Penalized maximum likelihood estimation with l^1 penalty

Loubes Jean-Michel

CNRS - Institut de Matématiques et de modélisation, UMR 5149
Equipe de Probabilités et Statistique - Université Montpellier 2
Place E. Bataillon, 34095 Montpellier cedex

Abstract

We focus on density estimation using penalized loglikelihood method. We aim at building an adaptive estimator in the sense that it converges at the optimal rate of convergence without prior knowledge of its regularity. For this, we penalize the log-likelihood by a function, which depends on the roughness of the density: the l^1 norm of the wavelet coefficients of the log-density. In this setting, we prove adaptivity for l^2 norm over a certain class of sets, Besov spaces.

Key words: Density estimation, Penalized Maximum Likelihood, Complexity regularization, Wavelet Bases

1 Introduction and Notations

Consider the estimation of a probability density f_0 on a bounded domain $\mathcal{X}$ based on independent samples X_i, $i = 1, \ldots, n$. In classical parametric estimation, some parametric model is often assumed of $f(x)$ and the model is fitted to the observations by maximum likelihood. But when few is known about the law of the data, nonparametric method are to be considered. In this work, we propose a new method, a penalized maximum likelihood with a l^1 penalty. We build an adaptive estimator, in the sense that it achieves the optimal rate of convergence while built without any prior smoothness assumption.

Email address: Jean-Michel.Loubes@math.univ-montp2.fr (Loubes Jean-Michel).

Nonparametric density estimation has been tackled by several authors. Indeed, there are many routes to density estimation, including kernel methods [4], smoothed histograms [5], spline methods in [23], wavelet basis in [15] or maximum likelihood methods in [17], [20] [9], [10] or [11]. In this work, we will concentrate on maximum penalized likelihood method. Good and Gaskins in [17] were the first to introduce the idea of roughness penalty estimation. Indeed, a naive application of maximum likelihood method leads to a too rough estimates since if we maximize without constraint $L_n(f) = \sum_{i=1}^n \log f(X_i)$, the maximizer often degenerates into a set of spikes at the data points. Hence, it is natural to add a penalty over the complexity of the estimator in order to restrict the set of admissible estimators. Complexity here means that we can either control the size of an approximation set or the regularity of the solution space. In a sieves type methodology, the penalty is of order the dimension of the approximating space, see for instance the work by Birgé and Massart in [3]. In our work, we will focus on smoothness type penalties. Given the observations $X_1, \ldots, X_n$ and assuming that the density f_0 belongs to a class of functions $\mathcal{F}$, the penalized loglikelihood estimator is defined as

$$\hat{f}_n = \arg\max_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \log f(X_i) - \lambda_n^2 I(f) \right), \tag{1}$$

where $I(.)$ is a penalty and λ_n^2 a sequence of positive number decreasing to zero, which balances the two terms of (1). The smaller λ_n^2, the closer to the data is the estimator. Hence the smoothing sequence must decrease to zero but not too fast, such that the regularization effect occurs.
Choosing the penalty as well as the decay of the smoothing sequence determine the asymptotic behavior of $\hat{f}_n$. Several choices have been investigated for the penalty: in the original paper by Good and Gaskings [17], the authors propose to use a flamboyancy functional such as $I(f) = 2 \int \left(f''\right)^2$ or $I(f) = 2 \int \left(\sqrt{f}'\right)^2$ in [14]. Silverman in [20] chose $I(f) = \int \left([\log f]^{(3)}\right)^2$. Other authors, see for instance Tapia and Thompson in [19], have discretized the problem. Projection methods onto different bases and penalty over the coefficients provide good estimators. Stone in [21] or Barron and Sheu in [1] consider log-splines bases. For a general review of penalization methods for density estimation, we refer to [8]. Van de Geer in [22] provide a theorem for general penalties. Under the assumption that there exists $m > 0$ such that f_0 lies in a Sobolev space $H^m([0,1])$ and for a choice of penalty $I(f) = \int_0^1 (f^{(m)})^2(t)dt$, the consistency in Hellinger distance $h(.,.)$ follows. More precisely it is proven that

$$h(\tilde{f}_n, f_0) = O_{\mathbf{P}}(\lambda_n)(1 + I(f_0))$$

provided that the smoothing parameter decreases at the following rate:

$$\lambda_n^{-1} = O_{\mathbf{P}}(n^{\frac{m}{2m+1}})(1 + I(f_0))^{\frac{1}{2}}.$$

