Optimal model selection in heteroscedastic regression using piecewise polynomials

Adrien Saumard

Optimal model selection in heteroscedastic regression using piecewise polynomials

A. Saumard
Department of Statistics, University of Washington, Seattle, WA 98195, USA
INRIA Saclay Île-de-France, France

January 17, 2013

Abstract

We consider the estimation of a regression function with random design and heteroscedastic noise in a non-parametric setting. More precisely, we address the problem of characterizing the optimal penalty when the regression function is estimated by using a penalized least-squares model selection method. In this context, we show the existence of a minimal penalty, defined to be the maximum level of penalization under which the model selection procedure totally misbehaves. 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. When the shape of the optimal penalty is known, this allows to apply the so-called slope heuristics initially proposed by Birgé and Massart [19], which further provides with a data-driven calibration of penalty procedure. The use of the results obtained by the author in [40], considering the least-squares estimation of a regression function on a fixed finite-dimensional linear model, allows us to go beyond the case of histogram models, which is already treated by Arlot and Massart in [7]. Finally, in cases where the shape of the ideal penalty is unknown, we propose a hold-out penalization procedure, proved to be asymptotically optimal under mild conditions on the data split.

Keywords: Non-parametric regression, Heteroscedastic noise, Random design, Optimal model selection, Slope heuristics, data-driven penalty.

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 [37], 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 [37], 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 to be 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 [38], various statistical situations require to let these quantities depend on the amount of data. Thus, pointing out the importance of Talagrand's type concentration inequalities in the non-asymptotic approach, Birgé and Massart [17], [18] and Barron, Birgé and Massart [12] have been able to build non-asymptotic 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 [31], Bartlett & *al.* [13], to the refined local Rademacher complexities in the case where some favorable noise conditions hold (see for instance Bartlett, Bousquet and Mendelson [14], Koltchinskii [32]). But as a prize to pay for generality, the above penalties

suffer from their dependence on unknown or unrealistic constants. They are very difficult to implement and calibrate in practice and satisfy oracle inequalities with possibly huge leading constants. Other general-purpose penalties have been proposed, such as the bootstrap penalties of Efron [28] and the resampling and V-fold penalties of Arlot [5] and [8]. These penalties are essentially resampling estimates of the difference between the empirical risk and the risk. They can be used in practice since, in particular, they avoid the practical drawbacks of the local Rademacher complexities. Arlot [5], [8] 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 conjectures that the restriction on histograms is mainly technical and that his results can be extended to more general situations.

In this paper, we address the problem of optimal model selection, in a bounded regression setting with heteroscedastic noise and random design. A penalty will be said to be optimal if it achieves a non-asymptotic oracle inequality with leading constant almost one, i.e. converging to one when the sample size tends to infinity. In the following, we restrict ourselves to "small" collections of models, where the number of models is not more than polynomial in the sample size, a case where such an optimal penalty can exist. In more general settings, when the collection of models can be large, one should gather the models of equal or equivalent complexity and derive an oracle inequality with respect to the infimum of the risk on the union of models with the same complexities, as explained in Birgé and Massart [19]. This would allow to consider optimal penalties for large collections of models, but this problem is beyond the scope of this paper.

It is also worth noticing that by "non-asymptotic oracle inequality" we essentially mean that the complexity of the collection of models as well as their dimensions are allowed to depend on the sample size n. Of course n fixed and dependencies on n in the residual terms are explicit in our results, but the latter are expected to be relevant when n is large. As a matter of fact, our results are stated for n greater than some unknown constant. In this approach, by a "constant" we understand a quantity independent from the sample size. In practice, natural trade-offs between the value of the constants and the sample size have to be taken into account in order to legitimate some assumptions, such as for instance the polynomial complexity of the collection of models.

Model selection *via* penalization is not the only method which provides sharp oracle inequalities for the estimation a non-parametric regression function. Indeed, aggregation techniques ([24], [27], [33]) and PAC-Bayesian bounds ([26], [27]) also allow to obtain nearly optimal constants in the derived oracle inequalities. A major difference between aggregation and model selection studies, is that in aggregation results, the estimators at hand are usually considered as deterministic functions. This implies in practice that the estimators are estimated beforehand. In [24], Bunea, Tsybakov and Wegkamp derive some sharp oracle inequalities for different aggregation tasks by means of a single unifying procedure. However, the authors ask for a fixed design and homoscedastic Gaussian noise. By using aggregation with exponential weights, Dalalyan and Tsybakov obtain in [27] oracle inequalities of a PAC-Bayesian flavor with leading constant 1 and optimal rate of the remainder term for the estimation of a regression function with deterministic design and homoscedastic errors. Moreover, the law of errors should be symmetric or n-divisible. PAC-Bayesian methods are systematically investigated in Catoni, [26]. The work of Lecué and Mendelson in [33] 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. 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 [26].

Birgé and Massart [19] have discovered, in a generalized linear Gaussian model setting, that the optimal penalty is closely related to the minimal one, defined to be the maximal penalty under which the procedure totally misbehaves. They prove sharp upper and lower bounds for the minimal penalty and show that the optimal penalty is twice the minimal one, both for small and large collections of models. These facts are called the *slope heuristics*. The authors also exhibit a jump in the dimension of the selected model occurring around the value of the minimal penalty, and use it to estimate the minimal penalty from the data. Taking a penalty equal to twice the previous estimate then gives a non-asymptotic 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, Giraud and Huet [11] 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 [25] obtained a lower bound

on the penalty term, in the case of small collections of models.

The slope phenomenon has been then extended by Arlot and Massart [7] in a bounded regression framework, with heteroscedastic noise and random design. The authors consider least-squares estimators on a "small" collection of histograms models. Their analysis differ from the one of Birgé and Massart in an important fact. Indeed, the authors do not assume a particular *shape* of the penalty term. As a matter of fact, the penalties considered by Birgé and Massart in [19] were known functions of the dimension of the models, whereas heteroscedasticity of the noise allows 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 propose to estimate the shape of the penalty by using Arlot's resampling or V-fold penalties, proved to be efficient in their regression framework by Arlot [8] and [5].

The approach developed in [7] is more general than the histogram case, except for some identified technical parts of the proofs, thus providing some quite general algebra that can be applied in other frameworks to derive sharp model selection results. The authors have also identified the minimal penalty as the mean of the empirical excess 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 [7] conjecture that this equivalence between the empirical and true excess loss is a quite general fact in M-estimation, as well as, by a rather direct consequence, the slope phenomenon for models not too badly chosen in terms of approximation properties. A general result supporting this conjecture is the high dimensional Wilks' phenomenon discovered by Boucheron and Massart [22] in the setting of bounded contrast minimization. The authors derive in [22] 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 moment of first order 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 [35] 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 also proves in the least-squares density estimation setting the efficiency of Arlot's resampling penalties, and generalizes these results to weakly dependent data, see [36]. Arlot and Bach [9] recently consider the problem of selecting among linear estimators in non-parametric regression. Their framework includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression or spline smoothing, and the choice of a kernel in multiple kernel learning. In such cases, the minimal penalty is not necessarily half the optimal one, but the authors propose to estimate the unknown variance by the minimal penalty and to use it in a plug-in version of Mallows' C_L. The latter penalty is proved to be optimal by establishing a non-asymptotic 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 a bounded heteroscedastic with random design regression framework, by considering a "small" collection of finite-dimensional linear models of piecewise polynomials. This setting extends the case of histograms already treated by Arlot and Massart [7]. An interesting consequence for the practicer is that piecewise polynomials are known to have good approximation properties in Besov spaces and can lead to minimax rates of convergence, see for instance [12] and [41]. As a matter of fact, histograms allow minimax procedures only on Hölderian spaces.

Two main assumptions concerning our models must be satisfied. First, we require that the models have a uniform localized orthonormal basis structure in $L_2\left(P^X\right)$, where P^X is the law of the explicative variable X. This kind of analytical property describing the L_∞-structure of the models has already been used in a model selection framework by Birgé and Massart [17] and Barron, Birgé and Massart [12] (see also Massart [38]). It is worth noticing that considering penalization procedures for regression with random design, Baraud [10] also used the localized basis structure to prove oracle inequalities in a setting very close from ours. However, Baraud's framework differs from ours by the fact that the risk considered by the author is given by the quadratic norm in $L_2(\nu)$ for a reference measure ν, known from the statistician.

Considering the unit cube of $\mathbb{R}^q$ and taking $P^X = \text{Leb}$ the Lebesgue measure on it, it is shown in Birgé and Massart [17] that the assumption of localized orthonormal basis are satisfied for some wavelet expansions and piecewise polynomials uniformly bounded in their degrees. It is also known, Massart [38], that in the case

of histograms the property of localized basis in $L_2\left(P^X\right)$ is equivalent to the lower regularity of the considered partition with respect to P^X, an assumption required by Arlot and Massart in [7]. Moreover, we show in [40] that if P^X has a density, with respect to the Lebesgue measure on the unit interval, that is uniformly bounded away from zero then, assuming the lower regularity of the partition defining the piecewise polynomials ensures that the assumption of localized basis is satisfied for such a model.

The second property that must be satisfied in our setting is that the least-squares estimators are uniformly consistent over the collection of models and converge toward the orthogonal projections of the unknown regression function. Again, such a property is shown in [40] to be satisfied for suitable histograms and more general piecewise polynomial models. This allows us to recover the results of Arlot and Massart [7] with the same set of assumptions when the noise is uniformly bounded from above and from below, and to extend it to models of piecewise polynomials uniformly bounded in their degrees. Taking advantage of the sharp estimates of the empirical and true excess losses for a fixed model given in [40], our proofs then rely on the same algebra of proofs as those given in Arlot and Massart [7].

If the noise is homoscedastic, then the shape of the ideal penalty is known, and is linear in the dimension of the models as it is the case in Mallows' C_p. However, if the noise is heteroscedastic, then the ideal penalty is not even a function of the linear dimensions of the models (see [6]). So, it remains to give a robust estimator of this shape. As emphasized by Arlot in [5] and [8], V-fold and resampling penalties are good, natural candidates for this task. In this paper, we show that a hold-out penalty - which is highly 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. This is an advantage of our penalization procedure compared to the classical hold-out procedure, which is likely to be asymptotically suboptimal in this case. It is worth noticing that hold-out type procedures have also been exploited in Chapter 8 of Massart [38] as simple tools in practice to overcome the margin adaptivity issue in classification.

The paper is organized as follows. We describe in Section 2 the statistical framework, the slope heuristics is treated in Section 3 and the hold-out penalization is considered in Section 4. The proofs are postponed to the remainder of the paper.

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

From (1), it comes that s_* is the 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 polynomials, 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 quadratic norm endowing $L_2\left(P^X\right)$ and 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 $P(f) = Pf = \mathbb{E}[f(\xi)]$. 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

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

where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ is the empirical measure built from the data.

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

$$\ell(s_*, s_M) := P(Ks_M - Ks_*) = \|s_M - s_*\|_2^2 \quad \text{and} \quad \ell(s_M, s_n(M)) := P(Ks_n(M) - Ks_M) \geq 0 .$$

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

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

and we prove sharp bounds for the latter quantity in [40], based on the expansion of the least-squares contrast to the sum of a linear part and a quadratic part.

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

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

The associated oracle estimator $s_n(M_*)$ thus achieves the best performance in terms of excess loss among the collection $\{s_n(M); M \in \mathcal{M}_n\}$. The oracle model is a random quantity because it depends on the data and it is also unknown as it depends on the law 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,

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

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

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

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 polynomials

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

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

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

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

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

Notice that models of histograms on the unit interval are exactly models of piecewise polynomials with degrees not larger than 0.

In [40], it is shown that models of piecewise polynomials have nice analytical and statistical properties. Let us recall two of them.

In Lemma 8 of [40], it is proved that if the law 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 $\mathcal{P}$ if lower regular with respect to Leb - that is there exists a positive constant c such that $\left|\mathcal{P}\right| \inf_{I \in \mathcal{P}} \mathrm{Leb}\left(I\right) \geq c > 0$ - then the associated model of piecewise polynomials 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, see [21], [20] and [38], the property of localized basis is known to play a center role in M-estimation and model selection using vector spaces or more general sieves.

