Modified Akaike's criterion for histogram density estimation

Gwénaëlle Castellan
Laboratoire de mathématiques - U.M.R. 8628
Université Paris XI, 91405 Orsay, France
gwenaelle.castellan@math.u-psud.fr

Abstract

We address the problem of choosing a histogram estimator from a sample of some unknown density. We consider regular as well as irregular partitions and propose a penalized maximum likelihood selection criterion. We define the penalty term involved in our criterion in order to minimize non asymptotically the Hellinger risk for the resulting penalized estimator of the unknown density. Our criterion appears either as a mild correction or as a substantial modification of Akaike's criterion, depending on the complexity of the collection of partitions to be considered.

AMS 1991 subject classifications: 62G05, 62G07.
Key words and phrases: Akaike's information criterion, Histogram, Model selection, Nonparametric density estimation, Penalization.

1 Introduction

We consider n independent identically distributed random variables X_1,...,X_n with values in some probability space $(\mathbb{X}, \mathcal{X}, \mu)$ and with common distribution $P = s\mu$. Typically, $\mathbb{X}$ could be the interval $[0, 1]$ and μ the Lebesgue measure on $[0, 1]$. In this case, one of the simplest estimation procedure is the histogram estimation. Given some partition m of intervals of $[0, 1]$, the histogram estimator is defined by

$$\hat{s}_m = \sum_{I \in m} \frac{P_n(I)}{\mu(I)} \mathbb{1}_I. \tag{1.1}$$

where P_n denotes the empirical probability measure associated with $X_1, ..., X_n$:

$$P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

for any function f. This estimator is known to be the maximum likelihood estimator over the class of densities which are piecewise constant on m (this set will be denoted by $\mathcal{S}_m$). The main problem is to choose an adequate partition m. This paper is devoted to the study of data driven procedures for choosing a "best" partition, based on penalized maximum likelihood model selection criteria.
If we consider a collection of partitions $\mathcal{M}$ and the Kullback-Leibler information $K(s, \hat{s}_m)$ as a loss function, an ideal partition would minimize the risk $\mathbb{E}[K(s, \hat{s}_m)]$ when m varies in $\mathcal{M}$. Of course, this "best" partition is unknown.

Akaike's Information Criterion (AIC) is a general penalized maximum likelihood model selection procedure, which, in the context of histogram estimation, can be written as

$$\mathrm{crit}\,(m) = P_n\,(-\log \hat{s}_m) + \mathrm{pen}\,(m)$$

where $\mathrm{pen}\,(m) = (|m| - 1)\,/n$. Akaike's heuristics, as described in Akaike (1973), is based on the hope that the minimizer $\hat{m}$ of $\mathrm{crit}\,(m)$ over $\mathcal{M}$ leads to an estimator $\tilde{s} = \hat{s}_{\hat{m}}$ which has the same performance as the best estimator $\hat{s}_m$ (in terms of Kullback-Leibler risk), at least asymptotically. It is essential to notice that Akaike's heuristics is fundamentally asymptotic in the sense that $\mathcal{M}$ is a given finite collection and n goes to infinity. Of course, such a restriction on the collection of partitions is not realistic for histogram estimation. Indeed, in this case, one of the first problem that one wants to study is the optimal cell width selection for regular histograms, which corresponds to the case where $\mathcal{M}$ is taken as the collection of regular partitions. Let us first describe the state of the art on this problem. Peter Hall (1990) has studied some properties of AIC. More precisely, he shows that the optimal cell width $\hat{h}_n$ (in the sense of minimizing AIC) is equivalent to the cell width obtained by the Kulback-Leibler cross validation at least for sufficiently smooth and compactly supported densities. Taylor (1987) analyzes the asymptotic behaviour of $\hat{h}_n$. He shows that $\hat{h}_n$ is a.s. equivalent to $n^{-1/3}$ within a factor which depends on s. Note that $n^{-1/3}$ corresponds to the asymptotic order of the optimal cell width in the sense of the minimization of the risk with respect to various loss functions, see Taylor (1987) for Kullback-Leibler information, Kanazawa (1993) for Hellinger, Devroye and Györfi (1985) for $\mathbb{L}_1$, Freedman and Diaconis (1981) and Scott (1979) for $\mathbb{L}_2$. From a rather different perspective, Yu and Speed (1992) has looked at another model selection procedure called minimum description lenght based on Rissanen's approach. Up to our knowledge, the comparison of the risk $\mathbb{E}[K\,(s,\tilde{s})]$ with $\inf_{m\in\mathcal{M}} \mathbb{E}[K\,(s,\hat{s}_m)]$ has not yet been treated while it is known (Rudemo 1982) that $\mathbb{E}\left(\|s-\hat{s}\|_2^2\right)$ is equivalent to $\inf_{m\in\mathcal{M}} \mathbb{E}\left(\|s-\hat{s}_m\|_2^2\right)$ when $\hat{s}$ is defined from the $\mathbb{L}_2$-unbiased cross validation procedure.
Much less is known about optimal selection of irregular histograms. In a first paper, Kanazawa (1988) built some irregular partition $\hat{m}_D$ with knots among the order statistics of the sample when the number of pieces D is given. He showed that this partition is asymptotically optimal for the Hellinger risk, under some restrictives conditions on the cell width. In a second paper (Kanazawa 1992), he studied the optimal number of pieces: the one that minimizes the Hellinger risk between the true density s and the estimator $\hat{s}_{m_D}$, he showed that this number is asymptotically proportional to $n^{1/3}$, if s is smooth.

In a multidimensinnal framework, Lugosi and Nobel (1996) present sufficient conditions for the almost sure $\mathbb{L}^1$-consistency of histograms density estimates based on data-dependent partitions. Their results are based on the well-konwn inequality of Vapnick and Chervonenkis.

Barron et al. (1999) give some general risk bounds for nonparametric estimation via model selection. Their approach allows the number as well as the dimension of the models to depend on n, and their bounds are non asymptotic. They prove that, with an adequate choice of the penalty function, the resulting penalized maximum likelihood estimator realizes the best trade-off between the bias term and the variance term. But, in the framework of density estimation, the form of the penalty that they provide for general families of linear models involves unknown constants.

In this paper, we first study the risk of a histogram estimator $\hat{s}_m$. We prove that this risk is the sum of a bias term equal to $K\,(s,\mathcal{S}_m)$ (the Kullback-Leibler distance between s and the model $\mathcal{S}_m$) and of a variance term proportional to D_m/n (where $D_m = |m| - 1$).

Next, we consider some family of histogram estimators $\{\hat{s}_m,\, m \in \mathcal{M}_n\}$, and we try to get good constants for the penalty function. We consider the regular and the irregular case. The difference between these two cases is the complexity of the family of partitions, $\mathcal{M}_n$: for the regular case, $\mathcal{M}_n$ contains only one partition for a given number of pieces whereas for the irregular case, many partitions of the family could have the same number of pieces. The complexity of the family of partitions is taken into account in the choice of the penalty function. Indeed, we introduce some family of weights $\{L_m,\, m \in \mathcal{M}_n\}$ such that $\sum_{m \in \mathcal{M}_n} \exp(-L_m D_m) \leq \Sigma$ for some constant Σ independent of n, and we define the penalty function as

$$\mathrm{pen}\,(m) \geq c_1 \left(1 + \sqrt{c_2 L_m}\right)^2 \frac{D_m}{n}$$

where c_1, c_2 are positive constants which do not depend on s. We prove that any constant $c_1 > 1/2$ is allowed to get a penalized maximum likelihood estimator $\tilde{s} = \hat{s}_{\hat{m}}$, the risk of which satisfies:

$$\mathbb{E}\left[h^2\,(s, \tilde{s})\right] \leq C\,(c_1) \inf_{m \in \mathcal{M}_n} \left[K\,(s, \mathcal{S}_m) + \mathrm{pen}\,(m)\right] + \frac{C\,(\rho, \Sigma, B)}{n}$$

where h^2 denotes the squared Hellinger distance. To obtain such a risk bound, we require that there exists some positive constants ρ, B such that $s > \rho > 0$ and that for all $m \in \mathcal{M}_n$, $\mu\,(I) \geq B\,(\log n)^2/n$, for all $I \in m$. This result recovers the AIC in the particular case of a family of regular partitions. We also provide some optimal value for the constant c_1 in order to get a better risk bound. The proofs of our results relie upon the concentration inequalities for suprema of bounded empirical processes due to Talagrand (1996) and recently refined in terms of constants by Massart (1998). Furthermore, some simulations of the different values of the penalty function confirmed our results (Birgé and Rozenholc 1999).

The organization of this paper is as follows: Section 2 is devoted to the presentation of the statistical framework of histogram estimation and, in particular, gives some evaluation of the risk of a histogram estimator with respect to Kullback-Leibler or Hellinger loss function. Section 3 presents our main results which are proved in section 5, and Section 4 contains the exponential bounds used in the demonstrations.

2 The statistical framework

2.1 Histograms

Definition 2.1. *Given some probability space $(\mathbb{X}, \mathcal{X}, \mu)$ and a measurable partition m of $\mathbb{X}$, one defines $\mathcal{S}_m$ to be the set of densities with respect to μ which are piecewise constant on the partition m,*

$$\mathcal{S}_m = \left\{ \sum_{I \in m} \frac{a_I}{\mu(I)} \mathbb{1}_I \text{ with } a_I \geq 0, \text{ for all } I \in m \text{ and } \sum_{I \in m} a_I = 1 \right\}. \tag{2.2}$$

We shall call $\mathcal{S}_m$ the histogram model based on m and denote by $D_m + 1$ the number of pieces of m, hence, D_m represents the dimension of the model $\mathcal{S}_m$. We also introduce the linear subspace S_m of $\mathbb{L}^2\,(\mu)$ defined by

$$S_m = \mathrm{Span}\,\{\mathbb{1}_I,\ I \in m\}\,.$$

We consider the log-likelihood criterion γ_n defined by

$$\gamma_n(t) = -\frac{1}{n} \sum_{i=1}^{n} \log t(X_i) = -P_n\,(\log t)$$

for every density t with respect to μ and we recall that $\hat{s}_m$ minimizes γ_n over $\mathcal{S}_m$. More precisely, $\hat{s}_m$ is unique and is characterized by the equations:

$$P_n(f) = \int f \hat{s}_m \, d\mu \ , \quad \text{for all } \ f \in S_m. \tag{2.3}$$

2.2 The loss functions

2.2.1 The Kullback-Leibler loss

Let P, resp Q, two probability distributions with density p, resp. q, with respect to μ, we denote the Kullback-Leibler information by

$$K(p,q) = \begin{cases} \int_{\mathbb{X}} p \log \frac{p}{q} \, d\mu & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

The Kullback-Leibler information is naturally linked to the maximum likelihood estimation procedure since

$$\mathbb{E}[\gamma_n(t) - \gamma_n(s)] = K(s,t).$$

This is the reason why we shall use it as a loss function when analyzing the performances of histogram estimators in terms of risk bounds.
It is well known that $K(p,q)$ is nonnegative and equal to zero if and only if $p = q$ a.e. The Kullback-Leibler information has a tendency to behave like the square of a Hilbertian distance even if, of course, it is not strictly speaking the square of a distance. This idea has been developed in some general framework by Csiszár (1975). In our context, it can be very simply explicited. Indeed, one can consider the information projection of s onto $\mathcal{S}_m$ to be the minimizer s_m of $K(s,t)$ when $t \in \mathcal{S}_m$. It is characterized by the equations:

$$P(f) = \int f s_m d\mu \ , \quad \text{for all } \ f \in S_m \tag{2.4}$$

and is given by

$$s_m = \sum_{I \in m} \frac{P(I)}{\mu(I)} \mathbb{1}_I. \tag{2.5}$$

In the sequel, we shall denote the distance between the density s and the model $\mathcal{S}_m$ by $K(s, \mathcal{S}_m)$, and by definition, $K(s, \mathcal{S}_m) = K(s, s_m)$. Furthermore, for any $t \in \mathcal{S}_m$

$$K(s,t) = \int s \log \frac{s}{s_m} \, d\mu + \int s \log \frac{s_m}{t} \, d\mu$$

and since the function $\log(s_m/t)$ belongs to S_m, (2.4) leads to the Pythagore type identity

$$K(s,t) = K(s, s_m) + K(s_m, t)$$

In particular, the following decomposition is available

$$K(s, \hat{s}_m) = K(s, s_m) + K(s_m, \hat{s}_m) \tag{2.6}$$

where $K(s, s_m)$ represents some approximation error and $K(s_m, \hat{s}_m)$ represents some estimation error within the model $\mathcal{S}_m$.

2.2.2 The Hellinger distance

Since the Kullback-Leibler information is unbounded, it may turn to be difficult to handle. This is the reason why we shall also need to use the squared Hellinger distance as a loss function. The squared Hellinger distance is denoted by

$$h^2(p,q) = \frac{1}{2} \int_{\mathbb{X}} (\sqrt{p} - \sqrt{q})^2 \, d\mu$$

for all probability densities p and q with respect to μ. It is well known that

$$K(p,q) \geq 2h^2(p,q) \tag{2.7}$$

and that when $\log(p/q)$ is not too large, $K(p,q)$ really behaves like $h^2(p,q)$, see Lemma 4.4 in Birgé (1983).

2.3 The Risk of the histogram estimator on a given partition

Our purpose here is to evaluate the risk of the maximum likelihood estimator on the histogram model $\mathcal{S}_m$ with respect to Kullback-Leibler information in a first part, and with respect to squared Hellinger distance in a second part. The motivation for these computations is to provide a benchmark in order to judge the performance of adaptive histograms estimators that will be studied in the next section.

Let $\hat{s}_m$ be the histogram estimator given by (1.1) and s_m be the Kullback-Leibler projection given by (2.5) on $\mathcal{S}_m$. In order to get risk bounds for the histogram estimator $\hat{s}_m$, we shall use the following notation:

$$\Omega_m(\varepsilon) = \{|P_n(I) - P(I)| \leq \varepsilon P(I) \ , \ \text{for all } \ I \in m\} \tag{2.8}$$

defined for all $\varepsilon \in]0,1[$. We can derive from Bernstein's inequality, Pollard (1984) p. 193, a control of $\mathbb{P}[\Omega_m^c(\varepsilon)]$:

$$\mathbb{P}[|P_n(I) - P(I)| \geq \varepsilon P(I)] \leq 2\exp[-nP(I)C(\varepsilon)]$$

where $C(\varepsilon) = \varepsilon^2 [2(1+\varepsilon/3)]^{-1}$. Consequently, if

$$\rho = \operatorname{essinf}\{s(x)\,,\, x \in \mathbb{X}\} \quad \text{and} \quad \Lambda_m = \inf\{\mu(I)\,,\, I \in m\}$$

are assumed to be positive then

$$\mathbb{P}[\Omega_m^c(\varepsilon)] \leq 2\Lambda_m^{-1} \exp[-n\rho\Lambda_m C(\varepsilon)]\,. \tag{2.9}$$

Furthermore, if we assume that there exists some constant B such that $\Lambda_m \geq B(\log n)^2/n$ then there exists a constant $C(\varepsilon,\rho,B)$ depending on ε, ρ and B such that

$$\mathbb{P}[\Omega_m^c(\varepsilon)] \leq \frac{C(\varepsilon,\rho,B)}{n}. \tag{2.10}$$

The reason for introducing the event $\Omega_m(\varepsilon)$ is that, on this set, we can control the ratio $\hat{s}_m/s_m$ and by this way we can link the Kullback-Leibler information or the Hellinger distance between s and $\hat{s}_m$ to the following statistics,

Definition 2.2. *Given some partition m of $\mathbb{X}$, one defines the statistics $\chi_n^2(m)$ by*

$$\chi_n^2(m) = \sum_{I \in m} \frac{[P_n(I) - P(I)]^2}{P(I)}.$$

One can remark that $n\chi_n^2(m)$ is the standard Chi-square statistics. We are now in position to approximate the risk of the histogram estimator on a given partition.

