Slope heuristics for heteroscedastic regression on a random design

Sylvain Arlot, Pascal Massart

Submitted to the Annals of Statistics

SLOPE HEURISTICS FOR HETEROSCEDASTIC REGRESSION ON A RANDOM DESIGN

By Sylvain Arlot and Pascal Massart

Universite Paris-Sud

In a recent paper [BM06], Birgé and Massart have introduced the notion of minimal penalty in the context of penalized least squares for Gaussian regression. They have shown that for several model selection problems, simply multiplying by 2 the minimal penalty leads to some (nearly) optimal penalty in the sense that it approximately minimizes the resulting oracle inequality. Interestingly, the minimal penalty can be evaluated from the data themselves which leads to a data-driven choice of the penalty that one can use in practice. Unfortunately their approach heavily relies on the Gaussian nature of the stochastic framework that they consider. Our purpose in this paper is twofold: stating a heuristics to design a data-driven penalty (the *slope heuristics*) which is not sensitive to the Gaussian assumption as in [BM06] and proving that it works for penalized least squares random design regression. As a matter of fact, we could prove some precise mathematical results only for histogram bin-width selection. For some technical reasons which are explained in the paper, we could not work at the level of generality that we were expecting but still this is a first step towards further results and even if the mathematical results hold in some specific framework, the approach and the method that we use are indeed general.

1. Introduction. Model selection has received much interest in the last decades. A very common approach is penalization. In a nutshell, it chooses the model which minimizes the sum of the empirical risk (how does the algorithm fit the data) and some complexity measure of the model (called the penalty). This is the case of FPE (Akaike [Aka70]), AIC (Akaike [Aka73]) and Mallows' C_p or C_L (Mallows [Mal73]).

In this article, we consider the question of the *efficiency* of such penalization procedures, *i.e.* that their quadratic risk is asymptotically equivalent to the risk of the oracle. This property is often called asymptotic optimality. It does not mean that the procedure finds out a "true model" (which may not even exist), which would be the *consistency* problem. A procedure is efficient when it makes the best possible use of the data in terms of the quadratic risk of the

AMS 2000 subject classifications: Primary 62G05; secondary 62J05

Keywords and phrases: non-parametric regression, non-asymptotic, model selection, penalization, heteroscedastic data, histogram, concentration inequalities

final estimator.

There is a huge amount of literature about this question. About the asymptotic optimality of Mallows' C_p, Akaike's FPE and AIC, we mention here the works of Shibata [Shi81] for Gaussian errors, Li [Li87] under suitable moment assumptions on the errors, and Polyak and Tsybakov [PT90] for sharper moment conditions in the Fourier case. Then, non-asymptotic oracle inequalities (with a constant $K > 1$) have been proven by Barron, Birgé and Massart [BBM99], Birgé and Massart [BM01] in the Gaussian case, and Baraud [Bar00, Bar02] under some moment conditions on the errors. In the gaussian case, non-asymptotic oracle inequalities with a constant K_n which goes to 1 when n goes to infinity have been obtained by Birgé and Massart [BM06a].

A related problem is how much should we penalize at least? In other words, is there a minimal penalty? In the framework of Gaussian regression on a fixed-design, this question has been addressed by Birgé and Massart [BM01, BM06a], and Baraud, Giraud and Huet [BGH07] (the latter considering the unknown variance case).

Apart from the theoretical understanding of penalization methods, this question is of much interest from the practical viewpoint. In Sect. 4 of [BM06a], Birgé and Massart describe their so-called "*slope heuristics*" (see also Massart [Mas07], Sect. 8.5.2). It relies on the fact that twice the minimal penalty is almost the optimal penalty. Then, if one knows that a good penalty has the form $\mathrm{pen}(m) = KF(D_m)$ (where D_m is the dimension of the model and $K > 0$ a tuning parameter), they propose the following strategy for choosing K from the data. Define $\widehat{m}(K)$ the selected model as a function of K. First, compute $K_{\min}$ such that $D_{\widehat{m}(K)}$ is huge for $K < K_{\min}$ and reasonable when $K \geq K_{\min}$. Secondly, define $\widehat{m} := \widehat{m}(2K_{\min})$. Such a method has been successfully applied for multiple change points detection by Lebarbier [Leb05].

However, all the results about minimal penalties concern the homoscedastic fixed-design framework, where the penalty is a function of the dimension, often linear. In this paper, we prove that a similar phenomenon occurs in the *heteroscedatic random-design* case: the optimal penalty is about twice the minimal one. Our main advance is that we no longer assume the penalties to be linear in the dimension, nor even to be functions of the dimension. For proving such a result, we have to assume that each model is the vector space of piecewise constant functions on some partition of the feature space. This is quite a restriction, but we conjecture that it is mainly technical, and that the slope heuristics stays valid at least in the general least-square regression framework. We provide some evidence for this by proving several key concentration inequalities without the restriction to histograms.

Another argument supporting this conjecture is that several simulation studies have shown recently that the slope heuristics could be used in several frameworks: mixture models (Maugis and Michel [MM07]), clustering (Baudry [Bau07]), spatial statistics (Verzelen [Ver07]), estimation of oil reserves (Lepez [Lep02]) and genomics (Villers [Vil07]). Our results do not give a formal proof for these applications of the slope heuristics. However, they are a first step towards such a result, by proving that it may be applied when the ideal penalty has a general shape.

This paper is organized as follows. We describe our framework and give some notations in Sect. 2. Our main theoretical results are stated in Sect. 3. We then discuss their practical consequences in Sect. 4. Appendix A is devoted to computational issues. All the proofs are given in Appendix B.

2. Framework.

2.1. *Regression.* We observe some data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, i.i.d. with common law P. Denoting by s the regression function, we have

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \tag{1}$$

where $\sigma : \mathcal{X} \mapsto \mathbb{R}$ is the heteroscedastic noise-level and ϵ_i are i.i.d. centered noise terms, possibly dependent from X_i, but with mean 0 and variance 1 conditionally to X_i. Typically, the feature space $\mathcal{X}$ is a compact set of $\mathbb{R}^d$. Throughout this paper, we make two main assumptions (which can also be relaxed, see Sect. 3):

- The data is bounded: $\|Y\|_\infty \leq A < \infty$.
- Uniform lower-bound on the noise-level: $\sigma(X) \geq \sigma_{\min} > 0$ a.s.

Given a predictor $t : \mathcal{X} \mapsto \mathcal{Y}$, its quality is measured by the (quadratic) prediction loss

$$\mathbb{E}_{(X,Y)\sim P}\left[\gamma(t,(X,Y))\right] =: P\gamma(t) \qquad \text{where} \quad \gamma(t,(x,y)) = \left(t(x)-y\right)^2$$

is the least-square contrast. Then, the Bayes predictor[1] is the regression function s, and we define the excess loss as

$$l(s,t) := P\gamma\left(t\right) - P\gamma\left(s\right) = \mathbb{E}_{(X,Y)\sim P}\left(t(X)-s(X)\right)^2 \quad .$$

Given a particular set of predictors S_m (called a *model*), we define the best predictor over S_m

$$s_m := \arg\min_{t\in S_m} \left\{ P\gamma(t) \right\} \quad ,$$

and its empirical counterpart

$$\widehat{s}_m := \arg\min_{t\in S_m} \left\{ P_n\gamma(t) \right\}$$

(when it exists and is unique), where $P_n = n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$. This estimator is the well-known *empirical risk minimizer*, also called least-square estimator since γ is the least-square contrast.

[1] *i.e.* the minimizer of $P\gamma(t)$ over the set of all predictors.

2.2. *Ideal model selection.* We now assume that we have a family of models $(S_m)_{m\in\mathcal{M}_n}$, hence a family of estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ (*via* empirical risk minimization). We are looking for some data-dependent $\widehat{m} \in \mathcal{M}_n$ such that $l(s,\widehat{s}_{\widehat{m}})$ is as small as possible. This is the model selection problem. For instance, we would like to prove some oracle inequality of the form

$$l(s,\widehat{s}_{\widehat{m}}) \leq K \inf_{m\in\mathcal{M}_n} \{\, l(s,\widehat{s}_m)\,\} + R_n$$

in expectation or on a set of large probability, with K close to 1 and $R_n = o(n^{-1})$.

General penalization procedures can be described as follows. Let $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$ be some penalty function, possibly data-dependent. Then, define

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \{\,\mathrm{crit}(m)\,\} \qquad \text{with} \qquad \mathrm{crit}(m) := P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)\ . \tag{2}$$

Since the ideal criterion crit is the true prediction error $P\gamma\,(\,\widehat{s}_m\,)$, the *ideal penalty* is

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

Of course, this quantity is unknown because it depends on the true distribution P. A natural idea is to choose pen as close as possible to $\mathrm{pen}_{\mathrm{id}}$ for every model $m \in \mathcal{M}_n$. We show below, in a very general setting, that when pen estimates well the ideal penalty $\mathrm{pen}_{\mathrm{id}}$, $\widehat{m}$ satisfies an oracle inequality with constant K close to 1.

By definition of $\widehat{m}$,

$$\forall m \in \mathcal{M}_n, \quad P_n\gamma(\widehat{s}_{\widehat{m}}) \leq P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m})\ .$$

For every $m \in \mathcal{M}_n$, we define

$$\begin{aligned} p_1(m) &= P\,(\gamma(\widehat{s}_m) - \gamma(s_m)\,) & \qquad p_2(m) &= P_n\,(\gamma(s_m) - \gamma(\widehat{s}_m)\,) \\ \delta(m) &= (P_n - P)\gamma(s_m) & \qquad \overline{\delta}(m) &= \delta(m) - (P_n - P)\gamma(s) \end{aligned}$$

so that

$$l(s,\widehat{s}_m) = P_n\gamma(\widehat{s}_m) + p_1(m) + p_2(m) - \overline{\delta}(m) - P_n\gamma(s)\ .$$

We then have, for every $m \in \mathcal{M}_n$,

$$\begin{aligned} & l(s,\widehat{s}_{\widehat{m}}) + (\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}})(\widehat{m}) \leq l(s,\widehat{s}_m) + (\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}})(m) \\ \text{where} \qquad & \mathrm{pen}'_{\mathrm{id}}(m) := p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}_{\mathrm{id}}(m) + (P_n - P)\gamma\,(\,s\,)\ . \end{aligned} \tag{3}$$

In order to derive an oracle inequality from (3), we have to show that for every $m \in \mathcal{M}_n$, $\mathrm{pen}(m)$ is close to $\mathrm{pen}'_{\mathrm{id}}(m)$ (or, equivalently, to $\mathrm{pen}_{\mathrm{id}}(m)$, since $\mathrm{pen}'_{\mathrm{id}} - \mathrm{pen}_{\mathrm{id}}$ does not depend on m). Actually, both $\mathrm{pen}_{\mathrm{id}}$ and $\mathrm{pen}'_{\mathrm{id}}$ are ideal penalties (they lead to select the same optimal model).

The reason why we prefer $\mathrm{pen}'_{\mathrm{id}}$ in (3) is that it has lower deviations around its expectation than $\mathrm{pen}_{\mathrm{id}}$.

When the penalty pen is too large, the left-hand side of (3) stays larger than $l(s, \widehat{s}_{\widehat{m}})$ so that we can still obtain an oracle inequality (possibly with a large constant K). On the contrary, when pen is too small, the left-hand side of (3) can become negligible in front of $l(s, \widehat{s}_{\widehat{m}})$ (which makes K explode) or — worse — can be nonpositive (so that we can no longer derive an oracle inequality from (3)). We shall see in the following that this corresponds to the existence of a "minimal penalty".

Consider first the case $\mathrm{pen}(m) = p_2(m)$ in (2). Then, $\mathbb{E}[\mathrm{crit}(m)] = \mathbb{E}[P_n \gamma(s_m)] = P\gamma(s_m)$, so that $\widehat{m}$ tends to be the model with the smallest bias, hence the more complex one. As a consequence, the risk of $\widehat{s}_{\widehat{m}}$ is very large. When $\mathrm{pen}(m) = Cp_2(m)$ with $C < 1$, $\mathrm{crit}(m)$ is a decreasing function of the complexity of m, so that $\widehat{m}$ is still one of the more complex models. On the contrary, when $\mathrm{pen}(m) > p_2(m)$, $\mathrm{crit}(m)$ starts to increase with the complexity of m (at least for the largest models), so that $\widehat{m}$ has a smallest complexity. This intuition supports the conjecture that the "minimal amount of penalty" required for the model selection procedure to work may be $p_2(m)$.

