Maximum Likelihood Estimation in Poisson Regression via Wavelet Model Selection

Frédérique Leblanc, Frédérique Letué

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

27th April 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 among 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 for 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.

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 counts, which can be modeled through Poisson regression. Many authors have already discussed various nonparametric estimation procedures for such models. Whatever the estimation method, the drawback is to give estimators depending on some unknown smoothing parameter. Hence, this parameter should also be estimated with respect to the given data. Among available adaptative devices we focus here on the model selection methods

developped in (Barron, Birgé & Massart 1999) and in (Birgé & Massart 2001). One of the advantages of this approach is to provide nonasymptotic risk upper bounds for the selected estimator. Our models are constructed using wavelet basis and we choose the best model with respect to the Kullback Leibler risk.

We consider n independent copies $(Y_i, x_i)_{1\leq i\leq n}$, where the Y_i are Poisson variables of mean μ_i and the $(x_i)_{1\leq i\leq n}$ is a deterministic design in $[0,1]$. In the nonparametric regression context we want to explain the unknown μ_i as some general function μ of the regressor x_i. Since for counts the mean is positive, the model $\mu = \exp(f)$ ensures that μ remains positive and lets f be unrestricted (see for instance (McCullagh & Nelder 1989)).

Our goal is to estimate f under mild conditions, over some subspace generated by wavelet basis. Let us denote $(\phi_\lambda)_\lambda$ a wavelet basis of $L_2[0,1]$ and S_Λ the subspace of $L_2[0,1]$ generated by the set of wavelets $\{(\phi_\lambda), \lambda \in \Lambda\}$.

We define our models as linear subspaces of S_Λ. We construct a collection of maximum likelihood estimators within each model and we select one model, which mimics the best one with respect to the Kullback-Leibler risk.

More precisely, for any subset m of the larger index set Λ we define

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

where D_m is the cardinal of the subset m.

For each model, 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:

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

We compare the estimators of the collection, with respect to the Kullback-Leibler loss between the distributions modelized by f and h, denoted by:

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

Let us set $\bar{f}_m$ the function in S_m minimizing the Kullback-Leibler loss function,

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

The functions $K(f,\cdot)$ and $\gamma_n(\cdot)$ may not attain their infimum over the space S_m. In such a case, $\bar{f}_m$ or $\hat{f}_m$ may be undefined. Nevertheless we prove that on a large probability set, if one of them exists the other one also (see Lemma 6.1). Hence, in the sequel, we only consider the models m for which $|\hat{f}_m|_\infty \leq B$ or $|\bar{f}_m|_\infty \leq B$ for some given positive B.

Like with the usual quadratic risk, we can show that:

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

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

Let us consider 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

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

The ideal model m^* 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

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

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

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

In (Reynaud-Bouret 2003) penalized projection estimators for the intensity of Poisson processes are proposed. Moreover, oracle inequalities for the L_2-risk are provided for various kinds of basis (histograms, piecewise polynomials, Fourier or wavelets). In our work we restrict to the particular Poisson regression framework and we use penalized maximum likelihood estimators. We also furnish an oracle inequality for the Kullback-Leibler risk of the logarithm of the mean and we only consider wavelets basis.

In a recent paper (Baraud & Birgé 2005), histograms type estimators for non-negative random variables are studied, including Poisson variables and oracle inequalities are given for the Hellinger risk.

In (Kolaczyk & Nowak 2004) and (Kolaczyk & Nowak 2005) complexity penalized likelihood estimators are proposed in frameworks that include the Poisson model. Adaptivity and minimax near-optimality of the Hellinger risk are also stated in these works.

A quite complete presentation of wavelets methods for estimating the intensity of a Poisson process is given in (Besbeas, De Feis & Sapatinas 2004). The performances of the proposed estimators are evaluated by numerical experiments. Among these methods, one can cite the Anscombe transformation used in (Donoho 1993) and the Fisz transformation used in (Fryzlewicz & Nason 2004), which are applied on the data set to recover almost Gaussian observations and then allow the use of standard wavelets methods.

In (Kolaczyk 1997), (Kolaczyk 1999*b*) and (Nowak & Baraniuk 1999) wavelet shrinkage techniques are proposed to be applied directly to the given Poisson

process. Bayesian procedures have been also proposed in (Kolaczyk 1999*a*) and (Timmermann & Nowak 1999). Such methods have been also applied to a larger family of distributions containing the Poisson one in (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).

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 that are required to obtain the oracle inequality presented in Section 3. Then in Section 4 a lower bound for the Kullback-leibler risk is studied, 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 and of the lower bound. Technical lemmas are proven in 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, which 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 $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbb{Z}$, then constitutes an orthonormal basis for $L^2(\mathbb{R})$. For fixed $j \in \mathbb{Z}$, the $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$, $k \in \mathbb{Z}$ form 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 $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), k \in \mathbb{Z}$ 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 $P_j f = \sum_{k\in\mathbb{Z}} < f, \varphi_{j,k} > \varphi_{j,k}$ the orthogonal projection of f on the approximation space V_j, we have $P_{j+1} f = P_j f + \sum_{k\in\mathbb{Z}} < f, \psi_{j,k} > \psi_{j,k}$.

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 by $H(\nu)$ 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

Afterwards, we shall use the following notations:

$$\begin{aligned}
\forall f \in L_2[0,1] \quad &: \quad \|f\|_2^2 = \int_{[0,1]} f^2(t)dt \quad \text{and} \quad \|f\|_\infty = \sup_{x\in[0,1]} |f(x)|. \\
\forall (a_k)_k \in \mathbb{R}^{\,q} \quad &: \quad |a|_2^2 = \sum_k a_k^2 \quad \text{and} \quad |a|_\infty = \sup_k |a_k|. \\
\forall (b_k)_k \in \mathbb{R}^{\,n}, \quad \forall (c_k)_k \in \mathbb{R}^{\,n} \quad &: \quad < b, c >_n = \frac{1}{n}\sum_{k=1}^n b_k c_k \quad \text{and} \quad |b|_n^2 = < b, b >_n .
\end{aligned}$$

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

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

$$\begin{aligned} \Lambda_l &= \{(-1,0)\} \cup \{\lambda = (j,k); j = 0, ..., l-1; k = 0, ..., 2^j - 1\} \ \ \forall 1 \leq l \leq J; \\ \Lambda_0 &= \{(-1,0)\} \ \text{ and } \ \Lambda = \Lambda_J. \end{aligned} \tag{2.2}$$

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 $a \in \mathbb{R}^{2^J}$:

$$\| \sum_{\lambda \in \Lambda_l} a_\lambda \phi_\lambda \|_\infty \leq c(\psi) 2^{l/2} |a|_\infty. \tag{2.3}$$

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 $(\beta_{j,k})_k \in \mathbb{R}^{2^j}$, $\| \sum_{k=0,...,2^j-1} \beta_{j,k} \psi_{j,k} \|_\infty \leq c(\psi)/(1+\sqrt{2}) 2^{j/2} |\beta|_\infty$. Assertion (2.3) immediately follows since $\sum_{j=0}^{l-1} 2^{j/2} \leq (1+\sqrt{2}) 2^{l/2}$.