2.3.1 Kullback-Leibler risk

The evaluation of Kullback-Leibler risk derives from the following lemma. This lemma is adapted from Lemma 1 of Barron and Sheu (1991) and will be used repeatedly in the sequel.

Lemma 2.3. *For all positive densities p and q*

$$\frac{1}{2}\int p\wedge q\left(\log\frac{p}{q}\right)^2 d\mu \leq K(p,q) \leq \frac{1}{2}\int p\vee q\left(\log\frac{p}{q}\right)^2 d\mu.$$

Proof. If we note $f=\log\frac{q}{p}$ then

$$K(p,q)=\int p\left(e^f-1-f\right)d\mu=\int p\,\phi(f)d\mu,$$

where ϕ is the function defined by $\phi(x)=e^x-1-x$, for all $x\in\mathbb{R}$. Then, we use the following inequality:

$$\frac{1}{2}x^2\,\left(1\wedge e^x\right)\leq\phi(x)\leq\frac{1}{2}x^2\,\left(1\vee e^x\right)$$

for all real number x. Thus

$$\frac{1}{2}\int f^2\left(1\wedge e^f\right)p\,d\mu\leq K(p,q)\leq\frac{1}{2}\int f^2\left(1\vee e^f\right)p\,d\mu$$

which leads to the result. □

This lemma allows to control the estimation error in the decomposition of the Kullback-Leibler information between the density s and the histogram estimator $\hat{s}_m$.
More precisely, applying this Lemma to the densities s_m and $\hat{s}_m$, one has, on the set $\Omega_m(\varepsilon)$ defined by (2.8)

$$\frac{1-\varepsilon}{2}\int s_m\left(\log\frac{s_m}{\hat{s}_m}\right)^2 d\mu\leq K(s_m,\hat{s}_m)\leq\frac{1+\varepsilon}{2}\int s_m\left(\log\frac{s_m}{\hat{s}_m}\right)^2 d\mu$$

since $(1-\varepsilon)\,s_m\leq\hat{s}_m\leq(1+\varepsilon)\,s_m$ on this set. Moreover,

$$\int s_m\left(\log\frac{s_m}{\hat{s}_m}\right)^2 d\mu=\int\frac{(\hat{s}_m-s_m)^2}{s_m}\left(\frac{\log\left(\hat{s}_m/s_m\right)}{\hat{s}_m/s_m-1}\right)^2 d\mu.$$

Thus, since

$$\frac{1}{1\vee x}\leq\frac{\log x}{x-1}\leq\frac{1}{1\wedge x},\quad\text{for all } x>0, \tag{2.11}$$

one has

$$\frac{1}{(1+\varepsilon)^2}\int\frac{(\hat{s}_m-s_m)^2}{s_m}d\mu\leq\int s_m\left(\log\frac{s_m}{\hat{s}_m}\right)^2 d\mu\leq\frac{1}{(1-\varepsilon)^2}\int\frac{(\hat{s}_m-s_m)^2}{s_m}d\mu.$$

According to Definition 2.2 we can write

$$\int\frac{(\hat{s}_m-s_m)^2}{s_m}d\mu=\chi_n^2\,(m)\,, \tag{2.12}$$

and we get on the set $\Omega_m(\varepsilon)$

$$\frac{1-\varepsilon}{2\,(1+\varepsilon)^2}\,\chi_n^2(m)\leq K\,(s_m,\hat{s}_m)\leq\frac{1+\varepsilon}{2\,(1-\varepsilon)^2}\,\chi_n^2(m). \tag{2.13}$$

This leads to the following proposition:

Proposition 2.4. *Let m be some partition of $\mathbb{X}$ with D_m+1 pieces, $\hat{s}_m$ be the histogram estimator defined by (1.1), and s_m be the Kullback-Leibler projection defined by (2.5).*
For all $\varepsilon \in]0,1[$, the Kullback-Leibler risk of $\hat{s}_m$ restricted to the set $\Omega_m(\varepsilon)$ defined by (2.8) satisfies

$$\mathbb{E}\left[K\left(s,\hat{s}_m\right) 1\!\!1_{\Omega_m(\varepsilon)}\right] \leq K(s,s_m) + \frac{1+\varepsilon}{2(1-\varepsilon)^2}\frac{D_m}{n}$$

and, if $\rho = \operatorname{essinf}\{s(x)\,,\, x\in\mathbb{X}\}$ and $\Lambda_m = \inf\{\mu(I)\,,\, I\in m\}$ are supposed to be positive and if $K\left(s,s_m\right)$ is finite then

$$\begin{aligned}\mathbb{E}\left[K\left(s,\hat{s}_m\right) 1\!\!1_{\Omega_m(\varepsilon)}\right] &\geq K(s,s_m) + \frac{1-\varepsilon}{2(1+\varepsilon)^2}\frac{D_m}{n} \\ &- \left[K(s,s_m) + \left(\rho\Lambda_m\right)^{-1}\right)\right]\mathbb{P}\left[\Omega_m^c(\varepsilon)\right].\end{aligned}$$

Furthermore, the probability of $\Omega_m^c(\varepsilon)$ is bounded by (2.9).

Proof. We control the term $K\left(s_m,\hat{s}_m\right)$ in (2.6) via (2.13) on the set $\Omega_m\left(\varepsilon\right)$.
On the one hand, we have

$$K(s,\hat{s}_m) 1\!\!1_{\Omega_m(\varepsilon)} \leq K(s,s_m) + \frac{1+\varepsilon}{2\left(1-\varepsilon\right)^2}\chi_n^2(m),$$

and taking expectation we obtain the upper bound of the proposition.
On the other hand, we get

$$K(s,\hat{s}_m) 1\!\!1_{\Omega_m(\varepsilon)} \geq \left[K(s,s_m) + \frac{1-\varepsilon}{2\left(1+\varepsilon\right)^2}\chi_n^2(m)\right] 1\!\!1_{\Omega_m(\varepsilon)}$$

Then the lower bound derives from $\chi_n^2(m) \leq 2\sup_{I\in m}\left(P\left(I\right)\right)^{-1} \leq 2\left(\rho\Lambda_m\right)^{-1}$. We conclude by the control of $\mathbb{P}\left[\Omega_m^c(\varepsilon)\right]$ given by (2.9). □

If we furthermore assume that the measures of the elements of m are not too small, we can write the previous proposition as follows:

Proposition 2.5. *Let m be some partition of $\mathbb{X}$, $\hat{s}_m$ be the histogram estimator defined by (1.1), and s_m the Kullback-Leibler projection defined by (2.5).*
Assume that $\rho = \operatorname{essinf}\{s(x)\,,\, x\in\mathbb{X}\}$ is positive and that there exists a constant B such that $\Lambda_m = \inf\{\mu(I)\,,\, I\in m\}$ satisfies $\Lambda_m \geq B\left(\log n\right)^2/n$ then

$$K\left(s,\hat{s}_m\right) = \left[1+o_P\left(1\right)\right]\left[K\left(s,s_m\right) + \frac{1}{2}\chi_n^2\left(m\right)\right]$$

where $\chi_n^2\left(m\right)$ is given by Definition 2.2.

Proof. We use the same notations as in the previous proof. Setting

$$Z_m = \frac{K\left(s_m,\hat{s}_m\right) - \frac{1}{2}\chi_n^2(m)}{\chi_n^2(m)}$$

we derive from (2.13) that on $\Omega_m(\varepsilon)$,

$$-\frac{3\varepsilon+\varepsilon^2}{2\left(1+\varepsilon\right)^2} \leq Z_m \leq \frac{3\varepsilon-\varepsilon^2}{2\left(1-\varepsilon\right)^2}.$$

Furthermore, the hypothesis on Λ_m implies the control of $\mathbb{P}\left[\Omega_m^c\left(\varepsilon\right)\right]$ given by (2.10). Consequently, Z_m converges to 0 in probability when n goes to infinity and this leads to

$$K\left(s_m,\hat{s}_m\right) = \left[\frac{1}{2} + o_P\left(1\right)\right]\chi_n^2(m).$$

We finish the proof using the Pythagore type identity (2.6). □

2.3.2 Hellinger risk

An upper bound for the Hellinger risk of an histogram estimator can be found in Birgé and Rozenholc (1999). They prove there that the Hellinger risk of $\hat{s}_m$ satisfies:

$$\mathbb{E}\left[h^2\left(s, \hat{s}_m\right)\right] \leq h^2\left(s, s_m\right) + \frac{D_m}{2n}.$$

We can improve on this upper bound and also get a lower bound if we consider the Hellinger risk restricted to the set $\Omega_m\left(\varepsilon\right)$.

Proposition 2.6. *Let m be some partition of $\mathbb{X}$ with $D_m + 1$ pieces, $\hat{s}_m$ be the histogram estimator defined by (1.1), and s_m the Kullback-Leibler projection defined by (2.5).*
For all $\varepsilon \in]0,1[$, the Hellinger risk of $\hat{s}_m$ restricted to the set $\Omega_m(\varepsilon)$, defined by (2.8), satisfies

$$\mathbb{E}\left[h^2\left(s, \hat{s}_m\right) \mathbb{1}_{\Omega_m(\varepsilon)}\right] \leq (1+\varepsilon)^{1/2} h^2(s, s_m) + (8-6\varepsilon)^{-1} \frac{D_m}{n}$$

and, if $\rho = \operatorname{essinf}\left\{s(x)\,,\, x \in \mathbb{X}\right\}$ and $\Lambda_m = \inf\left\{\mu(I)\,,\, I \in m\right\}$ are supposed to be positive then

$$\begin{aligned}\mathbb{E}\left[h^2\left(s, \hat{s}_m\right) \mathbb{1}_{\Omega_m(\varepsilon)}\right] &\geq (1-\varepsilon)^{1/2} h^2(s, s_m) + (8+4\varepsilon)^{-1} \frac{D_m}{n} \\ &- \left[(1-\varepsilon)^{1/2} h^2(s, s_m) + \left(4\rho\Lambda_m\right)^{-1}\right] \mathbb{P}\left[\Omega_m^c(\varepsilon)\right].\end{aligned}$$

Furthermore, the probability of $\Omega_m^c(\varepsilon)$ is bounded by (2.9).

Proof. We decompose the squared Hellinger distance as follows:

$$h^2\left(s, \hat{s}_m\right) = h^2\left(s, s_m\right) + h^2\left(s_m, \hat{s}_m\right) + \int \left(\sqrt{s} - \sqrt{s_m}\right)\left(\sqrt{s_m} - \sqrt{\hat{s}_m}\right) d\mu.$$

Furthermore, denoting by Π_{S_m} the orthogonal projection on S_m, one has

$$\int \left(\sqrt{s} - \sqrt{s_m}\right)\left(\sqrt{s_m} - \sqrt{\hat{s}_m}\right) d\mu = \int \left[\Pi_{S_m}\left(\sqrt{s_m} - \sqrt{s}\right)\right]\left(\sqrt{\hat{s}_m} - \sqrt{s_m}\right) d\mu.$$

But, by Jensen's inequality,

$$\Pi_{S_m}\left(\sqrt{s}\right) \leq \sqrt{\Pi_{S_m}\left(s\right)} = \sqrt{s_m}$$

thus $\Pi_{S_m}\left(\sqrt{s_m} - \sqrt{s}\right) \geq 0$ and we deduce that, on the set $\Omega_m\left(\varepsilon\right)$,

$$\int \left(\sqrt{s} - \sqrt{s_m}\right)\left(\sqrt{s_m} - \sqrt{\hat{s}_m}\right) d\mu \leq \left(\sqrt{1+\varepsilon} - 1\right) \int \left[\Pi_{S_m}\left(\sqrt{s_m} - \sqrt{s}\right)\right] \sqrt{s_m} d\mu$$

and

$$\int \left(\sqrt{s} - \sqrt{s_m}\right)\left(\sqrt{s_m} - \sqrt{\hat{s}_m}\right) d\mu \geq \left(\sqrt{1-\varepsilon} - 1\right) \int \left[\Pi_{S_m}\left(\sqrt{s_m} - \sqrt{s}\right)\right] \sqrt{s_m} d\mu.$$

since $\sqrt{\left(1-\varepsilon\right) s_m} \leq \sqrt{\hat{s}_m} \leq \sqrt{\left(1+\varepsilon\right) s_m}$ on this set.
Hence, $\int \left[\Pi_{S_m}\left(\sqrt{s_m} - \sqrt{s}\right)\right] \sqrt{s_m} d\mu = h^2\left(s, s_m\right)$ leads to

$$\left(\sqrt{1-\varepsilon} - 1\right) h^2\left(s, s_m\right) \leq \int \left(\sqrt{s} - \sqrt{s_m}\right)\left(\sqrt{s_m} - \sqrt{\hat{s}_m}\right) d\mu \leq \left(\sqrt{1+\varepsilon} - 1\right) h^2\left(s, s_m\right).$$

Consequently, on $\Omega_m(\varepsilon)$

$$(1-\varepsilon)^{\frac{1}{2}} h^2(s, s_m) + h^2(s_m, \hat{s}_m) \leq h^2(s, \hat{s}_m) \leq (1+\varepsilon)^{\frac{1}{2}} h^2(s, s_m) + h^2(s_m, \hat{s}_m). \qquad (2.14)$$

Moreover, on the set $\Omega_m(\varepsilon)$, we can link the term $h^2(s_m, \hat{s}_m)$ to the statistics $\chi_n^2(m)$ given by the Definition 2.2. Indeed, since

$$h^2(s_m, \hat{s}_m) = \frac{1}{2}\int \frac{(\hat{s}_m - s_m)^2}{\left(\sqrt{\hat{s}_m} + \sqrt{s_m}\right)^2} d\mu$$

one has, using (2.12),

$$\left[2\left((1+\varepsilon)^{\frac{1}{2}}+1\right)^2\right]^{-1} \chi_n^2(m) \leq h^2(s_m, \hat{s}_m) \leq \left[2\left((1-\varepsilon)^{\frac{1}{2}}+1\right)^2\right]^{-1} \chi_n^2(m).$$

Then, we can write

$$(8+4\varepsilon)^{-1}\chi_n^2(m) \leq h^2(s_m, \hat{s}_m) \leq (8-6\varepsilon)^{-1}\chi_n^2(m).$$

Finally, we get on the set $\Omega_m(\varepsilon)$, first,

$$h^2(s, \hat{s}_m) \leq (1+\varepsilon)^{\frac{1}{2}} h^2(s, s_m) + (8-6\varepsilon)^{-1}\chi_n^2(m). \qquad (2.15)$$

And, taking the expectation, we obtain the upper bound of the proposition. Secondly, one has on the set $\Omega_m(\varepsilon)$,

$$h^2(s, \hat{s}_m) \geq (1-\varepsilon)^{\frac{1}{2}} h^2(s, s_m) + (8+4\varepsilon)^{-1}\chi_n^2(m) \qquad (2.16)$$

Then, the lower bound derive from $\chi_n^2(m) \leq 2\sup_{I\in m}(P(I))^{-1} \leq 2(\rho\Lambda_m)^{-1}$. Next, the control of $\mathbb{P}[\Omega_m^c(\varepsilon)]$ given by (2.9) achieves the proof. □

Proposition 2.7. *Let m be some partition of $\mathbb{X}$, $\hat{s}_m$ be the histogram estimator defined by (1.1), and s_m the Kullback-Leibler projection defined by (2.5).*
Assume that $\rho = \operatorname{essinf}\{s(x), x \in \mathbb{X}\}$ is positive and that there exists a constant B such that $\Lambda_m = \inf\{\mu(I), I \in m\}$ satisfies $\Lambda_m \geq B(\log n)^2/n$ then

$$h^2(s, \hat{s}_m) = [1 + o_P(1)]\left[h^2(s, s_m) + \frac{1}{8}\chi_n^2(m)\right]$$

where $\chi_n^2(m)$ is given by Definition 2.2.

