}_{\min}\left(M\right) = \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \ ,$$

thus generalizing the results of Arlot and Massart in [9] to the case of piecewise polynomial functions.

4. Hold-out penalization

The conditions on the penalty given in Theorems 1 and 2 can not be directly checked in practice. Indeed, they are expressed in terms of the mean of the empirical excess loss on each model, which is an unknown quantity in general.

Nevertheless, in the homoscedastic case, it is easy to see that Mallows' penalty is a nonasymptotic quasi-optimal penalty. According to Theorem 2, such a penalty is given by twice the mean of the empirical excess loss. Now, using Theorem 10 of [36], we get (with an explicit control of the second order terms in the following equivalence),

$$2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \sim \frac{1}{2}\mathcal{K}_{1,M}^2 \frac{D_M}{n} \ ,$$

where $\mathcal{K}_{1,M}^2 = 1/D_M \sum_{k=1}^{D_M} \mathbb{E}\left(\left(\psi_{1,M}\left(X,Y\right)\cdot\varphi_k\left(X\right)\right)^2\right)$, $\psi_{1,M}\left(X,Y\right) = -2\left(Y - s_M\left(X\right)\right)$ and $\left(\varphi_k\right)_{k=1}^{D_M}$ is an orthonormal basis in $\left(M, \|\cdot\|_2\right)$. By easy computations, we deduce that if the noise is homoscedastic, that is $\sigma^2\left(X\right) \equiv \sigma^2 > 0$, it holds

$$\frac{1}{2}\mathcal{K}_{1,M}^2 \frac{D_M}{n} = 2\sigma^2 \frac{D_M}{n} + \mathbb{E}\left[\left(s_* - s_M\right)^2 \frac{\sum_{i=1}^{D_M}\varphi_k^2}{n}\right] \ . \tag{18}$$

The second term at the right of identity (18) being negligible for models of interest in the conditions of Theorem 2 (thanks to Lemma 7 in [36], which implies that $\sum_{i=1}^{D_M}\varphi_k^2 \leq LD_M$ for some constant $L > 0$), we conclude that an asymptotically optimal penalty is given by $2\sigma^2 D_M/n$, which is Mallows' classical penalty.

In the case where the noise level is homoscedastic but unknown, Mallows' penalty is only known through a constant, the noise level, which can be estimated *via* the slope heuristics (for practical issues about the slope heuristics, see Baudry et al. [14]). But in the common situation where the noise level is sufficiently heteroscedastic, the shape of the ideal penalty is not linear in the dimension of the models and not even a *function* of the linear dimensions. In such a case, Arlot [7] proved that any calibration of a linear penalty leads to a suboptimal procedure, but yet can achieve an oracle inequality with a leading constant more than one.

In order to achieve a nearly optimal selection procedure in the general situation, it remains to estimate the ideal penalty or, thanks to the slope heuristics, the shape of the ideal penalty. This section is devoted to this task. We propose a hold-out type penalty that automatically adapts to heteroscedasticity. Let us now detail our hold-out penalization procedure.

The ideal penalty is defined by

$$\operatorname{pen}_{\mathrm{id}}\left(M\right) := P\left(Ks_n\left(M\right)\right) - P_n\left(Ks_n\left(M\right)\right) \ ,$$

for all $M \in \mathcal{M}_n$. A natural idea is to divide the data into two groups, indexed by I_1 and I_2, satisfying $I_1 \cap I_2 = \emptyset$ and $I_1 \cup I_2 = \{1, \ldots, n\}$ and to propose the following hold-out type penalty,

$$\operatorname{pen}_{ho,C}\left(M\right) := C\left(P_{n_2}\left(Ks_{n_1}\left(M\right)\right) - P_{n_1}\left(Ks_{n_1}\left(M\right)\right)\right) \ ,$$

where $P_{n_i} = 1/n_i \sum_{j\in I_i}\delta_{\xi_j}$, $n_i = \operatorname{Card}(I_i)$, for $i = 1, 2$, $s_{n_1}\left(M\right) \in \arg\min_{s\in M} P_{n_1}\left(Ks\right)$ and $C > 0$ is a constant to be determined. Indeed,

if n_1 is not too small, $P_{n_1}(Ks_{n_1}(M))$ is likely to vary like $P_n(Ks_n(M))$ and $P_{n_2}(Ks_{n_1}(M))$ is, conditionally to $(\xi_j)_{j \in I_1}$, an unbiased estimate of $P(Ks_{n_1}(M))$, which again is likely to vary like $P(Ks_n(M))$. Moreover, we see from Theorem 10 in [36] that when the model M is fixed, the quantities $P_n(Ks_n(M))$ and $P(Ks_n(M))$ are almost inversely proportional to n, so a good constant in front of the hold-out penalty should be $C_{opt} = n_1/n$.

The previous observation is justified by the following theorem, where for the sake of clarity we fixed $n_1 = n_2 = n/2$. For a more general version of Theorem 3, see Section 5.3. We set

$$\text{pen}_{ho}(M) = \frac{1}{2}(P_{n_2}(Ks_{n_1}(M)) - P_{n_1}(Ks_{n_1}(M))) \tag{19}$$

and

$$\widehat{M}_{1/2} \in \arg\min_{M \in \mathcal{M}_n}\{P_n(Ks_n(M)) + \text{pen}_{ho}(M)\}.$$

Theorem 3. *Consider the procedure defined in (19), with $n_1 = n_2 = n/2$. Suppose that the assumptions (**SAPP**) of Section 3.2 hold. Then, for any $\eta \in (0, \beta_+/(1+\beta_+))$, there exist an integer n_0 only depending on η and on constants in (**SAPP**), a positive constant A_6 only depending on $c_{\mathcal{M}}$ given in (**SAPP**), two positive constants A_7 and A_8 only depending on constants in (**SAPP**) and a sequence $\theta_n \leq A_7(\ln n)^{-1/4}$ such that it holds for all $n \geq n_0$, with probability at least $1 - A_6 n^{-2}$,*

$$D_{\widehat{M}_{1/2}} \leq n^{\eta + 1/(1+\beta_+)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{1/2}\right)\right) \leq (1 + \theta_n)\ell(s_*, s_n(M_*)) + A_8\frac{(\ln n)^3}{n}. \tag{20}$$

*Assume that in addition (**Ap**) holds (see Theorem 2). Then it holds for all $n \geq n_0((\mathbf{SAPP}), C_-, \beta_-, \eta)$, with probability at least $1 - A_6 n^{-2}$,*

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{1/2}\right)\right) \leq (1 + \theta_n)\inf_{M \in \mathcal{M}_n}\{\ell(s_*, s_n(M))\}. \tag{21}$$

Theorem 3 shows the asymptotic optimality of the hold-out penalization procedure, for a half-and-half split of the data. This is a remarkable fact compared to the classical hold-out, defined by

$$\widehat{M}_{ho} \in \arg\min_{M \in \mathcal{M}_n}\{P_{n_2}(Ks_{n_1}(M))\}. \tag{22}$$

Indeed, the choice $n_1 = n/2$ in (22) is likely to lead to an asymptotically suboptimal procedure, as the criterion is close in expectation to $P(Ks_{n/2}(M))$, and so is close to the oracle, but for $n/2$ data points. The hold-out penalization allows us to overcome this difficulty. Arlot [5, 6] described similar advantages for resampling and V-fold penalties.

Notice also that the random hold-out penalty proposed by Arlot [6] is proportional to the mean along the splits of our hold-out penalty, providing thus a "stabilization effect" in practice. This should bring some improvement compared to our unique split, at the price of increased computational cost. However, the stabilization effect seems more difficult to study mathematically, and our results provide a first step toward the study of the more complicated resampling penalties.

5. Proofs

We first present in Section 5.1 some "structural" properties of models, denoted **(GSA)**, that are sufficient for our needs and that are satisfied for models of piecewise polynomial functions considered in (**SAPP**). Then in Sections 5.2 and 5.3 respectively, we prove the results stated in Sections 3.3 and 4, for **(GSA)** instead of (**SAPP**).

5.1. A more general setting

General set of assumptions: (GSA)

Assume (**P1**), (**P2**), (**P3**), (**An**) and (**Ap**$_u$) of (**SAPP**). Furthermore suppose that,

(Ab') A positive constant A exists, such that for all $M \in \mathcal{M}_n$, $|Y_i| \leq A < \infty$, $\|s_M\|_\infty \leq A < \infty$.

(Alb) 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, for all $(\beta_k)_{k=1}^{D_M} \in \mathbb{R}^{D_M}$,

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

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

(Ac$_\infty$) a positive integer n_1 exists such that, for all $n \geq n_1$, there exist a positive constant A_{cons} and an event Ω_∞ of probability at least $1 - n^{-2-\alpha_\mathcal{M}}$, on which for all $M \in \mathcal{M}_n$,

$$\|s_n(M) - s_M\|_\infty \leq A_{cons} \sqrt{\frac{D_M \ln n}{n}} \ . \tag{23}$$

Notice that the covariate space $\mathcal{X}$ is general in (**GSA**). Let us explain how assumptions (**Ab'**), (**Ad**$_{\mathrm{Leb}}$), (**Aud**) and (**Alr**) of (**SAPP**) allow to recover (**Ab**), (**Alb**) and (**Ac**$_\infty$) of (**GSA**) in the special case of models of piecewise polynomial functions.