3 Wavelet model selection

3.1 Wavelet models

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

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

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

   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 criterion, this choice should be compared to adaptive linear procedure.

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

$$\begin{aligned} \mathcal{M}(L_n) = \{\Lambda_0\} \ \cup \ & \{m_{(l,\mathcal{I}_l)} = \{\Lambda_{l-1} \cup \{(l-1,k), k \in \mathcal{I}_l \\ & \mid \ \mathcal{I}_l \subset \{0, ..., 2^{l-1}-1\} \ \text{ and } \ \mathcal{I}_l \neq \emptyset\}, 1 \leq l \leq L_n\}, \end{aligned} \tag{3.2}$$

   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 $\binom{2^{l-1}}{d}$ 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 $a \in \mathbb{R}^{D_m}$*

$$\| \sum_{\lambda \in m} a_\lambda \phi_\lambda \|_\infty \leq b^{loc} D_m^{1/2} |a|_\infty.$$

For the first collection it is an immediate application of (2.3) 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 $< \cdot, \cdot >_n$.*

This assumption holds when the Haar basis is considered and for any deterministic design such that $x_i \in [(i-1)/n, i/n]$.

Next, for technical reasons, we 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 only consider the models up to the level $L_n < J/2 = \log n/(2 \log 2)$. Nevertheless, this condition being purely technical, in practice, we consider all the models up to the level $J = \log n / \log 2$.

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

Proposition 3.1. *Let Assumptions 1 and 2 hold and let $\tau \in]0,1[$ and B be some constants. If $|f|_\infty \leq B$ then for any $n \geq 1$, there exists some event Ω_n such that*

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

and for any model $m \in \mathcal{M}(L_n)$ such that $|\bar{f}_m|_\infty \leq B$,

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

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

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

$$\sum_{m\in\mathcal{M}(L_n)} e^{-\mathcal{L}_m D_m} \leq \Sigma < +\infty. \tag{3.3}$$

Define the penalty function as:

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

where $c_1 = (1+\alpha)^4$ and $c_2 = (1+\alpha)^4(1+6/\alpha)$.

For any f such that $|f|_\infty \leq B$ and $n \geq 1$, there exists some set Ω_n such that

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

and such that:

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

The constant B involved in the definition of the penalty function should be choosen as an upper bound of $|f|_\infty$. On the one hand we would like to choose it as large as possible to consider the largest possible model collection. On the other hand, the constants in the penalty term and in the residual term increase with B. In practice we will take for B an estimator of $|f|_\infty$.

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

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

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 (3.3) , so that 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 3. *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

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

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 (3.3) $\Sigma = 1/(\exp \mathcal{L} - 1)$. Herein, we recover the usual bound D_m/n up to a constant for the variance term in the risk decomposition.

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

   Assumption 4. *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 (3.3). Nevertheless, we take them as small as possible to avoid a too large risk bound in the oracle inequality. We can choose $L_m = \log n$ for all models m since

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

   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

$$\binom{2^{l-1}}{d} \leq \left(\frac{e2^{l-1}}{d}\right)^d = e^{d(1+\log(2^{l-1}/d))} \leq e^{d(1+2^{l-1}/d)} = e^D.$$

   Hence, Assumption 4 holds for $a = 1$ and $A = 1$. Moreover, we get easily:

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

   as soon as $n \geq 3$. Then in (3.3), $\Sigma = \frac{e/3}{1-e/3}$. Herein, we recover the bound $(D_m \log n)/n$ up to a constant for the variance 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 $k \in \mathbb{N}$ and $0 \leq \alpha < 1$. Let us consider the Hölder class $\mathcal{F} = \Sigma(\nu, L)$ of functions f defined over the interval $[0, 1]$ that 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 which 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 $\mathcal{C}^{\infty}(B)$ the space of functions uniformly bounded by B, where B is a positive constant.

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. *Let B and L be positive constants and $\nu > 1/2$. There exists a constant C, which only depends on B, L and ν, such that:*

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

where $v_n = n^{-\frac{\nu}{2\nu+1}}$.

The lower bound over the Sobolev class $H(\nu, L)$ is a direct consequence of the lower bound over the Hölder class $\Sigma(\nu, L)$ since that class contains the latter 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 chosen 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 $n^{-2\nu/(2\nu+1)}$ over the Sobolev class $H(\nu, L)$ and hence is minimax. Indeed, since $K(f, \bar{f}_{m_l}) \leq K(f, P_l f)$ due to definition of $\bar{f}_{m_l}$ and since over a Sobolev class $K(f, P_l f)$ is of the same order as $\|f - P_l f\|_2^2 = \mathcal{O}(2^{-2l\nu})$ 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-off 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 $\mathbb{E}_f(K(f, \hat{f}_{\hat{m}})\, \mathbf{1}\mathbf{l}_{\Omega_n})$ is bounded from above by $\mathcal{O}(n^{-2\nu/(2\nu+1)})$.

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 $\mathcal{O}(n^{-2\nu'/(2\nu'+1)})$ 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 $\mathcal{O}(-n^{2\nu'/(2\nu'+1)})$. Due to the choice of weights $L_m = \log n$, the selected estimator can only reach the rate $\mathcal{O}(n^{-2\nu'/(2\nu'+1)})$ up to a $\log n$ factor which is the usual price to pay for adaptivity.

5 A simulation study

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

1. Does the $e^{|\hat{f}_m|_\infty + B}$ 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 in (Reynaud-Bouret 2003) ?

5.1 Choice of the penalty functions

In the proof of the Theorem, it can be seen that the term $|\hat{f}_m|_\infty + B$ 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 $[0,1]$, $f_4 = \mathbf{1\!l}_{[1/4,1/2]} - \mathbf{1\!l}_{[1/2,3/4]}$. We use the Haar basis. In this case, such models may be described, for $J \geq 0$, by:
$$S_J^H = \{\sum_{j=0}^{J} \sum_{k=0}^{j-1} \beta_{j,k}\, \mathbf{1\!l}_{[k2^{-j},(k+1)2^{-j}[}, \beta \in \mathbb{R}^{2^J-1}\}.$$

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 $n = 128 = 2^7$ independent 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 it 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.