As a result, the estimator achieves the optimal rate of convergence but the prior knowledge of its regularity m is needed for its construction.
Hence, the asymptotic behaviour of the estimates depends on an optimal choice of the smoothing parameter λ_n^2. In the previous works, the optimal value relies either on the prior knowledge of the regularity of the class of functions $\mathcal{F}$, or is found with cross validation techniques, which prevents adaptive estimation. That is the reason why we propose a l^1 penalty whose sparsity property is a key to adaptation, as in [18]. More precisely, under the assumption that the log density belongs to a Besov space with regularity s, choosing for penalty the l^1 norm of the wavelet coefficients of the log density, leads to an estimator converging at the minimax rate of convergence $\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}$ for the L^2 norm.

The article falls into three main parts. In the next section, Section 2, we provide the main theorem which describes the behavior of the estimator. All the technical lemmas are stated in Section 3, while the proofs are gathered in Section 4.

2 Main results

Consider an independent random sample $X_1, \ldots, X_n$ with unknown probability density with respect to Lebesgue measure λ, $f_0 = \frac{d\mathbf{P}}{d\lambda}$ on $[0,1]$. Assume that there is a functional set $\mathcal{F}$ such that $f_0 \in \mathcal{F}$. For a given penalty I, define the penalized maximum likelihood estimator $\tilde{f}_n$

$$\tilde{f}_n = \arg\max_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \log f(X_i) - \lambda_n^2 I(f) \right). \tag{2}$$

This estimator is well studied in [22] and is not adaptive. Hence we propose the following procedure. First set

$$\gamma_0 = \log(f_0) + b(\gamma_0),$$

with $b(\gamma_0) = -\int \log(f_0) d\mathbf{P}$. So, to every density $f \in \mathcal{F}$, we associate the variable $\gamma = \log f + b(\gamma)$ lying in the correspondent functional class Γ. Indeed, often in the literature of density estimation, it is more convenient to assume some regularity properties over the logarithm of the density, moreover it ensures positivity of the estimator.
Since the density integrates to one, we have the useful relation

$$b(\gamma) = \log \int e^{\gamma(x)} d\lambda(x). \tag{3}$$

Moreover we have

$$b(\gamma) - b(\gamma_0) = K(f,f_0) \tag{4}$$

where $K(.,.)$ is the Kullback-Leibler information. The two distance are linked by the following inequality:

$$h^2(f, f_0) \leqslant \frac{1}{2} K(f, f_0).$$

We shall now consider the penalty function $J : \Gamma \to \mathbb{R}^+$: $\forall \gamma \in \Gamma$, $J(\gamma) = I(f)$. Hence, the penalized maximum log-likelihood estimator (2) can be written as follows

$$\hat{\gamma}_n = \arg\max_{\gamma \in \Gamma} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(X_i) - b(\gamma) - \lambda_n^2 J(\gamma) \right). \tag{5}$$