Assumption (**Ab'**) only differs from (**Ab**) by the fact that the projections of the target onto the models are uniformly bounded in sup-norm. In the general

case, this is indeed not guaranteed, but considering piecewise polynomial functions uniformly bounded in their degrees, this follows from simple computations (see Section 5.3 in [36]). Then, assumption (**Alb**) requires the existence of a localized orthonormal basis for each model. In the case of piecewise polynomial functions, this is ensured by (**Ad**$_{\text{Leb}}$), (**Aud**)and (**Alr**), see Lemma 8 of [36]. Finally, assumption (**Ac**$_\infty$) states the consistency of each estimator for the supnorm. Again, this is satisfied for models of piecewise polynomial functions under assumptions (**Ad**$_{\text{Leb}}$), (**Aud**) and (**Alr**). This result is established in Lemma 9 of [36].

Let us now describe a set of assumptions, less restrictive than (**SAPP**), that allows to recover (**GSA**) when considering histogram models. Lemma 5 and 6 of [36] allow to recover (**GSA**) from (**SAH**) for models of histograms.

Set of assumptions for histogram models: (SAH)

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**), (**Ab**) and (**Ap**$_u$) from (**SAPP**). Assume moreover,

(**Alrh**) there exists a positive constant $c^h_{\mathcal{M},P}$ such that,

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

Theorems 1 and 2 would also be valid when replacing the set of assumptions (**SAPP**) by (**SAH**). This would lead to the (almost exact) recovering of the assumptions and results described in Theorems 2 and 3 of [9], concerning the selection of least-squares estimators among histogram models.

5.2. Proofs related to Section 3.3

The following remark will be useful.

Remark 1. Since constants in (**GSA**) are uniform over the collection $\mathcal{M}_n$, we deduce from Theorem 2 of [36] applied with $\alpha = 2 + \alpha_{\mathcal{M}}$ and $A_- = A_+ = A_{\mathcal{M},+}$ that if assumptions (**P2**), (**Ab'**), (**An**), (**Alb**) and (**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 (**GSA**), such that for all $M \in \mathcal{M}_n$ satisfying

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

by setting

$$\varepsilon_n (M) = A_0 \max \left\{ \left(\frac{\ln n}{D_M} \right)^{1/4} ; \ \left(\frac{D_M \ln n}{n} \right)^{1/4} \right\} \tag{24}$$

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

$$\mathbb{P}\left[(1 - \varepsilon_n(M))\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P\left(Ks_n\left(M\right) - Ks_M\right) \leq (1 + \varepsilon_n\left(M\right))\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right]$$
$$\geq 1 - 10n^{-2-\alpha_{\mathcal{M}}} \tag{25}$$

and

$$\mathbb{P}\left[\left(1 - \varepsilon_n^2(M)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq \left(1 + \varepsilon_n^2(M)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2\right]$$
$$\geq 1 - 5n^{-2-\alpha_{\mathcal{M}}} \tag{26}$$

where $\mathcal{K}_{1,M}^2 = 1/D_M \sum_{k=1}^{D_M} \mathbb{E}\left((\psi_{1,M}(X,Y) \cdot \varphi_k(X))^2\right)$, $\psi_{1,M}(X,Y) = -2(Y - s_M(X))$ and $(\varphi_k)_{k=1}^{D_M}$ is an orthonormal basis in $(M, \|\cdot\|_2)$. Moreover, for all $M \in \mathcal{M}_n$, we have by Theorem 3 of [36], for a positive constant A_u depending on $A, A_{cons}, r_{\mathcal{M}}$ and $\alpha_{\mathcal{M}}$ and for all $n \geq n_0\left(A_{cons}, n_1\right)$,

$$\mathbb{P}\left[P\left(Ks_n\left(M\right) - Ks_M\right) \geq A_u\frac{D_M \vee \ln n}{n}\right] \leq 3n^{-2-\alpha_{\mathcal{M}}} \tag{27}$$

and

$$\mathbb{P}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right) \geq A_u\frac{D_M \vee \ln n}{n}\right] \leq 3n^{-2-\alpha_{\mathcal{M}}} . \tag{28}$$

Two technical lemmas are needed. In the first lemma, we intend to evaluate the minimal penalty $\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$ for models of dimension not too small.

Lemma 4. *Assume (P2), (Ab'), (An), (Alb) and (Ac$_\infty$) of (GSA). Then, for every model $M \in \mathcal{M}_n$ of dimension D_M such that*

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

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

$$\left(1 - L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] \tag{29}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2\left(M\right)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 , \tag{30}$$

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

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

$$\left(1 - \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n\left(Ks_M - Ks_n\left(M\right)\right) \leq \left(1 + \varepsilon_n\left(M\right)\right)\frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 , \tag{31}$$

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

$$0 \leq P_n(Ks_M - Ks_n(M)) \leq P_n Ks_M = \frac{1}{n} \sum_{i=1}^n (Y_i - s_M(X_I))^2 \leq 4A^2 \quad (32)$$

and as $D_M \geq 1$, we have

$$\varepsilon_n(M) = A_0 \max \left\{ \left(\frac{\ln n}{D_M} \right)^{1/4} ; \left(\frac{D_M \ln n}{n} \right)^{1/4} \right\} \geq A_0 n^{-1/8} . \quad (33)$$

We also have

$$\mathbb{E}[P_n(Ks_M - Ks_n(M))]$$
$$= \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{(\Omega_1(M))^c}\right] . \quad (34)$$

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 (32) and (33) that

$$\mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{(\Omega_1(M))^c}\right] \leq 20A^2 n^{-2-\alpha_{\mathcal{M}}} \leq \frac{80A^2}{A_0^2 \sigma_{\min}^2} \varepsilon_n^2(M) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 . \quad (35)$$

Moreover, we have $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0, A_{\mathcal{M},+}, A_{cons})$, so by (31),

$$0 < \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 - \varepsilon_n^2(M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n(Ks_M - Ks_n(M)) \mathbf{1}_{\Omega_1(M)}\right] \quad (36)$$

$$\leq \left(1 + \varepsilon_n^2(M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 . \quad (37)$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2} \varepsilon_n^2(M)$ by (33), we use (35), (36) and (37) in (34) 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 (29) and (30), as A_0 only depends on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$, A, $\sigma_{\min}$ and $r_{\mathcal{M}}$. $\qquad\square$

Lemma 5. *Let $\alpha > 0$. Assume that (**Ab'**) of (**GSA**) is satisfied. Then there exists a positive constant A_d, 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 $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(|\bar{\delta}(M)| \geq A_d \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \right) \leq 2n^{-\alpha} . \quad (38)$$

If moreover, assumptions (**P2**), (**An**), (**Alb**) and (**Ac**$_\infty$) of (**GSA**) 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

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + A_d \frac{\ln n}{\sqrt{D_M}} \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right) \leq 2n^{-\alpha} \;, \tag{39}$$

where $\mathrm{p}_2\left(M\right) := P_n\left(K s_M - K s_n\left(M\right)\right) \geq 0.$

Proof. We set

$$A_d = \max\left\{4A\sqrt{\alpha}; \; \frac{8A^2}{3}\alpha; \; \frac{8A^2\alpha}{\sqrt{A_{\mathcal{M},+}}\sigma_{\min}^2} + \frac{16A^2\alpha}{3A_{\mathcal{M},+}\sigma_{\min}}\right\} \;. \tag{40}$$

Since by (**Ab'**) we have $|Y| \leq A$ a.s. and $\|s_M\|_\infty \leq A$, it holds $\|s_*\|_\infty = \|\mathbb{E}\left[Y|X\right]\|_\infty \leq A$, and so $\|s_M - s_*\|_\infty \leq 2A.$ Next, we apply Bernstein's inequality (see Proposition 2.9 of [34]) to $\bar{\delta}\left(M\right) = \left(P_n - P\right)\left(K s_M - K s_*\right).$ Notice that

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

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

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

and therefore, by Bernstein's inequality we have for all $x > 0$,

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

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

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \sqrt{\frac{16A^2\alpha\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{8A^2\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \;, \tag{41}$$

which gives the first part of Lemma 5 for A_d given in (40). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and using it in (41) with $a = \ell\left(s_*, s_M\right)$, $b = \frac{4A^2\alpha\ln n}{n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}\left(M\right)\right| \geq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + \left(4\sqrt{D_M} + \frac{8}{3}\right)\frac{A^2\alpha\ln n}{n}\right) \leq 2n^{-\alpha} \;. \tag{42}$$

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

$$\left(1 - L_{A_{\mathcal{M},-}, A, \sigma_{\min}, r_{\mathcal{M}}, \alpha_{\mathcal{M}}} \varepsilon_n^2 (M)\right) \frac{D_M}{4n} \mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[p_2 (M)\right] \tag{43}$$

where $\varepsilon_n (M) = A_0 \max\left\{\left(\frac{\ln n}{D_M}\right)^{1/4}; \left(\frac{D_M \ln n}{n}\right)^{1/4}\right\}$. Moreover, as $D_M \leq A_{\mathcal{M},+} n (\ln n)^{-2}$ by (**P2**) 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 (43), $\mathbb{E}\left[p_2 (M)\right] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n}$ 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 (42), to conclude the proof for the value of A_d given in (40) by simple computations. $\square$

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 5.1, we will note $L_{(\mathbf{GSA})}$ and $n_0 (\mathbf{GSA})$. The values of these constants may change from line to line.