Considering models of piecewise polynomials 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 [40] that the least-squares estimator $s_n\left(M\right)$ converges in sup-norm toward the linear projection s_M of the regression function s_*. More precisely, if $\left\|\cdot\right\|_\infty$ denotes the sup-norm on $\mathcal{X}$ and if there exists a constant A such that $\left|Y\right| \leq A$ *a.s.* then for any $\alpha > 0$, there exists a constant L and a positive integer n_0 such that, for all $n \geq n_0$,

$$\mathbb{P}\left(\left\|s_n\left(M\right) - s_M\right\|_\infty \geq L\sqrt{\frac{D_M \ln n}{n}}\right) \leq n^{-\alpha} \ .$$

As a matter of fact, 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 polynomials - see Section 3.3 below.

3 The slope heuristics

We state in Section 3.4 below our results that theoretically validate the slope heuristics in our bounded heteroscedastic regression setting, using a "small" collection of models of piecewise polynomials. In particular, we essentially extend the results stated in Theorems 2 and 3 of Arlot and Massart [7] for histogram models to the case models of piecewise polynomials uniformly bounded in their degrees. The proofs are postponed to the end of the paper, and heavily rely on results obtained in [40] where we consider a fixed model, as well as on the general algebra of proofs developed by Arlot and Massart [7]. The reader interested 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 toward the linear projections of the regression function onto the models.

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 [40], let us first briefly expose, at a heuristic level, 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

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

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

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

The penalty function $\mathrm{pen}_{\mathrm{id}}$ is called the *ideal penalty* - as it allows to select the oracle - and is unknown because it depends on the distribution of the data. As pointed out by Arlot and Massart [7], the leading idea of penalization in the efficiency problem is to give some sharp estimate, up to a constant, of the ideal penalty. This would allow to perform 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 constant almost one, converging to one when the sample size n tends to infinity.

Concerning the estimation of the optimal penalty, Arlot and Massart [7] conjectured that the mean of the empirical excess loss $\mathbb{E}[P_n(Ks_M - Ks_n(M))]$ 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 $M \in \mathcal{M}_n$,

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

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

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

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

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 $\mathrm{pen}_{\mathrm{min}}$:

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

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 3.4 precise results which validate the slope heuristics for models of piecewise polynomials.

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

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

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 $P_n(Ks_M - Ks_n(M))$. Indeed, as shown by Boucheron and Massart [22], 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, $\text{pen} \approx (1+\delta)\,\text{pen}_{\min}$ with $\delta > 0$, then by the same kind of manipulations, the selected model minimizes the following criterion, for all $M \in \mathcal{M}_n$,

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

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, $\text{pen} \approx 2 \times \text{pen}_{\min}$, and if we assume that the empirical excess loss is equivalent to the excess loss,

$$P_n\left(Ks_M - Ks_n\left(M\right)\right) \sim P\left(Ks_n\left(M\right) - Ks_M\right) \ , \tag{9}$$

then according to (8) the selected model almost minimizes

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

Hence,

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

and the procedure is nearly optimal.

We give in [40] some results showing that (9) is a quite general fact in least-squares regression and is in particular satisfied when considering models of piecewise polynomials. Thus, these results represent a preliminary material for the present study, and we shall base our arguments on the results exposed in [40].

3.2 A data-driven calibration of penalty algorithm

The slope heuristics stated in Section 3 predict that a jump in the dimensions of the selected models should occur around the minimal penalty. This jump can be used to estimate the minimal penalty and consequently, the optimal one. Let us denote by $\text{pen}_{\text{shape}}$ the shape of the minimal penalty which is, according to the slope heuristics, equal to the shape of the optimal penalty. Thus, for two unknown positive constants $A_{\min}$ and A_* depending on the unknown distribution of the data, we have

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

where $A_* = 2 \times A_{\min}$ whenever the optimal penalty is twice the minimal one. We assume now that the shape of the minimal penalty is known, from some prior knowledge or because it has been estimated from the data, for instance by using Arlot's resampling and V-fold penalties as suggested in [7]. In the latter paper, Arlot and Massart propose to *calibrate* the optimal penalty by the following procedure and by doing so, they extend to general penalty shapes a previous algorithm proposed by Birgé and Massart [19].

Algorithm of data-driven calibration of penalties:

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

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

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

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

Our aim in the present paper is not to apply the above algorithm in practice. We refer to Arlot and Massart [7] for a precise formalization of the algorithm and to Baudry, Maugis and Michel [16] for detailed discussions on implementation issues. Data-driven calibration of penalties algorithms have already been applied with some success in many statistical frameworks such as mixture models [39], clustering [15], spatial statistics [42], estimation of oil reserves [34] and genomics [43], to name but a few. These applications tend to support the conjecture of Arlot and Massart [7], which predicts that the slope heuristics are valid in a quite general framework.

3.3 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 polynomials. We denote by $\mathcal{P}_M$ the partition of $\mathcal{X}$ underlying the model M.

Set of assumptions for piecewise polynomials: (**SAPP**)

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

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

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in \left[n^{1/\left(1+\beta_+\right)}, c_{rich} n^{1/\left(1+\beta_+\right)}\right]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$, where β_+ is defined in $(\mathbf{Ap}_u)$.

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

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

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

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

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

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

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

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

(Alr) Lower regularity of the partitions: 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 to the general regression relation stated in (1). Thirdly, assumptions (**Aud**) and (**Alr**) specify some quantities related to the choice of the models of piecewise polynomials.

Assumption (**P1**) states that the collection of models has a "small" complexity, more precisely a polynomially increasing one with respect to the amount of data. For this kind of complexities, if one wants to perform a good model selection procedure for prediction, the chosen penalty should estimate the mean of the ideal one on each model, 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 [17] and [12]).

In assumption (**P2**) we restrict the dimensions of the models from above, in a way that is not too restrictive since we allow the dimension to be of the order of the sample size within a power of a logarithmic factor. We assume in (**P3**) that the collection of models contains a model M_0 of reasonably large dimension and a model M_1 of high dimension, which is necessary since we prove the existence of a jump between high and reasonably large dimensions. 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 practicer 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 ask 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 require a polynomially decreasing of excess loss of linear projections of the regression function onto the models. It is well-known that piecewise polynomials uniformly bounded in their degrees have good approximation properties in Besov spaces. More precisely, as stated in Lemma 12 of Barron, Birgé and Massart [12], 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 [12]), then taking models of piecewise polynomials of degree bounded by $r > \alpha - 1$ on regular partitions with respect to the Lebesgue measure Leb on $\mathcal{X}$, and assuming that P^X has a density with respect to Leb which is bounded in sup-norm, assumption ($\mathbf{Ap}_u$) is satisfied.

Assumption (**Ab'**) is rather restrictive, since it excludes the 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 use classical Bousquet and Klein-Rio's inequalities - recalled in Section 5.4 below. As a matter of fact, we do not know yet if an adaptation of our proofs (including results established in [40]) by using extensions of the latter inequalities to some unbounded cases - see for instance Adamczak's concentration inequalities in [1] - would be possible.

The noise restriction stated in (**An**) is needed to derive our results that optimal to the first order. More precisely, it allows in [40] 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 [7] concerning the case of histogram models. As it is noticed in Section 5.3 of [40], assumption (**An**) could be replaced by the following assumption, which states that the partitions underlying the models of piecewise polynomials are regular from above with respect to the Lebesgue measure on $[0,1]$.

(Aur) Upper-regularity of the partition: a positive constant $c^+_{\mathcal{M},\mathrm{Leb}}$ exists such that, for all $M \in \mathcal{M}_n$,

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

Now, assumptions ($\mathbf{Ad}_{\mathrm{Leb}}$), (**Aud**) and (**Alr**) essentially allow to recover some good analytical and statistical properties for the models of piecewise polynomials, 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.

3.4 Statement of the theorems

We are now able to state our main results dealing with the slope heuristics.

Theorem 2 *Under the set of assumptions (**SAPP**) of Section 3.3, for $A_{\mathrm{pen}} \in [0,1)$ and $A_p > 0$, we assume that with probability at least $1 - A_p n^{-2}$ we have*

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

*where the model M_1 is defined in assumption (**P3**) of (**SAPP**). 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 $(\boldsymbol{SAPP})$ such that, for all $n \geq n_0\left((\boldsymbol{SAPP}), A_{\text{pen}}\right)$, it holds with probability at least $1 - A_1 n^{-2}$,

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

and

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

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

Theorem 5 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 3 *Assume that the set of assumptions $(\boldsymbol{SAPP})$ of Section 3.3 holds. Moreover, for some $\delta \in [0,1)$ and $A_p, A_r > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}(\ln n)^3$, it holds*

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

together with

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

for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}(\ln n)^3$. Then, for any $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$, there exist an integer n_0 only depending on η, δ and β_+ and on constants in $(\boldsymbol{SAPP})$, a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in $(\boldsymbol{SAPP})$ and on A_p, two positive constants A_4 and A_5 only depending on constants in $(\boldsymbol{SAPP})$ and on A_r and a sequence

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

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

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

and

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

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

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

and

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

Theorem 6 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 ($\mathbf{Ap}_u$) with a polynomial lower bound of the bias of the models. On histogram models, Arlot shows in Section 8.10 of [4] that this lower bound is satisfied for non constant α-Hölderian, $\alpha \in (0,1]$, regression functions and for regular partitions. This result remains to be suitably extended to models of piecewise polynomials uniformly bounded in their degrees.

Finally, from Theorems 5 and 6, we identify the minimal penalty with the mean of the empirical excess loss on each model,

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

thus generalizing the results of Arlot and Massart in [7] to the case of piecewise polynomials.

4 Hold-out penalization

In Section 3.4, we gave theoretical guarantees for a penalized least-squares criterion to achieve a non-asymptotic oracle inequality (Theorem 3), that can be asymptotically optimal, as well as a sufficient condition leading to a bad behavior, close to the selection operated by the (unpenalized) empirical loss (Theorem 2).

But the conditions given in Theorems 2 and 3 can not be directly applied 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 non-asymptotic quasi-optimal penalty. According to Theorem 6, such a penalty is given by twice the mean of the empirical excess loss. Now, using Theorem 10 of [40], we get (with an explicit control of the second order terms in the following equivalence),

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

where $\mathcal{K}_{1,M}^2 = 1/D_M \sum_{k=1}^{D_M} \mathbb{E}\left(\left(\psi_{1,M}(X,Y)\cdot\varphi_k(X)\right)^2\right)$, $\psi_{1,M}(X,Y) = -2\left(Y - s_M(X)\right)$ and $(\varphi_k)_{k=1}^{D_M}$ is an orthonormal basis in $\left(M, \|\cdot\|_{L_2(P^X)}\right)$. By easy computations, we deduce that if the noise is homoscedastic, that is $\sigma^2(X) \equiv \sigma^2 > 0$, 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 6 (thanks to Lemma 7 in [40], 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, Mallow's penalty is only known through a constant, the noise level, which can be estimated *via* the slope heuristics algorithm described in Section 3.2 above. 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, any calibration of a linear penalty lead to a suboptimal procedure, but yet can achieve an oracle inequality with a leading constant more than one, see [6].

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 point. We propose a hold-out type penalty that automatically adapts to heteroscedasticity. Moreover, it appears that this penalty can be viewed as a special case of the resampling penalties proposed by Arlot and that have been studied in heteroscedastic regression using linear models of histograms, see [8]. A short discussion comparing our procedure with the classical hold-out procedure can be found at the end of this section. Let us now detail the hold-out type procedure.

We want to estimate the ideal penalty, up to a constant additive term. The ideal penalty is defined by

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

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 $I_1 \cup I_2 = \{1, ..., n\}$ and to propose the following hold-out type penalty,

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

where $P_{n_i} = 1/n_i \sum_{j \in I_i} \delta_{\xi_j}$, $n_i = \mathrm{Card}(I_i)$, for $i = 1, 2$, $s_{n_1}(M) \in \arg\min_{s \in M} P_{n_1}(Ks)$ and $C > 0$ is a constant to be determined. Indeed, if n_1 is not too small, $P_{n_1}\left(Ks_{n_1}(M)\right)$ is likely to vary like $P_n(Ks_n(M))$ and $P_{n_2}\left(Ks_{n_1}(M)\right)$ is, conditionally to $\left(\xi_j\right)_{j \in I_1}$, an unbiased estimate of $P\left(Ks_{n_1}(M)\right)$, which again is likely to vary like $P(Ks_n(M))$. Moreover, we see from Theorem 10 in [40] 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 preceding observation is justified by the following theorem, which is a corollary of Theorems 2 and 3. We denote by

$$\widehat{M}_{n_1} \in \arg\min_{M \in \mathcal{M}_n} \left\{P_n(Ks_n(M)) + \mathrm{pen}_{ho,C}(M)\right\}$$

the model selected by the hold-out penalization.

