Finite sample improvement of Akaike's Information Criterion

Adrien Saumard **Fabien Navarro**
CREST, CNRS-ENSAI, Université Bretagne Loire

Abstract

We emphasize that it is possible to improve the principle of unbiased risk estimation for model selection by addressing excess risk deviations in the design of penalization procedures. Indeed, we propose a modification of Akaike's Information Criterion that avoids overfitting, even when the sample size is small. We call this correction an over-penalization procedure. As proof of concept, we show the nonasymptotic optimality of our histogram selection procedure in density estimation by establishing sharp oracle inequalities for the Kullback-Leibler divergence. One of the main features of our theoretical results is that they include the estimation of unbounded log-densities. To do so, we prove several analytical and probabilistic lemmas that are of independent interest. In an experimental study, we also demonstrate state-of-the-art performance of our over-penalization criterion for bin size selection, in particular outperforming AICc procedure.

MSC 2010 subject classifications: 62G07, 62G09, 62G10.
Key words and phrases: model selection, testing, AIC, maximum likelihood, density estimation, bin size, over-penalization, AIC corrected, small sample size.

1 Introduction

Since its introduction by Akaike in the early seventies [2], the celebrated Akaike's Information Criterion (AIC) has been an essential tool for the statistician and its use is almost systematic in problems of model selection and estimator selection for prediction. By choosing among estimators or models constructed from finite degrees of freedom, the AIC recommends more specifically to maximize the log-likelihood of the estimators penalized by their corresponding degrees of freedom. This procedure has found pathbreaking applications in density estimation, regression, time series or neural network analysis, to name a few ([32]). Because of its simplicity and negligible computation cost—whenever the estimators are given—, it is also far from outdated and continues to serve as one of the most useful devices for model selection in high-dimensional statistics. For instance, it can be used to efficiently tune the Lasso ([55]).

Any substantial and principled improvement of AIC is likely to have a significant impact on the practice of model choices and we bring in this paper an efficient and theoretically grounded solution to the problem of overfitting that can occur when using AIC on small to medium sample sizes.

The fact that AIC tends to be unstable and therefore perfectible in the case of small sample sizes is well known to practitioners and and has long been noted. Suguira [52] and Hurvich and Tsai [36] have proposed the so-called AICc (for AIC corrected), which tends to penalize more than AIC. However, the derivation of AICc comes from an asymptotic analysis where the dimension of the models are considered fixed relative to the sample size. In fact, such an assumption does not fit the usual practice of model selection, where the largest models are of dimensions close to the sample size. Another drawback related to AICc is that it has been legitimated through a mathematical analysis only in the linear regression model and for autoregressive models ([32]). However, to the best of our knowledge, outside these frameworks, there is no theoretical ground for the use of AICc.

Building on considerations from the general nonasymptotic theory of model selection developed during the nineties (see for instance [11] and [43]) and in particular on Castellan's analysis [30], Birgé and Rozenholc [22] have considered an AIC modification specifically designed for the selection of the bin size in histogram selection for density estimation. Indeed, results of [30]—and more generally results of [11]—advocate to take into account in the design of penalty the number of models to be selected. The importance of the cardinality of the collection of models for model selection is in fact a very general phenomenon and one of the main outcomes of the nonasymptotic model selection theory. In the bin size selection problem, this corresponds to adding a small amount to AIC. Unfortunately, the theory does not specify uniquely the term to be added to AIC. On the contrary, infinitely many corrections are accepted by the theory and in order to choose a good one, intensive experiments were conducted in [22]. The resulting AIC correction therefore always has the disadvantage of being specifically designed for the task on which it has been tested.

We propose a general approach that goes beyond the limits of the unbiased risk estimation principle. The latter principle is indeed at the core of Akaike's model selection procedure and is more generally the main model selection principle, which underlies procedures such as Stein's Unbiased Risk Estimator (SURE, [50]) or cross-validation ([7]). We point out that it is more efficient to estimate a quantile of the risk of the estimators—the level of the quantile depending on the size of the collection of models—than its mean. We also develop a (pseudo-)testing point of view, where we find that unbiased risk estimation does not in general allow to control the sizes of the considered tests. This is thus a new, very general model selection principle that we put forward and formalize. We call it an over-penalization procedure, because it systematically involves adding small terms to traditional penalties such as AIC.

Our procedure consists in constructing an estimator from multiple pseudo-tests, built on some random events. From this perspective, our work shares strong connections recent advances in robust estimation by designing estimators from tests ([9, 20, 21, 10, 39]). However, our focus and perspectives are significantly different from this existing line of works. Indeed, estimators such as T-estimators ([20]) or ρ-estimators ([10]) are quasi-universal estimators in the sense that they have very strong statistical guarantees, but they have the drawback to be very difficult—if feasible—to compute. In particular, [9] also builds estimators from frequency histograms, but to our knowledge no implementation of such estimators exists and it seems to be an open question whether a polynomial time algorithm can effectively compute them or not. Here, we rather keep the tractability of AIC procedure, but we don't look particularly at robustness properties (for instance against outliers). We focus on improving AIC in the nonasymptotic regime.

In addition, it is worth noting that several authors have examined some links between multiple testing and model selection, in particular by making some modifications to classical criteria (see for instance [34, Chapter 7]). But these lines of research differ significantly from our approach. Indeed, the first and main use in the literature of multiple testing point of view for model selection concerns variable selection, i.e. the identification of models, particularly in the context of linear regression. ([13, 27, 49, 56, 1, 16]). It consists in considering simultaneously the testing of each variable being equal to zero or not. Instead, we consider model selection from a predictive perspective and do not focus on model identification. It should also be noted that multiple tests may be considered after model selection or at the same time as selective inference ([47, 37, 17]), but these questions are not directly related to the scope of our paper.

Lets us now detail our contributions.

- We propose a general formulation of the model selection task for prediction in terms of a (pseudo-)test procedure (understood in a non classical way which will be detailed in the Section 2.3.2), thus establishing a link between two major topics of contemporary research. In particular, we propose a generic property that the pseudo-tests collection should satisfy in order to ensure an oracle inequality for the selected model. We call this property the "transitivity property" and show that it generalizes penalization procedures together with T-estimators and ρ-estimators.
- Considering the problem of density estimation by selecting a histogram, we prove a sharp and fully nonasymptotic oracle inequality for our procedure. Indeed, we describe a control of Kullback-Leibler (KL) divergence—also called excess risk—of the selected histogram as soon as we have one observation. We emphasize that this very strong feature may not be possible when considering AIC. We also stressed that up to our knowledge, our oracle inequality is the first

nonasymptotic result comparing the KL divergence of the selected model to the KL divergence of the oracle in an unbounded setting. Indeed, oracle inequalities in density estimation are generally expressed in terms of Hellinger distance—which is much easier to handle than the KL divergence, because it is bounded—for the selected model.

- In order to prove our oracle inequality, we improve upon the previously best known concentration inequality for the chi-square statistics (Castellan [30], Massart [43]) and this allows us to gain an order of magnitude in the control of the deviations of the excess risks of the estimators. Our result on the chi-square statistics is general and of independent interest.
- We also prove new Bernstein-type concentration inequalities for log-densities that are unbounded. Again, these probabilistic results, which are naturally linked to information theory, are general and of independent interest.
- We generalize previous results of Barron and Sheu [12] regarding the existence of margin relations in maximum likelihood estimation (MLE). Indeed, related results of [12] where established under boundedness of the log-densities and we extend them to unbounded log-densities with moment conditions.
- Finally, from a practical point of view, we bring a nonasymptotic improvement of AIC that has, in its simplest form, the same computational cost as AIC. Furthermore, we show that our over-penalization procedure largely outperforms AIC on small and medium sample sizes, but also surpasses existing AIC corrections such as AICc or Birgé-Rozenholc's procedure.

Let us end this introduction by detailing the organization of the paper.

We present our over-penalization procedure in Section 2. More precisely, we detail in Sections 2.1 and 2.2 our model selection framework related to MLE via histograms. Then in Section 2.3 we define formally over-penalization procedures and highlight their generality. We explain the ideas underlying over-penalization from three different angles: estimation of the ideal penalty, pseudo-testing and a graphical point of view.

Section 3 is devoted to statistical guarantees related to over-penalization. In particular, as concentration properties of the excess risks are at the heart of the design of an over-penalization, we detail them in Section 3.1. We then deduce a general and sharp oracle inequality in Section 3.2 and highlight the theoretical advantages compared to an AIC analysis.

New mathematical tools of a probabilistic and analytical nature and of independent interest are presented in Section 4. Section 5 contains the experiments, with detailed practical procedures. We consider two different practical variations of over-penalization and compare them with existing penalization procedures. The superiority of our method is particularly transparent.

The proofs are gathered in Section 6, as well as in the Appendix A which provides further theoretical developments that extend the description of our over-penalization procedure.

2 Statistical Framework and Notations

The over-penalization procedure, described in Section 2.3, is legitimated at a heuristic level within a generic M-estimator selection framework. We put to emphasis in Section 2.1 on maximum likelihood estimation (MLE) since, as proof of concept for our over-penalization procedure, our theoretical and experimental results will address the case of bin size selection for maximum likelihood histogram selection in density estimation. In order to be able to discuss in Section 2.3 the generality of our approach in an M-estimation setting, our presentation of MLE brings notations which extend directly to M-estimation with a general contrast.

2.1 Maximum Likelihood Density Estimation

We are given n independent observations $(\xi_1, \ldots, \xi_n)$ with unknown common distribution P on a measurable space $(\mathcal{Z}, \mathcal{T})$. We assume that there exists a known probability measure μ on $(\mathcal{Z}, \mathcal{T})$ such that P admits a density f_* with respect to μ: $f_* = dP/d\mu$. Our goal is to estimate the density f_*.

For an integrable function f on $\mathcal{Z}$, we set $Pf = P(f) = \int_{\mathcal{Z}} f(z)\, dP(z)$ and $\mu f = \mu(f) = \int_{\mathcal{Z}} f(z)\, d\mu(z)$. If $P_n = 1/n \sum_{i=1}^{n} \delta_{\xi_i}$ denotes the empirical distribution associated to the sample $(\xi_1, \ldots, \xi_n)$, then we set $P_n f = P_n(f) = 1/n \sum_{i=1}^{n} f(\xi_i)$. Moreover, taking the conventions $\ln 0 = -\infty$, $0 \ln 0 = 0$ and defining the positive part as $(x)_+ = x \vee 0$, we set

$$\mathcal{S} = \left\{ f : \mathcal{Z} \longrightarrow \mathbb{R}_+ ; \ \int_{\mathcal{Z}} f d\mu = 1 \text{ and } P(\ln f)_+ < \infty \right\} .$$

We assume that the unknown density f_* belongs to $\mathcal{S}$.

Note that since $P(\ln f_*)_- = \int f_* \ln f_* \mathbb{1}_{f_* \leq 1} d\mu < \infty$, the fact that f_* belongs to $\mathcal{S}$ is equivalent to $\ln(f_*) \in L_1(P)$, the space of integrable functions on $\mathcal{Z}$ with respect to P.

We consider the MLE of the density f_*. To do so, we define the maximum likelihood contrast γ to be the following functional,

$$\gamma : f \in \mathcal{S} \longmapsto (z \in \mathcal{Z} \longmapsto -\ln(f(z))) \ .$$

Then the risk $P\gamma(f)$ associated to the contrast γ on a function $f \in \mathcal{S}$ is the following,

$$P\gamma(f) = P(\ln f)_- - P(\ln f)_+ \in \mathbb{R} \cup \{+\infty\} \ .$$

Also, the excess risk of a function f with respect to the density f_* is classically given in this context by the KL divergence of f with respect to f_*. Recall that for two probability distributions P_f and P_g on $(\mathcal{Z}, \mathcal{T})$ of respective densities f and g with respect to μ, the KL divergence of P_g with respect to P_f is defined to be

$$\mathcal{K}(P_f, P_g) = \begin{cases} \int_{\mathcal{Z}} \ln\left(\frac{dP_f}{dP_g}\right) dP_g = \int_{\mathcal{Z}} f \ln\left(\frac{f}{g}\right) d\mu & \text{if} \quad P_f \ll P_g \\ \infty & \text{otherwise.} \end{cases}$$

By a slight abuse of notation we denote $\mathcal{K}(f, t)$ rather than $\mathcal{K}(P_f, P_g)$ and by the Jensen inequality we notice that $\mathcal{K}(f, g)$ is a nonnegative quantity, equal to zero if and only if $f = g \quad \mu$-*a.s.* Hence, for any $f \in \mathcal{S}$, the excess risk of a function f with respect to the density f_* satisfies

$$P(\gamma(f)) - P(\gamma(f_*)) = \int_{\mathcal{Z}} \ln\left(\frac{f_*}{f}\right) f_* d\mu = \mathcal{K}(f_*, f) \geq 0 \tag{1}$$

and this nonnegative quantity is equal to zero if and only if $f_* = f \quad \mu - a.s.$ Consequently, the unknown density f_* is uniquely defined by

$$\begin{aligned} f_* &= \arg\min_{f \in \mathcal{S}} \{P(-\ln f)\} \\ &= \arg\min_{f \in \mathcal{S}} \{P\gamma(f)\} \ . \end{aligned}$$

For a model m, that is a subset $m \subset \mathcal{S}$, we define the maximum likelihood estimator on m, whenever it exists, by

$$\begin{aligned} \hat{f}_m &\in \arg\min_{f \in m} \{P_n \gamma(f)\} \qquad (2) \\ &= \arg\min_{f \in m} \left\{ \frac{1}{n} \sum_{i=1}^{n} -\ln(f(\xi_i)) \right\} \ . \end{aligned}$$

2.2 Histogram Models

The models m that we consider here to define the maximum likelihood estimators as in (2) are made of histograms defined on a fixed partition of $\mathcal{Z}$. More precisely, for a finite partition Λ_m of $\mathcal{Z}$ of cardinality $|\Lambda_m| = D_m + 1$, $D_m \in \mathbb{N}$, we set

$$m = \left\{ f = \sum_{I \in \Lambda_m} \beta_I \mathbb{1}_I \ ; \ (\beta_I)_{I \in \Lambda_m} \in \mathbb{R}_+^{D_m + 1}, \ f \geq 0 \text{ and } \sum_{I \in \Lambda_m} \beta_I \mu(I) = 1 \right\} \ .$$

Note that the smallest affine space contained in m is of dimension D_m. The quantity D_m can thus be interpreted as the number of degrees of freedom in the (parametric) model m. We assume that any element I of the partition Λ_m is of positive measure with respect to μ: for all $I \in \Lambda_m, \quad \mu(I) > 0$. As the partition Λ_m is finite, we have $P(\ln f)_+ < \infty$ for all $f \in m$ and so $m \subset \mathcal{S}$. We state in the next proposition some well-known properties that are satisfied by histogram models submitted to the procedure of MLE (see for example [43, Section 7.3]).

Proposition 2.1 *Let*

$$f_m = \sum_{I \in \Lambda_m} \frac{P(I)}{\mu(I)} \mathbf{1}_I \ .$$

Then $f_m \in m$ and f_m is called the KL projection of f_* onto m. Moreover, it holds

$$f_m = \arg\min_{f \in m} P(\gamma(f)) \ .$$

The following Pythagorean-like identity for the KL divergence holds, for every $f \in m$,

$$\mathcal{K}(f_*, f) = \mathcal{K}(f_*, f_m) + \mathcal{K}(f_m, f) \ . \tag{3}$$

The maximum likelihood estimator on m is well-defined and corresponds to the so-called frequencies histogram associated to the partition Λ_m. We also have the following formulas,

$$\hat{f}_m = \sum_{I \in \Lambda_m} \frac{P_n(I)}{\mu(I)} \mathbb{1}_I \ \text{and} \ P_n(\gamma(f_m) - \gamma(\hat{f}_m)) = \mathcal{K}(\hat{f}_m, f_m) \ .$$

Remark 2.2 *Histogram models are special cases of general exponential families exposed for example in Barron and Sheu [12] (see also Castellan [30] for the case of exponential models of piecewise polynomials). The projection property (3) can be generalized to exponential models (see [12, Lemma 3] and Csiszár [33]).*

Remark 2.3 *As by (1) we have*

$$P(\gamma(f_m) - \gamma(f_*)) = \mathcal{K}(f_*, f_m)$$

and for any $f \in m$,

$$P(\gamma(f) - \gamma(f_*)) = \mathcal{K}(f_*, f)$$

we easily deduce from (3) that the excess risk on m is still a KL divergence, as we then have for any $f \in m$,

$$P(\gamma(f) - \gamma(f_m)) = \mathcal{K}(f_m, f) \ .$$

2.3 Over-Penalization

Now let's define our model selection procedure. We propose three ways to understand the benefits of over-penalization. Of course, the three points of view are interrelated, but they provide different and complementary insights on the behavior of over-penalization.

2.3.1 Over-Penalization as Estimation of the Ideal Penalty

We are given a collection of histogram models denoted $\mathcal{M}_n$, with finite cardinality depending on the sample size n, and its associated collection of maximum likelihood estimators $\left\{\hat{f}_m; m \in \mathcal{M}_n\right\}$. By taking a (nonnegative) penalty function pen on $\mathcal{M}_n$,

$$\text{pen} : m \in \mathcal{M}_n \longmapsto \text{pen}(m) \in \mathbb{R}^+ \ ,$$

the output of the penalization procedure (also called the selected model) is by definition any model satisfying,

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n(\gamma(\hat{f}_m)) + \mathrm{pen}\,(m) \right\} \ . \tag{4}$$

We aim at selecting an estimator $\hat{f}_{\widehat{m}}$ with a KL divergence, pointed on the true density f_*, as small as possible. Hence, we want our selected model to have a performance as close as possible to the excess risk achieved by an oracle model (possibly non-unique), defined to be,

$$m_* \in \arg\min_{m\in\mathcal{M}_n} \left\{ \mathcal{K}(f_*, \hat{f}_m) \right\} \tag{5}$$

$$= \arg\min_{m\in\mathcal{M}_n} \left\{ P(\gamma(\hat{f}_m)) \right\} \ . \tag{6}$$

From (6), it is seen that an ideal penalty in the optimization task (4) is given by

$$\mathrm{pen}_{\mathrm{id}}\,(m) = P(\gamma(\hat{f}_m)) - P_n(\gamma(\hat{f}_m)) \ ,$$

since in this case, the criterion $\mathrm{crit}_{\mathrm{id}}\,(m) = P_n(\gamma(\hat{f}_m)) + \mathrm{pen}_{\mathrm{id}}\,(m)$ is equal to the true risk $P(\gamma(\hat{f}_m))$. However $\mathrm{pen}_{\mathrm{id}}$ is unknown and, at some point, we need to give some estimate of it. In addition, $\mathrm{pen}_{\mathrm{id}}$ is random, but we may not be able to provide a penalty, even random, whose fluctuations at a fixed model m would be positively correlated to the fluctuations of $\mathrm{pen}_{\mathrm{id}}\,(m)$. This means that we are rather searching for an estimate of a *deterministic functional* of $\mathrm{pen}_{\mathrm{id}}$. But which functional would be convenient? The answer to this question is essentially contained in the solution of the following problem.

Problem 1. *For any fixed* $\beta \in (0, 1)$ *find the* deterministic penalty $\mathrm{pen}_{\mathrm{id},\beta} : \mathcal{M}_n \to R_+$, *that minimizes the value of* C, *among constants* $C > 0$ *which satisfy the following oracle inequality,*

$$\mathbb{P}\left(\mathcal{K}(f_*, \hat{f}_{\widehat{m}}) \le C \inf_{m\in\mathcal{M}_n} \left\{ \mathcal{K}(f_*, \hat{f}_m) \right\} \right) \ge 1 - \beta \ . \tag{7}$$

The solution—or even the existence of a solution—to the problem given in (7) is not easily accessible and depends on assumptions on the law P of data and on approximation properties of the models, among other things. In the following, we give a reasonable candidate for $\mathrm{pen}_{\mathrm{id},\beta}$. Indeed, let us set $\beta_{\mathcal{M}} = \beta/\mathrm{Card}(\mathcal{M}_n)$ and define

$$\mathrm{pen}_{\mathrm{opt},\beta}\,(m) = q_{1-\beta_{\mathcal{M}}} \left\{ P(\gamma(\hat{f}_m) - \gamma(f_m)) + P_n(\gamma(f_m) - \gamma(\hat{f}_m)) \right\} \ , \tag{8}$$

where $q_\lambda \{Z\} = \inf \{q \in \mathbb{R}; \mathbb{P}(Z \le q) \ge \lambda\}$ is the quantile of level λ for the real random variable Z. Our claim is that $\mathrm{pen}_{\mathrm{opt},\beta}$ gives in (7) a constant C which is close to one, under some general assumptions (see Section 3 for precise results). Let us explain now why $\mathrm{pen}_{\mathrm{opt},\beta}$ should lead to a nearly optimal model selection.