In order to obtain an adaptive procedure, We do not consider a penalty depending directly on the regularity of the unknown function, but, using ideas analogous to the ones developed in Loubes and van de Geer [18], we consider the l^1-norm of coefficients of the function γ in a well chosen basis. More precisely, we assume that γ_0 lies in a Besov space $B^s_{p\infty}([0,1])$ with $s > \frac{1}{p}$. In the literature on density estimation, one often considers so-called Besov spaces $B^s_{p,q}([0,1])$. Such spaces are intrinsically connected to the analysis of curves since the scale of Besov spaces yields the opportunity to describe the regularity of functions, with more accuracy than the classical Hölder scale. General references about Besov spaces are Besov, Il'in and Nikol'skii [2], Edmund and Triebel [7] and DeVore and Lorentz [6]. The notation $B^s_{p,q}([0,1])$ refers to the case of functions on $[0,1]$, with "smoothness" s, and where p and q refer to L_p- and L_q-norms with respect to Lebesgue measure. In our framework, the main parameter is s, which stands for the regularity of the density to be estimated. Consider a compactly supported wavelet basis (ψ_{jk}), $j \geqslant 0$, $k = 0, \ldots, 2^j - 1$ of $B^s_{p\infty}([0,1])$ with respect to Lebesgue measure, with enough regularity $r > s$. Enough means here that the wavelet must have at least, $r > s$ vanishing moments, which corresponds to the regularity of the wavelet. We recall that a wavelet regularity is expressed through its number of vanishing moments, see e.g. Jaffard and Meyer [12] or Mallat [13]. Now for every function in this Besov space, there are coefficients $(\beta_{jk})_{j,k}$ called the wavelet coefficients such that we can write $f = \sum_{jk} \beta_{jk} \psi_{jk}$. Then a Besov norm for $s > 1/p$ is equivalent to an appropriate norm in the sequence space, that is, the space of the wavelet coefficients, see DeVore and Lorentz [6] or Donoho, Johnstone, Kerkyacharyan and Picard [16].

So decompose the log-density onto a wavelet basis and write $\gamma_0 = \sum_{jk} \beta^0_{jk} \psi_{jk}$. For all resolution level $j_1 = j_1(n)$, consider the approximation space V_{j_1} defined by $V_{j_1} = \text{Vect}\{\psi_{jk},\ j < j_1,\ k = 0, \ldots, 2^j - 1\}$. Write also $\gamma_1 = \sum_{j<j_1} \sum_{k=0}^{2^j-1} \beta^0_{jk} \psi_{jk}$, the projection of γ_0 onto the space V_{j_1}. Now we consider

the following penalty for all $\gamma \in \Gamma$

$$J(\gamma) = \sum_{j<j_1} \sum_k |\beta_{jk}|.$$

With that choice of penalty, we can prove the following theorem describing the asymptotic behaviour of the penalized M-estimator (5).

Theorem 1 *Assume that* $\exists 0 < C < \infty, \sup_{\gamma\in\Gamma} |\gamma| \leqslant C$. *For* j_1 *such that* $2^{j_1} = 0\left(\frac{n}{\log n}\right)$, *and the smoothing sequence such that* $\lambda_n^2 \geqslant c\sqrt{\frac{\log n}{n}}$ *for* c *a constant, hence the penalized log-likelihood estimator defined for* $\gamma_0 \in B^s_{p\infty}([0,1])$ *as*

$$\hat{\gamma}_n = \arg \max_{\gamma=\sum_{j<j_1}\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{jk}\in\Gamma} \left(\frac{1}{n}\sum_{i=1}^n \gamma(X_i) - b(\gamma) - \lambda_n^2 \sum_{j<j_1}\sum_k |\beta_{jk}| \right),$$

is such that there exists a finite positive constant C_2 *such that*

$$\|\hat{\gamma}_n - \gamma_0\|^2 = O_P\left(\frac{n}{\log n}\right)^{-\frac{2s}{2s+1}}.$$

The proof of this theorem relies on empirical process theory. It is postponed to the appendix, Section 4.

Remark 2 *The condition* $\sup_{\gamma\in\Gamma} |\gamma| \leqslant C$, *is similar to the usual condition in log-likelihood estimation where a lower bound for the density is required:* $\exists \eta_0 > 0,\ f = \frac{d\mathbf{P}}{d\lambda} \geqslant \eta_0^2$.