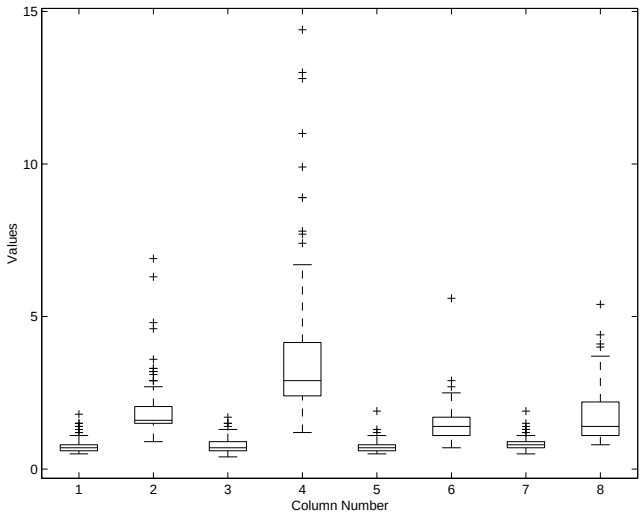

Figure 1: Distribution of the lowest constants selecting the "true" model: (1-2) $f = f_4$, (3-4) $f = 2f_4$, (5-6) $f = f_4/2$, (7-8) $f = f_{smooth}$ via (1,3,5,7) penalized maximum likelihood criterion, (2,4,6,8) penalized projection criterion.

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 the ones obtained in the PMLE case.

Therefore, we have decided to skip the $e^{|\hat{f}_m|_\infty + B}$ factor in the penalty term for the PMLE and to keep it for the PPE. More precisely, consequently, we shall take

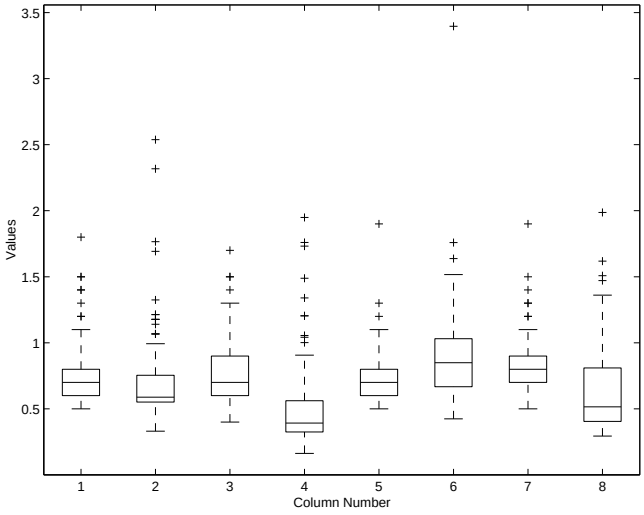

Figure 2: Distribution of the lowest constants selecting the "true" model: (1-2) $f = f_4$, (3-4) $f = 2f_4$, (5-6) $f = f_4/2$, (7-8) $f = f_{smooth}$ via (1,3,5,7) penalized maximum likelihood criterion, (2,4,6,8) penalized projection criterion, (2,4,6,8) contants are divided by $\exp(|f|_\infty)$.

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.

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 estimate f_{16} via the Haar basis on the models $S_J^H, 0 \leq J \leq 7$.

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.

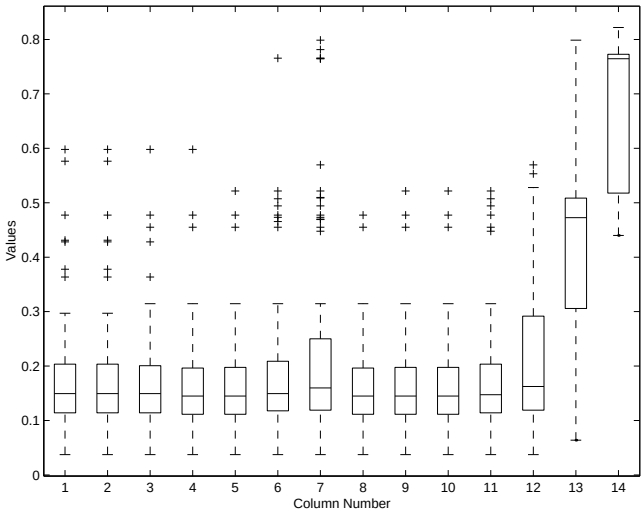

Figure 3: Distribution of the Average Square Error for $f = f_{16}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (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) $c_{P,1} = 3.4$.

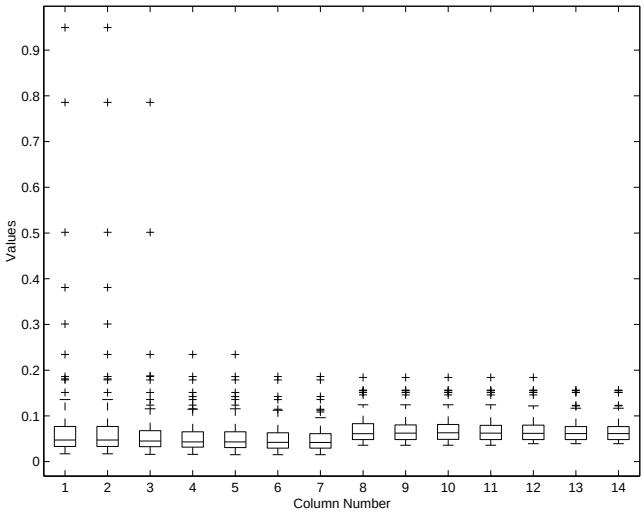

Figure 4: Distribution of the Average Square Error for $f = f_{smooth}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (5) $c_{ML,0.95} = 1.2$, (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) $c_{P,1} = 3.4$.

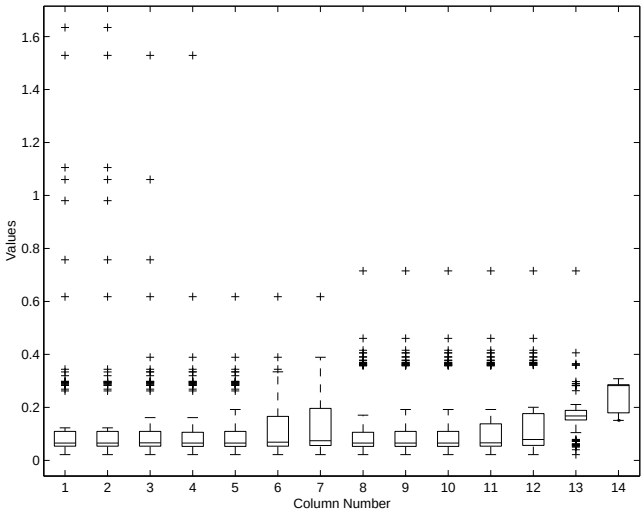

Figure 5: Distribution of the Kullback-Leibler divergence for $f = f_{16}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (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) $c_{P,1} = 3.4$.