Proof. We use the same notations as in the previous proof. Setting

$$Y_m = h^2(s, s_m) + \frac{1}{8}\chi_n^2(m)$$

and

$$Z_m = \frac{h^2(s, \hat{s}_m) - Y_m}{Y_m}$$

we derive from (2.15) and (2.16) that on $\Omega_m(\varepsilon)$, $|Z_m| \leq \varepsilon$. Furthermore, the hypothesis on Λ_m allows the control of $\mathbb{P}[\Omega_m^c(\varepsilon)]$ given by (2.10). Consequently, Z_m converges to 0 in probability when n goes to infinity. □

3 Model selection

Let $\mathcal{M}_n$ be some countable collection of partitions of $\mathbb{X}$. In this section, we consider the corresponding collection of histogram estimators $\{\hat{s}_m,\ m \in \mathcal{M}_n\}$ and we want to select a "best" estimator among this collection.

Ideally, one would like to minimize the Kullback-Leibler risk over all the partitions $m \in \mathcal{M}_n$, that is to choose the best partition $\bar{m} \in \mathcal{M}_n$. As we have seen in section 2.3, the Kullback-Leibler risk is in the order of $K(s, s_m) + D_m/2n$ (where s_m denotes the Kullback-Leibler projection over the histogram model $\mathcal{S}_m$). Unfortunately, the minimization of this risk requires the knowledge of s, and consequently, the ideal partition $\bar{m} \in \mathcal{M}_n$ also depends on s: thus $\hat{s}_{\bar{m}}$ cannot be used as an estimator of s. The purpose of model selection is to provide a criterion based only on the data which select a partition $\hat{m} \in \mathcal{M}_n$ behaving approximately as well as the ideal choice $\bar{m}$: we want to define a proper estimator $\tilde{s} = \hat{s}_{\hat{m}}$ of s, the risk of which realizes the best trade-off between $K(s, s_m)$ and $D_m/2n$. Actually, in favorable cases (regular cases) the risk of $\tilde{s}$, with an adequate choice of $\hat{m}$, will be within a constant factor (independent of n) of the ideal risk. The following heuristics leads to the model selection criterion which has first been proposed by Akaike (1973).

The minimization's procedure of the Kullback-Leibler risk amounts to the minimization of

$$K(s, s_m) + \frac{D_m}{2n} - \int s \log s \, d\mu = -\int s_m \log s_m \, d\mu + \frac{D_m}{2n}$$

over all the partitions $m \in \mathcal{M}_n$. Since the term $\int s_m \log s_m \, d\mu$ depends of s, consequently it must be estimated. One can think that $\int \hat{s}_m \log \hat{s}_m \, d\mu$ will provide such an estimator. Now,

$$\mathbb{E}\left(\int \hat{s}_m \log \hat{s}_m \, d\mu\right) = \mathbb{E}[K(\hat{s}_m, s_m)] + \int s_m \log s_m \, d\mu$$

and using (2.13) one can see that $K(\hat{s}_m, s_m)$ is approximately close to $\frac{1}{2}\chi_n^2(m)$, consequently $\mathbb{E}[K(\hat{s}_m, s_m)] \simeq D_m/2n$. Thus, $-\int \hat{s}_m \log \hat{s}_m \, d\mu + D_m/2n$ can be taken as an unbiaised estimator of $-\int s_m \log s_m \, d\mu$. Finally, one can hope that the partition $\hat{m}$ minimizing

$$-\int \hat{s}_m \log \hat{s}_m \, d\mu + \frac{D_m}{n}$$

provides a good estimator $\tilde{s} = \hat{s}_{\hat{m}}$. Actually, Akaike gave no proof of the properties of his method. The purpose of this paper is to study such a criterion: to valid it in certain cases or, possibly, to modify it in other cases. It is therefore necessary to consider more general criteria of the form $-\int \hat{s}_m \log \hat{s}_m \, d\mu + \mathrm{pen}(m)$ to be minimized with respect to $m \in \mathcal{M}_n$, where "pen" is a nonnegative function on $\mathcal{M}_n$ called penalty function.

More precisely, one can define the penalized maximum likelihood estimator as follows:

Definition 3.1. *Let $\mathcal{M}_n$ be a countable family of partitions of $\mathbb{X}$, and $\{\hat{s}_m,\ m \in \mathcal{M}_n\}$ the corresponding family of histogram estimators where $\hat{s}_m$ is defined by (1.1). Given a penalty function,* $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}_+$, *one defines the penalized maximum likelihood estimator $\tilde{s} \in \mathcal{S}_{\hat{m}}$ as:*

$$\tilde{s} = \hat{s}_{\hat{m}} \tag{3.17}$$

where $\hat{m}$ minimizes the penalized criterion over $\mathcal{M}_n$:

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m) = -\int \hat{s}_m \log \hat{s}_m \, d\mu + \mathrm{pen}(m). \tag{3.18}$$

One can remark that Akaike's criterion corresponds to a penalty function equal to

$$\mathrm{pen}(m) = \frac{D_m}{n}.$$

The value of the penalty function depends on the uniform control of the statistics $\chi_n^2(m)$ for all partitions $m \in \mathcal{M}_n$. This control takes into account the number of partitions of the family $\mathcal{M}_n$ having the same number of pieces. That is why Akaike's criterion turn out to be unsufficient for "large" families of partitions (like the family of all irregular partitions) whereas it works well for families which are not "too large" (like the family of regular partitions). Thus, we will see that a suitable penalty function is essentially the sum of a term proportional to D_m/n and of a term depending on the complexity of family $\mathcal{M}_n$.
This is the reason why we introduce the two following types of family of partitions $\mathcal{M}_n$:

$(\mathbf{H_1})$: *there exists an integer a and some fixed constant B' such that the number of partitions of the family $\mathcal{M}_n$ with the same number of pieces $D+1$ is bounded by $B'D^a$. Moreover, if we note*

$$\Lambda_n = \inf_{m\in\mathcal{M}_n} \inf_{I\in m} \{\mu(I)\}$$

then we assume that there exists some constant B such that $\Lambda_n \geq B(\log n)^2/n$.

A particular example of a family satisfying $(\mathbf{H_1})$ is the family of regular partitions. A partition m is regular if all its elements have the same measure with respect to μ, consequently, in this case, $B' = 1$ and $a = 0$, $\Lambda_n = \left[\sup_{m\in\mathcal{M}_n}\{D_m+1\}\right]^{-1}$ where $D_m + 1$ is the number of pieces of m and $D_m + 1 \leq n/B(\log n)^2$ for every partition m. Furthermore, this family contains only one partition for a given number of pieces. We will denote this family $\mathcal{M}_n^r$.

$(\mathbf{H_2})$: *we consider a partition m_n of $\mathbb{X}$ with N_n+1 pieces such that $\mu(I) = (N_n+1)^{-1}$ for all $I \in m_n$ and we assume that every partition of the family $\mathcal{M}_n$ is built from m_n, i.e. a partition m belongs to $\mathcal{M}_n$ if any element of m is the union of elements of m_n. In this case, we note*

$$\Lambda_n = (N_n+1)^{-1}$$

and we assume that there exists some constant B such that $\Lambda_n \geq B(\log n)^2/n$.

As an example, we can take $\mathbb{X} = [0,1]$ and consider a grid $\{k/(N_n+1), 0 \leq k \leq N_n+1\}$, then the family is composed of all partitions of $[0,1]$ with endpoints belonging to this grid. Furthermore, $N_n + 1 \leq n/B(\log n)^2$ and this family contains $C_{N_n}^D$ partitions with $D+1$ pieces. We will note this family $\mathcal{M}_n^{ir}$.

In this section, we show what kind of penalty function is allowed in order to select a "best" estimator among the collection $\{\hat{s}_m,\ m \in \mathcal{M}_n\}$. We first give the general form of the penalty function valid for the families of partitions satisfying $(\mathbf{H_1})$ or $(\mathbf{H_2})$ and next, we apply this result in our particular cases. Then we study the results we can add for the families satisfying the hypothesis $(\mathbf{H_1})$.

In the sequel, we shall use various quantities depending on the parameters involved in our theorems. Such a quantity will systematically be denoted by C, C_1, C_2 or more precisely $C(.,\ldots)$ to indicate the dependence on various parameters and especially those which are related to the unknown density s.

3.1 General results

The following result provides a model selection criterion, that is a data choice

$$(X_1, ..., X_n) \mapsto \hat{m}(X_1, ..., X_n)$$

based on a penalized maximum likelihood method, and gives some non asymptotic bounds on the performances of the resulting estimator $\tilde{s} = \hat{s}_{\hat{m}}$.
First, we find the general form of the penalty function, and secondly, we give the choice of the penalty function corresponding to the two different families (the regular and the irregular case).

3.1.1 The main theorems

Theorem 3.2. *Let $X_1, ..., X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P = s\mu$. Consider a countable family $\mathcal{M}_n$ of partitions of $\mathbb{X}$ and assume that this family satisfies either* $(\mathbf{H_1})$ *or* $(\mathbf{H_2})$.
Let $(L_m)_{m \in \mathcal{M}_n}$ be some family of non negative weights such that

$$\Sigma = \sum_{m \in \mathcal{M}_n} \exp\left(-L_m D_m\right) < +\infty. \tag{3.19}$$

Choose some real positive numbers $c_1 > 1/2$ and $c_2 = 2(\kappa + 2)$ ($\kappa = 4$ works), and a penalty function pen (.) *such that*

$$\operatorname{pen}(m) \geq c_1 \left(1 + \sqrt{c_2 L_m}\right)^2 \frac{D_m}{n}, \quad \textit{for all } m \in \mathcal{M}_n$$

where $D_m + 1$ is the number of pieces of the partition m.
If $\rho = \operatorname{essinf}\{s(x), \ x \in \mathbb{X}\}$ is positive and if $S = \int s (\log s)^2 \, d\mu$ is finite then there exists a constant $C(c_1, \rho, S, \Sigma, B)$ such that the penalized maximum likelihood estimator $\tilde{s}$ defined by (3.17) satisfies

$$\mathbb{E}\left[h^2(s, \tilde{s})\right] \leq \frac{(2c_1)^{1/5}}{(2c_1)^{1/5} - 1} \inf_{m \in \mathcal{M}_n} \left\{K(s, \mathcal{S}_m) + \operatorname{pen}(m)\right\} + \frac{C(c_1, \rho, S, \Sigma, B)}{n}$$

This theorem gives the general form of the penalty function:

$$\operatorname{pen}(m) = c_1 \left(1 + \sqrt{c_2 L_m}\right)^2 \frac{D_m}{n}$$

where the weights L_m satisfy (3.19) and, of course, the constant c_1 and c_2 are independant of the density s. The choice $c_1 > 1/2$ provides an upper bound for the Hellinger risk of the penalized maximum likelihood estimator:

$$\mathbb{E}\left[h^2(s, \tilde{s})\right] \leq C_1 \inf_{m \in \mathcal{M}_n} \left\{K(s, \mathcal{S}_m) + \operatorname{pen}(m)\right\} + \frac{C_2}{n} \tag{3.20}$$

where the constant C_1 does not depend on s whereas the constant C_2 depends on s (via ρ and S) and on the family of models (via Σ and B). Furthermore, the constant C_1, which depends only on c_1, converges to infinity when c_1 tends to $1/2$. The following theorem shows that the special choice $c_1 = 1$ seems to improve the risk bound (3.20) in the sens that C_1 converges to $1/2$ when C_2 goes to infinity. This special choice requires a proof which is more sophisticated.

Theorem 3.3. *Let $X_1, ..., X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P = s\mu$. Consider a countable family $\mathcal{M}_n$ of partitions of $\mathbb{X}$ and assume that this family satisfies either* $(\mathbf{H_1})$ *or* $(\mathbf{H_2})$.
Let $(L_m)_{m \in \mathcal{M}_n}$ be some family of non negative weights satisfying the condition (3.19).

Define the penalty function by

$$\text{pen}(m) = \left(1 + \sqrt{c_2 L_m}\right)^2 \frac{D_m}{n}\ , \quad \textit{for all}\ \ m \in \mathcal{M}_n$$

with $c_2 = 2(\kappa + 1)$ $(\kappa = 4$ *works) and where* $D_m + 1$ *is the number of pieces of the partition* m*.*
If $\rho = \operatorname{essinf}\{s(x)\ ,\ x \in \mathbb{X}\}$ *is positive and if* $S = \int s(\log s)^2 d\mu$ *is finite then for all* $\varepsilon \leq 0.23$ *there exists constants* $C(\varepsilon)$ *and* $C(\varepsilon, \rho, S, \Sigma, B)$ *such that the penalized maximum likelihood estimator* $\tilde{s}$ *defined by (3.17) satisfies*

$$\begin{aligned}\mathbb{E}\left[h^2(s, \tilde{s})\right] &\leq C(\varepsilon) \inf_{m \in \mathcal{M}_n} \left\{K(s, \mathcal{S}_m) + \frac{1}{2}\left(1 + \sqrt{2c_2 L_m}\right)^2 \frac{D_m}{n}\right\} \\ &+ \frac{C(\varepsilon, \rho, S, \Sigma, B)}{n}\end{aligned}$$

where the functions $C(\varepsilon)$ *and* $C(\varepsilon, \rho, S, \Sigma, B)$ *satisfies*

$$\lim_{\varepsilon \to 0} C(\varepsilon) = \frac{1}{2}\ ,\ \lim_{\varepsilon \to 0} C(\varepsilon, \rho, S, \Sigma, B) = \infty.$$

Remark. We get the following values for the constant $C(\varepsilon)$:

ε	0.23	0.2	0.15	0.1	0.05	0.01
$C(\varepsilon)$	48.5	4.60	1.74	1.02	0.69	0.53

3.1.2 Choice of the weights $\{L_m\ , m \in \mathcal{M}_n\}$

The penalty function depends of the family $\mathcal{M}_n$ through the choice of the weights L_m satisfying (3.19) for a reasonnable value of Σ. The number of histogram models $\mathcal{S}_m$ having the same dimension plays a fondamental role in the convergence of the serie (3.19). That is why we distinguish the two different types of family $(\mathbf{H_1})$ or $(\mathbf{H_2})$.

• Case $(\mathbf{H_1})$
By assumption, choosing the same weight L_D for the partitions of the family having the same number of pieces $D + 1$, we can write

$$\Sigma \leq B' \sum_{D \geq 1} D^a \exp(-L_D D)$$

since the family $(\mathbf{H_1})$ contains at most $B'D^a$ partitions for a given number of pieces $D + 1$.

Constant weights The simplest way of choosing the L_ms is to take them constant. Taking $L_m = L$ for all $m \in \mathcal{M}_n$ implies that Σ is bounded independently of n. Consequently, all penalties of the form

$$\text{pen}(m) = c\,\frac{D_m}{n}$$

with $c > 1/2$ are allowed, including that of Akaike, namely $c = 1$. Since $K(s, \mathcal{S}_m) + \text{pen}(m)$ in the risk bound (3.20) represents some evaluation of the Kullback-Leibler risk of an histogram estimator $\hat{s}_m$ and $K(s, \mathcal{S}_m) = K(s, s_m)$ is close to $4h^2(s, s_m)$ when the ratio s/s_m is small enought, the risk of $\tilde{s}$ realizes the infimum of the risk of $\hat{s}_m$ over the family $\mathcal{M}_n$ up to some multiplicative constant. To be more precise, if we assume that there exists some positive number M such that $\|\log(s/s_m)\|_\infty \leq M$ then the upper bound (3.20) can be stated as follows:

$$\mathbb{E}\left[h^2(s, \tilde{s})\right] \leq C \inf_{m \in \mathcal{M}_n} \left\{h^2(s, s_m) + \text{pen}(m)\right\}.$$

Then, we get

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leq C \inf_{m\in\mathcal{M}_n} \mathbb{E}\left[h^2(s,\hat{s}_m)\right]$$

the constant C being better for the values of c close to 1. Finally, in this case the AIC is valid.