We set

$$\Omega_0 = \bigcap_{m\in\mathcal{M}_n} \left\{ P(\gamma(\hat{f}_m) - \gamma(f_m)) + P_n(\gamma(f_m) - \gamma(\hat{f}_m)) \le \mathrm{pen}_{\mathrm{opt},\beta}\,(m) \right\} \ .$$

We see, by definition of $\mathrm{pen}_{\mathrm{opt},\beta}$ and by a simple union bound over the models $m \in \mathcal{M}_n$, that the event Ω_0 is of probability at least $1 - \beta$. Now, by definition of $\widehat{m}$, we have, for any $m \in \mathcal{M}_n$,

$$P_n(\gamma(\hat{f}_{\widehat{m}})) + \mathrm{pen}_{\mathrm{opt},\beta}(\widehat{m}) \le P_n(\gamma(\hat{f}_m)) + \mathrm{pen}_{\mathrm{opt},\beta}\,(m) \ . \tag{9}$$

By centering by $P(\gamma(f_*))$ and using simple algebra, Inequality (9) can be written as,

$$\begin{aligned} & P(\gamma(\hat{f}_{\widehat{m}}) - \gamma(f_*)) + \left[\mathrm{pen}_{\mathrm{opt},\beta}(\widehat{m}) - (P(\gamma(\hat{f}_{\widehat{m}}) - \gamma(f_{\widehat{m}})) + P_n(\gamma(f_{\widehat{m}}) - \gamma(\hat{f}_{\widehat{m}}))) \right] \\ \le & P(\gamma(\hat{f}_m) - \gamma(f_*)) + \left[\mathrm{pen}_{\mathrm{opt},\beta}\,(m) - (P(\gamma(\hat{f}_m) - \gamma(f_m)) + P_n(\gamma(f_m) - \gamma(\hat{f}_m))) \right] \\ & + (P_n - P)(\gamma(f_m) - \gamma(f_{\widehat{m}})) \ . \end{aligned}$$

Now, on Ω_0, we have $\mathrm{pen}_{\mathrm{opt},\beta}(\widehat{m}) - (P(\gamma(\hat{f}_{\widehat{m}}) - \gamma(f_{\widehat{m}})) + P_n(\gamma(f_{\widehat{m}}) - \gamma(\hat{f}_{\widehat{m}}))) \geq 0$, so we get on Ω_0,

$$\begin{aligned} & P(\gamma(\hat{f}_{\widehat{m}}) - \gamma(f_*)) \\ \leq \; & P(\gamma(\hat{f}_m) - \gamma(f_*)) + \left[\mathrm{pen}_{\mathrm{opt},\beta}(m) - (P(\gamma(\hat{f}_m) - \gamma(f_m)) + P_n(\gamma(f_m) - \gamma(\hat{f}_m)))\right] \\ & + (P_n - P)(\gamma(f_m) - \gamma(f_{\widehat{m}})) \ . \end{aligned}$$

Specifying to the MLE context, the latter inequality writes,

$$\begin{aligned} & \mathcal{K}(f_*, \hat{f}_{\widehat{m}}) \\ & \leq \mathcal{K}(f_*, \hat{f}_m) + \underbrace{\left[\mathrm{pen}_{\mathrm{opt},\beta}(m) - (\mathcal{K}(f_m, \hat{f}_m) + \mathcal{K}(\hat{f}_m, f_m))\right]}_{(a)} + \underbrace{(P_n - P)(\gamma(f_m) - \gamma(f_{\widehat{m}}))}_{(b)} . \end{aligned}$$

In order to get an oracle inequality as in (7), it remains to control (a) and (b) in terms of the excess risks $\mathcal{K}(f_*, \hat{f}_m)$ and $\mathcal{K}(f_*, \hat{f}_{\widehat{m}})$. Quantity (a) is related to deviations bounds for the true and empirical excess risks of the M-estimators $\hat{f}_m$ and quantity (b) is related to fluctuations of empirical bias around the bias of the models. Suitable controls of these quantities (as achieved in our proofs) will give sharp oracle inequalities (see Section 3 below).

Remark 2.4 *Notice that our reasoning is not based on the particular value of the contrast, so that to emphasize this point we choose to keep γ in most of our calculations rather than to specify to the KL divergence related to the MLE case. As a matter of fact, the penalty $\mathrm{pen}_{\mathrm{opt},\beta}$ given in (8) is a good candidate in the general context of M-estimation.*

We define an over-penalization procedure as follows.

Definition 2.5 *A penalization procedure as defined in (4) is said to be an over-penalization procedure if the penalty pen that is used satisfies $\mathrm{pen}(m) \geq \mathrm{pen}_{\mathrm{opt},\beta}(m)$ for all $m \in \mathcal{M}_n$ and for some $\beta \in (1/2, 1)$.*

Based on concentration inequalities for the excess risks (see Section 3.1) we propose the following over-penalization penalty for histogram selection,

$$\mathrm{pen}_+(m) = \left(1 + C \max\left\{\sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}; \frac{\ln(n+1)}{D_m}\right\}\right) \frac{D_m}{n} \ , \tag{10}$$

where C is a constant that should be either fixed a priori ($C = 1$ or 2 are typical choices) or estimated using data (see Section 5 for details about the choice of C). The logarithmic terms appearing in (10) are linked to the cardinal of the collection of models, since in our proofs we take a constant α such that $\ln \mathrm{Card}(\mathcal{M}_\mathrm{n}) + 2\ln(\mathrm{n}+1) \leq \alpha \ln(\mathrm{n}+1)$. The constant α then enters in the constant C of (10). We show below nonasymptotic accuracy of such procedure, both theoretically and practically.

2.3.2 Over-Penalization through a pseudo-testing approach

Let us now present a multiple test point of view on the model selection problem. The goal is to infer the oracle model (5) so that an oracle inequality of the type of (7) is ensured.

This task can be formulated by solving iterative pseudo-tests. Indeed, set the following collection of null and alternative hypotheses indexed by pairs of models: for $(m, m') \in \mathcal{M}_n^2$,

$$\begin{cases} H_0(m, m') : \mathcal{K}(f_*, \hat{f}_m) > C_n \cdot \mathcal{K}(f_*, \hat{f}_{m'}) \\ H_1(m, m') : \mathcal{K}(f_*, \hat{f}_m) \leq C_n \cdot \mathcal{K}(f_*, \hat{f}_{m'}) \end{cases} ,$$

where the constant $C_n > 1$ is uniform in $m, m' \in \mathcal{M}_n$. To each pair $(m, m') \in \mathcal{M}_n^2$, let us assume that we are given a test $T(m, m')$ that is equal to one if $H_0(m, m')$ is rejected and zero otherwise.

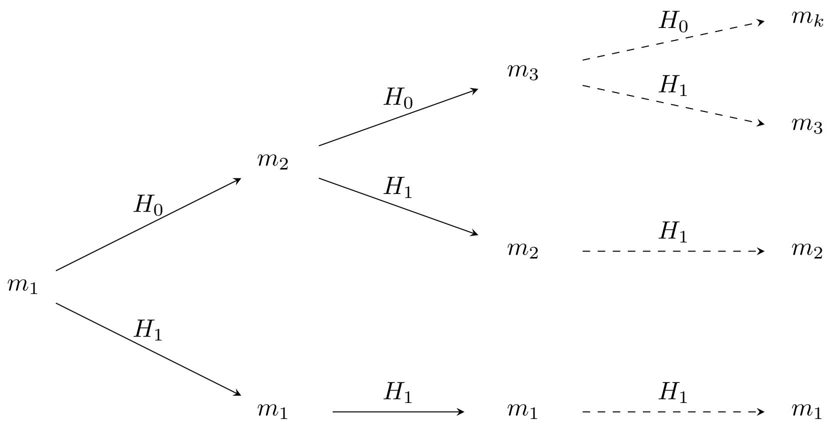


Figure 1: Iteration of the pseudo-tests along the collection of models.

It should be noted at this stage that what we have just called "pseudo-test" does not enter directly into the classical theory of statistical tests, since the null and alternative hypotheses that we consider are random events. However, as we will see, the only notion related to our pseudo-tests needed for our model selection study is the notion of the "size" of a pseudo-test, which we will give in the following and which will provide a mathematically consistent analysis of the statistical situation. Moreover,it seems that random assumptions naturally arise in some statistical frameworks. For instance, in the context of variable selection along along the Lasso path, Lockhart *et al.*[41] consider sequential random null hypotheses based on the active sets of the variable included up to a step on the Lasso path (see especially [41, Section 2.6] as well as the discussion of Bühlmann *et al.* [19, Section 2] and the rejoinder [40, Section 2]).

Finally, if the testing of random hypotheses disturbs the reader, we suggest taking our multiple pseudo-testing interpretation of the model selection task in its minimal sense, that is, as a mathematical description aiming at investigating tight conditions on the penalty, that would allow for near-optimal oracle guarantees for the penalization scheme (4).

In order to ensure an oracle inequality such as in (7), we want to avoid as far as possible selecting a model whose excess risk is far greater than the one of the oracle. In terms of the preceding tests, we will see that this exactly corresponds to controlling the "size" of the pseudo-tests $T(m, m')$.

Let us note $\mathcal{R}(m, m') = \{T(m, m') = 1\}$ the event where the pseudo-test $T(m, m')$ rejects the null hypothesis $H_0(m, m')$ and $\mathcal{T}(m, m') = \left\{\mathcal{K}(f_*, \hat{f}_m) > C_n \cdot \mathcal{K}(f_*, \hat{f}_{m'})\right\}$ the event where the hypothesis $H_0(m, m')$ is true. By extension with the classical theory of statistical testing, we denote $\alpha(m, m')$ the size of the pseudo-test $T(m, m')$, given by

$$\alpha(m, m') = \mathbb{P}\left(\mathcal{R}(m, m') \,|\mathcal{T}(m, m')\right) \ .$$

To explain how we select a model $\widehat{m}$ that is close to the oracle model m_*, let us enumerate the models of the collection: $\mathcal{M}_n = \{m_1, ..., m_k\}$. Note that we do not assume that the models are nested. Now, we test $T(m_1, m_j)$ for j increasing from 2 to k. If there exists $i_1 \in \{2, .., k\}$ such that $T(m_1, m_{i_1}) = 0$, then we perform the pseudo-tests $T(m_{i_1}, m_j)$ for j increasing from $i_1 + 1$ to k or choose m_k as our oracle candidate if $i_1 = k$. Otherwise, we choose m_1 as our oracle candidate. In general, we can thus define a finite increasing sequence m_{i_l} of models and a selected model $\widehat{m} = \widehat{m}(T)$ through the use of the collection of pseudo-tests $\{T(m, m'); m, m' \in \mathcal{M}_n\}$. Also, the number of pseudo-tests that are needed to define $\widehat{m}$ is equal to $\mathrm{Card}(\mathcal{M}_n) - 1$.

Let us denote $\mathcal{P}$ the set of pairs of models that have effectively been tested along the iterative procedure. We thus have $\mathrm{Card}(\mathcal{P}) = \mathrm{Card}(\mathcal{M}_n) - 1$. Now define the event Ω_T under which there is a first kind error along the collection of pseudo-tests in $\mathcal{P}$,

$$\Omega_T = \bigcup_{(m,m') \in \mathcal{P}} \left[\mathcal{R}(m, m') \bigcap \mathcal{T}(m, m')\right]$$

and assume that the sizes of the pseudo-tests are chosen such that

$$\beta = \sum_{(m,m')\in\mathcal{M}_n^2} \alpha\left(m, m'\right) < 1 \; . \tag{11}$$

Assume also that the selected model $\widehat{m}$ has the following Transitivity Property (**TP**):

(TP) If for any $(m,m') \in \mathcal{M}_n^2$, $T(m,m') = 1$ then $T(\widehat{m},m') = 1$.

We believe that the transitivity property (**TP**) is intuitive and legitimate as it amounts to assuming that there is no contradiction in choosing $\widehat{m}$ as a candidate oracle model. Indeed, if a model m is thought to be better than a model m', that is $T(m,m') = 1$, then the selected model $\widehat{m}$ should also be thought to be better that m', $T(\widehat{m},m') = 1$.

The power of the formalism we have just introduced lies in the fact that the combination of Assumptions (11) and (**TP**) automatically implies that an oracle inequality as in (7) is satisfied. Indeed, property (**TP**) ensures that $\mathcal{T}(\widehat{m},m_*) \subset \mathcal{R}(\widehat{m},m_*)$ because on the event $\mathcal{T}(\widehat{m},m_*)$, we have $\widehat{m} \neq m_*$, so there exists m such that $(m,m_*) \in \mathcal{P}$ and $T(m,m_*) = 1$ (otherwise $\widehat{m} = m_*$) which in turn ensures $T(\widehat{m},m_*) = 1$. Consequently, $\mathcal{T}(\widehat{m},m_*) = \mathcal{T}(\widehat{m},m_*) \cap \mathcal{R}(\widehat{m},m_*) \subset \Omega_T$, which gives

$$\mathbb{P}\left(\mathcal{K}(f_*,\hat{f}_{\widehat{m}}) > C_n \cdot \mathcal{K}(f_*,\hat{f}_{m_*})\right) = \mathbb{P}\left(\mathcal{T}(\widehat{m},m_*)\right) \leq \mathbb{P}\left(\Omega_T\right) \leq \beta \; , \tag{12}$$

that is equivalent to (7).

Let us turn now to a specific choice of pseudo-tests corresponding to model selection by penalization. If we define for a penalty pen, the penalized criterion $\mathrm{crit}_{\mathrm{pen}}(m) = P_n(\gamma(\hat{f}_m)) + \mathrm{pen}(m)$, $m \in \mathcal{M}_n$ and take the following pseudo-tests

$$T\left(m,m'\right) = T_{\mathrm{pen}}\left(m,m'\right) = \mathbb{1}_{\{\mathrm{crit}_{\mathrm{pen}}(m) \leq \mathrm{crit}_{\mathrm{pen}}(m')\}} \; ,$$

then it holds

$$\widehat{m}(T_{\mathrm{pen}}) \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n(\gamma(\hat{f}_m)) + \mathrm{pen}(m)\right\}$$

and property (**TP**) is by consequence satisfied.

It remains to choose the penalty pen such that the sizes of the pseudo-tests $T_{\mathrm{pen}}(m,m')$ are controlled. This is achieved by taking $\mathrm{pen} = \mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}$ as defined in (8), with $\tilde{\beta} = \beta/\mathrm{Card}(\mathcal{M}_n)$. Indeed, in this case,

$$\begin{aligned}
\alpha\left(m,m'\right) =& \mathbb{P}(P_n(\gamma(\hat{f}_m)) + \mathrm{pen}(m) \leq P_n(\gamma(\hat{f}_{m'})) + \mathrm{pen}\left(m'\right) \left|\mathcal{T}\left(m,m'\right)\right.) \\
\approx& \mathbb{P}(\mathcal{K}(f_*,\hat{f}_m) + \left[\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}(m) - (\mathcal{K}(f_m,\hat{f}_m) + \mathcal{K}(\hat{f}_m,f_m))\right] \\
& \leq \mathcal{K}(f_*,\hat{f}_{m'}) + \left[\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}\left(m'\right) - (\mathcal{K}(f_{m'},\hat{f}_{m'}) + \mathcal{K}(\hat{f}_{m'},f_{m'}))\right] \left|\mathcal{T}\left(m,m'\right)\right.) \\
\leq& \frac{\beta}{2\mathrm{Card}\left(\mathcal{M}_n^2\right)} + \mathbb{P}\left(\mathcal{K}(f_*,\hat{f}_m) \leq \mathcal{K}(f_*,\hat{f}_{m'})\right. \\
& \qquad + \left.\left[\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}\left(m'\right) - (\mathcal{K}(f_{m'},\hat{f}_{m'}) + \mathcal{K}(\hat{f}_{m'},f_{m'}))\right] \left|\mathcal{T}\left(m,m'\right)\right.\right) \; .
\end{aligned} \tag{13}$$

In line (13), the equality is only approximated since we neglected the centering of model biases by their empirical counterparts, as these centered random variables should be small compared to the other quantities for models of interest. Now assume that

$$\mathbb{P}\left(\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}\left(m'\right) - (\mathcal{K}(f_{m'},\hat{f}_{m'}) + \mathcal{K}(\hat{f}_{m'},f_{m'})) \geq \varepsilon_n \mathcal{K}(f_*,\hat{f}_{m'})\right) \leq \frac{\beta}{2\mathrm{Card}\left(\mathcal{M}_n^2\right)} \; ,$$

for some deterministic sequence ε_n not depending on m'. Such result is obtained in Section 3.1 and is directly related to the concentration behavior of the true and empirical excess risks. Then we get

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{K}(f_*,\hat{f}_m) \leq \mathcal{K}(f_*,\hat{f}_{m'}) + \left[\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}(m') - (\mathcal{K}(f_{m'},\hat{f}_{m'}) + \mathcal{K}(\hat{f}_{m'},f_{m'}))\right] \middle| \mathcal{T}(m,m')\right) \\
\leq& \frac{\beta}{2\mathrm{Card}\left(\mathcal{M}_n^2\right)} + \mathbb{P}\left(\mathcal{K}(f_*,\hat{f}_m) \leq (1+\varepsilon_n)\mathcal{K}(f_*,\hat{f}_{m'}) \middle| \mathcal{K}(f_*,\hat{f}_m) > C_n\mathcal{K}(f_*,\hat{f}_{m'})\right) \\
=& \frac{\beta}{2\mathrm{Card}\left(\mathcal{M}_n^2\right)} \ ,
\end{aligned}$$

where the last equality is valid if $C_n \geq 1+\varepsilon_n$. In this case,

$$\alpha\left(m,m'\right) \leq \frac{\beta}{\mathrm{Card}\left(\mathcal{M}_n^2\right)}$$

and inequality (11) is satisfied and so is the oracle inequality (12).

Remark 2.6 *There is a gap between the penalty* $\mathrm{pen}_{\mathrm{opt},\beta}$ *considered in Section 2.3.1 to ensure an oracle inequality and the penalty* $\mathrm{pen}_{\mathrm{opt},\tilde{\beta}/2}$ *defined in this section. This comes from the fact that using the pseudo-testing framework, we aim at controlling the probability of the event* Ω_T *under which there is a first kind error along the pseudo-tests performed in* $\mathcal{P}$*. Despite the fact that the set* $\mathcal{P}$ *consists in* $\mathrm{Card}(\mathcal{M}_\mathrm{n})-1$ *pseudo-tests, we give a bound that takes into account the* $\mathrm{Card}(\mathcal{M}_\mathrm{n}^2)$ *pseudo-tests defined from the pairs of models. There is a possible loss here that consists in inflating the set* $\mathcal{P}$ *in order to make the union bound valid. However, such loss would only affect the constant* C *in our over-penalization procedure (10) by a factor* 2*, since the modification of the order of the quantile affects the penalty through a logarithmic factor.*

Remark 2.7 *The Transitivity Property (**TP**) allows to unify most the selection rules. Indeed, as soon as we want to select an estimator* $\tilde{f}$ *(or a model) that optimizes a criterion,*

$$\tilde{f} \in \arg\min_{f\in\mathcal{F}}\left\{\mathrm{crit}\left(f\right)\right\} \ ,$$

where $\mathcal{F}$ *is a collection of candidate functions (or models), then* $\tilde{f}$ *is also defined by the collection of tests* $T\left(f,f'\right) = \mathbf{1}_{\{\mathrm{crit}(f)\leq\mathrm{crit}(f')\}}$*. In particular, in T-estimation ([20]) as well as in* ρ*-estimation [10]) the estimator is indeed a minimizer of a criterion that is interpreted as a diameter of a subset of functions in* $\mathcal{F}$*. Note that in T-estimation, the criterion is itself constructed through the use of some (robust) tests, that by consequence do not act at the same level as our tests in this case.*

2.3.3 Graphical insights on over-penalization

Finally, let us provide a graphic perspective on our over-penalization procedure.

If the penalty pen is chosen accordingly to the unbiased risk estimation principle, then it should satisfy, for any model $m \in \mathcal{M}_n$,

$$\mathbb{E}\left[P_n(\gamma(\hat{f}_m)) + \mathrm{pen}\left(m\right)\right] \sim \mathbb{E}\left[P\left(\gamma(\hat{f}_m)\right)\right] \ .$$

In other words, the curve $\mathcal{C}_n : m \mapsto P_n(\gamma(\hat{f}_m)) + \mathrm{pen}\left(m\right)$ fluctuates around its mean, which is essentially the curve $\mathcal{C}_P : m \mapsto \mathbb{E}\left[P\left(\gamma(\hat{f}_m)\right)\right]$, see Figure 2. Furthermore, the largest is the model m, the largest are the fluctuations of $P_n(\gamma(\hat{f}_m)) = \mathcal{K}\left(\hat{f}_m, f_m\right) + P_n(\gamma(f_m))$. This is seen for instance through the concentration inequalities established in Section A.1 below for the empirical excess risk $\mathcal{K}\left(\hat{f}_m, f_m\right)$. Consequently, it can happen that the curve $\mathcal{C}_n$ is rather flat for the largest models and that the selected model is among the largest of the collection, see Figure 2.

By using an over-penalization procedure instead of the unbiased risk estimation principle, we will compensate the deviations for the largest models and thus obtain a thinner region of potential selected models, see Figures 3 and 4. In other words, we will avoid overfitting and by doing so, we will ensure a reasonable performance of our over-penalization procedure in situations where unbiased risk estimation fails. This is particularly the case when the amount of data is small to moderate.

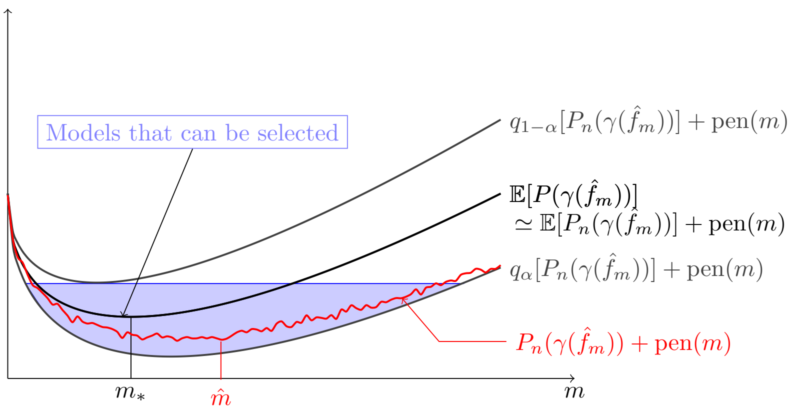


Figure 2: A schematic view of the situation corresponding to a selection procedure based on the unbiased risk principle. The penalized empirical risk (in red) fluctuates around the expectation of the true risk. The size of the deviations typically increase with the model size, making the shape of the curves possibly flat for the largest models of the collection. Consequently, the chosen model can potentially be very large and lead to overfitting.

