and Vishne [8] conjecture that most irreducible TSRs will be primitive. The influence of the primitivity status of a particular LFSR and transformation on the LFSRs pairs or related pairs from Theorems 2 and 3 would be a useful avenue of investigation. Furthermore, the use of involutions other than $m(x) = \frac{x}{x+1}$ to search for other connections between LFSRs could be interesting.

REFERENCES

[1] E. R. Berlekamp, *Algebraic Coding Theory*. Laguna Hills, CA: Aegean Park, 1984.

[2] M. Dewar and D. Panario. Tables for linear transformation shift registers. [Online]. Available: http://www.math.carleton.ca/~daniel/research/tsr/

[3] S. Golomb, *Shift-Register Sequences*. Laguna Hills, CA: Aegean Park, 1982.

[4] D. Jungnickel, *Finite Fields: Structures and Arithmetics*. Mannheim/Leipzig/Wien/Zurich, Germany/Austria/Switzerland: Wissenschaftsverlag, 1993.

[5] R. Lidl and H. Niederreiter, *Finite Fields*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[6] V. Shoup. NTL: Number Theory Library. [Online]. Available: http://www.shoup.net/ntl/.

[7] B. Preneel, "Introduction," in *Proc. Fast Software Encryption 1994 Workshop (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 1995, vol. 1008, pp. 1–5.

[8] B. Tsaban and U. Vishne, "Efficient linear feedback shift registers with maximal period," *Finite Fields Appl.*, vol. 8, pp. 256–267, 2002.

Density Estimation Via Exponential Model Selection

Gwénaëlle Castellan

***Abstract*—We address the problem of estimating some unknown density on a bounded interval using some exponential models of piecewise polynomials. We consider a finite collection of such models based on a family of partitions. And we study the maximum-likelihood estimator built on a data-driven selected model among this collection. In doing so, we validate Akaike's criterion if the partitions that we consider are regular and we modify it if the partitions are irregular. We deduce the rate of convergence of the squared Hellinger risk of our estimator in the regular case when the logarithm of the density belongs to some Besov space.**

***Index Terms*—Adaptive density estimation, Akaike's information criterion, exponential families, Kullback–Leibler information, model selection, penalization.**

I. INTRODUCTION

Let us consider n independent and identically distributed random variables $X_1, \ldots, X_n$ with common distribution P. We assume that $dP = s\, d\mu$, where μ denotes the Lebesgue measure on a bounded interval which, for simplicity, will be taken to be $[0, 1]$. Our purpose is to estimate s using a new data-driven selection procedure among maximum-likelihood estimators on exponential models of piecewise polynomials.

The main advantage of maximizing the log likelihood on an exponential model is that one can guarantee that the resulting estimator is positive and integrates to 1. The properties of regular exponential families have been thoroughly studied (see Brown [1] and Barndorff-Nielsen [2]). The approximation of log densities by polynomials have been studied by Neyman [3] and later by Crain [4]–[6]. In particular, conditions for the existence of the maximum-likelihood estimator on exponential families of polynomials have been given in Crain [4]–[6]. Moreover, Crain [7] also gave some approximation properties of these families. In a more general framework, Portnoy [8] has studied the asymptotic behavior of the maximum-likelihood estimator (see also Cencov [9] for compact subfamilies of exponential families). The study of the particular case of log-spline models has first been developed by Stone and Koo [10]. In this context, Stone, [11] and [12], has obtained some rates of convergence with respect to $\mathbb{L}_2$ and $\mathbb{L}_\infty$ norms for the estimation of a continuous positive density. In particular, Stone [11] has found an interesting bound for the $\mathbb{L}_\infty$ norm of the approximation error associated with the expected log likelihood. At the same time, independently of Stone's papers, Barron and Sheu [13] have studied the rates of convergence of the maximum-likelihood estimator on general exponential models and applied it to the exponential families of trigonometric series, polynomials, and splines. Koo and Kim [14] have used exponential families based on wavelets to get lower bounds for the rates of convergence when the log density is assumed to belong to some Besov space.

In this correspondence, we consider a collection of exponential models of piecewise polynomials $\{\mathcal{E}_\mathfrak{m},\ \mathfrak{m} \in \mathfrak{M}_n\}$ and the corresponding family $\{\hat{s}_\mathfrak{m},\ \mathfrak{m} \in \mathfrak{M}_n\}$ of maximum-likelihood estimators on $\mathcal{E}_\mathfrak{m}$. Given a penalty function pen_n on $\mathfrak{M}_n$ (independent of the true density s), the minimization of the following criterion:

$$\mathrm{crit}_n(\mathfrak{m}) = -P_n(\log \hat{s}_\mathfrak{m}) + \mathrm{pen}_n(\mathfrak{m})$$

where P_n denotes the empirical measure, leads to some data-dependent model $\hat{\mathfrak{m}}$ and to the penalized maximum-likelihood estimator $\hat{s}_{\hat{\mathfrak{m}}}$ on the model $\mathcal{E}_{\hat{\mathfrak{m}}}$. The purpose of our work is to design a penalty function such that the resulting estimator $\hat{s}_{\hat{\mathfrak{m}}}$ behaves as well as the "best" one among the collection.

The celebrated *A Information Criterion* (AIC) due to Akaike [15] corresponds to the penalty function $D_\mathfrak{m}/n$ where $D_\mathfrak{m}$ is the number of unknown parameters for the model $\mathfrak{m}$. Akaike's heuristics are based on asymptotic information-theoretic considerations. Barron *et al.* [16] have developed a method to get nonasymptotic risk bounds for general penalized minimum contrast estimators. Unfortunately, their method cannot be applied to exponential model selection. A more specific work on penalized maximum-likelihood estimators has been examined in Yang and Barron [17]. Their method is based on a metric dimension assumption and is valid for general exponential models (providing that all the log densities of a model are uniformly bounded). Nevertheless, the penalty function that they provide involves some unrealistic constants that have repercussions on the risk bound. In this correspondence, we focus our attention on the value of these constants (trying to get the optimal ones) taking advantage of the specificity of our context in order to validate the AIC in favorable cases or to modify it when necessary. In this way, we obtain a model-selection criterion which can be used in practice for the exponential families of piecewise polynomials.

II. STATISTICAL FRAMEWORK AND RESULTS

A. Exponential Models of Piecewise Polynomial

We want to choose an adequate partition to approximate the density s by some density whose logarithm is a piecewise polynomial based on this partition. We also aim at choosing an adequate degree for the polynomial on each interval of this partition.

Manuscript received May 22, 2000; revised January 10, 2003.

The author is with the Laboratoire de Mathématiques Appliquées, F.R.E. CNRS 2222, Bât. M2, U.S.T.L., 59655 Villeneuve d'Ascq Cedex, France (e-mail: gwenaelle.castellan@univ-lille1.fr).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2003.814485

Let us first define the family of exponential models of piecewise polynomials that we shall use. Given some family $\mathcal{M}_n$ of finite partitions of $[0, 1]$ and $\mathcal{N}$ some bounded part of $\mathbb{N}$, we denote

$$\mathfrak{M}_n = \{\mathfrak{m} = (m, \underline{r}),\ m \in \mathcal{M}_n,\ \underline{r} = (r_I)_{I \in m} \in \mathcal{N}^m\}.$$

For each $\mathfrak{m} = (m, \underline{r}) \in \mathfrak{M}_n$, we define $\mathcal{P}_\mathfrak{m}$ to be the linear space of functions f on $[0, 1]$ such that, for every interval I of m, $f\mathbb{1}_x$ is a polynomial with degree r_I on I. The set $\mathcal{E}_\mathfrak{m}$ of all positive densities on $([0, 1], \mu)$ whose logarithm belongs to $\mathcal{P}_\mathfrak{m}$ is the exponential model associated with $\mathcal{P}_\mathfrak{m}$ and we define the dimension of $\mathcal{E}_\mathfrak{m}$ by

$$D_\mathfrak{m} = \dim \mathcal{P}_\mathfrak{m} - 1 = \sum_{I \in m} (r_I + 1) - 1$$

(it is the number of free parameters of $\mathcal{E}_\mathfrak{m}$). We consider hereafter the family of models $\{\mathcal{E}_\mathfrak{m},\ \mathfrak{m} \in \mathfrak{M}_n\}$.

B. Maximum-Likelihood Estimators

The log-likelihood functional γ_n is given 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. Let $\mathfrak{m} = (m, \underline{r}) \in \mathfrak{M}_n$, it can be shown that the maximum-likelihood estimator $\hat{s}_\mathfrak{m}$ on the exponential model $\mathcal{E}_\mathfrak{m}$ (the minimizer of $\gamma_n(t)$ when t varies in $\mathcal{E}_\mathfrak{m}$) exists and is unique if and only if there are at least $\left[\frac{r_I}{2}\right]$ observations in each interval I ($[k]$ denotes the largest integer not greater than k). Without going into further detail, it is easy to check that $\hat{s}_\mathfrak{m}$ is the unique element of $\mathcal{E}_\mathfrak{m}$ which satisfies $\int_I \hat{s}_\mathfrak{m}\, d\mu = P_n(I)$ and $\int_I \hat{s}_\mathfrak{m}\phi\, d\mu = P_n(\phi\mathbb{1}_x)$ for every polynomial ϕ with degree not greater than r_I on I and for every interval $I \in m$. And the condition of existence of such a density on each interval is due to Crain [6, Theorem 3.1 and Theorem 3.2]. We define the maximum-likelihood estimator $\hat{s}_\mathfrak{m}$ on $\mathcal{E}_\mathfrak{m}$ by

$$\hat{s}_\mathfrak{m} = \begin{cases} \operatorname*{argmin}_{t \in \mathcal{E}_\mathfrak{m}} \{\gamma_n(t)\}, & \text{if } nP_n(I) > \left[\frac{r_I}{2}\right]\ \forall I \in m \\ 1, & \text{otherwise.} \end{cases}$$

This results in the collection of estimators $\{\hat{s}_\mathfrak{m},\ \mathfrak{m} \in \mathfrak{M}_n\}$.

