TECHNICAL
REPORT

0607

MAXIMUM LIKELIHOOD ESTIMATION IN POISSON REGRESSION VIA WAVELET MODEL SELECTION RUNNING TITLE : POISSON REGRESSION VIA MODEL SELECTION

LETUE F., and F. LEBLANC

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Maximum Likelihood Estimation in Poisson Regression via Wavelet Model Selection
Running title : Poisson regression via model selection

Frédérique Leblanc* and Frédérique Letué
LMC/SMS-UJF, LMC/SMS-UJF and LabSAD/UPMF
Tour IRMA
51, rue des Mathématiques, B.P. 53
38041 Grenoble cedex 9
FRANCE
Frederique.Leblanc@imag.fr, Frederique.Letue@imag.fr

21st February 2006

Abstract

In this work we estimate the regression function for Poisson variables, for a deterministic design in $[0, 1]$. Our final estimator which is adaptive to the data, is selected amoung a collection of maximum likelihood estimators with respect to a penalized empirical Kullback-Leibler risk. We obtain, an oracle inequality over the Kullback-Leibler risk for any fixed size n of the design. Moreover, we state an asymptotic lower bound on this risk over Sobolev spaces and prove that our estimator reaches this rate. Hence the selected estimator is asymptotically minimax over these spaces. We also present numerical experiments, including a strategy to adjust the constants involved in the penalty function which defines the selection criteria, which performs as well as the ideal one.

Keywords and phrases: Adaptive estimator, Kullback-Leibler risk, maximum likelihood estimator, minimax rate, model selection, penalization, Poisson regression, oracle inequality, wavelets.

1 Introduction

In many practical situations the collected data are observations of counting variables that are often modelized through Poisson regression. Many authors have already discussed different nonparametric estimation procedures and among them wavelets methods.

In (Besbeas, De Feis & Sapatinas 2004), the authors provide a quite complete presentation of wavelets methods for estimating the intensity of a Poisson process and they examine the performance of the proposed estimators via simulations. Among these methods, it is worth noticing Donoho's (Donoho 1993) using the Anscombe transformation and Fryżlewicz and Nason's (Fryzlewicz & Nason 2004) using the Fisz transformation in order to stabilize the variance. Other authors (Kolaczyk 1997), (Kolaczyk 1999*b*), (Nowak & Baraniuk 1999) propose wavelet shrinkage techniques to be applied directly to the given Poisson process. Bayesian procedures also have been proposed by (Kolaczyk 1999*a*) and (Timmermann & Nowak 1999). Finally, some wavelets techniques can be applied to a larger family of distributions containing the Poisson one (see (Antoniadis & Sapatinas 2001), (Antoniadis, Besbeas & Sapatinas 2001) and (Sardy, Antoniadis & Tseng 2004)). However the procedures used in the previous papers give asymptotic results and are based on penalized or shrinked estimators minimizing L_p risks (mainly the quadratic one).

In this paper, we adopt a model selection strategy following Birgé and Massart' ideas (see for instance, (Barron, Birgé & Massart 1999), (Birgé & Massart 2001)). One of the advantages of such approach is to provide non asymptotic risk upper bounds. It has already been applied to various frameworks by different authors. Among them, we shall cite regression in a fixed design

(Baraud 2000), density estimation via histograms (Castellan 1999) and via piecewise polynomials (Castellan 2003)) and Poisson process intensity estimation (Reynaud-Bouret 2003).

Herein, we observe n independent copies $(Y_i, x_i)_{1\leq i\leq n}$, where the Y_i are discrete random responses and the $(x_i)_{1\leq i\leq n}$ is a deterministic design in $[0,1]$. Each random variable Y_i is supposed to have a Poisson distribution with parameter $\mu_i = \exp f(x_i)$. This parametrization of the regression function is natural in the framework of Generalized Linear Models when using the canonical link function (see for instance (McCullagh & Nelder 1989)). Our aim is to estimate the function f in some large space S_Λ generated by wavelet basis.

We define our models as linear subspaces of a larger one S_Λ. We construct the collection of maximum likelihood estimators within each model and our goal is to select the "best" one amoung them in the sense of the Kullback-Leibler risk.

More precisely, let $(\varphi_\lambda)_{\lambda\in\Lambda}$ be a basis of S_Λ. For any subset m of the larger index set Λ whose cardinal, denoted $|\Lambda|$, is finite, the model S_m of dimension denoted D_m is defined as

$$S_m = \{\sum_{\lambda\in m} \beta_\lambda \phi_\lambda, (\beta_\lambda)_\lambda \in \mathbb{R}^{D_m}\}.$$

On each of these models, the maximum likelihood estimator on S_m is defined as

$$\hat{f}_m = \arg\min_{h\in S_m} \gamma_n(h), \tag{1.1}$$

where the contrast function γ_n is the opposite of the log-likelihood:

$$\gamma_n(h) = n^{-1}\sum_{i=1}^{n}(e^{h(x_i)} - Y_i h(x_i)).$$

In order to compare the estimators of the collection, we introduce the Kullback-Leibler loss between two functions f and h as:

$$K(f,h) = \mathbb{E}_f(\gamma_n(h) - \gamma_n(f)) = n^{-1}\sum_{i=1}^{n} e^{h(x_i)} - e^{f(x_i)} - e^{f(x_i)}(h(x_i) - f(x_i)).$$

Denoting by $\bar{f}_m$ the function in S_m minimizing the Kullback-Leibler loss function,

$$\bar{f}_m = \arg\min_{h\in S_m} K(f,h), \tag{1.2}$$

we can prove that

$$\mathbb{E}_f(K(f,\hat{f}_m)) = K(f,\bar{f}_m) + \mathbb{E}_f(K(\bar{f}_m,\hat{f}_m)).$$

In this decomposition, the first term represents a deterministic projection error, whereas the second term is an estimation error within the model S_m.

Considering the collection of estimators $\{\hat{f}_m, m\in\mathcal{M}_n\}$, the best estimator in this collection in the sense of the Kullback-Leibler loss is $\hat{f}_{m^*}$, where

$$m^* = \arg\min_{m\in\mathcal{M}_n} \mathbb{E}_f(K(f,\hat{f}_m)) = \arg\min_{m\in\mathcal{M}_n} (K(f,\bar{f}_m) + \mathbb{E}_f(K(\bar{f}_m,\hat{f}_m))).$$

The ideal model m^* is therefore the one which realizes the best trade-off between the approximation and the estimation errors. Unfortunately, this model is not available since it depends on the unknown function f to be estimated.

Consequently, we define the penalized maximum likelihood estimator as $\hat{f}_{\hat{m}}$ where

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} (\gamma_n(\hat{f}_m) + \mathrm{pen}(m)). \tag{1.3}$$

The aim of this paper is to propose penalty functions $\mathrm{pen}(\cdot)$ for which we are able to prove an oracle inequality for a given n, such as:

$$\mathbb{E}_f(K(f,\hat{f}_{\hat{m}})) \simeq \min_{m\in\mathcal{M}_n} \mathbb{E}_f(K(f,\hat{f}_m)).$$

The main difference between Reynaud-Bouret's work (Reynaud-Bouret 2003) and ours lies in that she uses penalized projection estimators and provides L_2 risk for her estimators. Her method can, like ours, be used with any wavelet basis, but also with histograms (other that Haar basis), piecewise polynomials and Fourier basis.

In a recent paper (Baraud & Birgé 2005), Baraud and Birgé develop histograms type estimators for nonnegative random variable, including Poisson variables. Here the method relies on the (not necessarily dyadic nor regular) histogram structure and cannot be simply adapted to other bases. They furnish the same kind of results as ours using a Hellinger type risk.

(Kolaczyk & Nowak 2004) and (Kolaczyk & Nowak 2005) also proposed complexity penalized likelihood estimators in frameworks that include the Poisson model. They prove adaptivity and minimax near-optimality of their estimators in the sense of the squared Hellinger distance. However, their method heavily depend on the ("unbalanced") Haar basis for (Kolaczyk & Nowak 2004) and on piecewise polynomials (Kolaczyk & Nowak 2005), whereas our is available for any wavelet basis with compact support. Furthermore, they give no oracle inequality.

The paper is organized as follows: In Section 2, we give the main definitions and tools about wavelets and Besov spaces and we describe the specific properties of wavelets (localisation for example) that are required to obtain the oracle inequality presented in Section 3. Then in Section 4 is studied a lower bound for the Kullback-leibler loss, over a ball of Hölder or Sobolev Space when an equispaced design is considered. These results provide the usual minimax rate for our final estimator over Sobolev balls. Section 5 is devoted to the numerical experiments and in Section 6 we give the proof of the Oracle inequality. To this end we use a concentration inequality due to Reynaud-Bouret (Reynaud-Bouret 2003). Proofs of the lower bound and technical lemmas are postponed to the Appendix.

2 Wavelets and Besov spaces

2.1 Orthogonal wavelets on $[0,1]$

We start this section by briefly reviewing some useful facts from basic wavelet theory, that will be used to derive our estimators. A general introduction to the theory of wavelets can be found in (Chui 1992), (Daubechies 1992), (Walter 1994) and (Vidakovic 1999). The construction of orthonormal wavelet bases for $L^2(\mathbb{R})$ is now well understood. There are many families of wavelets. Throughout this paper we will consider compactly supported wavelets such as Daubechies' orthogonal wavelets. For the construction of orthonormal bases of compactly supported wavelets for $L^2(\mathbb{R})$, one starts with a couple of special, compactly supported functions known as the scaling function φ and the wavelet ψ. The collection of functions $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx-k)$, $j,k\in\mathbb{Z}$, then constitutes an orthonormal basis for $L^2(\mathbb{R})$. For fixed $j\in\mathbb{Z}$, the $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^jx-k)$, $k\in\mathbb{Z}$ are an orthonormal basis for a subspace $V_j\subset L^2(\mathbb{R})$. The spaces V_j constitute a multiresolution analysis. The subspace generated by $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx-k), k\in\mathbb{Z}$ usually denoted W_j is the orthogonal complement of V_j in V_{j+1} and permits to describe the details at level j of the wavelet decomposition. Indeed, when denoting $P_jf = \sum_{k\in\mathbb{Z}} < f, \varphi_{j,k} > \varphi_{j,k}$ the orthogonal projection of f on the approximation space V_j, we have $P_{j+1}f = P_jf + \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}$.

The multiresolution analysis is said to be r-regular if φ is C^r, and if both φ and its derivatives, up to the order r, have a fast decay. One can prove that if a multiresolution analysis is r-regular, the wavelet ψ is also C^r and has vanishing moments up to the order r (see Corollary 5.2 in (Daubechies 1992)).

The smoother wavelets provide not only orthonormal bases for $L^2(\mathbb{R})$, but also unconditional bases for several function spaces including Besov spaces (see (Triebel 1983)).

Let us consider now orthogonal wavelets on the interval $[0,1]$. Adapting wavelets to a finite interval requires some modifications as described in (Cohen, Daubechies & Vial 1993). To summarize, for J_0 such that $2^{J_0}\geq 2r$, the construction in (Cohen et al. 1993) furnishes a finite set of 2^{J_0} scaling functions $\varphi_{J_0,k}$, and for each $j\geq J_0$, 2^j functions $\psi_{j,k}$, such that the collection of these functions forms a complete orthonormal system of $L_2[0,1]$. With this notation, the $L_2[0,1]$ reconstruction formula is

$$f(t) \quad = \quad \sum_{k=0}^{2^{J_0}-1} \alpha_{J_0,k}\varphi_{J_0,k}(t) + \sum_{j \geq J_0} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(t). \tag{2.1}$$

2.2 Besov spaces

In the following we will use Besov spaces on $[0,1]$, $B^{\nu}_{p,q}$ which are rather general and very well described in terms of sequences of wavelet coefficients. In particular for a suitable choice of the three parameters (ν, p, q) we can get Sobolev spaces or Hölder spaces. For the definition of Besov spaces, properties and functional inclusions we refer to (Triebel 1983). Let us just point out that the usual Sobolev space of regularity $\nu > 0$ denoted in the following $H(\nu)$ coincides with the Besov one $B^{\nu}_{2,2}$ and the Hölder space $\Sigma(\nu)$ with $B^{\nu}_{\infty,\infty}$ when $0 < \nu < 1$.

Here we just give the following characterization of the Besov space $B^{\nu}_{p,q}$ in terms of wavelet coefficients of its elements.

Lemma 2.1. *Let $0 < p$, $q \leq \infty$ and $\nu > \max\{(1/p-1), 0\}$. If the scaling function φ and the wavelet function ψ correspond to a multiresolution analysis of $L_2[0,1]$ that is $([\nu]+1)-$regular (here $[\cdot]$ stands for the integer part), then a function f in $L_p[0,1]$ belongs to the Besov space $B^{\nu}_{p,q}$ if and only if it admits the decomposition (2.1) such that*

$$\|f\|_{B^{\nu}_{p,q}} \equiv \|(\alpha_{J_0,k})_k\|_{l_p} + \left(\sum_{j \geq J_0} 2^{jq(\nu+1/2-1/p)} \|(\beta_{j,k})_k\|^q_{l_p} \right)^{1/q} < +\infty$$

for $J_0 \in \mathbb{N}$. The $\|f\|_{B^{\nu}_{p,q}}$ is equivalent to the Besov space norm.

For a proof see (Delyon & Juditsky 1995).

2.3 Notations and wavelet properties

In the sequel we shall use the following notations :

$$\forall f \in L_2[0,1] \quad : \quad \|f\|_2^2 = \int_{[0,1]} f^2(t)dt \quad \text{and} \quad \|f\|_\infty = \sup_{x\in[0,1]} |f(x)|.$$

$$\forall (a_k)_k \in \mathbb{R}^q \quad : \quad |a|_2^2 = \sum_k a_k^2 \quad \text{and} \quad |a|_\infty = \sup_k |a_k|.$$

$$\forall (b_k)_k \in \mathbb{R}^n \quad \text{and} \quad \forall (c_k)_k \in \mathbb{R}^n \quad : \quad < b, c >_n = \frac{1}{n}\sum_{k=1}^{n} b_k c_k \quad \text{and} \quad |b|_n^2 = < b, b >_n .$$

Moreover, the notation $|f|_2$ (resp. $|f|_\infty$, $|f|_n$, $< f, g >_n$) will abusively stand for $|(f(x_i))_i|_2$ (resp. $|(f(x_i))_i|_\infty$, $|(f(x_i))_i|_2/n, < (f(x_i))_i, (g(x_i))_i) >_n$).