In this article, we prove (in a particular framework, see Sect. 2.3) that

$$\forall m \in \mathcal{M}_n, \qquad p_1(m) \approx p_2(m) \ ,$$

so that the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m) \approx p_1(m) + p_2(m)$ is close to $2p_2(m)$. On the other hand, $p_2(m)$ is actually a "minimal penalty". So, we deduce that the optimal penalty is close to twice the minimal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m) \approx 2\,\mathrm{pen}_{\mathrm{min}}(m) \ .$$

This is the so-called "slope heuristics", which was first introduced by Birgé and Massart [BM06a] in a Gaussian setting. It is splitted into two main results. First, an oracle inequality with constant almost one when $\mathrm{pen} = 2\mathbb{E}[p_2]$ (Thm. 1), relying on (3) and the comparison $p_1 \approx p_2$. Second, lower bounds on $D_{\widehat{m}}$ and the risk of $\widehat{s}_{\widehat{m}}$ when pen is smaller than p_2 (Thm. 2).

These theorems rely on two kinds of results. First, both p_1, p_2 and $\overline{\delta}$ concentrate around their expectations (which can be done in a quite general framework, at least for p_2 and $\overline{\delta}$, see App. B.5). Second, $\mathbb{E}[p_1(m)] \approx \mathbb{E}[p_2(m)]$ for every $m \in \mathcal{M}_n$. This last point is quite hard in general, so that we must make a structural assumption on the models. In this article, we consider the histogram case, where S_m is the set of piecewise constant functions on some fixed partition $(I_\lambda)_{\lambda \in \Lambda_m}$. We describe this framework in the next subsection.

2.3. *Histograms.* A "model of histograms" S_m is the the set of piecewise constant functions (histograms) on some partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$. It is thus a vector space of dimension $D_m = \mathrm{Card}(\Lambda_m)$, spanned by the family $(\mathbb{1}_{I_\lambda})_{\lambda \in \Lambda_m}$. As this basis is orthogonal in $L^2(\mu)$ for

any probability measure on $\mathcal{X}$, computations are quite easy. This is the only reason why we assume that each S_m is a model of histograms in Sect. 3. The following notations will be useful throughout this paper.

$$p_\lambda := P(X \in I_\lambda) \qquad \widehat{p}_\lambda := P_n(X \in I_\lambda)$$
$$(\sigma_\lambda^r)^2 := \mathbb{E}\left[\sigma(X)^2 \,\middle|\, X \in I_\lambda\right] \qquad \left(\sigma_\lambda^d\right)^2 := \mathbb{E}\left[(s(X) - s_m(X))^2 \,\middle|\, X \in I_\lambda\right]$$

$$s_m := \arg\min_{t \in S_m} P\gamma(t) = \sum_{\lambda \in \Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \qquad \text{with} \quad \beta_\lambda = \mathbb{E}_P\left[Y \mid X \in I_\lambda\right]$$
$$\widehat{s}_m := \arg\min_{t \in S_m} P_n\gamma(t) = \sum_{\lambda \in \Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \text{with} \quad \widehat{\beta}_\lambda = \frac{1}{n\widehat{p}_\lambda} \sum_{X_i \in I_\lambda} Y_i$$

Remark that $\widehat{s}_m$ is uniquely defined if and only if each I_λ contains at least one of the X_i. Otherwise, we will consider that the model m can not be chosen. In order to make $\mathbb{E}\left[p_1(m)\right]$ well-defined and finite, we choose a convention for $p_1(m)$ when $\min_{\lambda \in \Lambda_m} \widehat{p}_\lambda = 0$ (see (34) in App. B).

In order to understand the main difference between our framework and the homoscedastic fixed-design one, let us compare the expectations of the ideal penalty.

In the homoscedastic fixed-design framework[2], it is quite straightforward to show that

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 D_m}{n} \ . \tag{4}$$

On the other hand, in our framework, we can prove the following (*cf.* [Arl07], Sect. 5.7.2). Denote by $\mathbb{E}^{\Lambda_m}[\cdot]$ the expectation conditionally to $(\mathbb{1}_{X_i \in I_\lambda})_{1 \leq i \leq n,\, \lambda \in \Lambda_m}$. If for every $\lambda \in \Lambda_m$, $\widehat{p}_\lambda > 0$, then

$$\mathbb{E}^{\Lambda_m}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n} \sum_{\lambda \in \Lambda_m} \left(\frac{p_\lambda}{\widehat{p}_\lambda} + 1\right) \left(\left(\sigma_\lambda^d\right)^2 + (\sigma_\lambda^r)^2\right) \ . \tag{5}$$

Apart from the difference between $p_\lambda / \widehat{p}_\lambda$ and 1 (which does not matter with large probability, see App. B.5), there are two main differences between (4) and (5). Firstly, the bias term $\left(\sigma_\lambda^d\right)^2$, which is due to the randomness of the design. If s is highly non-smooth, this term can be significant. Secondly, the variance term $(\sigma_\lambda^r)^2$ depends on $\lambda \in \Lambda_m$, whereas it is constant equal to σ^2 in the homoscedastic case. When $(p_\lambda)_{\lambda \in \Lambda_m}$ are far from the uniform weights, $n^{-1} \sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2$ is far from $D_m n^{-1} \mathbb{E}\left[\sigma(X)^2\right]$. As shown in Chap. 4 of [Arl07], in such cases, it may happen that any linear penalization procedure is suboptimal. Then, more general penalties than the ones considered in [BM06a] are required.

[2] in which the true distribution P gives weights n^{-1} to each of the (deterministic) design points $X_1, \ldots, X_n$. The unknown distribution is only the one of $(\epsilon_i)_{1 \leq i \leq n}$.

3. Theoretical results. In this section, we restrict ourselves to the histogram regression case. Remember that we do not consider histograms as a final goal. We only make this assumption in order to make explicit computations and obtain results from which we can derive heuristics for practical applications.

Let $(S_m)_{m\in\mathcal{M}}$ be a family of histogram models satisfying

(**P1**) Polynomial complexity of $\mathcal{M}_n$: $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.
(**P2**) Richness of $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ s.t. $D_{m_0} \in [\sqrt{n}, c_{\mathrm{rich}}\sqrt{n}]$.

Assumption (**P1**) is quite classical when one aims at proving the asymptotic optimality of a model selection procedure (it is for instance implicitly assumed by Li [Li87], in the homoscedastic fixed-design case).

For any penalty function $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}^+$, we define the following model selection procedure:

$$\widehat{m} \in \arg \min_{m\in\mathcal{M}_n,\, \min_{\lambda\in\Lambda_m}\{\widehat{p}_\lambda\}>0} \{ P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \} \quad . \tag{6}$$

3.1. *Optimal penalties.* Our first result is an oracle inequality. The following theorem shows that the penalization procedure (6) is efficient (*i.e.* satisfies a non-asymptotic oracle inequality, with constant C converging to 1 when n goes to infinity), provided that the penalty is well chosen.

THEOREM 1. *Assume that the data $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d. and satisfy the following:*

(**Ab**) *Bounded data: $\|Y_i\|_\infty \leq A < \infty$.*
(**An**) *Noise-level bounded from below: $\sigma(X_i) \geq \sigma_{\min} > 0$ a.s.*
(**Ap**) *Polynomial decreasing of the bias: there exists $\beta_1 \geq \beta_2 > 0$ and $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ such that*

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq l(s, s_m) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for $\mathcal{L}(X)$: $D_m \min_{\lambda\in\Lambda_m} \{ \mathbb{P}(X \in I_\lambda) \} \geq c_{\mathrm{r},\ell}^X$.*

For every $m \in \mathcal{M}_n$, consider the penalty

$$\mathrm{pen}(m) = 2\mathbb{E}\left[P_n \left(\gamma(s_m) - \gamma(\widehat{s}_m) \right) \right] \quad . \tag{7}$$

Then, if $\widehat{m}$ is defined by (6), there exists a constant K_1 and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_1 n^{-2}$,

$$l(s, \widehat{s}_{\widehat{m}}) \leq [1 + \epsilon_n] \inf_{m\in\mathcal{M}_n} \{ l(s, \widehat{s}_m) \} \quad . \tag{8}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right] \leq [1 + \epsilon_n] \, \mathbb{E}\left[\inf_{m\in\mathcal{M}_n} \{ l(s, \widehat{s}_m) \} \right] + \frac{A^2 K_1}{n^2} \quad . \tag{9}$$

The constant K_1 may depend on the constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) *and* ($\mathbf{Ar}_\ell^{\mathbf{X}}$), *but not on n. The small term ϵ_n depends only on n (it can for instance be upperbounded by $\ln(n)^{-1/5}$).*

The rationale behind this theorem is that the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ is close to its expectation, which is itself close to $2\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right]$. Then, (3) directly implies an oracle inequality like (8), hence (9).

Actually, Thm. 1 above is a corollary of a more general result, Thm. 3, that we state in App. B.2. In particular, if

$$\mathrm{pen}(m) = C\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \tag{10}$$

instead of (7), then the constant $1+\epsilon_n$ in (8) becomes $(C-1)^{-1}+\epsilon_n$ if $C \in (1,2]$ and $C-1+\epsilon_n$ if $C \geq 2$. This means that for every $C > 1$, the penalty defined by (10) is efficient, up to a multiplicative constant. This is well known in the homoscedastic case [BM01, Bar00, Bar02], but new in the heteroscedastic one.

We now make a few comments about the assumptions of Thm. 1:

- (**Ab**) and (**An**) are rather mild. In particular, they allow quite general heteroscedastic noises. They can also be relaxed, for instance thanks to results proven by Arlot [Arl07] (Chap. 6 and Sect. 8.3) , which allow the noise to vanish or to be unbounded.
- ($\mathbf{Ar}_\ell^{\mathbf{X}}$) is satisfied for "almost regular" histograms when X has a lower bounded density w.r.t. Leb.
- The upper bound in (**Ap**) holds when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and s α-hölderian with $\alpha \in (0,1]$. The lower bound is more surprising. Indeed, it is classical to assume that $l(s,s_m) > 0$ for every $m \in \mathcal{M}_n$ for proving the asymptotic optimality of Mallows' C_p (*e.g.* by Shibata [Shi81], Li [Li87] and Birgé and Massart [BM06a]). We here make a stronger assumption because we need a non-asymptotic lower bound on the dimension of both the oracle and selected models.
  The reason why this assumption is not too restrictive is that non-constant α-hölderian functions satisfy (**Ap**) when $(I_\lambda)_{\lambda\in\Lambda_m}$ is regular and X has a lower-bounded density w.r.t. the Lebesgue measure on $\mathcal{X} \subset \mathbb{R}^k$ (*cf.* Sect. 8.10 in [Arl07] for more details). Notice that Stone [Sto85] and Burman [Bur02] used the same assumption in the density estimation framework.

3.2. *Minimal penalties.* In the previous subsection, we have shown that the penalization procedure built upon $C\mathbb{E}\left[p_2(m)\right]$ with any $C > 1$ satisfies an oracle inequality with a constant $K(C)$. According to our analysis, $K(2)$ is close to 1, and $K(C)$ explodes when C goes either to 1 or to infinity. However, this is not sufficient to state that $C = 1$ corresponds to the "minimal amount of penalization" needed, since we only have upper bounds on the risk. Theorem 2 below shows that $C < 1$ actually induces an explosion of the risk, so that the condition $C \geq 1$ is necessary (we do not study the critical situation $C = 1$).

THEOREM 2. *Assume that the data $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d. and satisfy the following:*

(**Ab**) *Bounded data:* $\|Y_i\|_\infty \leq A < \infty$.
(**An**) *Noise-level bounded from below:* $\sigma(X_i) \geq \sigma_{\min} > 0$ *a.s.*
($\mathbf{Ap_u}$) *Polynomial upper bound on the bias: there exists* $\beta_2 > 0$ *and* $C_{\mathrm{b}}^+ > 0$ *such that*

$$l(s, s_m) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for* $\mathcal{L}(X)$*:* $D_m \min_{\lambda\in\Lambda_m} \{\mathbb{P}(X \in I_\lambda)\} \geq c_{\mathrm{r},\ell}^X$.

Let $C \in [0; 1)$ *and assume that for every* $m \in \mathcal{M}_n$,

$$0 \leq \mathrm{pen}(m) \leq C\mathbb{E}\left[P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \tag{11}$$

with probability at least $1 - Ln^{-2}$.

Then, if $\widehat{m}$ *is defined by* (6)*, there exists two constants* K_2, K_3 *such that, with probability at least* $1 - K_2 n^{-2}$,

$$D_{\widehat{m}} \geq K_3 n \ln(n)^{-1} \quad . \tag{12}$$

On the same event,

$$l(s, \widehat{s}_{\widehat{m}}) \geq \ln(n) \inf_{m\in\mathcal{M}_n} \{l(s, \widehat{s}_m)\} \quad . \tag{13}$$

The constants K_2 *and* K_3 *may depend on* C *and constants in* (**P1**)*,* (**P2**)*,* (**Ab**)*,* (**An**)*,* (**Ap**) *and* ($\mathbf{Ar}_\ell^{\mathbf{X}}$)*, but not on* n.