C. Squared Hellinger Risk and Kullback–Leibler Information

To compare the performances of estimators we shall use the squared Hellinger risk as a criterion of quality. We recall that the squared Hellinger loss between two densities p and q with respect to μ is defined by

$$h^2(p, q) = \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2\, d\mu.$$

An ideal model $\overline{\mathfrak{m}}$ among the family $\mathfrak{M}_n$ would minimize the risk $\boldsymbol{E}\left[h^2(s, \hat{s}_\mathfrak{m})\right]$ when $\mathfrak{m}$ varies in $\mathfrak{M}_n$. Unfortunately, such a model depends on s, and, consequently, $\hat{s}_{\overline{\mathfrak{m}}}$ cannot be used as an estimator of s. We aim at selecting some model $\hat{\mathfrak{m}} \in \mathfrak{M}_n$ from the data only in such a way that the resulting estimator $\hat{s}_{\hat{\mathfrak{m}}}$ on the model $\mathcal{E}_{\hat{\mathfrak{m}}}$ behaves almost as well as the ideal $\hat{s}_{\overline{\mathfrak{m}}}$. To do this, we define a data-driven criterion as follows. Given some penalty function $\mathrm{pen}_n\colon \mathfrak{M}_n \to \mathbb{R}_+$, the penalized maximum-likelihood criterion is defined by

$$\mathrm{crit}_n(\mathfrak{m}) = \gamma_n(\hat{s}_\mathfrak{m}) + \mathrm{pen}_n(\mathfrak{m})$$

and the minimization of this criterion over $\mathfrak{M}_n$ leads to an estimator $\tilde{s} = \hat{s}_{\hat{\mathfrak{m}}} \in \mathcal{E}_{\hat{\mathfrak{m}}}$, called penalized maximum-likelihood estimator.

In order to compare the squared Hellinger risk $\boldsymbol{E}\left[h^2(s, \tilde{s})\right]$ with $\inf_{\mathfrak{m} \in \mathfrak{M}_n} \boldsymbol{E}\left[h^2(s, \hat{s}_\mathfrak{m})\right]$, we need to evaluate the risk $\boldsymbol{E}\left[h^2(s, \hat{s}_\mathfrak{m})\right]$. We introduce then the Kullback–Leibler information to assess the qualities of approximation of $\mathcal{E}_\mathfrak{m}$. The Kullback–Leibler information between two densities p and q with respect to μ is given by

$$K(p, q) = \int p \log \frac{p}{q}\, d\mu.$$

It is well known that $K(p, q) \geq 2h^2(p, q)$ and that, when the sup-norm of $\log(p/q)$ is bounded by some constant M

$$K(p, q) \leq C(M)h^2(p, q)$$

(see Birgé [18, Lemma 4.4]). We can define the information projection of s onto $\mathcal{E}_\mathfrak{m}$ to be the minimizer $\overline{s}_\mathfrak{m}$ of $K(s, t)$ when t varies in $\mathcal{E}_\mathfrak{m}$. This projection always exists and we denote

$$K(s, \mathcal{E}_\mathfrak{m}) = \inf_{t \in \mathcal{E}_\mathfrak{m}} K(s, t) = K(s, \overline{s}_\mathfrak{m})$$

(see Crain [4, Theorem 3.1]). Moreover, the following decomposition is available:

$$K(s, \hat{s}_\mathfrak{m}) = K(s, \overline{s}_\mathfrak{m}) + K(\overline{s}_\mathfrak{m}, \hat{s}_\mathfrak{m}) \tag{1}$$

where $K(s, \overline{s}_\mathfrak{m})$ represents some approximation error and $K(\overline{s}_\mathfrak{m}, \hat{s}_\mathfrak{m})$ some estimation error within the model $\mathcal{E}_\mathfrak{m}$. The Kullback–Leibler loss has been studied by Barron and Sheu [13]. They proved that if $\lim_{n-+\infty} D_\mathfrak{m}/\sqrt{n} = 0$ then, with a probability tending to one as n tends to infinity, the maximum-likelihood estimator on $\mathcal{E}_\mathfrak{m}$ exists and satisfies

$$K(s, \hat{s}_\mathfrak{m}) = O_P\left(K(s, \mathcal{E}_\mathfrak{m}) + \frac{D_\mathfrak{m}}{n}\right).$$

Note that this result is valid for exponential families of trigonometric series, polynomials, and splines. It implies the same bound for the squared Hellinger loss of $\hat{s}_\mathfrak{m}$, since $K \geq 2h^2$. So, the squared Hellinger risk of $\hat{s}_\mathfrak{m}$ is bounded by some quantity of the order of $K(s, \mathcal{E}_\mathfrak{m}) + D_\mathfrak{m}/n$, at least asymptotically. We would have preferred to get an upper bound and a lower bound using the squared Hellinger distance to evaluate the approximation error (instead of the Kullback–Leibler information). Unfortunately, to our knowledge, such bounds have not been proven. Hence, we shall content ourselves to bound (up to a multiplicative constant) the squared Hellinger risk of $\tilde{s}$ by the infimum of $K(s, \mathcal{E}_\mathfrak{m}) + D_\mathfrak{m}/n$ over $\mathfrak{M}_n$, as was done in Barron *et al.* [16] and in Yang and Barron [17].

D. The Main Theorem

The choice of the penalty function and the corresponding risk bound will strongly depend on the complexity of the family of models. This is essentially the reason why we have to make one of the following assumptions about the family of partitions.

($\mathbf{H_1}$): There exist some integer a and some positive constant Γ' such that the number of partitions of the family $\mathcal{M}_n$ with the same number of pieces D is bounded by $\Gamma' D^a$. A particular example of such family is the one of regular partitions, denoted by $\mathcal{M}_n^r$. A partition m is regular if all its elements have the same measure with respect to μ. Such a family contains only one partition with a given number of pieces and thus satisfies ($\mathbf{H_1}$) (with $a = 0$ and $\Gamma' = 1$).

($\mathbf{H_2}$): The family of partitions is composed of irregular partitions with endpoints belonging to the grid $\left\{kN_n^{-1},\ 0 \leq k \leq N_n\right\}$ for some fixed integer N_n. This family is denoted by $\mathcal{M}_n^{ir}$. Note that $\mathcal{M}_n^{ir}$ contains $\binom{N_n - 1}{D-1}$ partitions with D pieces.

In both cases, we denote

$$\Gamma_n = \inf_{m \in \mathcal{M}_n} \inf_{I \in m} \{\mu(I)\}. \tag{2}$$

For these two types of families of exponential models of piecewise polynomials, the following theorem (to be proved in Section III) holds.

Theorem 2.1: Assume that the family of partitions $\mathcal{M}_n$ satisfies either $(\mathbf{H_1})$ or $(\mathbf{H_2})$. Let $(\rho_n)_{n\in\mathbb{N}^*}$ be some sequence of positive numbers such that

$$\lim_{n\to+\infty}\rho_n = 0 \quad \text{and} \quad \lim_{n\to+\infty} n\Gamma_n\rho_n^2/\log n = +\infty$$

where Γ_n is given by (2). Let $(L_{\mathfrak{m}})_{\mathfrak{m}\in\mathfrak{M}_n}$ be some family of nonnegative weights such that there exists some positive constant Σ independent on n satisfying

$$\sum_{\mathfrak{m}\in\mathfrak{M}_n}\exp\left(-L_{\mathfrak{m}}D_{\mathfrak{m}}\right)\le\Sigma. \tag{3}$$

Choose some positive number $c > 1/2$ and a penalty function such that

$$\mathrm{pen}_n(\mathfrak{m}) \ge c\left(1+\sqrt{2\left(1+1/c\right)L_{\mathfrak{m}}}\right)^2\frac{D_{\mathfrak{m}}}{n}$$

for all $\mathfrak{m}\in\mathfrak{M}_n$. Assume that the density s is positive with

$$\log s \in \mathbb{L}_\infty\left([0,1],\mu\right)$$

and for every $\mathfrak{m}\in\mathfrak{M}_n$ denote $M_{\mathfrak{m}}^{\infty} = \left\|\log\left(s/\overline{s}_{\mathfrak{m}}\right)\right\|_\infty$ where $\overline{s}_{\mathfrak{m}}$ is the information projection of s onto $\mathcal{E}_{\mathfrak{m}}$. Then, there exist some constants $C_1(c)$, $C_2(M_{\mathfrak{m}}^{\infty},c)$, and $C_3(s,c)$ such that the penalized maximum-likelihood estimator $\tilde{s}$ satisfies

$$\boldsymbol{E}\left[h^2\left(s,\tilde{s}\right)\mathbf{1}_{\{\tilde{s}\ge\rho_n\}}\right] \le C_1(c)\inf_{\mathfrak{m}\in\mathfrak{M}_n}\left\{2K\left(s,\mathcal{E}_{\mathfrak{m}}\right)+\mathrm{pen}_n(\mathfrak{m}) + C_2\left(M_{\mathfrak{m}}^{\infty},c\right)\frac{2\Sigma+1}{n}\right\}+\frac{C_3(s,c)}{n^2}$$

where

$$C_2\left(M_{\mathfrak{m}}^{\infty},c\right) = \left(\left(2\Phi\left(-M_{\mathfrak{m}}^{\infty}\right)\right)^{-1}+\frac{M_{\mathfrak{m}}^{\infty}}{3}+C_2'(c)\right)$$

with $\Phi(x) = \frac{e^x-1-x}{x^2}$ for all $x\in\mathbb{R}$, and $\{\tilde{s}\ge\rho_n\}$ denotes the event $\{\tilde{s}(x)\ge\rho_n,\ \forall\, x\in[0,1]\}$.

Remark 1: If the sequences $(\Gamma_n)_{n\in\mathbb{N}^*}$ and $(\rho_n)_{n\in\mathbb{N}^*}$ satisfy

$$\Gamma_n \ge \Gamma\frac{(\log n)^4}{n} \quad \text{and} \quad \rho_n \ge \frac{\Theta}{\log n}$$

for some positive constants Γ and Θ, they also satisfy the assumptions of the theorem.

Remark 2: Our penalized criterion generalizes the one that we have used in the case of histograms density estimation (see Castellan [19, Theorem 3.2]). However, the risk bound in the general case essentially differs from the particular case of histograms by the presence of $\mathbf{1}_{\{\tilde{s}\ge\rho_n\}}$ in the computation of the risk. This restriction is due to the bad behavior in $\mathbb{L}_\infty$ norm of the maximum-likelihood estimator on exponential families. But we can hope that the penalized maximum-likelihood estimator behaves better in practice. Nevertheless, this restriction can be avoided asymptotically in the case of regular partitions if $\log s$ is assumed to belong to some Besov space (see Proposition 2.1).

Remark 3: If the quantities $M_{\mathfrak{m}}^{\infty}$ are uniformly bounded with respect to $\mathfrak{m}\in\mathfrak{M}_n$, then the estimator $\tilde{s}$ realizes the best tradeoff between the "bias" term $K(s,\mathcal{E}_{\mathfrak{m}})$ (error of approximation) and the "variance" term (error of estimation) represented by $\mathrm{pen}_n(\mathfrak{m})$ (since this term can be proportional to $D_{\mathfrak{m}}/n$), the term $C_3(s,c)/n^2$ appearing as a remainder.

Remark 4: The quantity $C_1(c)$ involved in the risk bound of Theorem 2.1 tends to infinity when c goes to $1/2$. Thus, referring to the case of histograms where $\mathcal{N}=\{0\}$ (see Castellan [19]), this restriction is necessary, and it seems that we should recommend to take c close to 1 in order to get the best risk bound stable with respect to n.