Variable weights We can also take the weights L_m as function of the dimension D_m in order to get a smaller value for the constant Σ. For example, we can choose $L_m = LD_m^{-1/2}$ for a reasonnable value of L (for instance $L=1$). Hence, the penalty function can be written as:

$$\mathrm{pen}(m) = \frac{D_m}{n} + c\,\frac{D_m^{3/4}}{n}$$

and gives the following risk bound:

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leq C(\varepsilon) \inf_{m\in\mathcal{M}_n}\left\{K(s,\mathcal{S}_m) + \frac{1}{2}\,\mathrm{pen}(m)\right\} + \frac{C(\varepsilon,\rho,S,\Sigma,B)}{n}$$

This penalty function can be viewed as a modified version of Akaike's criterion which is equivalent to it when D_m converges to infinity.

In both situations (constant or variable weights) we have proven that, for suitable choices of the penalty function, these penalized maximum likelihood estimators are optimal in the sens that they realize the best trade-off between the bias term and the variance term, up to constants, and it seems that one should prefer variable weights to constant weights. These theorical results has been confirmed by the simulations of Birgé and Rozenholc (1999).

• **Case ($\mathbf{H_2}$)**

We consider the particular case of the family $\mathcal{M}_n^{ir}$ satisfying the hypothesis ($\mathbf{H_2}$). For this example, the number of histogram models having the same dimension D is bounded by $C_{N_n}^D$.

Constant weights If we take the weights L_m all equals to some constant L then

$$\Sigma \leq \sum_{D=1}^{N_n} C_{N_n}^D e^{-LD} \leq \left(1+e^{-L}\right)^{N_n} - 1.$$

Consequently, if we take $L = \log n$ then we get $\Sigma \leq e^{\frac{1}{B}} - 1$ and this leads to a penalty function of the form

$$\mathrm{pen}(m) = cD_m\,\frac{\log n}{n}.$$

Then we loose a $\log n$ factor in the risk bound:

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leq C\log n \inf_{m\in\mathcal{M}_n}\left\{K(s,s_m) + \frac{D_m}{n}\right\}.$$

Variable weights It is also possible to consider weights depending on the dimension: $L_m = L(D_m)$ for a partition m with D_m+1 pieces. This leads to

$$\Sigma = \sum_{D=1}^{N_n} C_{N_n}^D e^{-L(D)D} \leq \sum_{D\geq 1}\left(\frac{eN_n}{D}\right)^D e^{-L(D)D}$$

using Lemma 6 of Barron, Birgé, and Massart (1999) p. 399. Consequently, the choice $L(D) = L + \log(eN_n/D)$ gives $\Sigma \leq \left(e^L - 1\right)^{-1}$. This leads to a penalty function of the form

$$\mathrm{pen}(m) = c\frac{D_m}{n}\log\left(\frac{N_n}{D_m}\right)$$

and gives the following risk bound:

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \le C(\varepsilon)\left[\inf_{m\in\mathcal{M}_n}\left\{K(s,\mathcal{S}_m) + {c'}_2\frac{D_m}{n}\log\left(\frac{N_n}{D_m}\right)\right\}\right] + \frac{C(\varepsilon,\rho,S,\Sigma,B)}{n}.$$

Consequently, if s belongs to some histogram model $\mathcal{S}_m$ with dimension D then the Hellinger risk of s is bounded by $CD/n\log(n/D)$. We can remark that the $\log(n/D)$ factor is necessary since it follows from Proposition 2 of Birgé and Massart (1998) that the minimax risk over the set $S_D = \cup_{m:D_m=D}\mathcal{S}_m$ is in the same order $D/n\log(n/D)$ if $9 \le D \le n/5$. This remark also shows that the strategy of variable weights produces the right rate of convergence, that is the reason why one should prefer this strategy.

3.2 Case (H_1)

For the family satisfying the hypothesis $(\mathbf{H_1})$ we can get first a risk bound for the penalized maximum likelihood estimator with respect to Kullback-Leibler risk, and secondly, we can prove that the condition $c_1 > 1/2$ in Theorem 3.2 is necessary.

3.2.1 Kullback-Leibler risk for the penalized maximum likelihood estimator

If we apply Theorem 3.2 with constant weights, the penalty $\operatorname{pen}(m) = cD_m/n$ with $c > 1/2$ leads to the bound (3.20) for the Hellinger risk of the corresponding penalized maximum likelihood estimator. One should prefer to get bound for the Kullback-Leibler risk since (3.20) involves a bias term with respect to Kullback-Leibler information. Such a risk bound is possible if we assume that $\|\log(s/s_m)\|_\infty$ is bounded uniformly over the family $\mathcal{M}_n$.

Theorem 3.4. *Let $X_1,\dots,X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P = s\mu$. Consider the countable family of partitions of $\mathbb{X}$, $\mathcal{M}_n$ and assume that this family satisfies the hypothesis $(\mathbf{H_1})$.*
Choose some real number $c > 1/2$ and define the penalty function by

$$\operatorname{pen}(m) \ge c\,\frac{D_m}{n}\ , \quad \textit{for all}\ \ m\in\mathcal{M}_n$$

where $D_m + 1$ is the number of pieces of the partition m.
If $\rho = \operatorname{essinf}\{s(x)\ ,\ x\in\mathbb{X}\}$ is positive and if there exists $M > 0$ such that

$$\left|\log\left(\|s\|_\infty/\rho\right)\right| \le M$$

then, whatever $b > 0$, there exists $C(c,\rho,b,B)$, $C(c,\rho,M,B)$ and a set $\Omega_n(c)$ such that the penalized maximum likelihood estimator $\tilde{s}$ defined by (3.17) satisfies

$$\mathbb{E}\left[K(s,\tilde{s})\,1\!\!1_{\Omega_n(c)}\right] \le \frac{(2c)^{1/6}}{(2c)^{1/6}-1}\inf_{m\in\mathcal{M}_n}\left\{K(s,\mathcal{S}_m) + \operatorname{pen}(m)\right\} + \frac{C(c,\rho,M,B)}{n}$$

with $\mathbb{P}\left[\Omega_n^c(c)\right] \le \frac{C(c,\rho,b,B)}{n^b}$.

Remark. The Kullback-Leibler risk is restricted to a large set $\Omega_n(c)$ which warrants that $\hat{s}_m$ is positive for all $m\in\mathcal{M}_n$ (this implies that $\tilde{s}$ is also positive). Indeed, the estimator $\hat{s}_m$ can be zero on some set $I\in m$ if there is no data X_i in this set, and the Kullback-Leibler information between s and s_m is then infinite. We can modify the estimator $\hat{s}_m$ in order to get a positive estimator by considering a positive density $\hat{\hat{s}}_m$ such that

$$\gamma_n\left(\hat{\hat{s}}_m\right) \le \inf_{t\in\mathcal{S}_m}\gamma_n(t) + \frac{1}{n}.$$

Next, we can define the penalized estimator $\tilde{\hat{s}} = \hat{\hat{s}}_{\hat{m}}$ where $\hat{m}$ is the minimizer of

$$\gamma_n\left(\hat{\hat{s}}_m\right) + \text{pen}\,(m)$$

over the family $\mathcal{M}_n$. Then, we can get with $\text{pen}\,(m) = cD_m/n$ the following risk bound:

$$\mathbb{E}\left[K\left(s, \tilde{\hat{s}}\right)\right] \leq \frac{(2c)^{1/6}}{(2c)^{1/6} - 1}\left[\inf_{m \in \mathcal{M}_n}\left\{K\left(s, \mathcal{S}_m\right) + \text{pen}\,(m)\right\}\right] + \frac{C\left(c, \rho, M, B,\right)}{n}.$$

3.2.2 Lower bound for the penalty function

Applying Theorem 3.4 the penalty function can be taken as $\text{pen}\,(m) = cD_m/n$ for some arbitrary constant $c > 1/2$. A question which arises from this is whether the condition $c > 1/2$ is necessary. The following theorem shows that if one chooses for one partition $m_N \in \mathcal{M}_n$ with $N+1$ pieces, $\text{pen}\,(m_N) = cN/n$ with $c < 1/2$ then it will produce a bad estimation of the uniform density $s = 1\!\!1_{\mathbb{X}}$. Indeed, if N is large enough then this choice of the penalty will select automatically the models with large dimension.

Theorem 3.5. *Let $X_1, ..., X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P = s\mu$ with $s = 1\!\!1_{\mathbb{X}}$. Consider the countable family of partitions $\mathcal{M}_n$ and assume that this family satisfies the hypothesis* $(\mathbf{H_1})$.
Assume that there exists one partition $m_N \in \mathcal{M}_n$ with $N+1$ pieces such that

$$\text{pen}\,(m_N) = c\,\frac{N}{n}$$

with $c < 1/2$.
Let define $\hat{m}$ to be the corresponding minimizer over $\mathcal{M}_n$ of the penalized criterion $\text{crit}\,(m)$ *given by (3.18). Then, whatever the values of* $\text{pen}\,(m)$ *for $m \neq m_N$ there exists positive numbers N_0 and L depending only on c such that for all $N \geq N_0$*

$$\mathbb{P}\left(D_{\hat{m}} \geq \frac{1-4c^2}{4}N\right) \geq 1 - \beta\,(c)$$

where $\beta\,(c) = \Sigma\,(L)\exp\left[-\frac{L}{2}\left(\left(1-4c^2\right)N\right)^{\frac{1}{2}}\right] + \frac{C(c)}{n}$ *with* $\Sigma\,(L) = B'\sum_{D\geq 1} D^a e^{-L\sqrt{D}}$.
Moreover, if $\tilde{s} = \hat{s}_{\hat{m}}$,

$$\mathbb{E}\left[K\left(s, \tilde{s}\right)\right] \geq \delta\,(c)\left[1 - \beta\,(c)\right]\frac{N}{n}$$

where $\delta\,(c) = (1-2c)\,(1+2c)^2/16$.

It follows from Theorem 3.4 that a value of $c > 1/2$ for the penalty function leads to a Kullback-Leibler risk of the penalized estimator $\tilde{s}$ bounded by C/n when $s = 1\!\!1_{\mathbb{X}}$ while the value $c < 1/2$ leads to a risk larger than $C'N/n$ for large value of N, and consequently, to a bad estimation of s. The proof of this theorem is adapted from the one of the statistical Gaussian framework (Birgé and Massart 1999a).

3.3 Adaptive properties of the maximum likelihood estimator in the minimax sens

3.3.1 One-dimensional case

Let us now assume that s is a density with respect to Lebesgue measure μ on $[0,1]$. We study two cases: the first one is devoted to the α-Hölderian densities and the second one to the densities with bounded α-variation. In both cases, α is a real positive number such that $0 < \alpha \leq 1$.

- Let us define $\mathcal{H}(H, \alpha)$ to be the set of the functions f satisfying:

$$|f(x) - f(y)| \leq H |x - y|^{\alpha} \text{ for all } x, y \in \mathbb{X}.$$

Assume that s belongs to $\mathcal{H}(H, \alpha)$ and that $\rho = \text{essinf}\{s(x)\ ,\ x \in [0,1]\}$ is positive. We can consider the collection of histogram models $\{\mathcal{S}_m,\ m \in \mathcal{M}_n^r\}$ which has good approximation properties for the set $\mathcal{H}(H, \alpha)$. Indeed, Proposition 1 of Barron et al. (1999) p. 327, implies that $K(s, \mathcal{S}_m) \wedge 1$ is bounded $CR^2 D_m^{-2\alpha}$. Consequently, there exists a constant C' such that for any $H > 0$ and $\alpha \in]0,1]$, the penalized maximum likelihood estimator $\tilde{s}$ built from the family $\mathcal{M}_n^r$ and a penalty function given by Theorem 3.2 or Theorem 3.4 satisfies

$$\mathbb{E}\left[h^2(s, \tilde{s})\right] \leq C' H^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}}.$$

Even if H and α were known, one couldn't do better from the minimax point of vue.

- We can also use the rich collection of models $\left\{\mathcal{S}_m,\ m \in \mathcal{M}_n^{ir}\right\}$ to estimate much less homogeneous density with smoothness α. Indeed, assume that s has a bounded α-variation:

$$\sup_{k \geq 2} \sup_{x_1 \leq \ldots \leq x_k} \sum_{j=2}^{k} |s(x_j) - s(x_{j-1})|^{1/\alpha} = J_\alpha(s) < +\infty$$

where the supremum is taken over all increasing sequences $x_1 < \ldots < x_k$ of points in $[0,1]$, and that $\rho = \text{essinf}\{s(x)\ ,\ x \in [0,1]\}$ is positive. Then it follows from Theorem 3.2 that the penalized maximum likelihood estimator $\tilde{s}$ built from the family $\mathcal{M}_n^{ir}$ satisfies

$$\mathbb{E}\left[h^2(s, \tilde{s})\right] \leq C \left(\frac{J_\alpha(s) \log n}{n}\right)^{2\alpha/(2\alpha+1)} \wedge 1.$$

see Barron et al. (1999) p. 355.

3.3.2 Multidimensional case

Let us now assume that s is a density with respect to Lebesgue measure μ on $[0,1]^q$. Given $\underline{\alpha} = (\alpha_1, \ldots, \alpha_q)$ and $\underline{H} = (H_1, \ldots, H_q)$, we define $\mathbb{H}(\underline{H}, \underline{\alpha})$ to be the set of the functions f on $[0,1]^q$ such that, for every $i \in \{1, \ldots, q\}$,

$$y \mapsto f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_q)$$

belong to $\mathcal{H}(H_i, \alpha_i)$. We assume that s belongs to $\mathcal{S}(\underline{H}, \underline{\alpha})$ for some $\underline{H} \in \mathbb{R}^q$ and $\underline{\alpha} \in]0,1]^q$ and that $\rho = \text{essinf}\{s(x)\ ,\ x \in [0,1]^q\}$ is positive.
If we note m_D the partition of $[0,1]$ with D pieces then we consider the family of the partitions $m = m_{D_1} \times \ldots \times m_{D_q}$ of $[0,1]^q$ such that $\prod_{i=1}^q D_i$ is bounded by $n/B(\log n)^2$. This family satisfies the hypothesis $(\mathbf{H_1})$ since the number of partitions having the same number of pieces is polynomial. It follows from Theorem 3.2 that, if we define

$$\frac{q}{\alpha} = \sum_{i=1}^{q} \frac{1}{\alpha_i} \text{ and } H = \left[\prod_{i=1}^{q} H_i^{1/\alpha_i}\right]^{\alpha/q}$$

and assume that for any i

$$n^\alpha H_i^{2\alpha+q} \geq H^q$$

then the penalized maximum likelihood estimator $\tilde{s}$ satisfies

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leq C\left[\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)} + \frac{1}{n}\right].$$

see Proposition 5 Barron et al. (1999) p. 346.

4 Exponential bounds

In a first part, we recall some general results about concentration of the supremum of empirical processes. Next, in a second part, we use these results to derive exponential bounds for the statistics given by Definition 2.2. Finally, in a third part, we prove a proposition which will be used in the proofs of Theorem 3.2 and Theorem 3.3.

4.1 Concentration inequalities for empirical processes

First, we recall the concentration inequalities on empirical processes established by Massart (1998).

Theorem 4.1. *Let $X_1, ..., X_n$ be independent random variables with values in some measurable space $(\mathbb{X}, \mathcal{X})$. Let b be some positive number and $\mathcal{F}$ be some countable class of real valued measurable functions on $\mathbb{X}$, such that $\|f\|_\infty \leq b$ for any $f \in \mathcal{F}$. Let*

$$Z = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right|.$$

Setting

$$\sigma^2 = \sup_{f\in\mathcal{F}}\sum_{i=1}^n \mathrm{Var}\left[f(X_i)\right]$$

for any positive numbers x and ε the following inequalities hold:

$$\mathbb{P}\left[Z \geq (1+\varepsilon)\,\mathbb{E}(Z) + \sigma\sqrt{2\kappa x} + \kappa(\varepsilon)\,bx\right] \leq \exp(-x) \tag{4.21}$$

$$\mathbb{P}\left[Z \leq (1-\varepsilon)\,\mathbb{E}(Z) - \sigma\sqrt{2\kappa' x} - \kappa'(\varepsilon)\,bx\right] \leq \exp(-x) \tag{4.22}$$

where the constants are defined by

$$\begin{aligned} \kappa &= 4 \quad \textit{and} \quad \kappa(\varepsilon) = 2.5 + 32\varepsilon^{-1} \\ \kappa' &= 5.4 \quad \textit{and} \quad \kappa'(\varepsilon) = 3.5 + 43.2\varepsilon^{-1}. \end{aligned} \tag{4.23}$$

We also recall the following result due to Ledoux (1996). This proposition will be usefull to get lower bound for the expectation of the $\chi_n^2(m)$.