Theorem 4 *Assume that the set of assumptions (**SAPP**) of Section 3.3 holds and that there exists $c \in (0,1)$ such that $nc \leq n_1 < n$. Assume moreover that there exists $\tau \in (1,3)$ satisfying $n(\ln n)^{\tau}/D_M \leq n_2 \leq n(1-c)$ for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} n/(\ln n)^2 \geq D_M \geq A_{\mathcal{M},+}(\ln n)^3$ and take $n_2 = n(1-c)$ if $D_M \leq A_{\mathcal{M},+}(\ln n)^3$. For any $C > 0$, define for all $M \in \mathcal{M}_n$,*

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

*Favorable case: assume that $n_1/(2n) < C \leq 3n_1/(2n)$. Then, for any $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$, there exist an integer n_0 only depending on c, η, C 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 on c and a sequence*

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

such that it holds for all $n \geq n_0((\textbf{SAPP}), c, \eta, C)$, with probability at least $1 - A_6 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(\frac{1 + 2\left|C\frac{n}{n_1} - 1\right|}{1 - 2\left|C\frac{n}{n_1} - 1\right|} + \frac{12\theta_n}{\left(1 - 2\left|C\frac{n}{n_1} - 1\right|\right)^2}\right) \ell\left(s_*, s_n(M_*)\right) + A_8 \frac{(\ln n)^3}{n} \ . \tag{19}$$

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*

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

Then it holds for all $n \geq n_0\left((\textbf{SAPP}), c, C_-, \beta_-, \eta, C\right)$, with probability at least $1 - A_6 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(\frac{1+2\left|C\frac{n}{n_1}-1\right|}{1-2\left|C\frac{n}{n_1}-1\right|}+\frac{12\theta_n}{\left(1-2\left|C\frac{n}{n_1}-1\right|\right)^2}\right)\inf_{M\in\mathcal{M}_n}\left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{20}$$

Unfavorable case: Assume that $C < n_1/(2n)$. Then there exist a constant $A_1 > 0$ only depending on constants in $(\boldsymbol{SAPP})$, as well as an integer n_0 and a positive constant A_2 only depending on c, C and on constants in $(\boldsymbol{SAPP})$ such that, for all $n \geq n_0((\boldsymbol{SAPP}), c, C)$, it holds with probability at least $1 - A_1 n^{-2}$,

$$D_{\widehat{M}} \geq A_2 n \ln(n)^{-2} \tag{21}$$

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \geq \frac{n^{\beta_+/\left(1+\beta_+\right)}}{\left(\ln n\right)^3}\inf_{M\in\mathcal{M}_n}\left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{22}$$

Let us now give some comments about the previous result.
First of all, Theorem 4 justifies the choice in practice of a hold-out penalization procedure, since for the choice $C_{opt} = n_1/n$, this leads to a quasi-optimal procedure. Yet, the choice of the multiplying factor C is sensible since an under-penalization, when C is less than $n_1/(2n)$, leads to disastrous results. In the aim to give a robust, non-asymptotical and data-driven choice of C, one could use the slope heuristics, based on the shape of the penalty given by the hold-out penalty with $C = 1$. We do not know yet if this additional procedure would increase the performances of the hold-out penalization in practice.

Another interesting aspect of our results lies in the conditions imposed on the data splitting, taken to satisfy $nc \leq n_1 < n$ and for all $M \in \mathcal{M}_n$, $n(\ln n)^\tau / D_M \leq n_2 \leq n(1-c)$. Here n_1 is devoted to "learn" each estimator used in the penalty and n_2 is used to "validate" the performance of the estimators. We ask that n_1 is of the order of the amount of initial data, within a constant factor $c < 1$, which is rather mild. Moreover, we take n_2 to be greater than a lower bound depending on each model M. It is thus important to note that we do not exclude that $(n_1, n_2) = (n_1(M), n_2(M))$ is depending on M, that is we split the data independently for each model. For collections of models with cardinality of the order of n - for instance one model for each dimension -, such splits will only increase by a constant factor the complexity of a hold-out penalization with a unique split of the data. The advantage of conditioning the data splitting to the dimension of the considered model would then be to strengthen for large models the accuracy of the estimators used in the penalty. Indeed, it is natural to think that more data is needed in the "learning step" when one is considering large models.

It is also worth noticing that our hold-out penalty is highly linked to Arlot's resampling penalties with resampling random hold-out weight, see [8]. Indeed, the general form of the resampling penalties proposed by Arlot is

$$\mathrm{pen}^W(M) = C\mathbb{E}^W\left[P_n\left(K\left(s_n^W(M)\right)\right) - P_n^W\left(K\left(s_n^W(M)\right)\right)\right] \ ,$$

where C is a positive constant, $s_n^W(M)$ is the estimator learned with the resampled empirical measure $P_n^W = 1/n\sum_{i=1}^n W_i\delta_{(X_i,Y_i)}$ and $\mathbb{E}^W[\cdot]$ denotes the mean with respect to the real random variables $(W_1, ..., W_n)$. In the case of the random hold-out weights with n_1 learning data, it holds $W_i^{n_1} = (n/n_1)\mathbf{1}_{i\in I}$ with I uniform subset of cardinality n_1 of $\{1, ..., n\}$. As $P_n = (n_1/n)P_{n_1} + (n_2/n)P_{n_2}$ and $P_n^{W^{n_1}} = P_{n_1}$, we get

$$\begin{aligned}\mathrm{pen}^{W^{n_1}}(M) &= C\mathbb{E}^{W^{n_1}}\left[P_n\left(K\left(s_n^{W^{n_1}}(M)\right)\right) - P_n^{W^{n_1}}\left(K\left(s_n^{W^{n_1}}(M)\right)\right)\right] \\ &= C\frac{n_2}{n}\mathbb{E}^I\left[P_{n_2}\left(Ks_{n_1}(M)\right) - P_{n_1}\left(Ks_{n_1}(M)\right)\right] \ .\end{aligned} \tag{23}$$

Thus, the random hold-out penalty proposed by Arlot 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 a raise in the computational cost. However, the stabilization effect seems more difficult to invest mathematically, and our results provide a first step toward the study of the more complicated resampling penalties. Finally, according to Theorem 4 the choice $Cn_2/n = n_1/n$ in (23) should be optimal and we recover indeed the choice advocated by Arlot in [8] to take $C = n_1/n_2$.

Finally, a remarkable fact about the hold-out penalization, compared to the classical hold-out, is that hold-out penalization is asymptotically optimal for a choice $n_1 = nc = n - n_2$ with $c \in (0, 1)$ - the most

classical choice being $n_1 = n/2$. Moreover, it is straightforward to see that the computational complexity of the penalization procedure is, within a multiplying constant, the same as the classical hold-out, the latter being defined by

$$\widehat{M}_{ho} \in \arg\min_{M \in \mathcal{M}_n} \left\{ P_{n_2} \left(K s_{n_1} (M) \right) \right\} \ . \tag{24}$$

The choice $n_1 = n/2$ in (24) is likely to lead to an asymptotically suboptimal procedure, as the criterion is close in expectation of $P\left(K s_{n/2}(M)\right)$, and so is close to the oracle, but for $n/2$ data. The hold-out penalization allows to overcome this difficulty. For similar advantages concerning resampling and V-fold penalties, see Arlot [8] and [5].

5 Proofs

This section is divided in four sections. First of all, we present a setting which is slightly more general than the piecewise polynomials case. Structural properties of models, that are sufficient for our needs and that are satisfied for models of piecewise polynomials considered in (**SAPP**), are highlighted in Section 5.1. Then in Sections 5.2 and 5.3 respectively, we prove results exposed in Sections 3.4 and 4, concerning the slope heuristics and the hold-out penalization respectively. Finally, we recall in Section 5.4 some probabilistic tools that are recurrent in our proofs.

5.1 A more general setting

The following set of assumptions, denoted (**GSA**), can be considered as a generalization of the set of assumptions (**SAPP**) exposed in Section 3.3 and which is related models of piecewise polynomials. Here, piecewise polynomials are replaced by models that are equipped with an orthonormal basis in $L_2\left(P^X\right)$ and that satisfy an assumption of consistency in sup-norm concerning the least-squares estimators.

General set of assumptions: (GSA)

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

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

(P3) Richness of $\mathcal{M}_n$: there exist $M_0, M_1 \in \mathcal{M}_n$ such that $D_{M_0} \in \left[n^{1/\left(1+\beta_+\right)}, c_{rich} n^{1/\left(1+\beta_+\right)} \right]$ and $D_{M_1} \geq A_{rich} n \left(\ln n\right)^{-2}$.

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

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

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

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

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

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

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

($\mathbf{Ac}_\infty$) Consistency in sup-norm of least-squares estimators: 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 $M \in \mathcal{M}_n$ it holds,

$$\|s_n(M) - s_M\|_\infty \leq A_{cons}\sqrt{\frac{D_M \ln n}{n}} \ . \tag{25}$$

In (**GSA**), assumptions (**P1**), (**P2**), (**P3**), (**An**) and ($\mathbf{Ap}_u$) are shared with the set (**SAPP**), whereas assumptions (**Ab**), (**Alb**) and ($\mathbf{Ac}_\infty$) differ from it. More precisely, assumptions (**Ab'**), ($\mathbf{Ad}_{\text{Leb}}$), (**Aud**) and (**Alr**) allow to recover (**Ab**), (**Alb**) and ($\mathbf{Ac}_\infty$) in the special case of models of piecewise polynomials.

Firstly, assumption (**Ab**) only differs from (**Ab'**) from 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 polynomials uniformly bounded in their degrees, this follows from simple computations (see Section 5.3 in [40]). Then, assumption (**Alb**) requires the existence of a localized orthonormal basis for each model. In the case of piecewise polynomials, this is ensured by ($\mathbf{Ad}_{\text{Leb}}$), (**Aud**)and (**Alr**), see Lemma 8 of [40]. Finally, assumption ($\mathbf{Ac}_\infty$) states that each estimator is consistent in sup-norm toward the corresponding projection of the target. Again, this is satisfied for models of piecewise polynomials under assumptions ($\mathbf{Ad}_{\text{Leb}}$), (**Aud**) and (**Alr**). This result is established in Lemma 9 of [40].

Let us now describe a set of assumption, less restrictive than (**SAPP**), that allows to recover (**GSA**) when considering histogram models. Lemma 5 and 6 of [40] 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**) and ($\mathbf{Ap}_u$) from the general set of assumptions. Assume moreover that the following conditions hold true:

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

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

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

It is worth noticing that Theorems 2 and 3 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 exposed in Theorems 2 and 3 of [7], concerning the selection of least-squares estimators among histogram models.

5.2 Proofs related to Section 3.4

We will prove the two following theorems, that are based on (**GSA**). From the remarks exposed in Section 5.1, these theorems imply Theorems 2 and 3 of Section 3.4.

Theorem 5 *Under the general set of assumptions (**GSA**) of Section 5.1, for $A_{\text{pen}} \in [0,1)$ and $A_p > 0$, we assume that with probability at least $1 - A_p n^{-2}$ we have*

$$0 \leq \text{pen}(M_1) \leq A_{\text{pen}} \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M_1)\right)\right] \ , \tag{26}$$

*where the model M_1 is defined in assumption (**P3**) of (**GSA**). Then there exist a constant $A_1 > 0$ only depending on constants in (**GSA**), as well as an integer n_0 and a positive constant A_2 only depending on A_{pen} and on constants in (**GSA**) such that, for all $n \geq n_0((\textbf{GSA}), A_{\text{pen}})$, it holds with probability at least $1 - A_1 n^{-2}$,*

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{n^{\beta_+/\left(1+\beta_+\right)}}{\left(\ln n\right)^3} \inf_{M\in\mathcal{M}_n}\left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{27}$$

Theorem 6 *Assume that the general set of assumptions (**GSA**) of Section 5.1 holds. Moreover, for some $\delta \in [0,1)$, $A_p, A_r > 0$, assume that an event of probability at least $1 - A_p n^{-2}$ exists on which, for every model $M \in \mathcal{M}_n$ such that $D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^3$, it holds*

$$\left|\mathrm{pen}\left(M\right) - 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]\right| \leq \delta\left(\ell\left(s_*, s_M\right) + \mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]\right) \tag{28}$$

together with

$$\left|\mathrm{pen}\left(M\right)\right| \leq A_r\left(\frac{\ell\left(s_*, s_M\right)}{\left(\ln n\right)^2} + \frac{\left(\ln n\right)^3}{n}\right) \tag{29}$$

*for every model $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. Then, for any $\eta \in \left(0, \beta_+/\left(1+\beta_+\right)\right)$, there exist an integer n_0 only depending on η, δ and β_+ and on constants in (**GSA**), a positive constant A_3 only depending on $c_{\mathcal{M}}$ given in (**GSA**) and on A_p, a positive constant A_4 and A_5 only depending on constants in (**GSA**) and on A_r and on A_r and a sequence*

$$\theta_n \leq \frac{A_4}{\left(\ln n\right)^{1/4}} \tag{30}$$

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

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{\left(1-\delta\right)^2}\right)\ell\left(s_*, s_n\left(M_*\right)\right) + A_5\frac{\left(\ln n\right)^3}{n} \ . \tag{31}$$

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 $n \geq n_0\left(\left(\textbf{\textit{GSA}}\right), C_-, \beta_-, \beta_+, \eta, \delta\right)$, with probability at least $1 - A_3 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{\left(1-\delta\right)^2}\right)\ell\left(s_*, s_n\left(M_*\right)\right) \ . \tag{32}$$

The following remark will be useful.