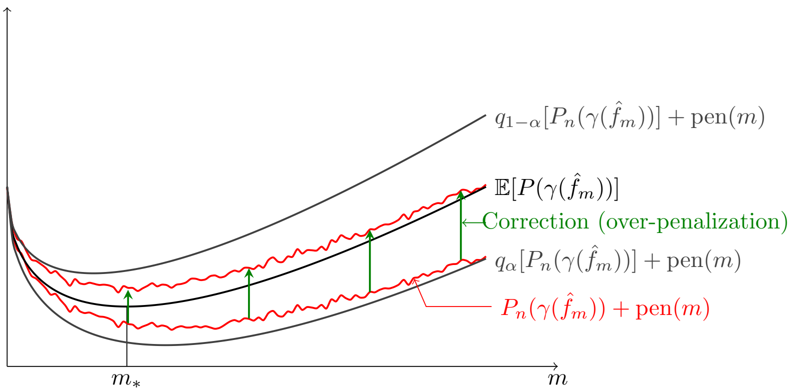


Figure 3: The correction that should be applied to an unbiased risk estimation procedure would ideally be of the size of the deviations of the risk for each model of the collection.

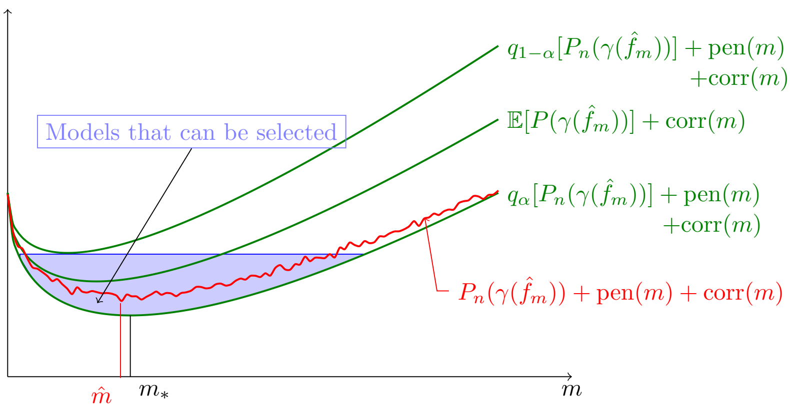


Figure 4: After a suitable correction, the minimum of the red curve has a better shape. In addition, the region of models that can be possibly selected is substantially smaller and in particular avoids the largest models of the collection.

3 Theoretical Guarantees

We state here our theoretical results related to the behavior of our over-penalization procedure.

As explained in Section 2.3, concentration inequalities for true and empirical excess risks are essential tools for understanding our model selection problem. As the theme of concentration inequalities for the excess risk constitutes a very recent and exciting area of research, these inequalities also have an interest in themselves and we state them out in Section 3.1.

In Section 3.2, we give a sharp oracle inequality proving the optimality of our procedure. We also compare our result to what would be obtained for AIC, which suggests the superiority of over-penalization in the nonasymptotic regime, i.e. for a small to medium sample size.

3.1 True and empirical excess risks' concentration

In this section, we fix the linear model m made of histograms and we are interested by concentration inequalities for the true excess risk $P\left(\gamma(\hat{f}_m) - \gamma(f_m)\right)$ on m and for its empirical counterpart $P_n\left(\gamma(f_m) - \gamma(\hat{f}_m)\right)$.

Theorem 3.1 *Let $n \in \mathbb{N}$, $n \geq 1$ and let α, A_+, A_- and A_Λ be positive constants. Take m a model of histograms defined on a fixed partition Λ_m of $\mathcal{Z}$. The cardinality of Λ_m is denoted by D_m. Assume that $1 < D_m \leq A_+ n/(\ln(n+1)) \leq n$ and*

$$0 < A_\Lambda \leq D_m \inf_{I \in \Lambda_m} \{P(I)\} \ . \tag{14}$$

If $(\alpha+1) A_+/A_\Lambda \leq \tau = \sqrt{3} - \sqrt{2} < 0.32$, then a positive constant A_0 exists, only depending on α, A_+ and A_Λ, such that by setting

$$\varepsilon_n^+(m) = \max\left\{\sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}; \frac{\ln(n+1)}{D_m}\right\} \tag{15}$$

and

$$\varepsilon_n^-(m) = \max\left\{\sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}\right\} \ ,$$

we have, on an event of probability at least $1 - 4(n+1)^{-\alpha}$,

$$\left(1 - A_0 \varepsilon_n^-(m)\right) \frac{D_m}{2n} \leq \mathcal{K}\left(f_m, \hat{f}_m\right) \leq \left(1 + A_0 \varepsilon_n^+(m)\right) \frac{D_m}{2n} \ , \tag{16}$$

$$\left(1 - A_0 \varepsilon_n^-(m)\right) \frac{D_m}{2n} \leq \mathcal{K}\left(\hat{f}_m, f_m\right) \leq \left(1 + A_0 \varepsilon_n^+(m)\right) \frac{D_m}{2n} \ . \tag{17}$$

In the previous theorem, we obtain sharp upper and lower bounds for true and empirical excess risk on m. They are optimal at the first order since the leading constants are equal in the upper and lower bounds. They show the concentration of the true and empirical excess risks around the value $D_m/(2n)$. Moreover, Theorem 3.1 establishes equivalence with high probability of the true and empirical excess risks for models of reasonable dimension.

Concentration inequalities for the excess risks as in Theorem3.1 is a new and exciting direction of research related to the theory of statistical learning and to high-dimensional statistics. Boucheron and Massart [25] obtained a pioneering result describing the concentration of the empirical excess risk around its mean, a property that they call a high-dimensional Wilks phenomenon. Then a few authors obtained results describing the concentration of the true excess risk around its mean [48], [31], [46] or around its median [14], [15] for (penalized) least square regression and in an abstract M-estimation framework [54]. In particular, recent results of [54] include the case of MLE on exponential models and as a matter of fact, on histograms. Nevertheless, Theorem 3.1 is a valuable addition to the literature on this line of research since we obtain here nonetheless the concentration around a fixed point, but an explicit value $D_m/2n$ for this point. On the contrary, the concentration point is available in [54] only through an implicit formula involving local suprema of the underlying empirical process.

The principal assumption in Theorem 3.1 is inequality (14) of lower regularity of the partition with respect to P. It is ensured as soon as the density f_* is uniformly bounded from below and the partition is lower regular with respect to the reference measure μ (which will be the Lebesgue measure in our experiments). No restriction on the largest values of f_* are needed. In particular, we do not restrict to the bounded density estimation setting.

Castellan [29] proved related, but weaker inequalities than in Theorem 3.1 above. She also asked for a lower regularity property of the partition, as in Proposition 2.5 [29], where she derived a sharp control of the KL divergence of the histogram estimator on a fixed model. More precisely, Castellan assumes that there exists a positive constant B such that

$$\inf_{I \in \Lambda_m} \mu(I) \geq B \frac{(\ln(n+1))^2}{n} \; . \tag{18}$$

This latter assumption is thus weaker than (14) for the considered model as its dimension D_m is less than the order $n(\ln(n+1))^{-2}$. We could assume (18) instead of (14) in order to derive Theorem 3.1. This would lead to less precise results for second order terms in the deviations of the excess risks but the first order bounds would be preserved. More precisely, if we replace assumption (14) in Theorem 3.1 by Castellan's assumption (18), a careful look at the proofs show that the conclusions of Theorem 3.1 are still valid for $\varepsilon_n = A_0 (\ln(n+1))^{-1/2}$, where A_0 is some positive constant. Thus assumption (14) is not a fundamental restriction in comparison to Castellan's work [29], but it leads to more precise results in terms of deviations of the true and empirical excess risks of the histogram estimator.

The proof of Theorem 3.1, that can be found in Section A.1, is based on an improvement of independent interest of the previously best known concentration inequality for the chi-square statistics. See Section 4.1 for the precise result.

3.2 An Oracle Inequality

First, let us state the set of five structural assumptions required to establish the nonasymptotic optimality of the over-penalization procedure. These assumptions will be discussed in more detail at the end of this section, following the statement of a sharp oracle inequality.

Set of assumptions (**SA**)

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

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $m \in \mathcal{M}_n$,

$$D_m \leq A_{\mathcal{M},+} \frac{n}{(\ln(n+1))^2} \leq n \; .$$

(Asm) The unknown density f_* satisfies some moment condition and is uniformly bounded from below: there exist some constants $A_{\min} > 0$ and $p > 1$ such that,

$$\int_{\mathcal{Z}} f_*^p \left[(\ln f_*)^2 \vee 1 \right] d\mu < +\infty$$

and

$$\inf_{z \in \mathcal{Z}} f_*(z) \geq A_{\min} > 0 \; . \tag{19}$$

(Alr) Lower regularity of the partition with respect to μ: there exists a positive finite constant A_Λ such that, for all $m \in \mathcal{M}_n$,

$$D_m \inf_{I \in \Lambda_m} \mu(I) \geq A_\Lambda \geq A_{\mathcal{M},+}(\alpha_{\mathcal{M}} + 6)/\tau \; ,$$

where $\tau = \sqrt{3} - \sqrt{2} > 0$.

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

$$C_- D_m^{-\beta_-} \leq \mathcal{K}\left(f_*, f_m\right) \leq C_+ D_m^{-\beta_+} \ .$$

We are now ready to state our main theorem related to optimality of over-penalization.

Theorem 3.2 *Take $n \geq 1$ and $r \in (0, p-1)$. For some $\Delta > 0$, consider the following penalty,*

$$\mathrm{pen}\left(m\right) = \left(1 + \Delta \varepsilon_n^+\left(m\right)\right) \frac{D_m}{n} \ , \qquad \textit{for all } m \in \mathcal{M}_n \ .$$

*Assume that the set of assumptions **(SA)** holds and that*

$$\beta_- < p\left(1+\beta_+\right)/\left(1+p+r\right) \quad \textit{or} \quad p/\left(1+r\right) > \beta_- + \beta_-/\beta_+ - 1 \ . \tag{20}$$

*Then there exists an event Ω_n of probability at least $1-(n+1)^{-2}$ and some positive constant A_1 depending only on the constants defined in **(SA)** such that, if $\Delta \geq A_1 > 0$ then we have on Ω_n,*

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq \left(1+\delta_n\right) \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}\left(f_*, \hat{f}_m\right) \right\} \ , \tag{21}$$

where $\delta_n = L_{\boldsymbol{(SA)},\Delta,r} \left(\ln(n+1)\right)^{-1/2}$ is convenient.

We derive in Theorem 3.2 a pathwise oracle inequality for the KL excess risk of the selected estimator, with constant almost one. Our result thus establishes the nonasymptotic quasi-optimality of over-penalization with respect to the KL divergence.

It is worth noting that three very strong features related to oracle inequality (21) significantly improve upon the literature. Firstly, inequality (21) expresses the performance of the selected estimator through its KL divergence and compare it to the KL divergence of the oracle. Nonasymptotic results pertaining to (robust) maximum likelihood based density estimation usually control the Hellinger risk of the estimator, [30], [43], [22], [20], [10]. The main reason is that the Hellinger risk is much easier to handle that the KL divergence from a mathematical point of view. For instance, the Hellinger distance is bounded by one while the KL divergence can be infinite. However, from a M-estimation perspective, the natural excess risk associated to likelihood optimization is indeed the KL divergence and not the Hellinger distance. These two risks are provably close to each other in the bounded setting [43], but may behave very differently in general.

Second, nonasymptotic results describing the performance of procedures based on penalized likelihood, by comparing more precisely the (Hellinger) risk of the estimator to the KL divergence of the oracle, all deal with the case where the log-density to be estimated is bounded ([30], [43]). Here, we substantially extend the setting by considering only the existence of a finite polynomial moment for the large values of the density to be estimated.

Finally, the oracle inequality (21) is valid with positive probability—larger than 3/4—as soon as one data is available. To our knowledge, any other oracle inequality describing penalization performance for maximum likelihood density estimation is valid with positive probability only when the sample size n is greater than an integer n_0 which depends on the constants defining the problem and that is thus unknown. We emphasize that we control the risk of the selected estimator *for any sample size* and that this property is essential in practice when dealing with small to medium sample sizes. Based on the arguments developed in Section 2.3, we believe that such a feature of Theorem 3.2 is accessible only through the use of over-penalization and we conjecture in particular that *it is impossible using AIC to achieve such a control the KL divergence of the selected estimator for any sample size.*

The oracle inequality (21) is valid under conditions (20) relating the values of the bias decaying rates $\beta_-$and β_+ to the order p of finite moment of the density f_* and the parameter r. In order to understand these latter conditions, let us assume for simplicity that $\beta_- = \beta_+ =: \beta$. Then the conditions (20) both reduce to $\beta < p/(1+r)$. As r can be taken as close to zero as we want, the latter inequality reduces to $\beta < p$. In particular, if the density to be estimated is bounded ($p = \infty$), then conditions (20) are automatically satisfied. If on the contrary the density f_* only has finite polynomial moment p, then the bias should not decrease too fast. In light of the following comments, if f_* is assumed to be α-Hölderian, $\alpha \in (0,1]$, then $\beta \le 2\alpha \le 2$ and the conditions (20) are satisfied, in the case where $\beta_- = \beta_+$, as soon as $p \ge 2$.

To conclude this section, let's now comment on the set of assumptions (**SA**). Assumption (**P1**) indicates that the collection of models has increasing polynomial complexity. This is well suited to bin size selection because in this case we usually select among a number of models which is strictly bounded from above by the sample size. In the same manner, Assumption (**P2**) is legitimate enough for us and corresponds to practice, where we aim at considering bin sizes for which each element of the partition contains a few sample points.

Assumption (**Asm**) imposes conditions on the density to be estimated. More precisely, Assumption (19) stating that the unknown density is uniformly bounded from below can also be found in the work of Castellan [29]. The author assumes moreover in Theorem 3.4 where she derives an oracle inequality for the (weighted) KL excess risk of the histogram estimator, that the target is of finite sup-norm. Furthermore, from a statistical perspective, the lower bound (19) is legitimate since, by Assumption (**Alr**), we use models of lower-regular partitions with respect to the Lebesgue measure. In the case where Inequality (19) would not hold, one would typically have to consider exponentially many irregular histograms to take into account the possibly vanishing mass of some elements of the partitions (for more details on this aspect that goes beyond the scope of the present paper, see for instance [43]).

We require in (**Ap**) that the quality of the approximation of the collection of models is good enough in terms of bias. More precisely, we require a polynomially decreasing of excess risk of KL projections of the unknown density onto the models. For a density f_* uniformly bounded away from zero, the upper bound on the bias is satisfied when for example, $\mathcal{Z}$ is the unit interval, $\mu = \text{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_m are regular and the density f_* belongs to the set $\mathcal{H}(H,\alpha)$ of α-hölderian functions for some $\alpha \in (0,1]$: if $f \in \mathcal{H}(H,\alpha)$, then for all $(x,y) \in \mathcal{Z}^2$

$$|f(x) - f(y)| \le H\,|x-y|^{\alpha} \ .$$

In that case, $\beta_+ = 2\alpha$ is convenient and AIC-type procedures are adaptive to the parameters H and α, see Castellan [29].

In assumption (**Ap**) of Theorem 3.2 we also assume that the bias $\mathcal{K}(f_*, f_m)$ is bounded from below by a power of the dimension D_m of the model m. This hypothesis is in fact quite classical as it has been used by Stone [51] and Burman [28] for the estimation of density on histograms and also by Arlot and Massart [8] and Arlot [4], [5] in the regression framework. Combining Lemma 1 and 2 of Barron and Sheu [12] - see also inequality (34) of Proposition 4.7 below – we can show that

$$\frac{1}{2} e^{-3\left\|\ln\left(\frac{f_*}{f_m}\right)\right\|_\infty} \int_{\mathcal{Z}} \frac{(f_m - f_*)^2}{f_*} d\mu \le \mathcal{K}(f_*, f_m)$$

and thus assuming for instance that the target is uniformly bounded, $\|f_*\|_\infty \le A_*$, we get

$$\frac{A_{\min}^3}{2A_*^4} \int_{\mathcal{Z}} (f_m - f_*)^2 \, d\mu \le \mathcal{K}(f_*, f_m) \ .$$

Now, since in the case of histograms the KL projection f_m is also the $L_2(\mu)$ projection of f_* onto m, we can apply Lemma 8.19 in Section 8.10 of Arlot [3] to show that assumption (**Ap**) is indeed satisfied for $\beta_- = 1 + \alpha^{-1}$, in the case where $\mathcal{Z}$ is the unit interval, $\mu = \text{Leb}$ is the Lebesgue measure on the unit interval, the partitions Λ_m are regular and the density f_* is a non-constant α-hölderian function.

The proof of Theorem 3.2 and further descriptions of the behavior of the procedure can be found in the supplementary material, Section A.2.

4 Probabilistic and Analytical Tools

In this section we set out some general results that are of independent interest and serve as tools for the mathematical description of our statistical procedure.

The first two sections contain new or improved concentration inequalities, for the chi-square statistics (Section 4.1) and for general log-densities (Section 4.2). We established in Section 4.3 some results that are related to the so-called margin relation in statistical learning and that are analytical in nature.

4.1 Chi-Square Statistics' Concentration

The chi-square statistics plays an essential role in the proofs related to Section 3.1. Let us recall its definition.

Definition 4.1 *Given some histogram model m, the statistics $\chi_n^2(m)$ is defined by*

$$\chi_n^2(m) = \int_{\mathcal{Z}} \frac{\left(\hat{f}_m - f_m\right)^2}{f_m} d\mu = \sum_{I \in m} \frac{(P_n(I) - P(I))^2}{P(I)} \ .$$

The following proposition provides an improvement upon the previously best known concentration inequality for the right tail of the chi-square statistics, available in [30]—see also [43, Proposition 7.8] and [24, Theorem 12.13].

Proposition 4.2 *For any $x, \theta > 0$, it holds*

$$\mathbb{P}\left(\chi_n(m)\mathbf{1}_{\Omega_m(\theta)} \geq \sqrt{\frac{D_m}{n}} + \left(1 + \sqrt{2\theta} + \frac{\theta}{6}\right)\sqrt{\frac{2x}{n}}\right) \leq \exp(-x) \ , \tag{22}$$

where we set $\Omega_m(\theta) = \bigcap_{I \in m} \{|P_n(I) - P(I)| \leq \theta P(I)\}$. More precisely, for any $x, \theta > 0$, it holds with probability at least $1 - e^{-x}$,

$$\chi_n(m)\mathbf{1}_{\Omega_m(\theta)} < \sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + 2\sqrt{\frac{\theta}{n}}\left(\sqrt{x} \wedge \left(\frac{xD_m}{2}\right)^{1/4}\right) + \frac{\theta}{3}\sqrt{\frac{x}{n}}\left(\sqrt{\frac{x}{D_m}} \wedge \frac{1}{\sqrt{2}}\right) \ . \tag{23}$$

The proof of Theorem 4.2 can be found in Section 6.1.

Let us detail its relationship with the bound in Proposition 7.8 in [43], which is: for any $x, \varepsilon > 0$,

$$\mathbb{P}\left(\chi_n(m)\mathbf{1}_{\Omega_m(\varepsilon^2/(1+\varepsilon/3))} \geq (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}}\right)\right) \leq \exp(-x) \ . \tag{24}$$

By taking $\theta = \varepsilon^2/(1+\varepsilon/3) > 0$, we get $\varepsilon = \theta/6 + \sqrt{\theta^2/36 + \theta} > \theta/6 + \sqrt{\theta} > 0$. Assume that $D_m \geq 2x$. We obtain by (22), with probability at least $1 - \exp(-x)$,

$$\begin{aligned}
\chi_n(m)\mathbf{1}_{\Omega_m(\theta)} &< \sqrt{\frac{D_m}{n}} + \left(1 + \sqrt{2\theta} + \frac{\theta}{6}\right)\sqrt{\frac{2x}{n}} \\
&\leq \sqrt{\frac{D_m}{n}} + \left(1 + \sqrt{2}\varepsilon\right)\sqrt{\frac{2x}{n}} \\
&< (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}}\right) \ .
\end{aligned}$$

So in this case, inequality (22) improves upon (24). Notice that in our statistical setting (see Theorem 3.1) the restriction on D_m is as follows $D_m \leq A_+ n(\ln(n+1))^{-1}$. Furthermore, in our proofs we apply (22) with x proportional to $\ln(n+1)$ (see the proof of Theorem 3.1). Hence, for a majority of models, we have $x \ll D_m$ and so

$$\sqrt{\frac{2x}{n}} \ll \sqrt{\frac{D_m}{n}} \ .$$

As a result, the bounds that we obtain in Theorem 3.1 by the use of Inequality (22) are substantially better than the bounds we would obtain by using Inequality (24) of [43]. More precisely, the term $\sqrt{D_m \ln(n+1)/n}$ in (15) would be replaced by $(D_m \ln(n+1)/n)^{1/4}$, thus changing the order of magnitude for deviations of the excess risks.

Now, if $D_m \leq 2x$ then (22) gives that with probability at least $1-e^{-x}$,

$$\begin{aligned}
\chi_n(m)\,\mathbf{1}_{\Omega_m(\theta)} &< \sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + 2\sqrt{\frac{\theta}{n}}\left(\frac{xD_m}{2}\right)^{1/4} + \frac{\theta}{6}\sqrt{\frac{2x}{n}} \\
&\leq \left(1+\frac{\sqrt{\theta}}{2^{1/4}}\right)\sqrt{\frac{D_m}{n}} + \left(1+\frac{\sqrt{\theta}}{2^{3/4}}+\frac{\theta}{6}\right)\sqrt{\frac{2x}{n}} \\
&< (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}}\right) \ .
\end{aligned}$$

So in this case again, inequality (23) improves upon (24), with an improvement that can be substantial depending on the value of the ratio x/D_m.