Proposition 4.2. *Let $X_1, ..., X_n$ be independent random variables with values in some measurable space $(\mathbb{X}, \mathcal{X})$ and $\mathcal{F}$ be some countable class of real valued measurable functions on $\mathbb{X}$. Let*

$$Z = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right|.$$

Let $(X'_1, ..., X'_n)$ be independent from $(X_1, ..., X_n)$ and with the same distribution. Setting

$$v = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\left(f\left(X_i\right) - f\left(X'_i\right)\right)^2\right]$$

then,

$$\mathbb{E}\left(Z^2\right) - \left[\mathbb{E}\left(Z\right)\right]^2 \leq v.$$

Remark. Using the notations of Theorem 4.1 and Proposition 4.2, we have the following inequality:

$$v \leq 2\sigma^2 + 32b\mathbb{E}\left(Z\right). \tag{4.24}$$

4.2 Study of $\chi_n^2\left(m\right)$

We recall that the statistics $\chi_n^2\left(m\right)$ is the random variable given by Definition 2.2. We define the centered empirical measure, ν_n associated with X_1,...,X_n by

$$\nu_n(f) = P_n\left(f\right) - P\left(f\right) \tag{4.25}$$

for every P-integrable function f. According to the definition of the histogram $\hat{s}_m$ and of the Kullback-Leibler projection s_m on a given partition m, we can write

$$\nu_n\left(f\right) = \int f\left(\hat{s}_m - s_m\right)d\mu \tag{4.26}$$

for every function $f \in S_m$.
Furthermore, if we note $\varphi_I = P\left(I\right)^{-\frac{1}{2}}\mathbb{1}_I$, for all $I \in m$, we get

$$\chi_n^2\left(m\right) = \sum_{I\in m}\nu_n^2\left(\varphi_I\right).$$

4.2.1 Exponential bound for $\chi_n\left(m\right)$

All the theorems relies upon the following proposition which is an application of the concentration inequalities due to Talagrand and improved in terms of constants by Massart (1998) (see Theorem 4.1 above).

Proposition 4.3. *Let $X_1, ..., X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P = s\mu$. Let m be some partition of $\mathbb{X}$ with $D_m + 1$ pieces. Let $\chi_n^2\left(m\right)$ be the statistics given by Definition 2.2. Then, for any positive real numbers ε and x:*

$$\mathbb{P}\left[\chi_n(m)\mathbb{1}_{\Omega_m(\varepsilon)} \geq \left(1+\varepsilon\right)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2\kappa x}{n}}\right)\right] \leq \exp\left(-x\right) \tag{4.27}$$

where $\Omega_m\left(\varepsilon\right) = \left\{\forall I \in m, \left|P_n\left(I\right) - P\left(I\right)\right| \leq \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}P\left(I\right)\right\}$.
Moreover, one also has:

$$\mathbb{P}\left[\chi_n(m) \leq \left(1-\varepsilon\right)\left[\mathbb{E}\left(\chi_n(m)\right)\right] - \sqrt{\frac{2\kappa' x}{n}} - \frac{\kappa'\left(\varepsilon\right)b_m x}{n}\right] \leq \exp\left(-x\right) \tag{4.28}$$

where $b_m = \sup_{I\in m} P\left(I\right)^{-\frac{1}{2}}$ and the constants κ, κ', $\kappa(\varepsilon)$ and $\kappa'(\varepsilon)$ are given by (4.23).

Proof. Let ε and x be some positive numbers. We can write

$$\chi_n(m) = \left[\sum_{I \in m} \nu_n^2(\varphi_I)\right]^{\frac{1}{2}} = \sup_{a \in B_m} \left|\nu_n\left(\sum_{I \in m} a_I \varphi_I\right)\right| \tag{4.29}$$

with $B_m = \left\{a \in \mathbb{R}^m : \sum_{I \in m} a_I^2 = 1\right\}$.
The proof of the deviation inequality on the right tail is adapted from Birgé and Massart (1999b). The main point is to apply the inequality (4.21) to a smaller class of functions than the natural class $\left\{\sum_{I \in m} a_I \varphi_I \, , a \in B_m\right\}$. Indeed, the supremun in the equality (4.29) is achieved at $a \in \mathbb{R}^m : a_I = \nu_n(\varphi_I)\left[\chi_n(m)\right]^{-1}$. Setting $\xi = \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$, the set $\Omega_m(\varepsilon)$ can be written as:

$$\Omega_m(\varepsilon) = \left\{|\nu_n(\varphi_I)| \leq \xi P(I)^{\frac{1}{2}}, \text{ for all } I \in m\right\}.$$

Consequently, defining A_m by:

$$A_m = \left\{a \in \mathbb{R}^m : |a|_2 = 1; \, |a_I| \leq \frac{\xi}{z} P(I)^{\frac{1}{2}}, \, \forall I \in m\right\}$$

where z is some positive number which will be choosen later, we get on the event $\Omega_m(\varepsilon) \cap \{\chi_n(m) \geq z\}$:

$$\chi_n(m) = \sup_{a \in A_m} \left|\nu_n\left(\sum_{I \in m} a_I \varphi_I\right)\right|$$

Taking A'_m to be a countable and dense subset of A_m:

$$\sup_{a \in A_m} \left|\nu_n\left(\sum_{I \in m} a_I \varphi_I\right)\right| = \sup_{a \in A'_m} \left|\nu_n\left(\sum_{I \in m} a_I \varphi_I\right)\right|$$

Hence, an application of (4.21) to the countable family of functions

$$\left\{\sum_{I \in m} a_I \varphi_I \, , a \in A'_m\right\}$$

leads to

$$\mathbb{P}\left[\sup_{a \in A_m} \left|\nu_n\left(\sum_{I \in m} a_I \varphi_I\right)\right| \geq (1+\varepsilon) E_m + \sqrt{\frac{2\kappa\sigma^2 x}{n}} + \kappa(\varepsilon)\frac{bx}{n}\right] \leq \exp(-x)$$

where

$$E_m = \mathbb{E}[\chi_n(m))] \leq \mathbb{E}^{\frac{1}{2}}\left[\chi_n^2(m)\right] \leq \sqrt{\frac{D_m}{n}}$$

and

$$\sigma^2 = \sup_{a \in A'_m}\left[\mathrm{Var}\left(\sum_{I \in m} a_I \varphi_I(X_1)\right)\right] \leq 1$$

and

$$b = \sup_{a \in A'_m}\left\|\sum_{I \in m} a_I \varphi_I\right\|_\infty = \sup_{a \in A'_m}\left\{\sup_{I \in m}\left|\frac{a_I}{P(I)^{\frac{1}{2}}}\right|\right\} \leq \frac{\xi}{z}.$$

Consequently,

$$\mathbb{P}\left[\chi_n(m)\,\mathbb{1}_{\Omega_m(\varepsilon)\cap\{\chi_n(m)\geq z\}}\geq(1+\varepsilon)\sqrt{\frac{D_m}{n}}+\sqrt{\frac{2\kappa x}{n}}+\kappa(\varepsilon)\frac{\xi x}{zn}\right]\leq\exp(-x).$$

Then, if we choose $z=\sqrt{\frac{2\kappa x}{n}}$ we get

$$\mathbb{P}\left[\chi_n(m)\,\mathbb{1}_{\Omega_m(\varepsilon)}\geq(1+\varepsilon)\left(\sqrt{\frac{D_m}{n}}+\sqrt{\frac{2\kappa x}{n}}\right)\right]\leq\exp(-x).$$

The deviation inequality on the left tail is a direct application of (4.22) to the countable set of functions $\left\{\sum_{I\in m}a_I\varphi_I\,,a\in B'_m\right\}$ where B'_m is a countable and dense subset of B_m:

$$\mathbb{P}\left[\sup_{a\in B'_m}\left|\nu_n\left(\sum_{I\in m}a_I\varphi_I\right)\right|\leq(1-\varepsilon)E_m-\sqrt{\frac{2\kappa'\sigma^2x}{n}}-\kappa'(\varepsilon)\frac{bx}{n}\right]\leq\exp(-x)$$

where

$$E_m=\mathbb{E}[\chi_n(m))]$$

and

$$\sigma^2=\sup_{a\in B'_m}\left[\mathrm{Var}\left(\sum_{I\in m}a_I\varphi_I(X_1)\right)\right]\leq 1$$

and

$$b=\sup_{a\in B'_m}\left\|\sum_{I\in m}a_I\varphi_I\right\|_\infty=\sup_{a\in B'_m}\left\{\sup_{I\in m}\left|\frac{a_I}{P(I)^{\frac{1}{2}}}\right|\right\}\leq\sup_{I\in m}P(I)^{-\frac{1}{2}}\leq b_m.$$

Consequently,

$$\mathbb{P}\left[\chi_n(m)\leq(1-\varepsilon)\,\mathbb{E}[\chi_n(m))]-\sqrt{\frac{2\kappa' x}{n}}-\frac{\kappa'(\varepsilon)\,b_m x}{n}\right]\leq\exp(-x)$$

which achieves the proof. □

4.2.2 Control of the expectation of $\chi_n(m)$

In order to use the deviation inequality on the left tail for $\chi_n(m)$ above, it is necessary to evaluate $\mathbb{E}[\chi_n(m)]$. Our purpose is to show that the upper bound which derives from Jensen's inequality:

$$\mathbb{E}^2[\chi_n(m)]\leq\mathbb{E}\left[\chi_n^2(m)\right]\leq\frac{D_m}{n}$$

is sharp. This is again a consequence of the concentration of measure phenomenon and more precisely of the Poincaré type inequality due to Ledoux (see Proposition 4.2 above).

Proposition 4.4. *Let $X_1,...,X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P=s\mu$. Let m be a partition of $\mathbb{X}$ with D_m+1 pieces. Let $\chi_n^2(m)$ be the statistics given by Definition 2.2.*
If ρ is positive and if $\Lambda_m=\inf\{\mu(I),I\in m\}$ is supposed to be larger than $1/n$, then for all $\varepsilon>0$,

$$(1+\varepsilon)^{-1}\frac{D_m-C(\varepsilon,\rho)}{n}\leq\mathbb{E}^2[\chi_n(m)]\leq\frac{D_m}{n}$$

where $C(\varepsilon,\rho)=2+16^2\varepsilon^{-1}\rho^{-1}$.

Proof. The upper bound is obvious. Let us now focus on the lower bound. Using (4.29), we can apply Proposition 4.2 and (4.24) to the countable set of functions $\left\{\sum_{I\in m} a_I\varphi_I\,, a\in B_m'\right\}$ where we recall that B_m' is a countable and dense subset of the set B_m. Here, $\sigma^2\leq 1$ and

$$b=\sup_{a\in B_m'}\left\|\sum_{I\in m} a_I\varphi_I\right\|_\infty \leq \sup_{I\in m} P\left(I\right)^{-\frac{1}{2}} \leq \left(\frac{n}{\rho}\right)^{\frac{1}{2}}$$

thus, one has

$$\mathbb{E}\left[\chi_n^2\left(m\right)\right]-\left[\mathbb{E}\left(\chi_n\left(m\right)\right)\right]^2\leq 2n^{-1}+32(n\rho)^{-1/2}\mathbb{E}[\chi_n\left(m\right)].$$

Let ε be some positive number, the following inequality $2ab\leq\varepsilon a^2+\varepsilon^{-1}b^2$ leads to

$$\mathbb{E}^2\left[\chi_n\left(m\right)\right]\geq(1+\varepsilon)^{-1}\left[\mathbb{E}\left[\chi_n^2\left(m\right)\right]-16^2\varepsilon^{-1}\left(\rho n\right)^{-1}-2n^{-1}\right].$$

And, the result follows from the identity:

$$\mathbb{E}\left[\chi_n^2\left(m\right)\right]=\frac{D_m}{n}.$$

□

4.3 Exponential bound for $\nu_n\left(\log f/s\right)$

The following proposition will be used for the control of the Hellinger risk in Theorem 3.2 and Theorem 3.3.

Proposition 4.5. *Let $X_1,...,X_n$ be some independent $\mathbb{X}$-valued random variables with common distribution $P=s\mu$. Then, for all positive density f and any positive x:*

$$\mathbb{P}\left[\nu_n\left(\log\frac{f}{s}\right)\geq K\left(s,f\right)-2h^2\left(s,f\right)+2\frac{x}{n}\right]\leq\exp\left(-x\right)$$

Proof. The result derives from Markov's inequality :

$$\begin{aligned}\mathbb{P}\left[\nu_n\left(\log\frac{f}{s}\right)\geq a\right] &= \mathbb{P}\left[\exp\left(\frac{n}{2}\nu_n\left(\log\frac{f}{s}\right)\right)\geq\exp\left(\frac{na}{2}\right)\right]\\ &\leq \exp\left[-n\left(\frac{a}{2}-L\left(\frac{1}{2}\right)\right)\right]\end{aligned}$$

where $L\left(\frac{1}{2}\right)=\log\mathbb{E}\left[\exp\left(\frac{1}{2}\log\frac{f}{s}\left(X_1\right)+\frac{1}{2}K\left(s,f\right)\right)\right]$. Since,

$$\begin{aligned}L\left(\frac{1}{2}\right) &= \frac{1}{2}K\left(s,f\right)+\log\int\sqrt{sf}d\mu=\frac{1}{2}K\left(s,f\right)+\log\left[1-h^2\left(s,f\right)\right]\\ &\leq \frac{1}{2}K\left(s,f\right)-h^2\left(s,f\right).\end{aligned}$$

we take $a=K\left(s,f\right)-2h^2\left(s,f\right)+2\frac{x}{n}$ to conclude. □

5 Proofs of the theorems

We introduce some set $\Omega_n[A]$ that will turn to be usefull several times in the proofs. The set $\Omega_n[A]$ is defined by:

$$\Omega_n[A] = \bigcap_{m\in\mathcal{M}_n} \{|P_n(I) - P(I)| \le AP(I),\ \forall I \in m\} \tag{5.30}$$

where A is a positive number in $]0,1[$ to be chosen adequately later. Given some partition $m \in \mathcal{M}_n$, we recall that s_m denotes the Kullback-Leibler projection, defined by (2.5) and $\hat{s}_m$ the histogram estimator, defined by (1.1) on $\mathcal{S}_m$. The main point is that, on $\Omega_n[A]$ one has

$$(1-A)\,s_m \le \hat{s}_m \le (1+A)\,s_m$$

for every partition $m \in \mathcal{M}_n$. Consequently, on this set, we can link the Kullback-Leibler information between s_m and $\hat{s}_m$ and the statistics given by Definition 2.2 via (2.13):

$$\frac{1-A}{2(1+A)^2}\chi_n^2(m) \le K(\hat{s}_m, s_m) \le \frac{1+A}{2(1-A)^2}\chi_n^2(m) \tag{5.31}$$

uniformly over $m \in \mathcal{M}_n$. This will be usefull in the sequel.
Furthermore, we can control the probability of this set by Bernstein's inequality:
- for the family satisfying $(\mathbf{H_1})$,

$$\begin{aligned}
\mathbb{P}[\Omega_n^c[A]] &\le \sum_{m\in\mathcal{M}_n}\sum_{I\in m}\mathbb{P}[|P_n(I)-P(I)| > AP(I)] \\
&\le 2\sum_{m\in\mathcal{M}_n}\sum_{I\in m}\exp\left[-nA'P(I)\right] \\
&\le 2|\mathcal{M}_n|\,\Lambda_n^{-1}\exp\left[-n\Lambda_n\rho A'\right]
\end{aligned}$$

where $A' = A[2(1+A/3)]^{-1}$. And, using $|\mathcal{M}_n| \le \Lambda_n^{-(a+1)}$ we obtain:

$$\mathbb{P}[\Omega_n^c[A]] \le 2\Lambda_n^{-(a+2)}\exp\left[-n\Lambda_n\rho A'\right]. \tag{5.32}$$

- for the family satisfying $(\mathbf{H_2})$,

$$\mathbb{P}[\Omega_n^c[A]] \le \sum_{I\in m_n}\mathbb{P}[|P_n(I)-P(I)| > AP(I)] \le 2\Lambda_n^{-1}\exp\left[-n\Lambda_n\rho A'\right] \tag{5.33}$$

Finally, in both cases, the hypothesis $\Lambda_n \ge B(\log n)^2/n$ leads to

$$\mathbb{P}[\Omega_n^c[A]] \le \frac{C(\rho, A, B)}{n}. \tag{5.34}$$

We now turn to the proofs.