Remark 3 *Throughout all the paper, we are not concerned with the issue of the existence of a solution to the maximization problem* (5), *and we make the assumption that a solution always exists. If not, consider the following approximation for a sequence* $\epsilon_n \to 0$*:*

$$\tilde{\gamma}_n = \arg\max_{\gamma\in\Gamma} \left(\frac{1}{n}\sum_{i=1}^n \gamma(X_i) - b(\gamma) - \lambda_n^2 J(\gamma) + \epsilon_n \right).$$

For a choice $\epsilon_n = O(\frac{1}{n})$, *the estimator* $\tilde{\gamma}_n$ *has the same asymptotic rate of convergence as* $\hat{\gamma}_n$.

The penalized log-likelihood estimator is pseudo-adaptive over the Besov class of functions $\{B^s_{p\infty}([0,1]),\ s > 1/p\}$ since it is convergent at the minimax rate of convergence up to a logarithmic factor for a quadratic loss and its unknown regularity is not used in its definition. The l^1 penalty provides adaptivity in density estimation in the same way as in the regression scheme, where the equivalence between l^1 penalty and a soft-thresholded estimator is obvious. In

density estimation, soft-thresholded estimators and penalized maximum likelihood estimator with l^1 penalty does not have the same expression. Nevertheless the sparsity constraint over the coefficients is of the same type, leading to similar asymptotic behaviours and turning maximum likelihood estimator into an adaptive estimator.

3 Technical Lemmas

In this section, we recall several results that are at the starting point of the proof of Theorem 1. The proofs can be found in [18].
Throughout all this paper, for a given set A, we will use the notation $\#A$ for the cardinality of the set A. For $m = (j, k)$, consider the penalty $J(\gamma) = \sum_{m\in\Lambda} |\beta_m|$, for Λ a finite set. Let $\mathcal{I}_n$ be any subset of Λ and define

$$N_n = \#\mathcal{I}_n, \quad J_N(\gamma) = \sum_{m\in\mathcal{I}_n} |\beta_m|, \quad J_M(\gamma) = \sum_{m\notin\mathcal{I}_n} |\beta_m|.$$

Hence we have

$$J(\gamma) = J_N(\gamma) + J_M(\gamma) = \sum_{m\in\Lambda} |\beta_m|.$$

Consider the set of functions $\gamma = \sum_{m\in\Lambda} \beta_m\psi_m$, for which

$$\sum_{m\in\Lambda} |\beta_m|^\rho \leqslant 1,$$

for some $0 \leqslant \rho \leqslant 2$. We may think of ρ as a *roughness* parameter: if $\rho = 0$, we assume the convention $x^0 = 1$ if x is non zero and $0^0 = 0$. As a consequence we get

$$\sum_{m\in\Lambda} |\beta_m|^0 = \#\{\beta_m,\ \beta_m \neq 0\}$$

So for $\rho = 0$ the function γ may have at most 1 non-zero coefficient, whereas, on the other extreme, $\rho = 2$ only requires that γ is within the n-dimensional unit ball. We can point out that the sets $\{\beta,\ \sum_{m\in\Lambda} |\beta_m|^p \leqslant 1\}$ increase for the inclusion as ρ becomes large. Thus, the smaller ρ the "smoother" γ will be.
This is also reflected by the entropy calculation: the smaller ρ, the smaller the entropy, see [18] for more comments on this topic.
The following lemma provides upper bound for the penalty term when the objective function γ belongs to a ball of space with roughness ρ.