The following result describes the concentration from the left of the chi-square statistics and is proved in Section 6.1.

Proposition 4.3 *Let α, $A_\Lambda > 0$. Assume $0 < A_\Lambda \leq D_m \inf_{I\in m}\{P(I)\}$. Then there exists a positive constant A_g depending only on A_Λ and α such that*

$$\mathbb{P}\left(\chi_n(m) \leq \left(1 - A_g\left(\sqrt{\frac{\ln(n+1)}{D_m}} \vee \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\right)\sqrt{\frac{D_m}{n}}\right) \leq (n+1)^{-\alpha} \ .$$

4.2 Bernstein type concentration inequalities for log-densities

The following proposition gives concentration inequalities for the empirical bias at the right of its mean.

Proposition 4.4 *Consider a density $f \in \mathcal{S}$. We have, for all $z \geq 0$,*

$$\mathbb{P}\left(P_n(\ln(f/f_*)) \geq \frac{z}{n}\right) \leq \exp(-z) \ . \tag{25}$$

Moreover, if we can take a finite quantity v which satisfies $v \geq \int (f \vee f_*)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu$, we have for all $z \geq 0$,

$$\mathbb{P}\left((P_n - P)(\ln(f/f_*)) \geq \sqrt{\frac{2vz}{n}} + \frac{2z}{n}\right) \leq \exp(-z) \ . \tag{26}$$

One can notice, with Inequality (25), that the empirical bias always satisfies some exponential deviations at the right of zero. In the Information Theory community, this inequality is also known as the "No Hyper-compression Inequality" ([35]).

Inequality (26) seems to be new and takes the form of a Bernstein-like inequality, even if the usual assumptions of Bernstein's inequality are not satisfied. In fact, we are able to recover such a behavior by inflating the usual variance to the quantity v.

We now turn to concentration inequalities for the empirical bias at the left of its mean.

Proposition 4.5 *Let $r>0$. For any density $f\in\mathcal{S}$ and for all $z\geq 0$, we have*

$$\mathbb{P}\left(P_n\left(\ln\left(f/f_*\right)\right)\leq -z/nr-(1/r)\ln\left(P\left[\left(f_*/f\right)^r\right]\right)\right)\leq \exp\left(-z\right)\ . \tag{27}$$

Moreover, if we can set a quantity w_r which satisfies $w_r\geq\int\left(\frac{f_*^{r+1}}{f^r}\vee f_*\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu$, then we get, for all $z\geq 0$,

$$\mathbb{P}\left(\left(P_n-P\right)\left(\ln\left(f/f_*\right)\right)\leq -\sqrt{\frac{2w_r z}{n}}-\frac{2z}{nr}\right)\leq \exp\left(-z\right)\ . \tag{28}$$

4.3 Margin-Like Relations

Our objective in this section is to control the variance terms v and w_r, appearing respectively in Lemma 4.4 and 4.5, in terms of the KL divergence pointed on the target f_*. This is done in Lemma 4.6 below under moment assumptions for f_*.

Proposition 4.6 *Let $p>1$ and $c_+,c_->0$. Assume that the density f_* satisfies*

$$J:=\int_{\mathcal{Z}} f_*^p\left(\left(\ln\left(f_*\right)\right)^2\vee 1\right)d\mu<+\infty \text{ and } Q:=\int_{\mathcal{Z}}\frac{\left(\ln\left(f_*\right)\right)^2\vee 1}{f_*^{p-1}}d\mu<+\infty \tag{29}$$

Take a density f such that $0<c_-\leq\inf_{z\in\mathcal{Z}}\{f(z)\}\leq\sup_{z\in\mathcal{Z}}\{f(z)\}\leq c_+<+\infty$. Then, for some $A_{MR,d}>0$ only depending on J,Q,p,c_+ and c_-, it holds

$$P\left[\left(\frac{f}{f_*}\vee 1\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right]\leq A_{MR,d}\times\mathcal{K}\left(f_*,f\right)^{1-\frac{1}{p}}\ . \tag{30}$$

More precisely, $A_{MR,d}=\left(4c_-^{1-p}\left(\left(\ln c_-\right)^2\vee 1\right)J+4c_+^p\left(\left(\ln c_+\right)^2\vee 1\right)Q\right)^{1/p}$ holds. For any $0<r\leq p-1$, we have the following inequality,

$$P\left[\left(\frac{f_*}{f}\vee 1\right)^r\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right]\leq A_{MR,g}\times\mathcal{K}\left(f_*,f\right)^{1-\frac{r+1}{p}}\ , \tag{31}$$

available with $A_{MR,g}=\left(4c_-^{1-p}\left(\left(\ln c_-\right)^2\vee 1\right)J+2\left(\left(\ln\left(c_+\right)\right)^2+J+Q\right)\right)^{\frac{r+1}{p}}$.

Theorem 4.6 states that the variance terms, appearing in the concentration inequalities of Section 4.2, are bounded from above, under moment restrictions on the density f_*, by a power less than one of the KL divergence pointed on f_*. The stronger are the moment assumptions, given in (29), the closer is the power to one. One can notice that J is a restriction on large values of f_*, whereas Q is related to values of f_* around zero.

We call these inequalities "margin-like relations" because of their similarity with the margin relations known first in binary classification ([42], [53]) and then extended to empirical risk minimization (see [44] and [6] for instance). Indeed, from a general point of view, margin relations relate the variance of contrasted functions (logarithm of densities here) pointed on the contrasted target to a function (in most cases, a power) of their excess risk.

Now we're tightening the restrictions on the values of f_* around zero. Indeed, we ask in the following lemma that the target is uniformly bounded away from zero.

Proposition 4.7 *Let $p>1$ and $A_{\min},c_+,c_->0$. Assume that the density f_* satisfies*

$$J:=\int_{\mathcal{Z}} f_*^p\left(\left(\ln\left(f_*\right)\right)^2\vee 1\right)d\mu<+\infty \text{ and } 0<A_{\min}\leq\inf_{z\in\mathcal{Z}} f_*\left(z\right)\ .$$

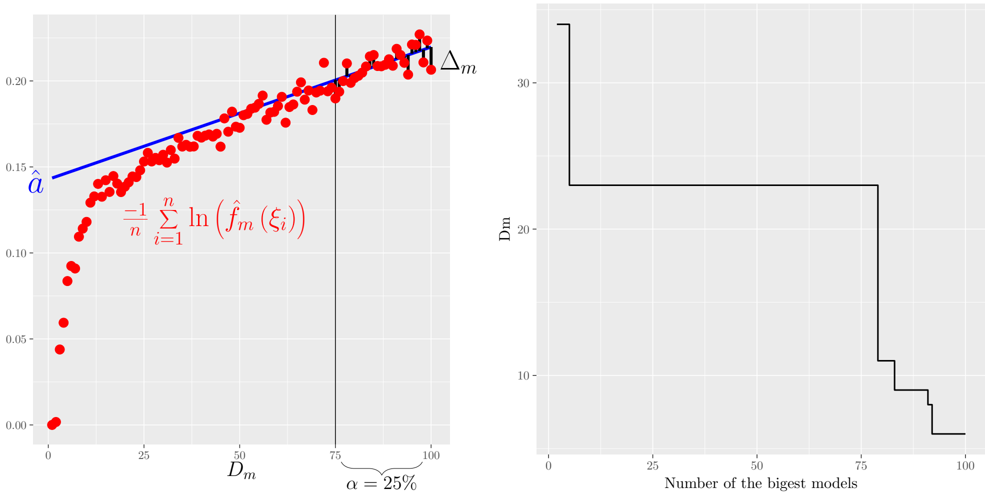


Figure 5: Estimation of the over-penalization constant.

Then there exists a positive constant $A_{MR,-}$ *only depending on* $A_{\min}, J, r$ *and* p *such that, for any* $m \in \mathcal{M}_n$*,*

$$P\left[\left(\frac{f_m}{f_*} \vee 1\right)\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq A_{MR,-} \times \mathcal{K}\left(f_*, f_m\right)^{1-1/p} \tag{32}$$

and for any $0 < r \leq p-1$*,*

$$P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq A_{MR,-} \times \mathcal{K}\left(f_*, f_m\right)^{1-\frac{r+1}{p}} \quad . \tag{33}$$

If moreover $\ln(f_*) \in L_\infty(\mu)$*, i.e.* $0 < A_{\min} \leq \inf_{z\in\mathcal{Z}} f_*(z) \leq \|f_*\|_\infty < +\infty$*, then there exists* $\tilde{A} > 0$ *only depending on* $r, A_{\min}$ *and* $\|f_*\|_\infty$ *such that, for any* $m \in \mathcal{M}_n$*,*

$$P\left[\left(\frac{f_m}{f_*} \vee 1\right)\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \vee P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq \tilde{A} \times \mathcal{K}\left(f_*, f_m\right) \quad . \tag{34}$$

It is worth noting that Lemma 4.7 is stated only for projections f_m because we actually take advantage of their special form (as local means of the target) in the proof of the lemma. The benefit, compared to results of Lemma 4.6, is that Inequalities (32), (33) and (34) do not involve assumptions on the values of f_m (and in particular they do not involve the sup-norm of f_m).

5 Experiments

A simulation study is conducted to compare the numerical performance of the model selection procedures we discussed. We demonstrate the usefulness of our procedure on simulated data examples. The numerical experiments were performed using R.

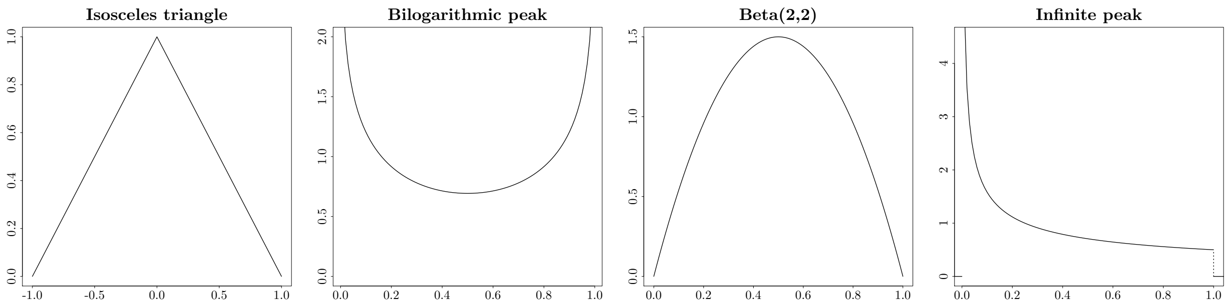
5.1 Experimental Setup

We have compared the numerical performance of our procedure with the classic methods of penalisation of the literature on several densities. In particular, we consider the estimator of [22] and AICc ([52, 36]). We also report on AIC's behaviour. In the following, we name the procedure of [22] by BR, and our criterion AIC_1 when the constant $C = 1$ in (10) and AIC_a for a fully adaptive procedure which will be detailed below. More specifically, the performance of the following four model selection methods were compared:

1. AIC:
$$\widehat{m}_{\mathrm{AIC}} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \mathrm{crit}_{\mathrm{AIC}}(m) \right\},$$
with
$$\mathrm{crit}_{\mathrm{AIC}}(m) = P_n(\gamma(\widehat{f}_m)) + \frac{D_m}{n}.$$
2. AICc:
$$\widehat{m}_{\mathrm{AICc}} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n(\gamma(\widehat{f}_m)) + \mathrm{pen}_{\mathrm{AICc}}(m) \right\},$$
with
$$\mathrm{pen}_{\mathrm{AICc}}(m) = \frac{D_m}{n - D_m - 1}.$$
3. BR:
$$\widehat{m}_{\mathrm{BR}} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n(\gamma(\widehat{f}_m)) + \mathrm{pen}_{\mathrm{BR}}(m) \right\},$$
with
$$\mathrm{pen}_{\mathrm{BR}}(m) = \frac{(\log D_m)^{2.5}}{n}$$
4. AIC_1:
$$\widehat{m}_{\mathrm{AIC}_1} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n(\gamma(\widehat{f}_m)) + \mathrm{pen}_{\mathrm{AIC}_1}(m) \right\},$$
with
$$\mathrm{pen}_{\mathrm{AIC}_1}(m) = 1 \times \max \left\{ \sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}; \frac{\ln(n+1)}{D_m} \right\} \frac{D_m}{n} \ ,$$
5. AIC_a:
$$\widehat{m}_{\mathrm{AIC}_a} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n(\gamma(\widehat{f}_m)) + \mathrm{pen}_{\mathrm{AIC}_a}(m) \right\},$$
with
$$\mathrm{pen}_{\mathrm{AIC}_a}(m) = \hat{C} \max \left\{ \sqrt{\frac{D_m \ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_m}}; \frac{\ln(n+1)}{D_m} \right\} \frac{D_m}{n} \ ,$$
where $\hat{C} = \mathrm{median}_{\alpha \in \mathcal{P}} \hat{C}_\alpha$, with $\hat{C}_\alpha = \mathrm{median}_{m \in \mathcal{M}_\alpha} |\hat{C}_m|$, where
$$\hat{C}_m = \frac{\Delta_m}{\max \left\{ \sqrt{\frac{D_m}{n}}; \sqrt{\frac{1}{D_m}} \right\} \frac{D_m}{2}} \ ,$$

Δ_m is the least-squares distance between the opposite of the empirical risk $-P_n(\gamma(\widehat{f}_m))$ and a fitted line of equation $y = xD_m/n + \hat{a}$ (see Figure 5), $\mathcal{P}$ is the set of proportions α corresponding to the longest plateau of equal selected models when using penalty (10) with constant $C = \hat{C}_\alpha$ and $\mathcal{M}_\alpha$ is the set of models in the collection associated to the proportion α of the largest dimensions.

Let us briefly explain the ideas underlying the design of the fully adaptive AIC_a procedure. According to the definition of penalty $\mathrm{pen}_{\mathrm{opt},\beta}$ given in (8) the constant C in our over-penalization procedure (10) should ideally estimate some normalized deviations of the sum of the excess risk and the empirical excess risk on the models of the collection. Based on Theorem 3.1,we can also assume that the deviations of excess risk and excess empirical risk are of the same order. Moreover, considering the largest models in the collection neglects questions of bias and, therefore, the median of the normalized deviations of the empirical risk around its mean for the largest models should be a reasonable estimator of the C constant. Now the problem is to give a tractable definition to the "largest models" in the collection. To do this, we choose a proportion α of the largest dimensions of the models at hand and calculate using these models an estimator $\hat{C}_\alpha$ of the constant C in (10).



(a) Isosceles triangle (b) Bilogarithmic peak (c) Beta (2,2) (d) Infinite peak

Figure 6: Test densities f.

We then proceed for each α in a grid of values between 0 and 1 to a model selection step by over-penalization using the constant $C = \hat{C}_\alpha$. This gives us a graph of the selected dimensions with respect to the proportions (see Figure 5). Finally we define our over-penalization constant $\hat{C}$ as the median of the values of the constants $\hat{C}_\alpha$, $\alpha \in \mathcal{P}$ where $\mathcal{P}$ is the largest plateau in the graph of the selected dimensions with respect to proportions α.

The models that we used along the experiments are made of histograms densities defined a on a regular partition of the interval $[0, 1]$ (with the exception of the density Isosceles triangle which is supported on $[-1, 1]$). We show the performance of the proposed method for a set of four test distributions (see Figure 6) and described in the *benchden*[1] R-package [45] which provides an implementation of the distributions introduced in [18].

5.2 Results

We compared procedures on $N = 100$ independent data sets of size n ranging from 50 to 1000. We estimate the quality of the model-selection strategies using the median KL divergence. Boxplots were made of the KL risk over the N trials. The horizontal lines of the boxplot indicate the $5\%, 25\%, 50\%, 75\%$, and 95% quantiles of the error distribution. The median value of AIC (horizontal black line) is also superimposed for visualization purposes.

It can be seen from Figure 7 that, as expected, for each method and in all cases, the KL divergence decreases as the sample size increases. We also see clearly that there is generally a substantial advantage to modifying AIC for sample sizes smaller than 1000.

Moreover, none of the methods is more effective than the others in all cases. However, in almost all cases, AIC_1 always had a lower median KL than the others. We therefore recommend using AIC_1 rather than AIC_c. Indeed, compared to AIC_1, it seems that AIC_c is not penalizing enough,which translates into a worse performance for samples equal to 50 and 100. Moreover, both procedures have exactly the same calculation costs.

On the contrary, it seems that the BR criterion penalizes too much. As a result, its performance deteriorates relative to other methods as the sample size increases.

Finally, the behavior of AIC_a is quite good in general, but it is often a little less efficient than AIC_1. We interpret this fact by postulating that the extra computations that we engage in AIC_a to remove any hyper-parameter from the procedure actually induces extra variance in the estimation.

Acknowledgements

The first author warmly thanks Matthieu Lerasle for instructive discussions on the topic of estimation by tests and Alain Célisse for a nice discussion at a early stage of this work. He is also grateful to Pascal Massart for having pushed him towards obtaining better oracle inequalities than in a previous version of this work. Finally, we owe thanks to Sylvain Arlot and Amandine Dubois for a careful reading that helped to correct some mistakes and improve the presentation of the paper.

[1]Available on the CRAN http://cran.r-project.org.

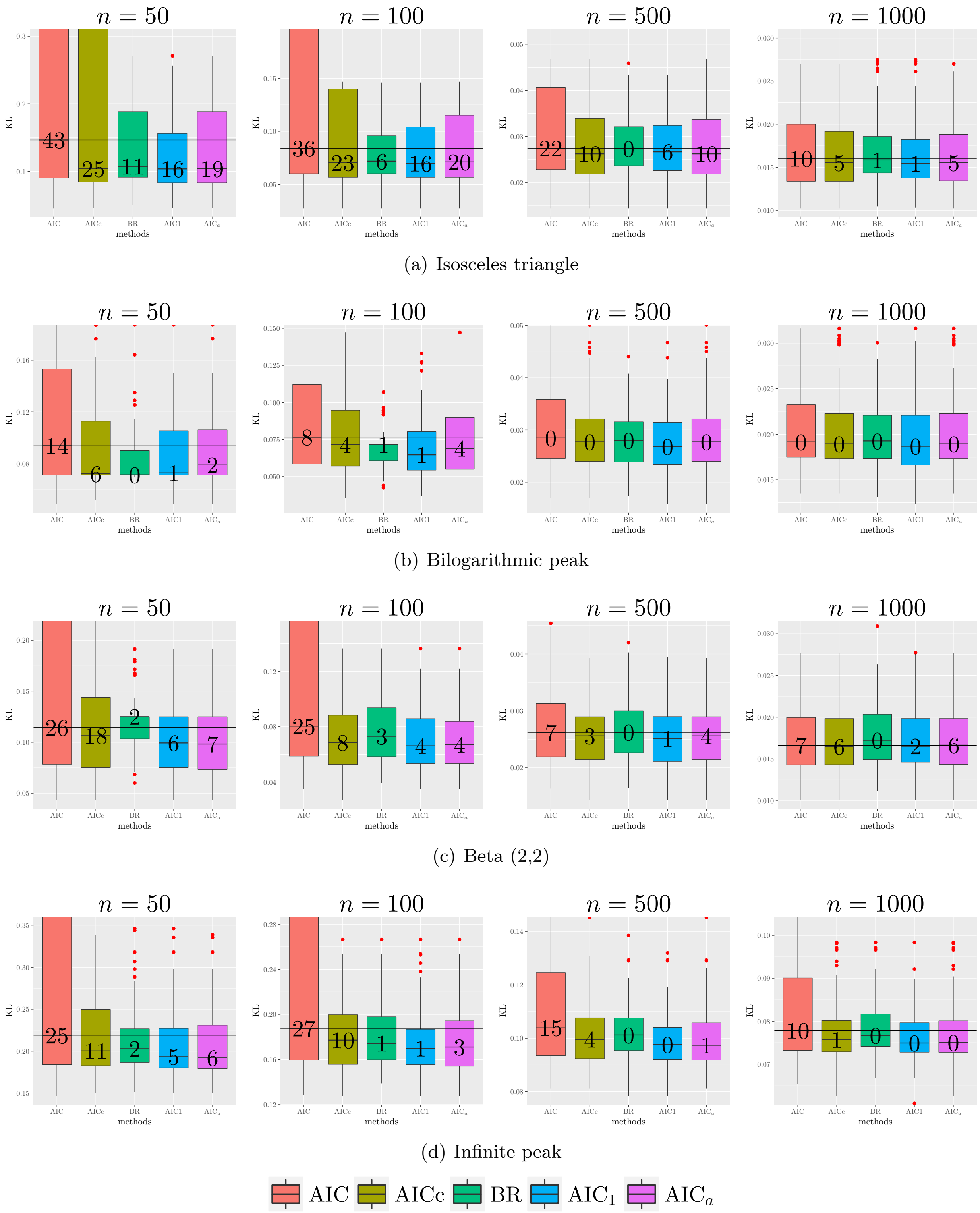


Figure 7: KL divergence results. Box plots of the KL divergence to the true distribution for the estimated distribution. The solid black line corresponds to the AIC KL divergence median. The term inside the box is the number of time the KL divergence equals ∞ out of 100.