We let J such that $2^J = n$ and the set of indices which permits to describe the space V_l is given by:

$$\Lambda_l = \{(-1,0)\} \cup \{\lambda = (j,k); j = 0, ..., l-1; k = 0, ..., 2^j-1\} \ \forall 1 \leq l \leq J; \Lambda_0 = \{(-1,0)\} \text{ and } \Lambda = \Lambda_J.$$

We put $\phi_{(-1,0)} = \varphi$ and for any $\lambda = (j,k) \neq (-1,0)$, $\phi_\lambda = \psi_{j,k}$. Our results are deeply based on the following crucial property of wavelets:

For any $0 \leq l \leq J$, the basis of the linear space S_{Λ_l} is localized in the following sense: there exists some constant $c(\psi)$ such that for any $a \in \mathbb{R}^{2^J}$:

$$\| \sum_{\lambda\in\Lambda_l} a_\lambda \phi_\lambda \|_\infty \leq c(\psi) 2^{l/2} |a|_\infty. \tag{2.2}$$

This property is a direct consequence of the localization of wavelets. Indeed, since the support of $\psi_{j,k}$ has a size proportionnal to $2r2^{-j}$, at any fixed level j, only a finite number of wavelets $\psi_{j,k}$ are overlapping. Hence there exist some constant $c(\psi)$ such that for any $(\beta_{j,k})_k \in \mathbb{R}^{2^j}$, $\|\sum_{k=0,\dots,2^j-1} \beta_{j,k}\psi_{j,k}\|_\infty \leq c(\psi)/(1+\sqrt{2})2^{j/2}|\beta|_\infty$. Assertion (2.2) immediately follows since $\sum_{j=0}^{l-1} 2^{j/2} \leq (1+\sqrt{2})2^{l/2}$.

A second important property of the wavelet basis is that $(\phi_\lambda)_\lambda$ is also an orthonormal family for the scalar product $< \cdot, \cdot >_n$ when the equispaced design is considered. It is also the case for any design where $x_i \in [(i-1)/n, i/n]$ when working with Haar basis since the support of the basis functions are not overlapping.

3 Wavelet model selection

3.1 Wavelet models

Among the three following collections of models, we concentrate over the two first one. Let $L_n \in \{0, ..., J\}$ and set $\Lambda_n^* = \Lambda_{L_n}$.

1. We want to select amoung the estimators whose all coefficients until a given level $l-1$ of details (i.e. estimate the projection over V_l) are kept, that is :

$$\mathcal{M}(L_n) \quad = \quad \{\Lambda_l, 0 \leq l \leq L_n\}, \tag{3.1}$$

   and in this case $m_l = \Lambda_l$. Here, the dimension of the model S_{m_l} is given by $D_{m_l} = 2^l$. With a least squared criterium, this choice should be compared to adaptive linear procedure.

2. We consider the estimators where all coefficients are kept up to a given level $(l-2)$ of details and only some of them at level $l-1$ (i.e. estimate the projection over V_{l-1} and some directions of W_{l-1}):

$$\begin{aligned}\mathcal{M}(L_n) = \{\Lambda_0\} \quad \cup \quad & \{m_{(l,\mathcal{I}_l)} = \{\Lambda_{l-1} \cup \{(l-1,k), k \in \mathcal{I}_l \\ & \mid \quad \mathcal{I}_l \subset \{0, ..., 2^{l-1}-1\} \ \text{ and } \ \mathcal{I}_l \neq \emptyset\}, 1 \leq l \leq L_n\},\end{aligned}$$

   and in this case $S_{m_{(l,\mathcal{I}_l)}} = V_{l-1} \oplus W_{l-1}^{\mathcal{I}_l}$ where $W_{l-1}^{\mathcal{I}_l} \subset W_{l-1}$. Here the dimension of $S_{m_{(l,\mathcal{I}_l)}}$ is $D_{m_{(l,\mathcal{I}_l)}} = 2^{l-1} + |\mathcal{I}_l|$ where $1 \leq |\mathcal{I}_l| \leq 2^{l-1}$. For any given l and $1 \leq d \leq 2^{l-1}$ there are $\begin{pmatrix} 2^{l-1} \\ d \end{pmatrix}$ models with dimension $2^{l-1}+d$. With this choice, our procedure should be compared, to usual procedures based on hard thresholding.

3. We could also define models built on the coefficients complete binary tree. In such a case, a model would be a sub-tree containing the root (corresponding to the V_0 space). This should be compared to soft threshold procedures.

Property 1. *For any $m \in \mathcal{M}(L_n)$, there exists some constant b^{loc} such that for any $a \in \mathbb{R}^{|m|}$*

$$\|\sum_{\lambda \in m} a_\lambda \phi_\lambda\|_\infty \leq b^{loc} D_m^{l/2} |a|_\infty.$$

For the first collection it is an immediate application of (2.2) with $b^{loc} = c(\psi)$, whereas for the second one we take $b^{loc} = \sqrt{2}c(\psi)$, since $2^{l-1} \leq D_{m_l,\mathcal{I}_l} \leq 2^l$.

3.2 Oracle inequality

Assumption 1. *The family $(\phi_\lambda)_{\lambda\in\Lambda}$ is orthonormal for the scalar product $< \cdot, \cdot >_n$.*

We have already noticed that this is fulfilled for wavelet basis and for the equispaced design.

Next, for technical reasons, we will need to bound the dimension of the largest model in the considered collection $\mathcal{M}(L_n)$.

Assumption 2. *Suppose that the maximal dimension* 2^{L_n} *is bounded by* $n^{1-\theta}$, *where* $1/2 < \theta < 1$.

This constraint imposes to visit the models only up to the level $L_n < J/2 = \ln n/(2\ln 2)$. Nevertheless, this condition being purely technical, in practice, we will visit all the models up to the level $J = \ln n/\ln 2$.

Assumption 3. *For any function* f *such that* $|f|_\infty < \infty$ *and such that for any model* $m \in \mathcal{M}(L_n)$ *we have* $|\bar{f}_m|_\infty \leq \bar{B}$ *and* $|f|_\infty \leq |\bar{f}_{\Lambda_n^*}|_\infty$.

Note that the condition $|f|_\infty \leq |\bar{f}_{\Lambda_n^*}|_\infty$ is fullfilled as soon as f is supposed to belong to $S_{\Lambda_n^*}$. Moreover, the first part of the assumption is satisfied for any function f when considering the Haar basis.

Before anouncing the main result we first give an upper bound for the Kullback-Leibler risk on a given model.

Proposition 3.1. *Suppose Assumptions 1 and 2 satisfied and let* $\tau \in]0,1[$ *be some constant. For any* n, *any function* f *satisfying Assumption 3, there exists some event* Ω_n *such that*

$$\mathbb{P}\left(\Omega_n^C\right) \leq \frac{c(|f|_\infty, \bar{B}, b^{loc}, \tau)}{n^2},$$

and for any model $m \in \mathcal{M}(L_n)$,

$$\mathbb{E}(K(f, \hat{f}_m)\, \mathbb{1}_{\Omega_n}) \leq K(f, \bar{f}_m) + 2e^{\tau/2+\bar{B}+|f|_\infty}\frac{D_m}{n}.$$

Next, we propose some penalty function which enables to select some model $\hat{m}$ which behaves as well as the ideal but unknown model m^*.

Theorem 3.1. *Let Assumptions 1 and 2 be satisfied,* α *be some positive constant and* $\tau \in]0,1[$. *Let* $\{\mathcal{L}_m\}_{m\in\mathcal{M}(L_n)}$ *be positive numbers such that*

$$\sum_{m\in\mathcal{M}(L_n)} e^{-\mathcal{L}_m D_m} \quad \leq \quad \Sigma < +\infty. \tag{3.2}$$

Define the penalty function as:

$$\text{pen}(m) = e^{|\hat{f}_m|_\infty + |\hat{f}_{\Lambda_n^*}|_\infty + \tau}\left(\frac{c_1}{2} + c_2\mathcal{L}_m\right)\frac{D_m}{n},$$

where $c_1 = (1+\alpha)^4$ *and* $c_2 = (1+\alpha)^4(1+6/\alpha)$. *For any* f *satisfying Assumption 3, there exists some set* Ω_n *such that*

$$\mathbb{P}\left(\Omega_n^C\right) \leq \frac{c(|f|_\infty, \bar{B}, b^{loc}, \alpha, \tau)}{n^2},$$

and such that for any model $m \in \mathcal{M}(L_n)$, *we have:*

$$\mathbb{E}(K(f, \hat{f}_{\hat{m}})\, \mathbb{1}_{\Omega_n}) \leq \frac{(1+\alpha)^2}{\alpha}\inf_{m\in\mathcal{M}(L_n)}\left(K(f, \bar{f}_m) + 2\,\mathbb{E}(\text{pen}(m)\, \mathbb{1}_{\Omega_n})\right) + \frac{3C(|f|_\infty, \bar{B}, \alpha, \tau)\Sigma}{n}.$$

The previous risk inequality can be seen as an oracle inequality: indeed the penalty term can be bounded by:

$$\mathbb{E}(\text{pen}(m)\, \mathbb{1}_{\Omega_n}) \leq e^{2(\tau+\bar{B})}\left(\frac{c_1}{2} + c_2\mathcal{L}_m\right)\frac{D_m}{n}.$$

3.3 Choice of the weights $\{\mathcal{L}_m, m \in \mathcal{M}(L_n)\}$

The choice of these weights is done in order to check the constraint (3.2), hence it depends on the complexity of the model family. Let us consider the following two cases:

1. **Family with a polynomial number of models per dimension**

Assumption 4. *There exist some integer r and some constant R such that the number of models with a given dimension D is bounded by RD^r.*

In this case, the weights can be choosen as constants $\mathcal{L}_m = \mathcal{L}$ for all models m since

$$\sum_{m \in \mathcal{M}(L_n)} e^{-\mathcal{L}_m D_m} \leq \sum_{D=1}^{+\infty} \sum_{m, D_m = D} e^{-\mathcal{L} D} \leq \sum_{D=1}^{+\infty} R D^r e^{-\mathcal{L} D} = \Sigma < +\infty.$$

This assumption is fulfilled when using the first collection of models (3.1). Indeed, in this case there is a single model per dimension $D \in \{1, ..., 2^{L_n}\}$ and the previous assumption holds for $r = 0$ and $R = 1$. Then in (3.2) $\Sigma = 1/(\exp \mathcal{L} - 1)$. Herein we recover the usual bound D_m/n up to a constant for the stochastic term in the risk decomposition.

2. **Family with an exponential number of models per dimension**

Assumption 5. *There exist some constants A and a such that the number of models with a given dimension D is bounded by Ae^{aD}.*

In this case, the weights have to be choosen larger than in the previous case in order to satisfy condition (3.2). Nevertheless, we take them as small as possible to avoid a too large risk bound in the oracle inequality. We can choose $L_m = \ln n$ for all models m since

$$\sum_{m \in \mathcal{M}(L_n)} e^{-L_m D_m} \leq \sum_{D=1}^{+\infty} \sum_{m, D_m = D} e^{-D \ln n} \leq \sum_{D=1}^{+\infty} A e^{aD} e^{-D \ln n} = \Sigma < +\infty.$$

This assumption is fulfilled when using the second collection of models (3.2). Indeed, in this case, each dimension $D \in \{2, ..., 2^{L_n}\}$ can be decomposed as $D = 2^{l-1} + d$ with $1 \leq l \leq L_n$ and $1 \leq d \leq 2^{l-1}$ and there are $\binom{2^{l-1}}{d}$ models with dimension D. Furthermore

$$\binom{2^{l-1}}{d} \leq \left(\frac{e 2^{l-1}}{d}\right)^d = e^{d(1+\ln(2^{l-1}/d))} \leq e^{d(1+2^{l-1}/d)} = e^D.$$

Hence, Assumption 5 holds for $a = 1$ and $A = 1$. Moreover, we get easily:

$$\sum_{D=1}^{\infty} e^D e^{-D \ln n} = \frac{e/n}{1 - e/n} \leq \frac{e/3}{1 - e/3},$$

as soon as $n \geq 3$. Then in (3.2), $\Sigma = \frac{e/3}{1-e/3}$. Herein we recover the bound $(D_m \ln n)/n$ up to a constant for the stochastic term in the risk decomposition. This is the usual price to pay for investigating a large collection of models, when the true function lies in a Besov space rather than in a Sobolev one.

4 Lower bounds on Besov spaces

Set $\nu \geq 0, \nu = k + \alpha$ with $k \in N$ and $0 \leq \alpha < 1$. Let us consider the Hölder class $\mathcal{F} = \Sigma(\nu, L)$ of functions f defined over the interval $[0, 1]$ which admit k derivatives and such that the k-th derivative satisfies:

$$|f^{(k)}(x) - f^{(k)}(y)| \leq L|x - y|^\alpha, \quad \forall (x, y) \in [0, 1]^2. \tag{4.1}$$

We also consider the Sobolev Class $H(\nu, L)$ of regularity $\nu \in \mathbb{N}^*$ over the interval $[0, 1]$ of functions whose Sobolev norm (i.e. the L_2-norm of the ν-th derivative of f) is bounded by L. Note that for any integer $\nu \geq 1$ such a class contains the Hölder class $\mathcal{F} = \Sigma(\nu, L)$. Furthermore we denote $\mathcal{C}^\infty(S)$ the space of functions uniformly bounded by S.

In this section we will state that the minimax rate of convergence for the estimation problem with Poisson response is the same as the usual minimax rate of convergence in nonparametric regression estimation. The following lower bound is stated in the case of a deterministic and equispaced design $(x_i)_{1\leq i\leq n}$ in $[0,1]$ and over a Hölder class.

Theorem 4.1. *Set $\nu > 1/2$, there exists a constant C such that*

$$\liminf_{n\to\infty} \inf_{\hat{f}_n\in\mathcal{C}^\infty(S)} \sup_{f\in\Sigma(\nu,L)\cap\mathcal{C}^\infty(S)} \mathbb{E}_f(K(f,\hat{f}_n)v_n^{-2}) \geq Ce^{-3S}L^2 > 0,$$

where $v_n = n^{-\frac{\nu}{2\nu+1}}$ and C is an explicit positive constant.

The lower bound over the Sobolev class $H(\nu,L)$ is a direct consequence of the previous one since this class contains the Hölder one when ν is a nonzero integer.

In the Gaussian regression case, it is now well known, that when the quadratic risk is considered, the linear wavelet estimator reaches the minimax rate of convergence $n^{-2\nu/(\nu+1)}$ over the Sobolev Class $H(\nu,L)$ as soon as the optimal resolution level j^* is choosen such that $2^{j^*} = \mathcal{O}(n^{1/(2\nu+1)})$.

Here when considering the collection (3.1), the selected estimator $\hat{f}_{\hat{m}}$ reaches the rate $n^{1/(2\nu+1)}$ over the Sobolev class $H(\nu,L)$ and hence is minimax. Indeed, since $K(f,\bar{f}_{m_l}) \leq K(f,P_lf)$ due to definition of $\bar{f}_{m_l}$ and since over a Sobolev class $K(f,P_lf)$ is of the same order as $\|f-P_lf\|_2^2 = \mathcal{O}(2^{-2l\nu})$ the bias term $K(f,\bar{f}_{m_l})$ is also of order $\mathcal{O}(2^{-2l\nu})$. Furthermore the dimension D_{m_l} of the model S_{m_l} is 2^l. Hence the trade of between the bias term and the penalization term in Theorem 3.1 is obtained for $2^l = \mathcal{O}(n^{1/(2\nu+1)})$. Moreover, the residual term in the oracle inequality being of order $1/n$ the risk $\mathbb{E}(K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\Omega_n})$ is estimated by $\mathcal{O}(n^{1/(2\nu+1)})$.

We guess that on Besov classes the obtained lower bound for the Kullback-Leibler risk should be the same as the usual one for quadratic risk, that is $\mathcal{O}(n^{1/(2\nu'+1)})$ with $\nu' = \nu - 1/p + 1/2$, $\nu \geq 1/2$ and $p \leq 2$. For this larger class of functions, the richest collection of models (3.2) should be considered, in order to obtain an upper bound for the bias term of order $\mathcal{O}(n^{1/(2\nu'+1)})$. Due to the choice of weights $L_m = \ln n$, the selected estimator can only reach the rate $\mathcal{O}(n^{1/(2\nu'+1)})$ up to a $\ln n$ factor which is the usual price to pay for adaptivity.