Lemma 4 *Suppose that*

$$\sum_{m\in\Lambda} |\beta_m|^\rho \leqslant 1,$$

for some $0 \leqslant \rho < 1$. *Take* $\mathcal{J}_n = \{m :\ |\beta_m| > \lambda_n^2\}$. *We obtain that*

$$N_n \leqslant \lambda_n^{-2\rho} \quad J_M \leqslant \lambda_n^{2(1-\rho)},$$

and as a result we get

$$\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n J_M^{\frac{1}{2}} \leqslant 2\lambda_n^{2-\rho}.$$

For $\gamma_0 = \sum_j \sum_k \beta_{jk}\psi_{jk} \in B^s_{pq}$, we get for $J > 0$:

$$\left(\sum_{j=1}^{J} 2^{j((2s+1)\frac{p}{2}-1)\frac{q}{p}} \left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \leqslant 1. \tag{6}$$

This quantity is equivalent to the Besov semi-norm. Throughout, we assume $s \geqslant 0$, $p \geqslant 1$, and $q \geqslant 1$. In the Besov space interpretation, $\mathcal{B}^s_{pq}$ (with $J = \infty$) corresponds (in the sense of norm equivalence) to a Besov ball in the space $B^s_{pq}([0,1])$.

Lemma 5 *Suppose that* $\beta = (\beta_{jk})$ *satisfies* (6), *with* $\rho = 2/(2s+1) \leqslant \min(p,q)$, *and* $J < \infty$. *Then*

$$\sum_{j=1}^{J} \sum_{k=1}^{2^j} |\beta_{j,k}|^\rho \leqslant J^{q-\frac{\rho}{q}}. \tag{7}$$

4 Appendix

Proof of Lemma 5:

By Hölder's inequality, for a sequence $a_1, \ldots, a_L$, and for $t \geqslant 1$,

$$\sum_{l=1}^{L} |a_l| \leqslant L^{\frac{t-1}{t}} \left(\sum_{l=1}^{L} |a_l|^t \right)^{\frac{1}{t}}. \tag{8}$$

Apply this first with $L = J$, $|a_j| = \sum_{k=1}^{2^j} |\beta_{jk}|^\rho$, and $t = q/\rho$. Then we find

$$\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^\rho \right\} \leqslant J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}. \tag{9}$$

Next, apply (8) with $L = 2^j$, $|a_{j,k}| = |\beta_{jk}|^\rho$, and $t = p/\rho$. This yields

$$\left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^\rho \right\} \leqslant \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\beta_{jk}|^p)^{\frac{\rho}{p}} \right\}.$$

Do this for each $j = 1, \dots J$, and insert the result in (9):

$$J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$\leqslant J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\beta_{jk}|^p)^{\frac{\rho}{p}} \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$= J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} 2^{j(\frac{p-\rho}{p})\frac{q}{\rho}} \left\{ \sum_{k=1}^{2^j} |\beta_{jk}|^p \right\}^{\frac{q}{p}} \right)^{\frac{\rho}{q}} \leqslant J^{\frac{q-\rho}{q}},$$

since

$$\left(\frac{p-\rho}{p} \right) \frac{q}{\rho} = \left((2s+1)\frac{p}{2} - 1 \right) \frac{q}{p}.$$

Proof of Theorem 1:

Set γ_1 the projection of γ_0 onto the approximation space V_{j_1}. The estimation error can be split in two terms: a stochastic term and an approximation error.

$$||\hat{\gamma}_n - \gamma_0|| \leqslant ||\hat{\gamma}_n - \gamma_1|| + ||\gamma_1 - \gamma_0||.$$

From the property of the wavelet basis and the choice of the level j_1 we have, since $\gamma_0 \in B^s_{p\infty}([0,1])$:

$$||\gamma_1 - \gamma_0|| \leqslant 2^{-j_1 s} \leqslant \left(\frac{n}{\log n} \right)^{-s/2}.$$

Such upper bound only involves the regularity of the log-density γ_0, measured in terms of Besov spaces.
Now we turn on the stochastic error. The proof involves a concentration inequality for the empirical process, as it is stated in [22]. The following upper bound stands for all $\gamma \in \Gamma$,

$$\begin{aligned} |\int (\gamma - \gamma_1) d(P_n - \mathbf{P})| &= |\int \sum_m (\beta_m - \beta_m^0)\psi_m d(P_n - \mathbf{P})| \\ &\leqslant \sup_m \left| \int \psi_m d(P_n - P) \right| J(\gamma - \gamma_1). \end{aligned}$$