Proof of Theorem 2. From the definition of the selected model $\widehat{M}$ given in (6), $\widehat{M}$ minimizes

$$\mathrm{crit} (M) := P_n (Ks_n (M)) + \mathrm{pen} (M) , \tag{44}$$

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

$$\mathrm{crit}' (M) := \mathrm{crit} (M) - P_n (Ks_*) , \tag{45}$$

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

$$\mathrm{p}_1 (M) = P (Ks_n (M) - Ks_M) ,$$

$$\mathrm{p}_2 (M) = P_n (Ks_M - Ks_n (M)) ,$$

$$\bar{\delta} (M) = (P_n - P) (Ks_M - Ks_*)$$

and

$$\mathrm{pen}_{\mathrm{id}}' (M) = \mathrm{p}_1 (M) + \mathrm{p}_2 (M) - \bar{\delta} (M) ,$$

we have

$$\ell (s_*, s_n (M)) = P_n (Ks_n (M)) + \mathrm{p}_1 (M) + \mathrm{p}_2 (M) - \bar{\delta} (M) - P_n (Ks_*) \tag{46}$$

and by (45),

$$\text{crit}'(M) = \ell(s_*, s_n(M)) + (\text{pen}(M) - \text{pen}'_{\text{id}}(M)) \ . \tag{47}$$

As $\widehat{M}$ minimizes crit$'$ over $\mathcal{M}_n$, it is therefore sufficient by (47), to control $\text{pen}(M) - \text{pen}'_{\text{id}}(M)$ - or equivalently $\text{crit}'(M)$ - in terms of the excess loss $\ell(s_*, s_n(M))$, 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 $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$, (12) holds and

$$|\mathrm{p}_1(M) - \mathbb{E}[\mathrm{p}_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\mathbb{E}[\mathrm{p}_2(M)] \tag{48}$$

$$|\mathrm{p}_2(M) - \mathbb{E}[\mathrm{p}_2(M)]| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\mathbb{E}[\mathrm{p}_2(M)] \tag{49}$$

$$\left|\bar{\delta}(M)\right| \leq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + L_{(\mathbf{GSA})}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}[\mathrm{p}_2(M)] \tag{50}$$

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{51}$$

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $D_M \leq A_{\mathcal{M},+}(\ln n)^3$, (13) holds together with

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{52}$$

$$\mathrm{p}_2(M) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{(\ln n)^3}{n} \tag{53}$$

$$\mathrm{p}_1(M) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{(\ln n)^3}{n} \tag{54}$$

By (25), (26), (27) and (28) in Remark 1, Lemma 4, Lemma 5 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, and since (12) 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 (50) implies by (24) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$, for all $n \geq n_0(\mathbf{GSA})$,

$$\begin{aligned}
\left|\bar{\delta}(M)\right| &\leq L_{(\mathbf{GSA})}\left(\frac{(\ln n)^3}{D_M} \cdot \frac{\ln n}{D_M}\right)^{1/4} \times \mathbb{E}[\ell(s_*, s_M) + \mathrm{p}_2(M)] \\
&\leq L_{(\mathbf{GSA})}\varepsilon_n(M)\mathbb{E}[\ell(s_*, s_M) + \mathrm{p}_2(M)] \ ,
\end{aligned}$$

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

$|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)|$

$\leq |\mathrm{p}_1 (M) + \mathrm{p}_2 (M) - \mathrm{pen} (M)| + |\bar{\delta} (M)|$

$\leq |\mathrm{p}_1(M) + \mathrm{p}_2(M) - 2\mathbb{E} [\mathrm{p}_2(M)]| + (L_{(\mathbf{GSA})} \varepsilon_n(M) + \delta) \mathbb{E} [\ell (s_*, s_M) + \mathrm{p}_2(M)]$

$\leq (\delta + L_{(\mathbf{GSA})} \varepsilon_n (M)) \mathbb{E} [\ell (s_*, s_M) + \mathrm{p}_2 (M)]$. $\qquad (55)$

Now notice that using (**P2**) in (24) 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 $\ell (s_*, s_n (M)) = \ell (s_*, s_M) + \mathrm{p}_1 (M)$, we thus have on Ω_n, for all $n \geq n_0 (\mathbf{GSA})$,

$$0 \leq \mathbb{E} [\ell (s_*, s_M) + \mathrm{p}_2 (M)]$$
$$\leq \ell (s_*, s_n (M)) + |\mathrm{p}_1 (M) - \mathbb{E} [\mathrm{p}_2 (M)]|$$
$$\leq \ell (s_*, s_n (M)) + \frac{L_{(\mathbf{GSA})} \varepsilon_n (M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n (M)} \mathrm{p}_1 (M) \qquad \text{by (48)}$$
$$\leq \frac{1 + L_{(\mathbf{GSA})} \varepsilon_n (M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n (M)} \ell (s_*, s_n (M))$$
$$\leq (1 + L_{(\mathbf{GSA})} \varepsilon_n (M)) \ell (s_*, s_n (M)) . \qquad (56)$$

Hence, using (56) in (55), 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 $n \geq n_0 (\mathbf{GSA})$,

$$|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)| \leq (\delta + L_{(\mathbf{GSA})} \varepsilon_n (M)) \ell (s_*, s_n (M)) . \qquad (57)$$

Consequently, 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 (47) and (57),

$$(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n (M)) \ell (s_*, s_n (M)) \leq \mathrm{crit}' (M)$$
$$\leq (1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n (M)) \ell (s_*, s_n (M)) . \qquad (58)$$

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 (13), (52) and (53), it holds on Ω_n, for any $\tau > 0$ and for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$|\mathrm{pen}'_{\mathrm{id}} (M) - \mathrm{pen} (M)|$

$\leq \mathrm{p}_1 (M) + \mathrm{p}_2 (M) + |\mathrm{pen} (M)| + |\bar{\delta} (M)|$

$\leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} + A_r \frac{\ell (s_*, s_M)}{(\ln n)^2} + A_r \frac{(\ln n)^3}{n}$

$\quad + L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell (s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right)$

$$\leq L_{(\mathbf{GSA}),A_r} \left(\frac{(\ln n)^3}{n} + \frac{\ell\left(s_*, s_M\right)}{(\ln n)^2} \right) + \tau \ell\left(s_*, s_M\right) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n}$$

$$\leq L_{(\mathbf{GSA}),A_r} \left(\frac{(\ln n)^3}{n} + \frac{\ell\left(s_*, s_M\right)}{(\ln n)^2} \right) + \tau \ell\left(s_*, s_n\left(M\right)\right) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \; .$$
$$(59)$$

Hence, by taking $\tau = (\ln n)^{-2}$ in (59) 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,

$$\left| \mathrm{pen}'_{\mathrm{id}}\left(M\right) - \mathrm{pen}\left(M\right) \right| \leq L_{(\mathbf{GSA}),A_r} \left(\frac{\ell\left(s_*, s_n\left(M\right)\right)}{(\ln n)^2} + \frac{(\ln n)^3}{n} \right) \; . \qquad (60)$$

Moreover, by (47) and (60), we have on the event Ω_n, for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

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

$$\leq \left(1 + L_{(\mathbf{GSA}),A_r} (\ln n)^{-2}\right) \ell\left(s_*, s_n\left(M\right)\right) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \; . \qquad (62)$$

Oracle inequalities:

Recall that by the definition given in (5), an oracle model satisfies

$$M_* \in \arg \min_{M \in \mathcal{M}_n} \left\{ \ell\left(s_*, s_n\left(M\right)\right) \right\} \; . \qquad (63)$$

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 (75) and (77), we have on Ω_n, for any $\eta \in (0, \beta_+ / (1 + \beta_+))$ and for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta\right)$,

$$D_{\widehat{M}} \leq n^{1/(1+\beta_+)+\eta} \; , \qquad (64)$$

$$D_{M_*} \leq n^{1/(1+\beta_+)+\eta} \; . \qquad (65)$$

Now, from (64) we distinguish two cases in order to control $\mathrm{crit}'(\widehat{M})$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/(1+\beta_+)+\eta}$, we get by (58), for all $n \geq n_0\left(\mathbf{GSA}\right)$,

$$\mathrm{crit}'\left(\widehat{M}\right) \geq \left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n\left(\widehat{M}\right)\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) \; . \qquad (66)$$