5.1 Minimal value for the penalty function

This section is devoted to the demonstrations of Theorem 3.2 and Theorem 3.4, since the two demonstrations follows the same stages.
Let $m \in \mathcal{M}_n$, and $\varepsilon > 0$ that will be chosen either as a function of c for the proof of Theorem 3.4 or as a function of c_1 for the proof of Theorem 3.2.

• The study of the risk of the penalized maximum likelihood estimator $\tilde{s}$ defined by (3.17) derives from an elementary inequality. By definition of $\tilde{s}$ we have:

$$\gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m) \leq \gamma_n(s_m) + \mathrm{pen}(m).$$

Furthermore, if $\int s \log s \, d\mu$ is finite, then for all density t such that $\int s \log t \, d\mu$ is finite:

$$\gamma_n(t) = -\nu_n(\log t) - \int s \log t \, d\mu = K(s,t) - \int s \log s \, d\mu - \nu_n(\log t).$$

where ν_n is the centered empirical measure defined by (4.25). This leads to the following inequality on the Kullback-Leibler information between s and $\tilde{s}$:

$$K(s,\tilde{s}) \leq K(s,s_m) + \nu_n(\log \tilde{s} - \log s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \tag{5.35}$$

We control the term $\nu_n\left(\log \frac{\tilde{s}}{s_m}\right)$ in (5.35) by decomposing

$$\nu_n\left(\log \frac{\tilde{s}}{s_m}\right) = \nu_n\left(\log \frac{\hat{s}_{\hat{m}}}{s_{\hat{m}}}\right) + \nu_n\left(\log \frac{s_{\hat{m}}}{s}\right) + \nu_n\left(\log \frac{s}{s_m}\right).$$

Our purpose is to control $\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right)$ and $\nu_n\left(\log \frac{s_{m'}}{s}\right)$ uniformly for all $m' \in \mathcal{M}_n$. For this purpose we define V_s by:

$$V_s^2(p,q) = \int s\left(\log \frac{p}{q}\right)^2 d\mu$$

for any densities p and q such that $\log \frac{p}{q} \in \mathbb{L}^2(sd\mu)$. And we introduce a sequence of positive weights $(x_m)_{m\in\mathcal{M}_n}$ which satisfies

$$\sum_{m\in\mathcal{M}_n} \exp(-x_m) < \infty.$$

Let $m' \in \mathcal{M}_n$.

• For the term $\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right)$, in order to apply proposition 4.3, we write:

$$\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) \leq \sup_{t\in\mathcal{S}_{m'}} \left|\nu_n\left(\frac{\log t - \log s_{m'}}{V_s(t,s_{m'})}\right)\right| V_s(\hat{s}_{m'}, s_{m'}).$$

and if we note for all $I \in m$, $\varphi_I = P(I)^{-\frac{1}{2}} 1\!\!1_I$ then

$$\sup_{t\in\mathcal{S}_{m'}} \left|\nu_n\left(\frac{\log t - \log s_{m'}}{V_s(t,s_{m'})}\right)\right| = \sup_{a\in\mathbb{R}^{m'},\, |a|_2=1} \left|\sum_{I\in m'} a_I \nu_n(\varphi_I)\right| = \left[\sum_{I\in m'} \nu_n^2(\varphi_I)\right]^{1/2} = \chi_n(m')$$

according to Definition 2.2 of $\chi_n^2(m')$.

Let A be some number in $]0,1[$ that will be chosen later as a function of ε.

Then we introduce the set

$$\Omega_1 = \left\{\chi_n(m')\, 1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)\left[\sqrt{\frac{D_{m'}}{n}} + \sqrt{\frac{2\kappa x_{m'}}{n}}\right],\ \forall m' \in \mathcal{M}_n\right\}$$

where $\Omega_n [A]$ is the set defined by (5.30). Provided that $A \leq \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$, we derive from proposition 4.3 that

$$\mathbb{P}\left(\Omega_1^c\right) \leq \sum_{m' \in \mathcal{M}_n} e^{-x_{m'}}. \tag{5.36}$$

On the set Ω_1 one has

$$\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) 1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)\left[\sqrt{\frac{D_{m'}}{n}} + \sqrt{\frac{2\kappa x_{m'}}{n}}\right] V_s\left(\hat{s}_{m'}, s_{m'}\right).$$

Moreover, since $\hat{s}_{m'} \geq (1-A)\, s_{m'}$ on the set $\Omega_n [A]$, by an application of Lemma 2.3 we get on this set and provided that $A < 1$,

$$K\left(s_{m'}, \hat{s}_{m'}\right) \geq \frac{1-A}{2} \int s_{m'} \left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right)^2 d\mu.$$

Moreover, the function $\left(\log \hat{s}_{m'}/s_{m'}\right)^2$ belongs to $S_{m'}$, thus (2.4) leads to

$$\int s_{m'} \left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right)^2 d\mu = V_s^2\left(\hat{s}_{m'}, s_{m'}\right).$$

Finally, on $\Omega_n [A]$, one has

$$K\left(s_{m'}, \hat{s}_{m'}\right) \geq \frac{1-A}{2} V_s^2\left(\hat{s}_{m'}, s_{m'}\right).$$

Thus on Ω_1,

$$\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) 1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)\left[\sqrt{\frac{D_{m'}}{n}} + \sqrt{\frac{2\kappa x_{m'}}{n}}\right] \left[\frac{2}{1-A} K\left(s_{m'}, \hat{s}_{m'}\right)\right]^{\frac{1}{2}}.$$

Now, we will use the following elementary inequalities which hold for any positive number η, and any numbers a and b:

$$2ab \leq \eta a^2 + \eta^{-1} b^2 \tag{5.37}$$

$$(a+b)^2 \leq (1+\eta)\, a^2 + \left(1+\eta^{-1}\right) b^2. \tag{5.38}$$

If we take

$$A = \frac{\varepsilon}{1+\varepsilon} \wedge \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$$

we obtain on Ω_1 (using (5.37) with $\eta = 1+\varepsilon$ and $b = K\left(s_{m'}, \hat{s}_{m'}\right)^{1/2}$)

$$\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) 1\!\!1_{\Omega_n[A]} \leq \frac{1}{2}(1+\varepsilon)^4 \left[\sqrt{\frac{D_{m'}}{n}} + \sqrt{\frac{2\kappa x_{m'}}{n}}\right]^2 + \frac{1}{1+\varepsilon} K\left(s_{m'}, \hat{s}_{m'}\right).$$

Given $\zeta > 0$, we choose

$$x_{m'} = L_{m'} D_{m'} + \zeta$$

where the $L_{m'}$ are chosen, either as some constant L that will depend on M and ε for the proof of Theorem 3.4, or such that the condition (3.19) holds for the proof of Theorem 3.2.

Finally, we get on Ω_1 (using (5.38) with $\eta = \varepsilon$)

$$\begin{aligned}\nu_n\left(\log\frac{\hat{s}_{m'}}{s_{m'}}\right)\mathbb{1}_{\Omega_n[A]} &\leq \frac{1}{2}(1+\varepsilon)^5\left(1+\sqrt{2\kappa L_{m'}}\right)^2\frac{D_{m'}}{n}+\frac{1}{1+\varepsilon}K\left(s_{m'},\hat{s}_{m'}\right)\\ &+ \frac{\kappa(1+\varepsilon)^5}{\varepsilon}\frac{\zeta}{n}.\end{aligned}\tag{5.39}$$

• For the Kullback-Leibler risk (Theorem 3.4), we can control $\nu_n\left(\log\frac{s_{m'}}{s}\right)$ by an application of Bernstein's inequality,

$$\mathbb{P}\left[\nu_n\left(\log\frac{s_{m'}}{s}\right)\geq V_s\left(s,s_{m'}\right)\sqrt{2\frac{x_{m'}}{n}}+\frac{M}{3}\frac{x_{m'}}{n}\right]\leq\exp\left(-x_{m'}\right)$$

since $\left\|\log\frac{s}{s_{m'}}\right\|_\infty\leq M$ and $\mathrm{Var}\left[\log\frac{s}{s_{m'}}(X_1)\right]\leq V_s^2\left(s,s_{m'}\right)$. Thus, if we consider the set

$$\Omega_2=\left\{\nu_n\left(\log\frac{s_{m'}}{s}\right)\leq V_s\left(s,s_{m'}\right)\sqrt{2\frac{x_{m'}}{n}}+\frac{M}{3}\frac{x_{m'}}{n},\ \forall m'\in\mathcal{M}_n\right\}$$

we get

$$\mathbb{P}\left(\Omega_2^c\right)\leq\sum_{m'\in\mathcal{M}_n}e^{-x_{m'}}.\tag{5.40}$$

Furthermore, by Lemma 2.3,

$$K\left(s,s_{m'}\right)\geq\frac{e^{-M}}{2}\int s\left(\log\frac{s}{s_{m'}}\right)^2 d\mu=\frac{e^{-M}}{2}V_s^2\left(s,s_{m'}\right)$$

since $s_{m'}\geq e^{-M}s$. Then, on Ω_2 we get,

$$\nu_n\left(\log\frac{s_{m'}}{s}\right)\leq\left[2e^M K\left(s,s_{m'}\right)\right]^{\frac{1}{2}}\sqrt{2\frac{x_{m'}}{n}}+\frac{M}{3}\frac{x_{m'}}{n}.$$

Using inequality (5.37) (with $\eta=1+\varepsilon$ and $b=K\left(s,s_{m'}\right)^{1/2}$) we obtain

$$\nu_n\left(\log\frac{s_{m'}}{s}\right)\leq\frac{1}{1+\varepsilon}K\left(s,s_{m'}\right)+\left[(1+\varepsilon)e^M+\frac{M}{3}\right]\left[L\frac{D_{m'}}{n}+\frac{\zeta}{n}\right]\tag{5.41}$$

Finally, combining inequality (5.35) with (5.39) and (5.41) applied to $m'=\hat{m}$, we deduce that on $\Omega_1\cap\Omega_2$:

$$\begin{aligned}K\left(s,\hat{s}_{\hat{m}}\right)\mathbb{1}_{\Omega_n[A]} &\leq K\left(s,s_m\right)+\mathrm{pen}\left(m\right)-\mathrm{pen}\left(\hat{m}\right)+\nu_n\left(\log\frac{s}{s_m}\right)\mathbb{1}_{\Omega_n[A]}\\ &+ \frac{1}{2}(1+\varepsilon)^5\left[\left(1+\sqrt{2\kappa L}\right)^2+\frac{2L\left((1+\varepsilon)e^M+M/3\right)}{(1+\varepsilon)^5}\right]\frac{D_{\hat{m}}}{n}\\ &+ \frac{1}{1+\varepsilon}\left[K\left(s,s_{\hat{m}}\right)+K\left(s_{\hat{m}},\hat{s}_{\hat{m}}\right)\right]\mathbb{1}_{\Omega_n[A]}+C\left(\varepsilon,M\right)\frac{\zeta}{n}\end{aligned}$$

where $C\left(\varepsilon,M\right)=(1+\varepsilon)e^M+M/3+\kappa(1+\varepsilon)^5/\varepsilon$.
Now, we choose L depending on ε and M such that:

$$\left(1+\sqrt{2\kappa L}\right)^2+\frac{2L\left[(1+\varepsilon)e^{2M}+M/3\right]}{(1+\varepsilon)^5}=1+\varepsilon.$$

And we take ε such that $c = \frac{1}{2}(1+\varepsilon)^6$ in order to get

$$\frac{1}{2}(1+\varepsilon)^6 \frac{D_{\hat{m}}}{n} - \text{pen}(\hat{m}) \leq 0.$$

Next, we deduce from inequalities (5.36) and (5.40) that

$$\mathbb{P}\left(\Omega_1^c \cup \Omega_2^c\right) \leq 2 \sum_{m' \in \mathcal{M}_n} e^{-x_{m'}} \leq 2B' \sum_{D \geq 1} D^a e^{-LD} e^{-\zeta}$$

thus $\mathbb{P}(\Omega_1 \cap \Omega_2) \geq 1 - 2B\Sigma(L) e^{-\zeta}$ where $\Sigma(L) = B' \sum_{D \geq 1} D^a e^{-LD}$.
And, using the Pythagore type identity (2.6), we get with a probability larger than $1 - 2\Sigma(L) e^{-\zeta}$,

$$\frac{\varepsilon}{1+\varepsilon} K(s, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_n(\varepsilon)} \leq K(s, s_m) + \text{pen}(m) + \nu_n\left(\log \frac{s}{s_m}\right) \mathbb{1}_{\Omega_n[A]} + C(\varepsilon, M) \frac{\zeta}{n}. \quad (5.42)$$

Now,

$$\left|\mathbb{E}\left[\nu_n\left(\log \frac{s}{s_m}\right) \mathbb{1}_{\Omega_n[A]}\right]\right| = \left|\mathbb{E}\left[\nu_n\left(\log \frac{s}{s_m}\right) \mathbb{1}_{\Omega_n^c[A]}\right]\right| \leq 2M\mathbb{P}\left[\Omega_n^c[A]\right].$$

Consequently, (5.34) and an integration of (5.42) allow to conclude that

$$\mathbb{E}\left[K(s, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_n[A]}\right] \leq \frac{1+\varepsilon}{\varepsilon}\left[K(s, s_m) + \text{pen}(m)\right] + \frac{C(\varepsilon, \rho, M, B)}{n}.$$

And the proof of Theorem 3.4 is complete.