Remark 5: The particular form of the penalty function derives from a Talagrand-type inequality. We actually use the version of Bousquet [20]. This type of inequality allows to control the supremum of empirical processes and is a fundamental tool for the analysis of penalized minimum contrast estimators established by Birgé and Massart [21].

Remark 6: We can compare our results with the results of Barron *et al.* [16] and Yang and Barron [17] who study penalized maximum-likelihood criteria in other frameworks. Similarly, the penalty function depends on the complexity of the family of models. In both references, the squared Hellinger risk is used to study the performances of the penalized maximum-likelihood estimator and the "bias" term is evaluated in terms of Kullback–Leibler information. However, the result of Barron *et al.* [16, Theorem 2, p. 327] cannot be applied in our context since the models introduced by Barron *et al.* [16] are sets of positive functions t belonging to finite-dimensional linear subspaces of $\mathbb{L}_2(\mu)$ such that t^2 is a density. Yang and Barron [17] have studied more specifically families of exponential models. Nevertheless, their results need a metric dimension assumption which forces in the applications to put a $\mathbb{L}_\infty$ upper bound on the log densities of each model. This leads to a penalty function depending on these bounds and on a universal constant which is unrealistic. In comparison, the advantage of our method (valid in a more specific framework) is to provide a penalized maximum-likelihood criterion which is directly applicable. The main drawback is the restriction $\mathbb{1}_{\{\tilde{s}\ge\rho_n\}}$ in the computation of the risk.

E. Applications

In this subsection, we develop possible choices for the penalty function and deduce the performance of the associated penalized maximum-likelihood estimator. These choices depend on the family of weights $(L_{\mathfrak{m}})_{\mathfrak{m}\in\mathfrak{M}_n}$ satisfying (3), and we have to consider separately the cases of Assumptions $(\mathbf{H_1})$ and $(\mathbf{H_2})$. In the sequel, we do not give the details of the elementary computations of the series in (3) (see Castellan [22] for more details).

1) Regular Case:

a) Choice of the penalty function: We assume here that the family of models satisfies Assumption $(\mathbf{H_1})$. We can choose then constant weights and take $L_{\mathfrak{m}} = L+\log\left(e(2r+1)\right)$ for any positive constant L, where $r=\max\mathcal{N}$.

If the degree is fixed for all the intervals of a given partition, that is, $\mathfrak{m}=(m,\underline{r})$ belongs to $\mathfrak{M}_n$ if $r_I = k$ for every $I\in m$ for some integer k which varies in $\mathcal{N}$, then positive constant weights work. Indeed, if $L_{\mathfrak{m}} = L$ for every $\mathfrak{m}\in\mathfrak{M}_n$ for some positive constant L then (3) is satisfied and all penalty function such that

$$\mathrm{pen}_n(\mathfrak{m}) = c\frac{D_{\mathfrak{m}}}{n}$$

for any constant $c > 1/2$ is allowed. In particular, the choice $\mathrm{pen}_n(\mathfrak{m}) = D_{\mathfrak{m}}/n$ corresponds to Akaike's criterion.

We can also take weights $L_{\mathfrak{m}}$ that depend on $|m|$ for $\mathfrak{m}=(m,\underline{r})$. For instance, $L_{\mathfrak{m}}=L(D)=D^{-1/2}$ for $|m|=D$ works and any penalty function satisfying $\mathrm{pen}_n(\mathfrak{m})\ge cD_{\mathfrak{m}}(1+c'D^{-1/4})^2/n$ can be taken (for any $c>1/2$ and a suitable constant c'). In particular, we could recommend to add a correction to Akaike's criterion since it improves on the risk bound nonasymptotically for the histograms models (see Castellan [19] and the simulation study made by Birgé and Rosenholc [23]). But there is no proof of this in the general case of exponential models of piecewise polynomials.

In both cases, the criterion we propose is the analog of Akaike's criterion, modified when we take variable weights, but equivalent at least asymptotically.

b) Rate of convergence of the penalized maximum-likelihood estimator: Let us now study the performances of the penalized maximum-likelihood estimator. We consider the family of models $\mathfrak{M}_n$ based on the family $\mathcal{M}_n^r$ of regular partitions and we assume that the log-density $\log s$ belongs to a Besov space. We refer to DeVore and Lorentz

[24, pp. 44 and 55] for the definition of Besov spaces. The proof of the following proposition will be given in Section III.

Proposition 2.1: Assume that $\log s$ belongs to some Besov space $\mathcal{B}_{\alpha,l,\infty}([0,1])$ for $l \geq 2$ and $\alpha > 1/l$. Assume that α is unknown in $(0, r+1)$ for some known integer r. Let $\{\mathcal{E}_{\mathfrak{m}}, \mathfrak{m} \in \mathfrak{M}_n\}$ be the collection of exponential models corresponding to the family of regular partitions $\mathcal{M}_n^r$ and to $\mathcal{N} = \{0, \ldots, r\}$. If $\tilde{s}$ is the penalized maximum-likelihood estimator associated with this collection with a penalty function $\mathrm{pen}_n(\mathfrak{m}) \geq cD_{\mathfrak{m}}/n$ for some constant $c > 1/2$ then we have

$$\limsup_{n-\infty} n^{\frac{2\alpha}{2\alpha+1}} \boldsymbol{E}\left[h^2(s, \tilde{s})\right] < +\infty.$$

Remark 1: Note that the rate of convergence $n^{-\frac{2\alpha}{2\alpha+1}}$ is known to be optimal for this class of functions, see Koo and Kim [14].

Remark 2: We can use the same argument as Yang and Barron [17, Sec. III-B] to estimate not strictly positive densities. Indeed, it is possible to remove the assumption that s is positive replacing s by $(s+1)/2$. But the price to pay is to obtain optimal rates of convergence for the $\mathbb{L}_1$ risk. For a nonasymptotic rate of convergence with respect to the $\mathbb{L}_1$ loss, we can refer to Devroye and Lugosi [25]: they use a data-based criterion to select the smoothing factor for Kernel density estimation.

2) Irregular Case: We consider the particular example of the family of partitions $\mathcal{M}_n^{ir}$, case $(\mathbf{H_2})$. If we take constant weights, it is necessary to choose them of the order of $\log n$ which leads to a penalty function satisfying

$$\mathrm{pen}_n(\mathfrak{m}) \geq \log n \frac{D_{\mathfrak{m}}}{n}.$$

The reason for this extra $\log n$ factor comes from the fact that the family $\mathfrak{M}_n$ contains many models of the same dimension. This penalty function leads to lose a $\log n$ factor in the risk bound. However, in the particular case $\mathcal{N} = \{0\}$, it can be shown that this $\log n$ factor in the risk bound is necessary (see Birgé and Massart [26, Proposition 2]).

We can also choose variable weights. Indeed, the choice

$$L_{\mathfrak{m}} = L_D = \log\left(\frac{c(N_n - 1)}{D-1}\right)$$

for a suitable value of c when $\mathfrak{m} = (m, \underline{r})$ with $|m| = D$ is convenient for the convergence of the series (3). Thus, we can choose a penalty function equal to

$$\mathrm{pen}_n(\mathfrak{m}) = c\frac{D_{\mathfrak{m}}}{n}\log\frac{c' N_n}{D},$$

for every model $\mathfrak{m}$ such that $|m| = D$, with suitable constants c and c'. Choosing variable weights produces better bounds. In particular, it gives the right rate of convergence when $\mathcal{N} = \{0\}$ and s is assumed to be some piecewise-constant density with D pieces (see Birgé and Massart [26, Proposition 2]).

III. Proof of the Main Results

A. Proof of Theorem 2.1

The proof is similar to the one of the histograms case (see Castellan [19, proof of Theorem 3.2]). The main argument of both proofs is a uniform control of the ratio $\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}}$ over $\mathfrak{m} \in \mathfrak{M}_n$. In contrast to the histograms case, the estimators $\hat{s}_{\mathfrak{m}}$ and also the information projections $\overline{s}_{\mathfrak{m}}$ on an exponential model $\mathcal{E}_{\mathfrak{m}}$ are defined implicitly which makes this control very tricky and technical. Another troublesome point is the link between the variance of the empirical contrast (the log likelihood) and the Kullback–Leibler information which is more difficult to analyze for a piecewise-polynomial than for a piecewise-constant log density. We only give a sketch of proof in order to avoid too many technical details (see Castellan [22, Ch. 2, Sec. 2.4] for a complete proof).

1) Definition of the Estimators and of the Information Projection: Let $\mathfrak{m} = (m, \underline{r}) \in \mathfrak{M}_n$. In this part, we explain how the estimators $\hat{s}_{\mathfrak{m}}$ and the information projection $\overline{s}_{\mathfrak{m}}$ are defined in order to control their ratio. We consider what we shall call in the sequel the "Legendre basis" of $\mathcal{P}_{\mathfrak{m}}$ which can be constructed as follows. Let $\{Q_k, k \in \mathbb{N}\}$ be the sequence of Legendre polynomials, which is known to be orthonormal in $\mathbb{L}_2([-1,1])$ (see Whittaker and Watson [27, pp. 302–305] for details), then we define the "Legendre basis" $\{\psi_{I,k}, 0 \leq k \leq r_I, I \in m\}$ of $\mathcal{P}_{\mathfrak{m}}$ as follows:

$$\psi_{I,k}(x) = \sqrt{\frac{2k+1}{\mu(I)}}\, Q_k\left(2\frac{x - a_I}{\mu(I)} - 1\right)$$

where a_I represents the left endpoint of I. In the sequel, we shall also use the following notations.

- Given some measurable space $(\mathbb{X}, \mu)$, and a family of functions $\{\mathbb{1}, \varphi_\lambda, \lambda \in \Lambda\}$ in $\mathbb{L}_2(\mathbb{X}, \mu) \cap \mathbb{L}_\infty(\mathbb{X}, \mu)$ such that φ_λ is orthogonal to $\mathbb{1}$ for all $\lambda \in \Lambda$, we denote by $t(\beta)$ the density function defined by

  $$t(\beta) = \exp\left(\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda - \Psi(\beta)\right)$$

  where

  $$\Psi(\beta) = \log\int_{\mathbb{X}}\exp\left(\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\right)d\mu$$

  for all $\beta \in \mathbb{R}^\Lambda$. The gradient G of Ψ is

  $$G(\beta) = \left(\int_{\mathbb{X}}\varphi_\lambda t(\beta)\,d\mu\right)_{\lambda\in\Lambda}.$$

- For every $I \in m$ such that $r_I \geq 1$, the functions corresponding to the orthogonal family $\{\mathbb{1}_I, \psi_{I,k}, 1 \leq k \leq r_I\}$ in $\mathbb{L}_2(I, \mu)$ are denoted by t_I, Ψ_I, and G_I.
- For every interval $I \in m$, the space of the polynomials with degree r_I defined on I is denoted by $\mathcal{P}_I$, and $\mathcal{E}_I$ denotes the corresponding exponential model.