Otherwise, if $D_{\widehat{M}} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (61),

$$\left(1 - L_{(\mathbf{GSA}),A_r} (\ln n)^{-2}\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}'\left(\widehat{M}\right) \; . \qquad (67)$$

Let us denote $\mathcal{S}_n = \left\{ M \in \mathcal{M}_n; \ A_{\mathcal{M},+} (\ln n)^3 \leq D_M \leq n^{1/(1+\beta_+)+\eta} \right\}$. In all cases, we have by (66) and (67), for all $n \geq n_0 \, (\mathbf{GSA})$,

$$
\mathrm{crit}' \left(\widehat{M} \right) \geq \left(1 - \delta - L_{(\mathbf{GSA}),A_r} \left((\ln n)^{-2} + \sup_{M \in \mathcal{S}_n} \varepsilon_n (M) \right) \right) \ell \left(s_*, s_n \left(\widehat{M} \right) \right)
$$

$$
- L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{68}
$$

Similarly, from (65) we distinguish two cases in order to control $\mathrm{crit}' \, (M_*)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{M_*} \leq n^{1/(1+\beta_+)+\eta}$, we get by (58), for all $n \geq n_0 \, (\mathbf{GSA})$,

$$
\mathrm{crit}' \, (M_*) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n (M_*) \right) \ell \, (s_*, s_n \, (M_*)) \ . \tag{69}
$$

Otherwise, if $D_{M_*} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (62),

$$
\mathrm{crit}' \, (M_*) \leq \left(1 + L_{(\mathbf{GSA}),A_r} (\ln n)^{-2} \right) \ell \, (s_*, s_n \, (M_*)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{70}
$$

In all cases, we deduce from (69) and (70) that we have for all $n \geq n_0 \, ((\mathbf{GSA}),\delta)$,

$$
\mathrm{crit}' \, (M_*) \leq \left(1 + \delta + L_{(\mathbf{GSA}),A_r} \left((\ln n)^{-2} + \sup_{M \in \mathcal{S}_n} \varepsilon_n (M) \right) \right) \ell \, (s_*, s_n \, (M_*))
$$

$$
+ L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{71}
$$

Hence, by setting

$$
\theta_n = L_{(\mathbf{GSA}),A_r} \left((\ln n)^{-2} + \sup_{M \in \mathcal{S}_n} \varepsilon_n (M) \right) \ ,
$$

we have by (24), for all $n \geq n_0 \, ((\mathbf{GSA}), \eta, \delta)$,

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

and we deduce from (68) and (71), since $\frac{1}{1-x} \leq 1 + 2x$ for all $x \in \left[0, \frac{1}{2} \right)$, that for all $n \geq n_0 \, ((\mathbf{GSA}), \eta, \delta)$, it holds on Ω_n,

$$
\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \leq \left(\frac{1 + \delta + \theta_n}{1 - \delta - \theta_n} \right) \ell \, (s_*, s_n \, (M_*)) + \frac{L_{(\mathbf{GSA}),A_r}}{1 - \delta - \theta_n} \frac{(\ln n)^3}{n}
$$

$$
\leq \left(\frac{1 + \delta}{1 - \delta} + \frac{5 \theta_n}{(1 - \delta)^2} \right) \ell \, (s_*, s_n \, (M_*)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \ . \tag{72}
$$

Inequality (15) 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 $\frac{1}{2} > \eta > (1 - \beta_+)_+ / 2$ and for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta, \delta\right)$, it holds on Ω_n,

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta} \ , \tag{73}$$

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{74}$$

Now, using (66) and (69), by the same kind of computations leading to (72), we deduce that it holds on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta, \delta\right)$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta+\theta_n}{1-\delta-\theta_n}\right) \ell\left(s_*, s_n\left(M_*\right)\right)$$

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

Thus inequality (17) is proved and Theorem 2 follows. $\qquad\square$

Lemma 6 (Control on the dimension of the selected model). *Assume that (GSA) holds. Let $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$. If $n \geq n_0\left((\mathbf{GSA}), \eta, \delta\right)$ then, on the event Ω_n defined in the proof of Theorem 2, we have*

$$D_{\widehat{M}} \leq n^{1/(1+\beta_+)+\eta} \ . \tag{75}$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{1/(1+\beta_+)+\eta} \ . \tag{76}$$

Lemma 7 (Control on the dimension of oracle models). *Assume that (GSA) holds. Let $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$. If $n \geq n_0\left((\mathbf{GSA}), \eta\right)$ then, on the event Ω_n defined in the proof of Theorem 2, we have*

$$D_{M_*} \leq n^{1/(1+\beta_+)+\eta} \ . \tag{77}$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{M_*} \leq n^{1/(1+\beta_+)+\eta} \ . \tag{78}$$

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

$$\mathrm{crit}'\left(M\right) = \mathrm{crit}\left(M\right) - P_n K s_* = \ell\left(s_*, s_M\right) - \mathrm{p}_2\left(M\right) + \bar{\delta}\left(M\right) + \mathrm{pen}\left(M\right) \tag{79}$$

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**) holds: let $M \in \mathcal{M}_n$ be such that $D_M < A_{\mathcal{M},+}\left(\ln n\right)^3$. By (13) and (79), it holds

$$\mathrm{crit}'\left(M\right) \geq \left(1 - \frac{A_r}{\left(\ln n\right)^2}\right) \ell\left(s_*, s_M\right) - \mathrm{p}_2\left(M\right) + \bar{\delta}\left(M\right) - A_r\frac{\left(\ln n\right)^3}{n} \ .$$

We then have on Ω_n,

$$
\begin{aligned}
&\ell\left(s_*, s_M\right) \geq C_- A_{\mathcal{M},+}^{-\beta_-}\left(\ln n\right)^{-3\beta_-} && \text{by } (\mathbf{Ap}) \\
&\mathrm{p}_2\left(M\right) \leq L_{(\mathbf{GSA})}\frac{\left(\ln n\right)^3}{n} && \text{from } (53) \\
&\bar{\delta}\left(M\right) \geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) && \text{from } (52).
\end{aligned}
$$

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

$$
\mathrm{crit}'\left(M\right) \geq \frac{C_- A_{\mathcal{M},+}^{-\beta_-}}{2}\left(\ln n\right)^{-3\beta_-} . \tag{80}
$$

2. Lower bound for large models: let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/(1+\beta_+)+\eta}$. From (12) and (49) we have on Ω_n, for all $n \geq n_0\left(A_{\mathcal{M},+}\right)$,

$$
\begin{aligned}
&\mathrm{pen}\left(M\right) - \mathrm{p}_2\left(M\right) \\
&\quad \geq \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] - \left(\delta + L_{(\mathbf{GSA})}\varepsilon_n^2\left(M\right)\right)\left(\ell\left(s_*, s_M\right) + \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right) .
\end{aligned}
$$

Using $(\mathbf{P2})$ and the fact that $D_M \geq n^{1/(1+\beta_+)+\eta}$ in (24), we deduce that for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta, \beta_+\right)$, $L_{(\mathbf{GSA})}\varepsilon_n^2\left(M\right) \leq \frac{1}{2}\left(1 - \delta\right)$ and as by $(\mathbf{An})$, $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$, we also deduce from Lemma 4 that for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$, $\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$. Consequently, it holds for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta, \beta_+\right)$,

$$
\begin{aligned}
\mathrm{pen}\left(M\right) - \mathrm{p}_2\left(M\right) &\geq \frac{\sigma_{\min}^2}{4}\left(1 - \delta\right)\frac{D_M}{n} - C_+ D_M^{-\beta_+} \\
&\geq \left(1 - \delta\right)L_{(\mathbf{GSA})}n^{-\frac{\beta_+}{1+\beta_+}+\eta} \tag{81}
\end{aligned}
$$

From (51) it holds on Ω_n,

$$
\begin{aligned}
\bar{\delta}\left(M\right) &\geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \\
&\geq -L_{(\mathbf{GSA})}\left(n^{-\frac{1+2\beta_+}{2(1+\beta_+)}}\sqrt{\ln n} + \frac{\ln n}{n}\right) . \tag{82}
\end{aligned}
$$

Hence, we deduce from (79), (81) and (82) that we have on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta, \beta_+\right)$,

$$
\mathrm{crit}'\left(M\right) \geq \left(1 - \delta\right)L_{(\mathbf{GSA})}n^{-\frac{\beta_+}{1+\beta_+}+\eta} . \tag{83}
$$