We must derive a concentration inequality over $|\int \psi_m d(P_n - \mathbf{P})|$, from a Bernstein type inequality. Recall that for $\xi_1, \dots, \xi_n$ i.i.d bounded random variables such that $\mathbf{E}\xi_i = 0$, $\mathbf{E}\xi_i^2 \leqslant \sigma^2$, $|\xi_i| \leqslant \|\xi\|_\infty < \infty$, then:

$$\mathbf{P}\left(|\frac{1}{n}\sum_{i=1}^{n} \xi_i| > \lambda \right) \leqslant 2\exp\left(-\frac{n\lambda^2}{2(\sigma^2 + \|\xi\|_\infty \lambda/3)} \right), \ \forall \lambda > 0.$$

Previous inequality gives here the following upper bound:

$$\mathbf{P}\left(|\int \psi_m d(P_n - P)| > T_n\right) \leqslant 2\exp(-\frac{nT_n^2}{2(\sigma^2 + 3/2||Y||_\infty)})$$

where $Y_i = \psi_m(X_i) - E(\psi_m(X_i))$ are independent random variables with zero mean. Here $\sigma^2 \leqslant ||f_0||_\infty$ and $||Y||_\infty \leqslant 2^{j/2}M$. So, using Bernstein inequality, there exists a finite constant A such that,

$$\mathbf{P}(\sup_m |\int \psi_m d(P_n - \mathbf{P})| > T_n) \leqslant 2\exp(-A((nT_n^2) \wedge (nT_n))).$$

Using this inequality we obtain:

$$\begin{aligned}\mathbf{P}(\sum_{m\in\Lambda} |\int \psi_m d(P_n - \mathbf{P})| > T_n) &\leqslant \sum_{m\in\Lambda} 2\exp(-AnT_n^2) \\ &\leqslant |\Lambda| 2\exp(-AnT_n^2).\end{aligned}$$

If we choose $T_n = c\sqrt{\frac{\log n}{n}}$ then if the set of indices Λ is polynomial in n, for c large enough we have

$$\mathbf{P}(\sup_{m\in\Lambda} |\int |\psi_m| d(P_n - \mathbf{P})| > T_n) \leqslant 2\frac{|\Lambda|}{n^{Ac^2}} \to O.$$

Now recall our model: we consider a wavelet basis $m = (j,k)$ and we begin to approximate the log-density by its projection onto the space V_{j_1} for the convenient choice of j_1 that. Moreover, we have made the assumption that the log-density belongs to a Besov space $B^s_{p\infty}([0,1])$ and is bounded in the supremum norm. As a result,

$$\begin{aligned}\mathbf{P}(\sup_{0\leqslant j\leqslant j_1,k} |\int \psi_{jk} d(P_n - \mathbf{P})| \geqslant T_n) &\leqslant \sum_{j=0}^{j_1}\sum_k \mathbf{P}(|\int \psi_{jk} d(P_n - \mathbf{P})| \geqslant T_n) \\ &\leqslant \sum_{j=0}^{j_1}\sum_k 2\exp(-A_j((nT_n^2) \wedge (nT_n))) \\ &\leqslant 2\sum_{j=0}^{j_1} 2^j \exp(-AnT_n^2)\end{aligned}$$

for a choice of T_n and j_1 such that $T_n = c\sqrt{\frac{\log n}{n}}$ and $2^{j_1} \leqslant 1/c\sqrt{\frac{n}{\log n}}$, we have

$$\begin{aligned}\mathbf{P}(\sup_{0\leqslant j\leqslant j_1,k} |\int \psi_{jk} d(P_n - \mathbf{P})| \geqslant T_n) &\leqslant 22^{j_1}\exp(-AnT_n^2) \\ &\leqslant 2/c\frac{n^{1/2-Ac^2}}{\sqrt{\log n}}.\end{aligned}$$