We will also use the following property of $\mathcal{P}_I$: if $\|.\|_{\infty,I}$ denotes the sup-norm in $\mathbb{L}_\infty(I, \mu)$ and $\|.\|_2$ the $\mathbb{L}_2$-norm in $\mathbb{L}_2(I, \mu)$, then

$$\|f\|_{\infty,I} \leq (r_I + 1)\,\mu(I)^{-\frac{1}{2}}\|f\|_2, \qquad \text{for all } f \in \mathcal{P}_I. \tag{4}$$

We are now in a position to give the equations that define the estimator and the information projection. These two densities are solutions of similar optimization problems. These problems can be solved intervals by intervals. For every $I \in m$ such that $r_I \geq 1$, let us define

$$\hat{\delta}_I = (\hat{\delta}_{I,1}, \ldots, \hat{\delta}_{I,r_I}) \in \mathbb{R}^{r_I}$$

with $\hat{\delta}_{I,k} = P_n(\psi_{I,k})/P_n(I)$ and

$$\overline{\delta}_I = \left(\overline{\delta}_{I,1}, \ldots, \overline{\delta}_{I,r_I}\right) \in \mathbb{R}^{r_I}$$

with $\overline{\delta}_{I,k} = P(\psi_{I,k})/P(I)$. The vectors

$$\hat{\beta}_I = (\hat{\beta}_{I,1}, \ldots, \hat{\beta}_{I,r_I}) \in \mathbb{R}^{r_I} \quad \text{and} \quad \overline{\beta}_I = \left(\overline{\beta}_{I,1}, \ldots, \overline{\beta}_{I,r_I}\right) \in \mathbb{R}^{r_I}$$

are, respectively, the unique solutions of

$$G_I(\hat{\beta}_I) = \hat{\delta}_I \quad \text{and} \quad G_I\left(\overline{\beta}_I\right) = \overline{\delta}_I.$$

Then, for every $I \in m$, $\hat{s}_I$ and $\overline{s}_I$ are the functions of $\mathcal{E}_I$ given by

$$\hat{s}_I = \begin{cases} \mu(I)^{-1} \mathbb{1}_I, & \text{if } r_I = 0 \\ t_I(\hat{\beta}_I), & \text{if } r_I \geq 1 \end{cases}$$

and

$$\overline{s}_I = \begin{cases} \mu(I)^{-1} \mathbb{1}_I, & \text{if } r_I = 0 \\ t_I\left(\overline{\beta}_I\right), & \text{if } r_I \geq 1. \end{cases}$$

Finally, we define the estimator and the information projection by

$$\hat{s}_{\mathfrak{m}} = \sum_{I \in m} P_n(I)\, \hat{s}_I \mathbb{1}_I, \quad \text{and} \quad \overline{s}_{\mathfrak{m}} = \sum_{I \in m} P(I)\, \overline{s}_I \mathbb{1}_I.$$

2) Uniform Control of the Ratio $\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}}$: In the sequel, we set

$$\rho = \operatorname*{essinf}_{x \in [0,1]} \{s(x)\}$$

and denote by ν_n the centered empirical process $\nu_n = P_n - P$. We construct a set Ω_n (depending on the constant c of the penalty function) of high probability such that the ratio $\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}}$ is uniformly bounded (independently of s and of the dimension $D_{\mathfrak{m}}$) on this set. To this end, we apply Lemma 1.1. Indeed, a suitable control of $|\nu_n(\psi_{I,k})|$ for all $I \in m$ and for all $k \in \{0, \ldots, r_I\}$ allows to bound $|\hat{\delta}_I - \overline{\delta}_I|_2$ (where $|.|_2$ denotes the Euclidean norm) so as to satisfy assumption (15), and to deduce, by Lemma 1.1 with $\delta^0 = \hat{\delta}_I$ and $\delta = \overline{\delta}_I$ for all $I \in m$, a control of $\|\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}}\|_\infty$ provided that $\hat{s}_{\mathfrak{m}} \geq \rho_n$. This restriction here comes from the difficulties to invert the equations defining $\hat{s}_{\mathfrak{m}}$ and $\overline{s}_{\mathfrak{m}}$, and is the price to pay to obtain a bound which is independent of s. More precisely, the set Ω_n is the event satisfying

$$|\nu_n(\psi_{I,k})| \leq C(r,c)(\rho_n \wedge \rho)\sqrt{\mu(I)}, \quad \forall k \in \{0, \ldots, r\}, \ \forall I \in m \quad (5)$$

for all $m \in \mathcal{M}_n$, where $C(r,c)$ can be computed and depends on c and on $r = \max \mathcal{N}$. Over this set the ratio $\|\log(\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}})\|_\infty$ is bounded by some positive function of c, $\kappa(c) < 2$, for every model $\mathfrak{m} \in \mathfrak{M}_n$ such that $\hat{s}_{\mathfrak{m}} \geq \rho_n$. Now, we will work on this set. This is not too restrictive since we can derive that

$$\boldsymbol{P}(\Omega_n^c) \leq \frac{C_3(s,c)}{n^2}$$

from the assumption $\lim_{n \to +\infty} n\Gamma_n \rho_n^2 / \log n = +\infty$ and Bernstein's inequality (see Pollard [28, p. 193]).

3) Sketch of Proof: Let $\mathfrak{m} \in \mathfrak{M}_n$. It follows from the definition of $\tilde{s}$ that

$$K(s, \tilde{s}) \leq K(s, \overline{s}_{\mathfrak{m}}) + \nu_n(\log \tilde{s} - \log \overline{s}_{\mathfrak{m}}) + \operatorname{pen}_n(\mathfrak{m}) - \operatorname{pen}_n(\hat{\mathfrak{m}}). \quad (6)$$

We control the term $\nu_n(\log \tilde{s} - \log \overline{s}_{\mathfrak{m}})$ in (6) by setting

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

Let $\mathfrak{m}' = (m', \underline{r}) \in \mathfrak{M}_n$. We control $\nu_n(\log(\hat{s}_{\mathfrak{m}'}/\overline{s}_{\mathfrak{m}'}))$ and $\nu_n(\log(\overline{s}_{\mathfrak{m}'}/s))$ uniformly over all the models $\mathfrak{m}'$, in order to apply these controls to the random model $\hat{\mathfrak{m}}$. 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 $\log \frac{p}{q} \in \mathbb{L}_2(s d\mu)$. Note that $V_s^2(p,q)$ represents the "variance" of the contrast γ_n since

$$V_s^2(p,q) = n\boldsymbol{E}\left[(\gamma_n(p) - \gamma_n(q))^2\right].$$

c) Control of $\nu_n(\log \hat{s}_{\mathfrak{m}'} - \log \overline{s}_{\mathfrak{m}'})$: We write

$$\left|\nu_n\left(\log \frac{\hat{s}_{\mathfrak{m}'}}{\overline{s}_{\mathfrak{m}'}}\right)\right| \leq \sup_{t \in \mathcal{E}_{\mathfrak{m}'}} \left|\nu_n\left(\frac{\log(t/\overline{s}_{\mathfrak{m}'})}{V_s(t, \overline{s}_{\mathfrak{m}'})}\right)\right| V_s(\hat{s}_{\mathfrak{m}'}, \overline{s}_{\mathfrak{m}'}).$$

Setting $Z_{\mathfrak{m}'} = \sup\left\{|\nu_n(f)|, f \in \mathcal{P}_{\mathfrak{m}'}, \int f^2 s\, d\mu = 1\right\}$, we get

$$\left|\nu_n\left(\log \frac{\hat{s}_{\mathfrak{m}'}}{\overline{s}_{\mathfrak{m}'}}\right)\right| \leq Z_{\mathfrak{m}'} V_s(\hat{s}_{\mathfrak{m}'}, \overline{s}_{\mathfrak{m}'}).$$

We first apply Lemma 3.1 with $x = x_{\mathfrak{m}'} = L_{\mathfrak{m}'} D_{\mathfrak{m}'} + \zeta$ to deduce that $Z_{\mathfrak{m}'} \mathbb{1}_{\Omega_n}$ is upper-bounded by

$$C(c)\left(\sqrt{\frac{D_{\mathfrak{m}'}}{n}} + \sqrt{\frac{2L_{\mathfrak{m}'} D_{\mathfrak{m}'}}{n}}\right) + \sqrt{\frac{\zeta}{n}} \quad (7)$$

with a probability larger than $1 - e^{-L_{\mathfrak{m}'} D_{\mathfrak{m}'}} e^{-\zeta}$, where ζ is a positive number to be integrated at the end of the proof and $C(c) > 1$. Thus, $Z_{\mathfrak{m}'} \mathbb{1}_{\Omega_n}$ is upper-bounded by (7) uniformly with respect to $\mathfrak{m}' \in \mathfrak{M}_n$ with a probability larger than $1 - \Sigma e^{-\zeta}$. Now, we will use the following elementary inequality which holds for any positive number θ, and any numbers a and b,

$$2ab \leq \theta a^2 + \theta^{-1} b^2. \quad (8)$$

Hence, we derive that, for all $\mathfrak{m}' \in \mathfrak{M}_n$

$$\left|\nu_n\left(\log \frac{\hat{s}_{\mathfrak{m}'}}{\overline{s}_{\mathfrak{m}'}}\right)\right| \mathbb{1}_{\Omega_n} \leq C\left(1 + \sqrt{2L_{\mathfrak{m}'}}\right)^2 \frac{D_{\mathfrak{m}'}}{n} + C_2 V_s^2(\overline{s}_{\mathfrak{m}'}, \hat{s}_{\mathfrak{m}'}) + C_3 \frac{\zeta}{n} \quad (9)$$

on a set Ω_n^1 of probability larger than $1 - \Sigma e^{-\zeta}$. The values of C, C_2, C_3 depend on c. We can choose any value of $C > 1/2$ (using (8) with an adequate coefficient) and then $C_2 < 1/2$. Hence we can take $C = c$.

d) Control of $\nu_n(\log \overline{s}_{\mathfrak{m}'} - \log s)$: We use the exponential bound given by Castellan [19, Proposition 4.5] to derive that, for all $\mathfrak{m}' \in \mathfrak{M}_n$

$$\nu_n\left(\log \frac{\overline{s}_{\mathfrak{m}'}}{s}\right) \leq K(s, \overline{s}_{\mathfrak{m}'}) - 2h^2(s, \overline{s}_{\mathfrak{m}'}) + 2\frac{L_{\mathfrak{m}'} D_{\mathfrak{m}'}}{n} + 2\frac{\zeta}{n} \quad (10)$$

on a set Ω_n^2 of probability larger than $1 - \Sigma e^{-\zeta}$.