3. A better model exists for $\mathrm{crit}'\left(M\right)$: from $(\mathbf{P3})$, there exists $M_0 \in \mathcal{M}_n$ such that $n^{1/(1+\beta_+)} \leq D_{M_0} \leq c_{rich}n^{1/(1+\beta_+)}$. Then, for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$,

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

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

$$\ell\left(s_*, s_{M_0}\right) \leq C_+ n^{-\beta_+/(1+\beta_+)} . \tag{84}$$

By (50), we have on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$,

$$\left|\bar{\delta}(M_0)\right| \leq \frac{\ell(s_*, s_{M_0})}{\sqrt{D_{M_0}}} + L_{(\mathbf{GSA})} \frac{\ln n}{\sqrt{D_{M_0}}} \mathbb{E}\left[\mathrm{p}_2(M_0)\right] \leq L_{(\mathbf{GSA})} n^{-\frac{1+2\beta_+}{2(1+\beta_+)}} \ln(n) \tag{85}$$

and by (12),

$$\mathrm{pen}\left(M_0\right) \leq 3\left(\ell\left(s_*, s_{M_0}\right) + \mathbb{E}\left[\mathrm{p}_2\left(M_0\right)\right]\right) \leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} .$$

Consequently, we have on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$,

$$\begin{aligned}
\mathrm{crit}'\left(M_0\right) &\leq \ell\left(s_*, s_{M_0}\right) + \left|\bar{\delta}\left(M_0\right)\right| + \mathrm{pen}\left(M_0\right) \\
&\leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} .
\end{aligned} \tag{86}$$

To conclude, notice that the upper bound (86) is smaller than the lower bound given in (83) for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta\right)$. Hence, points 2 and 3 above yield inequality (75). Moreover, the upper bound (86) is smaller than lower bounds given in (80), derived by using $(\mathbf{Ap})$, and (83), for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta, \delta\right)$. This thus gives (76) and Lemma 6 is proved. □

Proof of Lemma 7. By definition, M_* minimizes

$$\ell\left(s_*, s_n\left(M\right)\right) = \ell\left(s_*, s_M\right) + \mathrm{p}_1\left(M\right)$$

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

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

$$\ell\left(s_*, s_n\left(M\right)\right) \geq \ell\left(s_*, s_M\right) \geq C_- A_{\mathcal{M},+}^{-\beta_-}\left(\ln n\right)^{-3\beta_-} \text{ by } (\mathbf{Ap}). \tag{87}$$

2. Lower bound of $\ell\left(s_*, s_n\left(M\right)\right)$ for large models: let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/(1+\beta_+)+\eta}$. From (48) we get on Ω_n,

$$\mathrm{p}_1\left(M\right) \geq \left(1 - L_{(\mathbf{GSA})} \varepsilon_n\left(M\right)\right) \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] .$$

Using $(\mathbf{P2})$ and the fact that $D_M \geq n^{1/(1+\beta_+)+\eta}$ in (24), we deduce that for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$, $L_{(\mathbf{GSA})} \varepsilon_n\left(M\right) \leq \frac{1}{2}$ and as by $(\mathbf{An})$, $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 4 that for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$, $\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n}$. Consequently, it holds for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$, on the event Ω_n,

$$\ell\left(s_*, s_n\left(M\right)\right) \geq \mathrm{p}_1\left(M\right) \geq \frac{\sigma_{\min}^2}{4} \frac{D_M}{n} \geq \frac{\sigma_{\min}^2}{4} n^{-\beta_+/(1+\beta_+)+\eta} . \tag{88}$$

3. A better model exists for $\ell\left(s_*, s_n\left(M\right)\right)$: from (**P3**), there exists $M_0 \in \mathcal{M}_n$ such that $n^{1/(1+\beta_+)} \leq D_{M_0} \leq c_{rich} n^{1/(1+\beta_+)}$. Moreover, for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$,

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

Using (**Ap**$_u$),

$$\ell\left(s_*, s_{M_0}\right) \leq C_+ n^{-\beta_+/(1+\beta_+)}$$

and by (48)

$$\mathrm{p}_1\left(M_0\right) \leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n\left(M\right)\right) \mathbb{E}\left[\mathrm{p}_2\left(M_0\right)\right] \ .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab'**) and as, by (24), 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 $n \geq n_0\left(\mathbf{GSA}\right)$, on the event Ω_n,

$$\mathrm{p}_1\left(M_0\right) \leq L_{(\mathbf{GSA})} \frac{D_M}{n} \leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} \ .$$

Consequently, on Ω_n, for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$,

$$\begin{aligned}
\ell\left(s_*, s_n\left(M_0\right)\right) &= \ell\left(s_*, s_{M_0}\right) + \mathrm{p}_1\left(M_0\right) \\
&\leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} \ .
\end{aligned} \tag{89}$$

The upper bound (89) is smaller than the lower bound (88) for all $n \geq n_0\left((\mathbf{GSA}), \eta\right)$, and this gives (77). If (**Ap**) holds, then the upper bound (89) is smaller than the lower bounds (87) and (88) for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \eta\right)$, which proves (78) and allows to conclude the proof of Lemma 7. $\qquad\square$

Proof of Theorem 1. As in 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 $n \geq n_0\left(\mathbf{GSA}\right)$, on which: (10) holds and

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

$$\left|\mathrm{p}_1\left(M\right) - \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right| \leq L_{(\mathbf{GSA})} \varepsilon_n\left(M\right) \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \ , \tag{90}$$

$$\left|\mathrm{p}_2\left(M\right) - \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right| \leq L_{(\mathbf{GSA})} \varepsilon_n^2\left(M\right) \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \ . \tag{91}$$

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

$$\mathrm{p}_2\left(M\right) \leq L_{(\mathbf{GSA})} \frac{\left(\ln n\right)^2}{n} \ . \tag{92}$$

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

$$\left|\bar{\delta}\left(M\right)\right| \leq L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell\left(s_*, s_M\right) \ln n}{n}} + \frac{\ln n}{n}\right) \ . \tag{93}$$

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

Lower bound on $D_{\widehat{M}}$. Let us recall that $\widehat{M}$ minimizes

$$\mathrm{crit}'\left(M\right) = \mathrm{crit}\left(M\right) - P_n K s_* = \ell\left(s_*, s_M\right) - \mathrm{p}_2\left(M\right) + \bar{\delta}\left(M\right) + \mathrm{pen}\left(M\right) \ . \quad (94)$$

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

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

We have

$$\ell\left(s_*, s_M\right) + \mathrm{pen}\left(M\right) \geq 0 \quad (95)$$

and from (93), as $\ell\left(s_*, s_M\right) \leq 4A^2$ by (**Ab'**), we get on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), d\right)$,

$$\begin{aligned}
\bar{\delta}\left(M\right) &\geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \\
&\geq -L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} \\
&\geq -d \times A^2 A_{rich}\left(\ln n\right)^{-2} \ . \quad (96)
\end{aligned}$$

Then, if $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^2$, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab'**) and as, by (24), for all $n \geq n_0\left(\mathbf{GSA}\right)$ it holds $L_{(\mathbf{GSA})}\varepsilon_n\left(M\right) \leq 1$, we deduce from (91) and Lemma 4 that for all $n \geq n_0\left((\mathbf{GSA}), d\right)$,

$$\mathrm{p}_2\left(M\right) \leq 2\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \leq 36A^2\frac{D_M}{n} \leq d \times 36A^2 A_{rich}\left(\ln n\right)^{-2} \ .$$

Whenever $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^2$, (92) gives that, for all $n \geq n_0\left((\mathbf{GSA}), d\right)$, on the event Ω'_n,

$$\mathrm{p}_2\left(M\right) \leq L_{(\mathbf{GSA})}\frac{\left(\ln n\right)^2}{n} \leq d \times 36A^2 A_{rich}\left(\ln n\right)^{-2} \ .$$

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

$$-\mathrm{p}_2\left(M\right) \geq -d \times 36A^2 A_{rich}\left(\ln n\right)^{-2} \ , \quad (97)$$

and finally, by using (95), (96) and (97) in (94), we deduce that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), d\right)$,

$$\mathrm{crit}'\left(M\right) \geq -d \times 37A^2 A_{rich}\left(\ln n\right)^{-2} \ . \quad (98)$$