• For the Hellinger risk (Theorem 3.2), in order to control the fluctuation $\nu_n(\log(s_{m'}/s))$ uniformly for all $m' \in \mathcal{M}_n$, without the hypothesis on the uniform control of $\|\log(s/s_{m'})\|_\infty$, we use the exponential bound described in Proposition 4.5. Thus, if we consider the set

$$\Omega'_2 = \left\{\nu_n\left(\log \frac{s_{m'}}{s}\right) \leq K(s, s_{m'}) - 2h^2(s, s_{m'}) + 2\frac{x_{m'}}{n}, \forall m' \in \mathcal{M}_n\right\}$$

we get

$$\mathbb{P}\left({\Omega'}_2^c\right) \leq \sum_{m' \in \mathcal{M}_n} \exp(-x_{m'}) \quad (5.43)$$

Now, using (5.35) and (5.39) together with $m' = \hat{m}$, we derive that on $\Omega_1 \cap \Omega'_2$,

$$\begin{aligned} K(s, \hat{s}_{\hat{m}}) \mathbb{1}_{\Omega_n[A]} &\leq K(s, s_m) + \text{pen}(m) - \text{pen}(\hat{m}) + \nu_n\left(\log \frac{s}{s_m}\right) \mathbb{1}_{\Omega_n[A]} \\ &+ \frac{1}{2}(1+\varepsilon)^5 \left[\left(1 + \sqrt{2\kappa L_{\hat{m}}}\right)^2 + 4L_{\hat{m}}\right] \frac{D_{\hat{m}}}{n} + C(\varepsilon) \frac{\zeta}{n} \\ &+ \left[\frac{1}{1+\varepsilon} K(s_{\hat{m}}, \hat{s}_{\hat{m}}) + K(s, s_{\hat{m}}) - 2h^2(s, s_{\hat{m}})\right] \mathbb{1}_{\Omega_n[A]} \end{aligned}$$

with $C(\varepsilon) = \kappa(1+\varepsilon)^5/\varepsilon + 2$.
Now, $\left(1 + \sqrt{2\kappa L_{\hat{m}}}\right)^2 + 4L_{\hat{m}} \leq \left(1 + \sqrt{c_2 L_{\hat{m}}}\right)^2$ and we choose ε such that $c_1 = \frac{1}{2}(1+\varepsilon)^5$ in such a way that

$$\frac{1}{2}(1+\varepsilon)^5 \left[\left(1 + \sqrt{2\kappa L_{\hat{m}}}\right)^2 + 4L_{\hat{m}}\right] \frac{D_{\hat{m}}}{n} - \text{pen}(\hat{m}) \leq 0.$$

Consequently, one has,

$$\begin{aligned}\left[2h^2\left(s,s_{\hat{m}}\right)+\frac{\varepsilon}{1+\varepsilon}K\left(s_{\hat{m}},\hat{s}_{\hat{m}}\right)\right]1\!\!1_{\Omega_n(\varepsilon)} &\le K\left(s,s_m\right)+\mathrm{pen}\left(m\right)+C\left(\varepsilon\right)\frac{\zeta}{n}\\ &+\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n[A]}.\end{aligned}$$

on $\Omega_1\cap\Omega'_2$. Moreover, we derive from (5.36) and (5.43) that

$$\mathbb{P}\left(\Omega_1^c\cup\Omega'^c_2\right)\le 2\sum_{m'\in\mathcal{M}_n}e^{-x_{m'}}\le 2\Sigma e^{-\zeta}$$

where Σ is defined by (3.19). Thus, using (2.7) we get with a probability larger than $1-2\Sigma e^{-\zeta}$,

$$\begin{aligned}\frac{2\varepsilon}{1+\varepsilon}\left[h^2\left(s,s_{\hat{m}}\right)+h^2\left(s_{\hat{m}},\hat{s}_{\hat{m}}\right)\right]1\!\!1_{\Omega_n[A]} &\le K\left(s,s_m\right)+\mathrm{pen}\left(m\right)+C\left(\varepsilon\right)\frac{\zeta}{n}\\ &+\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n[A]}.\end{aligned}\tag{5.44}$$

Now, in order to integrate this inequality we shall control $\mathbb{E}\left[\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n[A]}\right]$,

$$\begin{aligned}\left|\mathbb{E}\left[\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n[A]}\right]\right| &= \left|\mathbb{E}\left[\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n^c[A]}\right]\right|\\ &\le \left[\frac{1}{n}\int s\left(\log\frac{s}{s_m}\right)^2d\mu\right]^{\frac{1}{2}}\left[\mathbb{P}\left(\Omega_n^c\left[A\right]\right)\right]^{\frac{1}{2}}.\end{aligned}$$

And

$$\begin{aligned}\int s\left(\log\frac{s}{s_m}\right)^2d\mu &\le 2\left[\int s\left(\log s\right)^2d\mu+\int s\left(\log s_m\right)^2d\mu\right]\\ &\le 2\left[\int s\left(\log s\right)^2d\mu+\left(\log\left(\frac{1}{\rho}\vee n\right)\right)^2\right]\end{aligned}$$

since $\rho\le s_m\le n$. Finally, we derive from (5.32) or (5.33) and the hypothesis $\Lambda_n\ge B\left(\log n\right)^2/n$ that

$$\left|\mathbb{E}\left[\nu_n\left(\log\frac{s}{s_m}\right)1\!\!1_{\Omega_n[A]}\right]\right|\le\frac{C\left(\varepsilon,\rho,S,B\right)}{n}.\tag{5.45}$$

Then, by integration of (5.44)

$$\mathbb{E}\left[2\left(h^2\left(s,s_{\hat{m}}\right)+h^2\left(s_{\hat{m}},\hat{s}_{\hat{m}}\right)\right)1\!\!1_{\Omega_n[A]}\right]\le\frac{1+\varepsilon}{\varepsilon}\left[K\left(s,s_m\right)+\mathrm{pen}\left(m\right)\right]+\frac{C\left(\varepsilon,\rho,S,\Sigma,B\right)}{n}.$$

Moreover,

$$h^2\left(s,\hat{s}_{\hat{m}}\right)\le 2h^2\left(s,s_{\hat{m}}\right)+2h^2\left(s_{\hat{m}},\hat{s}_{\hat{m}}\right)$$

leads to the control of $\mathbb{E}\left[h^2\left(s,\hat{s}_{\hat{m}}\right)1\!\!1_{\Omega_n[A]}\right]$. Furthermore, with the control of $\mathbb{P}\left[\Omega_n^c\left[A\right]\right]$ given by (5.34), one also has

$$\mathbb{E}\left[h^2\left(s,\hat{s}_{\hat{m}}\right)1\!\!1_{\Omega_n^c[A]}\right]\le\frac{C\left(\varepsilon,\rho,B\right)}{n}$$

since $h^2\left(s,\hat{s}_{\hat{m}}\right)\le 1$. Hence, we conclude that

$$\mathbb{E}\left[h^2\left(s,\hat{s}_{\hat{m}}\right)\right]\le\frac{1+\varepsilon}{\varepsilon}\left[K\left(s,s_m\right)+\mathrm{pen}\left(m\right)\right]+\frac{C\left(\varepsilon,\rho,S,\Sigma,B\right)}{n}.$$

This achieves the proof of Theorem 3.2.

5.2 Optimizing the penalty function

This section is devoted to the demonstration of Theorem 3.3.
Let $m \in \mathcal{M}_n$, and $\varepsilon > 0$. We modify the basic inequality (5.35) using

$$\gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m) = \gamma_n(s_m) + \mathrm{pen}(m) - K(\hat{s}_m, s_m)$$

since $\gamma_n(\hat{s}_m) - \gamma_n(s_m) = -\int \log \frac{\hat{s}_m}{s_m} dP_n = -\int \hat{s}_m \log \frac{\hat{s}_m}{s_m} d\mu = -K(\hat{s}_m, s_m)$, using (2.3).
Consequently,

$$K(s, \tilde{s}) \leq K(s, s_m) + \nu_n(\log \tilde{s} - \log s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) - K(\hat{s}_m, s_m). \tag{5.46}$$

Let $\eta \in]0, 1[$ to be chosen later as a function of ε, one has

$$\begin{aligned} K(s, s_{\hat{m}}) + (1-\eta) K(s_{\hat{m}}, \hat{s}_{\hat{m}}) &\leq K(s, s_m) + \mathrm{pen}(m) + \nu_n\left(\log \frac{s}{s_m}\right) - K(\hat{s}_m, s_m) \\ &+ \nu_n\left(\log \frac{\hat{s}_{\hat{m}}}{s_{\hat{m}}}\right) - \eta K(s_{\hat{m}}, \hat{s}_{\hat{m}}) + \nu_n\left(\log \frac{s_{\hat{m}}}{s}\right) - \mathrm{pen}(\hat{m}). \end{aligned}$$

We consider again the set $\Omega_n[A]$ defined by (5.30) with some number $A \in]0, 1[$ which will be chosen later as a function of ε. Our aim is to control $\nu_n(\log(\hat{s}_{m'}/s_{m'})) - \eta K(s_{m'}, \hat{s}_{m'})$ uniformly over all $m' \in \mathcal{M}_n$, on the set $\Omega_n[A]$.
Let $m' \in \mathcal{M}_n$. On the set $\Omega_n[A]$, one has on the one hand, via (4.26) and (2.11),

$$\begin{aligned} \nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) &= \int \log \frac{\hat{s}_{m'}}{s_{m'}} (\hat{s}_{m'} - s_{m'}) \, d\mu \\ &= \int \left(\frac{\log(\hat{s}_{m'}/s_{m'})}{\hat{s}_{m'}/s_{m'} - 1}\right) \frac{(\hat{s}_{m'} - s_{m'})^2}{s_{m'}} d\mu \\ &\leq (1-A)^{-1} \int \frac{(\hat{s}_{m'} - s_{m'})^2}{s_{m'}} d\mu \end{aligned}$$

Hence, according to Definition 2.2 and (2.12),

$$\nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) \mathbb{1}_{\Omega_n[A]} \leq (1-A)^{-1} \chi_n^2(m') \, \mathbb{1}_{\Omega_n[A]} \tag{5.47}$$

And on the other hand, one has via (5.31)

$$K(s_{m'}, \hat{s}_{m'}) \, \mathbb{1}_{\Omega_n[A]} \geq \frac{1-A}{2(1+A)^2} \chi_n^2(m') \, \mathbb{1}_{\Omega_n[A]}. \tag{5.48}$$

Combining inequalities (5.46), (5.47) and (5.48) we get

$$\begin{aligned} \left[K(s, s_{\hat{m}}) + (1-\eta) K(s_{\hat{m}}, \hat{s}_{\hat{m}})\right] \mathbb{1}_{\Omega_n[A]} &\leq \left[\nu_n\left(\log \frac{s_{\hat{m}}}{s}\right) + B(\varepsilon, \eta) \chi_n^2(\hat{m})\right] \mathbb{1}_{\Omega_n[A]} \\ &+ K(s, s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + U_m \end{aligned} \tag{5.49}$$

where

$$B(\varepsilon, \eta) = \frac{1}{1-A} - \frac{\eta}{2} \frac{1-A}{(1+A)^2}$$

and U_m is the random variable given by

$$U_m = \left[\nu_n\left(\log \frac{s}{s_m}\right) - K(\hat{s}_m, s_m)\right] \mathbb{1}_{\Omega_n[A]}.$$

Next, applying Proposition 4.3 we derive that

$$\mathbb{P}\left[\chi_n\left(m'\right)1\!\!1_{\Omega_n[A]} \geq (1+\varepsilon)\left(\sqrt{\frac{D_{m'}}{n}}+\sqrt{\frac{2\kappa x_{m'}}{n}}\right)\right] \leq \exp\left(-x_{m'}\right)$$

provided that $A \leq \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$.
We choose

$$x_{m'} = L_{m'}D_{m'} + \zeta$$

and we introduce the set

$$\Omega_1 = \left\{\chi_n\left(m'\right)1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)\left[\left(1+\sqrt{2\kappa L_{m'}}\right)\sqrt{\frac{D_{m'}}{n}}+\sqrt{\frac{2\kappa\zeta}{n}}\right], \forall m' \in \mathcal{M}_n\right\}.$$

Then $\mathbb{P}\left(\Omega_1^c\right) \leq \Sigma e^{-\zeta}$ and on Ω_1,

$$\chi_n^2\left(m'\right)1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)^3\left(1+\sqrt{2\kappa L_{m'}}\right)^2\frac{D_{m'}}{n} + \frac{2\kappa(1+\varepsilon)^3}{\varepsilon}\frac{\zeta}{n}.$$

Moreover, defining

$$\Omega_2 = \left\{\nu_n\left(\log\frac{s_{m'}}{s}\right) \leq K\left(s, s_{m'}\right) - 2h^2\left(s, s_{m'}\right) + 2L_{m'}\frac{D_{m'}}{n} + 2\frac{\zeta}{n}, \forall m' \in \mathcal{M}_n\right\}$$

we derive from Proposition 4.5 that $\mathbb{P}\left(\Omega_2^c\right) \leq \Sigma e^{-\zeta}$. Finally, on $\Omega_1 \cap \Omega_2$, that is with a probability larger than $1 - 2\Sigma e^{-\zeta}$, (5.49) gives

$$\begin{aligned}\left[2h^2\left(s, s_{\hat{m}}\right) + (1-\eta)K\left(s_{\hat{m}}, \hat{s}_{\hat{m}}\right)\right]1\!\!1_{\Omega_n[A]} &\leq K\left(s, s_m\right) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \\ &+ \left[C_1(\varepsilon,\eta)\left(1+\sqrt{2\kappa L_{\hat{m}}}\right)^2 + 2L_{\hat{m}}\right]\frac{D_{\hat{m}}}{n} \\ &+ U_m + C_2(\varepsilon,\eta)\frac{\zeta}{n}. \end{aligned} \tag{5.50}$$

with

$$C_1(\varepsilon,\eta) = B(\varepsilon,\eta)(1+\varepsilon)^3 \text{ and } C_2(\varepsilon,\eta) = 2\kappa B(\varepsilon,\eta)(1+\varepsilon)^3/\varepsilon + 2.$$

Now, let us choose η depending on ε (via A) such that $B(\varepsilon,\eta) = (1+\varepsilon)^{-3}$ i.e.

$$\eta = 2\left(\frac{1+A}{1-A}\right)^2\left(1 - \frac{1-A}{(1+\varepsilon)^3}\right)$$

By elementary analysis, we see that $\eta \in]0,1[$ for all $\varepsilon \leq 0.23$. Furthermore, for these values of ε we have $\frac{2\kappa\varepsilon}{\kappa(\varepsilon)} < 1$, therefore we can take

$$A = \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}.$$

Moreover, on the set $\Omega_n[A]$, one has, via (2.14)

$$h^2\left(s, \hat{s}_{\hat{m}}\right) \leq (1+A)^{1/2}h^2\left(s, s_{\hat{m}}\right) + h^2\left(s_{\hat{m}}, \hat{s}_{\hat{m}}\right)$$

and $(1+A)^{1/2} \leq [1-\eta]^{-1}$. Then, we derive from (2.7), and (5.50) and the choice of the penalty function, that with a probability larger than $1 - 2\Sigma e^{-\zeta}$:

$$2h^2\left(s, \hat{s}_{\hat{m}}\right)1\!\!1_{\Omega_n[A]} \leq \frac{1}{1-\eta}\left[K\left(s, s_m\right) + \mathrm{pen}(m) + U_m + 2\left(\frac{\kappa}{\varepsilon}+1\right)\frac{\zeta}{n}\right]. \tag{5.51}$$

It remains to control $\mathbb{E}(U_m)$. Using Lemma 2.3 and (2.11), one has, see (5.48),

$$K(\hat{s}_m, s_m)\,\mathbb{1}_{\Omega_n[A]} \geq \frac{1-A}{2(1+A)^2}\chi_n^2(m)\,\mathbb{1}_{\Omega_n[A]}$$

Thus,

$$\mathbb{E}\left[K(\hat{s}_m, s_m)\,\mathbb{1}_{\Omega_n[A]}\right] \geq \frac{1-A}{2(1+A)^2}\left[\frac{D_m}{n} - \mathbb{E}\left(\chi_n^2(m)\,\mathbb{1}_{\Omega_n^c[A]}\right)\right]$$

and, using $\chi_n^2(m) \leq 2\sup_{I\in m} P(I)^{-1} \leq 2(\rho\Lambda_n)^{-1}$ together with (5.34) we get

$$\mathbb{E}\left[\chi_n^2(m)\,\mathbb{1}_{\Omega_n^c[A]}\right] \leq \frac{C(\varepsilon,\rho,B)}{n}.$$

Furthermore, we can show, as in the proof of Theorem 3.2, see (5.45), that

$$\left|\mathbb{E}\left[\nu_n\left(\log\frac{s}{s_m}\right)\mathbb{1}_{\Omega_n[A]}\right]\right| \leq \frac{C(\varepsilon,\rho,S,B)}{n}$$

Consequently,

$$\mathbb{E}(U_m) \leq -\frac{1-A}{2(1+A)^2}\frac{D_m}{n} + \frac{C(\varepsilon,\rho,S,B)}{n}.$$

Integrating (5.51), this leads to:

$$\begin{aligned}\mathbb{E}\left[2h^2(s,\hat{s}_{\hat{m}})\,\mathbb{1}_{\Omega_n[A]}\right] &\leq \frac{1}{1-\eta}\left[K(s,s_m) + \mathrm{pen}(m) - \frac{1-A}{2(1+A)^2}\frac{D_m}{n}\right]\\ &+ \frac{1}{1-\eta}\left[4\Sigma\left(\frac{\kappa}{\varepsilon}+1\right) + C(\varepsilon,\rho,S,B)\right]\frac{1}{n}.\end{aligned}$$

So setting

$$C(\varepsilon) = \frac{1}{1-\eta}\left(\frac{1+5A+2A^2}{(1+A)^2}\right)$$

we obtain for all $\varepsilon \leq 0.23$,

$$\mathbb{E}\left[2h^2(s,\hat{s}_{\hat{m}})\,\mathbb{1}_{\Omega_n[A]}\right] \leq C(\varepsilon)\left[K(s,s_m) + \left(\frac{1}{2}+\sqrt{2c_2L_m}\right)\frac{D_m}{n}\right] + \frac{C(\varepsilon,\rho,S,\Sigma,B)}{n}.$$

Since

$$\mathbb{E}\left[2h^2(s,\hat{s}_{\hat{m}})\,\mathbb{1}_{\Omega_n^c(\varepsilon)}\right] \leq 2\mathbb{P}\left[\Omega_n^c(\varepsilon)\right] \leq \frac{C(\varepsilon,\rho,B)}{n}$$

one derives

$$\mathbb{E}\left[2h^2(s,\hat{s}_{\hat{m}})\right] \leq C(\varepsilon)\left[K(s,s_m) + \left(\frac{1}{2}+\sqrt{2c_2L_m}\right)\frac{D_m}{n}\right] + \frac{C(\varepsilon,\rho,S,\Sigma)}{n}.$$

This achieves the proof of Theorem 3.2.