e) Control of $\nu_n(\log s - \log \overline{s}_{\mathfrak{m}})$: Using Bernstein's inequality, we get

$$\left|\nu_n\left(\log \frac{s}{\overline{s}_{\mathfrak{m}}}\right)\right| \leq V_s(s, \overline{s}_{\mathfrak{m}}) \sqrt{\frac{2\zeta}{n}} + \frac{M_{\mathfrak{m}}^\infty \zeta}{3n}$$

on a set Ω_n^3 of probability larger than $1 - e^{-\zeta}$. Now, we apply the results of Barron and Sheu [13, Inequality 3.1 of Lemma 1] to the densities s and $\overline{s}_{\mathfrak{m}}$. We get

$$K(s, \overline{s}_{\mathfrak{m}}) \geq \Phi(-M_{\mathfrak{m}}^\infty) V_s^2(s, \overline{s}_{\mathfrak{m}}).$$

And we derive that, over Ω_n^3 (using (8) with $\theta = \Phi(-M_{\mathfrak{m}}^\infty)$)

$$\left|\nu_n\left(\log \frac{s}{\overline{s}_{\mathfrak{m}}}\right)\right| \leq K(s, \overline{s}_{\mathfrak{m}}) + \left((2\Phi(-M_{\mathfrak{m}}^\infty))^{-1} + \frac{M_{\mathfrak{m}}^\infty}{3}\right)\frac{\zeta}{n}. \quad (11)$$

4) Conclusion: Finally, combining (6) with (9), (10) applied to $\mathfrak{m}' = \hat{\mathfrak{m}}$ and (11), we derive that, on a set $\Omega_n' = \Omega_n^1 \cap \Omega_n^2 \cap \Omega_n^3$ of probability larger than $1 - (2\Sigma + 1)\, e^{-\zeta}$

$$\begin{aligned} K(s, \hat{s}_{\hat{\mathfrak{m}}}) \leq\, & 2K(s, \overline{s}_{\mathfrak{m}}) + \mathrm{pen}_n(\mathfrak{m}) - \mathrm{pen}_n(\hat{\mathfrak{m}}) \\ & + \left[c\left(1 + \sqrt{2L_{\hat{\mathfrak{m}}}}\right)^2 + 2L_{\hat{\mathfrak{m}}} \right] \frac{D_{\hat{\mathfrak{m}}}}{n} \\ & + \left[K(s, \overline{s}_{\hat{\mathfrak{m}}}) - 2h^2(s, \overline{s}_{\hat{\mathfrak{m}}}) \right] \\ & + C_2 V_s^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) + C_3(M_{\mathfrak{m}}^{\infty}) \frac{\zeta}{n} \end{aligned}$$

where

$$C_3(M_{\mathfrak{m}}^{\infty}) = (2\Phi(-M_{\mathfrak{m}}^{\infty}))^{-1} + \frac{M_{\mathfrak{m}}^{\infty}}{3} + C_3$$

for some value of C_3 depending on c. Now, in the event $\{\hat{s}_{\hat{\mathfrak{m}}} \geq \rho_n\} \cap \Omega_n$, we have $\|\log(\hat{s}_{\hat{\mathfrak{m}}}/\overline{s}_{\hat{\mathfrak{m}}})\|_{\infty} \leq \kappa(c) < 2$ (from the construction of the set Ω_n). Thus, Lemma 2.1 in Appendix II allows to link the Kullback–Leibler distance $K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}})$ and the variance term $V_s^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}})$ up to an additive term $h^2(s, \overline{s}_{\hat{\mathfrak{m}}})$. This additive term is not present in the histograms case. Indeed, in this particular case, Lemma 1 of Barron and Sheu [13, Inequality 3.1] is sufficient to provide a link between $K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}})$ and

$$V_s^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) = \int \overline{s}_{\hat{\mathfrak{m}}} \left(\log(\overline{s}_{\hat{\mathfrak{m}}}/\hat{s}_{\hat{\mathfrak{m}}})\right)^2 d\mu.$$

But in the general case, this equality is false since $(\log(\overline{s}_{\hat{\mathfrak{m}}}/\hat{s}_{\hat{\mathfrak{m}}}))^2$ does not belong to $\mathcal{P}_{\mathfrak{m}}$, thus, we have to use Lemma 2.1. Hence,

$$V_s^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) \leq C_2^1 K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) + C_2^2 h^2(s, \overline{s}_{\hat{\mathfrak{m}}})$$

with some constants C_2^1 and C_2^2 depending on c such that $C_2^1 > 2$. Therefore, on the set $\{\hat{s}_{\hat{\mathfrak{m}}} \geq \rho_n\} \cap \Omega_n \cap \Omega_n'$ we have

$$\begin{aligned} K(s, \hat{s}_{\hat{\mathfrak{m}}}) \leq\, & 2K(s, \overline{s}_{\mathfrak{m}}) + \mathrm{pen}_n(\mathfrak{m}) + C_3(M_{\mathfrak{m}}^{\infty}) \frac{\zeta}{n} \\ & + \left[c\left(1 + \sqrt{2L_{\hat{\mathfrak{m}}}}\right)^2 + 2L_{\hat{\mathfrak{m}}} \right] \frac{D_{\hat{\mathfrak{m}}}}{n} - \mathrm{pen}_n(\hat{\mathfrak{m}}) \\ & + K(s, \overline{s}_{\hat{\mathfrak{m}}}) - \tilde{C}_2 h^2(s, \overline{s}_{\hat{\mathfrak{m}}}) + C_2' K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}). \end{aligned}$$

Then, we can take $C_2' < 1$. Our choice of penalty function implies that

$$\left[c\left(1 + \sqrt{2L_{\hat{\mathfrak{m}}}}\right)^2 + 2L_{\hat{\mathfrak{m}}} \right] \frac{D_{\hat{\mathfrak{m}}}}{n} - \mathrm{pen}_n(\hat{\mathfrak{m}}) \leq 0.$$

Then using the Pythagorean type identity (1), we get on $\{\hat{s}_{\hat{\mathfrak{m}}} \geq \rho_n\} \cap \Omega_n \cap \Omega_n'$

$$\begin{aligned} & \tilde{C}_2 h^2(s, \overline{s}_{\hat{\mathfrak{m}}}) + \left(1 - C_2'\right) K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) \\ & \qquad \leq 2K(s, \overline{s}_{\mathfrak{m}}) + \mathrm{pen}_n(\mathfrak{m}) + C_3(M_{\mathfrak{m}}^{\infty}) \frac{\zeta}{n}. \end{aligned}$$

Now, we use

$$K(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}}) \geq 2h^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}})$$

and

$$h^2(s, \hat{s}_{\hat{\mathfrak{m}}}) \leq 2h^2(s, \overline{s}_{\hat{\mathfrak{m}}}) + 2h^2(\overline{s}_{\hat{\mathfrak{m}}}, \hat{s}_{\hat{\mathfrak{m}}})$$

to derive that on Ω_n'

$$\tilde{C} h^2(s, \hat{s}_{\hat{\mathfrak{m}}}) \mathbb{1}_{\{\hat{s}_{\hat{\mathfrak{m}}} \geq \rho_n\} \cap \Omega_n} \leq 2K(s, \overline{s}_{\mathfrak{m}}) + \mathrm{pen}_n(\mathfrak{m}) + C_3(M_{\mathfrak{m}}^{\infty}) \frac{\zeta}{n}$$

and $\boldsymbol{P}(\Omega_n') \geq 1 - (2\Sigma + 1) e^{-\zeta}$ (for any positive number ζ). Integration with respect to ζ gives

$$\begin{aligned} & \boldsymbol{E}\left[h^2(s, \hat{s}_{\hat{\mathfrak{m}}}) \mathbb{1}_{\{\hat{s}_{\hat{\mathfrak{m}}} \geq \rho_n\} \cap \Omega_n}\right] \\ & \qquad \leq \frac{1}{\tilde{C}} \left\{ 2K(s, \overline{s}_{\mathfrak{m}}) + \mathrm{pen}_n(\mathfrak{m}) + C_3(M_{\mathfrak{m}}^{\infty}) \frac{2\Sigma + 1}{n} \right\}. \end{aligned}$$

We conclude the proof by the control of the probability of Ω_n.

B. Proof of Proposition 2.1

We first demonstrate that the quantities $M_{\mathfrak{m}}^{\infty} = \|\log(s/\overline{s}_{\mathfrak{m}})\|_{\infty}$ are uniformly bounded over $\mathfrak{M}_n$. Second, we can control the probability of the set $\{\tilde{s} \geq \rho_n\}^c$ which implies that the restriction $\mathbb{1}_{\{\tilde{s} \geq \rho_n\}}$ in the computation of the risk can be avoided. Thus, we can evaluate the rate of convergence of the squared Hellinger risk of the penalized maximum-likelihood estimator. Let us first remark that if $f = \log s$, since $f \in \mathcal{B}_{\alpha, l, \infty}([0, 1])$ with $\alpha > 1/l$, then f is bounded and Theorem 2.1 can be applied.

1) Uniform Control of $\|\log(s/\overline{s}_{\mathfrak{m}})\|_{\infty}$*:* We cannot directly prove such a uniform control and we have to introduce an auxiliary function $t_{\mathfrak{m}} \in \mathcal{E}_{\mathfrak{m}}$ such that $\|\log(s/t_{\mathfrak{m}})\|_{\infty}$ is uniformly bounded. Let us denote

$$f_{\mathfrak{m}} = \Pi_{S_{\mathfrak{m}}} f, \qquad C(f_{\mathfrak{m}}) = \log\left(\int e^{f_{\mathfrak{m}}}\, d\mu\right)$$

and

$$t_{\mathfrak{m}} = \exp(f_{\mathfrak{m}} - C(f_{\mathfrak{m}}))$$

where $S_{\mathfrak{m}}$ denotes the orthogonal complement of $\mathbb{1}$ in $\mathcal{P}_{\mathfrak{m}}$ and $\Pi_{S_{\mathfrak{m}}}$ the orthogonal projection onto $S_{\mathfrak{m}}$ in $\mathbb{L}_2(\mu)$. First, it is easy to check that $t_{\mathfrak{m}}$ is some function of $\mathcal{E}_{\mathfrak{m}}$ such that

$$\|f - \log t_{\mathfrak{m}}\|_{\infty} \leq 2\|f - \Pi_{\mathfrak{m}} f\|_{\infty} \tag{12}$$

where $\Pi_{\mathfrak{m}}$ denotes the orthogonal projection onto $\mathcal{P}_{\mathfrak{m}}$. Second, if $|f|_{\alpha, l}$ denotes the norm of f in $\mathcal{B}_{\alpha, l, \infty}([0, 1])$ (see DeVore and Lorentz [24, p. 55]) we derive from approximation theory that

$$\|f - \Pi_{\mathfrak{m}} f\|_2 \leq C_2(r) |f|_{\alpha, 2} D^{-\alpha} \tag{13}$$

$$\|f - \Pi_{\mathfrak{m}} f\|_{\infty} \leq C_{\infty}(r) |f|_{\alpha, l} D^{-\alpha + \frac{1}{l}} \tag{14}$$

for a model $\mathfrak{m} = (m, \underline{r})$ such that m is some regular partition of $[0, 1]$ with D pieces and $r_I = [\alpha]$ for all $I \in m$. In particular, (12) and (14) imply

$$\lim_{|m| \to +\infty} \|f - \log t_{\mathfrak{m}}\|_{\infty} = 0$$

and

$$\lim_{|m| \to +\infty} \|\log t_{\mathfrak{m}}\|_{\infty} = \|\log s\|_{\infty}.$$