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

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

We then have on Ω'_n,

$$
\begin{array}{ll}
\ell\left(s_*, s_{M_1}\right) \leq A_{rich}^{-\beta_+}\left(\ln n\right)^{2\beta_+} n^{-\beta_+} & \text{by } (\mathbf{Ap}_u) \\
\mathrm{p}_2\left(M_1\right) \geq \left(1 - L_{(\mathbf{GSA})}\varepsilon_n^2\left(M_1\right)\right)\mathbb{E}\left[\mathrm{p}_2\left(M_1\right)\right] & \text{by } (91) \\
\mathrm{pen}\left(M_1\right) \leq A_{\mathrm{pen}}\mathbb{E}\left[\mathrm{p}_2\left(M_1\right)\right] & \text{by } (10) \\
\left|\bar{\delta}\left(M_1\right)\right| \leq L_{(\mathbf{GSA})}\sqrt{\ln\left(n\right)/n} & \text{by } (93) \text{ and } (\mathbf{Ab'})
\end{array}
$$

and therefore,

$$
\begin{aligned}
\mathrm{crit}'\left(M_1\right) \quad \leq \quad & \left(-1 + A_{\mathrm{pen}} + L_{(\mathbf{GSA})}\varepsilon_n^2\left(M_1\right)\right)\mathbb{E}\left[\mathrm{p}_2\left(M_1\right)\right] \\
& + L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}} + A_{rich}^{-\beta_+}\frac{\left(\ln n\right)^{2\beta_+}}{n^{\beta_+}} \ . \quad (99)
\end{aligned}
$$

Hence, as $-1 + A_{\mathrm{pen}} < 0$, and as by (24), $(\mathbf{An})$ and Lemma 4 it holds for all $n \geq n_0\left((\mathbf{GSA}), A_{\mathrm{pen}}\right)$

$$
L_{(\mathbf{GSA})}\varepsilon_n^2\left(M_1\right) \leq \frac{1 - A_{\mathrm{pen}}}{2}
$$

and

$$
\mathbb{E}\left[\mathrm{p}_2\left(M_1\right)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n} \geq \frac{\sigma_{\min}^2 A_{rich}}{2}\left(\ln n\right)^{-2} \ ,
$$

we deduce from (99) that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), A_{\mathrm{pen}}\right)$,

$$
\mathrm{crit}'\left(M_1\right) \leq -\frac{1}{4}\left(1 - A_{\mathrm{pen}}\right)\sigma_{\min}^2 A_{rich}\left(\ln n\right)^{-2} \ . \quad (100)
$$

Now, by taking

$$
0 < d = \left(\frac{1}{149}\left(1 - A_{\mathrm{pen}}\right)\left(\frac{\sigma_{\min}}{A}\right)^2\right)\wedge\frac{1}{2} < 1 \quad (101)
$$

and by comparing (98) and (100), we deduce that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), A_{\mathrm{pen}}\right)$, for all $M \in \mathcal{M}_n$ such that $D_M \leq dA_{rich}n\left(\ln n\right)^{-2}$,

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

and so

$$
D_{\widehat{M}} > dA_{rich}n\left(\ln n\right)^{-2} \ . \quad (102)
$$

Excess Loss of $s_n\left(\widehat{M}\right)$. We take d with the value given in (101). 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 (24), $(\mathbf{P2})$, $(\mathbf{An})$ and Lemma 4, it holds on Ω'_n for all $n \geq n_0\left((\mathbf{GSA}), A_{\mathrm{pen}}\right)$, using (90),

$$
\ell\left(s_*, s_n\left(M\right)\right) \geq \mathrm{p}_1\left(M\right) \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n} \geq \frac{d\sigma_{\min}^2 A_{rich}}{2}\left(\ln n\right)^{-2} \ .
$$

By (102), we thus get that on Ω'_n, for all $n \geq n_0\left((\mathbf{GSA}), A_{\mathrm{pen}}\right)$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{d\sigma_{\min}^2 A_{rich}}{2}\left(\ln n\right)^{-2} . \tag{103}$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq n^{1/(1+\beta_+)} \leq D_{M_0} \leq c_{rich} n^{1/(1+\beta_+)}$$

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

$$\ell\left(s_*, s_{M_0}\right) \leq C_+ n^{-\beta_+/(1+\beta_+)} .$$

In addition, by (48),

$$\mathrm{p}_1\left(M\right) \leq \left(1 + L_{(\mathbf{GSA})}\varepsilon_n\left(M\right)\right)\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] .$$

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab'**) and as, by (24), for all $n \geq n_0\,(\mathbf{GSA})$ it holds $\varepsilon_n\left(M\right) \leq 1$, we deduce from Lemma 4 that for all $n \geq n_0\,(\mathbf{GSA})$,

$$\mathrm{p}_1\left(M\right) \leq L_{(\mathbf{GSA})}\frac{D_M}{n} \leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} .$$

Consequently, for all $n \geq n_0\,(\mathbf{GSA})$,

$$\ell\left(s_*, s_n\left(M_0\right)\right) \leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} \tag{104}$$

and the ratio between the two bounds (103) and (104) is larger than $n^{\beta_+/(1+\beta_+)}\left(\ln n\right)^{-3}$ for all $n \geq n_0\left(L_{(\mathbf{GSA})}, A_{\mathrm{pen}}\right)$, which yields (11). $\qquad\square$

5.3. Proofs related to Section 4

Theorem 3 is a straightforward consequence of the following result, that will be proved below.

Theorem 8. *Assume that (**GSA**) holds. With the notations of Section 4, assume moreover that there exist $c \in (0,1)$ such that $nc \leq n_1 < n$ and $\tau \in (1,3)$ satisfying $n\left(\ln n\right)^\tau/D_M \leq n_2 \leq n\left(1-c\right)$ for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M \leq A_{\mathcal{M},+}n/\left(\ln n\right)^2$. Take $n_2 = n\left(1-c\right)$ if $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. Define for all $M \in \mathcal{M}_n$,*

$$\mathrm{pen}_{ho}\left(M\right) = \frac{n_1}{n}\left(P_{n_2}\left(Ks_{n_1}\left(M\right)\right) - P_{n_1}\left(Ks_{n_1}\left(M\right)\right)\right) .$$

*Then, for any $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$, there exist an integer n_0 depending on c, η and on constants in (**GSA**), a positive constant A_6 only depending on $c_{\mathcal{M}}$ given in (**GSA**), two positive constants A_7 and A_8 only depending on constants in (**GSA**) and a sequence*

$$\theta_n \leq \frac{A_7}{\left(\ln n\right)^{1/4} \wedge \left(\ln n\right)^{(\tau-1)/2}}$$

such that it holds for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c, \eta\right)$, with probability at least $1 - A_6 n^{-2}$,

$$D_{\widehat{M}_{n_1}} \leq n^{\eta + 1/(1 + \beta_+)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(1 + \theta_n\right) \ell\left(s_*, s_n\left(M_*\right)\right) + A_8 \frac{\left(\ln n\right)^3}{n} \ . \tag{105}$$

*Assume that in addition (**Ap**) holds (see Theorem [2]). Then it holds for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), C_-, \beta_-, \eta, c\right)$, with probability at least $1 - A_6 n^{-2}$,*

$$A_{\mathcal{M}, +}\left(\ln n\right)^3 \leq D_{\widehat{M}_{n_1}} \leq n^{\eta + 1/(1 + \beta_+)}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(1 + \theta_n\right) \inf_{M \in \mathcal{M}_n} \left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{106}$$

Lemma 9. *Assume that (**GSA**) holds. Let $c \in (0, 1)$, $\tau \in (1, 3)$ and $(n_1, n_2) \in \mathbb{N}_*^2$. We assume that $nc \leq n_1 < n$ and set $n_2 = n - n_1$. Then there exists $L = L_{(\boldsymbol{GSA}), c} > 0$ such that for all $M \in \mathcal{M}_n$ satisfying $D_M \geq A_{\mathcal{M}, +}\left(\ln n\right)^2$, for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c\right)$, it holds with probability at least $1 - 12n^{-2 - \alpha_{\mathcal{M}}}$,*

$$\left|P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) - P\left(Ks_{n_1}(M) - Ks_M\right)\right|$$
$$\leq L \frac{\sqrt{\left(D_M \vee \ln n\right)\left(\ln n\right)\left(\left(\ln n\right)\left(\ln n_1\right) + n_2\right)}}{n_2 \sqrt{n_1}} \tag{107}$$

Now, let us assume that $n\left(\ln n\right)^\tau / D_M \leq n_2 \leq n\left(1 - c\right)$ if $A_{\mathcal{M}, +}\left(\ln n\right)^3 \leq D_M \leq A_{\mathcal{M}, +} n / \left(\ln n\right)^2$ and $n_2 = n\left(1 - c\right)$ if $D_M \leq A_{\mathcal{M}, +}\left(\ln n\right)^3$. If $A_{\mathcal{M}, +}\left(\ln n\right)^3 \leq D_M$, then by setting

$$\varepsilon_n^{1,2}(M) = L \frac{n\sqrt{\ln n\left(\left(\ln n\right)\left(\ln n_1\right) + n_2\right)}}{n_2 \sqrt{n_1 D_M}} \leq \frac{L}{\left(\ln n\right)^{(\tau - 1)/2}} \ , \tag{108}$$

we have for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c\right)$, with probability at least $1 - 12n^{-2 - \alpha_{\mathcal{M}}}$,

$$\left|P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) - P\left(Ks_{n_1}(M) - Ks_M\right)\right| \leq \varepsilon_n^{1,2}(M)\, \mathbb{E}\left[p_2(M)\right] \ . \tag{109}$$