Together with Thm. 1 (and the remarks below), this proves that $\mathbb{E}[p_2(m)]$ is a "minimal penalty": when $\mathrm{pen}(m) = C\mathbb{E}[p_2(m)]$, $\widehat{m}$ satisfies an oracle inequality if $C > 1$, but not if $C < 1$. This confirms the intuitive reasoning exposed at the end of Sect. 2.2.

As in the results of Birgé and Massart [BM06a], Thm. 2 points out two simultaneous phenomena when the penalty is too small. First, the dimension of the selected model explodes (12). Second, the efficiency of the model selection strongly decreases (13). This coupling is quite interesting. Indeed, we want to avoid underpenalization because of the second phenomenon, while the blow up of the dimension allows us to detect it more easily. This is crucial from the practical viewpoint, as we shall see in Sect. 4.

The novelty in Thm. 2 is that it does not make restrictive assumptions on the distribution of the noise, which may be nongaussian and heteroscedastic. Then, the minimal penalty may not be a function of the dimension D_m. Even more interesting consequences of Thm. 2 come from an accurate comparison with Thm. 1. This is the purpose of the next section.

3.3. *Comments.*

3.3.1. *Optimal penalty* vs. *minimal penalty.* First, Thm. 1 and 2 show that there is a link between the optimal penalty $\mathrm{pen}_{\mathrm{opt}}(m) = 2\mathbb{E}\left[p_2(m)\right]$ and the minimal penalty $\mathrm{pen}_{\mathrm{min}}(m) = \mathbb{E}\left[p_2(m)\right]$. Apart from the particular expression of $\mathbb{E}\left[p_2(m)\right]$, we can retain the following rule of thumb:

$$\textit{“optimal” penalty} \approx 2 \times \textit{“minimal” penalty} \ .$$

This has already been proposed by Birgé and Massart [BM06a], but their results were restricted to the Gaussian homoscedastic framework. In this paper, we extend them to a non-Gaussian and heteroscedastic setting.

From the practical viewpoint, this means that we can design an optimal penalty as soon as we can find the minimal one from the data. Of course, the ratio between the excess loss of $\widehat{s}_{\widehat{m}}$ and the one of the oracle is unknown, so that it is not straightforward to detect that a penalty is “minimal”. Interestingly, it appears from Thm. 1 and 2 and their proofs that the dimension of the selected model $D_{\widehat{m}}$ jumps exactly when the penalty is minimal. We detail this phenomenon in the next paragraph.

3.3.2. *Dimension jump.* In the statement of Thm. 2, we mention that $D_{\widehat{m}}$ is very large (proportionnal to $n/\ln(n)$) with a large probability when $\mathrm{pen}(m) \leq C\mathbb{E}\left[\mathrm{pen}_{\mathrm{min}}(m)\right]$ for some $C < 1$. This is not only the key of our proof that the risk of $\widehat{s}_{\widehat{m}}$ explodes when pen is too small.

Under the same assumptions, when $\mathrm{pen}(m) \approx C\mathbb{E}\left[\mathrm{pen}_{\mathrm{min}}(m)\right]$ for some $C > 1$, the proof of Thm. 1 shows that

$$\forall 1/2 > \alpha > \left(1 - \beta_2\right)_+ /2, \qquad \mathbb{P}\left(D_{\widehat{m}} \leq n^{1/2+\alpha}\right) \geq 1 - K_1'(\alpha) n^{-2} \ .$$

Denoting by $\widehat{m}(C)$ the selected model when $\mathrm{pen}(m) \approx C\,\mathrm{pen}_{\mathrm{min}}(m)$, we have proven that there is an event of large probability on which

$$\forall C \in [0,1), \qquad D_{\widehat{m}(C)} \geq \frac{K_3 n}{\ln(n)} \qquad \text{and} \qquad \forall C \in (1,2], \qquad D_{\widehat{m}(C)} \leq n^{1-\delta}$$

for some $\delta > 0$. In a nutshell, there is a large *dimension jump* around the minimal penalty.

Contrary to the explosion of the excess loss, the dimension jump can be observed from the data only. It is clearly observed in simulation studies (see Fig. 1 in App. A). This means that if one knows (at least approximately) the optimal penalty up to some multiplicative constant, the dimension jump allows to determine accurately the minimal penalty. Then, multiplying by two this estimate, one obtains an optimal penalty. In the next section, we discuss the resulting algorithm.

4. Practical use of the slope heuristics.

4.1. *Data-driven penalties.* We are now in position to define a data-driven calibration algorithm for penalization procedures. A similar method has already been proposed by Birgé and Massart [BM06a] (see also [BM06b]) and implemented by Lebarbier [Leb05].

ALGORITHM 1 (Data-driven penalization with slope heuristics).

1. Choose a shape of penalty $\mathrm{pen}_{\mathrm{shape}} : \mathcal{M}_n \mapsto \mathbb{R}^+$.
2. Compute the selected model $\widehat{m}(K)$ as a function of $K > 0$
$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma(\widehat{s}_m) + K \, \mathrm{pen}_{\mathrm{shape}}(m) \right\} \ .$$
3. Find $\widehat{K}_{\min} > 0$ such that $D_{\widehat{m}(K)}$ is too large for $K < \widehat{K}_{\min}$ and "reasonably small" for $K > \widehat{K}_{\min}$.
4. Select the model $\widehat{m} = \widehat{m}(2\widehat{K}_{\min})$.

Computational aspects are discussed in App. A. Let us now focus on another practical question, which is step 1 of Algorithm 1. In the homoscedastic framework, it is quite straightforward, the good shape of penalty being given by explicit formulas [BM06a]. If $\mathcal{M}_n$ has a polynomial complexity (*i.e.* satisfies (**P1**)), then $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$ is a good choice, since Mallows' C_p is asymptotically optimal. When the noise is highly heteroscedastic, this is no longer the case so that step 1 may become much harder. We study this question in the next subsection.

4.2. *Shape of the penalty.* In the heteroscedastic framework, the shape of the ideal penalty is unknown, because it does not depend only of the dimension of the models. In addition, it has been proved ([Arl07], Chap. 4) that any penalty of the form $\widehat{C} D_m$ is suboptimal for model selection in some heteroscedastic situation, even if $\widehat{C}$ is allowed to depend on the true distribution P. Then, optimal model selection with heteroscedastic data strongly requires to estimate the shape of $\mathrm{pen}_{\mathrm{id}}(m)$.

A natural idea for solving this problem is the use of resampling. As shown in [Arl07], resampling penalties (Chap. 6) or V-fold penalties (Chap. 5) provide good estimates of the shape of $\mathrm{pen}_{\mathrm{id}}$ in the heteroscedastic framework. Whereas these results are also restricted to the histogram case (for which the use of Algorithm 1 is unnecessary, because the optimal calibration constants are known), several theoretical results supports the conjecture that they are valid in a much more general situation.

Notice also that resampling does not give the exact shape of $\mathbb{E}\left[p_2(m) \right]$ (as required in Thm. 1), but only an approximation on an event of large probability. This why we state a more general result, Thm. 3, which allows pen to be only near the right penalty shape.

Combining Algorithm 1 with some resampling penalization algorithm (see [Arl07], Chap. 5 and 6), we now have a completely data-driven way for designing optimal penalties. The theoretical justification of this procedure allows the noise to be heteroscedastic and non-gaussian, which is a quite interesting point. Apart from the histogram case, we believe that it remains valid, but theoretical justification remains an open problem.

4.3. *Large number of models.* Contrary to Birgé and Massart [BM06a], we have restricted our study to the situation where the collection of models $\mathcal{M}_n$ is "small", *i.e.* has a size growing at most like a power of n. For several problems, such that complete variable selection, this assumption does not hold, and it is known from the homoscedastic case that the minimal penalty is much larger than $\mathbb{E}\left[p_2(m)\right]$.

Following (42) and the surrounding comments in [BM06a], we suggest to answer this question as follows. First, group the models according to some complexity index C_m (for instance their dimensions, or the approximate value of their resampling penalty): for $C \in \left\{1, \ldots, n^k\right\}$, define $\widetilde{S_C} = \bigcup_{C_m = C} S_m$. Then, replace the model selection problem with the family $(S_m)_{m \in \mathcal{M}_n}$ by a "complexity selection problem", *i.e.* model selection with the family $\left(\widetilde{S_C}\right)_{1 \leq C \leq n^k}$.

We believe that this grouping of the models is sufficient to take into account the richness of $\mathcal{M}_n$ for the optimal calibration of the penalty. A theoretical justification of this point may rely on the extension of our results to any kind of model, not only histogram ones (each $\widetilde{S_C}$ is not an "histogram model", since it is even not a vector space). As mentioned in the previous subsection, this remains an interesting open problem.

5. Conclusion. We have seen in this paper that it is possible to provide mathematical evidences that the method introduced in [BM06a] to design data-driven penalties remains efficient in a non Gaussian context. Our purpose in this conclusive section is to relate the heuristics that we have developped in Sect. 2 to the well known Mallows'C_p and Akaike's criteria and to the unbiased risk (or almost unbiased) estimation of the risk principle. To explain our idea which consists in guessing what is the right penalty to be used from the data themselves, let us come come back to Gaussian model selection. Towards this aim let us consider some empirical criterion γ_n (which can be the least squares criterion as in this paper but which could be the log-likelihood criterion as well). Let us also consider some collection of models $(S_m)_{m \in \mathcal{M}}$ and in each model S_m some minimizer s_m of $t \mapsto \mathbb{E}\left[\gamma_n(t)\right]$ over S_m (assuming that such a point does exist). Defining for every $m \in \mathcal{M}$,

$$\widehat{b}_m = \gamma_n(s_m) - \gamma_n(s) \text{ and } \widehat{v}_m = \gamma_n(s_m) - \gamma_n(\widehat{s}_m) \ ,$$

minimizing some penalized criterion

$$\gamma_n(\widehat{s}_m) + \mathrm{pen}(m)$$

over $\mathcal{M}$ amounts to minimize

$$\widehat{b}_m - \widehat{v}_m + \mathrm{pen}(m).$$

The point is that since $\widehat{b}_m$ is an unbiased estimator of the bias term $l(s, s_m)$. If we have in mind to use concentration arguments, one can hope that minimizing the quantity above will be approximately equivalent to minimize

$$l(s, s_m) - \mathbb{E}\left[\widehat{v}_m\right] + \mathrm{pen}(m) \ .$$

Since the purpose of the game is to minimize the risk $\mathbb{E}[l(s,\widehat{s}_m)]$, an ideal penalty would therefore be

$$\mathrm{pen}(m) = \mathbb{E}[\widehat{v}_m] + \mathbb{E}[\ell(s_m,\widehat{s}_m)] \quad .$$

In the Mallows' C_p case (for Gaussian fixed design regression least squares), the models S_m are linear and $\mathbb{E}[\widehat{v}_m] = \mathbb{E}[\ell(s_m,\widehat{s}_m)]$ are explicitly computable (at least if the level of noise is assumed to be known). For Akaike's penalized log-likelihood criterion, this is similar, at least asymptotically. More precisely, one uses the fact that

$$\mathbb{E}[\widehat{v}_m] \approx \mathbb{E}[\ell(s_m,\widehat{s}_m)] \approx \frac{D_m}{2n} \quad ,$$

where D_m stands for the number of parameters defining model S_m. The conclusion of these considerations is that Mallows' C_p as well as Akaike's criterion are indeed both based on the unbiased risk (or asymptotically unbiased) estimation principle.

The first idea that we are using in this paper is that one can go further in this direction and that the approximation $\mathbb{E}[\widehat{v}_m] \approx \mathbb{E}[\ell(s_m,\widehat{s}_m)]$ remains valid even in a non-asymptotic context. If one believes in it then a good penalty becomes $2\mathbb{E}[\widehat{v}_m]$ or equivalently (having still in mind concentration arguments) $2\widehat{v}_m$. This in some sense explains the rule of thumb which is given in [BM06a] and further studied in this paper and connect it to Mallows' C_p and Akaike's heuristics. Indeed, the minimal penalty is $\widehat{v}_m$ while the optimal penalty should be $\widehat{v}_m + \mathbb{E}[\ell(s_m,\widehat{s}_m)]$ and their ratio is approximately equal to 2. The second idea that we are using in this paper is that one can guess the minimal penalty from the data. There are indeed several ways to perform the estimation of the minimal penalty. Here we have studied a slope heuristics which amounts to consider that the shape of the minimal penalty is (at least approximately) of the form αD_m and estimate the unknown value α by the *slope* of the graph of $\gamma_n(\widehat{s}_m)$ for large enough values of D_m. It is easy to extend this method to other shapes of penalties, simply by replacing D_m by some (known!) function $f(D_m)$. For instance, Émilie Lebarbier has used $f(D_m) = D_m\left(2.5 + \ln\left(\frac{n}{D_m}\right)\right)$ for multiple change points detection from n noisy data. It is even possible to combine resampling ideas with the slope heuristics by taking a random function f which is built from a randomized empirical criterion. As shown in Arlot [Arl07] this approach turns out to be much more efficient than the rougher choice $f(D_m) = D_m$ for highly heteroscedastic random regression frameworks. Of course, the question of the optimality of the slope heuristics remains widely open but we believe that on the one hand this heuristics can be helpfull in practice and that on the other hand, proving its efficiency even on a toy model as we did in this paper is already something.