Consequently, the assumption (17) of Lemma 4.1 will be satisfied for the models $\mathfrak{m} \in \mathfrak{M}_n$ such that $|m| \geq N_0$ for some positive integer N_0 depending on s (assuming that n is large enough). We deduce then that for all models $\mathfrak{m} \in \mathfrak{M}_n$ such that $|m| \geq N_0$, $\|\log(s/\overline{s}_{\mathfrak{m}})\|_{\infty}$ is bounded (in fact, Lemma 4.1 gives a little more: $\lim_{|m| \to +\infty} \|\log(s/\overline{s}_{\mathfrak{m}})\|_{\infty} = 0$). We conclude that $\|\log \overline{s}_{\mathfrak{m}}\|_{\infty}$ is uniformly bounded over the family $\mathfrak{M}_n$ by some constant depending on s.

2) Consequence: Now, in order to apply Theorem 2.1, we evaluate the term $K(s, \mathcal{E}_{\mathfrak{m}}) = K(s, \overline{s}_{\mathfrak{m}})$. Using the definition of $\overline{s}_{\mathfrak{m}}$ and Lemma 1 of Barron and Sheu [13, Inequality 3.2] with $p = s$, $q = t_{\mathfrak{m}}$, and $c = C(f_{\mathfrak{m}}) - \int f\, d\mu$, we get

$$\begin{aligned} K(s, \overline{s}_{\mathfrak{m}}) & \leq K(s, t_{\mathfrak{m}}) \\ & \leq \Phi\left(\|f - \Pi_{\mathfrak{m}} f\|_{\infty}\right) \int s(f - \Pi_{\mathfrak{m}} f)^2\, d\mu. \end{aligned}$$

Finally, using (13), Theorem 2.1 yields

$$\begin{aligned} & \boldsymbol{E}\left[h^2(s, \tilde{s})\, \mathbb{1}_{\{\tilde{s} \geq \rho_n\}}\right] \\ & \leq C_1(c) \inf_{D \geq 1} \left\{ e^{C_{\infty}(r)|f|_{\alpha, l}} |f|_{\alpha, 2}^2 D^{-2\alpha} + \frac{D}{n} \right\} + C_3(f, c)/n^2 \\ & \leq C_1(c) \left(e^{C_{\infty}(r)|f|_{\alpha, l}} |f|_{\alpha, 2}^2 \right)^{\frac{-2\alpha}{2\alpha+1}} e^{C_{\infty}(r)|f|_{\alpha, l}} |f|_{\alpha, 2}^2 n^{-\frac{2\alpha}{2\alpha+1}} \\ & \quad + C_3(f, c)/n^2 \\ & \leq C_1(c)\, e^{\frac{1}{2\alpha+1} C_{\infty}(r)|f|_{\alpha, l}} |f|_{\alpha, 2}^{\frac{2}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} + C_3(f, c)/n^2. \end{aligned}$$

3) Control of the Set $\{\tilde{s} \geq \rho_n\}$: It remains to show that the set $\{\tilde{s} \geq \rho_n\}$ has high probability when n is large enough. We demonstrate that if n is large enough then we have $\tilde{s} \geq \rho_n$ on the set Ω_n defined by (5). We will prove, in fact, that $\|\log \hat{s}_{\mathfrak{m}}\|_\infty$ is uniformly bounded over $\mathfrak{M}_n$ on Ω_n. Since $\|\log \overline{s}_{\mathfrak{m}}\|_\infty$ is uniformly bounded over $\mathfrak{M}_n$ by some constant M, depending on s, we can apply Lemma 1.1 to $\delta^0 = \overline{\delta}_I$ and $\delta = \hat{\delta}_I$ with $b_0 = e^{-M}/P(I)$ for all $I \in m$. Then , if n is large enough (such that $\rho_n \leq \rho = \operatorname{essinf} s$ and $\rho_n \leq e^{-M}$), the assumption (15) of Lemma 1.1 will be satisfied over the set Ω_n, and we derive that $\|\log(\hat{s}_{\mathfrak{m}}/\overline{s}_{\mathfrak{m}})\|_\infty$ is uniformly bounded over Ω_n. Thus, if n is large enough, $\|\log \hat{s}_{\mathfrak{m}}\|_\infty$ is uniformly bounded over $\mathfrak{M}_n$ on the set Ω_n. Consequently, on Ω_n we have $\hat{s}_{\mathfrak{m}} \geq \rho_n$ for all the models $\mathfrak{m}$ of the family $\mathfrak{M}_n$ provided that n is large enough. We then conclude that $\Omega_n \subset \{\tilde{s} \geq \rho_n\}$ for all $n \geq n_0$ for some integer n_0 depending on s. Thus, for $n \geq n_0$, we get

$$\begin{aligned}\boldsymbol{E}\left[h^2(s,\tilde{s})\right] &\leq \boldsymbol{E}\left[h^2(s,\tilde{s})\mathbb{1}_{\{\tilde{s}\geq\rho_n\}}\right] + \boldsymbol{E}\left[h^2(s,\tilde{s})\mathbb{1}_{\Omega_n^c}\right] \\ &\leq \boldsymbol{E}\left[h^2(s,\tilde{s})\mathbb{1}_{\{\tilde{s}\geq\rho_n\}}\right] + \boldsymbol{P}(\Omega_n^c).\end{aligned}$$

We conclude with (15) and the control of $\boldsymbol{P}(\Omega_n^c)$.

Appendix I
Invertibility of the Parametrization of Exponential Families

We use the notations of Section III-A1 and we study the equations of the form $G(\beta) = \delta$. More precisely, the following lemma allows to quantify the fact that if δ is close enough to some $\delta^0 = G\left(\beta^0\right)$ then the solution β^δ of $G(\beta) = \delta$ is close to β^0. The proof of this lemma is adapted from Barron and Sheu [13, Lemma 5] (see also [22, Ch. 2, Lemma 2.5.6] for a complete proof).

Lemma 1.1: Let $\delta^0 \in G\left(\mathbb{R}^\Lambda\right)$ and β^0 such that $G\left(\beta^0\right) = \delta^0$. We denote

$$B_\Lambda = \sup_{c\in\mathbb{R}^\Lambda:\,|c|_2=1} \left\|\sum_{\lambda\in\Lambda} c_\lambda \varphi_\lambda\right\|_\infty$$

and we assume that $\inf_{\mathbb{X}} t\left(\beta^0\right) \geq b_0 > 0$. If $\delta \in \mathbb{R}^\Lambda$ satisfies

$$\left|\delta - \delta^0\right|_2 < \frac{b_0}{2B_\Lambda} \tag{15}$$

then there exists $\beta^\delta \in \mathbb{R}^\Lambda$ such that $G\left(\beta^\delta\right) = \delta$, furthermore

$$\left|\beta^\delta - \beta^0\right|_2 \leq \frac{1}{2B_\Lambda}\, g^{-1}\left(\frac{2B_\Lambda\left|\delta-\delta^0\right|_2}{b_0}\right)$$

and

$$\left\|\log\frac{t\left(\beta^\delta\right)}{t\left(\beta^0\right)}\right\|_\infty \leq g^{-1}\left(\frac{2B_\Lambda\left|\delta-\delta^0\right|_2}{b_0}\right)$$

where $g^{-1}\colon\,]0,\,1[\mapsto \mathbb{R}_+$ denotes the inverse of the function g defined on $(0,\,+\infty)$ by $g(x) = \left(e^{-x} - 1 + x\right)x^{-1}$.

Remark 1: For all $\varepsilon \in]0,\,1[$, the condition

$$\left|\delta - \delta^0\right|_2 \leq \frac{g(\varepsilon)\, b_0}{2B_\Lambda} = \frac{\varepsilon\Phi(-\varepsilon)\, b_0}{2B_\Lambda}$$

leads to

$$\left\|\log\frac{t\left(\beta^\delta\right)}{t\left(\beta^0\right)}\right\|_\infty \leq g^{-1}\left(\frac{2B_\Lambda\left|\delta-\delta^0\right|_2}{b_0}\right) \leq \varepsilon < 1.$$

Remark 2: We cannot give an explicit value for g^{-1}, but we can get the following upper bound:

$$g^{-1}(x) \leq \begin{cases} 4x, & \text{if } 0 < x \leq \frac{1}{2} \\ \frac{1}{1-x}, & \text{if } \frac{1}{2} < x < 1. \end{cases}$$

Consequently, this gives, if $0 \leq \left|\delta-\delta^0\right|_2 \leq \frac{b_0}{4B_\Lambda}$

$$\left\|\log\frac{t\left(\beta^\delta\right)}{t\left(\beta^0\right)}\right\|_\infty \leq \frac{8B_\Lambda\left|\delta-\delta^0\right|_2}{b_0}$$

and if $\frac{b_0}{4B_\Lambda} < \left|\delta-\delta^0\right|_2 < \frac{b_0}{2B_\Lambda}$

$$\left\|\log\frac{t\left(\beta^\delta\right)}{t\left(\beta^0\right)}\right\|_\infty \leq \frac{b_0}{b_0 - 2B_\Lambda\left|\delta-\delta^0\right|_2}.$$

Appendix II
Technical Lemma on Kullback–Leibler Information

The following lemma is used in the proof of Theorem 2.1 to connect the Kullback–Leibler information $K\left(\overline{s}_{\mathfrak{m}'},\,\hat{s}_{\mathfrak{m}'}\right)$ and the "variance" term $V_s^2\left(\overline{s}_{\mathfrak{m}'},\,\hat{s}_{\mathfrak{m}'}\right)$ for some model $\mathfrak{m}' \in \mathfrak{M}_n$.