If $D_M \leq A_{\mathcal{M}, +}\left(\ln n\right)^3$, we obtain

$$\mathbb{P}\left(\left|P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) - P\left(Ks_{n_1}(M) - Ks_M\right)\right| \geq L \frac{\left(\ln n\right)^2}{n}\right) \leq 12n^{-2 - \alpha_{\mathcal{M}}}. \tag{110}$$

Proof. By Bernstein's inequality (see Corollary 2.10 in [34]) applied to the sum of $\left(s_{n_1}(M)\right)\left(\xi_i\right)$ conditionally to $\left(\xi_j\right)_{j \in I_1}$, we get that for all $x > 0$, it holds

$$\mathbb{P}\left(\left|P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) - P\left(Ks_{n_1}(M) - Ks_M\right)\right| \geq x \,\middle|\, \left(\xi_j\right), \ j \in I_1\right)$$
$$\leq 2\exp\left(-\frac{nx^2}{2\left(v_1 + b_1 x/3\right)}\right) \ , \tag{111}$$

where
$$v_1 = \mathbb{E}_\xi \left[\left(K s_{n_1}(M)(\xi) - K s_M(\xi) \right)^2 \right]$$

and $b_1 = \| K s_{n_1}(M) - K s_M \|_\infty$. We have

$$
\begin{aligned}
v_1 &= \mathbb{E} \left[\left(2Y - s_M(X) - s_{n_1}(M)(X) \right)^2 \left(s_{n_1}(M)(X) - s_M(X) \right)^2 \right] \\
&\leq \left(4A + \| s_{n_1}(M) - s_M \|_\infty \right)^2 \mathbb{E}_X \left[\left(s_{n_1}(M)(X) - s_M(X) \right)^2 \right] \\
&= \left(4A + \| s_{n_1}(M) - s_M \|_\infty \right)^2 P \left(K s_{n_1}(M) - K s_M \right) \qquad (112)
\end{aligned}
$$

and

$$
\begin{aligned}
b_1 &= \left\| \left(2 \left(Y - s_M(X) \right) - s_{n_1}(M)(X) + s_M(X) \right) \left(s_{n_1}(M)(X) - s_M(X) \right) \right\|_\infty \\
&\leq 4A \| s_{n_1}(M) - s_M \|_\infty + \| s_{n_1}(M) - s_M \|_\infty^2 . \qquad (113)
\end{aligned}
$$

Now, we set $\Omega_v = \{ v_1 \leq L_v \left(D_M \vee \ln n_1 \right) / n_1 \}$ and $\Omega_b = \{ b_1 \leq L_b \sqrt{D_M \ln n_1 / n_1} \}$. By integrating (111), it comes for all $x > 0$,

$$
\begin{aligned}
&\mathbb{P} \left(\left| P_{n_2} \left(K s_{n_1}(M) - K s_M \right) - P \left(K s_{n_1}(M) - K s_M \right) \right| \geq x \right) \\
&\leq 2\mathbb{E} \left[\exp \left(- \frac{n_2 x^2}{2 \left(v_1 + b_1 x / 3 \right)} \right) \mathbf{1}_{\Omega_v \cap \Omega_b} \right] + 2\mathbb{P} \left(\Omega_v^c \right) + 2\mathbb{P} \left(\Omega_b^c \right) \\
&\leq 2 \exp \left(- \frac{n_2 x^2}{2 \left(L_v \left(D_M \vee \ln n_1 \right) / n_1 + L_b x \sqrt{D_M \ln n_1 / n_1} \right)} \right) \\
&\quad + 2\mathbb{P} \left(\Omega_v^c \right) + 2\mathbb{P} \left(\Omega_b^c \right)
\end{aligned}
$$

From assumption $(\mathbf{Ac}_\infty)$ and inequality (27), it is possible to choose L_v and L_b, depending among other constants on c, such that for all $n \geq n_0 \left((\mathbf{GSA}), c \right)$, $2\mathbb{P} \left(\Omega_v^c \right) + 2\mathbb{P} \left(\Omega_b^c \right) \leq 10 n^{-2 - \alpha_\mathcal{M}}$. Thus, we get for $L > 0$ large enough and for all $x > 0$,

$$
\begin{aligned}
&\mathbb{P} \left(\left| P_{n_2} \left(K s_{n_1}(M) - K s_M \right) - P \left(K s_{n_1}(M) - K s_M \right) \right| \geq x \right) \\
&\leq 2 \exp \left(- \frac{n_2 x^2}{L \left(\left(D_M \vee \ln n_1 \right) / n_1 + x \sqrt{D_M \ln n_1 / n_1} \right)} \right) + 10 n^{-2 - \alpha_\mathcal{M}} . \quad (114)
\end{aligned}
$$

By taking $x = \sqrt{L \alpha \ln n \left(D_M \vee \ln n_1 \right) \left(L \alpha \left(\ln n \right) \left(\ln n_1 \right) + 4 n_2 \right)} / \left(n_2 \sqrt{n_1} \right) > 0$ in the latter inequality, it comes, with probability at least $1 - 12 n^{-2 - \alpha_\mathcal{M}}$,

$$
\begin{aligned}
&\left| P_{n_2} \left(K s_{n_1}(M) - K s_M \right) - P \left(K s_{n_1}(M) - K s_M \right) \right| \\
&\leq L \frac{\sqrt{\left(D_M \vee \ln n_1 \right) \left(\ln n \right) \left(\left(\ln n \right) \left(\ln n_1 \right) + n_2 \right)}}{n_2 \sqrt{n_1}} ,
\end{aligned}
$$

where $L > 0$ depends on the constants in $(\mathbf{GSA})$ and on c. Inequalities (109) and (110) then follow from simple calculations. $\qquad \square$

Remark 2. It is easy to see that by using the assumption of consistency in sup-norm for a fixed model, stated as (**H5**) in [36], instead of (**Ac**$_\infty$) and by using Theorem 4 of [36] instead of inequality (27), the results established in Lemma 9 are valid with probability bounds proportional to $n^{-\alpha}$, for any $\alpha > 0$ (in Lemma 9, we only derive the case $\alpha = 2 + \alpha_\mathcal{M}$ for convenience).

Proof of Theorem 8. We set $\mathrm{pen}_0(M) = \mathrm{pen}_{ho}(M) - (n_1/n) \cdot (P_{n_2}(Ks_*) - P_{n_1}(Ks_*))$. It is worth noting that $P_{n_2}(Ks_*) - P_{n_1}(Ks_*)$ is a quantity independent of M, when M varies in $\mathcal{M}_n$. Hence, the procedure defined by pen_0 gives the same result as the hold-out procedure defined by pen_{ho}. It will be convenient for our analysis to consider pen_0 instead of pen_{ho}. As a matter of fact, we derive Theorem 8 as a corollary of Theorem 2 applied with $\mathrm{pen} \equiv \mathrm{pen}_0$, through the use of Lemma 9.

We get for all $M \in \mathcal{M}_n$,

$$
\begin{aligned}
\mathrm{pen}_0(M) &= \frac{n_1}{n}\left(P_{n_2}\left(Ks_{n_1}(M) - Ks_*\right) - P_{n_1}\left(Ks_{n_1}(M) - Ks_*\right)\right) \\
&= \frac{n_1}{n}\left(P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) - P_{n_1}\left(Ks_{n_1}(M) - Ks_M\right)\right) \\
&\quad + \frac{n_1}{n}\left(\left(P_{n_2} - P\right)\left(Ks_M - Ks_*\right) - \left(P_{n_1} - P\right)\left(Ks_M - Ks_*\right)\right) \\
&= \frac{n_1}{n}\left(\mathrm{p}_1^{n_2}(M) + \mathrm{p}_2^{n_1}(M) + \bar{\delta}^{n_2}(M) - \bar{\delta}^{n_1}(M)\right)
\end{aligned}
$$

where

$$
\mathrm{p}_1^{n_2}(M) = P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) , \ \mathrm{p}_2^{n_1}(M) = P_{n_1}\left(Ks_M - Ks_{n_1}(M)\right)
$$

and

$$
\bar{\delta}^{n_i}(M) = \left(P_{n_i} - P\right)\left(Ks_M - Ks_*\right) .
$$

Let Ω_n be the event on which:

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

$$
|\mathrm{p}_1(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{115}
$$

$$
|\mathrm{p}_2(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{116}
$$

together with

$$
\left|\mathrm{p}_1^{n_2}(M) - \frac{n}{n_1}\mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA}),c}\left[\varepsilon_n^{1,2}(M) + \varepsilon_n(M)\right]\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{117}
$$

$$
\left|\mathrm{p}_2^{n_1}(M) - \frac{n}{n_1}\mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA}),c}\varepsilon_n^2(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{118}
$$

$$
\left|\bar{\delta}^{n_1}(M)\right| \leq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + L_{(\mathbf{GSA}),c}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{119}
$$

$$
\left|\bar{\delta}^{n_2}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n_2}{n_2}} + \frac{\ln n_2}{n_2}\right) \tag{120}
$$