6 Proofs

6.1 Proofs Related to Section 4.1

Proof of Theorem 4.2. We fix $x, \theta > 0$ and we set $z > 0$ to be chosen later. Let us set for any $I \in m$, $\varphi_I = (P(I))^{-1/2} \mathbb{1}_I$. The family of functions $(\varphi_I)_{I \in m}$ forms an orthonormal basis of

$(m, \|\cdot\|_2)$. By Cauchy-Schwarz inequality, we have

$$\chi_n(m) = \sup_{(a_I)_{I \in m} \in B_2} \left| (P_n - P)\left(\sum_{I \in m} a_I \varphi_I \right) \right| ,$$

where $B_2 := \left\{ (a_I)_{I \in m} \ ; \ \sum_{I \in m} a_I^2 \leq 1 \right\}$. Furthermore, the case of equality in Cauchy-Schwarz inequality gives us

$$\chi_n(m) = (P_n - P)\left(\sum_{I \in m} a_I^{\infty} \varphi_I \right) \text{ with } a_I^{\infty} = \frac{(P_n - P)(\varphi_I)}{\chi_n(m)} .$$

Hence, by setting $\mathcal{A}(s) := B_2 \bigcap \left\{ (a_I)_{I \in m} \ ; \ \sup_{I \in m} \left| a_I (P(I))^{-1/2} \right| \leq s \right\}$ for any $s \geq 0$, we get

$$\chi_n(m) \mathbf{1}_{\Omega_m(\theta)} \mathbf{1}_{\{\chi_n(m) \geq z\}} \leq \sup_{(a_I)_{I \in m} \in \mathcal{A}(\theta/z)} \left| (P_n - P)\left(\sum_{I \in m} a_I \varphi_I \right) \right| . \tag{35}$$

By applying Bousquet's inequality ([26]) to the supremum in the right-hand side of (35) then gives for any $\delta > 0$,

$$\begin{aligned} &\mathbb{P}\left(\chi_n(m) \mathbf{1}_{\Omega_m(\theta)} \mathbf{1}_{\{\chi_n(m) \geq z\}} \geq (1+\delta) E_m + \sqrt{\frac{2\sigma_m^2 x}{n}} + \left(\frac{1}{\delta} + \frac{1}{3} \right) \frac{b_m x}{n} \right) \qquad (36) \\ \leq &\mathbb{P}\left(\sup_{(a_I)_{I \in m} \in \mathcal{A}(\theta/z)} \left| (P_n - P)\left(\sum_{I \in m} a_I \varphi_I \right) \right| \geq (1+\delta) E_m + \sqrt{\frac{2\sigma_m^2 x}{n}} + \left(\frac{1}{\delta} + \frac{1}{3} \right) \frac{b_m x}{n} \right) \\ \leq &\exp(-x) \ , \end{aligned}$$

with

$$E_m \leq \mathbb{E}[\chi_n(m)] \leq \mathbb{E}^{1/2}\left[\chi_n^2(m) \right] = \sqrt{\frac{D_m}{n}} \quad ; \quad \sigma_m^2 \leq \sup_{(a_I)_{I \in m} \in B_2} \mathbb{V}\left(\sum_{I \in m} a_I \varphi_I(\xi_1) \right) \leq 1$$

and

$$b_m = \sup_{(a_I)_{I \in m} \in \mathcal{A}(\theta/z)} \left\| \sum_{I \in m} a_I \varphi_I \right\|_{\infty} \leq \frac{\theta}{z} .$$

Injecting the latter bounds in (36) and taking $z = \sqrt{(D_m)/n} + \sqrt{2x/n}$, we obtain that with probability at least $1 - \exp(-x)$,

$$\chi_n(m) \mathbf{1}_{\Omega_m(\theta)} < (1+\delta) \sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + \left(\frac{1}{\delta} + \frac{1}{3} \right) \frac{\theta x}{\left(\sqrt{D_m} + \sqrt{2x} \right) \sqrt{n}} . \tag{37}$$

By taking $\delta = \sqrt{\theta x} \left(D_m + \sqrt{2x D_m} \right)^{-1/2}$, the right-hand side of (37) becomes

$$\begin{aligned} &\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + 2\sqrt{\frac{\theta x}{n}} \frac{\sqrt{D_m}}{\sqrt{D_m + \sqrt{2x D_m}}} + \frac{\theta x}{3\left(\sqrt{D_m} + \sqrt{2x} \right) \sqrt{n}} \\ \leq &\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2x}{n}} + 2\sqrt{\frac{\theta}{n}} \left(\sqrt{x} \wedge \left(\frac{x D_m}{2} \right)^{1/4} \right) + \frac{\theta}{3} \sqrt{\frac{x}{n}} \left(\sqrt{\frac{x}{D_m}} \wedge \frac{1}{\sqrt{2}} \right) , \end{aligned}$$

which gives (23). Inequality (22) is a direct consequence of (23). ■

Proof of Theorem 4.3. Let us set for any $I \in m$, $\varphi_I = (P(I))^{-1/2} \mathbb{1}_I$. The family of functions $(\varphi_I)_{I \in m}$ forms an orthonormal basis of $(m, \|\cdot\|_2)$. By Cauchy-Schwarz inequality, we have

$$\chi_n(m) = \sup_{(a_I)_{I \in m} \in B_2} \left| (P_n - P)\left(\sum_{I \in m} a_I \varphi_I \right) \right| ,$$

where $B_2 = \left\{ (a_I)_{I\in m} ; \sum_{I\in m} a_I^2 \leq 1 \right\}$. As

$$\sigma_m^2 := \sup_{(a_I)_{I\in m}\in B_2} \mathbb{V}\left(\sum_{I\in m} a_I \varphi_I (\xi_1) \right) \leq 1$$

and

$$b_m = \sup_{(a_I)_{I\in m}\in B_2} \left\| \sum_{I\in m} a_I \varphi_I \right\|_\infty \leq \sup_{I\in m} \|\varphi_I\|_\infty = \sup_{I\in m} \frac{1}{\sqrt{P(I)}} ,$$

we get by Klein-Rio's inequality (see [38]), for every $x, \delta > 0$,

$$\mathbb{P}\left(\chi_n (m) \leq (1-\delta) \mathbb{E}[\chi_n (m)] - \sqrt{\frac{2x}{n}} - \left(\frac{1}{\delta}+1\right)\frac{b_m x}{n} \right) \leq \exp(-x) \ . \tag{38}$$

Note that we have $b_m \leq \sqrt{D_m A_\Lambda^{-1}}$. Now, we bound $\mathbb{E}[\chi_n (m)]$ by below. By Theorem 11.10 in [24] applied to $\chi_n (m)$, we get, for any $\zeta > 0$,

$$\begin{aligned} \mathbb{E}\left[\chi_n^2 (m)\right] - \mathbb{E}^2 [\chi_n (m)] &\leq \frac{\sigma_m^2}{n} + 4\frac{b_m}{n}\mathbb{E}[\chi_n (m)] \\ &\leq \frac{1}{n} + 4\frac{\sqrt{D_m A_\Lambda^{-1}}}{n}\mathbb{E}[\chi_n (m)] \\ &\leq \frac{1}{n} + 4\zeta^{-1}\frac{D_m A_\Lambda^{-1}}{n^2} + \zeta \mathbb{E}^2 [\chi_n (m)] \ . \end{aligned} \tag{39}$$

The latter inequality results from $2ab \leq \zeta a^2 + \zeta^{-1} b^2$ applied with $a = \mathbb{E}[\chi_n (m)]$ and $b = 2\sqrt{D_m A_\Lambda^{-1}}/n$. As $\mathbb{E}\left[\chi_n^2 (m)\right] = D_m/n$, (39) applied with $\zeta = n^{-1/2}$ gives

$$\begin{aligned} \mathbb{E}[\chi_n (m)] &\geq \sqrt{\frac{1}{1+n^{-1/2}}\left(\frac{D_m}{n} - 4A_\Lambda^{-1}\frac{D_m}{n^{3/2}} - \frac{1}{n}\right)_+} \\ &\geq \sqrt{\frac{D_m}{n}}\left(1 - L_{A_\Lambda} D_m^{-1/2} \vee n^{-1/4}\right) \ . \end{aligned} \tag{40}$$

Hence, by using (40) and taking $x = \alpha \ln(n+1)$ and $\delta = n^{-1/4}\sqrt{\ln(n+1)}$ in (38), we obtain with probability at least $1-(n+1)^{-\alpha}$,

$$\begin{aligned} &\chi_n (m) \\ &\geq \left(1 - \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\sqrt{\frac{D_m}{n}}\left(1 - L_{A_\Lambda}\frac{1}{\sqrt{D_m}} \vee \frac{1}{n^{1/4}}\right) \\ &\quad - \sqrt{\frac{2\alpha\ln(n+1)}{n}} - \left(\frac{n^{1/4}}{\sqrt{\ln(n+1)}} + 1\right)\frac{\alpha\sqrt{D_m}\ln(n+1)}{n} \\ &\geq \sqrt{\frac{D_m}{n}}\left(1 - L_{A_\Lambda}\frac{\sqrt{\ln(n+1)}}{n^{1/4}} \vee \frac{1}{\sqrt{D_m}} - \sqrt{\frac{2\alpha\ln(n+1)}{D_m}} - L_\alpha\frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right) \\ &\geq \sqrt{\frac{D_m}{n}}\left(1 - L_{A_\Lambda,\alpha}\frac{\sqrt{\ln(n+1)}}{n^{1/4}} \vee \sqrt{\frac{\ln(n+1)}{D_m}}\right) \ , \end{aligned}$$

which concludes the proof. ■

6.2 Proofs related to Section 4.2

In this section, our aim is to control for any histogram model m, the deviations of the empirical bias from its mean, which is the true bias. Concentration inequalities for the centered empirical bias are provided in Section 4.2. In order to prove oracle inequalities for the KL divergence, we relate the magnitude of these deviations to the true bias. This is done in Section 4.3. Since we believe that the results in this section are of interest in themselves, we have stated them, when possible, for a general density in $\mathcal{S}$ rather for the projections f_m.

The results in this section are based on the Cramèr-Chernoff method (see [23] for instance). Let us recall that if we set $S := \sum_{i=1}^n X_i - \mathbb{E}[X_i]$, where $(X_i)_{i=1}^n$ are n i.i.d. real random variables, and for any $\lambda \geq 0$ and $y \in \mathbb{R}_+$,

$$\varphi_S(\lambda) := \ln(\mathbb{E}[\exp(\lambda S)]) = n(\ln(\mathbb{E}[\exp(\lambda X_1)]) - \lambda\mathbb{E}[X_1])$$

and

$$\varphi_S^*(y) := \sup_{\lambda\in\mathbb{R}_+}\{\lambda y - \varphi_S(\lambda)\}\ ,$$

then

$$\mathbb{P}(S \geq y) \leq \exp(-\varphi_S^*(y))\ . \tag{41}$$

Proof of Proposition 4.4. We first prove concentration inequality (25). We set $X_i := \ln(f/f_*)(\xi_i)$ and use Inequality (41). For $\lambda \in [0,1]$, as $\mathbb{E}[X_1] = -\mathcal{K}(f_*, f)$, we have

$$\begin{aligned}\varphi_S(\lambda)/n =& \ln\left(P\left[(f/f_*)^\lambda\right]\right) + \lambda\mathcal{K}(f_*, f)\\ \leq& \lambda\ln(P[f/f_*]) + \lambda\mathcal{K}(f_*, f) = \lambda\mathcal{K}(f_*, f)\ ,\end{aligned}$$

where the inequality derives from the concavity of the function $x \mapsto x^\lambda$. By setting $\mathcal{K} = \mathcal{K}(f_*, f)$, we thus get

$$\varphi_S^*(y) \geq \sup_{\lambda\in[0,1]}\{\lambda y - \varphi_S(\lambda)\} \geq (y - n\mathcal{K})_+\ ,$$

which implies,

$$\mathbb{P}((P_n - P)(\ln(f/f_*)) \geq x) \leq \exp(-n(x-\mathcal{K})_+)\ . \tag{42}$$

Inequality (25) is a direct consequence of (42). Moreover, we notice that for any $u \in \mathbb{R}$, $\exp(u) \leq 1 + u + \frac{u^2}{2}\exp(u_+)$ and $\ln(1+u) \leq u$, where $u_+ = u \vee 0$. By consequence, for $\lambda \in [0,1]$, it holds

$$\begin{aligned}\varphi_S(\lambda)/n &= \ln(\mathbb{E}[\exp(\lambda X_1)]) - \lambda\mathbb{E}[X_1]\\ &\leq \ln\left(1 + \lambda\mathbb{E}[X_1] + \frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left(\lambda(X_1)_+\right)\right]\right) - \lambda\mathbb{E}[X_1]\\ &\leq \frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left(\lambda(X_1)_+\right)\right] \leq \frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left((X_1)_+\right)\right] \leq \frac{\lambda^2}{2}v\ .\end{aligned}$$

Now, we get, for any $y \geq 0$,

$$\varphi_S^*(y) \geq \sup_{\lambda\in[0,1]}\{\lambda y - \varphi_S(\lambda)\} \geq \sup_{\lambda\in[0,1]}\left\{\lambda y - n\frac{\lambda^2 v}{2}\right\} = \left(\frac{y^2}{2nv}\mathbf{1}_{y\leq nv} + \left(y - \frac{nv}{2}\right)\mathbb{1}_{y>nv}\right)\ . \tag{43}$$

So, by using (41) with (43) taken with $x = y/n$, it holds

$$\mathbb{P}((P_n - P)(\ln(f/f_*)) \geq x) \leq \exp\left(-n\left(\frac{x^2}{2v}\mathbb{1}_{x\leq v} + \left(x - \frac{v}{2}\right)\mathbb{1}_{x>v}\right)\right)\ . \tag{44}$$

To obtain (26), we notice that Inequality (44) implies by simple calculations, for any $z \geq 0$,

$$\mathbb{P}\left((P_n - P)(\ln(f/f_*)) \geq \sqrt{\frac{2vz}{n}}\mathbb{1}_{z\leq nv/2} + \left(\frac{z}{n} + \frac{v}{2}\right)\mathbb{1}_{z>nv/2}\right) \leq \exp(-z)\ .$$

To conclude the proof, it suffices to remark that $\sqrt{2vz/n}\mathbb{1}_{z\leq nv/2}+(z/n+v/2)\,\mathbb{1}_{z>nv/2}\leq\sqrt{2vz/n}+2z/n$. ■

Proof of Proposition 4.5. Let us first prove the inequality of concentration (27). We set $X_i := \ln(f_*/f)(\xi_i)$ and use (41). For $\lambda\in[0,r]$, we have by Hölder's inequality, $P\left[(f_*/f)^\lambda\right]\leq P\left[(f_*/f)^r\right]^{\lambda/r}$. Hence,

$$\begin{aligned}\varphi_S(\lambda)/n &= \ln\left(P\left[(f_*/f)^\lambda\right]\right)-\lambda\mathcal{K}(f_*,f)\\ &\leq\lambda\left(\frac{1}{r}\ln\left(P\left[(f_*/f)^r\right]\right)-\mathcal{K}(f_*,f)\right)\ .\end{aligned}$$

Let us notice that by concavity of ln, we have $\frac{1}{r}\ln\left(P\left[(f_*/f)^r\right]\right)-\mathcal{K}(f_*,f)\geq 0$. Now we get, for any $y\geq 0$,

$$\begin{aligned}\varphi_S^*(y) &\geq \sup_{\lambda\in[0,r]}\{\lambda y-\varphi_S(\lambda)\}\\ &\geq \sup_{\lambda\in[0,r]}\left\{\lambda\left(y-n\left(\frac{1}{r}\ln\left(P\left[(f_*/f)^r\right]\right)-\mathcal{K}(f_*,f)\right)\right)\right\}\\ &= \left(ry-n\left(\ln\left(P\left[(f_*/f)^r\right]\right)-r\mathcal{K}(f_*,f)\right)\right)_+\ .\end{aligned}$$

Using (41), we obtain, for any $x\geq 0$,

$$\mathbb{P}\left((P_n-P)(\ln(f/f_*))\leq -x\right)\leq\exp\left(-n(rx-\ln\left[(f_*/f)^r\right]+r\mathcal{K}(f_*,f))_+\right)\ . \tag{45}$$

Inequality (27) is a straightforward consequence of (45). As in the proof of Theorem 4.4, we notice that for any $u\in\mathbb{R}$, $\exp(u)\leq 1+u+\frac{u^2}{2}\exp(u_+)$ and $\ln(1+u)\leq u$, where $u_+=u\vee 0$. By consequence, for $\lambda\in[0,r]$, it holds

$$\begin{aligned}\varphi_S(\lambda)/n &= \lambda\mathbb{E}[X_1]+\ln\left(\mathbb{E}\left[\exp(-\lambda X_1)\right]\right)\\ &\leq \ln\left(1-\lambda\mathbb{E}[X_1]+\frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left(\lambda(-X_1)_+\right)\right]\right)+\lambda\mathbb{E}[X_1]\\ &\leq \frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left(\lambda(-X_1)_+\right)\right]\leq\frac{\lambda^2}{2}\mathbb{E}\left[X_1^2\exp\left(r(-X_1)_+\right)\right]\leq\frac{\lambda^2}{2}w_r\ .\end{aligned}$$

Now we get, for any $y\geq 0$,

$$\begin{aligned}\varphi_S^*(y) &\geq \sup_{\lambda\in[0,r]}\{\lambda y-\varphi_S(\lambda)\}\\ &\geq \sup_{\lambda\in[0,r]}\left\{\lambda y-\frac{n\lambda^2 w_r}{2}\right\}=\frac{y^2}{2nw_r}\mathbb{1}_{y\leq rnw_r}+r\left(y-\frac{rnw_r}{2}\right)\mathbb{1}_{y>rnw_r}\ ,\end{aligned}$$

which gives

$$\mathbb{P}\left((P_n-P)(\ln(f/f_*))\leq -x\right)\leq\exp\left(-n\left(\frac{x^2}{2nw_r}\mathbb{1}_{x\leq rnw_r}+r\left(x-\frac{rnw_r}{2}\right)\mathbb{1}_{x>rnw_r}\right)\right) \tag{46}$$

Inequality (28) is again a consequence of (46), by the same kind of arguments as those involved in the proof of 26 in Lemma 4.4. ■

6.3 Proofs related to Section 4.3

Proof of Proposition 4.6. Let us take $q > 1$ such that $1/p + 1/q = 1$. It holds

$$
\begin{aligned}
P\left[\left(\frac{f}{f_*}\vee 1\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right] &= \int (f\vee f_*)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu \\
&= \int \frac{f_*\vee f}{f_*\wedge f}\left((f_*\wedge f)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right)^{\frac{1}{p}+\frac{1}{q}} d\mu \\
&= \int \left(\frac{f_*\vee f}{(f_*\wedge f)^{\frac{1}{q}}}\left|\ln\left(\frac{f}{f_*}\right)\right|^{\frac{2}{p}}\right)\left((f_*\wedge f)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right)^{\frac{1}{q}} d\mu \\
&\leq \left(\int (f_*\wedge f)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu\right)^{\frac{1}{q}}\left(\underbrace{\int \frac{(f_*\vee f)^p}{(f_*\wedge f)^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu}_{:=I}\right)^{\frac{1}{p}}
\end{aligned}
$$

where in the last step we used Hölder's inequality. Now, by Lemma 7.24 of Massart [43], it also holds

$$
\frac{1}{2}\int (f_*\wedge f)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu \leq \mathcal{K}(f_*, f)\ .
$$

In order to prove (30), it thus remains to bound I in terms of p, c_+ and c_- only. First, we decompose I into two parts,

$$
\int \frac{(f_*\vee f)^p}{(f_*\wedge f)^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu = \int \frac{f_*^p}{f^{p-1}}\left(\ln\left(\frac{f_*}{f}\right)\right)^2 \mathbb{1}_{f_*\geq f} d\mu + \int \frac{f^p}{f_*^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 \mathbb{1}_{f\geq f_*} d\mu. \tag{47}
$$

For the first term in the right-hand side of (47), we get

$$
\begin{aligned}
\int \frac{f_*^p}{f^{p-1}}\left(\ln\left(\frac{f_*}{f}\right)\right)^2 \mathbb{1}_{f_*\geq f} d\mu &\leq c_-^{1-p}\int f_*^p\left(\ln\left(\frac{f_*}{c_-}\right)\right)^2 d\mu \\
&\leq 4c_-^{1-p}\left((\ln c_-)^2\vee 1\right)\int f_*^p\left((\ln f_*)^2\vee 1\right) d\mu < +\infty\ ,
\end{aligned} \tag{48}
$$

where in the second inequality we used the following fact: $(a+b)^2 \leq 4\left(a^2\vee 1\right)\left(b^2\vee 1\right)$, for any real numbers a and b. The finiteness of the upper bound is guaranteed by assumption (29). For the second term in the right-hand side of (47), it holds by same kind of arguments that lead to (48),

$$
\begin{aligned}
\int \frac{f^p}{f_*^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 \mathbb{1}_{f\geq f_*} d\mu &\leq c_+^p\int \frac{1}{f_*^{p-1}}\left(\ln\left(\frac{c_+}{f_*}\right)\right)^2 d\mu \\
&\leq 4c_+^p\left((\ln c_+)^2\vee 1\right)\int f_*^{1-p}\left((\ln f_*)^2\vee 1\right) d\mu < +\infty\ ,
\end{aligned} \tag{49}
$$

where in the last inequality we used the following fact: $(a-b)^2 \leq 4\left(a^2\vee 1\right)\left(b^2\vee 1\right)$. Again, the finiteness of the upper bound is guaranteed by (29). Inequality (30) then follows from combining (47), (48) and (49).