5 A simulation study

In this part, we present some results in order to illustrate our results. Our aim is to compare our procedure with the projection procedure proposed by Patricia Reynaud (Reynaud-Bouret 2003). More precisely, we want to answer the following questions:

1. Does the $e^{|\hat{f}_m|_\infty + |\hat{f}_{\Lambda_n^*}|_\infty}$ factor in the penalty make any sense in practice ?

2. How to choose the constants involved in the penalty term in practice ?

3. How much the penalized maximum likelihood estimator is preferable to the penalized projection estimator as defined by (Reynaud-Bouret 2003) ?

5.1 Choice of the penalty functions

In the proof of the Theorem, it can be seen that the term $|\hat{f}_m|_\infty + |\hat{f}_{\Lambda_n^*}|_\infty$ in the penalty term comes from an estimation of $|f|_\infty$. Therefore, in order to see the sensibility of the penalty function to $|f|_\infty$, we choose functions that only differ from their infinity norms.

More precisely, we choose $n = 2^7 = 128$, and we choose the functions f and regular models so that the function f belongs to one of the following models:

a. $f = f_4$ is a regular piecewise constant function on $[0,1]$, $f_4 = \mathbb{1}_{[1/4,1/2]} - \mathbb{1}_{[1/2,3/4]}$, that we try to estimate using the Haar basis. In this case, such models may be described, for $J \geq 0$, by:

$$S_J^H = \{\sum_{j=0}^{J}\sum_{k=0}^{j-1} \beta_{j,k}\,\mathbb{1}_{[k2^{-j},(k+1)2^{-j}[}, \beta \in \mathbb{R}^{2^J-1}\}.$$

b. $f = 2f_4$ and the models are the same as above.

c. $f = f_4/2$ and the models are the same as above.

d. Let g be defined by $g(x) = a(x^2(1-x))^3 - 1$, where a is some positive constant such that $|g|_\infty = 1$. We define the models for $J \geq 0$ by:

$$S_J^\psi = \{\sum_{j=0}^{J}\sum_{k=0}^{j-1}\beta_{j,k}\psi_{j,k}, \beta \in \mathbb{R}^{2^J-1}\},$$

where the $\psi_{j,k}$ are the Symmlet basis with 4 vanishing moments (see (Daubechies 1992) and (Wickerhauser 1994)). The true function is then defined as:

$$f_{smooth} = P_2(g).$$

In these 4 cases, the true function belongs to the model S_2^H which dimension is $2^2 = 4$.

For $L = 100$ simulations, we generate $n = 128 = 2^7$ independant random variables Y_i with Poisson distribution with parameter $e^{f(i/n)}$. For each simulation, we calculate, on each model S_J, the maximum likelihood estimator $\hat{f}_J$ and the projection estimator $\hat{e}_J$, which is simply the L_2-projection of Y onto the model S_J. Since there is only one model with a given dimension, we then select the "best" model with the following penalized criteria:

$$\hat{J}_{ML} = \arg\min_{0\leq J\leq 7}(\gamma_n(\hat{f}_J) + c2^J/n), \quad \hat{J}_P = \arg\min_{0\leq J\leq 7}(n^{-1}\sum_{i=1}^{n}(Y_i - \hat{e}_{J,i})^2 + c2^J/n).$$

The final estimators are the penalized maximum likelihood estimator (PMLE) $\hat{f}_{\hat{J}_{ML}}$ and the penalized projection estimator (PPE) $\hat{e}_{\hat{J}_P}$. Note that, in the Haar basis case, the maximum likelihood estimator and the projection estimator coincide in each model S_J^H ($\hat{e}_J = \exp\hat{f}_J$) whereas this is not the case in the Symmlet case. Nevertheless, the chosen model is not necessarily the same since the selection criteria are not the same.

The constant c in the penalty term is first chosen equal to 0.1 and then grows by steps of 0.1. For lower values of c, the chosen dimension is the maximum one (here 2^7) and for a particular value of c suddenly jumps down to lower dimensions. For each simulation, we detect the lowest constant c selecting the true "model" ($J = 2$). Figure 1 shows the dispersion of these constants over the $L = 100$ simulations.

We can remark that these constants seem more stable with the PMLE than with the PPE. In particular, we see that the distribution of the PMLE constants is of the same order for the four functions whereas it seems to depend on $|f|_\infty$ for the PPE. If we divide the constants by $e^{|f|_\infty}$ in the PPE case, as described in Figure 2, we recover constants of the same order as ones obtained in the PMLE case.

Therefore, we have decided to skip the $e^{|\hat{f}_m|_\infty + |\hat{f}_{\Lambda_n^*}|_\infty}$ factor in the penalty term for the PMLE and to keep it for the PPE. More precisely, in the sequel, we shall take a penalty term of the form $\text{pen}_{ML}(J) = c_{ML}2^J/n$ for the PMLE and $\text{pen}_P(J) = c_P|\hat{e}_J|_\infty 2^J/n$ for the PPE. Furthermore, note that we choose $|\hat{e}_J|_\infty$ rather than $|\hat{e}_\Lambda|_\infty = |\hat{e}_7|_\infty$, as our Theorem would suggest, since we suspect that this latter would over-estimate $\exp(|f|_\infty)$.

5.2 Choice of the constant in the penalty functions

Next, we consider the constants $c_{KL,p}, c_{MC,p}, p = 0.75, 0.80, 0.85, 0.90, 0.95, 0.99, 1$ corresponding to the $0.75, 0.80, 0.85, 0.90, 0.95, 0.99, 1$ quantiles of the former constants for each procedure. We still choose the functions f so that they belong to one of the models:

a. $f = f_{16}$ is a regular piecewise constant function on $[0,1]$, equal to 1 on intervals $[1/16, 2/16[, [5/16, 6/16[, [9/16, 10/16[, [13/16, 14/16[$ and to -1 on intervals $[2/16, 3/16[, [6/16, 7/16[, [10/16, 11/16[, [14/16, 15/16[$ and 0 elsewhere. The true dimension is then $2^4 = 16$. We try to estimate f_{16} via the Haar basis on the models $S_J^H, 0 \leq J \leq 7$.

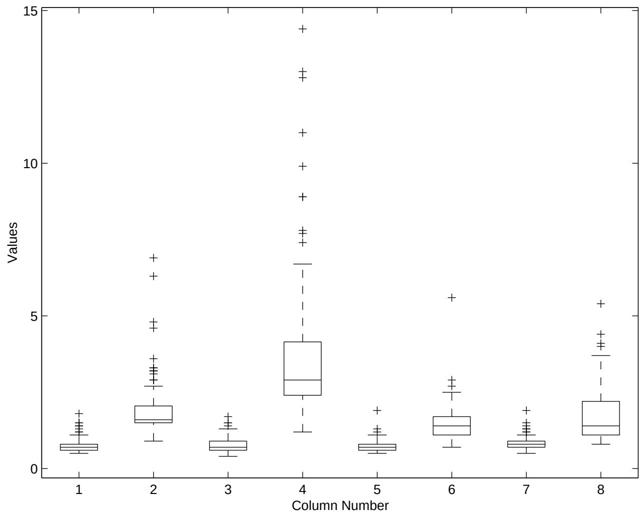

Figure 1: Distribution of the lowest constants selecting the "true" model: (1-2) $f = f_4$, (3-4) $f = 2f_4$, (5-6) $f = f_4/2$, (7-8) $f = f_{smooth}$ via (1,3,5,7) penalized maximum likelihood criterium, (2,4,6,8) penalized projection criterium.

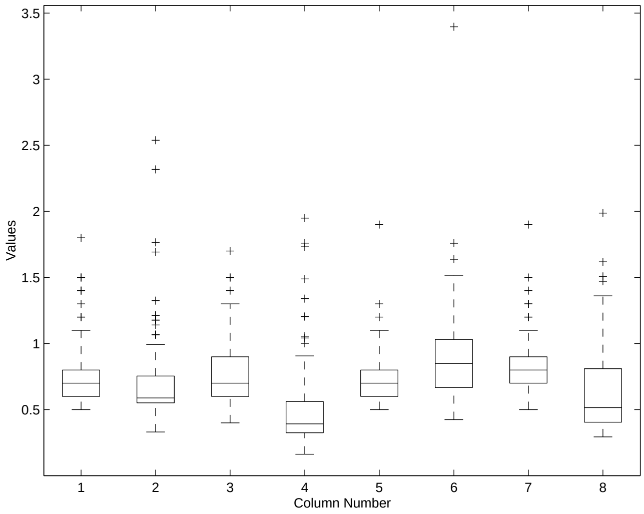

Figure 2: Distribution of the lowest constants selecting the "true" model: (1-2) $f = f_4$, (3-4) $f = 2f_4$, (5-6) $f = f_4/2$, (7-8) $f = f_{smooth}$ via (1,3,5,7) penalized maximum likelihood criterium, (2,4,6,8) penalized projection criterium, (2,4,6,8) contants are divided by $\exp(|f|_\infty)$.

b. $f = f_{smooth}$ like described in case d, the models are the $S_J^{\psi}, 0 \leq J \leq 7$, so that the true dimension still is $4 = 2^2$.

We perform $L = 100$ new simulations of $n = 128$ random variables Y_i and for each simulation, we calculate the penalized maximum likelihood estimator and penalized projection estimator, calculated with the previous seven constants. We present in Table 1 the distribution of the selected dimensions over the 100 simulations. We also present in Figures 3 and 4 the distribution of the Average Square Error and in Figures 5 and 6 the Kullback-Leibler divergence of both estimators over the $L = 100$ simulations and for each of the seven constants.

(a)

J	$\hat{J}_{ML}$							$\hat{J}_P$						
p	0.75	0.80	0.85	0.90	0.95	0.99	1	0.75	0.80	0.85	0.90	0.95	0.99	1
c_p	0.9	0.9	1.0	1.1	1.2	1.6	1.9	0.85	0.96	1.04	1.18	1.4	1.97	3.4
0	0	0	0	0	0	1	5	0	0	0	0	0	7	59
1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	1	2	6	10	1	2	2	5	21	66	39
4	94	94	96	98	98	93	85	99	98	98	95	79	27	0
5	6	6	3	1	0	0	0	0	0	0	0	0	0	0
2	72	72	79	84	89	96	98	66	75	82	87	94	100	100
3	21	21	18	15	10	4	2	34	25	18	13	6	0	0
4	4	4	1	1	1	0	0	0	0	0	0	0	0	0
5	1	1	1	0	0	0	0	0	0	0	0	0	0	0
6	2	2	1	0	0	0	0	0	0	0	0	0	0	0

(b)

J	$\hat{J}_{ML}$							$\hat{J}_P$						
p	0.75	0.80	0.85	0.90	0.95	0.99	1	0.75	0.80	0.85	0.90	0.95	0.99	1
c_p	0.9	0.9	1.0	1.1	1.2	1.6	1.9	0.85	0.96	1.04	1.18	1.4	1.97	3.4
0	0	0	0	0	0	1	5	0	0	0	0	0	7	59
1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	1	2	6	10	1	2	2	5	21	66	39
4	94	94	96	98	98	93	85	99	98	98	95	79	27	0
5	6	6	3	1	0	0	0	0	0	0	0	0	0	0
2	72	72	79	84	89	96	98	66	75	82	87	94	100	100
3	21	21	18	15	10	4	2	34	25	18	13	6	0	0
4	4	4	1	1	1	0	0	0	0	0	0	0	0	0
5	1	1	1	0	0	0	0	0	0	0	0	0	0	0
6	2	2	1	0	0	0	0	0	0	0	0	0	0	0

Table 1: Distribution of the selected dimensions over the 100 simulations: (a) $f = f_{16}$, Haar basis, (b) $f = f_{smooth}$, Symmlet basis.

From these results, it seems reasonable to keep, among the seven quantiles, for each procedure the 0.95 quantile, namely $c_{ML} = 1.2$ and $c_P = 1.4$ in the penalty term for the next simulations.

5.3 Comparison with the penalized projection estimator

In this part, we compare our penalized maximum likelihood procedure with the penalized projection estimator for different values of n and for two criteria, namely Average Square Error and Kullback-Leibler divergence. For that purpose, we choose

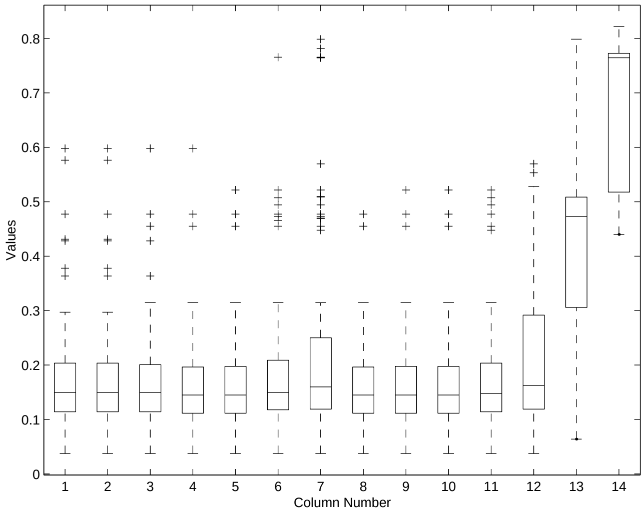

Figure 3: Distribution of the Average Square Error for $f = f_{16}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (6) $c_{ML,0.99} = 1.6$, (7) $c_{ML,1} = 1.9$, (8) $c_{P,0.75} = 0.85$, (9) $c_{P,0.80} = 0.96$, (10) $c_{P,0.85} = 1.04$, (11) $c_{P,0.90} = 1.18$, (12) $c_{P,0.95} = 1.4$, (13) $c_{P,0.99} = 1.97$, (14) $c_{P,1} = 3.4$.

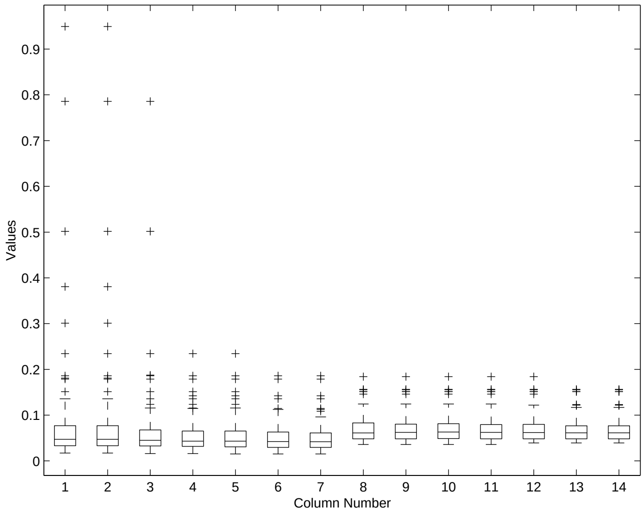

Figure 4: Distribution of the Average Square Error for $f = f_{smooth}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (6) $c_{ML,0.99} = 1.6$, (7) $c_{ML,1} = 1.9$, (8) $c_{P,0.75} = 0.85$, (9) $c_{P,0.80} = 0.96$, (10) $c_{P,0.85} = 1.04$, (11) $c_{P,0.90} = 1.18$, (12) $c_{P,0.95} = 1.4$, (13) $c_{P,0.99} = 1.97$, (14) $c_{P,1} = 3.4$.