APPENDIX A: COMPUTATIONAL ASPECTS OF THE SLOPE HEURISTICS

With Algorithm 2 (possibly combined with resampling penalties for step 1), we have a completely data-driven and optimal model selection procedure. From the practical viewpoint, the last two problems may be steps 2 and 3. First, at step 2, how can we compute exactly $\widehat{m}(K)$

for every $K \in (0, +\infty)$, this latter set being uncountable? The answer is that the whole trajectory $(\widehat{m}(K))_{K \geq 0}$ can be described with a small number of parameters, which can be computed fastly. This point is the object of Sect. A.1. Second, at step 3, how can the jump of dimension be detected automatically in practice. In other words, how should $\widehat{K}_{\min}$ be defined exactly, as a function of $(\widehat{m}(K))_{K \geq 0}$? We try to answer this question in Sect. A.2.

A.1. Computation of $(\widehat{m}(K))_{K \geq 0}$. For every model $m \in \mathcal{M}_n$, define

$$f(m) = P_n \gamma\left(\widehat{s}_m\right) \qquad g(m) = \mathrm{pen}_{\mathrm{shape}}(m)$$

$$\text{and} \qquad \forall K \geq 0, \qquad \widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ f(m) + K g(m) \right\} \ .$$

Since the latter definition can be ambiguous, we choose any total ordering $\preceq$ on $\mathcal{M}_n$ such that g is non-decreasing. Then, $\widehat{m}(K)$ is defined as the smallest element of

$$E(K) := \arg\min_{m \in \mathcal{M}_n} \left\{ f(m) + K g(m) \right\}$$

for $\preceq$. The main reason why the whole trajectory $(\widehat{m}(K))_{K \geq 0}$ can be computed efficiently is its very particular shape.

Indeed, the results below (mostly Lemma 2) show that $K \mapsto \widehat{m}(K)$ is piecewise constant, and non-increasing for $\preceq$. We then have

$$\forall i \in \{0, \ldots, i_{\max}\}, \quad \forall K \in [K_i, K_{i+1}), \qquad \widehat{m}(K) = m_i \ ,$$

and the whole trajectory $(\widehat{m}(K))_{K \geq 0}$ can be represented by:

- a non-negative integer $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n) - 1$ (the number of jumps),
- an increasing sequence of positive reals $(K_i)_{0 \leq i \leq i_{\max}+1}$ (the location of the jumps, with $K_0 = 0$ and $K_{i_{\max}+1} = +\infty$)
- a non-increasing sequence of models $(m_i)_{0 \leq i \leq i_{\max}}$.

We are now in position to give an efficient algorithm for step 2 in Algorithm 2. The point is that the K_i and the m_i can be computed sequentially, each step having a complexity proportional to $\mathrm{Card}(\mathcal{M}_n)$. This means that its overall complexity is lower than a constant times $\mathrm{Card}(\mathcal{M}_n)^2$. Notice also that Algorithm 2 can be stopped earlier if the only goal is to identify $\widehat{K}_{\min}$ (which may be done only with the first m_i).

Algorithm 2 (Step 2 of Algorithm 1). For every $m \in \mathcal{M}_n$, define $f(m) = P_n \gamma\left(\widehat{s}_m\right)$ and $g(m) = \mathrm{pen}_{\mathrm{shape}}(m)$. Choose $\preceq$ any total ordering on $\mathcal{M}_n$ such that g is non-decreasing.

- Init: $K_0 = 0$, $m_0 = \arg\min_{m \in \mathcal{M}_n} \{ f(m) \}$ (when this minimum is attained several times, m_0 is defined as the smallest one for $\preceq$).

- Step i, $i \geq 1$: Let

$$G(m_{i-1}) := \{ m \in \mathcal{M}_n \text{ s.t. } f(m) > f(m_{i-1}) \quad \text{and} \quad g(m) < g(m_{i-1}) \} \ .$$

If $G(m_{i-1}) = \emptyset$, then put $K_i = +\infty$, $i_{\max} = i - 1$ and stop.
Otherwise, define

$$K_i := \inf \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \text{ s.t. } m \in G(m_{i-1}) \right\} \tag{14}$$

and m_i the smallest element (for $\preceq$) of

$$F_i := \arg \min_{m \in G(m_{i-1})} \left\{ \frac{f(m) - f(m_{i-1})}{g(m_{i-1}) - g(m)} \right\} \ .$$

The validity of Algorithm 2 is justified by the following proposition, showing that these K_i and m_i are the same as the ones describing $(\widehat{m}(K))_{K \geq 0}$.

PROPOSITION 1. *If $\mathcal{M}_n$ is finite, algorithm 2 terminates and $i_{\max} \leq \mathrm{Card}(\mathcal{M}_n) - 1$. Using the notations of Algorithm 2, and defining $\widehat{m}(K)$ as the smallest element (for $\preceq$) of*

$$E(K) := \arg \min_{m \in \mathcal{M}_n} \{ f(m) + K g(m) \} \ ,$$

$(K_i)_{0 \leq i \leq i_{\max}+1}$ is increasing and $\forall i \in \{0, \ldots, i_{\max} - 1\}$, $\forall K \in [K_i, K_{i+1})$, $\widehat{m}(K) = m_i$.

It is proven in Sect. A.3.

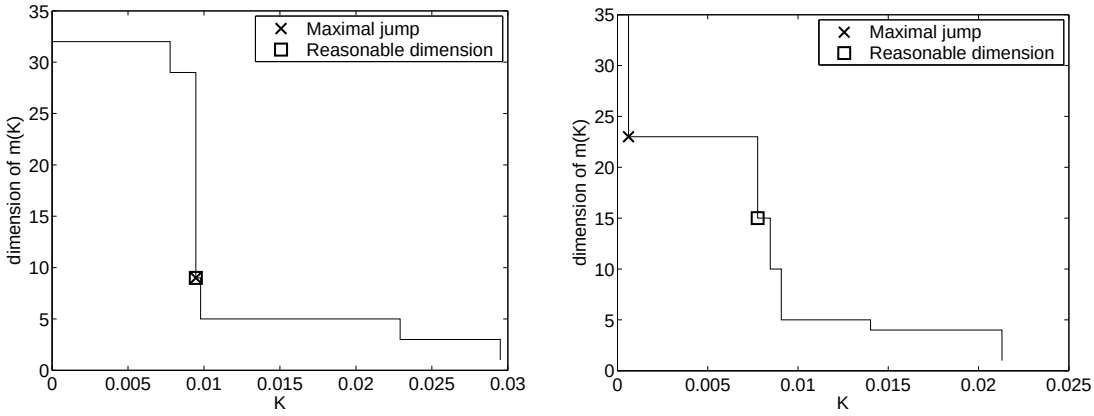
A.2. Definition of $\widehat{K}_{\min}$. We now come to the question of defining $\widehat{K}_{\min}$ as a function of $(\widehat{m}(K))_{K>0}$. As we have mentioned in Sect. 3.3.2, it corresponds to a "dimension jump", which should be observable since the whole trajectory of $\left(D_{\widehat{m}(K)} \right)_{K \geq 0}$ is known.

On Fig. 1, we represented $D_{\widehat{m}(K)}$ as a function of K for two simulated data sets. On the left (a), the dimension jump is quite clear, and we expect a formal definition of $\widehat{K}_{\min}$ to find this jump. The same picture holds for approximately 85% of the data sets. On the right (b), there seems to be several jumps, and a proper definition of $\widehat{K}_{\min}$ is problematic. What is sure is the necessity to find some automatic choice for $\widehat{K}_{\min}$, that is defining it properly.

We now propose two definitions that seem reasonable to us. For the first one, choose a threshold $D_{\text{reas.}}$, of order $n/(\ln(n))$, corresponding to the largest "reasonable" dimension for the selected model. Then, define

$$\widehat{K}_{\min} := \inf \left\{ K \text{ s.t. } D_{\widehat{m}(K)} \leq D_{\text{reas.}} \right\} \ .$$

With this definition, one can stop Algorithm 2 as soon as the threshold is reached. However, $\widehat{K}_{\min}$ may depend strongly on the choice of the threshold, which may not be quite obvious in the non-asymptotic situation (where $n/\ln(n)$ is not so far from n).



(a) One clear jump.

(b) Two jumps, two values for $\widehat{K}_{\min}$.

FIG 1. $D_{\widehat{m}(K)}$ *as a function of* K *for two different samples. Data are simulated with* $X \sim \mathcal{U}([0,1])$, $\epsilon \sim \mathcal{N}(0,1)$, $s(x) = \sin(\pi x)$, $\sigma \equiv 1$, $n = 200$. $(S_m)_{m \in \mathcal{M}_n}$ *is the collection of regular histogram models with dimension between 1 and* $n/(\ln(n))$. $\mathrm{pen}_{\mathrm{shape}}(m) = D_m$. *"Reasonable dimensions" are below* $n/(2\ln(n)) \approx 19$. *See Sect. 5.4 in [Arl07] for details (experiment (S1)).*

Our second idea is that $\widehat{K}_{\min}$ should match with the largest dimension jump, *i.e.*

$$\widehat{K}_{\min} := K_{i_{\max.\mathrm{jump}}} \qquad \text{with} \qquad i_{\max.\mathrm{jump}} = \arg \max_{i \in \{0,\dots,i_{\max}-1\}} \left\{ D_{m_{i+1}} - D_{m_i} \right\} \ .$$

Although this definition may seem less arbitrary than the previous one, it still depends strongly on $\mathcal{M}_n$, which may not contain so many large models for computational reasons. In order to ensure that there is a clear jump, an idea may be to add a few models of dimension $\approx n/2$, so that at least one[3] has a well-defined empirical risk minimizer $\widehat{s}_m$. This modification has the default of being quite arbitrary.

We compared the two definitions above ("reasonable dimension" *vs.* "maximal jump") on one thousand simulated data sets similar to the one of Fig. 1. Three cases occured:

1. The values of $\widehat{K}_{\min}$ do not differ (about 85% of the data sets; this is the (a) situation).
2. The values of $\widehat{K}_{\min}$ differ, but the selected models $\widehat{m}\left(\widehat{K}_{\min}\right)$ are still equal (about 8.5% of the data sets).
3. The finally selected models are different (about 6.5% of the data sets; this is the (b) situation).

Hence, in this non-asymptotic framework, the formal definition of $\widehat{K}_{\min}$ does not matter in general, but stays problematic in a few cases.

[3]several huge models may be necessary in practice, in order to decrease the variability of $\widehat{K}_{\min}$.

In terms of prediction error, we have compared the two methods by estimating the constant C_{or} that would appear in some oracle inequality:

$$C_{\text{or}} := \frac{\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\} \right]} \ .$$

With the "reasonable dimension" definition, $C_{\text{or}} \approx 1.88$. With the "maximal jump" definition, $C_{\text{or}} \approx 2.01$. As a comparison, Mallows' C_p (with a classical estimator of the variance σ^2) has a performance of $C_{\text{or}} \approx 1.93$ on the same data. For the three procedures, the standard deviation of the estimator of C_{or} is about 0.04. See [Arl07], Chap. 4, for more details. This preliminary simulation study shows that Algorithm 1 works efficiently (it is competitive with Mallows' C_p in a situation where this one is also optimal). It also suggests that the "reasonable dimension" definition may be better, but without very convincing evidence.

In order to make the choice of $\widehat{K}_{\min}$ as automatic as possible, we suggest to use simultaneously the two methods. When the selected models are not the same, then, send a warning to the final user, advising him to look at the curve $K \mapsto D_{\widehat{m}(K)}$ himself. Otherwise, stay confident in the automatic choice of $\widehat{m}(2\widehat{K}_{\min})$.

A.3. Proof of Prop. 1. First of all, since $\mathcal{M}_n$ is finite, the infimum in (14) is attained as soon as $G(m_{i-1}) \neq \emptyset$, so that m_i is well defined for every $i \leq i_{\max}$. Moreover, by construction, $g(m_i)$ decreases with i, so that all the $m_i \in \mathcal{M}_n$ are distinct. Hence, Algorithm 2 terminates and $i_{\max} + 1 \leq \text{Card}(\mathcal{M}_n)$. We now prove by induction the following property for every $i \in \{0, \ldots, i_{\max}\}$:

$$\mathcal{P}_i : \qquad K_i < K_{i+1} \quad \text{and} \quad \forall K \in [K_i, K_{i+1}), \quad \widehat{m}(K) = m_i \ .$$

Notice also that K_i can always be defined by (14) with the convention $\inf \emptyset = +\infty$.