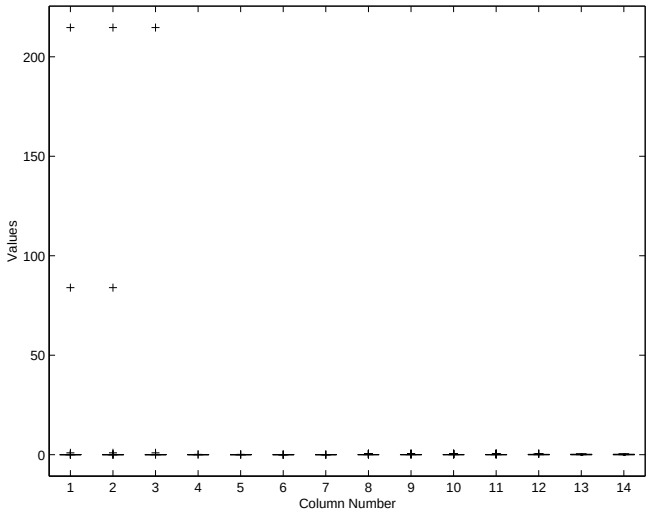

Figure 6: Distribution of the Kullback-Leibler divergence for $f = f_{smooth}$ of the (1-7) penalized maximum likelihood estimator, (8-14) penalized projection estimator, with constant in the penalty term: (1) $c_{ML,0.75} = 0.9$, (2) $c_{ML,0.80} = 0.9$, (3) $c_{ML,0.85} = 1.0$, (4) $c_{ML,0.90} = 1.1$, (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) $c_{P,1} = 3.4$.

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

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

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,

	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 criterion and by the Projection criterion over the 100 simulations for different sample sizes ($n = 128, 256, 512$): (a) $f = f_4$, Haar basis, (b) f polynomial, Symmlet basis.

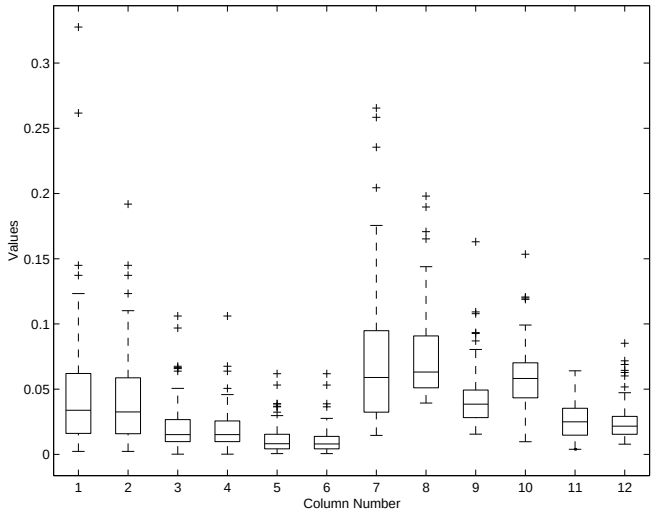

Figure 7: Distribution of the Average Square Error for (1-6) $f = f_4$ and (7-12) $f = f_{smooth}$ for different sample sizes: (1,2,7,8) $n = 128$, (3,4,9,10) $n = 256$, (5,6,11,12) $n = 512$, (1,3,5,7,9,11) PMLE, (2,4,6,8,10,12) PPE.

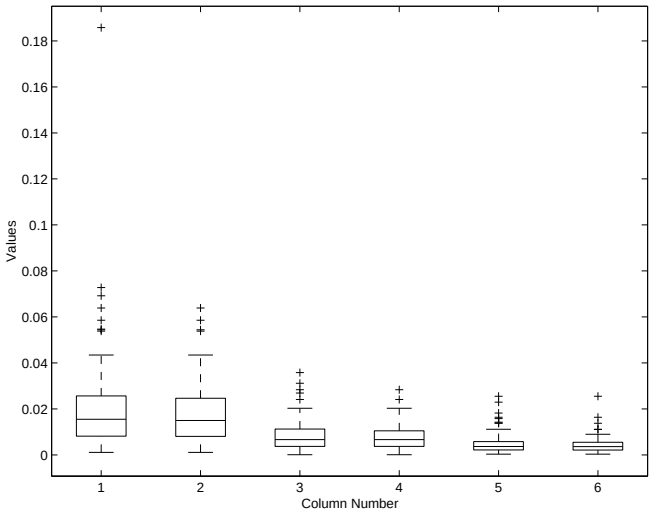

Figure 8: Distribution of the Kullback-Leibler divergence for $f = f_4$ for different sample sizes: (1,2) $n = 128$, (3,4) $n = 256$, (5,6) $n = 512$, (1,3,5) PMLE, (2,4,6) PPE.

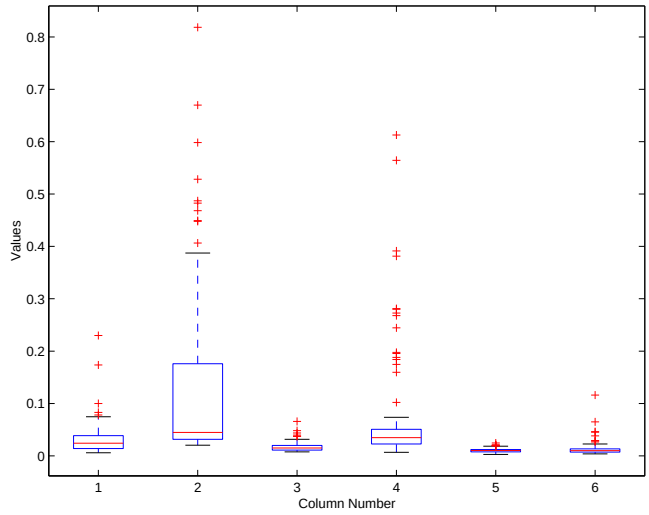

Figure 9: Distribution of the Kullback-Leibler divergence for $f = f_{smooth}$ for different sample sizes: (1,2) $n = 128$, (3,4) $n = 256$, (5,6) $n = 512$, (1,3,5) PMLE, (2,4,6) PPE.

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 higher in the maximum likelihood case than in the projection case, since the latter provides an explicit estimator whereas the first one requires the 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 criterion.

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

Let us denote $\varepsilon_i = Y_i - \mathbb{E}(Y_i) = Y_i - e^{f_i}$. By elementary algebric manipulation we have:

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

Next, due to definitions (1.3) of $\hat{m}$ and (1.1) of $\hat{f}_m$, we have:

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

Hence $\gamma_n(\hat{f}_{\hat{m}}) - \gamma_n(\bar{f}_m)$ is bounded by $\mathrm{pen}(m) - \mathrm{pen}(\hat{m})$. Thus, when substituting $\gamma_n(\hat{f}_{\hat{m}}) - \gamma_n(\bar{f}_m)$ by this latter upper bound in the decomposition of $K(f, \hat{f}_{\hat{m}})$, we get:

$$K(f, \hat{f}_{\hat{m}}) \leq K(f, \bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + < \hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}, \varepsilon >_n + < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n \quad (6.1)$$

Furthermore, since for any numbers a, b and any positive θ, $2ab \leq \theta a^2 + \frac{1}{\theta} b^2$, we have

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

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 θ_1 is a positive number. Next, substituting in (6.1) the upper bound given in (7.5) for $|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_n^2$, we obtain:

$$\begin{aligned} K(f, \hat{f}_{\hat{m}}) \leq K(f, \bar{f}_m) &+ \text{pen}(m) - \text{pen}(\hat{m}) \\ &+ \frac{\theta_1}{2}\chi_n^2(\hat{m}) + \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) + < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n . \end{aligned} \tag{6.2}$$