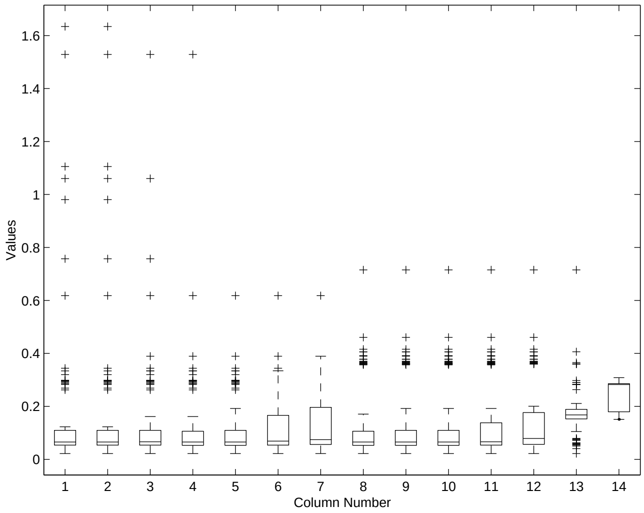

Figure 5: Distribution of the Kullback-Leibler divergence for $f = f_{16}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (6) $c_{ML,0.99} = 1.6$, (7) $c_{ML,1} = 1.9$, (8) $c_{P,0.75} = 0.85$, (9) $c_{P,0.80} = 0.96$, (10) $c_{P,0.85} = 1.04$, (11) $c_{P,0.90} = 1.18$, (12) $c_{P,0.95} = 1.4$, (13) $c_{P,0.99} = 1.97$, (14) $c_{P,1} = 3.4$.

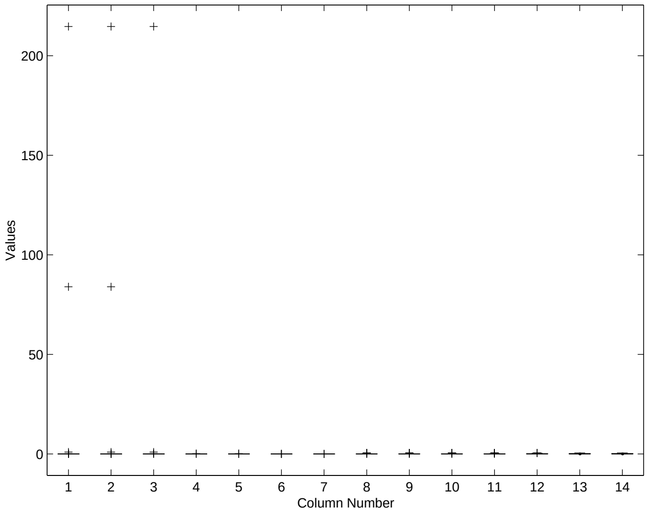

Figure 6: Distribution of the Kullback-Leibler divergence for $f = f_{smooth}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (6) $c_{ML,0.99} = 1.6$, (7) $c_{ML,1} = 1.9$, (8) $c_{P,0.75} = 0.85$, (9) $c_{P,0.80} = 0.96$, (10) $c_{P,0.85} = 1.04$, (11) $c_{P,0.90} = 1.18$, (12) $c_{P,0.95} = 1.4$, (13) $c_{P,0.99} = 1.97$, (14) $c_{P,1} = 3.4$.

a. $f = f_4$, the true dimension is then $2^2 = 4$. We estimate f_4 via the Haar basis on the models S_J^H (case a).

b. $f = g$ like described in case d, the models are the S_J^ψ. Hence, the true function belongs to none of the models.

We perform $L = 100$ new simulations of $n = 128 = 2^7, n = 256 = 2^8, n = 512 = 2^9$ random variables Y_i and the collection of models are defined by

$$\mathcal{M}_{128} = \{S_J, 0 \leq J \leq 7\}, \mathcal{M}_{256} = \{S_J, 0 \leq J \leq 8\}, \mathcal{M}_{512} = \{S_J, 0 \leq J \leq 9\}.$$

For each simulation, we calculate the penalized maximum likelihood estimator and the penalized projection estimator, computed with the constants determined in the previous section. We describe in Table 2 the distributions of the selected dimensions over the 100 simulations and in Table 3 the number of simulations for which the maximum likelihood procedure selects a lower, resp. equal, resp. higher dimension than the projection procedure. We also present in Figure 7 the distributions of the Average Square Error and in Figures 8 and 9 the Kullback-Leibler divergence of both estimators over the $L = 100$ simulations.

	J	$n = 128$		$n = 256$		$n = 512$	
		$\hat{J}_{ML}$	$\hat{J}_P$	$\hat{J}_{ML}$	$\hat{J}_P$	$\hat{J}_{ML}$	$\hat{J}_P$
(a)	2	96	100	97	100	93	98
	3	4	0	3	0	7	2
(b)	2	80	92	78	60	39	6
	3	18	8	21	40	60	94
	4	2	0	1	0	1	0

Table 2: Distribution of the selected dimensions over the 100 simulations for different sample sizes ($n = 128, 256, 512$): (a) $f = f_4$, Haar basis, (b) f polynomial, Symmlet basis.

	n	$\hat{J}_{ML} < \hat{J}_P$	$\hat{J}_{ML} = \hat{J}_P$	$\hat{J}_{ML} > \hat{J}_P$
(a)	128	0	96	4
	256	0	97	3
	512	0	95	5
(b)	128	1	86	13
	256	19	79	2
	512	33	66	1

Table 3: Comparison of the selected dimensions by the penalized Maximum Likelihood criterium and by the Projection criterium over the 100 simulations for different sample sizes ($n = 128, 256, 512$): (a) $f = f_4$, Haar basis, (b) f polynomial, Symmlet basis.

5.4 Conclusion

From a statistical point of view, this simulation study suggests that the penalized maximum likelihood estimator behaves better than the projection estimator. Indeed, for the first one, an estimation of $|f|_\infty$ is not required in the procedure although it is for the second one. Secondly, even if both procedures are equivalent when estimating a piecewise constant function (f_4), the penalized maximum likelihood estimator performs better than the penalized projection estimator when estimating a smooth function (here, a polynomial) and this, with both loss functions, Average Square Error and Kullback-Leibler divergence.

Nevertheless, the computing cost is much heavier in the maximum likelihood case than in the projection case, since the latter provides an explicit estimator whereas the first one requires the

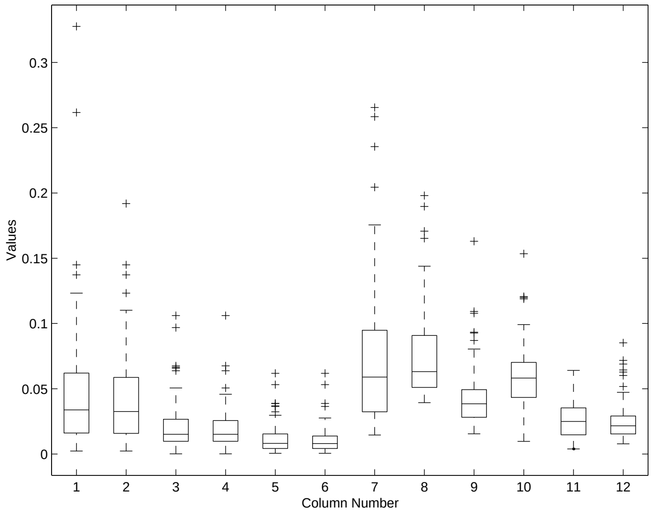

Figure 7: Distribution of the Average Square Error for (1-6) $f = f_4$ and (7-12) $f = f_{smooth}$ for different sample size: (1,2,7,8) $n = 128$, (3,4,9,10) $n = 256$, (5,6,11,12) $n = 512$, (1,3,5,7,9,11) PMLE, (2,4,6,8,10,12) PPE.

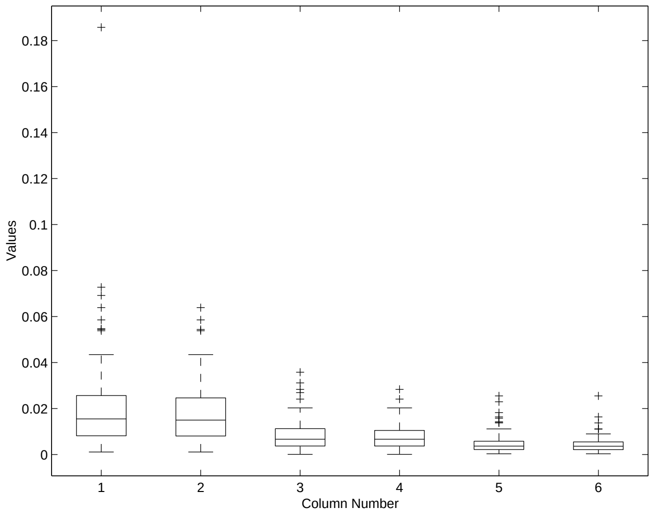

Figure 8: Distribution of the Kullback-Leibler divergence for $f = f_4$ for different sample size: (1,2) $n = 128$, (3,4) $n = 256$, (5,6) $n = 512$, (1,3,5) PMLE, (2,4,6) PPE.

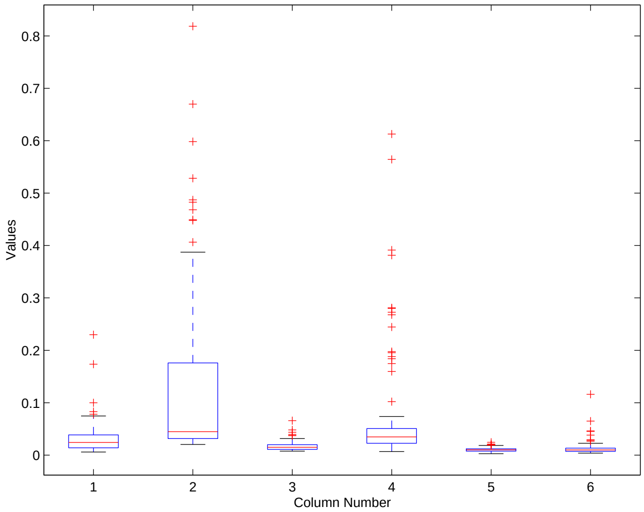

Figure 9: Distribution of the Kullback-Leibler divergence for $f = f_{smooth}$ for different sample size: (1,2) $n = 128$, (3,4) $n = 256$, (5,6) $n = 512$, (1,3,5) PMLE, (2,4,6) PPE.

minimization of a function, except in the particular case of the Haar basis: in this case indeed, just compute the estimator on each model by projection, and then select the best one using our penalized maximum likelihood criterium.

The constants 1.2 and 1.4 are calibrated for piecewise constants and smooth functions. For other kinds of functions (for instance, functions with bumps or angles), our constant calibration method should be applied with an adapted wavelet basis. Thus, the constants may change.

6 Proofs

6.1 Proof of the oracle inequality given in Theorem 3.1

We aim at proving that $\hat{f}_{\hat{m}}$ is a better estimator than $\hat{f}_m$ in the sense of the Kullback-Leibler risk.

By definition (1.3) of $\hat{m}$ and (1.1) of $\hat{f}_m$, we have:

$$\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(\hat{f}_m) + \text{pen}(m) \leq \gamma_n(\bar{f}_m) + \text{pen}(m). \tag{6.1}$$

Furthermore, with the notation $\varepsilon_i = Y_i - \mathbb{E}(Y_i) = Y_i - e^{f_i}$.

$$K(f, \hat{f}_{\hat{m}}) = K(f, \bar{f}_m) + \gamma_n(\hat{f}_{\hat{m}}) - \gamma_n(\bar{f}_m) + < \hat{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n .$$

Using (6.1), we get:

$$K(f, \hat{f}_{\hat{m}}) \leq K(f, \bar{f}_m) + \text{pen}(m) - \text{pen}(\hat{m}) + < \hat{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n .$$

The latter term can be splitted in two parts:

$$< \hat{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n = < \hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}, \varepsilon >_n + < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n .$$

Furthermore, using that for any numbers, a, b and any positive θ

$$2ab \leq \theta a^2 + \frac{1}{\theta} b^2, \tag{6.2}$$

we can write

$$\begin{aligned} < \hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}, \varepsilon >_n & \leq \sup_{h \in S_{\hat{m}}} \frac{< h, \varepsilon >_n}{|h|_n} |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_n \\ & \leq \frac{\theta_1}{2} \chi_n^2(\hat{m}) + \frac{1}{2\theta_1} |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_n^2, \end{aligned}$$

where

$$\chi_n(m') = \sup_{h \in S_{m'}} \frac{< h, \varepsilon >_n}{|h|_n} = \sup_{h \in S_{m'}, |h|_n \leq 1} < h, \varepsilon >_n .$$

and where θ_1 is a positive number to be defined later. Next we link $|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_n^2$ with the Kullback-Leibler divergence $K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}})$ thanks to (7.5), and we have:

$$K(f, \hat{f}_{\hat{m}}) \leq K(f, \bar{f}_m) + \text{pen}(m) - \text{pen}(\hat{m}) + \frac{\theta_1}{2} \chi_n^2(\hat{m}) + \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) + < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n . \tag{6.3}$$

In order to control the term $\chi_n^2(\hat{m})$, we need to introduce some set $\Omega_n[A]$, for some positive constant A, which will become the set Ω_n of the theorem for a certain value of A. This set will also allow the control of $|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty$.

Let A be some positive number and ρ such that

$$1 - \theta < 2\rho < \theta, \tag{6.4}$$

where θ is defined in Assumption 2. We define

$$\Omega_n[A] = \left\{ \sup_{\lambda \in \Lambda_n^*} | < \varphi_\lambda, \varepsilon >_n | \leq \frac{A n^{-\rho}}{b^{loc} |\Lambda_n^*|^{1/2}} \right\}.$$

The next proposition, which is the key point of this proof, states the control of the term $\chi_n^2(\hat{m})$.

Proposition 6.1. *Let $(x_{m'})_{m' \in \mathcal{M}(L_n)}$ be some positive numbers and suppose that $A \leq \frac{12\alpha e^{-|f|_\infty}}{\kappa(\alpha)}$, where $\kappa(\alpha)$ is defined in (6.17). Then, there exists some set Ω_n^1 such that $\mathbb{P}\left({\Omega_n^1}^C\right) \leq \sum_{m' \in \mathcal{M}(L_n)} e^{-x_{m'}}$, and on the set Ω_n^1*

$$\chi_n(\hat{m}) \, \mathbb{1}_{\Omega_n[A]} \leq (1+\alpha) e^{|f|_\infty/2} \left((\frac{D_{\hat{m}}}{n})^{1/2} + (\frac{12 x_{\hat{m}}}{n})^{1/2} \right). \tag{6.5}$$

The proof of this proposition, which can be found in Section 6.3.2, relies on P. Reynaud's concentration inequality.

In the next proposition, we control the latter term $< \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n$.