$\mathcal{P}_0$ *holds true.* By definition of K_1, it is clear that $K_1 > 0$ (it may be equal to $+\infty$ if $G(m_0) = \emptyset$). For $K = K_0 = 0$, the definition of m_0 is the one of $\widehat{m}(0)$, so that $\widehat{m}(K) = m_0$. For $K \in (0, K_1)$, Lemma 2 shows that either $\widehat{m}(K) = \widehat{m}(0) = m_0$ or $\widehat{m}(K) \in G(0)$. In the latter case, by definition of K_1,

$$\frac{f(\widehat{m}(K)) - f(m_0)}{g(m_0) - g(\widehat{m}(K))} \geq K_1 > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_0) + Kg(m_0)$$

which is contradictory with the definition of $\widehat{m}(K)$. Hence, $\mathcal{P}_0$ holds true.

$\mathcal{P}_i \Rightarrow \mathcal{P}_{i+1}$ *for every* $i \in \{0, \dots, i_{\max} - 1\}$. Assume that $\mathcal{P}_i$ holds true. First, we have to prove that $K_{i+2} > K_{i+1}$. Since $K_{i_{\max}+1} = +\infty$, this is clear if $i = i_{\max} - 1$. Otherwise, $K_{i+2} < +\infty$ and m_{i+2} exists. Then, by definition of m_{i+2} and K_{i+2} (resp. m_{i+1} and K_{i+1}), we have

$$f(m_{i+2}) - f(m_{i+1}) = K_{i+2}(g(m_{i+1}) - g(m_{i+2})) \tag{15}$$

$$f(m_{i+1}) - f(m_i) = K_{i+1}(g(m_i) - g(m_{i+1})) \ . \tag{16}$$

Moreover, $m_{i+2} \in G(m_{i+1}) \subset G(m_i)$, and $m_{i+2} \prec m_{i+1}$ (because g is non-decreasing). Using again the definition of K_{i+1}, we have

$$f(m_{i+2}) - f(m_i) > K_{i+1}(g(m_i) - g(m_{i+2})) \tag{17}$$

(otherwise, we would have $m_{i+2} \in F_{i+1}$ and $m_{i+2} \prec m_{i+1}$, which is not possible). Combining the difference of (17) and (16) with (15), we have

$$K_{i+2}(g(m_{i+1}) - g(m_{i+2})) > K_{i+1}(g(m_{i+1}) - g(m_{i+2})) \ ,$$

so that $K_{i+2} > K_{i+1}$ (since $g(m_{i+1}) > g(m_{i+2})$).

Second, we prove that $\widehat{m}(K_{i+1}) = m_{i+1}$. From $\mathcal{P}_i$, we know that for every $m \in \mathcal{M}_n$, for every $K \in [K_i, K_{i+1})$, $f(m_i) + Kg(m_i) \leq f(m) + Kg(m)$. Taking the limit when K goes to K_{i+1}, we obtain that $m_i \in E(K_{i+1})$. By (16), we then have $m_{i+1} \in E(K_{i+1})$. On the other hand, if $m \in E(K_{i+1})$, Lemma 2 shows that either $f(m) = f(m_i)$ and $g(m) = g(m_i)$ or $m \in G(m_i)$. In the first case, $m_{i+1} \prec m$ (because g is non-decreasing). In the second one, $m \in F_{i+1}$, so $m_{i+1} \preceq m$. Since $\widehat{m}(K_{i+1})$ is the smallest element of $E(K_{i+1})$, we have proven that $m_{i+1} = \widehat{m}(K_{i+1})$.

Last, we have to prove that $\widehat{m}(K) = m_{i+1}$ for every $K \in (K_1, K_2)$. From the last statement of Lemma 2, we have either $\widehat{m}(K) = \widehat{m}(K_1)$ or $\widehat{m}(K_1) \in G(\widehat{m}(K))$. In the latter case (which is only possible if $K_{i+2} < \infty$), by definition of K_{i+2},

$$\frac{f(\widehat{m}(K)) - f(m_{i+1})}{g(m_{i+1}) - g(\widehat{m}(K))} \geq K_{i+2} > K$$

so that

$$f(\widehat{m}(K)) + Kg(\widehat{m}(K)) > f(m_{i+1}) + Kg(m_{i+1})$$

which is contradictory with the definition of $\widehat{m}(K)$. □

LEMMA 2. *Use the notations of Prop. 1 and its proof. If* $0 \leq K < K'$, $m \in E(K)$ *and* $m' \in E(K')$, *then we have either*

(a) $f(m) = f(m')$ *and* $g(m) = g(m')$.
(b) $f(m) < f(m')$ *and* $g(m) > g(m')$.

In particular, we have either $\widehat{m}(K) = \widehat{m}(K')$ *or* $\widehat{m}(K') \in G(\widehat{m}(K))$.

PROOF OF LEMMA 2. By definition of $E(K)$ and $E(K')$, we have

$$f(m) + Kg(m) \leq f(m') + Kg(m') \tag{18}$$

$$f(m') + K'g(m') \leq f(m) + K'g(m) \ . \tag{19}$$

Summing (18) and (19) gives $(K' - K)g(m') \leq (K' - K)g(m)$ so that

$$g(m') \leq g(m) \ . \tag{20}$$

Since $K \geq 0$, (18) and (20) give $f(m) + Kg(m) \leq f(m') + Kg(m)$, *i.e.*

$$f(m) \leq f(m') \ . \tag{21}$$

Moreover, using (19), $g(m) = g(m')$, implies $f(m') \leq f(m)$, *i.e.* $f(m) = f(m')$ by (21). In the same way, (18) and (20) show that $f(m) = f(m')$ imply $g(m) = g(m')$. In both cases, (a) is satisfied. Otherwise, $f(m) < f(m')$ and $g(m) > g(m')$, *i.e.* (b) is satisfied.

The last statement follows by taking $m = \widehat{m}(K)$ and $m' = \widehat{m}(K')$, because g is non-decreasing, so that the minimum of g in $E(K)$ is attained by $\widehat{m}(K)$. □

APPENDIX B: PROOFS

B.1. Conventions and notations. In the following, when we do not want to write explicitly some constants, we use the letter L. It means "some absolute constant, possibly different from a line to another, or even within the same line". When L is not numerical, but depends on some parameters $p_1, \ldots, p_k$, it is written $L_{p_1,\ldots,p_k}$. $L_{(\mathbf{SH1})}$ (resp. $L_{(\mathbf{SH2})}$) denotes a constant that depends only on the set of assumptions of Thm. 3 (resp. Thm. 2), including (**P1**) and (**P2**).

We also make use of the following notations: for every $a, b \in \mathbb{R}$, $a \wedge b$ is the minimum of a and b, $a \vee b$ is the maximum of a and b, $a_+ = a \vee 0$ is the positive part of a and $a_- = a \wedge 0$ is its negative part.

B.2. A general oracle inequality. First of all, let us state a general theorem, from which Thm. 1 is an obvious corollary.

THEOREM 3. *Assume that the data $(X_i, Y_i)_{1 \leq i \leq n}$ are i.i.d. and satisfy the following:*

(**Ab**) *Bounded data: $\|Y_i\|_\infty \leq A < \infty$.*

(**An**) *Noise-level bounded from below: $\sigma(X_i) \geq \sigma_{\min} > 0$ a.s.*

(**Ap**) *Polynomial decreasing of the bias: there exists $\beta_1 \geq \beta_2 > 0$ and $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ such that*

$$C_{\mathrm{b}}^- D_m^{-\beta_1} \leq l(s, s_m) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \ .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Lower regularity of the partitions for $\mathcal{L}(X)$: $D_m \min_{\lambda \in \Lambda_m} p_\lambda \geq c_{\mathrm{r},\ell}^X$.*

Let $c_1, c_2, C_1, C_2 \geq 0$ such that $c_2 > 1$ and assume that for every $m \in \mathcal{M}_n$,

$$\begin{aligned} &\mathbb{E}\left[c_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) + c_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \\ &\quad \leq \operatorname{pen}(m) \leq \mathbb{E}\left[C_1 P\left(\gamma(\widehat{s}_m) - \gamma(s_m)\right) + C_2 P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)\right] \end{aligned} \tag{22}$$

with probability at least $1 - Ln^{-2}$.

Then, if $\widehat{m}$ is defined by (6), there exists a constant K_1 and a sequence ϵ_n converging to zero at infinity such that, with probability at least $1 - K_1 n^{-2}$,

$$l(s, \widehat{s}_{\widehat{m}}) \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n\right] \inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\} \quad . \tag{23}$$

Moreover, we have the oracle inequality

$$\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] \leq \left[\frac{1 + (C_1 + C_2 - 2)_+}{(c_1 + c_2 - 1) \wedge 1} + \epsilon_n\right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\}\right] + \frac{A^2 K_1}{n^2} \quad . \tag{24}$$

The constant K_1 may depend on c_1, c_2 and constants in (**P1**), (**P2**), (**Ab**), (**An**), (**Ap**) *and* ($\mathbf{Ar}_\ell^{\mathbf{X}}$), *but not on n. The small term ϵ_n depends only on n (it can for instance be upperbounded by $\ln(n)^{-1/5}$).*

The particular form of condition (22) on the penalty is motivated by the fact that the ideal shape of penalty $\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m)\right]$ (or equivalently $\mathbb{E}\left[2p_2(m)\right]$) is unknown in general. Then, it has to be estimated from the data, for instance by resampling. Notice also that (22) can be assumed only for the models of dimension larger than $\ln(n)^\xi$ (for some $\xi \geq 0$), at the price of making K_1 depend on $\xi > 0$. Under the assumptions of Thm. 3, it has been proven ([Arl07], Chap. 5 and 6; see also [Arl08b, Arl08a]) that resampling penalties satisfy condition (22) with constants $c_1 + c_2 = 2 - \delta_n$ and $C_1 + C_2 = 2 + \delta_n$ (for some absolute sequence δ_n converging to zero at infinity), at least for models of dimension larger than $\ln(n)^\xi$ (where ξ depends on the constants in the assumptions on the data).

In such a situation (obtained by resampling or not), (23) shows that we have an *asymptotically optimal* model selection procedure.

The rationale behind this theorem is that if pen is close to $c_1 p_1 + c_2 p_2$, then $\operatorname{crit}(m) = l(s, s_m) + c_1 p_1(m) + (c_2 - 1) p_2(m)$. If $c_1 = c_2 = 1$, this is exactly the ideal criterion $l(s, \widehat{s}_m)$. If $c_1 + c_2 = 2$ with $c_1 \geq 0$ and $c_2 > 1$, we obtain the same result because $p_1(m)$ and $p_2(m)$ are quite close (at least when D_m is large). This closeness between p_1 and p_2 is the keystone of the slope heuristics. Notice that if $\max_{m \in \mathcal{M}_n} D_m \leq K_1'(\ln(n))^{-1} n$ (for some constant K_1' depending only on the assumptions of Thm. 1, like K_1), one can replace the condition $c_2 > 1$ by $c_1 + c_2 > 1$ and $c_1, c_2 \geq 0$.

B.3. Proof of Thm. 3. This proof is very similar to the one of Thm. 5.1 of [Arl07]. We give it for the sake of completeness.

From (3), we have for each $m \in \mathcal{M}_n$ such that $A_n(m) := \min_{\lambda \in \Lambda_m} \{ n\widehat{p}_\lambda \} > 0$

$$l(s, \widehat{s}_{\widehat{m}}) - \left(\mathrm{pen}'_{\mathrm{id}}(\widehat{m}) - \mathrm{pen}(\widehat{m}) \right) \leq l(s, \widehat{s}_m) + \left(\mathrm{pen}(m) - \mathrm{pen}'_{\mathrm{id}}(m) \right) \quad . \tag{25}$$

with $\mathrm{pen}'_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \overline{\delta}(m) = \mathrm{pen}(m) + (P - P_n)\gamma(s)$. It is sufficient to control $\mathrm{pen} - \mathrm{pen}'_{\mathrm{id}}$ for every $m \in \mathcal{M}_n$.