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

$$\left| \bar{\delta}^{n_1} \left(M \right) \right| \leq L_{(\mathbf{GSA}),c} \left(\sqrt{\frac{\ell \left(s_*, s_M \right) \ln n}{n}} + \frac{\ln n}{n} \right) \tag{121}$$

$$\left| \bar{\delta}^{n_2} \left(M \right) \right| \leq L_{(\mathbf{GSA}),c} \left(\sqrt{\frac{\ell \left(s_*, s_M \right) \ln n}{n}} + \frac{\ln n}{n} \right) \tag{122}$$

$$\mathrm{p}_2^{n_1} \left(M \right) \leq L_{(\mathbf{GSA}),c} \frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA}),c} \frac{\left(\ln n \right)^3}{n} \tag{123}$$

$$\mathrm{p}_1^{n_2} \left(M \right) \leq L_{(\mathbf{GSA}),c} \frac{\left(\ln n \right)^3}{n} \tag{124}$$

By (25), (26), (27) and (28) in Remark 1, Lemma 4 and Lemma 9, we get for all $n \geq n_0 \left((\mathbf{GSA}), c \right)$,

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

We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} \left(\ln n \right)^3 \leq D_M$. Notice that (119) implies by (24) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} \left(\ln n \right)^3 \leq D_M$,

$$\left| \bar{\delta}^{n_1} \left(M \right) \right| \leq L_{(\mathbf{GSA}),c} \left(\frac{\left(\ln n \right)^3}{D_M} \cdot \frac{\ln n}{D_M} \right)^{1/4} \times \left(\ell \left(s_*, s_M \right) + \mathbb{E} \left[p_2 \left(M \right) \right] \right)$$

$$\leq L_{(\mathbf{GSA}),c} \varepsilon_n \left(M \right) \left(\ell \left(s_*, s_M \right) + \mathbb{E} \left[p_2 \left(M \right) \right] \right) .$$

In addition, from (120), Lemma 4 and the fact that $n \left(\ln n \right)^\tau / D_M \leq n_2$, we get that for all $n \geq n_0 \left(\mathbf{GSA} \right)$,

$$\left| \bar{\delta}^{n_2} \left(M \right) \right| \leq L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell \left(s_*, s_M \right) \ln n_2}{n_2}} + \frac{\ln n_2}{n_2} \right)$$

$$\leq L_{(\mathbf{GSA})} \left(\frac{\ell \left(s_*, s_M \right)}{\left(\ln n \right)^{(\tau-1)/2}} + \frac{\ln n_2}{n_2} \left(\ln n \right)^{(\tau-1)/2} \right)$$

$$\leq L_{(\mathbf{GSA})} \left(\ln n \right)^{(1-\tau)/2} \left(\ell \left(s_*, s_M \right) + \mathbb{E} \left[p_2 \left(M \right) \right] \right) .$$

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

$$\left| \mathrm{pen}_0 \left(M \right) - 2\mathbb{E} \left[p_2 \left(M \right) \right] \right|$$

$$\leq \frac{n_1}{n} \left(\left| \mathrm{p}_1^{n_2} \left(M \right) - \frac{n}{n_1} \mathbb{E} \left[p_2 \left(M \right) \right] \right| + \left| \mathrm{p}_2^{n_1} \left(M \right) - \frac{n}{n_1} \mathbb{E} \left[p_2 \left(M \right) \right] \right| \right)$$

$$+ \left| \bar{\delta}^{n_1} \left(M \right) \right| + \left| \bar{\delta}^{n_2} \left(M \right) \right|$$

$$\leq L_{(\mathbf{GSA}),c} \left(\varepsilon_n^{1,2} \left(M \right) + \varepsilon_n \left(M \right) + \left(\ln n \right)^{(1-\tau)/2} \right) \left(\ell \left(s_*, s_M \right) + \mathbb{E} \left[p_2 \left(M \right) \right] \right)$$

$$\tag{125}$$

Hence, inequality (12) of Theorem 2 is satisfied on Ω_n by taking

$$\delta = L_{(\mathbf{GSA}),c}\left(\varepsilon_n^{1,2}\left(M\right) + \varepsilon_n\left(M\right) + \left(\ln n\right)^{(1-\tau)/2}\right) .$$

Moreover, we have $\delta \in [0,1)$ for all $n \geq n_0\left((\mathbf{GSA}),c,\tau\right)$.

Let us now consider models $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. By (121), (122), (124) and (123), we have on Ω_n,

$$
\begin{aligned}
\left|\mathrm{pen}_0\left(M\right)\right| &= \frac{n_1}{n}\left|\mathrm{p}_1^{n_2}\left(M\right) + \mathrm{p}_2^{n_1}\left(M\right) + \bar{\delta}^{n_2}\left(M\right) - \bar{\delta}^{n_1}\left(M\right)\right| \\
&\leq L_{(\mathbf{GSA}),c}\left(\sqrt{\frac{\ell\left(s_*,s_M\right)\ln n}{n}} + \frac{\left(\ln n\right)^3}{n}\right) \\
&\leq L_{(\mathbf{GSA}),c}\left(\frac{\ell\left(s_*,s_M\right)}{\left(\ln n\right)^2} + \frac{\left(\ln n\right)^3}{n}\right) \quad (126)
\end{aligned}
$$

Inequality (126) implies that inequality (13) of Theorem 2 is satisfied with $A_r = L_{(\mathbf{GSA}),c}$. From (125) and (126), we thus apply Theorem 2 with $A_p = L_{A_p,c_{\mathcal{M}}}$, and this gives Theorem 8 with

$$
\begin{aligned}
\theta_n &= L_{(\mathbf{GSA}),c}\left(\left(\ln n\right)^{-2} + \left(\ln n\right)^{(1-\tau)/2}\right. \\
&\quad \left. + \sup_{M \in \mathcal{M}_n}\left\{\varepsilon_n\left(M\right) + \varepsilon_n^{1,2}\left(M\right),\; A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M \leq n^{\eta+1/(1+\beta_+)}\right\}\right) .
\end{aligned}
$$

$\square$

Acknowledgements

References

[1] R. ADAMCZAK. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:1000–1034, 2008. MR2424985

[2] H. AKAIKE. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970. MR0286233

[3] 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. MR0483125

[4] S. ARLOT. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[5] S. ARLOT. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[6] S. ARLOT. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009. MR2519533

[7] S. ARLOT. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv:0812.3141.

[8] S. ARLOT AND F. BACH. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[9] S. ARLOT AND P. MASSART. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[10] Y. BARAUD, C. GIRAUD, AND S. HUET. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009. MR2502646

[11] A. BARRON, L. BIRGÉ, AND P. MASSART. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. MR1679028

[12] P.L. BARTLETT, S. BOUCHERON, AND G. LUGOSI. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[13] P.L. BARTLETT, O. BOUSQUET, AND S. MENDELSON. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005. MR2166554

[14] J.-P. BAUDRY, C. MAUGIS, AND B. MICHEL. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012. MR2865029

[15] L. BIRGÉ AND P. MASSART. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993. MR1240719

[16] L. BIRGÉ AND P. MASSART. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997. MR1462939

[17] L. BIRGÉ AND P. MASSART. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998. MR1653272

[18] L. BIRGÉ AND P. MASSART. Gaussian model selection. *J.Eur.Math.Soc.*, 3(3):203–268, 2001. MR1848946

[19] L. BIRGÉ AND P. MASSART. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. MR2288064

[20] S. BOUCHERON AND P. MASSART. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011. MR2824862

[21] F. BUNEA, A. B. TSYBAKOV, AND M. H. WEGKAMP. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007. MR2351101

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

[23] O. CATONI. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920

[24] A. S. DALALYAN AND J. SALMON. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 2012.

[25] A. S. DALALYAN AND A. B. TSYBAKOV. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007. MR2397581

[26] B. EFRON. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983. MR0711106

[27] V. KOLTCHINSKII. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001. MR1842526

[28] V. KOLTCHINSKII. Local Rademacher complexities and oracle inequalities in risk minimisation. *Ann. Statist.*, 34(6):2593–2656, 2006. MR2329442

[29] G. LECUÉ AND S. MENDELSON. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009. MR2529440

[30] M. LERASLE. Optimal model selection for density estimation of stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011. MR2893855

[31] M. LERASLE. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908, 2012. MR2976568

[32] G. LEUNG AND A. R. BARRON. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. MR2242356

[33] COLIN L. MALLOWS. Some comments on C$_p$. *Technometrics*, 15:661–675, 1973.

[34] P. MASSART. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

[35] P. RIGOLLET AND A.B. TSYBAKOV. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

[36] A. SAUMARD. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6(1-2):579–655, 2012. MR2988421

[37] A. B. TSYBAKOV. *Introduction à l'estimation non-paramétrique.* Springer-Verlag, Berlin, 1996.