Remark 7 *Since constants in the general set of assumptions (**GSA**) made above are uniform over the collection $\mathcal{M}_n$, we deduce from Theorem 2 of [40] 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 the general set of assumptions, such that for all $M \in \mathcal{M}_n$ satisfying*

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

by setting

$$\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \ \left(\frac{D_M \ln n}{n}\right)^{1/4} \right\} \tag{33}$$

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

$$\mathbb{P}\left[(1-\varepsilon_n(M)) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P(Ks_n(M) - Ks_M) \leq (1+\varepsilon_n(M)) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \right] \geq 1 - 10n^{-2-\alpha_{\mathcal{M}}} \tag{34}$$

and

$$\mathbb{P}\left[\left(1-\varepsilon_n^2(M)\right) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq \left(1+\varepsilon_n^2(M)\right) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \right] \geq 1 - 5n^{-2-\alpha_{\mathcal{M}}} \tag{35}$$

where $\mathcal{K}_{1,M}^2 = 1/D_M \sum_{k=1}^{D_M} \mathbb{E}\left(\left(\psi_{1,M}(X,Y)\cdot\varphi_k(X)\right)^2 \right)$, $\psi_{1,M}(X,Y) = -2(Y - s_M(X))$ *and* $(\varphi_k)_{k=1}^{D_M}$ *is an orthonormal basis in* $\left(M, \|\cdot\|_{L_2(P^X)}\right)$. *Moreover, for all* $M \in \mathcal{M}_n$, *we have by Theorem 3 of [40], for a positive constant* A_u *depending on* $A, A_{cons}, r_{\mathcal{M}}$ *and* $\alpha_{\mathcal{M}}$ *and for all* $n \geq n_0(A_{cons}, n_1)$,

$$\mathbb{P}\left[P(Ks_n(M) - Ks_M) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3n^{-2-\alpha_{\mathcal{M}}} \tag{36}$$

and

$$\mathbb{P}\left[P_n(Ks_M - Ks_n(M)) \geq A_u \frac{D_M \vee \ln n}{n} \right] \leq 3n^{-2-\alpha_{\mathcal{M}}} \ . \tag{37}$$

Before stating the proofs of Theorems 6 and 5, we need two technical lemmas. In the first lemma, we intend to evaluate the minimal penalty $\mathbb{E}[P_n(Ks_M - Ks_n(M))]$ for models of dimension not too large and not too small.

Lemma 8 *Assume (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac**$_\infty$) of the general set of assumptions defined in Section 5.1. Then, for every model* $M \in \mathcal{M}_n$ *of dimension* D_M *such that*

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

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

$$\left(1 - L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2(M)\right) \frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}[P_n(Ks_M - Ks_n(M))] \tag{38}$$

$$\leq \left(1 + L_{A_{\mathcal{M},+},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2(M)\right) \frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ , \tag{39}$$

where $\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \left(\frac{D_M \ln n}{n}\right)^{1/4} \right\}$ *is defined in Remark 7.*

Proof. As explained in Remark 7, 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 $\Omega_1(M)$ of probability at least $1 - 5n^{-2-\alpha_{\mathcal{M}}}$,

$$(1-\varepsilon_n(M)) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \leq P_n(Ks_M - Ks_n(M)) \leq (1+\varepsilon_n(M)) \frac{1}{4}\frac{D_M}{n}\mathcal{K}_{1,M}^2 \ , \tag{40}$$

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 $|Y_i| \leq A$ *a.s.* and $\|s_M\|_\infty \leq A$ by (**Ab**), it holds

$$0 \leq P_n(Ks_M - Ks_n(M)) \leq P_n Ks_M = \frac{1}{n}\sum_{i=1}^n (Y_i - s_M(X_I))^2 \leq 4A^2 \tag{41}$$

and as $D_M \geq 1$, we have

$$\varepsilon_n(M) = A_0 \max\left\{ \left(\frac{\ln n}{D_M}\right)^{1/4} ; \ \left(\frac{D_M \ln n}{n}\right)^{1/4} \right\} \geq A_0 n^{-1/8} \ . \tag{42}$$

We also have

$$\begin{aligned} &\mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\right] \\ &= \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] + \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \ . \end{aligned} \tag{43}$$

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 (41) and (42) that

$$0 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{(\Omega_1(M))^c}\right] \leq 20A^2 n^{-2-\alpha_{\mathcal{M}}} \leq \frac{80A^2}{A_0^2\sigma_{\min}^2}\varepsilon_n^2(M)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{44}$$

Moreover, we have $\varepsilon_n(M) < 1$ for all $n \geq n_0(A_0, A_{\mathcal{M},+}, A_{cons})$, so by (40),

$$0 < \left(1 - 5n^{-2-\alpha_{\mathcal{M}}}\right)\left(1 - \varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[P_n\left(Ks_M - Ks_n(M)\right)\mathbf{1}_{\Omega_1(M)}\right] \tag{45}$$

$$\leq \left(1 + \varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \ . \tag{46}$$

Finally, noticing that $n^{-2-\alpha_{\mathcal{M}}} \leq A_0^{-2}\varepsilon_n^2(M)$ by (42), we use (44), (45) and (46) in (43) 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 (38) and (39), as A_0 only depends on $\alpha_{\mathcal{M}}$, $A_{\mathcal{M},+}$, A, $\sigma_{\min}$ and $r_{\mathcal{M}}$. ■

Lemma 9 *Let $\alpha > 0$. Assume that (**Ab**) of Section 5.1 is satisfied. Then a positive constant A_d exists, depending only in A, $A_{\mathcal{M},+}$, $\sigma_{\min}$ and α such that, by setting $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$, we have for all $M \in \mathcal{M}_n$,*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq A_d\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right)\right) \leq 2n^{-\alpha} \ . \tag{47}$$

*If moreover, assumptions (**P2**), (**Ab**), (**An**), (**Alb**) and (**Ac$_\infty$**) of the general set of assumptions defined in Section 5.1 hold, then for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$ and for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha)$, we have*

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + A_d\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[\mathrm{p}_2(M)\right]\right) \leq 2n^{-\alpha} \ , \tag{48}$$

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

Proof. We set

$$A_d = \max\left\{4A\sqrt{\alpha};\ \frac{8A^2}{3}\alpha;\ \frac{8A^2\alpha}{\sqrt{A_{\mathcal{M},+}}\sigma_{\min}^2} + \frac{16A^2\alpha}{3A_{\mathcal{M},+}\sigma_{\min}}\right\} \ . \tag{49}$$

Since by (**Ab**) we have $|Y| \leq A$ *a.s.* and $\|s_*\|_\infty \leq A$, it holds $\|s_*\|_\infty = \|\mathbb{E}[Y|X]\|_\infty \leq A$, and so $\|s_M - s_*\|_\infty \leq 2A$. Next, we apply Bernstein's inequality (137) to $\bar{\delta}(M) = (P_n - P)(Ks_M - Ks_*)$. Notice that

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

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

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

and therefore, by (137) we have for all $x > 0$,

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

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

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \sqrt{\frac{16A^2\alpha\ell(s_*, s_M)\ln n}{n}} + \frac{8A^2\alpha\ln n}{3n}\right) \leq 2n^{-\alpha} \ , \tag{50}$$

which gives the first part of Lemma 9 for A_d given in (49). Now, by noticing the fact that $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$, and using it in (50) with $a = \ell(s_*, s_M)$, $b = \frac{4A^2\alpha\ln n}{n}$ and $\eta = D_M^{-1/2}$, we obtain

$$\mathbb{P}\left(\left|\bar{\delta}(M)\right| \geq \frac{\ell(s_*, s_M)}{\sqrt{D_M}} + \left(4\sqrt{D_M} + \frac{8}{3}\right)\frac{A^2\alpha\ln n}{n}\right) \leq 2n^{-\alpha} \ . \tag{51}$$

Then, for a model $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^2 \leq D_M$, we apply Lemma 8 and by (38), it holds for all $n \geq n_0(A_{\mathcal{M},+}, A, A_{cons}, n_1, r_{\mathcal{M}}, \sigma_{\min}, \alpha_{\mathcal{M}})$,

$$\left(1 - L_{A_{\mathcal{M},-},A,\sigma_{\min},r_{\mathcal{M}},\alpha_{\mathcal{M}}}\varepsilon_n^2(M)\right)\frac{D_M}{4n}\mathcal{K}_{1,M}^2 \leq \mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{52}$$

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 (52), $\mathbb{E}\left[\mathrm{p}_2(M)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$ 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 (51), to conclude the proof for the value of A_d given in (49) by simple computations. ■

In order to avoid cumbersome notations in the proofs of Theorems 6 and 5, 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 $n_0((\mathbf{GSA}))$.

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

$$\mathrm{crit}(M) := P_n(Ks_n(M)) + \mathrm{pen}(M) \ , \tag{53}$$

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

$$\mathrm{crit}'(M) := \mathrm{crit}(M) - P_n(Ks_*) - A_t \ . \tag{54}$$

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

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

By setting

$$\begin{aligned}\mathrm{p}_1\left(M\right) &= P\left(Ks_n\left(M\right) - Ks_M\right) \ , \\ \mathrm{p}_2\left(M\right) &= P_n\left(Ks_M - Ks_n\left(M\right)\right) \ , \\ \bar{\delta}\left(M\right) &= \left(P_n - P\right)\left(Ks_M - Ks_*\right)\end{aligned}$$

and

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

we have

$$\ell\left(s_*, s_n\left(M\right)\right) = P_n\left(Ks_n\left(M\right)\right) + \mathrm{p}_1\left(M\right) + \mathrm{p}_2\left(M\right) - \bar{\delta}\left(M\right) - P_n\left(Ks_*\right) \tag{55}$$

and by (54),

$$\mathrm{crit}'\left(M\right) = \ell\left(s_*, s_n\left(M\right)\right) + \left(\mathrm{pen}\left(M\right) - \mathrm{pen}'_{\mathrm{id}}\left(M\right)\right) \ . \tag{56}$$

As $\widehat{M}$ minimizes crit' over $\mathcal{M}_n$, it is therefore sufficient by (56), to control $\mathrm{pen}\left(M\right) - \mathrm{pen}'_{\mathrm{id}}\left(M\right)$ - or equivalently $\mathrm{crit}'\left(M\right)$ - in terms of the excess loss $\ell\left(s_*, s_n\left(M\right)\right)$, 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 $A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_M$, (28) hold and

$$\left|\mathrm{p}_1\left(M\right) - \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n\left(M\right)\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \tag{57}$$

$$\left|\mathrm{p}_2\left(M\right) - \mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n^2\left(M\right)\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \tag{58}$$

$$\left|\bar{\delta}\left(M\right)\right| \leq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + L_{(\mathbf{GSA})}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \tag{59}$$

$$\left|\bar{\delta}\left(M\right)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{60}$$

- For all models $M \in \mathcal{M}_n$ of dimension D_M such that $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$, (29) holds together with

$$\left|\bar{\delta}\left(M\right)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{61}$$

$$\mathrm{p}_2\left(M\right) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{\left(\ln n\right)^3}{n} \tag{62}$$

$$\mathrm{p}_1\left(M\right) \leq L_{(\mathbf{GSA})}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA})}\frac{\left(\ln n\right)^3}{n} \tag{63}$$

By (34), (35), (36) and (37) in Remark 7, Lemma 8, Lemma 9 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, and since (28) holds with probability at least $1 - A_p n^{-2}$, we get for all $n \geq n_0\left((\mathbf{GSA})\right)$,