Inequality (31) follows from the same kind of computations. Indeed, we have by the use of

Hölder's inequality,

$$P\left[\left(\frac{f_*}{f}\vee 1\right)^r\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right]=\int\left(\frac{f_*^{r+1}}{f^r}\vee f_*\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu$$

$$=\int\left(\frac{f_*^{r+1}\vee f_*f^r}{f^r\left(f_*\wedge f\right)^{1-\frac{r+1}{p}}}\left|\ln\left(\frac{f}{f_*}\right)\right|^{\frac{2(r+1)}{p}}\right)\left(\left(f_*\wedge f\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2\right)^{1-\frac{r+1}{p}} d\mu$$

$$\leq\left(\int\left(f_*\wedge f\right)\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu\right)^{1-\frac{r+1}{p}}\left(\underbrace{\int\frac{f_*^p\vee f_*^{\frac{p}{r+1}}f^{\frac{rp}{r+1}}}{f^{\frac{rp}{r+1}}f_*^{\frac{p}{r+1}-1}\wedge f^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu}_{:=I_r}\right)^{\frac{r+1}{p}} .$$

In order to prove (31), it thus remains to bound I_r in terms of p, c_+ and c_- only. Again, we split I_r into two parts,

$$\int\frac{f_*^p\vee f_*^{\frac{p}{r+1}}f^{\frac{rp}{r+1}}}{f^{\frac{rp}{r+1}}f_*^{\frac{p}{r+1}-1}\wedge f^{p-1}}\left(\ln\left(\frac{f}{f_*}\right)\right)^2 d\mu=\int\frac{f_*^p}{f^{p-1}}\left(\ln\left(\frac{f_*}{f}\right)\right)^2\mathbb{1}_{f_*\geq f}d\mu+\int f_*\left(\ln\left(\frac{f}{f_*}\right)\right)^2\mathbb{1}_{f\geq f_*}d\mu . \tag{50}$$

The first term in the right-hand side of (50) is given by (48) above. For the second term in the right-hand side of (50), it holds

$$\begin{aligned}\int f_*\left(\ln\left(\frac{f}{f_*}\right)\right)^2\mathbb{1}_{f\geq f_*}d\mu &\leq\int f_*\left(\ln\left(\frac{c_+}{f_*}\right)\right)^2 d\mu\\ &\leq 2\left(\ln\left(c_+\right)\right)^2+2P\left(\left(\ln f_*\right)^2\right) .\end{aligned}\tag{51}$$

Furthermore we have $f_*^{p-1}+f_*^{-p}\geq 1$, so $P\left(\left(\ln f_*\right)^2\right)\leq P\left(f_*^{p-1}\left(\ln f_*\right)^2\right)+P\left(f_*^{-p}\left(\ln f_*\right)^2\right)\leq J+Q$ and by (51),

$$\int f_*\left(\ln\left(\frac{f}{f_*}\right)\right)^2\mathbb{1}_{f\geq f_*}d\mu\leq 2\left(\left(\ln\left(c_+\right)\right)^2+J+Q\right)<+\infty ,$$

where the finiteness of the upper bound comes from (29). Inequality (31) then easily follows. ■

Proof of Proposition 4.7. Let us first prove Inequality (32). Considering the proof of Inequality (30) of Lemma 4.6 given above, we see that it is sufficient to bound the second term in the right-hand side of (47), applied with $f=f_m$, in terms of $A_{\min}, J$ and p only. It holds

$$\int\frac{f_m}{f_*^{p-1}}\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\mathbb{1}_{f_m\geq f_*}d\mu\leq A_{\min}\int\left(\frac{f_m}{A_{\min}}\right)^p\left(\ln\left(\frac{f_m}{A_{\min}}\right)\right)^2 d\mu . \tag{52}$$

Now, we set h an auxilliary function, defined by $h\left(x\right)=x^p\left(\ln x\right)^2$ for any $x\geq 1$. It is easily seen that h is convex. So it holds, for any $I\in\Lambda_m$,

$$h\left(\frac{f_m\left(I\right)}{A_{\min}}\right)=h\left(\int_I\frac{f_*}{A_{\min}}\frac{d\mu}{\mu\left(I\right)}\right)\leq\int_I h\left(\frac{f_*}{A_{\min}}\right)\frac{d\mu}{\mu\left(I\right)} .$$

From the latter inequality and from (52), we deduce

$$\int\frac{f_m}{f_*^{p-1}}\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\mathbb{1}_{f_m\geq f_*}d\mu\leq A_{\min}\int h\left(\frac{f_m}{A_{\min}}\right)d\mu\leq A_{\min}\int h\left(\frac{f_*}{A_{\min}}\right)d\mu\leq 4A_{\min}^{1-p}\left(\left(\ln A_{\min}\right)^2\vee 1\right)J .$$

Inequality (32) is thus proved.

In the same manner, to establish Inequality (33) it suffices to adapt the proof of inequality (31) given above by controlling the second term in the right-hand side of (50), applied with $f = f_m$, in terms of $A_{\min}, p$ and $P(\ln f_*)^2$ only. Let us notice that the function f defined on $[1, +\infty)$ by $f(x) = x(\ln x)^2$ is convex. We have

$$\int f_* \left(\ln\left(\frac{f_m}{f_*}\right)\right)^2 \mathbb{1}_{f_m \geq f_*} d\mu \leq \int f_m \left(\ln\left(\frac{f_m}{A_{\min}}\right)\right)^2 d\mu = A_{\min} \int f\left(\frac{f_m}{A_{\min}}\right) d\mu \ .$$

Now, for any $I \in \Lambda_m$, it holds $f\left(\frac{f_m(I)}{A_{\min}}\right) = f\left(\int_I \frac{f_*}{A_{\min}} \frac{d\mu}{\mu(I)}\right) \leq \int_I f\left(\frac{f_*}{A_{\min}}\right) \frac{d\mu}{\mu(I)}$. Hence,

$$\int f_* \left(\ln\left(\frac{f_m}{f_*}\right)\right)^2 \mathbb{1}_{f_m \geq f_*} d\mu \leq A_{\min} \int f\left(\frac{f_*}{A_{\min}}\right) d\mu \leq 2P(\ln f_*)^2 + 2(\ln A_{\min})^2 \ ,$$

which gives the desired upper-bound and proves (33). In the event that $f_* \in L_\infty(\mu)$, we have to prove (34). We have $\inf_{z \in \mathcal{Z}} f_*(z) \leq \inf_{z \in \mathcal{Z}} f_m(z) \leq \|f_m\|_\infty \leq \|f_*\|_\infty$, so it holds

$$P\left[\left(\frac{f_m}{f_*} \vee 1\right)\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \vee P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r \left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq \left(\frac{\|f_*\|_\infty}{A_{\min}}\right)^{r \vee 1} P\left[\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \ .$$

Now, Inequality (34) is a direct consequence of Lemma 1 of Barron and Sheu [12], which contains the following inequality,

$$P\left[\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \leq 2\exp\left(\left\|\ln\left(\frac{f_m}{f_*}\right)\right\|_\infty\right) \mathcal{K}(f_*, f_m) \ .$$

This finishes the proof of Lemma 4.7. ∎

References

[1] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

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

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

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

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

[6] S. Arlot and P. L. Bartlett. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 05 2011.

[7] S. Arlot and A. Célisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

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

[9] Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2):239–284, 2009.

[10] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

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

[12] A. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[13] P. Bauer, B. M. Pötscher, and P. Hackl. Model selection by multiple test procedures. *Statistics*, 19(1):39–44, 1988.

[14] P. Bellec and A. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In V. Panov, editor, *Modern Problems of Stochastic Analysis and Statistics*, pages 315–333, Cham, 2017. Springer International Publishing.

[15] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Towards the study of least squares estimators with convex penalty. *arXiv preprint arXiv:1701.09120*, 2017.

[16] Y. Benjamini and Y. Gavrilov. A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.*, 3(1):179–198, 2009.

[17] Y. Benjamini and D. Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, 100(469):71–93, 2005. With comments and a rejoinder by the authors.

[18] A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38(3):3–59, 1994.

[19] P. Bühlmann, L. Meier, and S. van de Geer. Discussion: "a significance test for the lasso". *Ann. Statist.*, 42(2):469–477, 04 2014.

[20] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[21] L. Birgé. Robust tests for model selection. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 47–64. Inst. Math. Statist., Beachwood, OH, 2013.

[22] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10:24–45 (electronic), 2006.

[23] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities.* Oxford University Press, 2013.

[24] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford, 2013.

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

[26] 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.

[27] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 379–391. Springer, Berlin, 2006.

[28] P. Burman. Estimation of equifrequency histogram. *Statist. Probab. Lett.*, 56(3):227–238, 2002.

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

[30] G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.

[31] S. Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 12 2014.

[32] G. Claeskens and N. L. Hjort. *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2008.

[33] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[34] T. Dickhaus. *Simultaneous statistical inference*. Springer, Heidelberg, 2014. With applications in the life sciences.

[35] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[36] C. M. Hurvich and C.-L. Tsai. Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.*, 9(3):259–265, 1990.

[37] J. T. G. Hwang and Z. Zhao. Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *J. Amer. Statist. Assoc.*, 108(502):607–618, 2013.

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

[39] G. Lecué and M. Lerasle. Learning from MOM's principles: Le Cam's approach. 2017.

[40] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. Rejoinder: "a significance test for the lasso". *Ann. Statist.*, 42(2):518–531, 04 2014.

[41] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2):413–468, 04 2014.

[42] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann.Stat.*, 27:1808–1829, 1999.

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

[44] P. Massart and E. Nédélec. Risks bounds for statistical learning. *Ann.Stat.*, 34(5):2326–2366, 2006.

[45] T. Mildenberger and H. Weinert. The benchden package: Benchmark densities for nonparametric density estimation. *Journal of Statistical Software, Articles*, 46(14):1–14, 2012.

[46] A. Muro and S. van de Geer. Concentration behavior of the penalized least squares estimator. *arXiv preprint arXiv:1511.08698*, 2015.

[47] J. Qiu and J. T. G. Hwang. Sharp simultaneous confidence intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, 63(3):767–776, 2007.

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

[49] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 32, 28, 2005.

[50] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[51] C. 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.

[52] N. Sugiura. Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.

[53] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann.Stat.*, 32:135–166, 2004.

[54] S. van de Geer and M. J. Wainwright. On concentration for (regularized) empirical risk minimization. *Sankhya A*, 79(2):159–200, Aug 2017.

[55] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007.

[56] V. Zuber and K. Strimmer. High-dimensional regression and variable selection using CAR scores. *Stat. Appl. Genet. Mol. Biol.*, 10:Art. 34, 29, 2011.

A Further Proofs and Theoretical Results

A.1 Proofs Related to Section 3.1

Most of the arguments given in the proofs of this section are borrowed from Castellan [29]. We essentially rearrange these arguments in a more efficient way, thus obtaining better concentration bounds than in [29] (or [43], see also [23]).
We also set, for any $\varepsilon > 0$, the event $\Omega_m(\varepsilon)$ where some control of $\hat{f}_m$ in sup-norm is achieved,

$$\Omega_m(\varepsilon) = \left\{ \left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty \leq \varepsilon \right\} .$$

As we have the following formulas for the estimators and the projections of the target,

$$\hat{f}_m = \sum_{I \in \Lambda_m} \frac{P_n(I)}{\mu(I)} \mathbb{1}_I \quad , \quad f_m = \sum_{I \in \Lambda_m} \frac{P(I)}{\mu(I)} \mathbb{1}_I ,$$

we deduce that,

$$\left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty = \sup_{I \in \Lambda_m} \frac{|(P_n - P)(I)|}{P(I)} . \tag{53}$$

Hence, it holds $\Omega_m(\varepsilon) = \bigcap_{I \in m} \{|P_n(I) - P(I)| \leq \varepsilon P(I)\}$.

Let us state the main result of this section, concerning upper and lower bounds for the true and empirical excess risks of the histograms on each model. These bounds are optimal to the first order.

Before giving the proof of Theorem 3.1, the following lemma will be useful. It describes the consistency in sup-norm of the histogram estimators, suitably normalized by the projections of the target on each model.

Lemma A.1 *Let α, A_+ and A_Λ be positive constants. Consider a finite partition m of $\mathcal{Z}$, with cardinality D_m. Assume*

$$0 < A_\Lambda \leq D_m \inf_{I \in m} \{P(I)\} \quad \text{and} \quad 0 < D_m \leq A_+ \frac{n}{\ln(n+1)} \leq n . \tag{54}$$

Then by setting

$$R_n^\infty(m) = \sqrt{\frac{2(\alpha+1) D_m \ln(n+1)}{A_\Lambda n}} + \frac{(\alpha+1) D_m \ln(n+1)}{A_\Lambda n} , \tag{55}$$

we get

$$\mathbb{P}\left(\left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty \leq R_n^\infty(m) \right) \geq 1 - 2(n+1)^{-\alpha} . \tag{56}$$

In other words, $\mathbb{P}(\Omega_m(R_n^\infty(m))) \geq 1 - 2(n+1)^{-\alpha}$. In addition, there exists a positive constant A_c, only depending on α, A_+ and A_Λ, such that $R_n^\infty(m) \leq A_c \sqrt{\frac{D_m \ln(n+1)}{n}}$. Furthermore, if

$$\frac{(\alpha+1) A_+}{A_\Lambda} \leq \tau = \sqrt{3} - \sqrt{2} < 0.32 ,$$

or if $n \geq n_0(\alpha, A_+, A_\Lambda)$, then $R_n^\infty(m) \leq 1/2$.

Proof of Lemma A.1. Let $\beta > 0$ to be fixed later. Recall that, by (53) we have

$$\left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty = \sup_{I \in \Lambda_m} \frac{|(P_n - P)(I)|}{P(I)} . \tag{57}$$

By Bernstein's inequality (see Proposition 2.9 in [43]) applied to the random variables $\mathbb{1}_{\xi_i \in I}$ we get, for all $x > 0$,

$$\mathbb{P}\left[\left|(P_n - P)(I)\right| \geq \sqrt{\frac{2P(I)\,x}{n}} + \frac{x}{3n} \right] \leq 2\exp(-x) \ .$$

Taking $x = \beta \ln(n+1)$ and normalizing by the quantity $P(I) > 0$ we get

$$\mathbb{P}\left[\frac{\left|(P_n - P)(I)\right|}{P(I)} \geq \sqrt{\frac{2\beta \ln(n+1)}{P(I)\,n}} + \frac{\beta \ln(n+1)}{P(I)\,3n} \right] \leq 2(n+1)^{-\beta} \ . \tag{58}$$

Now, by the first inequality in (54), we have $0 < P(I)^{-1} \leq A_\Lambda^{-1} D_m$. Hence, using (58) we get

$$\mathbb{P}\left[\frac{\left|(P_n - P)(I)\right|}{P(I)} \geq \sqrt{\frac{2\beta D_m \ln(n+1)}{A_\Lambda n}} + \frac{\beta D_m \ln(n+1)}{A_\Lambda n} \right] \leq 2(n+1)^{-\beta} \ , \tag{59}$$

We then deduce from (57) and (59) that

$$\mathbb{P}\left[\left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty \geq R_n^\infty(m) \right] \leq \frac{2D_m}{(n+1)^\beta}$$

and, since $D_m \leq n$, taking $\beta = \alpha + 1$ yields Inequality (56). The other facts of Lemma A.1 then follow from simple computations. ■

Proof of Theorem 3.1. Recall that α is fixed. By Inequality (22) in Proposition 4.2 applied with $\theta = R_n^\infty(m)$—where $R_n^\infty(m)$ is defined in (55) with our fixed value of α—and $x = \alpha \ln(n+1)$, it holds with probability at least $1 - (n+1)^{-\alpha}$,

$$\chi_n(m)\,\mathbf{1}_{\Omega_m(R_n^\infty(m))} \leq \sqrt{\frac{D_m}{n}} + \left(1 + \sqrt{2R_n^\infty(m)} + \frac{R_n^\infty(m)}{6}\right)\sqrt{\frac{2\alpha \ln(n+1)}{n}} \ . \tag{60}$$

As $R_n^\infty(m) \leq L_{\alpha,A_+,A_\Lambda}\sqrt{D_m \ln(n+1)/n} \leq L_{\alpha,A_+,A_\Lambda}$, (60) gives

$$\begin{aligned}\chi_n(m)\,\mathbf{1}_{\Omega_m(R_n^\infty(m))} &\leq \sqrt{\frac{D_m}{n}} + L_{\alpha,A_+,A_\Lambda}\sqrt{\frac{\ln(n+1)}{n}} \\ &= \sqrt{\frac{D_m}{n}}\left(1 + L_{\alpha,A_+,A_\Lambda}\sqrt{\frac{\ln(n+1)}{D_m}}\right) \ .\end{aligned} \tag{61}$$

We set the event Ω_0 on which we have

$$\left\| \frac{\hat{f}_m - f_m}{f_m} \right\|_\infty \leq R_n^\infty(m) \ ,$$

$$\chi_n(m) \leq \sqrt{\frac{D_m}{n}}\left(1 + L_{\alpha,A_+,A_\Lambda}\sqrt{\frac{\ln(n+1)}{D_m}}\right) \ , \tag{62}$$

and

$$\chi_n(m) \geq \left(1 - A_g\left(\sqrt{\frac{\ln(n+1)}{D_m}} \vee \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\right)\sqrt{\frac{D_m}{n}} \ .$$

In particular, $\Omega_0 \subset \Omega_m(R_n^\infty(m))$. By (61), Lemma A.1 and Proposition 4.3, it holds $\mathbb{P}(\Omega_0) \geq 1 - 4(n+1)^{-\alpha}$. It suffices to prove the inequalities of Theorem 3.1 on Ω_0. The following inequalities, between the excess risk on m and the chi-square statistics, are shown in [29] (Inequalities (2.13)). For any $\varepsilon \in (0,1)$, on $\Omega_m(\varepsilon)$,

$$\frac{1-\varepsilon}{2(1+\varepsilon)^2}\chi_n^2(m) \le \mathcal{K}\left(f_m, \hat{f}_m\right) \le \frac{1+\varepsilon}{2(1-\varepsilon)^2}\chi_n^2(m) \ . \tag{63}$$

Under the condition $(\alpha+1)A_+A_\Lambda^{-1} < \tau$, Proposition A.1 gives $R_n^\infty(m) < 1$. Hence, by applying the right-hand side of (63) with $\varepsilon = R_n^\infty(m) \le 1/2$, using (62) and the fact that $(1-\epsilon)^{-1} \le 1+2\epsilon$ for $\epsilon \le 1/2$, we get on Ω_0,

$$\begin{aligned}\mathcal{K}\left(f_m, \hat{f}_m\right) \le& \frac{1+R_n^\infty(m)}{2(1-R_n^\infty(m))^2}\chi_n^2(m)\\ \le& \left(\frac{1}{2} + L_{\alpha,A_+,A_\Lambda}\sqrt{\frac{D_m\ln(n+1)}{n}}\right)\frac{D_m}{n}\left(1+L_{\alpha,A_+,A_\Lambda}\sqrt{\frac{\ln(n+1)}{D_m}}\right)^2 .\end{aligned}$$

Then simple computations allow to get the right-hand side inequality in (16).

By applying the left-hand side of (63) with $\varepsilon = R_n^\infty(m) \le 1/2$ and using the fact that $(1+\epsilon)^{-1} \ge 1-\epsilon$, we also get on Ω_0,

$$\begin{aligned}\mathcal{K}\left(f_m, \hat{f}_m\right) \ge& \frac{1-R_n^\infty(m)}{2(1+R_n^\infty(m))^2}\chi_n^2(m)\\ \ge&(1-R_n^\infty(m))^3\left(1-A_g\left(\sqrt{\frac{\ln(n+1)}{D_m}} \vee \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\right)^2\frac{D_m}{2n}\\ \ge&(1-(3R_n^\infty(m)\wedge 1))\left(1-2A_g\left(\sqrt{\frac{\ln(n+1)}{D_m}} \vee \frac{\sqrt{\ln(n+1)}}{n^{1/4}}\right)\right)\frac{D_m}{2n}\end{aligned}$$

The left-hand side inequality in (16) then follows by simple computations, noticing in particular that $n^{-1/4}\sqrt{\ln(n+1)} \le \sqrt{\ln(n+1)/D_m} \vee \sqrt{D_m\ln(n+1)/n}$.

Inequalities in (17) follow from the same kind of arguments as those involved in the proofs of inequalities in (16). Indeed, from [43, Lemma 7.24]—or [29, Lemma 2.3] —, it holds

$$\frac{1}{2}\int\left(\hat{f}_m \wedge f_m\right)\left(\ln\frac{\hat{f}_m}{f_m}\right)^2 d\mu \le \mathcal{K}\left(\hat{f}_m, f_m\right) \le \frac{1}{2}\int\left(\hat{f}_m \vee f_m\right)\left(\ln\frac{\hat{f}_m}{f_m}\right)^2 d\mu \ .$$

We deduce that for $\varepsilon \in (0,1)$, we have on $\Omega_m(\varepsilon)$,

$$\frac{1-\varepsilon}{2}\int f_m\left(\ln\frac{\hat{f}_m}{f_m}\right)^2 d\mu \le \mathcal{K}\left(\hat{f}_m, f_m\right) \le \frac{1+\varepsilon}{2}\int f_m\left(\ln\frac{\hat{f}_m}{f_m}\right)^2 d\mu \ . \tag{64}$$

As for every $x>0$, we have $(1\vee x)^{-1} \le (x-1)^{-1}\ln x \le (1\wedge x)^{-1}$, (64) leads by simple computations to the following inequalities

$$\frac{1-\varepsilon}{2(1+\varepsilon)^2}\chi_n^2(m) \le \mathcal{K}\left(\hat{f}_m, f_m\right) \le \frac{1+\varepsilon}{2(1-\varepsilon)^2}\chi_n^2(m) \ .$$

We thus have the same upper and lower bounds, in terms of the chi-square statistic $\chi_n^2(m)$, for the empirical excess risk as for the true excess risk. ■

A.2 Oracle inequalities and dimension guarantees

Using the notations of Section 3.2, we define the set of assumptions ($\mathbf{SA}_0$) to be the conjunction of assumptions (**P1**), (**P2**), (**Asm**) and (**Alr**). The set of assumptions (**SA**) of Section 3.2 thus consists on assuming ($\mathbf{SA}_0$) together with (**Ap**).

For some of the following results, we will also need the following assumptions.

(P3) Richness of $\mathcal{M}_n$: there exist $m_0, m_1 \in \mathcal{M}_n$ such that $D_{m_0} \in [\sqrt{n}, c_{rich}\sqrt{n}]$ and $D_{m_1} \geq A_{rich} n \left(\ln(n+1)\right)^{-2}$.