Lemma 2.1: If $p,\ qands$ are some positive densities with respect to μ such that

$$\left\|\log\frac{p}{q}\right\|_\infty \leq \eta$$

for some positive number η, then, for all positive θ we have

$$K(p,\,q) \geq \Phi(-\eta)\left(1-\theta\eta\sqrt{2}\right)\int s\left(\log\frac{p}{q}\right)^2 d\mu - \Phi(-\eta)\,\frac{\eta}{\theta}\,\sqrt{2}h^2(s,\,p)$$

where Φ is given by $\Phi(x) = \left(e^x - 1 - x\right)x^{-2}$ for all $x \in \mathbb{R}$.

Proof: Since $\|\log(p/q)\|_\infty \leq \eta$, [13, Lemma 1 (3.1)] of Barron and Sheu leads to

$$K(p,\,q) \geq \Phi(-\eta)\int p\left(\log\frac{q}{p}\right)^2 d\mu.$$

Moreover

$$\begin{aligned}&\int p\left(\log\frac{q}{p}\right)^2 d\mu \\ &\quad\geq \int_{p\geq s} s\left(\log\frac{q}{p}\right)^2 d\mu + \int_{p<s} p\left(\log\frac{q}{p}\right)^2 d\mu \\ &\quad= \int s\left(\log\frac{q}{p}\right)^2 d\mu - \int (s-p)_+\left(\log\frac{q}{p}\right)^2 d\mu\end{aligned}$$

and

$$\begin{aligned}&\int (s-p)_+\left(\log\frac{q}{p}\right)^2 d\mu \\ &\quad\leq \eta\int\frac{(s-p)_+}{\sqrt{s}+\sqrt{p}}\left(\sqrt{s}+\sqrt{p}\right)\left|\log\frac{q}{p}\right| d\mu \\ &\quad\leq \eta\left[\int\left(\frac{s-p}{\sqrt{s}+\sqrt{p}}\right)^2 d\mu\right]^{\frac{1}{2}} \\ &\qquad\times\left[\int_{s>p}\left(\sqrt{s}+\sqrt{p}\right)^2\left(\log\frac{q}{p}\right)^2 d\mu\right]^{\frac{1}{2}} \\ &\quad\leq \eta 2\sqrt{2}\left[h^2(s,\,p)\right]^{\frac{1}{2}}\left[\int s\left(\log\frac{q}{p}\right)^2 d\mu\right]^{\frac{1}{2}} \\ &\quad\leq \eta\sqrt{2}\left[\frac{1}{\theta}h^2(s,\,p) + \theta\int s\left(\log\frac{q}{p}\right)^2 d\mu\right] \\ &\quad= \frac{\eta}{\theta}\sqrt{2}h^2(s,\,p) + \theta\eta\sqrt{2}\int s\left(\log\frac{q}{p}\right)^2 d\mu.\end{aligned}$$

Consequently, we conclude with

$$\int p\left(\log\frac{q}{p}\right)^2 d\mu \geq \left(1-\theta\eta\sqrt{2}\right)\int s\left(\log\frac{q}{p}\right)^2 d\mu - \frac{\eta}{\theta}\sqrt{2}h^2(s,p). \quad \square$$

Appendix III
Control of a Chi-Square-Type Statistics

This appendix is devoted to the following lemma which will be used in the proof of Theorem 2.1 to control $\nu_n(\log(\hat{s}_{\mathfrak{m}'}/\overline{s}_{\mathfrak{m}'}))$ uniformly over $\mathfrak{M}_n$. The proof of this lemma is based on a Talagrand-type inequality and, more precisely, we use the version of Bernstein's inequality due to Bousquet [20].

Lemma 3.1: Let $\mathfrak{m}'=(m',\underline{r})$ be some model of $\mathfrak{M}_n$, $\mathcal{P}_{\mathfrak{m}'}$ be the linear space of piecewise polynomials based on $\mathfrak{m}'$ (defined in Section II-A), and $Z_{\mathfrak{m}'}$ be the random variable defined by

$$Z_{\mathfrak{m}'}=\sup\left\{|\nu_n(f)|,\ f\in\mathcal{P}_{\mathfrak{m}'},\ \int f^2 s\,d\mu=1\right\}$$

where ν_n is the centered empirical measure. For all $\varepsilon\in\,]0,1[$ and all $x>0$

$$\boldsymbol{P}\left[Z_{\mathfrak{m}'}\mathbb{1}_{\Omega_{\mathfrak{m}'}}\geq(1+\varepsilon)\left(\sqrt{\frac{D_{\mathfrak{m}'}}{n}}+\sqrt{\frac{2x}{n}}\right)\right]\leq\exp(-x)$$

with $D_{\mathfrak{m}'}=\dim\mathcal{P}_{\mathfrak{m}'}-1$ and

$$\Omega_{\mathfrak{m}'}=\bigcap_{\substack{k\in\{0,\ldots,r_I\}\\ I\in m'}}\left\{|\nu_n(\psi_{I,k})|\leq C(r,\varepsilon)\rho\sqrt{\mu(I)}\right\}$$

where $\{\psi_{I,k},\ 0\leq k\leq r_I,\ I\in m'\}$ is the "Legendre basis" of $\mathcal{P}_{\mathfrak{m}'}$ and $C(r,\varepsilon)=6\varepsilon^2(r+1)^{-\frac{5}{2}}(3+\varepsilon)^{-1}$.

Proof: We will use an orthonormal basis

$$\left\{\phi_{I,k}, I\in m',\ 0\leq k\leq r_I\right\}\ \text{of}\ \mathcal{P}_{\mathfrak{m}'}\ \text{in}\ \mathbb{L}_2([0,1],P)$$

where for every $I\in m'$ the family $\{\phi_{I,k},\ 0\leq k\leq r_I\}$ is an orthonormal basis of $\mathcal{P}_I$ in $\mathbb{L}_2(I,P)$ and is derived from the "Legendre basis" of $\mathcal{P}_I$. We choose $\phi_{I,0}=P(I)^{-1/2}\mathbb{1}_I$, and for every $k\in\{1,\ldots,r_I\}$, the function $\phi_{I,k}$ is a linear combination of $\psi_{I,l}$ for $0\leq l\leq k$ with coefficients depending on s. Using this basis, we can write

$$Z_{\mathfrak{m}'}=\sup_{a\in\mathcal{B}_{\mathfrak{m}'}}\left|\sum_{I\in m'}\sum_{0\leq k\leq r_I}a_{I,k}\nu_n(\phi_{I,k})\right| =\left[\sum_{I\in m'}\sum_{0\leq k\leq r_I}\nu_n^2(\phi_{I,k})\right]^{\frac{1}{2}} \tag{16}$$

where $\mathcal{B}_{\mathfrak{m}'}=\left\{a\in\prod_{I\in m'}\mathbb{R}^{r_I+1}:\ |a|_2=1\right\}$.

We first control, on the set $\Omega_{\mathfrak{m}'}$, the fluctuations $|\nu_n(\phi_{I,k})|$ for all $I\in m'$ and all $k\in\{0,\ldots,r_I\}$. On $\Omega_{\mathfrak{m}'}$, for all $I\in m'$ and $k\in\{0,\ldots,r_I\}$, we control $|\nu_n(\psi_{I,k})|$, and this allows us to bound $|\nu_n(\phi_{I,k})|$. Indeed, there exists $c\in\mathbb{R}^{k+1}$ (which depends on s, I, and k) such that $\phi_{I,k}=\sum_{0\leq l\leq k}c_l\psi_{I,l}$. Consequently, we have

$$\nu_n(\phi_{I,k})=\sum_{0\leq l\leq k}c_l\nu_n(\psi_{I,l}).$$

Moreover, $\int\phi_{I,k}^2 s\,d\mu\geq\rho\int\phi_{I,k}^2\,d\mu$ leads to $|c|_2\leq\rho^{-1/2}$. Thus, on the set $\Omega_{\mathfrak{m}'}$

$$|\nu_n(\phi_{I,k})|\leq|c|_2\left[\sum_{0\leq l\leq k}\nu_n(\psi_{I,l})^2\right]^{\frac{1}{2}} \leq(k+1)^{\frac{1}{2}}C(r,\varepsilon)\sqrt{\rho\mu(I)}.$$

Next, the supremum in (16) is achieved at $a\in\prod_{I\in m'}\mathbb{R}^{r_I+1}$ such that $a_{I,k}=\nu_n(\phi_{I,k})/Z_{\mathfrak{m}'}$. We define

$$\mathcal{A}_{\mathfrak{m}'}=\left\{a\in\prod_{I\in m'}\mathbb{R}^{r_I+1}:\ |a|_2=1,\ |a_I|_2\leq\sqrt{\rho\mu(I)}(r_I+1)^{\frac{3}{2}}C(r,\varepsilon)/z,\ \forall I\in m'\right\}$$

where z is some positive number which will be chosen later and $|a_I|_2$ denotes the Euclidean norm of the vector $a_I=(a_{I,0},\ldots,a_{I,r_I})$. Thus, we obtain on the set $\Omega_{\mathfrak{m}'}\cap\{Z_{\mathfrak{m}'}\geq z\}$

$$Z_{\mathfrak{m}'}=\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left|\sum_{I\in m'}\sum_{0\leq k\leq r_I}a_{I,k}\nu_n(\phi_{I,k})\right|$$

where $\mathcal{A}'_{\mathfrak{m}'}$ is a countable and dense subset of $\mathcal{A}_{\mathfrak{m}'}$. We are now in a position to apply the deviation inequality due to Bousquet [20, see Theorem 2.3] to the random variable $Z_{\mathfrak{m}'}$. This leads to

$$\boldsymbol{P}\left[\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left|\nu_n\left(\sum_{I\in m'}\sum_{0\leq k\leq r_I}a_{I,k}\phi_{I,k}\right)\right|\geq(1+\varepsilon)E_{\mathfrak{m}'} +\sqrt{\frac{2\sigma^2x}{n}}+\left(\frac{1}{3}+\frac{1}{\varepsilon}\right)\frac{bx}{n}\right]\leq e^{-x}$$

where

$$E_{\mathfrak{m}'}=\boldsymbol{E}(Z_{\mathfrak{m}'})\leq\boldsymbol{E}^{\frac{1}{2}}\left(\sum_{I\in m'}\sum_{0\leq k\leq r_I}\nu_n^2(\phi_{I,k})\right)\leq\sqrt{\frac{D_{\mathfrak{m}'}}{n}}$$

and

$$\sigma^2=\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left(\operatorname{Var}\left(\sum_{I\in m'}\sum_{0\leq k\leq r_I}a_{I,k}\phi_{I,k}(X_1)\right)\right)\leq 1$$

and, using (4)

$$\begin{aligned}b&=\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left\|\sum_{I\in m'}\sum_{0\leq k\leq r_I}a_{I,k}\phi_{I,k}\right\|_\infty\\ &=\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left\{\max_{I\in m'}\left\|\sum_{0\leq k\leq r_I}a_{I,k}\phi_{I,k}\right\|_\infty\right\}\\ &\leq\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left\{\max_{I\in m'}\left\{\frac{r_I+1}{\sqrt{\mu(I)}}\left\|\sum_{0\leq k\leq r_I}a_{I,k}\phi_{I,k}\right\|_\infty\right\}\right\}\\ &\leq\sup_{a\in\mathcal{A}'_{\mathfrak{m}'}}\left\{\max_{I\in m'}\left\{(r_I+1)[\rho\mu(I)]^{-\frac{1}{2}}|a_I|_2\right\}\right\}\\ &\leq\frac{(r+1)^{\frac{5}{2}}C(r,\varepsilon)}{z}.\end{aligned}$$