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

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

We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$. Notice that (59) implies by (33) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$, for all $n \geq n_0$ (($\mathbf{GSA}$)),

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

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

$$\begin{aligned} &\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \\ &\leq \left|\mathrm{p}_1(M) + \mathrm{p}_2(M) - \mathrm{pen}(M)\right| + \left|\bar{\delta}(M)\right| \\ &\leq \left|\mathrm{p}_1(M) + \mathrm{p}_2(M) - 2\mathbb{E}\left[\mathrm{p}_2(M)\right]\right| + \left(L_{(\mathbf{GSA})} \varepsilon_n(M) + \delta\right) \mathbb{E}\left[\ell(s_*, s_M) + \mathrm{p}_2(M)\right] \\ &\leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \mathbb{E}\left[\ell(s_*, s_M) + \mathrm{p}_2(M)\right] \ . \end{aligned} \tag{64}$$

Now notice that using (**P2**) in (33) 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 $n \geq n_0$ (($\mathbf{GSA}$)),

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\ell(s_*, s_M) + \mathrm{p}_2(M)\right] \\ &\leq \ell(s_*, s_n(M)) + \left|\mathrm{p}_1(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \\ &\leq \ell(s_*, s_n(M)) + \frac{L_{(\mathbf{GSA})} \varepsilon_n(M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n(M)} \mathrm{p}_1(M) \qquad \text{by (57)} \\ &\leq \frac{1 + L_{(\mathbf{GSA})} \varepsilon_n(M)}{1 - L_{(\mathbf{GSA})} \varepsilon_n(M)} \ell(s_*, s_n(M)) \\ &\leq \left(1 + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \end{aligned} \tag{65}$$

Hence, using (65) in (64), we have on Ω_n for all models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_M$ and for all $n \geq n_0$ (($\mathbf{GSA}$)),

$$\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \leq \left(\delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \tag{66}$$

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 (56) and (66),

$$\left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \leq \mathrm{crit}'(M) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n(M)\right) \ell(s_*, s_n(M)) \ . \tag{67}$$

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 (29), (61) and (62), it holds on Ω_n, for any $\tau > 0$ and for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\begin{aligned} &\left|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)\right| \\ &\leq \mathrm{p}_1(M) + \mathrm{p}_2(M) + \left|\mathrm{pen}(M)\right| + \left|\bar{\delta}(M)\right| \\ &\leq L_{(\mathbf{GSA})} \frac{(\ln n)^3}{n} + A_r \frac{\ell(s_*, s_M)}{(\ln n)^2} + A_r \frac{(\ln n)^3}{n} + L_{(\mathbf{GSA})} \left(\sqrt{\frac{\ell(s_*, s_M) \ln n}{n}} + \frac{\ln n}{n} \right) \\ &\leq L_{(\mathbf{GSA}), A_r} \left(\frac{(\ln n)^3}{n} + \frac{\ell(s_*, s_M)}{(\ln n)^2} \right) + \tau \ell(s_*, s_M) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \\ &\leq L_{(\mathbf{GSA}), A_r} \left(\frac{(\ln n)^3}{n} + \frac{\ell(s_*, s_M)}{(\ln n)^2} \right) + \tau \ell(s_*, s_n(M)) + \left(\tau^{-1} + 1\right) L_{(\mathbf{GSA})} \frac{\ln n}{n} \ . \end{aligned} \tag{68}$$

Hence, by taking $\tau = (\ln n)^{-2}$ in (68) we get that for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$, it holds on Ω_n,

$$|\mathrm{pen}'_{\mathrm{id}}(M) - \mathrm{pen}(M)| \leq L_{(\mathbf{GSA}),A_r} \left(\frac{\ell(s_*, s_n(M))}{(\ln n)^2} + \frac{(\ln n)^3}{n} \right) . \tag{69}$$

Moreover, by (56) and (69), we have on the event Ω_n, for all $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+} (\ln n)^3$,

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

$$\leq \left(1 + L_{(\mathbf{GSA}),A_r} (\ln n)^{-2}\right) \ell(s_*, s_n(M)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} . \tag{71}$$

Oracle inequalities:

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

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \{\ell(s_*, s_n(M))\} . \tag{72}$$

By Lemmas 10 and 11 below, we control on Ω_n the dimensions of the selected model $\widehat{M}$ and the oracle model M_*. More precisely, by (84) and (86), we have on Ω_n, for any $\eta \in \left(0, \beta_+ / \left(1+\beta_+\right)\right)$ and for all $n \geq n_0((\mathbf{GSA}), \eta, \delta)$,

$$D_{\widehat{M}} \leq n^{1/\left(1+\beta_+\right)+\eta} , \tag{73}$$

$$D_{M_*} \leq n^{1/\left(1+\beta_+\right)+\eta} . \tag{74}$$

Now, from (73) we distinguish two cases in order to control $\mathrm{crit}'\left(\widehat{M}\right)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{M}} \leq n^{1/\left(1+\beta_+\right)+\eta}$, we get by (67), for all $n \geq n_0((\mathbf{GSA}))$,

$$\mathrm{crit}'\left(\widehat{M}\right) \geq \left(1 - \delta - L_{(\mathbf{GSA})} \varepsilon_n\left(\widehat{M}\right)\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) . \tag{75}$$

Otherwise, if $D_{\widehat{M}} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (70),

$$\left(1 - L_{(\mathbf{GSA}),A_r} (\ln n)^{-2}\right) \ell\left(s_*, s_n\left(\widehat{M}\right)\right) - L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} \leq \mathrm{crit}'\left(\widehat{M}\right) . \tag{76}$$

In all cases, we have by (75) and (76), for all $n \geq n_0((\mathbf{GSA}))$,

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

$$- L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} . \tag{77}$$

Similarly, from (74) we distinguish two cases in order to control $\mathrm{crit}'(M_*)$. If $A_{\mathcal{M},+} (\ln n)^3 \leq D_{M_*} \leq n^{1/\left(1+\beta_+\right)+\eta}$, we get by (67), for all $n \geq n_0((\mathbf{GSA}))$,

$$\mathrm{crit}'(M_*) \leq \left(1 + \delta + L_{(\mathbf{GSA})} \varepsilon_n(M_*)\right) \ell(s_*, s_n(M_*)) . \tag{78}$$

Otherwise, if $D_{M_*} \leq A_{\mathcal{M},+} (\ln n)^3$, we get by (71),

$$\mathrm{crit}'(M_*) \leq \left(1 + L_{(\mathbf{GSA}),A_r} (\ln n)^{-2}\right) \ell(s_*, s_n(M_*)) + L_{(\mathbf{GSA}),A_r} \frac{(\ln n)^3}{n} . \tag{79}$$

In all cases, we deduce from (78) and (79) that we have for all $n \geq n_0\left((\mathbf{GSA}),\delta\right)$,

$$\mathrm{crit}'\left(M_*\right) \leq \left(1+\delta+L_{(\mathbf{GSA}),A_r}\left(\left(\ln n\right)^{-2}+\sup_{M\in\mathcal{M}_n,\ A_{\mathcal{M},+}(\ln n)^3\leq D_M\leq n^{1/\left(1+\beta_+\right)+\eta}}\varepsilon_n\left(M\right)\right)\right)\ell\left(s_*,s_n\left(M_*\right)\right)$$
$$+L_{(\mathbf{GSA}),A_r}\frac{\left(\ln n\right)^3}{n}\ . \tag{80}$$

Hence, by setting

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

we have by (33), for all $n \geq n_0\left((\mathbf{GSA}),\eta,\delta\right)$,

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

and we deduce from (77) and (80), since $\frac{1}{1-x}\leq 1+2x$ for all $x\in\left[0,\frac{1}{2}\right)$, that for all $n \geq n_0\left((\mathbf{GSA}),\eta,\delta\right)$, it holds on Ω_n,

$$\begin{aligned}\ell\left(s_*,s_n\left(\widehat{M}\right)\right) &\leq \left(\frac{1+\delta+\theta_n}{1-\delta-\theta_n}\right)\ell\left(s_*,s_n\left(M_*\right)\right)+\frac{L_{(\mathbf{GSA}),A_r}}{1-\delta-\theta_n}\frac{\left(\ln n\right)^3}{n}\\ &\leq \left(\frac{1+\delta}{1-\delta}+\frac{5\theta_n}{\left(1-\delta\right)^2}\right)\ell\left(s_*,s_n\left(M_*\right)\right)+L_{(\mathbf{GSA}),A_r}\frac{\left(\ln n\right)^3}{n}\ .\end{aligned} \tag{81}$$

Inequality (31) is now proved.
It remains to prove the second part of Theorem 6. We assume that assumption (**Ap**) holds. From Lemmas 10 and 11, we have that for any $\frac{1}{2}>\eta>\left(1-\beta_+\right)_+/2$ and for all $n\geq n_0\left((\mathbf{GSA}),C_-,\beta_-,\eta,\delta\right)$, it holds on Ω_n,

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta}\ , \tag{82}$$
$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{M_*} \leq n^{1/2+\eta}\ . \tag{83}$$

Now, using (75) and (78), by the same kind of computations leading to (81), we deduce that it holds on Ω_n, for all $n\geq n_0\left((\mathbf{GSA}),C_-,\beta_-,\eta,\delta\right)$,

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

Thus inequality (32) is proved and Theorem 6 follows. ■

Lemma 10 (Control on the dimension of the selected model) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta>\left(1-\beta_+\right)_+/2$. If $n\geq n_0\left((\mathbf{GSA}),\eta,\delta\right)$ then, on the event Ω_n defined in the proof of Theorem 6, it holds*

$$D_{\widehat{M}} \leq n^{1/2+\eta}\ . \tag{84}$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{\widehat{M}} \leq n^{1/2+\eta}\ . \tag{85}$$

Lemma 11 (Control on the dimension of oracle models) *Assume that the general set of assumptions (**GSA**) hold. Let $\eta > \left(1-\beta_+\right)_+/2$. If $n \geq n_0\left((\textbf{GSA}),\eta\right)$ then, on the event Ω_n defined in the proof of Theorem 6, it holds*

$$D_{M_*} \leq n^{1/2+\eta} \ . \tag{86}$$

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

$$A_{\mathcal{M},+}\left(\ln n\right)^3 \leq D_{M_*} \leq n^{1/2+\eta} \ . \tag{87}$$

Proof of Lemma 10. Recall that $\widehat{M}$ minimizes

$$\mathrm{crit}'\left(M\right) = \mathrm{crit}\left(M\right) - P_n K s_* = \ell\left(s_*, s_M\right) - \mathrm{p}_2\left(M\right) + \bar{\delta}\left(M\right) + \mathrm{pen}\left(M\right) \tag{88}$$

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

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

$$\begin{aligned}
\ell\left(s_*, s_M\right) &\geq C_- A_{\mathcal{M},+}^{-\beta_-}\left(\ln n\right)^{-3\beta_-} \ \text{ by } (\textbf{Ap}) \\
\mathrm{pen}\left(M\right) &\geq -A_r\left(\frac{\ell\left(s_*, s_M\right)}{\left(\ln n\right)^2} + \frac{\left(\ln n\right)^3}{n}\right) \\
\mathrm{p}_2\left(M\right) &\leq L_{(\textbf{GSA})}\frac{\left(\ln n\right)^3}{n} \qquad\qquad \text{from (62)} \\
\bar{\delta}\left(M\right) &\geq -L_{(\textbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \ \text{from (61)}.
\end{aligned}$$

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

$$\mathrm{crit}'\left(M\right) \geq \frac{C_- A_{\mathcal{M},+}^{-\beta_-}}{2}\left(\ln n\right)^{-3\beta_-} \ . \tag{89}$$

2. Lower bound for large models: let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/\left(1+\beta_+\right)+\eta}$. From (28) and (58) we have on Ω_n,

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

   Using (**P2**) and the fact that $D_M \geq n^{1/\left(1+\beta_+\right)+\eta}$ in (33), we deduce that for all $n \geq n_0\left((\textbf{GSA}), \eta, \delta, \beta_+\right)$, $L_{(\textbf{GSA})}\varepsilon_n^2\left(M\right) \leq \frac{1}{2}\left(1-\delta\right)$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 8 that for all $n \geq n_0\left((\textbf{GSA}), \eta\right)$, $\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right] \geq \frac{\sigma_{\min}^2}{2}\frac{D_M}{n}$. Consequently, it holds for all $n \geq n_0\left((\textbf{GSA}), \eta, \delta, \beta_+\right)$,

$$\mathrm{pen}\left(M\right) - \mathrm{p}_2\left(M\right) \geq \frac{\sigma_{\min}^2}{4}\left(1-\delta\right)\frac{D_M}{n} - C_+ D_M^{-\beta_+} \geq \left(1-\delta\right)L_{(\textbf{GSA})} n^{-\frac{\beta_+}{1+\beta_+}+\eta} \tag{90}$$