As soon as we have chosen c large enough, the last quantity tends to zero as n increases. The condition over the choice of the constant c can be written as:

$$c^2 \geqslant \max(||f_0||_\infty, 2/3||\psi||_\infty).$$

Then on an event of probability one we can write that for every $\lambda_n^2 \geqslant c\sqrt{\frac{\log n}{n}}$ we have

$$\sup_{(j,k)\in\Lambda} |\int \psi_{jk} d(P_n - \mathbf{P})| \leqslant \lambda_n^2.$$

The following inequality is a direct consequence of the definition of the M-estimator (5).

$$b(\hat{\gamma}_n) - b(\gamma_1) + \lambda_n^2 J(\hat{\gamma}_n) \leqslant \int (\hat{\gamma}_n - \gamma_1) d(P_n - \mathbf{P}) + \lambda_n^2 J(\gamma_1).$$

As a matter of fact, recalling the definition of $\hat{\gamma}_n$ and using the fact that γ is, by construction, a centered variable, we get that:

$$\begin{aligned}
\forall \gamma \in \Gamma, \quad \int \hat{\gamma}_n dP_n - b(\hat{\gamma}_n) - \lambda_n^2 J(\hat{\gamma}_n) &\geqslant \int \gamma dP_n - b(\gamma) - \lambda_n^2 J(\gamma) \\
\int \hat{\gamma}_n dP_n - b(\hat{\gamma}_n) - \lambda_n^2 J(\hat{\gamma}_n) &\geqslant \int \gamma_1 dP_n - b(\gamma_1) - \lambda_n^2 J(\gamma_1) \\
\int (\hat{\gamma}_n - \gamma_1) dP_n + \lambda_n^2 J(\gamma_1) &\geqslant b(\hat{\gamma}_n) - b(\gamma_1) + \lambda_n^2 J(\hat{\gamma}_n) \\
\int (\hat{\gamma}_n - \gamma_1) d(P_n - \mathbf{P}) + \lambda_n^2 J(\gamma_1) &\geqslant b(\hat{\gamma}_n) - b(\gamma_1) + \lambda_n^2 J(\hat{\gamma}_n)
\end{aligned}$$

Now, consistency of the estimator and a Taylor's expansion of $b(\gamma)$ lead to:

$$b(\hat{\gamma}_n) - b(\gamma_1) = E\|\hat{\gamma}_n(X_1) - \gamma_1(X_1)\|^2/(1 + O(1)). \tag{10}$$

But since $\frac{d\mathbf{P}}{d\lambda} \geqslant \eta^2$, we have

$$\frac{||\hat{\gamma}_n - \gamma_1||^2}{1 + O_{\mathbf{P}}(1)} + \lambda_n^2 J(\hat{\gamma}_n) \leqslant \int (\hat{\gamma}_n - \gamma_1) dP_n + \lambda_n^2 J(\gamma_1). \tag{11}$$

As a result, for the stochastic term, we have the following inequality:

$$\frac{||\hat{\gamma}_n - \gamma_1||^2}{1 + O_{\mathbf{P}}(1)} + \lambda_n^2 J(\hat{\gamma}_n) \leqslant \lambda_n^2 J(\hat{\gamma}_n - \gamma_1) + \lambda_n^2 J(\gamma_1).$$

Set $\Lambda = \{j < j_1,\ k = 0, \ldots, 2^j - 1\}$. Or

$$\begin{aligned}
\frac{||\hat{\gamma}_n - \gamma_1||^2}{1 + O_{\mathbf{P}}(1)} + \lambda_n^2 J_M(\hat{\gamma}_n) &\leqslant \lambda_n^2 J_N(\hat{\gamma}_n - \gamma_1) + \lambda_n^2 J_M(\hat{\gamma}_n - \gamma_1) \\
&\quad + \lambda_n^2 (J_N(\gamma_1) - J_N(\hat{\gamma}_n)) + \lambda_n^2 J_M(\gamma_1) \\
&\leqslant 2\|\hat{\gamma}_n - \gamma_1\| \lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n^2 J_M(\hat{\gamma}_n) + 2\lambda_n^2 J_M(\gamma_1),
\end{aligned}$$

or

$$\|\hat{\gamma}_n - \gamma_1\|^2 \leqslant 2\|\hat{\gamma}_n - \gamma_1\|\lambda_n^2 N_n^{\frac{1}{2}} + 2\lambda_n^2 J_M(\gamma_0).$$