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

$$\mathcal{K}\left(f_*, f_m\right) \leq C_+ D_m^{-\beta_+} \ .$$

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

$$C_- D_m^{-\beta_-} \leq \mathcal{K}\left(f_*, f_m\right) \leq C_+ D_m^{-\beta_+} \ .$$

Theorem 3.2 is a direct corollary of the following theorem.

Theorem A.2 *Take $n \geq 1$ and $r \in (0, p-1)$. Assume that the set of assumptions ($\boldsymbol{SA}_0$) holds and that for some $\theta \in (1/2, +\infty)$ and $\Delta > 0$,*

$$\operatorname{pen}(m) = \left(\theta + \Delta \varepsilon_n^+(m)\right) \frac{D_m}{n} \ ,$$

for every model $m \in \mathcal{M}_n$. Then there exists an event Ω_n of probability $1 - (n+1)^{-2}$ and some positive constants A_1, A_2 depending only on the constants defined in ($\boldsymbol{SA}_0$) such that, if $\Delta \geq (\theta - 1)_- A_1 + A_2 > 0$ then we have on Ω_n,

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq \frac{1 + 2(\theta-1)_+ + L_{(\boldsymbol{SA}),\theta,r} \left(\ln(n+1)\right)^{-1/2}}{1 - 2(\theta-1)_-} \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}\left(f_*, \hat{f}_m\right) \right\} + L_{(\boldsymbol{SA}),\theta,r} \frac{\left(\ln(n+1)\right)^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}} \ . \tag{65}$$

Assume furthermore that Assumption ($\boldsymbol{Ap}_u$) holds. Then it holds on Ω_n,

$$D_{\widehat{m}} \leq L_{(\boldsymbol{SA}),\Delta,\theta,r} n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \sqrt{\ln(n+1)} \quad , \quad D_{m_*} \leq L_{(\boldsymbol{SA})} n^{\frac{1}{1+\beta_+}} \ .$$

In particular, if we are in the case where $p < \beta_+$ then Inequality (65) reduces to

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq L_{(\boldsymbol{SA}),\theta,r} \frac{\left(\ln(n+1)\right)^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}} \ . \tag{66}$$

Grant Assumption ($\boldsymbol{Ap}$). Then it holds on Ω_n,

$$L^{(1)}_{\Delta,\theta,(\boldsymbol{SA})} \frac{n^{\frac{\beta_+}{\beta_-(1+\beta_+)}}}{\left(\ln(n+1)\right)^{\frac{1}{\beta_-}}} \leq D_{\widehat{m}} \leq L^{(2)}_{\Delta,\theta,(\boldsymbol{SA})} n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \sqrt{\ln(n+1)} \ ,$$

$$L^{(1)}_{(\boldsymbol{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq D_{m_*} \leq L^{(2)}_{(\boldsymbol{SA})} n^{\frac{1}{1+\beta_+}}$$

and

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq \frac{1 + 2(\theta-1)_+ + L_{(\boldsymbol{SA}),\theta,r} n^{-\frac{\beta_+}{(1+\beta_+)\beta_-}} \sqrt{\ln(n+1)}}{1 - 2(\theta-1)_-} \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}\left(f_*, \hat{f}_m\right) \right\} + L_{(\boldsymbol{SA}),\theta,r} \frac{\left(\ln(n+1)\right)^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}} \tag{67}$$

Furthermore, if $\beta_- < p\left(1+\beta_+\right)/\left(1+p+r\right)$ or $p/(1+r) > \beta_- + \beta_-/\beta_+ - 1$, then we have on Ω_n,

$$\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq \frac{1 + 2(\theta-1)_+ + L_{(\boldsymbol{SA}),\theta,r} \left(\ln(n+1)\right)^{-1/2}}{1 - 2(\theta-1)_-} \inf_{m \in \mathcal{M}_n} \left\{ \mathcal{K}\left(f_*, \hat{f}_m\right) \right\} \ . \tag{68}$$

We obtain in Theorem A.2 oracle inequalities and dimension bounds for the oracle and selected models.

In order to avoid cumbersome notations in the proofs of Theorem A.2, when generic constants L and n_0 depend on constants defined in the sets of assumptions $(\mathbf{SA}_0)$ or $(\mathbf{SA})$, we will note $L_{(\mathbf{SA})}$ and $n_0\left((\mathbf{SA})\right)$. The values of the these constants may change from line to line, or even whithin one line.

Proof of Theorem 3.2

- Proof of oracle Inequality (65):

From the definition of the selected model $\widehat{m}$ given in (4), $\widehat{m}$ minimizes

$$\operatorname{crit}(m) := P_n\left(\gamma(\hat{f}_m)\right) + \operatorname{pen}(m) \ ,$$

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

$$\operatorname{crit}'(m) := \operatorname{crit}(m) - P_n\left(\gamma\left(f_*\right)\right) \tag{69}$$

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

$$\begin{aligned}
\mathcal{K}\left(f_*, \hat{f}_m\right) &= P\left(\gamma(\hat{f}_m) - \gamma\left(f_*\right)\right) \\
&= P_n\left(\gamma(\hat{f}_m)\right) + P_n\left(\gamma(f_m) - \gamma(\hat{f}_m)\right) + \left(P_n - P\right)\left(\gamma\left(f_*\right) - \gamma(f_m)\right) \\
&\quad + P\left(\gamma(\hat{f}_m) - \gamma(f_m)\right) - P_n\left(\gamma\left(f_*\right)\right) \ .
\end{aligned}$$

By setting

$$\begin{aligned}
p_1(m) &= P\left(\gamma(\hat{f}_m) - \gamma(f_m)\right) = \mathcal{K}\left(f_m, \hat{f}_m\right) \ , \\
p_2(m) &= P_n\left(\gamma(f_m) - \gamma(\hat{f}_m)\right) = \mathcal{K}\left(\hat{f}_m, f_m\right) \ , \\
\bar{\delta}(m) &= \left(P_n - P\right)\left(\gamma\left(f_*\right) - \gamma(f_m)\right) = \left(P_n - P\right)\left(\ln\left(f_m / f_*\right)\right)
\end{aligned}$$

and

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

we have

$$\mathcal{K}\left(f_*, \hat{f}_m\right) = P_n\left(\gamma(\hat{f}_m)\right) + p_1(m) + p_2(m) + \bar{\delta}(m) - P_n\left(\gamma\left(f_*\right)\right)$$

and by (69),

$$\operatorname{crit}'(m) = \mathcal{K}\left(f_*, \hat{f}_m\right) + \left(\operatorname{pen}(m) - \operatorname{pen}'_{\mathrm{id}}(m)\right) \ . \tag{70}$$

As $\widehat{m}$ minimizes crit' over $\mathcal{M}_n$, it is therefore sufficient by (70) to control $\operatorname{pen}(m) - \operatorname{pen}'_{\mathrm{id}}(m)$ in terms of the excess risk $\mathcal{K}\left(f_*, \hat{f}_m\right)$, for every $m \in \mathcal{M}_n$, in order to derive oracle inequalities. We further set

$$\mathcal{K}_m = \mathcal{K}\left(f_*, f_m\right) \quad , \quad v_m = P\left[\left(\frac{f_m}{f_*} \vee 1\right)\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \quad \text{and} \quad w_m = P\left[\left(\frac{f_*}{f_m} \vee 1\right)^r\left(\ln\left(\frac{f_m}{f_*}\right)\right)^2\right] \ .$$

Let Ω_n be the event on which:

- For all models $m \in \mathcal{M}_n$, we set $z_n = (2 + \alpha_{\mathcal{M}})\ln(n+1) + 2\ln 2$ and we have,

$$\bar{\delta}(m) \leq \sqrt{\frac{2 v_m z_n}{n}} + \frac{2 z_n}{n} \tag{71}$$

$$-\bar{\delta}(m) \leq \sqrt{\frac{2 w_m z_n}{n}} + \frac{2 z_n}{n r} \tag{72}$$

$$-L_{(\mathbf{SA}),\alpha}\varepsilon_n^-(m)\frac{D_m}{2n} \leq p_1(m) - \frac{D_m}{2n} \leq L_{(\mathbf{SA}),\alpha}\varepsilon_n^+(m)\frac{D_m}{2n} \tag{73}$$

$$-L_{(\mathbf{SA}),\alpha}\varepsilon_n^-(m)\frac{D_m}{2n} \leq p_2(m) - \frac{D_m}{2n} \leq L_{(\mathbf{SA}),\alpha}\varepsilon_n^+(m)\frac{D_m}{2n} \tag{74}$$

By Theorem 3.1 applied with $\alpha = 5 + \alpha_{\mathcal{M}}$ and Propositions 4.4 and 4.5 applied with $z = z_n$, we get

$$\begin{aligned}\mathbb{P}\left(\Omega_n\right) &\geq 1 - \sum_{m \in \mathcal{M}_n} \left[4(n+1)^{-5-\alpha_{\mathcal{M}}} + \frac{(n+1)^{-2-\alpha_{\mathcal{M}}}}{2}\right] \\ &= 1 - \sum_{m \in \mathcal{M}_n} (n+1)^{-2-\alpha_{\mathcal{M}}} \geq 1 - (n+1)^{-2} \; .\end{aligned}$$

The following simple remark will be used along the proof: for any $m \in \mathcal{M}_n$, $z_n/n \leq L_{(\mathbf{SA})} \varepsilon_n^+ (m) \frac{D_m}{n}$. Notice also that $\varepsilon_n^- (\widehat{m}) \leq \varepsilon_n^+ (\widehat{m})$ (see Theorem 3.1).

By using (32), (70), (71), (73) and (74), we get that on Ω_n, for Δ of the form $(\theta - 1)_- L_{(\mathbf{SA})}^{(1)} + L_{(\mathbf{SA})}^{(2)}$ with $L_{(\mathbf{SA})}^{(1)}$ and $L_{(\mathbf{SA})}^{(2)}$ sufficiently large,

$$\begin{aligned}&\mathrm{crit}'\left(\widehat{m}\right) \\ \geq& \mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) + \mathrm{pen}\left(\widehat{m}\right) - p_1\left(\widehat{m}\right) - p_2\left(\widehat{m}\right) - \sqrt{\frac{2 v_{\widehat{m}} z_n}{n}} - \frac{2 z_n}{n} \\ \geq& \mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) + (\theta - 1) \frac{D_{\widehat{m}} - 1}{n} + \left(\Delta - L_{(\mathbf{SA})}\right) \varepsilon_n^+ \left(\widehat{m}\right) \frac{D_{\widehat{m}}}{n} - \sqrt{\frac{2 A_{MR,-} \mathcal{K}_{\widehat{m}}^{1-1/p} z_n}{n}} \\ \geq& \mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) - 2\left(\theta - 1\right)_- \left(\frac{D_{\widehat{m}} - 1}{2n} - L_{(\mathbf{SA})} \varepsilon_n^- \left(\widehat{m}\right) \frac{D_{\widehat{m}}}{n}\right) \\ &+ \left(\Delta - \left(L_{(\mathbf{SA})}^{(1)} \left(\theta - 1\right)_- + L_{(\mathbf{SA})}^{(2)}\right)\right) \varepsilon_n^+ \left(\widehat{m}\right) \frac{D_{\widehat{m}}}{n} - \sqrt{\frac{2 A_{MR,-} \mathcal{K}_{\widehat{m}}^{1-1/p} z_n}{n}} \\ \geq& \left(1 - 2\left(\theta - 1\right)_-\right) \mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) - \sqrt{\frac{2 A_{MR,-} \mathcal{K}_{\widehat{m}}^{1-1/p} z_n}{n}} \; . \qquad (75)\end{aligned}$$

Note that $1 - 2\left(\theta - 1\right)_- > 0$. Let us take $\eta \in \left(0, 1/2 - (\theta - 1)_-\right)$, so that

$$1 - 2\left(\theta - 1\right)_- - \eta > 1/2 - \left(\theta - 1\right)_- > 0 \; .$$

By Lemma A.5 applied with $a = \left(\eta \mathcal{K}_{\widehat{m}}\right)^{\frac{p-1}{2p}}$, $b = \eta^{-\frac{p-1}{2p}} \sqrt{2 A_{MR,-} z_n / n}$, $u = \frac{2p}{p-1}$ and $v = \frac{2p}{p+1}$, we have

$$\sqrt{\frac{2 A_{MR,-} \mathcal{K}_{\widehat{m}}^{1-1/p} z_n}{n}} \leq \eta \mathcal{K}_{\widehat{m}} + L_{(\mathbf{SA}),\alpha} \left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1}} \left(\frac{1}{\eta}\right)^{\frac{p-1}{p+1}} \; .$$

By using the latter inequality in (75) we obtain,

$$\mathrm{crit}'\left(\widehat{m}\right) \geq \left(1 - 2\left(\theta - 1\right)_- - \eta\right) \mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) - L_{(\mathbf{SA}),\alpha} \left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1}} \left(\frac{1}{\eta}\right)^{\frac{p-1}{p+1}} \; . \qquad (76)$$

We compute now an upper bound on crit′ for each model m. By Lemma A.5 applied with $a = \left(\eta \mathcal{K}_m\right)^{\frac{p-r-1}{2p}}$, $b = \eta^{-\frac{p-1-r}{2p}} \sqrt{2 A_{MR,-} z_n / n}$, $u = \frac{2p}{p-1-r}$ and $v = \frac{2p}{p+1+r}$, we have

$$\sqrt{\frac{2 A_{MR,-} \mathcal{K}_m^{1-\frac{r+1}{p}} z_n}{n}} \leq \eta \mathcal{K}_m + L_{(\mathbf{SA}),r} \left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1+r}} \left(\frac{1}{\eta}\right)^{\frac{p-1-r}{p+1+r}} \; .$$

By (72), (70), (73), (74), (33) and by using Lemma A.5 we have on Ω_n,

$$
\begin{aligned}
\operatorname{crit}'(m) &= \mathcal{K}\left(f_*, \hat{f}_m\right) + \operatorname{pen}(m) - p_1(m) - p_2(m) - \bar{\delta}(m) \\
&\leq \mathcal{K}\left(f_*, \hat{f}_m\right) + 2(\theta-1)_+ \left(\frac{D_m}{2n} - L_{(\mathbf{SA})}\varepsilon_n^-(m)\frac{D_m}{n}\right) \\
&\quad + \left(\Delta + (\theta-1)_+ L^{(1)}_{(\mathbf{SA})} + L^{(2)}_{(\mathbf{SA})}\right)\varepsilon_n^+(m)\frac{D_m}{n} + \sqrt{\frac{2A_{MR,-}\mathcal{K}_m^{1-\frac{r+1}{p}} z_n}{n}} + \frac{2z_n}{nr} \\
&\leq \left(1 + 2(\theta-1)_+ + \eta\right)\mathcal{K}\left(f_*, \hat{f}_m\right) + \left(\Delta + (\theta-1)_+ L^{(1)}_{(\mathbf{SA})} + L^{(2)}_{(\mathbf{SA}),r}\right)\varepsilon_n^+(m)\frac{D_m}{n} \\
&\quad + L_{(\mathbf{SA}),r}\left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1+r}}\left(\frac{1}{\eta}\right)^{\frac{p-1-r}{p+1+r}} .
\end{aligned}
$$

Recall that we took $\Delta = L^{(1)}_{(\mathbf{SA})}(\theta-1)_- + L^{(2)}_{(\mathbf{SA})}$ for some positive constants sufficiently large, so

$$
\Delta + (\theta-1)_+ L^{(1)}_{(\mathbf{SA})} + L^{(2)}_{(\mathbf{SA}),r} \leq |\theta-1|\, L^{(1)}_{(\mathbf{SA})} + L^{(2)}_{(\mathbf{SA}),r}
$$

and we finally get,

$$
\begin{aligned}
\operatorname{crit}'(m) &\leq \left(1 + 2(\theta-1)_+ + \eta\right)\mathcal{K}\left(f_*, \hat{f}_m\right) + \left(|\theta-1|\, L^{(1)}_{(\mathbf{SA})} + L^{(2)}_{(\mathbf{SA}),r}\right)\varepsilon_n^+(m)\frac{D_m}{n} \\
&\quad + L_{(\mathbf{SA}),r}\left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1+r}}\left(\frac{1}{\eta}\right)^{\frac{p-1-r}{p+1+r}} .
\end{aligned}
\tag{77}
$$

Now, as $\widehat{m}$ minimizes crit' we get from (76) and (77), on Ω_n,

$$
\begin{aligned}
\mathcal{K}\left(f_*, \hat{f}_{\widehat{m}}\right) \leq \frac{1 + 2(\theta-1)_+ + \eta}{1 - 2(\theta-1)_- - \eta}&\left(\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) + L_{(\mathbf{SA}),\theta,r}\varepsilon_n^+(m_*)\frac{D_{m_*}}{n}\right) \\
&+ L_{(\mathbf{SA}),\theta,r}\left(\frac{\ln(n+1)}{n}\right)^{\frac{p}{p+1+r}}\left(\frac{1}{\eta}\right)^{\frac{p-1-r}{p+1+r}} .
\end{aligned}
\tag{78}
$$

We distinguish two cases. If $D_{m_*} \geq L_{(\mathbf{SA})}(\ln(n+1))^2$ with a constant $L_{(\mathbf{SA})}$ chosen such that

$$
A_0\varepsilon_n^+(m_*) \leq \frac{1}{2\sqrt{\ln(n+1)}} \leq \frac{1}{2\sqrt{\ln 2}} < 1 ,
$$

where A_0 and $\varepsilon_n^+(m)$ are defined in Theorem 3.1, then by Theorem 3.1 it holds on Ω_n,

$$
\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq \left(1 - \frac{1}{2\sqrt{\ln(n+1)}}\right)\frac{D_{m_*}}{2n} .
$$

On the other hand, if $D_{m_*} \leq L_{(\mathbf{SA})}(\ln(n+1))^2$ then by Theorem 3.1 we have on Ω_n,

$$
\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) + L_{(\mathbf{SA}),\theta,r}\varepsilon_n^+(m_*)\frac{D_{m_*}}{n} \leq L_{(\mathbf{SA}),\theta,r}\frac{(\ln(n+1))^3}{n} .
$$

Hence, in any case we always have on Ω_n,

$$
\begin{aligned}
&\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) + L_{(\mathbf{SA}),\theta,r}\varepsilon_n^+(m_*)\frac{D_{m_*}}{n} \\
\leq &L_{(\mathbf{SA}),\theta,r}\frac{(\ln(n+1))^3}{n} + \left(1 + L_{(\mathbf{SA}),\theta,r}(\ln(n+1))^{-1/2}\right)\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) .
\end{aligned}
$$

By taking $\eta = (\ln(n+1))^{-1/2}\left(1/2 - (\theta-1)_-\right)$ and using the fact that in this case,

$$\frac{1+2(\theta-1)_+ + \eta}{1-2(\theta-1)_- - \eta} \leq \frac{1+2(\theta-1)_+ + L_\theta \eta}{1-2(\theta-1)_-} \ ,$$

we deduce that Inequality (78) gives,

$$\mathcal{K}\left(f_*, \hat{f}_{\hat{m}}\right) \leq \frac{1+2(\theta-1)_+ + L_{(\mathbf{SA}),\theta,r}(\ln(n+1))^{-1/2}}{1-2(\theta-1)_-} \inf_{m\in\mathcal{M}_n}\left\{\mathcal{K}\left(f_*, \hat{f}_m\right)\right\} + L_{(\mathbf{SA}),\theta,r}\frac{(\ln(n+1))^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}}$$

which is Inequality (65).

- Proof of Inequality (66):

By Lemma A.4 we know that $D_{m_*} \leq L_{(\mathbf{SA})} n^{\frac{1}{1+\beta_+}}$ on Ω_n. Furthermore, we have on Ω_n, by simple computations,

$$\begin{aligned}
\mathcal{K}\left(f_*, \hat{f}_m\right) &= \mathcal{K}(f_*, f_m) + \mathcal{K}\left(f_m, \hat{f}_m\right) \\
&\leq C_+ D_m^{-\beta_+} + \left(1 + L_{(\mathbf{SA})}\varepsilon_n^+(m)\right)\frac{D_m}{2n} \\
&\leq L_{(\mathbf{SA})}\left(D_m^{-\beta_+} + \frac{D_m}{n} + \frac{\ln(n+1)}{n}\right) \ .
\end{aligned}$$

This yields

$$\begin{aligned}
\inf_{m\in\mathcal{M}_n}\left\{\mathcal{K}\left(f_*, \hat{f}_m\right)\right\} &\leq L_{(\mathbf{SA})} \inf\left\{D_m^{-\beta_+} + \frac{D_m}{n} + \frac{\ln(n+1)}{n} \ ; \ m\in\mathcal{M}_n, \ D_m \leq L_{(\mathbf{SA})} n^{\frac{1}{1+\beta_+}}\right\} \\
&\leq L_{(\mathbf{SA})} n^{-\frac{\beta_+}{1+\beta_+}} \ .
\end{aligned}$$

To conclude, it suffices to notice that if $\beta_+ > p$ then

$$n^{-\frac{\beta_+}{1+\beta_+}} \leq n^{-\frac{p}{p+1}} \leq n^{-\frac{p}{p+1+r}} \ ,$$

which finally gives

$$\frac{1+2(\theta-1)_+ + L_{(\mathbf{SA}),\theta,r}(\ln(n+1))^{-1/2}}{1-2(\theta-1)_-} \inf_{M\in\mathcal{M}_n}\left\{\mathcal{K}\left(f_*, \hat{f}_m\right)\right\} \leq L_{(\mathbf{SA}),\theta,r}\frac{(\ln(n+1))^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}} \ .$$

- Proof of Inequality (67):

From **(Ap)**, we know by Lemma A.4 that there exist $L^{(1)}_{(\mathbf{SA})}, L^{(2)}_{(\mathbf{SA})} > 0$ such that

$$L^{(1)}_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq D_{M_*} \leq L^{(2)}_{(\mathbf{SA})} n^{\frac{1}{1+\beta_+}}$$

and so

$$\begin{aligned}
\varepsilon_n^+(m_*) &\leq \max\left\{\sqrt{\frac{D_{m_*}\ln(n+1)}{n}}; \sqrt{\frac{\ln(n+1)}{D_{m_*}}}; \frac{\ln(n+1)}{D_{m_*}}\right\} \\
&\leq L_{(\mathbf{SA})} \max\left\{n^{-\frac{\beta_+}{2(1+\beta_+)}}; n^{-\frac{\beta_+}{2(1+\beta_+)\beta_-}}\right\}\sqrt{\ln(n+1)} \\
&= L_{(\mathbf{SA})} n^{-\frac{\beta_+}{(1+\beta_+)\beta_-}}\sqrt{\ln(n+1)} \ .
\end{aligned} \tag{79}$$

Assume for now that we also have

$$A_0\varepsilon_n^+(m_*) \leq (\ln(n+1))^{-1/2} \ , \tag{80}$$

where A_0 and $\varepsilon_n^+(m)$ are defined in Theorem 3.1. Then by Theorem 3.1 it holds on Ω_n,

$$\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq \left(1 - (\ln(n+1))^{-1/2}\right) \frac{D_{m_*}}{2n} \ .$$

In this case, we deduce that

$$\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) + L_{(\mathbf{SA}),\theta,r}\varepsilon_n^+(m_*)\frac{D_{m_*}}{n} \leq \left(1 + L_{(\mathbf{SA}),\theta,r} n^{-\frac{\beta_+}{(1+\beta_+)\beta_-}} \sqrt{\ln(n+1)}\right) \mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \ , \tag{81}$$

and Inequality (67) simply follows from using Inequality (78).