   From (60) it holds on Ω_n,

$$\bar{\delta}\left(M\right) \geq -L_{(\textbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \geq -L_{(\textbf{GSA})}\left(n^{-\frac{1+2\beta_+}{2\left(1+\beta_+\right)}}\sqrt{\ln n} + \frac{\ln n}{n}\right) \ . \tag{91}$$

   Hence, we deduce from (88), (90) and (91) that we have on Ω_n, for all $n \geq n_0\left((\textbf{GSA}), \eta, \delta, \beta_+\right)$,

$$\mathrm{crit}'\left(M\right) \geq \left(1-\delta\right)L_{(\textbf{GSA})} n^{-\frac{\beta_+}{1+\beta_+}+\eta} \ . \tag{92}$$

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

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

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

$$\ell(s_*, s_{M_0}) \leq C_+ n^{-\beta_+/(1+\beta_+)} . \tag{93}$$

By (59), we have on Ω_n, for all $n \geq n_0((\mathbf{GSA}), \eta)$,

$$\left|\bar{\delta}(M_0)\right| \leq \frac{\ell(s_*, s_{M_0})}{\sqrt{D_{M_0}}} + L_{(\mathbf{GSA})} \frac{\ln n}{\sqrt{D_{M_0}}} \mathbb{E}[\mathrm{p}_2(M_0)] \leq L_{(\mathbf{GSA})} n^{-\frac{1+2\beta_+}{2(1+\beta_+)}} \ln(n) \tag{94}$$

and by (28),

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

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

$$\begin{aligned} \mathrm{crit}'(M_0) &\leq \ell(s_*, s_{M_0}) + \left|\bar{\delta}(M_0)\right| + \mathrm{pen}(M_0) \\ &\leq L_{(\mathbf{GSA})} n^{-\beta_+/(1+\beta_+)} . \end{aligned} \tag{95}$$

To conclude, notice that the upper bound (95) is smaller than the lower bound given in (92) for all $n \geq n_0\left((\mathbf{GSA}), \eta, \delta, \beta_+\right)$. Hence, points 2 and 3 above yield inequality (84). Moreover, the upper bound (95) is smaller than lower bounds given in (89), derived by using (**Ap**), and (92), for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \beta_+, \eta, \delta\right)$. This thus gives (85) and Lemma 10 is proved. ■

Proof of Lemma 11. By definition, M_* minimizes

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

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

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

$$\ell(s_*, s_n(M)) \geq \ell(s_*, s_M) \geq C_- A_{\mathcal{M},+}^{-\beta_-} (\ln n)^{-3\beta_-} \text{ by } (\mathbf{Ap}). \tag{96}$$

2. Lower bound of $\ell(s_*, s_n(M))$ for large models: let $M \in \mathcal{M}_n$ be such that $D_M \geq n^{1/(1+\beta_+)+\eta}$. From (57) we get on Ω_n,

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

Using (**P2**) and the fact that $D_M \geq n^{1/(1+\beta_+)+\eta}$ in (33), we deduce that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $L_{(\mathbf{GSA})} \varepsilon_n(M) \leq \frac{1}{2}$ and as by (**An**), $\mathcal{K}_{1,M} \geq 2\sigma_{\min}$ we also deduce from Lemma 8 that for all $n \geq n_0((\mathbf{GSA}), \eta)$, $\mathbb{E}[\mathrm{p}_2(M)] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n}$. Consequently, it holds for all $n \geq n_0((\mathbf{GSA}), \eta)$, on the event Ω_n,

$$\ell(s_*, s_n(M)) \geq \mathrm{p}_1(M) \geq \frac{\sigma_{\min}^2}{4} \frac{D_M}{n} \geq \frac{\sigma_{\min}^2}{4} n^{-\beta_+/(1+\beta_+)+\eta} . \tag{97}$$

3. A better model exists for $\ell(s_*, s_n(M))$: 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 $n \geq n_0((\mathbf{GSA}), \eta)$,

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

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

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

and by (57)

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

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (33), for all $n \geq n_0((\mathbf{GSA}))$ it holds $\varepsilon_n(M) \leq 1$, we deduce from Lemma 8 that for all $n \geq n_0((\mathbf{GSA}))$, on the event Ω_n,

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

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

$$\begin{aligned}\ell(s_*, s_n(M_0)) &= \ell(s_*, s_{M_0}) + \mathrm{p}_1(M_0)\\ &\leq L_{(\mathbf{GSA})}n^{-\beta_+/(1+\beta_+)} .\end{aligned} \tag{98}$$

The upper bound (98) is smaller than the lower bound (97) for all $n \geq n_0((\mathbf{GSA}), \eta)$, and this gives (86). If (**Ap**) hold, then the upper bound (98) is smaller than the lower bounds (96) and (97) for all $n \geq n_0\left((\mathbf{GSA}), C_-, \beta_-, \beta_+, \eta\right)$, which proves (87) and allows to conclude the proof of Lemma 11. ■

Proof of Theorem 5. As in the proof of Theorem 6, we consider the event Ω'_n of probability at least $1 - L_{c_{\mathcal{M}}, A_p}n^{-2}$ for all $n \geq n_0((\mathbf{GSA}))$, on which: (26) holds and

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

$$\left|\mathrm{p}_1(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] , \tag{99}$$

$$\left|\mathrm{p}_2(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] . \tag{100}$$

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

$$\mathrm{p}_2(M) \leq L_{(\mathbf{GSA})}\frac{(\ln n)^2}{n} . \tag{101}$$

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

$$\left|\bar{\delta}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right) . \tag{102}$$

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

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

$$\mathrm{crit}'(M) = \mathrm{crit}(M) - P_nKs_* = \ell(s_*, s_M) - \mathrm{p}_2(M) + \bar{\delta}(M) + \mathrm{pen}(M) . \tag{103}$$

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

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

We have

$$\ell(s_*, s_M) + \mathrm{pen}(M) \geq 0 \tag{104}$$

and from (102), as $\ell(s_*, s_M) \leq 4A^2$ by (**Ab**), we get on Ω'_n, for all $n \geq n_0((\mathbf{GSA}),d)$,

$$\begin{aligned}\bar{\delta}(M) &\geq -L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_M)\ln n}{n}} + \frac{\ln n}{n}\right)\\ &\geq -L_{(\mathbf{GSA})}\sqrt{\frac{\ln n}{n}}\\ &\geq -d \times A^2A_{rich}(\ln n)^{-2} .\end{aligned} \tag{105}$$

Then, if $D_M \geq A_{\mathcal{M},+} (\ln n)^2$, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (33), for all $n \geq n_0$ ((**GSA**)) it holds $L_{(\mathbf{GSA})} \varepsilon_n (M) \leq 1$, we deduce from (100) and Lemma 8 that for all $n \geq n_0$ ((**GSA**),d),

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

Whenever $D_M \leq A_{\mathcal{M},+} (\ln n)^2$, (101) gives that, for all $n \geq n_0$ ((**GSA**),d), on the event Ω'_n,

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

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

$$-\mathrm{p}_2 (M) \geq -d \times 36A^2 A_{rich} (\ln n)^{-2} \ , \tag{106}$$

and finally, by using (104), (105) and (106) in (103), we deduce that on Ω'_n, for all $n \geq n_0$ ((**GSA**), d),

$$\mathrm{crit}' (M) \geq -d \times 37A^2 A_{rich} (\ln n)^{-2} \ . \tag{107}$$

2. There exists a better model for $\mathrm{crit}' (M)$: By (**P3**), for all $n \geq n_0 (A_{\mathcal{M},+}, A_{rich})$ a model $M_1 \in \mathcal{M}_n$ exists such that
$$A_{\mathcal{M},+} (\ln n)^2 \leq \frac{A_{rich} n}{(\ln n)^2} \leq D_{M_1} \ .$$
We then have on Ω'_n,
$$\begin{aligned}
\ell (s_*, s_{M_1}) &\leq A_{rich}^{-\beta_+} (\ln n)^{2\beta_+} n^{-\beta_+} && \text{by } (\mathbf{Ap}_u) \\
\mathrm{p}_2 (M_1) &\geq \left(1 - L_{(\mathbf{GSA})} \varepsilon_n^2 (M_1)\right) \mathbb{E} [\mathrm{p}_2 (M_1)] && \text{by (100)} \\
\mathrm{pen} (M_1) &\leq A_{\mathrm{pen}} \mathbb{E} [\mathrm{p}_2 (M_1)] && \text{by (26)} \\
\left|\bar{\delta} (M_1)\right| &\leq L_{(\mathbf{GSA})} \sqrt{\frac{\ln n}{n}} && \text{by (102) and (}\mathbf{Ab}\text{)}
\end{aligned}$$
and therefore,
$$\mathrm{crit}' (M_1) \leq \left(-1 + A_{\mathrm{pen}} + L_{(\mathbf{GSA})} \varepsilon_n^2 (M_1)\right) \mathbb{E} [\mathrm{p}_2 (M_1)] + L_{(\mathbf{GSA})} \sqrt{\frac{\ln n}{n}} + A_{rich}^{-\beta_+} \frac{(\ln n)^{2\beta_+}}{n^{\beta_+}} \ . \tag{108}$$
Hence, as $-1 + A_{\mathrm{pen}} < 0$, and as by (33), (**An**) and Lemma 8 it holds for all $n \geq n_0$ ((**GSA**), A_{pen})
$$L_{(\mathbf{GSA})} \varepsilon_n^2 (M_1) \leq \frac{1 - A_{\mathrm{pen}}}{2} \quad \text{and} \quad \mathbb{E} [\mathrm{p}_2 (M_1)] \geq \frac{\sigma_{\min}^2}{2} \frac{D_M}{n} \geq \frac{\sigma_{\min}^2 A_{rich}}{2} (\ln n)^{-2} \ ,$$
we deduce from (108) that on Ω'_n, for all $n \geq n_0$ ((**GSA**), A_{pen}),
$$\mathrm{crit}' (M_1) \leq -\frac{1}{4} (1 - A_{\mathrm{pen}}) \sigma_{\min}^2 A_{rich} (\ln n)^{-2} \ . \tag{109}$$

Now, by taking
$$0 < d = \left(\frac{1}{149} (1 - A_{\mathrm{pen}}) \left(\frac{\sigma_{\min}}{A}\right)^2\right) \wedge \frac{1}{2} < 1 \tag{110}$$
and by comparing (107) and (109), we deduce that on Ω'_n, for all $n \geq n_0$ ((**GSA**), A_{pen}), for all $M \in \mathcal{M}_n$ such that $D_M \leq d A_{rich} n (\ln n)^{-2}$,
$$\mathrm{crit}' (M_1) < \mathrm{crit}' (M)$$
and so
$$D_{\widehat{M}} > d A_{rich} n (\ln n)^{-2} \ . \tag{111}$$

Excess Loss of $s_n\left(\widehat{M}\right)$. We take d with the value given in (110). 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 (33), (**P2**), (**An**) and Lemma 8, it holds on Ω_n' for all $n \geq n_0\left((\mathbf{GSA}), A_{\rm pen}\right)$, using (99),

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

By (111), we thus get that on Ω_n', for all $n \geq n_0\left((\mathbf{GSA}), A_{\rm pen}\right)$,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \geq \frac{d\sigma_{\min}^2 A_{rich}}{2}\left(\ln n\right)^{-2} \ . \tag{112}$$

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

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

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

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

In addition, by (57),

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

Hence, as $\mathcal{K}_{1,M} \leq 6A$ by (**Ab**) and as, by (33), for all $n \geq n_0\left((\mathbf{GSA})\right)$ it holds $\varepsilon_n\left(M\right) \leq 1$, we deduce from Lemma 8 that for all $n \geq n_0\left((\mathbf{GSA})\right)$

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

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

$$\ell\left(s_*, s_n\left(M_0\right)\right) \leq L_{(\mathbf{GSA})} n^{-\beta_+/\left(1+\beta_+\right)} \tag{113}$$

and the ratio between the two bounds (112) and (113) is larger than $n^{\beta_+/\left(1+\beta_+\right)}\left(\ln n\right)^{-3}$ for all $n \geq n_0\left(L_{(\mathbf{GSA})}, A_{\rm pen}\right)$, which yields (27). ■

5.3 Proofs related to Section 4

As in Section 5.2 above, we state the following theorem which, by Section 5.1, is more general than Theorem 4.