Proposition 6.2. *Let $(y_{m'})_{m' \in \mathcal{M}(L_n)}$ be some positive numbers and θ_2 and θ_3 be some positive constants. Then, there exists some set Ω_n^2 such that $\mathbb{P}\left({\Omega_n^2}^C\right) \leq 2 \sum_{m' \in \mathcal{M}(L_n)} e^{-y_{m'}}$, and on the set Ω_n^2*

$$\begin{aligned} < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n & \leq \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} (1 + \frac{1}{\theta_2}) \frac{y_{\hat{m}}}{n} + \theta_2 K(f, \bar{f}_{\hat{m}}) \\ & + \frac{e^{|\bar{f}_m - f|_\infty}}{2} (1 + \frac{1}{\theta_3}) \frac{y_m}{n} + \theta_3 K(f, \bar{f}_m). \end{aligned} \tag{6.6}$$

The proof of this proposition which relies on a Bernstein type inequality can be found in section 6.3.3.

Gathering (6.3), (6.5) and (6.6), we obtain that, on the set $\Omega_n^1 \cap \Omega_n^2$,

$$\begin{aligned} K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\Omega_n[A]} \leq\ & \mathbb{1}_{\Omega_n[A]} \Big\{ K(f,\bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \\ & + \frac{\theta_1}{2}(1+\alpha)^2 e^{|f|_\infty} \left(\left(\frac{D_{\hat{m}}}{n}\right)^{1/2} + (\frac{12x_{\hat{m}}}{n})^{1/2} \right)^2 + \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}},\hat{f}_{\hat{m}}) \\ & + (1+\frac{1}{\theta_2}) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty} y_{\hat{m}}}{2n} + \theta_2 K(f,\bar{f}_{\hat{m}}) + (1+\frac{1}{\theta_3}) \frac{e^{|\bar{f}_m - f|_\infty} y_m}{2n} + \theta_3 K(f,\bar{f}_m) \Big\} \end{aligned} \tag{6.7}$$

Let us now choose

$$x_{m'} = y_{m'} = \mathcal{L}_{m'} D_{m'} + \zeta.$$

When using the following consequence of (6.2),

$$(a+b)^2 \leq (1+\theta)a^2 + (1+\frac{1}{\theta})b^2,$$

we get

$$\begin{aligned} \left(\left(\frac{D_{\hat{m}}}{n}\right)^{1/2} + (\frac{12x_{\hat{m}}}{n})^{1/2} \right)^2 \leq\ & (1+\theta_4)\frac{D_{\hat{m}}}{n} + (1+\frac{1}{\theta_4}) \frac{12(\mathcal{L}_{\hat{m}} D_{\hat{m}} + \zeta)}{n} \\ & = (1+\theta_4)(1+\frac{12\mathcal{L}_{\hat{m}}}{\theta_4})\frac{D_{\hat{m}}}{n} + (1+\frac{1}{\theta_4})\frac{12\zeta}{n}, \end{aligned} \tag{6.8}$$

for some positive θ_4 to be choosen later. Hence, when substituting (6.8) in inequality (6.7) and factorizing the terms $K(f,\bar{f}_m)$, $\frac{D_{\hat{m}}}{n}$ and $\frac{\zeta}{n}$ we obtain:

$$\begin{aligned} K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\Omega_n[A]} \leq \mathbb{1}_{\Omega_n[A]} \Bigg[& (1+\theta_3)K(f,\bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (1+\frac{1}{\theta_3}) \frac{e^{|\bar{f}_m - f|_\infty} \mathcal{L}_m D_m}{2n} \\ & + \left\{ \frac{\theta_1}{2}(1+\alpha)^2 e^{|f|_\infty} (1+\theta_4)(1+\frac{12\mathcal{L}_{\hat{m}}}{\theta_4}) + (1+\frac{1}{\theta_2}) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} \mathcal{L}_{\hat{m}} \right\} \frac{D_{\hat{m}}}{n} \\ & + \left\{ 6\theta_1(1+\alpha)^2(1+\frac{1}{\theta_4})e^{|f|_\infty} + (1+\frac{1}{\theta_2}) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} \right\} \frac{\zeta}{n} + \\ & + \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}},\hat{f}_{\hat{m}}) + \theta_2 K(f,\bar{f}_{\hat{m}}) \Bigg]. \end{aligned}$$

Now we choose $0 < \theta_2 < 1$ and $\theta_1 = \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_2}$. Since $K(\bar{f}_{\hat{m}},\hat{f}_{\hat{m}}) + K(f,\bar{f}_{\hat{m}}) = K(f,\hat{f}_{\hat{m}})$, we get

$$\frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}},\hat{f}_{\hat{m}}) + \theta_2 K(f,\bar{f}_{\hat{m}}) = \theta_2 K(f,\hat{f}_{\hat{m}}).$$

Substituting this expression in the previous inequality and noticing that $(1+1/\theta_2) \leq 2/\theta_2$, we have:

$$\begin{aligned} (1-\theta_2)K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\Omega_n[A]} \leq \mathbb{1}_{\Omega_n[A]} \Big\{ & (1+\theta_3)K(f,\bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \\ & + (1+\frac{1}{\theta_3}) \frac{e^{|\bar{f}_m - f|_\infty} \mathcal{L}_m D_m}{2n} + T_1(\hat{m})\frac{D_{\hat{m}}}{n} + T_2(\hat{m})\frac{\zeta}{n} \Big\}, \end{aligned} \tag{6.9}$$

where

$$\begin{aligned} T_1(\hat{m}) &= \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{2\theta_2}(1+\alpha)^2 e^{|f|_\infty}(1+\theta_4)(1+\frac{12\mathcal{L}_{\hat{m}}}{\theta_4}) + \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{\theta_2}\mathcal{L}_{\hat{m}} \\ T_2(\hat{m}) &= \left\{ 12\frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_2}(1+\alpha)^2(1+\frac{1}{\theta_4})e^{|f|_\infty} + \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{\theta_2} \right\} \end{aligned}$$

Now, we need on the one hand to bound the quantities $|f|_\infty$ and $|\bar{f}_{\hat{m}} - f|_\infty$ in $T_1(\hat{m})$ and to choose the constant θ_2, θ_4 in such a way that

$$(-\operatorname{pen}(\hat{m}) + T_1(\hat{m})\frac{D_{\hat{m}}}{n})\, \mathbb{1}_{\Omega_n[A]} \quad \leq \quad 0, \tag{6.10}$$

and on the other hand, to bound $T_2(\hat{m})$ by a deterministic constant. To this end, the following proposition enables us to bound the terms of the type $|\bar{f}_{m'} - f|_\infty$ for any $m' \in \mathcal{M}(L_n)$.

Proposition 6.3. *Suppose Assumptions 1 and 3 satisfied. Set $\tau \in]0,1[$. If*

$$A \leq \frac{\tau}{4e^{1+\bar{B}}}, \tag{6.11}$$

then, on the set $\Omega_n[A]$, for any model $m' \in \mathcal{M}(L_n)$ we have :

$$|\hat{f}_{m'} - \bar{f}_{m'}|_\infty \quad \leq \quad \tau/2$$

Since the results in Proposition 6.3 are given for any $m' \in \mathcal{M}(L_n)$, they obviously hold true for the particular $m' = \hat{m}$.

In the sequel of the proof, we take $A = \inf(\frac{12\alpha e^{-|f|_\infty}}{\kappa(\alpha)}, \frac{\tau}{4e^{1+\bar{B}}})$ and we put $\Omega_n = \Omega_n[A]$ for this choice of A.

On Ω_n we can bound $T_1(\hat{m})$ using that:

$$\begin{aligned} (|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty)\, \mathbb{1}_{\Omega_n} &\leq (|\hat{f}_{\hat{m}}|_\infty + 2|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty)\, \mathbb{1}_{\Omega_n} \leq (|\hat{f}_{\hat{m}}|_\infty + \tau)\, \mathbb{1}_{\Omega_n} \\ |\bar{f}_{\hat{m}} - f|_\infty\, \mathbb{1}_{\Omega_n} &\leq (|f|_\infty + |\bar{f}_{\hat{m}}|_\infty)\, \mathbb{1}_{\Omega_n} \leq (|\bar{f}_{\Lambda_n^*}|_\infty + |\hat{f}_{\hat{m}}|_\infty + \tau/2)\, \mathbb{1}_{\Omega_n} \\ &\leq (|\hat{f}_{\Lambda_n^*}|_\infty + |\hat{f}_{\hat{m}}|_\infty + \tau)\, \mathbb{1}_{\Omega_n}\,. \end{aligned} \tag{6.12}$$

On Ω_n we also bound $T_2(\hat{m})$ using that $|\bar{f}_{\hat{m}}|_\infty \, \mathbb{1}_{\Omega_n} \leq \bar{B}\, \mathbb{1}_{\Omega_n}$, $|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty \, \mathbb{1}_{\Omega_n[A]} \leq \tau/2\, \mathbb{1}_{\Omega_n}$ and furthermore due to Assumption 3 we also have

$$|\bar{f}_{\hat{m}} - f|_\infty \quad \leq \quad |\bar{f}_{\hat{m}}|_\infty + |f|_\infty \leq \bar{B} + |\bar{f}_{\Lambda_n^*}|_\infty \leq 2\bar{B}. \tag{6.13}$$

Now, we choose $\theta_2 = 1/(1+\alpha)$, $\theta_4 = \alpha$, and α such that

$$\begin{aligned} (1+\alpha)^4 &= c_1 \\ (1+\alpha)^4(\frac{6}{\alpha} + 1) &= c_2, \end{aligned}$$

where c_1 and c_2 are the constants in the penalty term. With these choices of θ_2 and θ_4, substituting the bounds given in (6.12) and in (6.13) in expressions of $T_1(\hat{m})$ and of $T_2(\hat{m})$, we check (6.10) and we bound $T_2(\hat{m})$ with some constant $C'(\bar{B}, \alpha)$.

Hence inequality (6.9) over $\Omega_n^1 \cap \Omega_n^2$ gives :

$$\frac{\alpha}{1+\alpha}K(f, \hat{f}_{\hat{m}})\, \mathbb{1}_{\Omega_n} \leq \mathbb{1}_{\Omega_n} \left\{ (1+\theta_3)K(f, \bar{f}_m) + \operatorname{pen}(m) + (1+\frac{1}{\theta_3})\frac{e^{|\bar{f}_m - f|_\infty}\mathcal{L}_m D_m}{2n} + C'(\bar{B}, \alpha)\frac{\zeta}{n} \right\}.$$

Next, when choosing $\theta_3 = \alpha$ we bound the third term of the previous right hand side as follows

$$\mathbb{1}_{\Omega_n}(1+\frac{1}{\theta_3})\frac{e^{|\bar{f}_m - f|_\infty}\mathcal{L}_m D_m}{2n} \leq \mathbb{1}_{\Omega_n}(1+\frac{1}{\alpha})\frac{e^{|\hat{f}_m|+|\hat{f}_{\Lambda_n^*}|_\infty+1}\mathcal{L}_m D_m}{2n} \leq \mathrm{pen}(m)\,\mathbb{1}_{\Omega_n}\,.$$

Moreover, since Ω_n^1 and Ω_n^2 satisfy

$$\mathbb{P}\left({\Omega_n^1}^C \cup {\Omega_n^2}^C\right) \leq 3\sum_{m\in\mathcal{M}_n} e^{-\mathcal{L}_m D_m - \zeta} \leq 3\Sigma e^{-\zeta},$$

when applying lemma 7.6 with

$$\begin{aligned}\kappa_1 &= \frac{1+\alpha}{\alpha}\frac{C'(\bar{B},\alpha)}{n} = \frac{C(\bar{B},\alpha)}{n}\\ \kappa_2 &= 3\Sigma\end{aligned}$$

we get the oracle inequality.

It remains to prove that the set Ω_n has a great probability, which is given in the following proposition.

Proposition 6.4. *Under Assumptions 1, 2 and 3, for any A there exists some positive constant c which only depends on $|f|_\infty$, b^{loc} such that*

$$\mathbb{P}\left(\Omega_n[A]^C\right) \leq \frac{c(|f|_\infty, A, b^{loc})}{n^2}.$$

Since we have already choosen $A = \inf(\frac{12\alpha e^{-|f|_\infty}}{\kappa(\alpha)}, \frac{\tau}{4e^{1+B}})$ the control of Ω_n^C only depends on $|f|_\infty$, b^{loc}, $\bar{B}$, α and τ.

6.2 Proof of Proposition 3.1

This proof is a simplier version of the preceding one, since we only have to deal with one single fixed model m, rather than an random model $\hat{m}$. With the same notations, we easily have that, for any model m,

$$\begin{aligned}K(f,\hat{f}_m) &= K(f,\bar{f}_m) + \gamma_n(\hat{f}_m) - \gamma_n(\bar{f}_m) + <\varepsilon, \hat{f}_m - \bar{f}_m>_n\\ &\leq K(f,\bar{f}_m) + <\varepsilon, \hat{f}_m - \bar{f}_m>_n\\ &\leq K(f,\bar{f}_m) + \frac{\theta_1}{2}\chi_n^2(m) + \frac{1}{2\theta_1}|\hat{f}_m - \bar{f}_m|_2^2\\ &\leq K(f,\bar{f}_m) + \frac{\theta_1}{2}\chi_n^2(m) + \frac{e^{|\hat{f}_m - \bar{f}_m|_\infty + |\bar{f}_m|_\infty}}{\theta_1}K(\bar{f}_m,\hat{f}_m),\end{aligned}$$

for any positive θ_1. Therefore, bounding $|\hat{f}_m - \bar{f}_m|_\infty + |\bar{f}_m|_\infty$ by $\tau/2 + \bar{B}$ on the set Ω_n and setting $\theta_2 = \frac{e^{\tau/2+\bar{B}}}{\theta_1}$, we have

$$(1-\theta_2)K(f,\hat{f}_m)\,\mathbb{1}_{\Omega_n} \leq (1-\theta_2)K(f,\bar{f}_m) + \frac{e^{\tau/2+\bar{B}}}{2\theta_2}\chi_n^2(m).$$

Now, choose $\theta_2 = 1/2$ and since $\mathbb{E}(\chi_n^2(m)) \leq e^{|f|_\infty}D_m/n$:

$$\mathbb{E}(K(f,\hat{f}_m)\,\mathbb{1}_{\Omega_n}) \leq K(f,\bar{f}_m) + 2e^{|f|_\infty+\bar{B}+\tau/2}\frac{D_m}{n}.$$

6.3 Proofs of the propositions involded in the proof of the Theorem

6.3.1 Concentration inequalities

The proofs of Propositions 6.1, 6.2 and 6.4 heavily depend on concentration inequalities established by (Reynaud-Bouret 2003). Her results are announced in terms of Poisson processes but we can translate them in our framework in the following way:

Let $\mathbb{X} =]0, n]$ and $I_i =]i-1, i], 1 \leq i \leq n$. Let μ denote the Lebesgue measure on $\mathbb{R}$ and let define $d\nu = \sum_{i=1}^{n} e^{f(x_i)}\, \mathbb{1}_{I_i}\, d\mu$. Let N be a Poisson process with inhomogeneous intensity $d\nu$. Then, the random variables $\int \mathbb{1}_{I_i}\, dN$ have Poisson distributions with parameter $\nu(I_i) = e^{f(x_i)}$.