If Inequality (80) is not satisfied, that is

$$A_0\varepsilon_n^+(m_*) > (\ln(n+1))^{-1/2} \ , \tag{82}$$

then by (79), this means that there exists a positive constant $L_{(\mathbf{SA})}$ such that

$$L_{(\mathbf{SA})} n^{-\frac{\beta_+}{(1+\beta_+)\beta_-}} \sqrt{\ln(n+1)} > (\ln(n+1))^{-1/2} \ .$$

Consequently, this ensures that in the case where (82) is true, we also have $n \leq n_0((\mathbf{SA}))$. Hence, as

$$\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq C_- D_{m_*}^{-\beta_-} \geq L_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} > 0 \ ,$$

this yields Inequality (81) with a positive constant $L_{(\mathbf{SA}),\theta,r}$ in the right-hand term sufficiently large and then the result easily follows from using Inequality (78).

- Proof of Inequality (68):

If $D_{m_*} \geq L_{(\mathbf{SA})}(\ln(n+1))^2$ with a constant $L_{(\mathbf{SA})}$ chosen such that

$$A_0\varepsilon_n^+(m_*) \leq 1/2 \ ,$$

where A_0 and $\varepsilon_n^+(m)$ are defined in Theorem 3.1, then by Theorem 3.1 it holds on Ω_n,

$$\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq C_- D_{m_*}^{-\beta_-} + \frac{D_{m_*}}{4n} \ .$$

By Lemma A.4 we know that $L^{(1)}_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq D_{m_*} \leq L^{(2)}_{(\mathbf{SA})} n^{\frac{1}{1+\beta_+}}$ on Ω_n. This gives

$$\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq L_{(\mathbf{SA})} n^{-\frac{\beta_-}{1+\beta_+}} + L_{(\mathbf{SA})} n^{-1+\frac{\beta_+}{(1+\beta_+)\beta_-}}$$

and we deduce by simple algebra that if $\beta_- < p(1+\beta_+)/(1+p+r)$ or $p/(1+r) > \beta_+/(\beta_-(1+\beta_+)) - 1$, then

$$L_{(\mathbf{SA})}(\ln(n+1))^{-1/2}\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq \frac{(\ln(n+1))^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}} \ .$$

On the other hand, if $D_{m_*} \leq L_{(\mathbf{SA})}(\ln(n+1))^2$, this implies in particular

$$L^{(1)}_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq L_{(\mathbf{SA})}(\ln(n+1))^2 \ .$$

Hence, there exists an integer $n_0((\mathbf{SA}))$ such that $n \leq n_0((\mathbf{SA}))$. In this case, we can find a constant $L_{(\mathbf{SA})}$ such that

$$L_{(\mathbf{SA})}(\ln(n+1))^{-1/2}\mathcal{K}\left(f_*, \hat{f}_{m_*}\right) \geq \frac{(\ln(n+1))^{\frac{3p-1-r}{2(p+1+r)}}}{n^{\frac{p}{p+1+r}}}$$

and through the use of inequality (67), this conclude the proof of Inequality (68).

Lemma A.3 (Control on the dimension of the selected model) *Assume that* $(\boldsymbol{SA}_0)$ *holds together with* $(\boldsymbol{Ap}_u)$*. If* $\beta_+ \leq \frac{p}{r+1}$ *then, on the event* Ω_n *defined in the proof of Theorem A.2, we have*

$$D_{\widehat{m}} \leq L_{\Delta,\theta,r,(\boldsymbol{SA})} n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \sqrt{\ln(n+1)} \ . \tag{83}$$

If moreover $(\boldsymbol{Ap})$ *holds, then we get on the event* Ω_n*,*

$$L^{(1)}_{\Delta,\theta,(\boldsymbol{SA})} \frac{n^{\frac{\beta_+}{\beta_-(1+\beta_+)}}}{(\ln(n+1))^{\frac{1}{\beta_-}}} \leq D_{\widehat{m}} \leq L^{(2)}_{\Delta,\theta,(\boldsymbol{SA})} n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \sqrt{\ln(n+1)} \ . \tag{84}$$

Lemma A.4 (Control over the dimension of oracle models) *Assume that* $(\boldsymbol{SA}_0)$ *holds together with* $(\boldsymbol{Ap}_u)$*. We have on the event* Ω_n *defined in the proof of Theorem A.2,*

$$D_{m_*} \leq L_{(\boldsymbol{SA})} n^{\frac{1}{1+\beta_+}} \ .$$

If moreover $(\boldsymbol{Ap})$ *holds, then we get on the event* Ω_n*,*

$$L^{(1)}_{(\boldsymbol{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq D_{m_*} \leq L^{(2)}_{(\boldsymbol{SA})} n^{\frac{1}{1+\beta_+}} \ . \tag{85}$$

Proof of Lemma A.3. Recall that $\widehat{m}$ minimizes

$$\mathrm{crit}'(m) = \mathrm{crit}(m) - P_n\gamma(f_*) = \mathcal{K}_m - p_2(m) - \bar{\delta}(m) + \mathrm{pen}(m)$$

over the models $m \in \mathcal{M}_n$. Moreover, $\mathrm{pen}(m) = (\theta + \Delta\varepsilon_n^+(m)) D_m/n$. The analysis is restricted on Ω_n.

1. Upper bound on $\mathrm{crit}'(m)$:

$$p_2(m) \geq \left(\frac{1}{2} - L_{(\mathbf{SA})}\varepsilon_n^+(m)\right)\frac{D_m}{n}$$
$$-\bar{\delta}(m) \leq \sqrt{\frac{2w_m z_n}{n}} + \frac{2z_n}{nr} \ .$$

Moreover, by Lemma 4.6, we have $w_m \leq A_{MR,-}\mathcal{K}_m^{1-\frac{r+1}{p}}$ and so,

$$\begin{aligned}\mathrm{crit}'(m) \leq& \mathcal{K}_m + \left(\theta - \frac{1}{2} + L_{\Delta,(\mathbf{SA})}\varepsilon_n^+(m)\right)\frac{D_m}{n} + \sqrt{\frac{2A_{MR,-}\mathcal{K}_m^{1-\frac{r+1}{p}} z_n}{n}} \\ \leq& L_{\Delta,\theta,(\mathbf{SA})}\left(D_m^{-\beta_+} + \frac{D_m}{n} + \frac{\ln(n+1)}{n} + \sqrt{\frac{D_m^{-\beta_+\left(1-\frac{r+1}{p}\right)}\ln(n+1)}{n}}\right) \ .\end{aligned}$$

Now, if $\beta_+ \leq \frac{p}{r+1}$, then for m_0 such that $D_{m_0} = \left\lceil n^{\frac{1}{1+\beta_+}} \right\rceil$ we have

$$\frac{D_{m_0}}{n} \leq 2n^{-\frac{\beta_+}{1+\beta_+}} \ ; \ D_{m_0}^{-\beta_+} \leq n^{-\frac{\beta_+}{1+\beta_+}} \ ; \ \sqrt{\frac{D_{m_0}^{-\beta_+\left(1-\frac{r+1}{p}\right)}}{n}} \leq n^{-\frac{\beta_+\left(2-\frac{r+1}{p}\right)+1}{2(1+\beta_+)}} \leq n^{-\frac{\beta_+}{1+\beta_+}} \ ,$$

so we get

$$\mathrm{crit}'(m_0) \leq L_{\Delta,\theta,(\mathbf{SA})} n^{-\frac{\beta_+}{1+\beta_+}} \sqrt{\ln(n+1)} \ . \tag{86}$$

Otherwise, if $\beta_+ > \frac{p}{r+1}$, then for m_1 such that $D_{m_1} = \left\lceil n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \right\rceil$, we have

$$\frac{D_{m_1}}{n} \leq 2n^{-\frac{1+\beta_+\left(1-\frac{r+1}{p}\right)}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \; ; \; \sqrt{\frac{D_{m_1}^{-\beta_+\left(1-\frac{r+1}{p}\right)}}{n}} \leq n^{-\frac{1+\beta_+\left(1-\frac{r+1}{p}\right)}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \; ;$$

$$D_{m_1}^{-\beta_+} \leq n^{-\frac{\beta_+}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \leq n^{-\frac{1+\beta_+\left(1-\frac{r+1}{p}\right)}{2+\beta_+\left(1-\frac{r+1}{p}\right)}}$$

and

$$\mathrm{crit}'(m_1) \leq L_{\Delta,\theta,(\mathbf{SA})} n^{-\frac{1+\beta_+\left(1-\frac{r+1}{p}\right)}{2+\beta_+\left(1-\frac{r+1}{p}\right)}} \sqrt{\ln(n+1)} \; . \tag{87}$$

2. Lower bound on $\mathrm{crit}'(m)$: we have

$$p_2(m) \leq \left(\frac{1}{2} + L_{(\mathbf{SA})} \varepsilon_n^+(m)\right) \frac{D_m}{n}$$

$$-\bar{\delta}(m) \geq -\sqrt{\frac{2v_m z_n}{n}} - \frac{2z_n}{n} \; .$$

Moreover, by Lemma 4.6, we have for some constant $A_{MR,-} > 0$, $v_m \leq A_{MR,-} \mathcal{K}_m^{1-\frac{1}{p}}$. For Δ large enough we thus get,

$$\mathrm{crit}'(m) \geq \mathcal{K}_m + \left(\theta - \frac{1}{2}\right) \frac{D_m}{n} - \sqrt{\frac{2A_{MR,-}\mathcal{K}_m^{1-\frac{1}{p}} z_n}{n}} \tag{88}$$

Assume that $\beta_+ \leq \frac{p}{r+1}$. We take $D_m \leq L\left(n/\ln(n+1)\right)^{\frac{p}{(p+1)\beta_-}}$ for some constant $L > 0$. If L is small enough, we have by $(\mathbf{Ap})$ $\mathcal{K}_m \geq (8A_{MR,-}z_n/n)^{p/(p+1)}$ and by (88), $\mathrm{crit}'(m) \geq \mathcal{K}_m/2$. Now if

$$D_m \leq L\left(\left(\frac{n}{\ln(n+1)}\right)^{\frac{p}{(1+p)\beta_-}} \wedge \frac{n^{\frac{\beta_+}{\beta_-(1+\beta_+)}}}{(\ln(n+1))^{\frac{1}{2\beta_-}}}\right) \tag{89}$$

with L sufficiently small, only depending on Δ, θ and constants in $(\mathbf{SA})$, then by (86) we obtain $\mathrm{crit}'(m) > \mathrm{crit}'(m_0)$. As $\beta_+ \leq \frac{p}{r+1} < p$, the upper bound in (89) reduces to $D_m \leq Ln^{\frac{\beta_+}{\beta_-(1+\beta_+)}} / (\ln(n+1))^{\frac{1}{\beta_-}}$. This proves the left-hand side of (84).
Assume now that $\beta_+ > \frac{p}{r+1}$. If

$$D_m \leq L\left(\left(\frac{n}{\ln(n+1)}\right)^{\frac{p}{(1+p)\beta_-}} \wedge \frac{n^{\frac{1+\beta_+\left(1-\frac{r+1}{p}\right)}{\beta_-\left(2+\beta_+\left(1-\frac{r+1}{p}\right)\right)}}}{(\ln(n+1))^{\frac{1}{2\beta_-}}}\right)$$

with L sufficiently small, only depending on Δ, θ and constants in $(\mathbf{SA})$, then by (87) we obtain $\mathrm{crit}'(m) > \mathrm{crit}'(m_1)$. As $\beta_+ \leq \frac{p}{r+1} < p$, the upper bound in (89) reduces to $D_m \leq Ln^{\frac{\beta_+}{\beta_-(1+\beta_+)}} / (\ln(n+1))^{\frac{1}{\beta_-}}$. This proves the left-hand side of (84).

Assume that $\beta_+ \leq \frac{p}{r+1}$. We take $D_m \leq L\left(n\ln(n+1)\right)^{\frac{1}{2+\beta_+\left(1-\frac{1}{p}\right)}}$ for some constant $L > 0$. If L is large enough, then we get by (88) and simple calculations, $\mathrm{crit}'(m) \geq (\theta/2 - 1/4)\, D_m/n$. Furthermore, if

$$D_m \geq L\left(\left(n\ln(n+1)\right)^{\frac{1}{2+\beta_+\left(1-\frac{1}{p}\right)}} \wedge n^{\frac{1}{1+\beta_+}} \sqrt{\ln(n+1)}\right) \tag{90}$$

with L sufficiently small, only depending on Δ, θ and constants in (**SA**), then by (86) we obtain $\mathrm{crit}'(m) > \mathrm{crit}'(m_0)$. As $\beta_+ \leq \frac{p}{r+1} < p$, the lower bound in (90) reduces to $D_m \leq L\left(n\ln(n+1)\right)^{\frac{1}{2+\beta_+\left(1-\frac{1}{p}\right)}}$. This proves the left-hand side of (84).

Assume now that $\beta_+ > \frac{p}{r+1}$. If

$$D_m \geq L\left(\left(n\ln(n+1)\right)^{\frac{1}{2+\beta_+\left(1-\frac{1}{p}\right)}} \vee n^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}}\sqrt{\ln(n+1)}\right) \tag{91}$$

with L sufficiently large, only depending on Δ, θ and constants in (**SA**), then by (86) we obtain $\mathrm{crit}'(m) > \mathrm{crit}'(m_1)$. As $\beta_+ \leq \frac{p}{r+1} < p$, the upper bound in (89) reduces to $D_m \leq Ln^{\frac{\beta_+}{\beta_-(1+\beta_+)}} / \left(\ln(n+1)\right)^{\frac{1}{\beta_-}}$. This proves the left-hand side of (84). As $r > 0$, (91) reduces to $D_m \geq Ln^{\frac{1}{2+\beta_+\left(1-\frac{r+1}{p}\right)}}\sqrt{\ln(n+1)}$, which proves (83) and the right-hand side of (84).

Proof of Lemma A.4. By definition, m_* minimizes

$$\mathcal{K}\left(f_*, \hat{f}_m\right) = \mathcal{K}_m + p_1(m)$$

over the models $m \in \mathcal{M}_n$. The analysis is restricted on Ω_n.

1. Upper bound on $\mathcal{K}\left(f_*, \hat{f}_m\right)$: we have

$$\begin{aligned}\mathcal{K}\left(f_*, \hat{f}_m\right) &\leq C_+ D_m^{-\beta_+} + \left(\frac{1}{2} + L_{(\mathbf{SA})}\varepsilon_n^+(m)\right)\frac{D_m}{n} \\ &\leq L_{(\mathbf{SA})}\left(D_m^{-\beta_+} + \frac{D_m}{n} + \frac{\ln(n+1)}{n}\right) .\end{aligned}$$

Hence, if m_0 is such that $D_{m_0} = n^{\frac{1}{1+\beta_+}}$, then

$$\mathcal{K}\left(f_*, \hat{f}_{m_0}\right) \leq L_{(\mathbf{SA})} n^{-\frac{\beta_+}{1+\beta_+}} . \tag{92}$$

2. Lower bound on $\mathcal{K}\left(f_*, \hat{f}_m\right)$: there exists a constant A_0, only depending on constants in (**SA**), such that

$$\begin{aligned}\mathcal{K}\left(f_*, \hat{f}_m\right) &\geq \mathcal{K}_m + \left(\frac{1}{2} - A_0\varepsilon_n^-(m)\right)\frac{D_m}{n} \\ &\geq \mathcal{K}_m + \frac{D_m}{2n} - A_0 \max\left\{\left(\frac{D_m}{n}\right)^{3/2}\sqrt{\ln(n+1)}; \frac{\sqrt{D_m\ln(n+1)}}{n}\right\} .\end{aligned} \tag{93}$$

If (**Ap**) holds, then for

$$D_m \leq L_{(\mathbf{SA})} \min\left\{\frac{n^{\frac{3}{3+2\beta_-}}}{\left(\ln(n+1)\right)^{\frac{1}{3+2\beta_-}}}; \frac{n^{\frac{2}{1+2\beta_-}}}{\left(\ln(n+1)\right)^{\frac{1}{1+2\beta_-}}}\right\}$$

with $L_{(\mathbf{SA})}$ sufficiently small, we have

$$\mathcal{K}_m/2 \geq C_- D_m^{-\beta_-}/2 \geq A_0 \max\left\{\left(\frac{D_m}{n}\right)^{3/2}\sqrt{\ln(n+1)}; \frac{\sqrt{D_m\ln(n+1)}}{n}\right\} .$$

In this case, we have by (93), $\mathcal{K}\left(f_*, \hat{f}_m\right) \geq \mathcal{K}_m/2 + (D_m)/2n \geq C_- D_m^{-\beta_-}/2$. Moreover, if m is such that $D_m \leq L_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}}$ with $L_{(\mathbf{SA})}$ sufficiently small, we also have

$$D_m \leq L_{(\mathbf{SA})} n^{\frac{\beta_+}{(1+\beta_+)\beta_-}} \leq L_{(\mathbf{SA})} \min\left\{n^{\frac{\beta_+}{(1+\beta_+)\beta_-}}; \frac{n^{\frac{3}{3+2\beta_-}}}{\left(\ln(n+1)\right)^{\frac{1}{3+2\beta_-}}}; \frac{n^{\frac{2}{1+2\beta_-}}}{\left(\ln(n+1)\right)^{\frac{1}{1+2\beta_-}}}\right\}$$

and by (92) we get $\mathcal{K}\left(f_*, \hat{f}_{m_0}\right) < \mathcal{K}\left(f_*, \hat{f}_m\right)$, which gives the left-hand side of (85).
We turn now to the proof of the right-hand side of (85). Let $m \in \Updownarrow_n$ be such that $D_m \geq L_1 n^{\frac{1}{1+\beta_+}}$. By (92) we deduce that if L_1 is large enough, depending only on constants in (**SA**), then we have

$$\frac{D_m}{4n} > \mathcal{K}\left(f_*, \hat{f}_{m_0}\right) \ .$$

In addition, if $D_m \geq L_2 \left(\ln(n+1)\right)^2$ and $D_m \leq L_2^{-1} n / \ln(n+1)$ for some constant L_2 sufficiently large, then

$$\frac{D_m}{4n} \geq A_0 \max\left\{ \left(\frac{D_m}{n}\right)^{3/2} \sqrt{\ln(n+1)}; \frac{\sqrt{D_m \ln(n+1)}}{n} \right\}$$

and by (93), we deduce that $\mathcal{K}\left(f_*, \hat{f}_m\right) > \mathcal{K}\left(f_*, \hat{f}_{m_0}\right)$. The latter inequality implies that $D_{m_*} \leq L_1 n^{\frac{1}{1+\beta_+}}$. Our reasoning is valid if n is such that $L_2 \left(\ln(n+1)\right)^2 \leq L_1 n^{\frac{1}{1+\beta_+}} \leq L_2^{-1} n / \ln(n+1)$. At the price of enlarging L_1, we can always achieve $L_2 \left(\ln(n+1)\right)^2 \leq L_1 n^{\frac{1}{1+\beta_+}}$, with L_1 not depending on n. Then if $L_2^{-1} n / \ln(n+1) < L_1 n^{\frac{1}{1+\beta_+}}$, we still have

$$D_{m_*} \leq \max_{m \in \mathcal{M}_n} D_m \leq A_{\mathcal{M},+} \frac{n}{\left(\ln(n+1)\right)^2} \leq A_{\mathcal{M},+} L_2 L_1 n^{\frac{1}{1+\beta_+}} \ .$$

In every case, there exists $L > 0$ only depending on constants in (**SA**) such that $D_{m_*} \leq L n^{\frac{1}{1+\beta_+}}$.

Lemma A.5 *Let $(a,b) \in \mathbb{R}_+^2$ and $(u,v) \in [1, \infty]$ such that $1/u + 1/v = 1$. Then*

$$ab \leq \max\left\{a^u; b^v\right\} \leq a^u + b^v \ .$$

Proof. By symmetry, we can assume $a^u \geq b^v$. Then $b = \left(b^v\right)^{1/v} \leq a^{u/v} = a^{u-1}$ and so, $ab \leq a a^{u-1} = a^u \leq a^u + b^v$. ■