Theorem 12 *Assume that the general set of assumptions (**GSA**) of Section 5.1 holds and that there exists $c \in (0,1)$ such that $nc \leq n_1 < n$. Assume moreover that there exists $\tau \in (1,3)$ 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 $A_{\mathcal{M},+} n/\left(\ln n\right)^2 \geq D_M \geq A_{\mathcal{M},+}\left(\ln n\right)^3$ and take $n_2 = n\left(1-c\right)$ if $D_M \leq A_{\mathcal{M},+}\left(\ln n\right)^3$. For any $C > 0$, define for all $M \in \mathcal{M}_n$,*

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

*<u>Favorable case:</u> assume that $n/\left(2n_1\right) < C \leq 3n/\left(2n_1\right)$. Then, for $\frac{1}{2} > \eta > \left(1-\beta_+\right)_+/2$, there exist an integer n_0 depending on c, η, C 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 the set of assumptions (**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 $n \geq n_0\left((\boldsymbol{GSA}), c, \eta, C\right)$, with probability at least $1 - A_6 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(\frac{1+2\left|C\frac{n}{n_1}-1\right|}{1-2\left|C\frac{n}{n_1}-1\right|} + \frac{12\theta_n}{\left(1-2\left|C\frac{n}{n_1}-1\right|\right)^2}\right)\ell\left(s_*, s_n\left(M_*\right)\right) + A_8\frac{\left(\ln n\right)^3}{n} \ . \tag{114}$$

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 $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c, C_-, \beta_-, \eta, C\right)$, with probability at least $1 - A_6 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \leq \left(1+5\theta_n\right)\inf_{M\in\mathcal{M}_n}\left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{115}$$

Unfavorable case: Assume that $C < n/\left(2n_1\right)$. Then there exist a constant $A_1 > 0$ only depending on constants in $\left(\boldsymbol{GSA}\right)$, as well as an integer n_0 and a positive constant A_2 only depending on c, C and on constants in $\left(\boldsymbol{GSA}\right)$ such that, for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c, C\right)$, it holds with probability at least $1 - A_1 n^{-2}$,

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

and

$$\ell\left(s_*, s_n\left(\widehat{M}_{n_1}\right)\right) \geq \frac{n^{\beta_+/\left(1+\beta_+\right)}}{\left(\ln n\right)^3}\inf_{M\in\mathcal{M}_n}\left\{\ell\left(s_*, s_n\left(M\right)\right)\right\} \ . \tag{116}$$

The end of this section is devoted to the proof of Theorem 12.

Lemma 13 *Assume that the set of assumptions $\left(\boldsymbol{GSA}\right)$, stated in Section 5.1, 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 $L = L_{\left(\boldsymbol{GSA}\right),c} > 0$ 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*

$$\mathbb{P}\left(\left|P_{n_2}\left(Ks_{n_1}\left(M\right) - Ks_M\right) - P\left(Ks_{n_1}\left(M\right) - Ks_M\right)\right| \geq L\frac{\sqrt{\left(D_M \vee \ln n\right)\left(\ln n\right)\left(\left(\ln n\right)\left(\ln n_1\right) + n_2\right)}}{n_2\sqrt{n_1}}\right) \leq 12n^{-2-\alpha_{\mathcal{M}}} \ . \tag{117}$$

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

$$\varepsilon_n^{1,2}\left(M\right) = L\frac{n\sqrt{\ln n\left(\left(\ln n\right)\left(\ln n_1\right) + n_2\right)}}{n_2\sqrt{n_1 D_M}} \leq \frac{L}{\left(\ln n\right)^{(\tau-1)/2}} \ , \tag{118}$$

we have for all $n \geq n_0\left(\left(\boldsymbol{GSA}\right), c\right)$,

$$\mathbb{P}\left(\left|P_{n_2}\left(Ks_{n_1}\left(M\right) - Ks_M\right) - P\left(Ks_{n_1}\left(M\right) - Ks_M\right)\right| \geq \varepsilon_n^{1,2}\left(M\right)\mathbb{E}\left[\mathrm{p}_2\left(M\right)\right]\right) \leq 12n^{-2-\alpha_{\mathcal{M}}} \ . \tag{119}$$

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

$$\mathbb{P}\left(\left|P_{n_2}\left(Ks_{n_1}\left(M\right) - Ks_M\right) - P\left(Ks_{n_1}\left(M\right) - Ks_M\right)\right| \geq L\frac{\left(\ln n\right)^2}{n}\right) \leq 12n^{-2-\alpha_{\mathcal{M}}} \ . \tag{120}$$

Proof. By inequality (138) applied with $v_i = (s_{n_1}(M))(\xi_i)$ conditionally to $\left(\xi_j\right)_{j\in I_1}$, we get that for all $x>0$, it holds

$$\mathbb{P}\left(|P_{n_2}(Ks_{n_1}(M)-Ks_M)-P(Ks_{n_1}(M)-Ks_M)|\geq x\,\middle|\left(\xi_j\right),\ j\in I_1\right)\leq 2\exp\left(-\frac{nx^2}{2(v_1+b_1x/3)}\right)\ , \quad (121)$$

where

$$v_1=\mathbb{E}_\xi\left[\left(Ks_{n_1}(M)(\xi)-Ks_M(\xi)\right)^2\right]$$

and $b_1=\|Ks_{n_1}(M)-Ks_M\|_\infty$. We have

$$\begin{aligned} v_1 &= \mathbb{E}_{(X,Y)}\left[\left(2(Y-s_M(X))-s_{n_1}(M)(X)+s_M(X)\right)^2\left(s_{n_1}(M)(X)-s_M(X)\right)^2\right] \\ &\leq (4A+\|s_{n_1}(M)-s_M\|_\infty)^2\,\mathbb{E}_X\left[\left(s_{n_1}(M)(X)-s_M(X)\right)^2\right] \\ &= (4A+\|s_{n_1}(M)-s_M\|_\infty)^2\,P(Ks_{n_1}(M)-Ks_M) \end{aligned} \quad (122)$$

and

$$\begin{aligned} b_1 &= \|(2(Y-s_M(X))-s_{n_1}(M)(X)+s_M(X))(s_{n_1}(M)(X)-s_M(X))\|_\infty \\ &\leq 4A\|s_{n_1}(M)-s_M\|_\infty+\|s_{n_1}(M)-s_M\|_\infty^2\ . \end{aligned} \quad (123)$$

Now, we set $\Omega_v=\{v_1\leq L_v(D_M\vee\ln n_1)/n_1\}$ and $\Omega_b=\left\{b_1\leq L_b\sqrt{D_M\ln n_1/n_1}\right\}$. By integrating (121), it comes for all $x>0$,

$$\begin{aligned} &\mathbb{P}(|P_{n_2}(Ks_{n_1}(M)-Ks_M)-P(Ks_{n_1}(M)-Ks_M)|\geq x) \\ \leq\ & 2\mathbb{E}\left[\exp\left(-\frac{n_2x^2}{2(v_1+b_1x/3)}\right)\mathbf{1}_{\Omega_v\cap\Omega_b}\right]+2\mathbb{P}(\Omega_v^c)+2\mathbb{P}(\Omega_b^c) \\ \leq\ & 2\exp\left(-\frac{n_2x^2}{2\left(L_v(D_M\vee\ln n_1)/n_1+L_bx\sqrt{D_M\ln n_1/n_1}\right)}\right)+2\mathbb{P}(\Omega_v^c)+2\mathbb{P}(\Omega_b^c) \end{aligned}$$

From assumption $(\mathbf{Ac}_\infty)$ and inequality (36), it is possible to choose L_v and L_b, depending among other constants on c, such that for all $n\geq n_0((\mathbf{GSA}),c)$, $2\mathbb{P}(\Omega_v^c)+2\mathbb{P}(\Omega_b^c)\leq 10n^{-2-\alpha_{\mathcal{M}}}$. Thus, we get for $L>0$ large enough and for all $x>0$,

$$\begin{aligned} &\mathbb{P}(|P_{n_2}(Ks_{n_1}(M)-Ks_M)-P(Ks_{n_1}(M)-Ks_M)|\geq x) \\ \leq\ & 2\exp\left(-\frac{n_2x^2}{L\left((D_M\vee\ln n_1)/n_1+x\sqrt{D_M\ln n_1/n_1}\right)}\right)+10n^{-2-\alpha_{\mathcal{M}}}\ . \end{aligned} \quad (124)$$

By taking $x=\sqrt{L\alpha\ln n(D_M\vee\ln n_1)(L\alpha(\ln n)(\ln n_1)+4n_2)}/\left(n_2\sqrt{n_1}\right)>0$ in the latter inequality, it comes

$$\begin{aligned} &\mathbb{P}\left(|P_{n_2}(Ks_{n_1}(M)-Ks_M)-P(Ks_{n_1}(M)-Ks_M)|\geq L\frac{\sqrt{(D_M\vee\ln n_1)(\ln n)((\ln n)(\ln n_1)+n_2)}}{n_2\sqrt{n_1}}\right) \\ &\leq 12n^{-2-\alpha_{\mathcal{M}}}\ , \end{aligned}$$

where $L>0$ depends on the constants in $(\mathbf{GSA})$ and on c. Inequalities (119) and (120) then follow from simple calculations.

■

Remark 14 *It is easy to see that by using the assumption of consistency in sup-norm for a fixed model, stated as (**H5**) in [40], instead of (**Ac**$_\infty$) and by using Theorem 4 of [40] instead of inequality (36), the results established in Lemma 13 are valid with probability bounds proportional to $n^{-\alpha}$, for any $\alpha>0$ (in Lemma 13, we only derive the case of $\alpha=2+\alpha_{\mathcal{M}}$ for convenience).*

Proof of Theorem 12. For any $C > 0$, we have $\mathrm{pen}_{ho,C}(M) = C\left(P_{n_2}\left(Ks_{n_1}(M)\right) - P_{n_1}\left(Ks_{n_1}(M)\right)\right)$ and we set

$$\mathrm{pen}_C(M) = \mathrm{pen}_{ho,C}(M) - C\left(P_{n_2}\left(Ks_*\right) - P_{n_1}\left(Ks_*\right)\right) \ .$$

It is worth noting that $C\left(P_{n_2}\left(Ks_*\right) - P_{n_1}\left(Ks_*\right)\right)$ is a quantity independent of M, when M varies in $\mathcal{M}_n$. Hence, the procedure defined by pen_C gives the same result as the hold-out procedure defined by $\mathrm{pen}_{ho,C}$. It will be convenient for our analysis to consider pen_C instead of $\mathrm{pen}_{ho,C}$. As a matter of fact, we derive Theorem 12 as a corollary of Theorems 5 and 6 applied with $\mathrm{pen} \equiv \mathrm{pen}_C$, through the use of Lemma 13.