Next, let A be some positive number and ρ such that

$$1 - \theta < 2\rho < \theta, \tag{6.3}$$

where θ is defined in Assumption 2. In order to control the terms $\chi_n^2(\hat{m})$ and $|\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty$, we introduce the set $\Omega_n[A]$:

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

The set Ω_n of the theorem will be defined later as $\Omega_n[A]$ for a particular value of A. The following proposition, which is the key to state the oracle inequality, gives an upper bound for 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 (6.18). 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*

$$\chi_n(\hat{m})\, \mathbb{1}\!\!1_{\Omega_n[A]} \leq (1+\alpha) e^{|f|_\infty/2} \left(\left(\frac{D_{\hat{m}}}{n} \right)^{1/2} + \left(\frac{12 x_{\hat{m}}}{n} \right)^{1/2} \right). \tag{6.4}$$

The proof of this proposition is postponed in Section 6.3.2. It is an application of a concentration inequality for Poisson process that can be found in (Reynaud-Bouret 2003), to the case of independant Poisson variables.

The following proposition provides an upper bound for the last term of inequality (6.2).

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*

$$\begin{aligned} < \bar{f}_{\hat{m}} - \bar{f}_m, \varepsilon >_n \;\leq\; & \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} \left(1 + \frac{1}{\theta_2} \right) \frac{y_{\hat{m}}}{n} + \theta_2 K(f, \bar{f}_{\hat{m}}) \\ & + \frac{e^{|\bar{f}_m - f|_\infty}}{2} \left(1 + \frac{1}{\theta_3} \right) \frac{y_m}{n} + \theta_3 K(f, \bar{f}_m). \end{aligned} \tag{6.5}$$

The proof of the Bernstein type inequality (6.5) is given in Section 6.3.3.

Gathering (6.2), (6.4) and (6.5), on the set $\Omega_n^1 \cap \Omega_n^2$ we have:

$$
\begin{aligned}
& K(f, \hat{f}_{\hat{m}})\, \mathbf{1\!l}_{\Omega_n[A]} \leq \\
& \mathbf{1\!l}_{\Omega_n[A]} \left\{ K(f, \bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \right. \\
+ \quad & \frac{\theta_1}{2}(1+\alpha)^2 e^{|f|_\infty} \left(\left(\frac{D_{\hat{m}}}{n} \right)^{1/2} + \left(\frac{12 x_{\hat{m}}}{n} \right)^{1/2} \right)^2 + \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) \\
+ \quad & \left. \left(1 + \frac{1}{\theta_2} \right) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty} y_{\hat{m}}}{2n} + \theta_2 K(f, \bar{f}_{\hat{m}}) + \left(1 + \frac{1}{\theta_3} \right) \frac{e^{|\bar{f}_m - f|_\infty} y_m}{2n} + \theta_3 K(f, \bar{f}_m) \right\}
\end{aligned}
\tag{6.6}
$$

Let us choose $x_{m'} = y_{m'} = \mathcal{L}_{m'} D_{m'} + \zeta$. Since for any positive θ and any a and b $(a+b)^2 \leq (1+\theta)a^2 + (1+1/\theta)b^2$, for any positive θ_4 we get:

$$
\begin{aligned}
\left(\left(\frac{D_{\hat{m}}}{n} \right)^{1/2} + \left(\frac{12 x_{\hat{m}}}{n} \right)^{1/2} \right)^2 & \leq (1+\theta_4) \frac{D_{\hat{m}}}{n} + \left(1 + \frac{1}{\theta_4} \right) \frac{12(\mathcal{L}_{\hat{m}} D_{\hat{m}} + \zeta)}{n} \\
& \leq (1+\theta_4) \left(1 + \frac{12 \mathcal{L}_{\hat{m}}}{\theta_4} \right) \frac{D_{\hat{m}}}{n} + \left(1 + \frac{1}{\theta_4} \right) \frac{12\zeta}{n}.
\end{aligned}
\tag{6.7}
$$

Hence, when substituting (6.7) in inequality (6.6) and factorizing the terms $K(f, \bar{f}_m)$, $\frac{D_{\hat{m}}}{n}$ and $\frac{\zeta}{n}$, we obtain:

$$
\begin{aligned}
& K(f, \hat{f}_{\hat{m}})\, \mathbf{1\!l}_{\Omega_n[A]} \leq \\
& \mathbf{1\!l}_{\Omega_n[A]} \left[(1+\theta_3) K(f, \bar{f}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \left(1 + \frac{1}{\theta_3} \right) \frac{e^{|\bar{f}_m - f|_\infty} \mathcal{L}_m D_m}{2n} \right. \\
+ \quad & \left\{ \frac{\theta_1}{2}(1+\alpha)^2 e^{|f|_\infty} (1+\theta_4) \left(1 + \frac{12\mathcal{L}_{\hat{m}}}{\theta_4} \right) + \left(1 + \frac{1}{\theta_2} \right) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} \mathcal{L}_{\hat{m}} \right\} \frac{D_{\hat{m}}}{n} \\
+ \quad & \left\{ 6\theta_1 (1+\alpha)^2 \left(1 + \frac{1}{\theta_4} \right) e^{|f|_\infty} + \left(1 + \frac{1}{\theta_2} \right) \frac{e^{|\bar{f}_{\hat{m}} - f|_\infty}}{2} \right\} \frac{\zeta}{n} \\
+ \quad & \left. \frac{e^{|\bar{f}_{\hat{m}}|_\infty + |\hat{f}_{\hat{m}} - \bar{f}_{\hat{m}}|_\infty}}{\theta_1} K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) + \theta_2 K(f, \bar{f}_{\hat{m}}) \right].
\end{aligned}
\tag{6.8}
$$

Now, we take $0 < \theta_2 < 1$ 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 inequality (6.8) and noticing that $(1 + 1/\theta_2) \leq 2/\theta_2$, we have:

$$(1-\theta_2)K(f,\hat{f}_{\hat{m}})\,\mathbf{1l}_{\Omega_n[A]} \le \mathbf{1l}_{\Omega_n[A]}\left\{(1+\theta_3)K(f,\bar{f}_m)+\mathrm{pen}(m)-\mathrm{pen}(\hat{m})\right.$$
$$\left.+(1+\frac{1}{\theta_3})\frac{e^{|\bar{f}_m-f|_\infty}\mathcal{L}_m D_m}{2n}+T_1(\hat{m})\frac{D_{\hat{m}}}{n}+T_2(\hat{m})\frac{\zeta}{n}\right\}, \quad (6.9)$$

where

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

Next, on the one hand, we have 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

$$(-\mathrm{pen}(\hat{m})+T_1(\hat{m})\frac{D_{\hat{m}}}{n})\,\mathbf{1l}_{\Omega_n[A]} \le 0, \quad (6.10)$$

and on the other hand, we have to bound $T_2(\hat{m})$ by a deterministic constant. To this end, the following proposition enables us to bound the term $|\bar{f}_{m'}-f|_\infty$ for $m'=\hat{m}$ or $m'=m$.