For any $h \in \mathbb{R}^n$, let define $f_h = \sum_{i=1}^{n} h_i\, \mathbb{1}_{I_i}$. Then, $\int f dN = \sum_{i=1}^{n} h_i \int \mathbb{1}_{I_i}\, dN$ has the same distribution as $\sum_{i=1}^{n} h_i Y_i$. So, Reynaud-Bouret's inequalities (Reynaud-Bouret 2003) can be re-enunced in this way:

Theorem 6.1. Bernstein's inequality :
For any $\xi > 0$ and any $h \in \mathbb{R}^n$,

$$\begin{aligned} \mathbb{P}(\sum_{i=1}^{n} h_i\varepsilon_i \geq \xi) &\leq \exp\left(-\frac{\xi^2}{2\sum_{i=1}^{n} e^{f(x_i)}h_i^2 + \frac{2}{3}\xi|h|_\infty}\right) \\ \mathbb{P}(|\sum_{i=1}^{n} h_i\varepsilon_i| \geq \xi) &\leq 2\exp\left(-\frac{\xi^2}{2\sum_{i=1}^{n} e^{f(x_i)}h_i^2 + \frac{2}{3}\xi|h|_\infty}\right) \qquad (6.14) \end{aligned}$$

For any $u > 0$ and any $h \in \mathbb{R}^n$,

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{n} h_i\varepsilon_i \geq (2u\sum_{i=1}^{n} e^{f(x_i)}h_i^2)^{1/2} + |h|_\infty u/3\right) &\leq e^{-u}, \\ \mathbb{P}\left(|\sum_{i=1}^{n} h_i\varepsilon_i| \geq (2u\sum_{i=1}^{n} e^{f(x_i)}h_i^2)^{1/2} + |h|_\infty u/3\right) &\leq 2e^{-u}. \qquad (6.15) \end{aligned}$$

We will also need the following theorem:

Theorem 6.2. *Let S be some finite dimensional linear subspace of $\mathbb{L}_2$ and $(\varphi_\lambda)_{\lambda=1,\ldots,D}$ be some orthonormal basis of S for the inner product $<,>_n$. Let χ_n be the following Chi-square statistics:*

$$\chi_n(S) = \sup_{f\in S, |f|_n=1} < f, \varepsilon >_n = \left(\sum_{\lambda=1,\ldots,D} < \phi_\lambda, \varepsilon >_n^2\right)^{1/2}.$$

Let

$$M_S = \sup_{h\in S, |h|_n=1} n^{-1}\sum_{i=1}^{n} e^{f(x_i)}h_i^2$$

and assume that this quantity is finite. Let $\Omega_S(\alpha)$ be the event

$$\Omega_S(\alpha) = \left\{ |\sum_{\lambda=1,\ldots,D} < \phi_\lambda, \varepsilon >_n \varphi_\lambda|_\infty \leq \frac{12\alpha M_S}{\kappa(\alpha)} \right\}, \qquad (6.16)$$

where

$$\kappa(\alpha) = 5/4 + 32/\alpha. \qquad (6.17)$$

Then, for any positive α and x

$$\mathbb{P}\left(\chi_n(S)\,\mathbb{1}_{\Omega_S(\alpha)} \geq (1+\alpha)\left(\mathbb{E}(\chi_n^2(S))^{1/2} + (12M_S x/n)^{1/2}\right)\right) \leq e^{-x}. \qquad (6.18)$$

6.3.2 Proof of Proposition 6.1

For sake of simplicity we use here the notations $M_{m'} = M_{S_{m'}}$ and $M_\Lambda = M_{S_{\Lambda_n^*}}$. Define for any model $m' \in \mathcal{M}(L_n)$,

$$\begin{aligned}
\Omega_n^1(m') &= \left\{\chi_n(m')\,\mathbb{1}_{\Omega_{S_{m'}}} \leq (1+\alpha)\left(\mathbb{E}(\chi_n^2(m'))^{1/2} + (12M_{m'}x_{m'}/n)^{1/2}\right)\right\} \\
\Omega_n^1 &= \bigcap_{m'\in\mathcal{M}(L_n)} \Omega_n^1(m'),
\end{aligned}$$

where $\Omega_{S_{m'}}$ is defined by (6.16). From (6.18), we have

$$\mathbb{P}\left({\Omega_n^1}^C\right) \leq \sum_{m'\in\mathcal{M}(L_n)} \mathbb{P}\left(\Omega_n^1(m')^C\right) \leq \sum_{m'\in\mathcal{M}(L_n)} e^{-x_{m'}}.$$

Using Property 1, since $m' \subset \Lambda_n^*$ we have

$$|\sum_{\lambda\in m'} < \phi_\lambda, \varepsilon >_n \phi_\lambda|_\infty \leq b^{loc} D_{m'}^{1/2} \sup_{\lambda\in m'} |< \phi_\lambda, \varepsilon >_n| \leq b^{loc} D_{m'}^{1/2} \sup_{\lambda\in\Lambda_n^*} |< \phi_\lambda, \varepsilon >_n|.$$

Furthermore, for any m', $A \leq \frac{12\alpha e^{-|f|_\infty}}{\kappa(\alpha)} \leq \frac{12\alpha M_{m'}}{\kappa(\alpha)}$. Thus on the set $\Omega_n[A]$ we have

$$|\sum_{\lambda\in m'} < \phi_\lambda, \varepsilon >_n \phi_\lambda|_\infty \leq \frac{12\alpha n^{-\rho} M_{m'}}{\kappa(\alpha)} \leq \frac{12\alpha M_{m'}}{\kappa(\alpha)}.$$

Therefore, for any model m', $\Omega_n[A] \subset \Omega_{S_{m'}}$, so that on the set Ω_n^1,

$$\chi_n(m')\,\mathbb{1}_{\Omega_n[A]} \leq \chi_n(m')\,\mathbb{1}_{\Omega_{S_{m'}}} \leq (1+\alpha)\left(\mathbb{E}(\chi_n^2(m'))^{1/2} + (12M_{m'}x_{m'}/n)^{1/2}\right).$$

Moreover, we have:

$$\mathbb{E}(\chi_n^2(m')) = \sum_{\lambda\in m'} \mathbb{E} < \varphi_\lambda, \varepsilon >_n^2 \leq \sum_{\lambda\in m'} \mathbb{V}ar < \varphi_\lambda, \varepsilon >_n \leq \sum_{\lambda\in m'} \frac{M_{m'}}{n} = \frac{M_{m'}D_{m'}}{n}.$$

Noticing that $M_{m'} \leq e^{|f|_\infty}$ for any model $m' \in \mathcal{M}(L_n)$, (6.5) holds true for any model m'. Hence it is true for $m' = \hat{m}$.

6.3.3 Proof of Proposition 6.2

Let $\Omega_n^2(m')$ be defined for any model $m' \in \mathcal{M}(L_n)$ by

$$\begin{aligned}
\Omega_n^2(m') &= \left\{|< \bar{f}_{m'} - f, \varepsilon >_n| \leq \left(\frac{2y_{m'}}{n^2}\sum_{i=1}^n e^{f(x_i)}(\bar{f}_{m',i} - f_i)^2\right)^{1/2} + |\bar{f}_{m'} - f|_\infty \frac{y_{m'}}{3n}\right\}, \\
\Omega_n^2 &= \bigcap_{m'} \Omega_n^2(m').
\end{aligned}$$

Applying Bernstein's inequality (6.15) for $h = \bar{f}_{m'} - \bar{f}_m$, we deduce that $\mathbb{P}({\Omega_n^2}^C) \leq 2\sum_{m'} e^{-y_{m'}}$. Next, using (7.4),

$$\frac{\sum_{i=1}^n e^{f(x_i)}(\bar{f}_{m',i} - f_i)^2}{n^2} = \frac{V_f(f, \bar{f}_{m'})}{n} \leq \frac{2e^{|\bar{f}_{m'} - f|_\infty}}{n} K(f, \bar{f}_{m'}),$$

so that on the set Ω_n^2,

$$|< \bar{f}_{m'} - f, \varepsilon >_n| \leq \left(\frac{e^{|\bar{f}_{m'} - f|_\infty} y_{m'}}{n} 2K(f, \bar{f}_{m'})\right)^{1/2} + |\bar{f}_{m'} - f|_\infty \frac{y_{m'}}{3n},$$

using (6.2) with $a = (2K(f, \bar{f}_{m'}))^{1/2}$ and $b = (\frac{e^{|\bar{f}_{m'}-f|_\infty} y_{m'}}{n})^{1/2}$, for any model m'

$$\begin{aligned}
| < \bar{f}_{m'} - f, \varepsilon >_n | &\leq \frac{1}{2\theta_2} \frac{e^{|\bar{f}_{m'}-f|_\infty} y_{m'}}{n} + \theta_2 K(f, \bar{f}_{m'}) + |\bar{f}_{m'} - f|_\infty \frac{y_{m'}}{3n} \\
&\leq \frac{e^{|\bar{f}_{m'}-f|_\infty}}{2} (1 + \frac{1}{\theta_2}) \frac{y_{m'}}{n} + \theta_2 K(f, \bar{f}_{m'}),
\end{aligned}$$

for some positive constant θ_2. Since this is true for all models m', this is in particular true for $m' = \hat{m}$ and for $m' = m$ with θ_2 replaced by θ_3. To conclude, (6.6) follows from

$$| < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n | \leq | < \bar{f}_{\hat{m}} - f, \varepsilon >_n | + | < f - \bar{f}_m, \varepsilon >_n |.$$

6.3.4 Proof of Proposition 6.3

On the set $\Omega_n[A]$, for any $m' \subset \Lambda_n^*$, we have:

$$|\hat{\delta}_{m'} - \bar{\delta}_{m'}|_n^2 = \sum_{\lambda \in m'} < \phi_\lambda, \varepsilon_\lambda >^2 \leq \sum_{\lambda \in m'} \frac{A^2 n^{-2\rho}}{(b^{loc})^2 |\Lambda_n^*|} \leq \frac{A^2 n^{-2\rho}}{(b^{loc})^2}.$$

Due to Assumption 2 and to (6.4), we have that for any model m', $n^{-\rho} \leq n^{-(1-\theta)/2} \leq \frac{1}{D_{m'}^{1/2}}$. Since A satisfies (6.11), using Assumption 3, we have for any model m', $A \leq \frac{\tau}{4e^{1+|\bar{f}_{m'}|_\infty}}$, and thus

$$|\hat{\delta}_{m'} - \bar{\delta}_{m'}|_n \leq \frac{A n^{-\rho}}{b^{loc}} \leq \frac{\tau}{4 b^{loc} D_{m'}^{1/2} e^{1+|\bar{f}_{m'}|_\infty}}.$$

Hence, we can apply Lemma 7.5 on every model m' to get the result.

6.3.5 Proof of Proposition 6.4

From the definition of $\Omega_n[A]$, we have

$$\mathbb{P}\left(\Omega_n[A]^C\right) \leq \sum_{\lambda \in \Lambda_n^*} \mathbb{P}\left(| < \phi_\lambda, \varepsilon >_n | \geq \frac{A n^{-\rho}}{b^{loc} |\Lambda_n^*|^{1/2}} \right).$$

Using Bernstein's inequality (6.14) and setting $\xi(A) = \frac{A n^{-\rho}}{b^{loc} |\Lambda_n^*|^{1/2}}$, we get

$$\mathbb{P}\left(| < \phi_\lambda, \varepsilon >_n | \geq \xi(A) \right) \leq 2 \exp\left(- \frac{n^2 \xi^2}{2 \sum_{\lambda \in \Lambda_n^*} e^{f(x_i)} \phi_{\lambda,i}^2 + \frac{2}{3} n \xi |\phi_\lambda|_\infty} \right).$$

Since Assumption 1 gives orthonormality of the basis (ϕ_λ) for the $<, >_n$ inner product,

$$\sum_{\lambda \in \Lambda_n^*} e^{f(x_i)} \phi_{\lambda,i}^2 \leq e^{|f|_\infty} n |\varphi_\lambda|_n^2 = n e^{|f|_\infty}.$$

Furthermore, due to Property 1, for any $\lambda \in \Lambda_n^*$:

$$|\varphi_\lambda|_\infty \leq b^{loc} |\Lambda_n^*|^{1/2},$$

so that

$$\mathbb{P}\left(| < \phi_\lambda, \varepsilon >_n | \geq \xi(A) \right) \leq 2 \exp\left(-\eta(A) \frac{n^{1-2\rho}}{e^{|f|_\infty} {b^{loc}}^2 |\Lambda_n^*|} \right),$$

where $\eta(A) = \frac{A^2}{2+2A/3}$. Now, using Assumption 2, we get

$$\mathbb{P}\left(\Omega_n[A]^C\right) \leq 2|\Lambda_n^*| \exp(-\eta(A) \frac{n1 - 2\rho}{e^{|f|_\infty} {b^{loc}}^2 |\Lambda_n^*|}) \leq 2n^{1-\theta} \exp(-\eta(A) \frac{n^{\theta-2\rho}}{e^{|f|_\infty} {b^{loc}}^2})$$
$$= \frac{2}{n^2} n^{3-\theta} \exp(-Cn^{\theta-2\rho}),$$

where C is a positive constant depending on A, $|f|_\infty$ and b^{loc} but not on n. Since $\theta - 2\rho > 0$ from (6.4), $n^{3-\theta} \exp(-Cn^{\theta-2\rho})$ tends to 0 when n tends to infinity, so that the sequence remains bounded, which yields the result.

6.4 Proof of the lower bound given in Theorem 4.1

Let $\mathcal{F}_M$ denotes a finite subset of cardinality $M+1$ of $\mathcal{F} \cap \mathcal{C}^\infty(S)$, then we have for any estimator $\hat{f}_n$ of f:

$$\sup_{f \in \mathcal{F} \cap \mathcal{C}^\infty(S)} \mathop{\mathbb{E}}_f(K(f, \hat{f}_n) v_n^{-2}) \geq \sup_{f \in \mathcal{F}_M} \mathop{\mathbb{E}}_f((K(f, \hat{f}_n) v_n^{-2}).$$

Next, due to inequality (7.5) which provides a lower bound in discrete quadratic norm for the Kullback Leibler distance, we obtain that for any $f \in \mathcal{F}_M$ and any $\hat{f}_n \in \mathcal{C}^\infty(S)$:

$$\mathop{\mathbb{E}}_f((K(f, \hat{f}_n) v_n^{-2}) \geq \frac{e^{-3S}}{2} \mathop{\mathbb{E}}_f(|\hat{f}_n - f|_n^2 v_n^{-2}) \geq \frac{e^{-3S}}{2} \mathop{\mathbb{P}}_f(|\hat{f}_n - f|_n v_n^{-1} > \xi) \xi^2.$$

Hence, for any $\xi > 0$ and any $\hat{f}_n \in \mathcal{C}^\infty(S)$, when denoting f_k the elements of $\mathcal{F}_M$:

$$\sup_{f \in \mathcal{F}_M} E_f((K(f, \hat{f}_n) v_n^{-2}) \quad \geq \quad \frac{e^{-3S}}{2} \max_{k=0,\ldots,M} P_{f_k}(|\hat{f}_n - f_k|_n v_n^{-1} > \xi) \xi^2. \tag{6.19}$$

Therefore, the assertion of the proposition will follow from a non negative lower bound which does not depend on $\hat{f}_n$, for the probability in the right hand side quantity of this last inequality.

For the convenience of the reader we recall the basic tool (Theorem 2.5 in (Tsybakov 2004) p.85) we use to obtain such a bound. Note that for sake of simplicity we use a simplified version of the one given in (Tsybakov 2004), since we only wish to obtain optimal rate and do not investigate the more difficult problem of an optimal constant in the lower bound.