Consequently, using $C(r,\varepsilon)=6\varepsilon^2(r+1)^{-\frac{5}{2}}(3+\varepsilon)^{-1}$

$$\boldsymbol{P}\left[Z_{\mathfrak{m}'}\mathbb{1}_{\Omega_{\mathfrak{m}'}\cap\{Z_{\mathfrak{m}'}\geq z\}}\geq(1+\varepsilon)\sqrt{\frac{D_{\mathfrak{m}'}}{n}}+\sqrt{\frac{2x}{n}}+\frac{2\varepsilon x}{zn}\right]\leq\exp(-x).$$

Take $z=\sqrt{\frac{2x}{n}}$ to conclude. $\square$

Appendix IV
Control of $\|\log(s/\overline{s}_{\mathfrak{m}})\|_\infty$

Lemma 4.1: Let s be some positive density on $[0,1]$, m be some partition, $\underline{r}\in\mathbb{N}^m$, and $\mathfrak{m}=(m,\underline{r})$. Let $\overline{s}_{\mathfrak{m}}$ be the information projection of the density s on the exponential model $\mathcal{E}_{\mathfrak{m}}$. We assume that there exists some density $t_{\mathfrak{m}}$ in $\mathcal{E}_{\mathfrak{m}}$ such that $\log t_{\mathfrak{m}}$ is bounded and satisfies

$$\left\|\log\frac{s}{t_{\mathfrak{m}}}\right\|_\infty\leq\frac{1}{2}\log\left(1+\left(4(r+1)^{\frac{5}{2}}e^{2\|\log t_{\mathfrak{m}}\|_\infty}\right)^{-1}\right) \tag{17}$$

then $\|\log(s/\overline{s}_{\mathtt{m}})\|_\infty \leq 2\left(\|\log(s/t_{\mathtt{m}})\|_\infty + 1\right)$.

Proof: Let us consider the measure $d\nu = t_{\mathtt{m}}\, d\mu$, the exponential model $\mathcal{E}^\nu_{\mathtt{m}}$ given by

$$\mathcal{E}^\nu_{\mathtt{m}} = \left\{ t \in \mathbb{L}^1([0,\,1],\,\nu) : t > 0,\, \log t \in \mathcal{P}_{\mathtt{m}},\, \int t\, d\nu = 1 \right\}$$

and the Kullback–Leibler information $K^\nu(p,\,q)$ defined for any probability distributions P, resp., Q, with density p, resp., q, with respect to ν. Now, we consider the information projection $\overline{g}^{\,\nu}_{\mathtt{m}}$, with respect to K^ν, of the density $g = s/t_{\mathtt{m}}$ with respect to ν. We can prove that $\overline{g}^{\,\nu}_{\mathtt{m}} = \overline{s}_{\mathtt{m}}/t_{\mathtt{m}}$. And it remains to show that the logarithm of the information projection of g, $\|\log \overline{g}^{\,\nu}_{\mathtt{m}}\|_\infty$ is small provided that $\|\log g\|_\infty$ is small enough. This can be done using Lemma 1.1.

Let us consider the orthonormal basis of $\mathcal{P}_{\mathtt{m}}$ in $\mathbb{L}^2([0,\,1],\,\nu)$, $\{\psi^\nu_{I,\,k},\, I \in m,\, 0 \leq k \leq r_I\}$, where the family $\left\{\psi^\nu_{I,\,k},\, 0 \leq k \leq r_I\right\}$ is an orthonormal basis of $\mathcal{P}_I$ in $L^2(I,\,\nu)$ for every $I \in m$. We choose $\psi^\nu_{I,\,0} = \nu(I)^{-1/2}\,\mathbb{1}_I$. And t^ν_I, G^ν_I are the function given in Section III-A1 associated with the orthonormal family of functions $\left\{\psi^\nu_{I,\,k},\, 0 \leq k \leq r_I\right\}$ in $\mathbb{L}^2(I,\,\nu)$. For $I \in m$ such that $r_I \geq 1$, we denote $\overline{\beta}^{\,\nu}_I \in \mathbb{R}^{r_I}$ the unique solution of the equation $G^\nu_I\left(\overline{\beta}^{\,\nu}_I\right) = \overline{\delta}^{\,\nu}_I$ where $\overline{\delta}^{\,\nu}_I \in \mathbb{R}^{r_I}$ is defined by

$$\overline{\delta}^{\,\nu}_{I,\,k} = \frac{\int_I \psi^\nu_{I,\,k} g\, d\nu}{\int_I g\, d\nu}, \qquad \text{for all } k \in \{1,\,\ldots,\,r_I\}.$$

Then, by definition of the information projection, we can write

$$\overline{g}^{\,\nu}_{\mathtt{m}} = \sum_{I \in m} \left(\int_I g\, d\nu \right) \overline{g}^{\,\nu}_I\, \mathbb{1}_I$$

where

$$\overline{g}^{\,\nu}_I = \begin{cases} \nu(I)^{-1}\,\mathbb{1}_I, & \text{if } r_I = 0 \\ t^\nu_I\left(\overline{\beta}^{\,\nu}_I\right), & \text{if } r_I \geq 1. \end{cases}$$

Now, we apply Lemma 1.1 (Remark 2.) on the measurable space $(I,\,\nu)$ with the orthogonal family in $\mathbb{L}^2(I,\,\nu)$, $\{\mathbb{1}_I,\, \psi^\nu_{I,\,k},\, 1 \leq k \leq r_I\}$ to the vectors $\delta^0 = 0$ $(\beta^0 = 0)$ and $\delta = \overline{\delta}^{\,\nu}_I$ $(\beta = \overline{\beta}^{\,\nu}_I)$. Under assumption (17), we can prove that $\left|\overline{\delta}^{\,\nu}_I\right|_2$ is small enough to satisfy the condition (15) of Lemma 1.1, and this leads to a control of $\|\log(\overline{g}^{\,\nu}_I \nu(I))\|_\infty$ for every $I \in m$. Thus, we can derive a control of

$$\|\log \overline{g}^{\,\nu}_{\mathtt{m}}\|_\infty = \|\log(\overline{s}_{\mathtt{m}}/t_{\mathtt{m}})\|_\infty$$

and conclude using (17). □

REFERENCES

[1] L. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. Math. Statist., 1986.

[2] O. Barndorff-Nielsen, *Exponential Families, Exact Theory*, ser. no. 19. Aarhus, Denmark: Aarhus Univ., Various publications, 1970.

[3] J. Neyman, "'Smooth' test for goodness of fit," *Skand. Akt.*, vol. 20, pp. 149–199, 1937.

[4] B. R. Crain, "Estimation of distributions using orthogonal expansions," *Ann. Statist.*, vol. 2, pp. 453–463, 1974.

[5] ——, "Exponential models, maximum likelihood estimation and the Haar condition," *J. Amer. Statist. Assoc.*, vol. 71, pp. 737–740, 1976.

[6] ——, "More on estimation of distributions using orthogonal expansions," *J. Amer. Statist. Assoc.*, vol. 71, pp. 741–745, 1976.

[7] ——, "An information theoretic approach to approximating a probability distribution," *SIAM J. Appl. Math.*, vol. 32, pp. 339–346, 1977.

[8] S. Portnoy, "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity," *Ann. Statist.*, vol. 16, pp. 356–366, 1988.

[9] N. N. Cencov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: Amer. Math. Soc., 1982, vol. 53. Amer. Math. Soc. Translations.

[10] C. Stone and C. Koo, "Logspline density estimation," in *Contemporary Mathematics*. Providence, RI: Amer. Math. Soc., 1986, vol. 59, pp. 1–15.

[11] C. Stone, "Uniform error bounds involving logspline models," in *Probability, Statistics and Mathematics: Paper in Honor of Samuel Karlin*, T. Anderson and K. Athreya, Eds. Boston, MA: Academic, 1989, pp. 335–355.

[12] ——, "Large sample inference for logspline models," *Ann. Statist.*, vol. 18, pp. 717–741, 1990.

[13] A. Barron and C. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, pp. 1347–1369, 1991.

[14] J.-Y. Koo and W. Kim, "Wavelet density estimation by approximation of log-densities," *Statist. Probab. Lett.*, vol. 21, pp. 271–278, 1996.

[15] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory*, P. Petrov and E. Csaki, Eds. Budapest, Hungary: Akademia Kiado, 1973, pp. 267–281.

[16] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–415, 1999.

[17] Y. Yang and A. Barron, "An asymptotic property of model selection criteria," *IEEE Trans. Inform. Theory*, vol. 44, pp. 95–116, Jan. 1998.

[18] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, vol. 65, pp. 181–237, 1983.

[19] G. Castellan, "Modified Akaike's criterion for histogram density estimation," Univ. Paris-Sud, Orsay, France, Preprint 99.61, 1999.

[20] O. Bousquet, "A Bennett concentration inequality and its application to suprema of empirical processes," *C. R. Math. Acad. Sci. Paris*, vol. 334, no. 6, pp. 495–500, 2002.

[21] L. Birgé and P. Massart, "From model selection to adaptive estimation," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. Yang, Eds. New York: Springer-Verlag, 1997, pp. 55–87.

[22] G. Castellan, "Sélection d'histogrammes ou de modèles exponentiels de polynômes par morceaux à l'aide d'un critère de type Akaike," Ph.D. dissertation, Univ. Paris-Sud, Orsay, France, 2000.

[23] L. Birgé and Y. Rozenholc, "How many bins should be put in a regular histogram," Universités Paris VI et Paris VII, Preprint 721, 2002.

[24] R. DeVore and G. Lorentz, *Constructive Approximation*. Berlin: Springer-Verlag, 1993.

[25] L. Devroye and G. Lugosi, "Nonasymptotic universal smoothing factors, kernel complexity and yatracos classes," *Ann. Statist.*, vol. 25, no. 6, pp. 2626–2637, 1997.

[26] L. Birgé and P. Massart, "Minimum contrast estimators on sieves: Exponential bounds and rates of convergence," *Bernoulli*, vol. 4, pp. 329–375, 1998.

[27] E. Whittaker and G. Watson, *A Course of Modern Analysis*. London, U.K.: Cambridge Univ. Press, 1927.

[28] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.