Proposition 6.3. *Let Assumption 2 holds. Set $\tau \in]0,1[$. If*

$$A \le \frac{\tau}{4e^{1+B}}, \quad (6.11)$$

then, on the set $\Omega_n[A]$, for any model $m' \in \mathcal{M}(L_n)$ such that $|\hat{f}_{m'}|_\infty \le B$ or $|\bar{f}_{m'}|_\infty \le B$, we have :

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

Due to definition of $\hat{m}$ (1.3), the previous result can be applied to $\hat{m}$. In the sequel of the proof, we put $A=\inf(\frac{12\alpha e^{-|f|_\infty}}{\kappa(\alpha)}, \frac{\tau}{4e^{1+B}})$ 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:

$$\begin{aligned} (|\bar{f}_{\hat{m}}|_\infty+|\hat{f}_{\hat{m}}-\bar{f}_{\hat{m}}|_\infty)\,\mathbf{1l}_{\Omega_n} &\le (|\hat{f}_{\hat{m}}|_\infty+2|\hat{f}_{\hat{m}}-\bar{f}_{\hat{m}}|_\infty)\,\mathbf{1l}_{\Omega_n} \le (|\hat{f}_{\hat{m}}|_\infty+\tau)\,\mathbf{1l}_{\Omega_n} \\ |\bar{f}_{\hat{m}}-f|_\infty\,\mathbf{1l}_{\Omega_n} &\le (|f|_\infty+|\bar{f}_{\hat{m}}|_\infty)\,\mathbf{1l}_{\Omega_n} \le (B+|\hat{f}_{\hat{m}}|_\infty+\tau/2)\,\mathbf{1l}_{\Omega_n} \\ &\le (B+|\hat{f}_{\hat{m}}|_\infty+\tau)\,\mathbf{1l}_{\Omega_n}. \end{aligned} \quad (6.12)$$

Moreover, on Ω_n we also have $|\bar{f}_{\hat{m}}|_\infty\,\mathbf{1l}_{\Omega_n} \le (|\hat{f}_{\hat{m}}-\bar{f}_{\hat{m}}|_\infty+|\hat{f}_{\hat{m}}|_\infty \le (\tau/2+B)\,\mathbf{1l}_{\Omega_n}$. Hence we obtain:

$$|\bar{f}_{\hat{m}}-f|_\infty\,\mathbf{1l}_{\Omega_n} \le (|\bar{f}_{\hat{m}}|_\infty+|f|_\infty)\,\mathbf{1l}_{\Omega_n} \le (\tau/2+2B)\,\mathbf{1l}_{\Omega_n}, \quad (6.13)$$

which provides an upper bound for $T_2(\hat{m})$. Now, we choose $\theta_2=1/(1+\alpha)$, $\theta_4=\alpha$,

and α such that

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

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 $C'(B,\alpha)$.

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

$$\begin{aligned}&\frac{\alpha}{1+\alpha}K(f,\hat{f}_{\hat{m}})\,\mathbf{1l}_{\Omega_n} \leq\\ &\quad\mathbf{1l}_{\Omega_n}\left\{(1+\theta_3)K(f,\bar{f}_m)+\mathrm{pen}(m)+\left(1+\frac{1}{\theta_3}\right)\frac{e^{|\bar{f}_m-f|_\infty}\mathcal{L}_m D_m}{2n}+C'(\bar{B},\alpha)\frac{\zeta}{n}\right\}.\end{aligned} \tag{6.14}$$

Let us now suppose that $|\bar{f}_m|_\infty \leq B$ and choose $\theta_3 = \alpha$. Then the third term of the previous right hand side can be bounded as follows:

$$\mathbf{1l}_{\Omega_n}\left(1+\frac{1}{\theta_3}\right)\frac{e^{|\bar{f}_m-f|_\infty}\mathcal{L}_m D_m}{2n} \leq \mathbf{1l}_{\Omega_n}\left(1+\frac{1}{\alpha}\right)\frac{e^{|\hat{f}_m|+B+\tau/2}\mathcal{L}_m D_m}{2n} \leq \mathrm{pen}(m)\,\mathbf{1l}_{\Omega_n}.$$

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 7.5 with

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

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. *Let Assumptions 1 and 2 hold. For any positive A, there exists some positive constant c, which depends only on $|f|_\infty$, b^{loc} and A, 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}, 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 a 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 $\tau/2+B$ on the set Ω_n and setting $\theta_2=\frac{e^{\tau/2+\bar{B}}}{\theta_1}$, we have

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

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

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

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

6.3.1 Concentration inequalities

The proofs of Propositions 6.1, 6.2 and 6.4 highly depend on concentration inequalities established in (Reynaud-Bouret 2003). When applying these results to the following Poisson process N we obtain concentration inequalities for sequences of Poisson variables of mean $e^{f(x_i)}$.

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 $d\nu=\sum_{i=1}^n e^{f(x_i)}\,\mathbb{1}\!\!1_{I_i}\,d\mu$. Let N be a Poisson process with inhomogeneous intensity $d\nu$. Then, the random variables $\int \mathbb{1}\!\!1_{I_i}\,dN$ have Poisson distributions with parameter $\nu(I_i)=e^{f(x_i)}$.

For any $h\in\mathbb{R}^n$, let us define $f_h=\sum_{i=1}^n h_i\,\mathbb{1}\!\!1_{I_i}$. Then, $\int f dN=\sum_{i=1}^n h_i\int \mathbb{1}\!\!1_{I_i}\,dN$ has the same distribution as $\sum_{i=1}^n h_i Y_i$.

So, inequalities given in (Reynaud-Bouret 2003) can be re-enunced in this way:

Theorem 6.1. *Bernstein's inequality :*

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

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

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

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n h_i\varepsilon_i \geq (2u\sum_{i=1}^n e^{f(x_i)}h_i^2)^{1/2} + |h|_\infty u/3\right) &\leq e^{-u}, \\
\mathbb{P}\left(|\sum_{i=1}^n h_i\varepsilon_i| \geq (2u\sum_{i=1}^n e^{f(x_i)}h_i^2)^{1/2} + |h|_\infty u/3\right) &\leq 2e^{-u}. \qquad (6.16)
\end{aligned}$$

We will also need the following theorem:

Theorem 6.2. *Let S be some finite dimensional linear subspace of $\mathbb{L}_2$ and $(\phi_\lambda)_{\lambda=1,\ldots,D}$ 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

$$\Omega_S(\alpha) = \left\{|\sum_{\lambda=1,\ldots,D} < \phi_\lambda, \varepsilon >_n \phi_\lambda|_\infty \leq \frac{12\alpha M_S}{\kappa(\alpha)}\right\}, \qquad (6.17)$$

where

$$\kappa(\alpha) = 5/4 + 32/\alpha. \qquad (6.18)$$

Then, for any positive α and x,

$$\mathbb{P}\left(\chi_n(S)\,\mathbb{1}\!\mathrm{l}_{\Omega_S(\alpha)} \geq (1+\alpha)\left(\mathbb{E}(\chi_n^2(S))^{1/2} + (12M_S x/n)^{1/2}\right)\right) \leq e^{-x}. \qquad (6.19)$$