5.3 Proof of theorem 3.5

We will show that, with high probability, the quantity $\mathrm{crit}\,(m) - \mathrm{crit}\,(m_N)$ is positive for every partition $m \in \mathcal{M}_n$ with a number of pieces bounded by D with D of order N. This will imply that $\hat{m}$ has cardinality larger than D. The proof is based on the results of section 4.2.
Since $s = 1\!\!1_{\mathbb{X}}$, we can write for every partition $m \in \mathcal{M}_n$:

$$\mathrm{crit}\,(m) = -K\,(\hat{s}_m, s_m) + \mathrm{pen}\,(m)$$

and, therefore

$$\mathrm{crit}\,(m) - \mathrm{crit}\,(m_N) \geq K\,(\hat{s}_{m_N}, s_{m_N}) - K\,(\hat{s}_m, s_m) - c\,\frac{N}{n}$$

Let ε be some number of $]0,1[$ that will be chosen later as a function of c. All along the proof we will use the constants denoted by κ, κ', $\kappa\,(\varepsilon)$, and $\kappa'\,(\varepsilon)$ which are defined in section 4.
Let $\Omega_n\,[A]$ be defined by (5.30) with $P = \mu$, where A is some real positive number to be chosen later as a function of ε. Thus, applying (5.31), on $\Omega_n\,[A]$ we get for all $m \in \mathcal{M}_n$,

$$2\,[\mathrm{crit}\,(m) - \mathrm{crit}\,(m_N)] \geq \frac{1-A}{(1+A)^2}\,\chi_n^2\,(m_N) - \frac{1+A}{(1-A)^2}\,\chi_n^2\,(m) - 2c\,\frac{N}{n}. \tag{5.52}$$

Now, we will use the concentration of the variables $\chi_n^2\,(m)$ around their expectation (see Proposition 4.3). We define

$$D = \frac{1-4c^2}{4}N \text{ and } x_m = L\left(\sqrt{D_m} + \sqrt{D}\right), \quad \text{for all } m \in \mathcal{M}_n.$$

We consider the following events:

$$\Omega_1\,(D) = \bigcap_{m \in \mathcal{M}_n:\, D_m \leq D} \left\{ \chi_n\,(m)\,1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2\kappa x_m}{n}}\right)\right\}$$

and,

$$\Omega_2\,(D) = \bigcap_{m \in \mathcal{M}_n:\, D_m > D} \left\{ \chi_n\,(m) \geq (1-\varepsilon)\,\mathbb{E}[\chi_n\,(m)] - \sqrt{\frac{2\kappa' x_m}{n}} - \frac{\kappa'\,(\varepsilon)\,x_m}{\sqrt{n}}\right\}.$$

The controls of $\mathbb{P}\,[\Omega_1^c\,(D)]$ and $\mathbb{P}\,[\Omega_2^c\,(D)]$ derive from Proposition 4.3.
First, it follows from (4.27) that

$$\mathbb{P}\left[\chi_n\,(m)\,1\!\!1_{\Omega_n[A]} \geq (1+\varepsilon)\left(\sqrt{\frac{D_m}{n}} + \sqrt{\frac{2\kappa x_m}{n}}\right)\right] \leq \exp\,(-x_m)$$

provided that $A \leq \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$. Hence,

$$\mathbb{P}\,[\Omega_1^c\,(D)] \leq \sum_{m \in \mathcal{M}_n:\, D_m \leq D} \exp\,(-x_m)\,.$$

Secondly, (4.28) implies that

$$\mathbb{P}\left[\chi_n\,(m) \leq (1-\varepsilon)\,\mathbb{E}[\chi_n\,(m)] - \sqrt{\frac{2\kappa' x_m}{n}} - \frac{\kappa'\,(\varepsilon)\,x_m}{\sqrt{n}}\right] \leq \exp\,(-x_m)$$

since $b_m = \sup_{I \in m} P(I)^{-\frac{1}{2}} \leq \sqrt{n}$. Consequently,

$$\mathbb{P}\left[\Omega_2^c(D)\right] \leq \sum_{m \in \mathcal{M}_n : D_m > D} \exp(-x_m).$$

Next, defining $\Omega(D) = \Omega_1(D) \cap \Omega_2(D)$ we get

$$\mathbb{P}\left[\Omega^c(D)\right] \leq \sum_{m \in \mathcal{M}_n} \exp(-x_m) \leq \Sigma(L) \exp\left(-L\sqrt{D}\right).$$

Now, on $\Omega(D)$, on the one hand the following inequality holds true for every partition $m \in \mathcal{M}_n$ such that $D_m \leq D$,

$$\chi_n^2(m)\, 1\!\!1_{\Omega_n[A]} \leq (1+\varepsilon)^2 \left(1 + 2\sqrt{\kappa L}\right)^2 \frac{D}{n} \tag{5.53}$$

since $x_m \leq 2L\sqrt{D} \leq 2LD$. On the other hand, for the partition m_N, one has on $\Omega(D)$,

$$\chi_n^2(m_N) \geq \left[(1-\varepsilon)\, \mathbb{E}\left[\chi_n(m_N)\right] - \left(2\sqrt{L\kappa'} + 2L\kappa'(\varepsilon)\right)\sqrt{\frac{N}{n}}\right]^2$$

since $x_{m_N} \leq 2L\sqrt{N} \leq 2LN$. Furthermore, if $L < 1$ then

$$\chi_n^2(m_N) \geq \left[(1-\varepsilon)\, \mathbb{E}\left[\chi_n(m_N)\right] - C_1(\varepsilon)\sqrt{\frac{LN}{n}}\right]^2$$

with $C_1(\varepsilon) = 2\left[\sqrt{\kappa'} + \kappa'(\varepsilon)\right]$.
Now, in order to control $\mathbb{E}\left[\chi_n(m_N)\right]$, we apply Proposition 4.4, hence for any positive number ε',

$$\mathbb{E}\left[\chi_n(m_N)\right] \geq \left(1+\varepsilon'\right)^{-\frac{1}{2}} \left(\frac{N - C(\varepsilon')}{n}\right)^{\frac{1}{2}}$$

with $C(\varepsilon') = 2 + 16^2/\varepsilon'$. If we choose $\varepsilon' = \frac{\varepsilon(2-\varepsilon)}{(1-\varepsilon)^2}$ then

$$\chi_n^2(m_N) \geq \left[C_0(\varepsilon)\left(\frac{N - C_2(\varepsilon)}{n}\right)^{\frac{1}{2}} - C_1(\varepsilon)\sqrt{\frac{LN}{n}}\right]^2$$

with $C_0(\varepsilon) = (1-\varepsilon)^2$ and $C_2(\varepsilon) = C(\varepsilon')$. Consequently, on $\Omega(D)$ we get

$$\chi_n^2(m_N) \geq C_0^2(\varepsilon)\left(1 - C_1(\varepsilon)\sqrt{L}\right)^2 \frac{N}{n}. \tag{5.54}$$

Finally, using (5.53) and (5.54), on the set $\Omega(D) \cap \Omega_n(\varepsilon)$, (5.52) becomes

$$\begin{aligned} 2\left[\operatorname{crit}(m) - \operatorname{crit}(m_N)\right] \;\geq\; & \frac{1-A}{(1+A)^2}(1-\varepsilon)^4\left(1 - C_1(\varepsilon)\sqrt{L}\right)^2 \frac{N}{n} \\ & - \frac{1+A}{(1-A)^2}(1+\varepsilon)^2\left(1+2\sqrt{\kappa L}\right)^2 \frac{D}{n} - 2c\frac{N}{n} \end{aligned}$$

valid for every partition $m \in \mathcal{M}_n$ such that $D_m \leq D$.

Now, we choose $A \leq \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}$ such that $(1-A)(1+A)^{-2} \geq 1-\varepsilon$, that is,

$$A = \frac{2\kappa\varepsilon}{\kappa(\varepsilon)} \wedge \frac{-3+2\varepsilon+(9-8\varepsilon)^{\frac{1}{2}}}{2(1-\varepsilon)}.$$

In particular, this implies that $(1+A)(1-A)^{-2} \leq (1-\varepsilon)^{-3}$.
Hence, it follows that

$$\begin{aligned} 2\left[\operatorname{crit}(m) - \operatorname{crit}(m_N)\right] &\geq (1-\varepsilon)^5 \left(1 - C_1(\varepsilon)\sqrt{L}\right)^2 \frac{N}{n} \\ &- (1-\varepsilon)^{-5}\left(1+2\sqrt{\kappa L}\right)^2 \frac{D}{n} - 2c\frac{N}{n}. \end{aligned}$$

Next, we choose L such that $\sqrt{L} = \varepsilon / C_1(\varepsilon)$. And, $C_1(\varepsilon) \geq 2\sqrt{\kappa'} \geq 2\sqrt{\kappa}$ implies $2\sqrt{\kappa L} \leq \varepsilon$. This allows us to conclude that, on the set $\Omega(D) \cap \Omega_n[A]$ and for every partition $m \in \mathcal{M}_n$ such that $D_m \leq D$:

$$2\left[\operatorname{crit}(m) - \operatorname{crit}(m_N)\right] \geq \left[(1-\varepsilon)^7 - 2c\right]\frac{N}{n} - (1-\varepsilon)^{-7}\frac{D}{n}$$

Hence, choosing ε depending on c such that $(1-\varepsilon)^7 - 2c = (1-2c)/2$ leads to

$$2\left[\operatorname{crit}(m) - \operatorname{crit}(m_N)\right] \geq \frac{1-2c}{2}\frac{N}{n} - \frac{2}{1+2c}\frac{D}{n} = 0.$$

on the set $\Omega(D) \cap \Omega_n[A]$, whenever $D = \frac{1-4c^2}{4}N$. Finally, on the set $\Omega(D) \cap \Omega_n[A]$ one has $D_{\hat{m}} \geq D$ which means that

$$\mathbb{P}\left(D_{\hat{m}} \geq \frac{1-4c^2}{4}N\right) \geq 1 - \beta(c)$$

with $\beta(c) = \mathbb{P}\left[\Omega^c(D) \cup \Omega_n^c[A]\right]$. Since, by (5.34) there exists a constant $C(\varepsilon, B)$ such that $\mathbb{P}\left[\Omega_n^c[A]\right] \leq C(\varepsilon, B)/n$, one derives that

$$\beta(c) \leq \Sigma(L)\exp\left(-L\sqrt{D}\right) + \frac{C(\varepsilon, B)}{n}.$$

Furthermore, on the set $\Omega_n[A]$ one has on the one hand,

$$K(s, \hat{s}_m) \geq \frac{1-\varepsilon}{2}\chi_n^2(m)$$

for every partition $m \in \mathcal{M}_n$ and on the set $\Omega(D)$ one has on the other hand

$$\chi_n^2(m) \geq [1-\varepsilon]^6 \frac{D_m}{n}$$

for all the partitions $m \in \mathcal{M}_n$ such that $D_m > D$. Consequently, on the set $\Omega(D) \cap \Omega_n[A]$ and for every partition $m \in \mathcal{M}_n$ such that $D_m \geq D$ we get

$$K(s, \hat{s}_m) \geq \frac{1}{2}[1-\varepsilon]^7 \frac{D_m}{n} \geq \frac{(1-2c)(1+2c)^2 N}{16n} = \delta(c)\frac{N}{n}.$$

This leads to

$$\mathbb{E}\left[K(s, \hat{s}_{\hat{m}})\right] \geq \delta(c)\,\mathbb{P}\left[\Omega(D) \cap \Omega_n[A]\right]\frac{N}{n} \geq \delta(c)\left[1-\beta(c)\right]\frac{N}{n}.$$

And the proof of Theorem 3.5 is complete.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In P. Petrov and F. Csaki (Eds.), *Proceedings 2nd International Symposium on Information Theory*, pp. 267–281. Akademia Kiado, Budapest.

Barron, A., L. Birgé, and P. Massart (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields 113*, 301–415.

Barron, A. and C. Sheu (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist. 19*, 1054–1347.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb. 65*, 181–237.

Birgé, L. and P. Massart (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli 4*, 329–375.

Birgé, L. and P. Massart (1999a). Gaussian model selection. Technical report, Université Paris-Sud.

Birgé, L. and P. Massart (1999b). A generalized cross validation criterion for density estimation. Technical report, Université Paris-Sud.

Birgé, L. and Y. Rozenholc (1999). How many bins should be put in a regular histogram. Technical Report.

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

Devroye, L. and L. Gyorfi (1985). *Nonparametric density estimation: the L_1 view.* New York: Wiley.

Freedman, D. and P. Diaconis (1981). On the histogram as a density estimator: l_2 theory. *Z. Wahrscheinlichkeitstheorie Verw. Geb. 57*, 453–476.

Hall, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Relat. Fields 85*, 449–476.

Kanazawa, Y. (1988). An optimal variable cell histogram. *Commun. Statist. Theory Meth. 17(5)*, 1401–1422.

Kanazawa, Y. (1992). An optimal variable cell histogram based on the sample spacings. *Ann. Statist. 20(1)*, 291–304.

Kanazawa, Y. (1993). Hellinger distance and Akaike's information criterion for the histogram. *Statist. Probab. Letters 17*, 293–298.

Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics 1*, 63–87.

Lugosi, G. and A. Nobel (1996). Consistency of data–driven histogram methods for density estimation and classification. *Ann. Statist. 24*, 687–706.

Massart, P. (1998). About the constants in Talagrand's concentration inequalities for empirical processes. To appear in *Ann. probab.*

Pollard, D. (1984). *Convergence of stochastic processes.* New York: Springer–Verlag.

Rudemo, M. (1982). Empirical choice of histograms and Kernel density estimators. *Scand. J. Statist. 9*, 65–78.

Scott, D. (1979). On optimal and data-based histograms. *Biometrika 66*, 605–610.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math. 126(3)*, 505–563.

Taylor, C. (1987). Akaike's information criterion and the histogram. *Biometrika 74*, 636–639.

Yu, B. and T. Speed (1992). Data compression and histograms. *Probab. Theory Relat. Fields 92*, 195–229.