We will thus use the concentration inequalities of Sect. B.5 with $x = \gamma \ln(n)$ and $\gamma = 2 + \alpha_{\mathcal{M}}$. Define $B_n(m) = \min_{\lambda \in \Lambda_m} \{ np_\lambda \}$. Let Ω_n be the event on which

- for every $m \in \mathcal{M}_n$, (22) holds
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$:

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m) \right] - L_{(\mathbf{SH1})} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n(m)} \right] \mathbb{E}\left[p_2(m) \right] \tag{35}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m) \right] + L_{(\mathbf{SH1})} \left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m} e^{-LB_n(m)} \right] \mathbb{E}\left[p_2(m) \right] \tag{36}$$

- for every $m \in \mathcal{M}_n$ such that $B_n(m) > 0$:

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1) B_n(m)^{-1} \ln(n)} - \frac{L_{(\mathbf{SH1})} \ln(n)^2}{\sqrt{D_m}} \right) \mathbb{E}\left[p_2(m) \right] \quad . \tag{37}$$

$$|p_2(m) - \mathbb{E}\left[p_2(m) \right]| \leq \frac{L_{(\mathbf{SH1})} \ln(n)}{\sqrt{D_m}} \left[l(s, s_m) + \mathbb{E}\left[p_2(m) \right] \right] \tag{33}$$

$$\left| \overline{\delta}(m) \right| \leq \frac{l(s, s_m)}{\sqrt{D_m}} + L_{(\mathbf{SH1})} \frac{\ln(n)}{\sqrt{D_m}} \mathbb{E}\left[p_2(m) \right] \tag{31}$$

From Prop. 5 (for $\widetilde{p_1}$), Prop. 4 (for p_2), Prop. 3 (for $\overline{\delta}(m)$), we have

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

For every $m \in \mathcal{M}_n$ such that $D_m \leq L_{c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$, $(\mathbf{Ar}^{\mathbf{X}}_\ell)$ implies that $B_n(m) \geq L^{-1} \ln(n) \geq 1$. As a consequence, on Ω_n, if $\ln(n)^7 \leq D_m \leq L_{c^X_{\mathrm{r},\ell}} n \ln(n)^{-1}$:

$$\begin{aligned} \max \left\{ \left| \widetilde{p_1}(m) - \mathbb{E}\left[\widetilde{p_1}(m) \right] \right|, \left| p_2(m) - \mathbb{E}\left[p_2(m) \right] \right|, \left| \overline{\delta}(m) \right| \right\} \\ \leq \frac{L_{(\mathbf{SH1})} \mathbb{E}\left[l(s, s_m) + p_2(m) \right]}{\ln(n)} \end{aligned}$$

Using (38) (in Prop. 6) and the fact that $B_n(m) \geq L^{-1} \ln(n)$,

$$\frac{(c_1+c_2)\left(1-\widetilde{\delta_n}\right)}{2} \leq \mathbb{E}\left[\mathrm{pen}(m)\right] \leq \frac{(C_1+C_2)\left(1+\widetilde{\delta_n}\right)}{2}\mathbb{E}\left[\widetilde{p_1}(m)+p_2(m)\right]$$

with $0 \leq \widetilde{\delta_n} \leq L\ln(n)^{-1/4}$. We deduce: if $n \geq L_{(\mathbf{SH1})}$, for every $m \in \mathcal{M}_n$ such that $\ln(n)^7 \leq D_m \leq L_{c_{\mathrm{r},\ell}^X} n \ln(n)^{-1}$, on Ω_n,

$$\begin{aligned}\left[(c_1+c_2-2)_- - \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right] p_1(m) &\leq (\mathrm{pen}-\mathrm{pen}'_{\mathrm{id}})(m) \\ &\leq \left[(C_1+C_2-2)_+ + \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right] p_1(m) \ .\end{aligned}$$

We need to assume that n is large enough in order to upper bound $\mathbb{E}\left[p_2(m)\right]$ in terms of $p_1(m)$, since we only have

$$p_1(m) \geq \left[1 - \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right]_+ \mathbb{E}\left[p_2(m)\right]$$

in general.

Combined with (25), this gives: if $n \geq L_{(\mathbf{SH1})}$,

$$\begin{aligned} l(s,\widehat{s}_{\widehat{m}})\mathbb{1}_{\ln(n)^5 \leq D_{\widehat{m}} \leq L_{c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \leq \left[\frac{1+(C_1+C_2-2)_+}{(c_1+c_2-1)\wedge 1} + \frac{L_{(\mathbf{SH1})}}{\ln(n)^{1/4}}\right] \\ \times \inf_{m\in\mathcal{M}_n \ \text{s.t.} \ \ln(n)^7 \leq D_m \leq L_{\alpha_{\mathcal{M}},c_{\mathrm{r},\ell}^X} n\ln(n)^{-1}} \left\{ l(s,\widehat{s}_m)\right\} \ . \end{aligned} \tag{26}$$

Define the oracle model $m^\star \in \arg\min\left\{ l(s,\widehat{s}_m)\right\}$. We prove below that for any $c > 0$ and $\alpha > (1-\beta_2)_+/2$, if $n \geq L_{(\mathbf{SH1}),c,\alpha}$, then, on Ω_n:

$$\ln(n)^7 \leq D_{\widehat{m}} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1} \tag{27}$$

$$\ln(n)^7 \leq D_{m^\star} \leq n^{1/2+\alpha} \leq cn\ln(n)^{-1} \ . \tag{28}$$

The result follows since $L_{(\mathbf{SH1})}\ln(n)^{-1/4} \leq \epsilon_n = \ln(n)^{-1/5}$ for $n \geq L_{(\mathbf{SH1})}$. We finally remove the condition $n \geq n_0 = L_{(\mathbf{SH1})}$ by choosing $K_1 = L_{(\mathbf{SH1})}$ such that $K_1 n_0^{-2} \geq 1$.

Proof of (27). By definition, $\widehat{m}$ minimizes $\mathrm{crit}(m)$ over $\mathcal{M}_n$. It thus also minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s,s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $\mathcal{M}_n$.

1. Lower bound on $\mathrm{crit}'(m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. We then have

$$\begin{aligned} & l(s, s_m) \geq C_{\mathrm{b}}^- \, (\ln(n))^{-7\beta_1} \qquad \text{from } (\mathbf{Ap}) \\ & \mathrm{pen}(m) \geq 0 \\ & p_2(m) \leq L_{(\mathbf{SH1})} \sqrt{\frac{\ln(n)}{n}} + L_{(\mathbf{SH1})} \frac{D_m}{n} \leq L_{(\mathbf{SH1})} \sqrt{\frac{\ln(n)}{n}} \qquad \text{from (32)} \end{aligned}$$

and from (31) (in Prop. 3),

$$\overline{\delta}(m) \geq -L_A \sqrt{\frac{l(s, s_m) \ln(n)}{n}} + L_A \frac{\ln(n)}{n} \geq -L_A \sqrt{\frac{\ln(n)}{n}} \ .$$

We then have

$$\mathrm{crit}'(m) \geq L_{(\mathbf{SH1})} \, (\ln(n))^{-L_{\beta_1}} \ .$$

2. Lower bound for large models: let $m \in \mathcal{M}_n$ such that $D_m \geq n^{1/2+\alpha}$. From (22) and (32) (in Prop. 4),

$$\begin{aligned} \mathrm{pen}(m) - p_2(m) &\geq (c_2 - 1) \, \mathbb{E}\left[p_2(m)\right] - L_A \sqrt{\frac{\ln(n)}{n}} \\ &\geq \frac{(c_2 - 1)\sigma_{\min}^2 D_m}{n} - L_A \sqrt{\frac{\ln(n)}{n}} \end{aligned}$$

and from (29),

$$\overline{\delta}(m) \geq -L_{(\mathbf{SH1})} \sqrt{\frac{\ln(n)}{n}} \ .$$

Hence, if $D_m \geq n^{1/2+\alpha}$ and $n \geq L_{(\mathbf{SH1}),\alpha}$

$$\mathrm{crit}'(m) \geq \mathrm{pen}(m) + \overline{\delta}(m) - p_2(m) \geq L_{(\mathbf{SH1}),\alpha} n^{-1/2+\alpha} \ .$$

3. There exists a better model for $\mathrm{crit}(m)$: from (**P2**), there exists $m_0 \in \mathcal{M}_n$ such that $\sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}} \sqrt{n}$. If moreover $n \geq L_{c_{\mathrm{rich}},\alpha}$, then

$$\ln(n)^7 \leq \sqrt{n} \leq D_{m_0} \leq c_{\mathrm{rich}} \sqrt{n} \leq n^{1/2+\alpha} \ .$$

By (39) in Lemma 7, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.
Using (**Ap**),

$$l(s, s_{m_0}) \leq C_{\mathrm{b}}^+ c_{\mathrm{rich}}^{\beta_2} n^{-\beta_2/2}$$

so that, when $n \geq L_{(\mathbf{SH1})}$,

$$\begin{aligned}\operatorname{crit}'(m_0) &\leq l(s, s_{m_0}) + \left|\overline{\delta}(m)\right| + \operatorname{pen}(m) \\ &\leq L_{(\mathbf{SH1})} \left(n^{-\beta_2/2} + n^{-1/2} \right) \quad .\end{aligned}$$

If $n \geq L_{(\mathbf{SH1}),\alpha}$, this upper bound is smaller than the previous lower bounds for small and large models.

Proof of (28). Recall that $m^\star$ minimizes $l(s, \widehat{s}_m) = l(s, s_m) + p_1(m)$ over $m \in \mathcal{M}_n$, with the convention $l(s, \widehat{s}_m) = \infty$ if $A_n(m) = 0$.

1. Lower bound on $l(s, \widehat{s}_m)$ for small models: let $m \in \mathcal{M}_n$ such that $D_m < (\ln(n))^7$. From (**Ap**), we have
$$l(s, \widehat{s}_m) \geq l(s, s_m) \geq C_{\mathrm{b}}^- \left(\ln(n)\right)^{-7\beta_1} \quad .$$
2. Lower bound on $l(s, \widehat{s}_m)$ for large models: let $m \in \mathcal{M}_n$ such that $D_m > n^{1/2+\alpha}$. From (37), for $n \geq L_{(\mathbf{SH1}),\alpha}$,
$$\begin{gathered}\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1) \left(c_{\mathrm{r},\ell}^X \right)^{-1} \ln(n)} - \frac{L_{(\mathbf{SH1}),\alpha}}{n^{1/4}} \right) \mathbb{E}\left[\widetilde{p_2}(m)\right] \\ \text{so that} \quad l(s, \widehat{s}_m) \geq \widetilde{p_1}(m) \geq L_{(\mathbf{SH1}),\alpha} n^{-1/2+\alpha} \quad .\end{gathered}$$
3. There exists a better model for $l(s, \widehat{s}_m)$: let $m_0 \in \mathcal{M}_n$ be as in the proof of (27) and assume that $n \geq L_{c_{\mathrm{rich}},\alpha}$. Then,
$$p_1(m_0) \leq L_{(\mathbf{SH1})} \mathbb{E}\left[p_2(m)\right] \leq L_{(\mathbf{SH1})} n^{-1/2}$$
and the arguments of the previous proof show that
$$l(s, \widehat{s}_{m_0}) \leq L_{(\mathbf{SH1})} \left(n^{-\beta_2/2} + n^{-1/2} \right)$$
which is smaller than the previous upper bounds for $n \geq L_{(\mathbf{SH1}),\alpha}$.

Classical oracle inequality. Let Ω_n be the event on which (23) holds true. Then,

$$\begin{aligned}\mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}})\right] &= \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\Omega_n}\right] + \mathbb{E}\left[l(s, \widehat{s}_{\widehat{m}}) \mathbb{1}_{\Omega_n^c}\right] \\ &\leq \left[2\eta - 1 + \epsilon_n\right] \mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ l(s, \widehat{s}_m) \right\}\right] + A^2 K_1 \mathbb{P}\left(\Omega_n^c\right)\end{aligned}$$

which proves (24).

B.4. Proof of Thm. 2. Similarly to the proof of Thm. 3, we consider the event Ω'_n, of probability at least $1 - L_{c_{\mathcal{M}}} n^{-2}$, on which:

- for every $m \in \mathcal{M}_n$, (11) (for pen), (37) (for $\widetilde{p_1}$), (32)–(33) (for p_2, with $x = \gamma \ln(n)$ and $\theta = \sqrt{\ln(n)/n}$) and (29)–(31) (for $\overline{\delta}$, with $x = \gamma \ln(n)$ and $\eta = \sqrt{\ln(n)/n}$) hold true.
- for every $m \in \mathcal{M}_n$ such that $B_n(m) \geq 1$, (35) and (36) hold (for $\widetilde{p_1}$).

Lower bound on $D_{\widehat{m}}$. By definition, $\widehat{m}$ minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(s) = l(s, s_m) - p_2(m) + \overline{\delta}(m) + \mathrm{pen}(m)$$

over $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$. As in the proof of Thm. 3, we define $c = L_{c^X_{\mathrm{r},\ell}} > 0$ such that for every model of dimension $D_m \leq cn \ln(n)^{-1}$, $B_n(m) \geq L^{-1} \ln(n) \geq 1$. Let $d < 1$ to be chosen later.