6.3.2 Proof of Proposition 6.1

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

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

where $\Omega_{S_{m'}}$ is defined by (6.17). From (6.19), 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,

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

Moreover, we have:

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

Noticing that $M_{m'} \leq e^{|f|\infty}$ for any model $m' \in \mathcal{M}(L_n)$, (6.4) 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 (6.16) 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 that $ab \leq \theta_2 a^2/2 + b^2/(2\theta_2)$ with $a = (2K(f, \bar{f}_{m'}))^{1/2}$ and $b = (\frac{e^{|\bar{f}_{m'} - f|_\infty} y_{m'}}{n})^{1/2}$, we get:

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

for some positive constant θ_2. Since this is true for any model m', this is in particular true for $m' = \hat{m}$ and for $m' = m$ with θ_2 replaced by θ_3. To conclude, (6.5) 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

To prove Proposition 6.3 we state the following preliminary lemma adapted from the results given in (Barron & Sheu 1991) and (Castellan 2003). Let G be the function from $\mathbb{R}^{D_m}$ to $\mathbb{R}^{D_m}$ which $\lambda - th$ 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 $\bar{\delta}_m$ be vectors in $\mathbb{R}^{D_m}$ with λ-th coordinates

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

Lemma 6.1. *1. Suppose that there exists $\bar{\beta} \in \mathbb{R}^{D_m}$ such that $G(\bar{\beta}) = \bar{\delta}_m$ and let us denote $\bar{f}_m = \sum_{\lambda \in m} \bar{\beta}_\lambda \phi_\lambda$. For any $\tau \in]0,1[$, if*

$$|\hat{\delta}_m - \bar{\delta}_m|_2 \leq \frac{\tau}{(4b^{loc} D_m^{1/2} e^{1+|\bar{f}_m|_\infty})}, \tag{6.20}$$

then equation $G(\beta) = \hat{\delta}_m$ admits a solution $\hat{\beta}$. Hence when setting $\hat{f}_m =$

$\sum_{\lambda\in m}\hat{\beta}_\lambda\phi_\lambda$ *we get*

$$|\hat{f}_m - \bar{f}_m|_\infty \leq \tau/2.$$

2. *Suppose that there exists* $\hat{\beta} \in \mathbb{R}^{D_m}$ *such that* $G(\hat{\beta}) = \hat{\delta}_m$ *and let us denote* $\hat{f}_m = \sum_{\lambda\in m}\hat{\beta}_\lambda\phi_\lambda$. *For any* $\tau \in]0,1[$, *if*

$$|\hat{\delta}_m - \bar{\delta}_m|_2 \;\leq\; \frac{\tau}{(4b^{loc}D_m^{1/2}e^{1+|\hat{f}_m|_\infty})}, \tag{6.21}$$

then equation $G(\beta) = \bar{\delta}_m$ *admits a solution* $\bar{\beta}$. *Hence when setting* $\bar{f}_m = \sum_{\lambda\in m}\bar{\beta}_\lambda\phi_\lambda$ *we get*

$$|\hat{f}_m - \bar{f}_m|_\infty \leq \tau/2.$$

Proof. For any $\delta \in \mathbb{R}^{D_m}$, define F_δ as the function from $\mathbb{R}^{D_m}$ to $\mathbb{R}$ which 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 $G(\beta) = \hat{\delta}_m$ comes to minimize $F_{\hat{\delta}_m}$. Now for any $\beta \in \mathbb{R}^{D_m}$,

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

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

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

Due to (6.21) and since $0 < \tau < 1$, $2b^{loc}D_m^{1/2}e^{1+|\bar{f}_m|_\infty}|\hat{\delta}_m - \bar{\delta}_m|_2 < \tau < 1$, hence for any β on the sphere $F_{\hat{\delta}_m}(\beta) - F_{\hat{\delta}_m}(\bar{\beta}) > 0$. Moreover, the function $F_{\hat{\delta}_m}(\cdot) - F_{\hat{\delta}_m}(\bar{\beta})$ 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 $|\hat{\beta} - \bar{\beta}|_2 < 2e^{\tau + |\bar{f}_m|_\infty}|\hat{\delta}_m - \bar{\delta}_m|_2$.

Thus, from Lemma 7.4,

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

The proof of the second point is similar, exchanging $\hat{\delta}_m$ (resp. $\hat{f}_m$) with $\bar{\delta}_m$ (resp. $\bar{f}_m$). □

We let us now prove the proposition. On the set $\Omega_n[A]$, for any $m' \subset \Lambda_n^*$, we have:

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

Due to Assumption 2 and to (6.3), 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), if m' is such that $|\hat{f}_{m'}|_\infty \leq B$, then $A \leq \frac{\tau}{4e^{1+|\hat{f}_{m'}|_\infty}}$. Hence,

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

If m' is such that $|\bar{f}_{m'}|_\infty \leq B$, then

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

In both cases, when applying Lemma 6.1, we get the result.

6.3.5 Proof of Proposition 6.4

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

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

Using Bernstein's inequality (6.15) and setting $\xi(A) = \frac{An^{-\rho}}{b^{loc} D_{\Lambda_n^*}^{1/2}}$, 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 $\lambda \in \Lambda_n^*$:

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

so that

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

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

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

where C is a positive constant depending on A, $|f|_\infty$ and b^{loc} but not on n. Since $\theta - 2\rho > 0$ from (6.3), $n^{3-\theta} \exp(-Cn^{\theta-2\rho})$ 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 $\mathcal{F}_M$ denote a finite subset of cardinality $M+1$ of $\mathcal{F} \cap \mathcal{C}^\infty(B)$, then we have for any estimator $\hat{f}_n$ of f:

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

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 $\hat{f}_n \in \mathcal{C}^\infty(B)$:

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

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

$$\sup_{f \in \mathcal{F}_M} \mathbb{E}_f((K(f, \hat{f}_n) v_n^{-2}) \quad \geq \quad \frac{e^{-3B}}{2} \max_{k=0,\ldots,M} \mathbb{P}_{f_k}(|\hat{f}_n - f_k|_n v_n^{-1} > \xi)\xi^2. \quad (6.22)$$

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

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 the sake of simplicity, we use a simplified version of that 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 6.2. *Suppose that the elements $f_0, \ldots, f_M \in \mathcal{F}_M, M \geq 2$ are such that*
a) For all k, k' such that $0 \leq k < k' \leq M$, the following inequality holds:

$$|f_k - f_{k'}|_n \geq 2s_n > 0; \quad (6.23)$$

b) For any $k = 1, \ldots, M$ the Kullback-Leibler divergence between the likelihoods

under f_k and f_0 satisfies

$$\frac{1}{M}\sum_{k=1}^{M} nK(f_k, f_0) \leq a\log(M) \tag{6.24}$$

where $0 < a < 1/10$.