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

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

where

$$\mathrm{p}_1^{n_2}(M) = P_{n_2}\left(Ks_{n_1}(M) - Ks_M\right) \ , \ \mathrm{p}_2^{n_1}(M) = P_{n_1}\left(Ks_M - Ks_{n_1}(M)\right) \ , \ \bar{\delta}^{n_i}(M) = \left(P_{n_i} - P\right)\left(Ks_M - Ks_*\right) \ .$$

Let Ω_n be the event on which:

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

$$\left|\mathrm{p}_1(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{125}$$

$$\left|\mathrm{p}_2(M) - \mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA})}\varepsilon_n^2(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{126}$$

  together with

$$\left|\mathrm{p}_1^{n_2}(M) - \frac{n}{n_1}\mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA}),c}\left[\varepsilon_n^{1,2}(M) + \varepsilon_n(M)\right]\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{127}$$

$$\left|\mathrm{p}_2^{n_1}(M) - \frac{n}{n_1}\mathbb{E}\left[\mathrm{p}_2(M)\right]\right| \leq L_{(\mathbf{GSA}),c}\varepsilon_n^2(M)\,\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{128}$$

$$\left|\bar{\delta}^{n_1}(M)\right| \leq \frac{\ell\left(s_*, s_M\right)}{\sqrt{D_M}} + L_{(\mathbf{GSA}),c}\frac{\ln n}{\sqrt{D_M}}\mathbb{E}\left[\mathrm{p}_2(M)\right] \tag{129}$$

$$\left|\bar{\delta}^{n_2}(M)\right| \leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n_2}{n_2}} + \frac{\ln n_2}{n_2}\right) \tag{130}$$

- 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

$$\left|\bar{\delta}^{n_1}(M)\right| \leq L_{(\mathbf{GSA}),c}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{131}$$

$$\left|\bar{\delta}^{n_2}(M)\right| \leq L_{(\mathbf{GSA}),c}\left(\sqrt{\frac{\ell\left(s_*, s_M\right)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{132}$$

$$\mathrm{p}_2^{n_1}(M) \leq L_{(\mathbf{GSA}),c}\frac{D_M \vee \ln n}{n} \leq L_{(\mathbf{GSA}),c}\frac{\left(\ln n\right)^3}{n} \tag{133}$$

$$\mathrm{p}_1^{n_2}(M) \leq L_{(\mathbf{GSA}),c}\left(\frac{\left(\ln n\right)^2}{n} + \frac{D_M \vee \ln n}{n}\right) \leq L_{(\mathbf{GSA}),c}\frac{\left(\ln n\right)^3}{n} \tag{134}$$

By (34), (35), (36) and (37) in Remark 7, Lemma 8 and Lemma 13, we get for all $n \geq n_0\left((\mathbf{GSA}),c\right)$,

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

Let us assume that $n_1/(2n) < C \leq 3n_1/(2n)$ in order to prove the favorable case of Theorem 12.
We consider models $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$. Notice that (129) implies by (33) that, for all $M \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_M$,

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

In addition, from (130), Lemma 8 and the fact that $n(\ln n)^\tau/D_M \leq n_2$, we get that for all $n \geq n_0((\mathbf{GSA}))$,

$$\begin{aligned}\left|\bar{\delta}^{n_2}(M)\right| &\leq L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*,s_M)\ln n_2}{n_2}}+\frac{\ln n_2}{n_2}\right)\\ &\leq L_{(\mathbf{GSA})}\left(\frac{\ell(s_*,s_M)}{(\ln n)^{(\tau-1)/2}}+\frac{\ln n_2}{n_2}(\ln n)^{(\tau-1)/2}\right)\\ &\leq L_{(\mathbf{GSA})}(\ln n)^{(1-\tau)/2}\left(\ell(s_*,s_M)+\mathbb{E}[p_2(M)]\right)\ .\end{aligned}$$

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

$$\begin{aligned}&|\mathrm{pen}_C(M)-2\mathbb{E}[\mathrm{p}_2(M)]|\\ &\left|\mathrm{pen}_C(M)-2C\frac{n}{n_1}\mathbb{E}[\mathrm{p}_2(M)]\right|+2\left|C\frac{n}{n_1}-1\right|\mathbb{E}[\mathrm{p}_2(M)]\\ \leq\ & C\left(\left|\mathrm{p}_1^{n_2}(M)-\frac{n}{n_1}\mathbb{E}[\mathrm{p}_2(M)]\right|+\left|\mathrm{p}_2^{n_1}(M)-\frac{n}{n_1}\mathbb{E}[\mathrm{p}_2(M)]\right|\right)\\ &+\left|\bar{\delta}^{n_1}(M)\right|+\left|\bar{\delta}^{n_2}(M)\right|+2\left|C\frac{n}{n_1}-1\right|\mathbb{E}[\mathrm{p}_2(M)]\\ \leq\ & \left(L_{(\mathbf{GSA}),c}\left(\varepsilon_n^{1,2}(M)+\varepsilon_n(M)+(\ln n)^{(1-\tau)/2}\right)+2\left|C\frac{n}{n_1}-1\right|\right)\left(\ell(s_*,s_M)+\mathbb{E}[p_2(M)]\right)\end{aligned}\quad(135)$$

Hence, inequality (28) of Theorem 6 is satisfied on Ω_n by taking

$$\delta = 2\left|C\frac{n}{n_1}-1\right|+L_{(\mathbf{GSA}),c}\left(\varepsilon_n^{1,2}(M)+\varepsilon_n(M)+(\ln n)^{(1-\tau)/2}\right)\ .$$

Moreover, we have $\delta \in [0,1)$ for all $n \geq n_0((\mathbf{GSA}),c,\tau,C)$.
Let us now consider models $M \in \mathcal{M}_n$ such that $D_M \leq A_{\mathcal{M},+}(\ln n)^3$. By (131), (132), (134) and (133), we have on Ω_n,

$$\begin{aligned}|\mathrm{pen}_C(M)| &= C\left|\mathrm{p}_1^{n_2}(M)+\mathrm{p}_2^{n_1}(M)+\bar{\delta}^{n_2}(M)-\bar{\delta}^{n_1}(M)\right|\\ &\leq L_{(\mathbf{GSA}),c}\left(\sqrt{\frac{\ell(s_*,s_M)\ln n}{n}}+\frac{(\ln n)^3}{n}\right)\\ &\leq L_{(\mathbf{GSA}),c}\left(\frac{\ell(s_*,s_M)}{(\ln n)^2}+\frac{(\ln n)^3}{n}\right)\end{aligned}\quad(136)$$

Inequality (136) implies that inequality (29) of Theorem 6 is satisfied with $A_r = L_{(\mathbf{GSA}),c}$. From (135) and (136), we thus apply Theorem 6 with $A_p = L_{A_p,c_{\mathcal{M}}}$, and this gives the favorable case of Theorem 12, where

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

Let us now assume that $C \in (0, n_1/(2n))$ so that we can prove the unfavorable case of Theorem 12. Let M_1 be the model defined on assumption $(\mathbf{P}_3)$. We have

$$\mathrm{pen}_C(M_1) = C\left(\mathrm{p}_1^{n_2}(M_1) + \mathrm{p}_2^{n_1}(M_1) + \bar{\delta}^{n_2}(M_1) - \bar{\delta}^{n_1}(M_1)\right) .$$

So, on Ω_n, it holds

$$\begin{aligned}\mathrm{pen}_C(M_1) \quad \geq \quad & \left(2C\frac{n}{n_1} - L_{(\mathbf{GSA}),c}\left(\varepsilon_n^{1,2}(M_1) + \varepsilon_n(M_1) + \frac{\ln n}{\sqrt{D_{M_1}}}\right)\right)\mathbb{E}\left[\mathrm{p}_2(M_1)\right] \\ & -\frac{\ell(s_*, s_{M_1})}{\sqrt{D_{M_1}}} - L_{(\mathbf{GSA})}\left(\sqrt{\frac{\ell(s_*, s_{M_1})\ln n_2}{n_2}} + \frac{\ln n_2}{n_2}\right) .\end{aligned}$$

As $D_{M_1} \geq A_{rich} n (\ln n)^{-2}$, it holds $\mathbb{E}\left[\mathrm{p}_2(M_1)\right] \geq L_{(\mathbf{GSA})}(\ln n)^{-2}$, $\ell(s_*, s_{M_1}) \leq C_+\left(A_{rich} n (\ln n)^{-2}\right)^{-\beta_+}$ and we also have

$$\sqrt{\frac{\ell(s_*, s_{M_1})\ln n_2}{n_2}} + \frac{\ln n_2}{n_2} \leq \frac{\ell(s_*, s_{M_1})}{(\ln n)^{(\tau-1)/2}} + \frac{D_{M_1}}{n(\ln n)^{(\tau-1)/2}} .$$

Hence, for all $n \geq n_0((\mathbf{GSA}), c, C)$, we have

$$\mathrm{pen}_C(M_1) \geq 0 .$$

Moreover, on Ω_n,

$$\begin{aligned}\mathrm{pen}_C(M_1) \quad \leq \quad & \left(2C\frac{n}{n_1} + L_{(\mathbf{GSA}),c}\left(\varepsilon_n^{1,2}(M_1) + \varepsilon_n(M_1) + \frac{\ln n}{\sqrt{D_{M_1}}}\right)\right)\mathbb{E}\left[\mathrm{p}_2(M_1)\right] \\ & + \frac{\ell(s_*, s_{M_1})}{\sqrt{D_{M_1}}} + L_{(\mathbf{GSA})}\left(\frac{\ell(s_*, s_{M_1})}{(\ln n)^{(\tau-1)/2}} + \frac{D_{M_1}}{n(\ln n)^{(\tau-1)/2}}\right) \\ \leq \quad & \left(\frac{1}{2} + C\frac{n}{n_1}\right)\mathbb{E}\left[\mathrm{p}_2(M_1)\right] ,\end{aligned}$$

where the latter inequality holds for all $n \geq n_0((\mathbf{GSA}), c, C)$. Condition (26) of Theorem 5 is thus satisfied with $A_{\mathrm{pen}} = \frac{1}{2} + C\frac{n}{n_1} < 1$. By applying Theorem 5, we get the unfavorable case of Theorem 12, and so Theorem 12 is proved.

5.4 Probabilistic Tools

We recall here the main probabilistic results that are instrumental in our proofs.
The following tool is the well known Bernstein's inequality, that can be found for instance in [38], Proposition 2.9.

Theorem 15 *(Bernstein's inequality) Let $(\upsilon_1, ..., \upsilon_n)$ be independent real valued random variables and define*

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

Assuming that

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

and

$$|\upsilon_i| \leq b \quad a.s.$$

we have, for every $x > 0$,

$$\mathbb{P}\left(|S| \geq \sqrt{2v\frac{x}{n}} + \frac{bx}{3n}\right) \leq 2\exp\left(-x\right) \tag{137}$$

and also

$$\mathbb{P}\left(|S| \geq x\right) \leq 2\exp\left(-\frac{nx^2}{2\left(v + bx/3\right)}\right) \ . \tag{138}$$

We now turn to concentration inequalities for the empirical process around its mean. Bousquet's inequality [23] provides optimal constants for the deviations above the mean. Klein-Rio's inequality [29] gives sharp constants for the deviations below the mean, that slightly improve Klein's inequality [30].

Theorem 16 *Let* $(v_1, ..., v_n)$ *be* n *i.i.d. random variables having common law* P *and taking values in a measurable space* $\mathcal{Z}$*. If* $\mathcal{F}$ *is a class of measurable functions from* $\mathcal{Z}$ *to* $\mathbb{R}$ *satisfying*

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

then, by setting

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

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

Bousquet's inequality*:*

$$\mathbb{P}\left[\left\|P_n - P\right\|_{\mathcal{F}} - \mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{3n}\right] \leq \exp\left(-x\right) \tag{139}$$

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

$$\mathbb{P}\left[\left\|P_n - P\right\|_{\mathcal{F}} - \mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + \frac{1}{3}\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{140}$$

Klein-Rio's inequality:

$$\mathbb{P}\left[\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] - \left\|P_n - P\right\|_{\mathcal{F}} \geq \sqrt{2\left(\sigma_{\mathcal{F}}^2 + 2b\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right]\right)\frac{x}{n}} + \frac{bx}{n}\right] \leq \exp\left(-x\right) \tag{141}$$

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

$$\mathbb{P}\left[\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] - \left\|P_n - P\right\|_{\mathcal{F}} \geq \sqrt{2\sigma_{\mathcal{F}}^2\frac{x}{n}} + \varepsilon\mathbb{E}\left[\left\|P_n - P\right\|_{\mathcal{F}}\right] + \left(\frac{1}{\varepsilon} + 1\right)\frac{bx}{n}\right] \leq \exp\left(-x\right). \tag{142}$$

Acknowledgements

The author gratefully thanks the associate editor and two anonymous referees for their comments and suggestions, that greatly improved the quality of the paper.

References

[1] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

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

[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.

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

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

[8] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[9] Sylvain Arlot and Francis 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.

[10] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[11] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[12] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[13] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[14] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[15] Jean-Patrick Baudry, Gilles Celeux, and Jean-Michel Marin. Selecting models focussing on the modeller's purpose. In *COMPSTAT 2008—Proceedings in Computational Statistics*, pages 337–348. Physica-Verlag/Springer, Heidelberg, 2008.

[16] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

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

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

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

[20] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[21] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97:113–150, 1993.

[22] Stéphane Boucheron and Pascal Massart. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011.

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

[24] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

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

[26] Olivier 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.

[27] Arnak S. Dalalyan and Alexandre 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.

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

[29] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 2005.

[30] Thierry Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C. R. Math. Acad. Sci. Paris*, 334(6):501–504, 2002.

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

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

[33] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009.

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

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

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

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

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

[39] Cathy Maugis and Bertrand Michel. Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:320–339, 2011.

[40] Adrien 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.

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

[42] Nicolas Verzelen. High-dimensional Gaussian model selection on a Gaussian design. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(2):480–524, 2010.

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