1. Lower bound on $\mathrm{crit}'(m)$ for "small" models: assume that $m \in \mathcal{M}_n$ and $D_m \leq dcn \ln(n)^{-1}$. Then, $l(s, s_m) + \mathrm{pen}(m) \geq 0$ and from (29),

$$\overline{\delta}(m) \geq -L_A \sqrt{\frac{\ln(n)}{n}} \ .$$

   If $D_m \geq \ln(n)^4$, (33) implies that

$$p_2(m) \leq \left(1 + \frac{L_{(\mathbf{SH2})}}{\ln(n)}\right) \mathbb{E}\left[p_2(m)\right] \leq \frac{L_{(\mathbf{SH2})} D_m}{n} \leq \frac{cdL_{(\mathbf{SH2})}}{\ln(n)} \ .$$

   On the other hand, if $D_m < \ln(n)^4$, (32) implies that

$$p_2(m) \leq L_{(\mathbf{SH2})} \sqrt{\frac{\ln(n)}{n}} \ .$$

   We then have

$$\mathrm{crit}'(m) \geq -dL_{(\mathbf{SH2})} \left(\ln(n)\right)^{-1} \ .$$

2. There exists a better model for $\mathrm{crit}(m)$: let $m_1 \in \mathcal{M}_n$ such that

$$\ln(n)^4 \leq \frac{cdn}{c_{\mathrm{rich}} \ln(n)} \leq D_{m_1} \leq \frac{cn}{\ln(n)} \leq n \ .$$

   From (**P2**), this is possible as soon as $n \geq L_{c_{\mathrm{rich}},c,d}$. By (39) in Lemma 7, $A_n(m_0) \geq 1$ with probability at least $1 - Ln^{-2}$.

We then have

$$\begin{aligned} l(s,s_{m_1}) &\leq L_{(\mathbf{SH2}),c}\ln(n)^{\beta_2}n^{-\beta_2} \qquad \text{by } (\mathbf{Ap}) \\ p_2(m_1) &\leq \left(1+\frac{L_{(\mathbf{SH2})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m_1)\right] \qquad \text{by (33)} \\ \mathrm{pen}(m_1) &\leq C_2\mathbb{E}\left[p_2(m_1)\right] \qquad \text{by (11)} \\ \left|\overline{\delta}(m_1)\right| &\leq L_A\sqrt{\frac{\ln(n)}{n}} \qquad \text{by (29)} \end{aligned}$$

so that

$$\begin{aligned} \mathrm{crit}'(m_1) &\leq L_{(\mathbf{SH2}),c}\ln(n)^{\beta_2}n^{-\beta_2} + \left(C_2 - 1 - \frac{L_{(\mathbf{SH2})}}{\ln(n)}\right)\mathbb{E}\left[p_2(m_1)\right] + L_A\sqrt{\frac{\ln(n)}{n}} \\ &\leq \frac{(C_2-1)\sigma_{\min}^2 c}{2\ln(n)} \end{aligned}$$

if $n \geq L_{(\mathbf{SH2}),c}$.
We now choose d such that the constant $dL_{(\mathbf{SH2})}$ appearing in the lower bound on $\mathrm{crit}'(m)$ for "small" models is smaller than $(1-C_2)\sigma_{\min}^2 c/2$, i.e. $d \leq L_{(\mathbf{SH2}),c}$. Then, we assume that $n \geq n_0 = L_{(\mathbf{SH2}),c,d} = L_{(\mathbf{SH2})}$. Finally, we remove this condition as before by enlarging K_2.

Risk of $D_{\widehat{m}}$. The proof of (13) is quite similar to the one of (28). First, for every model $m \in \mathcal{M}_n$ such that $A_n(m) \geq 1$ and $D_m \geq K_3 n\ln(n)^{-1}$, we have

$$l(s,\widehat{s}_m) \geq \widetilde{p_1}(m) \geq L_{(\mathbf{SH2})}K_3\ln(n)^{-2} \qquad \text{by (37)} \ .$$

Then, the model $m_0 \in \mathcal{M}_n$ defined previously satisfies $A_n(m) \geq 1$, and

$$l(s,\widehat{s}_{m_0}) \leq L_{(\mathbf{SH2})}\left(n^{-\beta_2/2} + n^{-1/2}\right) \ .$$

If $n \geq L_{(\mathbf{SH2})}$, the ratio between these two bounds is larger than $\ln(n)$, so that (13) holds.

B.5. Concentration inequalities used in the main proofs. We do not always assume in this section that models are made of histograms, but only that they are bounded by some finite A. First, we can control $\overline{\delta}(m)$ with general models and bounded data.

PROPOSITION 3. *Assume that $\|Y\|_\infty \leq A < \infty$. Then for all $x \geq 0$, on an event of probability at least $1-2e^{-x}$:*

$$\forall \eta > 0, \quad \left|\overline{\delta}(m)\right| \leq \eta l(s,s_m) + \left(\frac{4}{\eta}+\frac{8}{3}\right)\frac{A^2 x}{n} \ . \tag{29}$$

If moreover

$$Q_m^{(p)} := \frac{n\mathbb{E}\left[p_2(m)\right]}{D_m} > 0 \ , \tag{30}$$

on the same event,

$$\left|\overline{\delta}(m)\right| \leq \frac{l(s,s_m)}{\sqrt{D_m}} + \frac{20}{3}\frac{A^2}{Q_m^{(p)}}\frac{\mathbb{E}[p_2(m)]}{\sqrt{D_m}}x \ . \tag{31}$$

Remark 1. In the histogram case,

$$Q_m^{(p)} = \frac{1}{D_m}\sum_{\lambda\in\Lambda_m}\left[(\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right] \geq (\sigma_{\min})^2 > 0 \ .$$

Then, we derive a concentration inequality for $p_2(m)$ in the histogram case from a general result of [BM04] (Thm. 2.2 in a preliminary version).

Proposition 4. *Let S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\|Y\|_\infty \leq A$ and define $p_2(m) = P_n\left(\gamma(s_m) - \gamma(\widehat{s}_m)\right)$.*

Then, for every $x \geq 0$, there exists an event of probability at least $1 - e^{1-x}$ on which for every $\theta \in (0;1)$,

$$|p_2(m) - \mathbb{E}\left[p_2(m)\right]| \leq C\left[\theta l(s,s_m) + \frac{A^2\sqrt{D_m}\sqrt{x}}{n} + \frac{A^2 x}{\theta n}\right] \tag{32}$$

for some absolute constant C. If moreover $\sigma(X) \geq \sigma_{\min} > 0$ a.s., we have on the same event:

$$|p_2(m) - \mathbb{E}\left[p_2(m)\right]| \leq \frac{C}{\sqrt{D_m}}\left[l(s,s_m) + \frac{A^2\mathbb{E}\left[p_2(m)\right]}{\sigma_{\min}^2}\left(\sqrt{x} + x\right)\right] \ . \tag{33}$$

Finally, we recall a concentration inequality for $p_1(m)$ that comes from [Arl07]. Its proof is particular to the histogram case. Moreover, since $\mathbb{E}\left[p_1\right]$ is not well-defined (because of the event $\{\min_{\lambda\in\Lambda_m}\{\widehat{p}_\lambda\} = 0\}$), we have to take the following convention

$$p_1(m) = \widetilde{p_1}(m) = \sum_{\lambda\in\Lambda_m \text{ s.t. } \widehat{p}_\lambda > 0} p_\lambda\left(\beta_\lambda - \widehat{\beta}_\lambda\right)^2) + \sum_{\lambda\in\Lambda_m \text{ s.t. } \widehat{p}_\lambda = 0} p_\lambda\left((\sigma_\lambda^r)^2 + \left(\sigma_\lambda^d\right)^2\right) \ . \tag{34}$$

Remark that $p_1(m) = \widetilde{p_1}(m)$ when $\min_{\lambda\in\Lambda_m}\{\widehat{p}_\lambda\} > 0$), so that this convention has no consequences on the final results (Thm. 3 and 2).

PROPOSITION 5 (Prop. 5.8, [Arl07]). *Let $\gamma > 0$ and S_m be the model of histograms associated with the partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\|Y\|_\infty \leq A < \infty$, $\sigma(X) \geq \sigma_{\min} > 0$ a.s. and $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B_n > 0$. Then, if $B_n \geq 1$, on an event of probability at least $1 - Ln^{-\gamma}$,*

$$\widetilde{p_1}(m) \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - L_{A,\sigma_{\min},\gamma}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \tag{35}$$

$$\widetilde{p_1}(m) \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + L_{A,\sigma_{\min},\gamma}\left[\frac{\ln(n)^2}{\sqrt{D_m}} + \sqrt{D_m}e^{-LB_n}\right]\mathbb{E}\left[p_2(m)\right] \quad . \tag{36}$$

If we only have a lower bound $B_n > 0$, then, with probability at least $1 - Ln^{-\gamma}$,

$$\widetilde{p_1}(m) \geq \left(\frac{1}{2 + (\gamma+1)B_n^{-1}\ln(n)} - \frac{L_{A,\sigma_{\min},\gamma}\ln(n)^2}{\sqrt{D_m}}\right)\mathbb{E}\left[p_2(m)\right] \quad . \tag{37}$$

PROOF. We changed a little the assumptions of the original proposition. The result still holds since $P_m^\ell(q) \leq 4A^2\sigma_{\min}^{-2}$ (with the notations of [Arl07]). □

B.6. Additional results needed. A crucial result in the proofs of Thm. 3 and 2 is that $p_1(m)$ and $p_2(m)$ are close in expectation. This comes from [Arl07] (Sect. 5.7.2).

PROPOSITION 6 (Lemma 5.6, [Arl07]). *Let S_m be a model of histograms adapted to some partition $(I_\lambda)_{\lambda\in\Lambda_m}$. Assume that $\min_{\lambda\in\Lambda_m}\{np_\lambda\} \geq B > 0$. Then,*

$$\begin{aligned}\left(1 - e^{-B}\right)^2\mathbb{E}\left[p_2(m)\right] &\leq \mathbb{E}\left[\widetilde{p_1}(m)\right] \\ &\leq \left[2 \wedge \left(1 + 5.1 \times B^{-1/4}\right) + (B \vee 1)\, e^{-(B\vee 1)}\right]\mathbb{E}\left[p_2(m)\right] \quad .\end{aligned} \tag{38}$$

Finally, we need the following technical lemma in the proof of the main theorems.

LEMMA 7. *Let $(p_\lambda)_{\lambda\in\Lambda_m}$ be non-negative real numbers of sum 1, $(n\widehat{p}_\lambda)_{\lambda\in\Lambda_m}$ a multinomial vector of parameters $(n; (p_\lambda)_{\lambda\in\Lambda_m})$. Then, for all $\gamma > 0$,*

$$\min_{\lambda\in\Lambda_m}\{n\widehat{p}_\lambda\} \geq \frac{\min_{\lambda\in\Lambda_m}\{np_\lambda\}}{2} - 2(\gamma+1)\ln(n) \tag{39}$$

with probability at least $1 - 2n^{-\gamma}$.

PROOF OF LEMMA 7. By Bernstein inequality ([Mas07], Prop. 2.9), for all $\lambda \in \Lambda_m$,

$$\mathbb{P}\left(n\widehat{p}_\lambda \geq (1-\theta)np_\lambda - \sqrt{2npx} - \frac{x}{3}\right) \geq 1 - e^{-x} \quad .$$

Take $x = (\gamma+1)\ln(n)$ above, and remark that $\sqrt{2npx} \leq \frac{np}{2} + x$. The union bound gives the result since $\mathrm{Card}(\Lambda_m) \leq n$. □

B.7. Proof of Prop. 3. Since $\|Y\|_\infty \leq A$, we have $\|s\|_\infty \leq A$ and $\|s_m\|_\infty \leq A$. In fact, everything happens as if $S_m \cup \{s\}$ was bounded by A in L^∞.

We have

$$\overline{\delta}(m) = \frac{1}{n}\sum_{i=1}^{n} \left(\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i)) - \mathbb{E}\left[\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right]\right)$$

and assumptions of Bernstein inequality ([Mas07], Prop. 2.9) are fulfilled with

$$c = \frac{8A^2}{3n} \quad \text{and} \quad v = \frac{8A^2 l(s,s_m)}{n}$$

since

$$\left\|\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i)) - \mathbb{E}\left[\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right]\right\|_\infty \leq 8A^2$$

and

$$\begin{aligned} \operatorname{var}\left(\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right) &\leq \mathbb{E}\left[\left(\gamma(s_m,(X_i,Y_i)) - \gamma(s,(X_i,Y_i))\right)^2\right] \\ &\leq 8A^2 l(s,s_m) \end{aligned} \tag{40}$$

because $\|s_m - s\|_\infty \leq 2A$ and

$$\left(\gamma(t,\cdot) - \gamma(s,\cdot)\right)^2 = \left(t(X) - s(X)\right)^2 \left(2(Y - s(X)) - t(X) + s(X)\right)^2$$
$$\text{and} \quad \mathbb{E}\left[(Y - s(X))^2 \,\middle|\, X\right] \leq \frac{(2A)^2}{4} = A \ .$$

We obtain that, with probability at least $1 - 2e^{-x}$,

$$\left|\overline{\delta}(m)\right| \leq \sqrt{2vx} + c = \sqrt{\frac{16A^2 l(s,s_m)x}{n}} + \frac{8A^2 x}{3n}$$

and (29) follows since $2\sqrt{ab} \leq a\eta + b\eta^{-1}$ for all $\eta > 0$. Taking $\eta = D_m^{-1/2} \leq 1$ and using $Q_m^{(p)}$ defined by (30), we deduce (31).