So we obtain, using Lemma 4 and Lemma 5 with $\rho = 2/(2s+1)$ and $\mathcal{I}_n = \{m = (j,k),\ |\beta_m| > \lambda_n^2\}$:

$$\begin{aligned} ||\hat{\gamma}_n - \gamma_1|| &\leqslant O\left(\lambda_n^2 N_n^{1/2} + \lambda_n J_M^{1/2}\right) \\ &\leqslant O\left(\frac{\log n}{n}\right)^{2-\rho} \\ &\leqslant O\left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}}. \end{aligned}$$

And, by comparison of the two rates of convergence, we have, for C_2 a positive finite constant:

$$||\hat{\gamma}_n - \gamma_0|| \leqslant C_2 \left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}},$$

which concludes the proof.

References

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

[2] Besov, O. V. and Ilin, V. P. and Nikolskiĭ, S. M. Integral representations of functions and imbedding theorems. Vol. I. *V. H. Winston & Sons*,1978.

[3] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[4] D. Bosq. *Nonparametric statistics for stochastic processes*. Springer-Verlag, New York, 1996. Estimation and prediction.

[5] Gwénaëlle Castellan. Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8):729–732, 2000.

[6] DeVore, R. A. and Lorentz, G. G. Constructive approximation. *Springer-Verlag*, 1993.

[7] Edmunds, D. E. and Triebel, H. Entropy numbers and approximation numbers in function spaces. *Proc. London Math. Soc. (3)*, 64(1), 153–169, 1992.

[8] P. P. B. Eggermont and V. N. LaRiccia, *Maximum penalized likelihood estimation. Vol. I*, Springer Series in Statistics, Springer-Verlag, New York, 2001, Density estimation.

[9] Gu, C. and Qiu, C. Smoothing spline density estimation: theory *Ann. Statist.*, 21 (1), 217–234, 1993.

[10] Gu, C. Smoothing spline density estimation: a dimensionless automatic algorithm. *J. Amer Statist. Assoc*, 88, 495–504, 1993.

[11] Gu, C and Wang, J. Penalized likelihood density estimation: direct cross-validation and scalabe approximation. *Statistica Sinica*, 13, 811–826, 2003.

[12] Jaffard, S. and Meyer, Y. Les ondelettes. *Harmonic analysis and partial differential equations (El Escorial, 1987)*,182–192, 1989.

[13] Mallat, S. A wavelet tour of signal processing. *Academic Press Inc.*,1998.

[14] G. F. de Montricher, R. A. Tapia, and J. R. Thompson. Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Ann. Statist.*, 3(6):1329–1348, 1975.

[15] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors.

[16] Donoho, D. L. and Johnstone, I. M. and Kerkyacharian, G. and Picard, D. Density estimation by wavelet thresholding. *Ann. Statist.*,24 (2), 508–539, 1996.

[17] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

[18] J-M. Loubes and S. van de Geer. Adaptive estimation using thresholding type penalties. *Statistica Neerlandica*, 56:1–26, 2002.

[19] D. W. Scott, R. A. Tapia, and J. R. Thompson. Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *Ann. Statist.*, 8(4):820–832, 1980.

[20] B. W. Silverman. Some remarks on roughness penalty density estimators. In *Limit theorems in probability and statistics, Vol. I, II (Veszprém, 1982)*, pages 957–979. North-Holland, Amsterdam, 1984.

[21] C. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990.

[22] Sara A. van de Geer. *Applications of empirical process theory*. Cambridge University Press, Cambridge, 2000.

[23] G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.