Then for any estimator $\hat{f}_n$

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

We construct now a convenient set of functions $\mathcal{F}_M$ that will verify Assumptions (6.23) and (6.24) for some large enough M, which 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 6.3. 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 the sequence 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 to construct a subset $\mathcal{F}_M$ of functions satisfying (6.24) and (6.23).

Lemma 6.3. *Let $\Phi \in \Sigma(\nu, 1)$ be compactly supported over $[-1/2, 1/2]$, such that $\|\Phi\|_\infty \leq 8^\nu B/L$ and $\|\Phi\|_2^2 < \log 2/(60L^2)$. 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) $f_{jn} \in \mathcal{F} \cap \mathcal{C}^\infty(B)$.

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 $\|\Phi\|_\infty \leq 8^\nu B/L$ and due to *i)*, f_{jn} is obviously bounded by B.

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 possible binary 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 $0 \leq k < k' \leq M$

$$\begin{aligned} \rho_H(\bar{w}^k, \bar{w}^{k'}) = card\{l : 1 \leq l \leq m, w_l^k \neq w_l^{k'}\} &\geq m/8 \\ \text{and} \quad 8\log(M)/\log(2) &\geq m. \end{aligned} \tag{6.25}$$

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 $k = 0, ..., M$, $f_k \in \mathcal{F} \cap \mathcal{C}^\infty(B)$ and $\|f_k\|_\infty \leq Lm^{-\nu}\|\Phi\|_\infty$. Let us check now that functions f_k also satisfy conditions (6.23) and (6.24), for n and M large enough.

When using Lemma 6.3 and the Varshanov-Gilbert lower bound for ρ_H given in (6.25), we get for any $0 \leq k < k' \leq M$, and for any n and $m \geq 8$:

On the one hand,

$$\begin{aligned} |f_k - f_{k'}|_n^2 &= \sum_{j=1}^m (w_j^k - w_j^{k'})^2 |f_{jn}|_n^2 \geq \sum_{j=1}^m (w_j^k - w_j^{k'})^2 (\|f_{jn}\|^2 - \|f_{jn}\|_\infty \|f'_{jn}\|_\infty n^{-1}) \\ &= \rho_H(\bar{w}_k, \bar{w}_{k'})(L^2\|\Phi\|_2^2 m^{-(2\nu+1)} - L^2 m^{-2\nu+1}\|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\ &\geq m^{-2\nu} L^2 \frac{\|\Phi\|_2^2}{8} R_{m,n} \quad \text{with} \quad R_{m,n} = 1 - \frac{\|\Phi\|_\infty \|\Phi'\|_\infty m^2}{L^2\|\Phi\|_2^2 n}; \end{aligned} \tag{6.26}$$

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

$$\begin{aligned} nK(f_k, f_0) &\leq \frac{1}{2} e^{\|f_k\|_\infty + \|f_k - f_0\|_\infty} n |f_k - f_0|_n^2 \leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \left(\sum_{j=1}^{m} (w_j^k)^2 |f_{jn}|_n^2 \right) \\ &\leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \sum_{j=1}^{m} (\|f_{jn}\|_2^2 + L^2 m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\ &\leq \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} n \sum_{j=1}^{m} (L^2 \|\Phi\|_2^2 m^{-(2\nu+1)} + L^2 m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty n^{-1}) \\ &\leq L^2 \|\Phi\|_2^2 \frac{1}{2} e^{2\|\Phi\|_\infty L m^{-\nu}} m \left(n m^{-(2\nu+1)} + \frac{m^{-2\nu+1} \|\Phi\|_\infty \|\Phi'\|_\infty}{\|\Phi\|_2^2} \right) \\ &\leq L^2 \|\Phi\|_2^2 \frac{4 \log M}{\log 2} P_{m,n} \end{aligned} \tag{6.27}$$

with

$$P_{m,n} = e^{2\|\Phi\|_\infty L m^{-\nu}} \left(\frac{n}{m^{2\nu+1}} + \frac{\|\Phi\|_\infty \|\Phi'\|_\infty}{\|\Phi\|_2^2 m^{2\nu-1}} \right).$$

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 (6.26) and (6.27):

$$|f_k - f_{k'}|_n \geq n^{\frac{-\nu}{2\nu+1}} \frac{L\|\Phi\|_2}{4\sqrt{2}} \quad \text{and} \quad nK(f_k, f_0) \leq 3L^2 \|\Phi\|_2^2 \log M.$$

Hence Assumptions (6.23) and (6.24) are obtained for $s_n = n^{-\nu/(2\nu+1)}(L\|\Phi\|_2)/(8\sqrt{2})$ and $a = 3L^2\|\Phi\|_2^2$, for n and M large enough, and Lemma 6.2 provides the estimate:

$$\max_{k=0,\ldots,M} P_{f_k}(|\hat{f}_n - f_k|_n > \xi v_n) \geq 0.04 \quad \text{for} \quad \xi = L\|\Phi\|_2/(8\sqrt{2}).$$

To end the proof, we substitute the previous lower bound in (6.22) which provides the result given in Theorem 4.1 with $C = 0.04 e^{-3B} L^2 \|\Phi\|_2^2/256$.

7 Appendix

7.1 Technical lemmas

Afterwards, for the sake of simplicity, we put $D = D_m$.

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 $h \in S_m$,*

$$\sum_{i=1}^{n} Y_i h_i = \sum_{i=1}^{n} e^{\hat{f}_{m,i}} h_i \tag{7.1}$$

$$\sum_{i=1}^{n} e^{f_i} h_i = \sum_{i=1}^{n} e^{\bar{f}_{m,i}} h_i. \tag{7.2}$$

In particular,

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

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 minimizer 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,*

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

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

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 $x \in \mathbb{R}$, $\frac{x^2}{2}e^{-|x|} \leq e^x - 1 - x \leq \frac{x^2}{2}e^{|x|}$, we have :

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

Moreover, for any i, $\exp(-|h_i - f_i|) \geq \exp(-|h-f|_\infty)$ and $\exp(|h_i - f_i|) \leq \exp(|h-f|_\infty)$. Hence, substituting these bounds in (7.6) we obtain (7.5). Next, since $\exp(-|f|_\infty) \leq \exp(f_i) \leq \exp(|f|_\infty)$ we have $\exp(-|f|_\infty)|h-f|_n^2 \leq V_f(f,h) \leq \exp(|f|_\infty)|h-f|_n^2$ 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 $h = \sum_{\lambda \in m}\beta_\lambda\phi_\lambda \in S_m$:*

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

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

Proof. Since $|\beta-\bar{\beta}|_\infty \leq |\beta-\bar{\beta}|_2$, 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:

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

The lower bound is deduced in the same way. □

7.1.4 Integration lemma

Lemma 7.5. *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. & 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. (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.