Lemma 6.1. *Suppose that the elements $f_0, ..., f_M \in \mathcal{F}_M, M \geq 2$ are such that*

a) For all k, k', $\quad 0 \leq k < k' \leq M$ inequality holds:

$$|f_k - f_{k'}|_n \geq 2s_n > 0; \tag{6.20}$$

b) For any $k = 1, ..., M$ the Kullback-Leibler divergence between the likelihoods under f_k and f_0 satisfy

$$\frac{1}{M} \sum_{k=1}^{M} nK(f_k, f_0) \leq a \ln(M) \tag{6.21}$$

where $0 < a < 1/10$.

Then for any estimator $\hat{f}_n \in \mathcal{C}^\infty(S)$

$$\max_{0 \leq k \leq M} P_{f_k}(|\hat{f}_n - f_k|_n \geq s_n) \geq c > 0, \qquad \text{with} \qquad c = 0.04$$

We construct now a convenient set of functions $\mathcal{F}_M$, that will verify Assumptions (6.20) and (6.21) for M large enough that will be choosen as an increasing funtion of n.

Let us consider a real positive function $\Phi(\cdot)$ (called basic function for the class $\Sigma(\nu, 1)$, with $\nu = k + \alpha$) satisfying assumptions given in Lemma 6.2. Set $m \in \mathbb{N}$ with $m \geq 8$ and consider the sequence of points $b_j = (j - 1/2)/m$ for all $j = 1, ..., m$, and of functions f_{jn} defined as :

$$b_j = \frac{j - 1/2}{m} \qquad f_{jn} = L \left(\frac{1}{m} \right)^{\nu} \Phi \left(\frac{x - b_j}{1/m} \right).$$

In the following lemma, we state the properties of functions f_{jn} that are necessary for constructing a subset $\mathcal{F}_M$ of functions satisfying (6.21) and (6.20).

Lemma 6.2. *Let $\Phi \in \Sigma(\nu, 1)$ be compactly supported over $[-1/2, 1/2]$, such that $\|\Phi\|_\infty \leq 8^\nu S/L$ and $\|\Phi\|_2^2 < \ln 2/(60L^2)$. Moreover we suppose that Φ has all its derivatives up to order $k+1$ with its $k+1$-th derivative uniformly bounded by 1. Let $m \geq 8$ then for any $j = 1...m$:*

i) f_{jn} is compactly supported over $[(j-1)/m, j/m]$, such that $\|f_{jn}\|_\infty = Lm^{-\nu}\|\Phi\|_\infty$, $\|f_{jn}\|_2^2 = L^2\|\Phi\|_2^2 m^{-(2\nu+1)}$ and $\|f'_{jn}\|_\infty = Lm^{-\nu+1}\|\Phi'\|_\infty$.

ii) $f_{jn} \in \mathcal{F} \cap \mathcal{C}^\infty(S)$.

iii) $|\|f_{jn}\|_2^2 - |f_{jn}|_n^2| \leq \|f_{jn}\|_\infty \|f'_{jn}\|_\infty n^{-1}$

Proof. The first point of the lemma is a straightforward consequence of required assumptions on the basic function Φ.

The Kernel Φ being compactly supported and having its $k+1$-th derivative uniformly bounded by 1, the k-th derivative of f_{jn} satisfies condition (4.1). Moreover, since $\|\Phi\|_\infty \leq 8^\nu S/L$ and due to i), f_{jn} is obviously bounded by S.

The third point is an application of Taylor expansion of f_{jn} at order one around each point $x_i = i/n$ of the design, indeed:

$$\begin{aligned}
|\|f_{jn}\|_2^2 - |f_{jn}|_n^2| &= |\sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f_{jn}^2(x) - f_{jn}^2(x_i))dx| \leq \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} |f_{jn}^2(x) - f_{jn}^2(x_i)|dx \\
&\leq 2\|f_{jn}\|_\infty \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} |f_{jn}(x) - f_{jn}(x_i)|dx \\
&\leq 2\|f_{jn}\|_\infty \|f'_{jn}\|_\infty \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} |x - x_i|dx = \|f_{jn}\|_\infty \|f'_{jn}\|_\infty n^{-1}
\end{aligned}$$

□

Consider now the set of all binary possible vectors $\bar{w} = (w_1, ..., w_m)$, $w_l \in \{0, 1\}$, $l = 1, ..., m$. Due to Varshanov-Gilbert Lemma (1962) (see (Tsybakov 2004) p. 89), if $m \geq 8$, there exists a subset $\mathcal{W} = (\bar{w}_0, ..., \bar{w}_M)$ such that $\bar{w}_0 = (0, ..., 0)$ and for any $0 \leq k < k' \leq M$

$$\rho_H(\bar{w}^k, \bar{w}^{k'}) = card\{l : 1 \leq l \leq m, w_l^k \neq w_l^{k'}\} \geq m/16 \quad \text{and} \quad 8\ln(M)/\ln(2) \geq m. \qquad (6.22)$$

Next, for each binary sequences $\bar{w}_k \in \mathcal{W}$, we define the function

$$f_k(x) = \sum_{j=1}^{m} w_j^k f_{jn}(x).$$

Since the supports of f_{jn} are non-overlapping, we have for any $k = 0, ..., M$, $f_k \in \mathcal{F} \cap \mathcal{C}^\infty(S)$ and $\|f_k\|_\infty \leq Lm^{-\nu}\|\Phi\|_\infty$. Let us check now that functions f_k also satisfy Assumptions (6.20) and (6.21), for n and M large enough.

When using Lemma 6.2 and the Varshanov-Gilbert upper bound for ρ_H given in (6.22), we get for any $0 \leq k < k' \leq M$, and for any n and $m \geq 8$:

On one hand,

$$\begin{aligned}
|f_k - f_{k'}|_n^2 &= \sum_{j=1}^{m} (w_j^k - w_j^{k'})^2 |f_{jn}|_n^2 \geq \sum_{j=1}^{m} (w_j^k - w_j^{k'})^2 (\|f_{jn}\|^2 - \|f_{jn}\|_\infty \|f'_{jn}\|_\infty n^{-1}) \\
&= \rho_H(\bar{w}_k, \bar{w}_{k'})(L^2 \|\Phi\|_2^2 m^{-(2\nu+1)} - L^2 m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\
&\geq m^{-2\nu} L^2 \frac{\|\Phi\|_2^2}{16} R_{m,n} \quad \text{with} \quad R_{m,n} = 1 - \frac{\|\Phi\|_\infty \|\Phi'\|_\infty m^2}{L^2 \|\Phi\|_2^2 n}; \qquad (6.23)
\end{aligned}$$

on the other hand, when also using inequality (7.5):

$$\begin{aligned}
nK(f_k, f_0) &\leq \frac{1}{2} e^{\|f_k\|_\infty + \|f_k - f_0\|_\infty} n |f_k - f_0|_n^2 \leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \left(\sum_{j=1}^{m} (w_j^k)^2 |f_{jn}|_n^2 \right) \\
&\leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \sum_{j=1}^{m} (\|f_{jn}\|_2^2 + L^2 m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\
&\leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \sum_{j=1}^{m} (L^2 \|\Phi\|_2^2 m^{-(2\nu+1)} + L^2 m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\
&\leq L^2 \|\Phi\|_2^2 \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} m \left(n m^{-(2\nu+1)} + \frac{m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty}{\|\Phi\|_2^2} \right) \\
&\leq L^2 \|\Phi\|_2^2 \frac{4 \ln M}{\ln 2} P_{m,n} \;\; \text{with} \;\; P_{m,n} = e^{2\|\Phi\|_\infty L m^{-\nu}} \left(\frac{n}{m^{2\nu+1}} + \frac{\|\Phi\|_\infty \|\Phi'\|_\infty}{\|\Phi\|_2^2 m^{2\nu-1}} \right)
\end{aligned} \tag{6.24}$$

Now we put $m = n^{1/(2\nu+1)}$. For such a choice, $R_{m,n}$ increases and tends to one, and $P_{m,n}$ decreases and tends to one when n tends to infinity. Hence for n large enough $\sqrt{R_{m,n}} \geq 1/2$ and $P_{m,n} \leq 3/2$ and we have when substituting these bounds in (6.23) and (6.24):

$$|f_k - f_{k'}|_n \geq n^{\frac{-\nu}{2\nu+1}} \frac{L\|\Phi\|_2}{8} \quad \text{and} \quad nK(f_k, f_0) \leq 3L^2 \|\Phi\|_2^2 \ln M.$$

Hence Assumptions (6.20) and (6.21) are obtained for $s_n = n^{-\nu/(2\nu+1)}(L\|\Phi\|_2)/16$ and $a = 3L^2\|\Phi\|_2^2$, for n and M large enough, and Lemma 6.1 provides the estimate:

$$\max_{k=0,\ldots,M} P_{f_k}(|\hat{f}_n - f_k|_n > \xi v_n) \geq 0.04 \quad \text{for} \;\; \xi = L\|\Phi\|_2/8.$$

To end the proof, we substitute the previous lower bound in (6.19) which provides the result given in Theorem 4.1 with $C = 0.04\|\Phi\|_2^2/128$.

7 Appendix

7.1 Technical lemmas

In the sequel, for sake of simplicity we put $D = D_m$ and for any h and f we will denote :

$$\begin{aligned}
\beta &= (\beta_\lambda)_{\lambda \in m} = (< h, \phi_\lambda >_n)_{\lambda \in m} \\
\bar{\beta} &= (\bar{\beta}_\lambda)_{\lambda \in m} = (< \bar{f}_m, \phi_\lambda >_n)_{\lambda \in m} \\
\hat{\beta} &= (\hat{\beta}_\lambda)_{\lambda \in m} = (< \hat{f}_m, \phi_\lambda >_n)_{\lambda \in m}.
\end{aligned}$$

7.1.1 Estimator and projection on a given model

Due to their definitions, (1.1) and (1.2), $\hat{f}_m$ and $\bar{f}_m$ have no simple analytical expression. Nevertheless, they satisfy the following relations :

Lemma 7.1. *For any $m \in \mathcal{M}(L_n)$ and any function $h \in S_m$,*

$$\begin{aligned}\sum_{i=1}^{n} Y_i h_i &= \sum_{i=1}^{n} e^{\hat{f}_{m,i}} h_i \qquad (7.1)\\ \sum_{i=1}^{n} e^{f_i} h_i &= \sum_{i=1}^{n} e^{\bar{f}_{m,i}} h_i. \qquad (7.2)\end{aligned}$$

Consequently,

$$\sum_{i=1}^{n} e^{\bar{f}_{m,i}} h_i = \mathbb{E}_f(\sum_{i=1}^{n} e^{\hat{f}_{m,i}} h_i). \qquad (7.3)$$

Proof. Since $h \in S_m$ we have $h = \sum_{\lambda=1}^{D} \beta_\lambda \phi_\lambda$ and

$$\gamma_n(h) = n^{-1} \sum_{i=1}^{n} (e^{h_i} - Y_i h_i) = n^{-1} \sum_{i=1}^{n} (\exp(\sum_{\lambda=1}^{D} \beta_\lambda \phi_{\lambda,i}) - Y_i \sum_{\lambda=1}^{D} \beta_\lambda \phi_{\lambda,i}).$$

Deriving with respect to β_{λ_0}, and $\hat{f}_m = \sum_{\lambda \in m} \beta_\lambda \phi_\lambda$ being a minimiser of the contrast function $\gamma_n(h)$ we get for any $\lambda_0 = 1, ..., D$:

$$\sum_{i=1}^{n} (\exp(\sum_{\lambda=1}^{D} \hat{\beta}_\lambda \varphi_{\lambda,i}) \phi_{\lambda_0,i} - Y_i \phi_{\lambda_0,i}) = 0.$$

Hence, for any function $\phi_{\lambda_0,i}$ of the basis of S_m relation (7.1) being satisfied, it holds also true for any linear combination of them. The proof of the second assertion (7.2) is analogous, so it is omitted. The third assertion obviously follows when noticing that expectation of the left hand side of (7.1) is equal to the left hand side of (7.2). □

7.1.2 Pythagoras Equality

Lemma 7.2. *For any $m \in \mathcal{M}(L_n)$ and any function $h \in S_m$, we have:*

$$K(f, h) = K(f, \bar{f}_m) + K(\bar{f}_m, h).$$

Proof.

$$\begin{aligned}K(f,h) &= n^{-1} \sum_{i=1}^{n} e^{h_i} - e^{f_i} - e^{f_i}(h_i - f_i)\\ &= n^{-1} \sum_{i=1}^{n} e^{h_i} - e^{\bar{f}_{m,i}} + e^{\bar{f}_{m,i}} - e^{f_i} - e^{f_i}(h_i - \bar{f}_{m,i} + \bar{f}_{m,i} - f_i)\\ &= K(f, \bar{f}_m) + n^{-1} \sum_{i=1}^{n} e^{h_i} - e^{\bar{f}_{m,i}} - e^{f_i}(h_i - \bar{f}_{m,i})\end{aligned}$$

The functions h et $\bar{f}_m$ are both in S_m, so is their difference. Therefore, when applying relation (7.2) we obtain:

$$\sum_{i=1}^{n} e^{f_i}(h_i - \bar{f}_{m,i}) = \sum_{i=1}^{n} e^{\bar{f}_{m,i}}(h_i - \bar{f}_{m,i}).$$

Then we get:

$$K(f,h) = K(f, \bar{f}_m) + n^{-1} \sum_{i=1}^{n} e^{h_i} - e^{\bar{f}_{m,i}} - e^{\bar{f}_{m,i}}(h_i - \bar{f}_{m,i}) = K(f, \bar{f}_m) + K(\bar{f}_m, h).$$

□

7.1.3 Links between distances

Lemma 7.3. *For any functions f and h,*

$$e^{-|h-f|_\infty}\frac{V_f(f,h)}{2} \le \quad K(f,h) \quad \le e^{|h-f|_\infty}\frac{V_f(f,h)}{2}, \tag{7.4}$$

$$e^{-|f|_\infty-|h-f|_\infty}\frac{|h-f|_n^2}{2} \le \quad K(f,h) \quad \le e^{|f|_\infty+|h-f|_\infty}\frac{|h-f|_n^2}{2}, \tag{7.5}$$

where

$$V_f(f,h) = n^{-1}\sum_{i=1}^n e^{f(x_i)}(h(x_i)-f(x_i))^2.$$

Proof. Recall the definition of the Kullback-Leibler divergence given in the introduction :

$$K(f,h) = \mathbb{E}_f(\gamma_n(h)-\gamma_n(f)) = n^{-1}\sum_{i=1}^n e^{f_i}(e^{h_i-f_i}-1-(h_i-f_i)).$$

Since for any $x \in \mathbb{R}$, $\frac{x^2}{2}e^{-|x|} \le e^x - 1 - x \le \frac{x^2}{2}e^{|x|}$, we have :

$$n^{-1}\sum_{i=1}^n e^{f_i}(e^{-|h_i-f_i|}\frac{(h_i-f_i)^2}{2}) \le K(f,h) \le n^{-1}\sum_{i=1}^n e^{f_i}(e^{+|h_i-f_i|}\frac{(h_i-f_i)^2}{2}). \tag{7.6}$$

Moreover, for any i, $\exp(-|h_i-f_i|) \ge \exp(-|h-f|_\infty)$ and $\exp(|h_i-f_i|) \le \exp(|h-f|_\infty)$. Hence, substituting these bounds in (7.6) we obtain (7.5). Next, since $\exp(-|f|_\infty) \le \exp(f_i) \le \exp(|f|_\infty)$ we have $\exp(-|f|_\infty)|h-f|_n^2 \le V_f(f,h) \le \exp(|f|_\infty)|h-f|_n^2$ and (7.4) follows. □

The next lemma deals with links between norms of functions and norms of the coefficient vectors in an orthonormalized basis, for the $<,>_n$ inner product.