B.8. Proof of Prop. 4. We apply here a result from [BM04] (Thm. 2.2 in a preliminary version), in which it is only assumed that γ takes its values in $[0;1]$. This is satisfied when $\|Y\|_\infty \leq A = 1/2$. When $A \neq 1/2$, we apply this result to $(2A)^{-1}Y$ and recover the general result by homogeneity.

First, we recall this result in the bounded least-square regression framework. For every $t : \mathcal{X} \mapsto \mathbb{R}$ and $\epsilon > 0$, we define

$$d^2(s,t) = 2l(s,t) \qquad \text{and} \qquad w(\epsilon) = \sqrt{2}\epsilon \ .$$

Let ϕ_m belong to the class of nondecreasing and continuous functions $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $x \mapsto f(x)/x$ is nonincreasing on $(0; +\infty)$ and $f(1) \geq 1$. Assume that for every $u \in S_m$ and $\sigma > 0$ such that $\phi_m(\sigma) \leq \sqrt{n}\sigma^2$,

$$\sqrt{n}\mathbb{E}\left[\sup_{t \in S_m,\, d(u,t) \leq \sigma} |\overline{\gamma}_n(u) - \overline{\gamma}_n(t)|\right] \leq \phi_m(\sigma) \ . \tag{41}$$

Let $\varepsilon_{\star,m}$ be the unique positive solution of the equation

$$\sqrt{n}\varepsilon_{\star,m}^2 = \phi_m(w(\varepsilon_{\star,m})) \ .$$

Then, there exists some absolute constant C such that for every real number $q \geq 2$ one has

$$\|p_2(m) - \mathbb{E}[p_2(m)]\|_q \leq \frac{C}{\sqrt{n}}\left[\sqrt{2q}\left(\sqrt{l(s, s_m)} \vee \varepsilon_{\star,m}\right) + q\frac{2}{\sqrt{n}}\right] \ . \tag{42}$$

For every model S_m of histograms, of dimension D_m as a vector space, we can take

$$\phi_m(\sigma) = 3\sqrt{2}\sqrt{D_m} \times \sigma \qquad \text{in (41)}. \tag{43}$$

The proof of this statement is made below. Then, $\varepsilon_{\star,m} = 6\sqrt{D_m}n^{-1/2}$.

Combining (42) with the classical link between moments and concentration (for instance Lemma 8.9 of [Arl07]), the first result follows. The second result is obtained by taking $\theta = D_m^{-1/2}$, as in Prop. 3.

PROOF OF (43). Let $u \in S_m$ and $d(u,t) = \sqrt{2}\,\|u(X) - t(X)\|_2$ for every $t : \mathcal{X} \mapsto \mathbb{R}$. Define $\psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ by

$$\psi(\sigma) = \mathbb{E}\left[\sup_{d(u,t) \leq \sigma,\, t \in S_m} |(P_n - P)(\gamma(u,\cdot) - \gamma(t,\cdot))|\right] \ .$$

We are looking for some nondecreasing and continuous function $\phi_m : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\phi_m(x)/x$ is nonincreasing, $\phi_m(1) \geq 1$ and for every $u \in S_m$,

$$\forall \sigma > 0 \quad \text{such that} \quad \phi_m(\sigma) \leq \sqrt{n}\sigma^2 \ , \qquad \phi_m(\sigma) \geq \sqrt{n}\psi(\sigma) \ .$$

We first look at a general upperbound on ψ.

Assume that $u = s_m$. If this is not the case, the triangular inequality shows that $\psi_{\text{general } u} \leq 2\psi_{u=s_m}$. Let us write

$$t = \sum_{\lambda \in \Lambda_m} t_\lambda \mathbb{1}_{I_\lambda} \qquad u = s_m = \sum_{\lambda \in \Lambda_m} \beta_\lambda \mathbb{1}_{I_\lambda} \ .$$

Computation of $P(\gamma(t,\cdot)-\gamma(s_m,\cdot))$. for some general $t \in S_m$:

$$\begin{aligned} P(\gamma(t,\cdot)-\gamma(s_m,\cdot)) &= \mathbb{E}\left[(t(X)-Y)^2-(s_m(X)-Y)^2\right] \\ &= \mathbb{E}\left[(t(X)-s_m(X))^2\right] + 2\mathbb{E}\left[(t(X)-s_m(X))(s_m(X)-s(X))\right] \\ &= \mathbb{E}\left[(t(X)-s_m(X))^2\right] \\ &= \sum_{\lambda\in\Lambda_m} p_\lambda (t_\lambda-\beta_\lambda)^2 \end{aligned}$$

since for every $\lambda \in \Lambda_m$, $\mathbb{E}\left[s(X) \mid X \in I_\lambda\right] = \beta_\lambda$.

Computation of $P_n(\gamma(t,\cdot)-\gamma(s_m,\cdot))$. for some general $t \in S_m$: with $\eta_i = Y_i - s_m(X_i)$, we have

$$\begin{aligned} P(\gamma(t,\cdot)-\gamma(s_m,\cdot)) &= \frac{1}{n}\sum_{i=1}^n \left[(t(X_i)-Y_i)^2-(u(X_i)-Y_i)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n (t(X_i)-u(X_i))^2 - \frac{2}{n}\sum_{i=1}^n \left[(t(X_i)-u(X_i))\eta_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda-u_\lambda)^2 \mathbb{1}_{X_i\in I_\lambda} - \frac{2}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda-u_\lambda)\mathbb{1}_{X_i\in I_\lambda}\eta_i \ . \end{aligned}$$

Back to $(P_n - P)$. We sum the two inequalities above and use the triangular inequality:

$$\begin{aligned} |(P_n-P)(\gamma(t,\cdot)-\gamma(u,\cdot))| &\le \left|\frac{1}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda-u_\lambda)^2(\mathbb{1}_{X_i\in I_\lambda}-p_\lambda)\right| \\ &\quad + \left|\frac{2}{n}\sum_{i=1}^n \sum_{\lambda\in\Lambda_m} (t_\lambda-u_\lambda)\mathbb{1}_{X_i\in I_\lambda}\eta_i\right| \\ &\le \frac{2A}{n}\sum_{\lambda\in\Lambda_m}\left[(\sqrt{p_\lambda}\,|t_\lambda-u_\lambda|)\,\frac{|\sum_{i=1}^n(\mathbb{1}_{X_i\in I_\lambda}-p_\lambda)|}{\sqrt{p_\lambda}}\right] \\ &\quad + \frac{2}{n}\sum_{\lambda\in\Lambda_m}\left[(\sqrt{p_\lambda}\,|t_\lambda-u_\lambda|)\,\frac{|\sum_{i=1}^n \mathbb{1}_{X_i\in I_\lambda}\eta_i|}{\sqrt{p_\lambda}}\right] \end{aligned}$$

since $|t_\lambda - u_\lambda| \le 2A$ for every $t \in S_m$.

We now assume that $d(u,t) \le \sigma$ for some $\sigma > 0$, *i.e.*

$$d(u,t)^2 = 2\sum_{\lambda\in\Lambda_m} p_\lambda (t_\lambda-u_\lambda)^2 \le \sigma^2 \ .$$

From Cauchy-Schwarz inequality, we obtain for every $t \in S_m$ such that $d(u,t) \leq \sigma$

$$|(P_n - P)(\gamma(t,\cdot) - \gamma(u,\cdot))| \leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda}} + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \frac{\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda}}$$

Back to ψ. The upper bound above does not depend on t, so that the left-hand side of the inequality can be replaced by a supremum over $\{\, t \in S_m \text{ s.t. } d(u,t) \leq \sigma \,\}$. Taking expectations and using Jensen's inequality ($\sqrt{\cdot}$ being concave), we obtain an upper bound on ψ:

$$\psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2}n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2}{p_\lambda}\right]} + \frac{\sqrt{2}\sigma}{n} \sqrt{\sum_{\lambda \in \Lambda_m} \mathbb{E}\left[\frac{\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2}{p_\lambda}\right]} \tag{44}$$

For every $\lambda \in \Lambda_m$, we have

$$\mathbb{E}\left(\sum_{i=1}^n (\mathbb{1}_{X_i \in I_\lambda} - p_\lambda)\right)^2 = \sum_{i=1}^n \mathbb{E}\left(\mathbb{1}_{X_i \in I_\lambda} - p_\lambda\right)^2 = n p_\lambda \left(1 - p_\lambda\right) \tag{45}$$

which simplifies the first term. For the second term, notice that

$$\forall i \neq j, \qquad \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{1}_{X_j \in I_\lambda} \eta_i \eta_j\right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i\right] \mathbb{E}\left[\mathbb{1}_{X_j \in I_\lambda} \eta_j\right]$$
$$\text{and} \quad \forall i, \qquad \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i\right] = \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \mathbb{E}\left[\eta_i \mid \mathbb{1}_{X_i \in I_\lambda}\right]\right] = 0$$

since η_i is centered conditionally to $\mathbb{1}_{X_i \in I_\lambda}$. Then,

$$\mathbb{E}\left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_\lambda} \eta_i\right)^2 = \sum_{i=1}^n \mathbb{E}\left[\mathbb{1}_{X_i \in I_\lambda} \eta_i^2\right] \leq n p_\lambda \left\|\eta\right\|_\infty^2 \leq n p_\lambda (2A)^2 \ . \tag{46}$$

Combining (44) with (45) and (46), we deduce that

$$\psi(\sigma) \leq \frac{2A\sigma}{\sqrt{2}\sqrt{n}} \sqrt{D_m - 1} + \frac{2\sqrt{2}A\sigma}{\sqrt{n}} \sqrt{D_m}$$
$$\leq 3A\sqrt{2} \frac{\sqrt{D_m}}{\sqrt{n}} \times \sigma \ .$$

As already noticed, we have to multiply this bound by 2 so that it is valid for every $u \in S_m$ and not only $u = s_m$.

The resulting upper bound (multiplied by $\sqrt{n}$) has all the desired properties for ϕ_m since $6A\sqrt{2}\sqrt{D_m} = 3\sqrt{2D_m} \geq 1$. The result follows. □

REFERENCES

[Aka70] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[Aka73] Hirotugu 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.

[Arl07] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[Arl08a] Sylvain Arlot. Model selection by resampling penalization. In preparation, 2008.

[Arl08b] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization. Preprint. arXiv:0802.0566, February 2008.

[Bar00] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[Bar02] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[Bau07] Jean-Patrick Baudry. Clustering through model selection criteria. , 2007. Poster session at One Day Statistical Workshop in Lisieux. `http://www.math.u-psud.fr/~baudry`, June 2007.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BGH07] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with unknown variance. Preprint. Arxiv:math.ST/0701250, January 2007.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BM04] Stéphane Boucheron and Pascal Massart. Data-driven penalties: heuristics and results. Personal communication, February 2004.

[BM06a] Lucien Birgé and Pascal Massart. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields*, 134(3), 2006.

[BM06b] Gilles Blanchard and Pascal Massart. Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6):2664–2671, 2006.

[Bur02] Prabir Burman. Estimation of equifrequency histograms. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

[Leb05] Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

[Lep02] Vincent Lepez. *Some estimation problems related to oil reserves*. PhD thesis, University Paris XI, 2002.

[Li87] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.

[Mal73] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[Mas07] 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.

[MM07] Cathy Maugis and Bertrand Michel. A nonasymptotic penalized criterion for gaussian mixture model selection. a variable selection and clustering problems. In preparation, September 2007.

[PT90] B. T. Polyak and A. B. Tsybakov. Asymptotic optimality of the C_p-test in the projection estimation of a regression. *Teor. Veroyatnost. i Primenen.*, 35(2):305–317, 1990.

[Shi81] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[Sto85] Charles J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520, Belmont, CA, 1985. Wadsworth.

[Ver07] Nicolas Verzelen. Model selection for graphical models. In preparation, September 2007.

[Vil07] Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques.* PhD thesis, University Paris XI, December 2007.

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

Pascal Massart
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathematiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select
E-mail: pascal.massart@math.u-psud.fr