Lemma 7.4. *Suppose Assumption 1 satisfied. Then for any $h \in S_m$ of dimension D_m :*

$$|h-\bar{f}_m|_\infty \quad \le \quad b^{loc}D_m^{1/2}|\beta-\bar{\beta}|_2, \tag{7.7}$$

$$\frac{e^{-|\bar{f}_m|_\infty}}{2}e^{-b^{loc}D_m^{1/2}|\beta-\bar{\beta}|_2}|\beta-\bar{\beta}|_2^2 \le K(\bar{f}_m,h) \quad \le \quad \frac{e^{|\bar{f}_m|_\infty}}{2}e^{b^{loc}D_m^{1/2}|\beta-\bar{\beta}|_2}|\beta-\bar{\beta}|_2^2.$$

Proof. Since $|\beta-\bar{\beta}|_\infty \le |\beta-\bar{\beta}|_2^2$, Assertion (7.7) follows immediately from Property 1.

For the second one we have :

$$K(\bar{f}_m,h) = n^{-1}\sum_{i=1}^n e^{h_i} - e^{\bar{f}_{m,i}} - e^{\bar{f}_{m,i}}(h_i-\bar{f}_{m,i}) = n^{-1}\sum_{i=1}^n e^{\bar{f}_{m,i}}(e^{h_i-\bar{f}_{m,i}}-1-(h_i-\bar{f}_{m,i})).$$

Applying inequalities (7.5), (7.7) and noticing that for any $h \in S_m$, $|h|_n^2 = |\beta|_2^2$, we obtain:

$$K(\bar{f}_m,h) \quad \le \quad e^{|\bar{f}_m|_\infty}e^{|h-\bar{f}_m|_\infty}\frac{|h-\bar{f}_m|_n^2}{2} \le e^{|\bar{f}_m|_\infty}e^{b^{loc}D_m^{-1/2}|\beta-\bar{\beta}|_2}\frac{|\beta-\bar{\beta}|_2^2}{2}.$$

The lower bound is deduced in the same way. □

7.1.4 Control of $K(\bar{f}_m,\hat{f}_m)$

In this section, we aim at controlling some distances between $\bar{f}_m$ and $\hat{f}_m$ in any model $m \in \Lambda_n^*$. Remark that the vectors $(\hat{\beta}_\lambda)_{\lambda\in m}$ and $(\bar{\beta}_\lambda)_{\lambda\in m}$ of coefficients of $\hat{f}_m$ and $\bar{f}_m$ in the basis $(\phi_\lambda)_{\lambda\in m}$

satisfy

$$G(\bar{\beta}) = \bar{\delta}_m \quad \text{and} \quad G(\hat{\beta}) = \hat{\delta}_m,$$

where the G is the function from $\mathbb{R}^{D_m}$ to $\mathbb{R}^{D_m}$ whose $\lambda - th$ component is given by :

$$G_\lambda(\beta) = n^{-1} \sum_{i=1}^{n} e^{\sum_\lambda \beta_\lambda \phi_{\lambda,i}} \phi_{\lambda,i}$$

and $\hat{\delta}_m$ and $\bar{\delta}_m$ are vectors in $\mathbb{R}^{D_m}$ with λ-th coordinates

$$\hat{\delta}_{m,\lambda} = n^{-1} \sum_{i=1}^{n} Y_i \phi_{\lambda,i} \text{ and } \bar{\delta}_{m,\lambda} = n^{-1} \sum_{i=1}^{n} e^{f_i} \phi_{\lambda,i}.$$

The following lemma is adapted from (Barron & Sheu 1991) and (Castellan 2003).

Lemma 7.5. *For any* $\tau \in]0,1[$*, if*

$$|\hat{\delta}_m - \bar{\delta}_m|_2 \quad \leq \quad \frac{\tau}{(4b^{loc} D_m^{1/2} e^{1+|\bar{f}_m|})}, \tag{7.8}$$

then equation $G(\beta) = \hat{\delta}_m$ *admits a solution* $\hat{\beta}$ *which satisfies*

$$|\hat{f}_m - \bar{f}_m|_\infty \leq 2e^{|\bar{f}_m| + \tau} b^{loc} D_m^{-1/2} |\hat{\delta}_m - \bar{\delta}_m|_2 \leq \tau/2.$$

Proof. For any $\delta \in \mathbb{R}^{D_m}$, define F_δ as the function from $\mathbb{R}^{D_m}$ to $\mathbb{R}$ whose derivative with respect to β_λ is $G_\lambda(\beta) - \delta_\lambda$:

$$F_\delta(\beta) = n^{-1} \sum_{i=1}^{n} e^{\sum_\lambda \beta_\lambda \phi_{\lambda,i}} - \sum_\lambda \delta_\lambda \beta_\lambda.$$

Due to definition of F_δ, solving equation $G(\beta) = \hat{\delta}$ comes to minimize $F_{\hat{\delta}}$. Now for any $\beta \in \mathbb{R}^{D_m}$,

$$\begin{aligned}
F_{\hat{\delta}}(\beta) - F_{\hat{\delta}}(\bar{\beta}) &= n^{-1} \sum_{i=1}^{n} e^{\sum_\lambda \beta_\lambda \phi_{\lambda,i}} - \sum_\lambda \hat{\delta}_\lambda \beta_\lambda - n^{-1} \sum_{i=1}^{n} e^{\sum_\lambda \bar{\beta}_\lambda \phi_{\lambda,i}} + \sum_\lambda \hat{\delta}_\lambda \bar{\beta}_\lambda \\
&= K(\bar{f}_m, h(\beta)) + n^{-1} \sum_{i=1}^{n} e^{\bar{f}_{m,i}} \sum_\lambda (\beta_\lambda - \bar{\beta}_\lambda) \phi_{\lambda,i} - < \hat{\delta}, \beta - \bar{\beta} > \\
&= K(\bar{f}_m, h(\beta)) + < \bar{\delta}, \beta - \bar{\beta} > - < \hat{\delta}, \beta - \bar{\beta} > \\
&= K(\bar{f}_m, h(\beta)) - < \hat{\delta} - \bar{\delta}, \beta - \bar{\beta} > \\
&\geq \frac{e^{-|\bar{f}_m|_\infty}}{2} e^{-b^{loc} D_m^{1/2} |\beta - \bar{\beta}|_2} |\beta - \bar{\beta}|_2^2 - |\hat{\delta} - \bar{\delta}|_2 |\beta - \bar{\beta}|_2.
\end{aligned}$$

Let τ be some number in $]0,1[$ and consider the sphere $\{\beta, |\beta - \bar{\beta}|_2 = 2e^\tau e^{|\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2\}$. For any β on the sphere,

$$F_{\hat{\delta}}(\beta) - F_{\hat{\delta}}(\bar{\beta}) > (e^{\tau - 2b^{loc} D_m^{1/2} e^\tau e^{|\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2} - 1) 2e^\tau e^{|\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2^2.$$

Due to (7.8) and since $0 < \tau < 1$, $2b^{loc} D_m^{1/2} e^{1+|\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2 < \tau < 1$, hence for any β on the sphere $F_{\hat{\delta}}(\beta) - F_{\hat{\delta}}(\bar{\beta}) > 0$. Moreover the function $F_{\hat{\delta}}(\cdot) - F_{\hat{\delta}}(\bar{\beta})$ being continuous and equal to zero in the center of the sphere $\bar{\beta}$ it admits a minimizer inside the sphere, say $\hat{\beta}$, such that $|\hat{\beta} - \bar{\beta}|_2 < 2e^{\tau + |\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2$.

Thus, from Lemma 7.4,

$$|\hat{f}_m - \bar{f}_m|_\infty \leq b^{loc} D_m^{1/2} |\hat{\beta} - \bar{\beta}|_2 \leq 2b^{loc} D_m^{1/2} e^{\tau + |\bar{f}_m|_\infty} |\hat{\delta} - \bar{\delta}|_2 \leq \tau/2.$$

□

7.1.5 Integration lemma

Lemma 7.6. *Let X and Y be positive random variables defined on the probability space $(\Omega, \mathbb{P})$. Assume that there exist positive constants κ_1 and κ_2 such that $\mathbb{P}(X \geq Y + \kappa_1\zeta) \leq \kappa_2 e^{-\zeta}$, then $\mathbb{E}(X) \leq \mathbb{E}(Y) + \kappa_1\kappa_2$.*

Proof. By definition, using Fubini,

$$\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X \geq x)\, dx = \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{X \geq x\}}\, dx d\,\mathbb{P}.$$

The latter event can be decomposed as

$$\mathbb{1}_{\{X \geq x\}} = \mathbb{1}_{\{X \geq x, Y \geq x\}} + \mathbb{1}_{\{X \geq x, Y < x\}} \leq \mathbb{1}_{\{Y \geq x\}} + \mathbb{1}_{\{Y < x \leq X\}},$$

so that

$$\mathbb{E}(X) \leq \mathbb{E}(Y) + \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{Y < x \leq X\}}\, dx d\,\mathbb{P}.$$

Now, changing variable x for ζ in the previous integral with $x = Y + \kappa_1\zeta$ we obtain

$$\int_\Omega \int_0^{+\infty} \mathbb{1}_{\{Y + \kappa_1\zeta \leq X\}}\, \kappa_1 d\zeta d\,\mathbb{P} = \int_0^{+\infty} \mathbb{P}(Y + \kappa_1\zeta \leq X)\, \kappa_1 d\zeta \leq \int_0^{+\infty} \kappa_1\kappa_2 e^{-\zeta} d\zeta = \kappa_1\kappa_2.$$

□

Acknowledgements

The authors would like to thank Anestis Antoniadis for helpfull discussions. This work was supported by Interuniversity Attraction Pole (IAP) research network in Statistics (http://www.stat.ucl.ac.be/IAP/).

References

Antoniadis, A., Besbeas, P. & Sapatinas, T. (2001), 'Wavelet shrinkage for natural exponential families with cubic variance functions', *Sankhya* **63**, 309–327.

Antoniadis, A. & Sapatinas, T. (2001), 'Wavelet shrinkage for natural exponential families with quadratic variance functions', *Biometrika* **88**, 805–820.

Baraud, Y. (2000), 'Model selection for regression on a fixed design.', *Probab. Theory Relat. Fields* **117**(4), 467–493.

Baraud, Y. & Birgé, L. (2005), Histogram type estimators based on nonnegative random variables, Technical report, http://math1.unice.fr/~baraud/publication/Poisson.pdf.

Barron, A., Birgé, L. & Massart, P. (1999), 'Risk bounds for model selection via penalization.', *Probab. Theory Relat. Fields* **113**(3), 301–413.

Barron, A. R. & Sheu, C.-H. (1991), 'Approximation of density functions by sequences of exponential families.', *Annals of Statistics* **19**(3), 1347–1369.

Besbeas, P., De Feis, I. & Sapatinas, T. (2004), 'A comparative simulation study of wavelet shrinkage estimators for Poisson counts.', *International Statistical Review* **72**, 209–237.

Birgé, L. & Massart, P. (2001), 'Gaussian model selection.', *J. Eur. Math. Soc. (JEMS)* **3**(3), 203–268.

Castellan, G. (1999), Modified Akaike's criterion for histogram density estimation, Preprint 61, Université Paris-Sud, http://www.math.u-psud.fr/~biblio/pub/1999/abs/ppo1999_61.html.

Castellan, G. (2003), 'Density estimation via exponential model selection.', *IEEE Transactions in Information Theory* **49**(8), 2052–2060.

Chui, C. (1992), *An introduction to Wavelets*, Academic Press, Boston.

Cohen, A., Daubechies, I. & Vial, P. (1993), 'Wavelets on the interval and fast wavelet transforms', *Appl. Comput. Harm. Analysis* **1**, 54–81.

Daubechies, I. (1992), Ten lectures on wavelets, *in* 'CBMS-NSF Regional Conference Series in Applied Mathematics', Vol. 61 of *SIAM*, Philadelphia.

Delyon, B. & Juditsky, A. (1995), Estimating wavelet coefficients, *in* Antoniadis & Oppenheim, eds, 'Wavelets and Statistics', Vol. 103 of *Lecture Notes in Statistics*, Springer-Verlag.

Donoho, D. (1993), Non-linear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data, *in* I. Daubechies, ed., 'Proceedings of Symposia in Applied Mathematics: Different Perspectives on Wavelets', Vol. 47, American Mathematical Society, San Antonio, pp. 173–205.

Fryzlewicz, P. & Nason, G. P. (2004), 'A Haar-Fisz algorithm for Poisson intensity estimation', *Journal of Computational and Graphical Statistics* **13**, 621–638.

Kolaczyk, E. (1997), 'Non-parametric estimation of Gamma-Ray burst intensities using Haar wavelets', *Astrophysical Journal* **493**, 340–349.

Kolaczyk, E. (1999*a*), 'Bayesian multiscale models for Poisson processes', *Journal of the American Statistical Association* **94**, 920–933.

Kolaczyk, E. (1999*b*), 'Wavelet shrinkage estimation of certain Poisson intensity signals using corrected threshold', *Statistica Sinica* **9**, 119–135.

Kolaczyk, E. & Nowak, R. (2004), 'Multiscale likelihood analysis and complexity penalized estimation', *Annals of Statistics* **32**(2), 500–527.

Kolaczyk, E. & Nowak, R. (2005), 'Multiscale Generalized Linear Models for nonparametric function estimation', *Biometrika* .

McCullagh, P. & Nelder, J. (1989), *Generalized linear models. 2nd ed.*

Nowak, R. & Baraniuk, R. (1999), 'Wavelet domain filtering for photon imaging systems', *IEEE Trans. Image Proc.* **8**, 666–678.

Reynaud-Bouret, P. (2003), 'Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities.', *Probability Theory and Related Fields* **126**(1), 103–153.

Sardy, S., Antoniadis, A. & Tseng, P. (2004), 'Automatic smoothing with wavelets for a wide class of distributions', *Journal of Computational and Graphical Statistics* **13**(2), 399–421.

Timmermann, K. & Nowak, R. (1999), 'Multiscale modeling and estimation of Poisson processes with applications to photon-limited imaging', *IEEE Trans. Info. Theor.* **133**, 846–862.

Triebel, H. (1983), *Theory of Function Spaces II*, Vol. 84 of *Monographs in Mathematics*, Birkhauser Verlag, Basel.

Tsybakov, A. (2004), *Introduction à l'estimation non-paramétrique*, Springer.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, John Wiley & Sons, New York.

Walter, G. G. (1994), *Wavelets and Other Orthogonal Systems with Applications*, CRC Press, Boca Raton, FL.

Wickerhauser, M. V. (1994), *Adapted Wavelet Analysis from Theory to Software*, AK Peters, Boston, MA, USA.