Probab. Theory Relat. Fields 97, 113–150 (1993)

Probability Theory and Related Fields

Rates of convergence for minimum contrast estimators

Lucien Birgé[1] **and Pascal Massart**[2]

[1]URA 1321 "Statistique et modèles aléatoires," 45-55 3e ét boîte 158, Université Paris VI, 4 Place Jussieu, F-75252 Paris Cedex 05, France
[2]URA 743 "Modélisation stochastique et Statistique", Bât. 425, Université Paris Sud, F-91405 Orsay Cedex, France

Received July 3, 1991; in revised form March 8, 1993

Summary. We shall present here a general study of minimum contrast estimators in a nonparametric setting (although our results are also valid in the classical parametric case) for independent observations. These estimators include many of the most popular estimators in various situations such as maximum likelihood estimators, least squares and other estimators of the regression function, estimators for mixture models or deconvolution... The main theorem relates the rate of convergence of those estimators to the entropy structure of the space of parameters. Optimal rates depending on entropy conditions are already known, at least for some of the models involved, and they agree with what we get for minimum contrast estimators as long as the entropy counts are not too large. But, under some circumstances ("large" entropies or changes in the entropy structure due to local perturbations), the resulting rates are only suboptimal. Counterexamples are constructed which show that the phenomenon is real for non-parametric maximum likelihood or regression. This proves that, under purely metric assumptions, our theorem is optimal and that minimum contrast estimators happen to be suboptimal.

Mathematics Subject Classification (1980): 62G05, 62J02

1 Introduction

Our purpose in this paper will be to study some general properties of "minimum contrast estimators" (M.C.E.), which include maximum likelihood estimators (M.L.E.) for i.i.d. variables, least squares (L.S.E.) and minimum $\mathbb{L}^1$-norm (M.$\mathbb{L}$1.N.E.) estimators in a regression setting and several other related estimators. This approach leads to a unified framework for studying various situations which are usually treated separately. This greater generality is, as is often the case, at the price of more restricted assumptions which could be improved in some particular circumstances but it has the advantage of providing a global approach to the problem. In order to motivate the subsequent study of M.C.E., let us first recall two desirable properties of estimators of a parameter belonging to a compact parameter space.

1) The estimator should have a risk which is close to the minimax.

2) It should not be strongly affected by small errors on the model, "small" meaning of the order of the expected risk.

To be more specific, assume that we deal with n variables with joint distribution $\mathbb{P}_{\theta,n}$, $\theta \in \Theta$, where (Θ, d) is a compact metric space. For some estimator T_n we shall consider its maximal risk

$$R(T_n) = \sup_{\theta \in \Theta} \mathbb{E}_{\theta,n}[d^2(\theta, T_n)]$$

and we would like to find a procedure θ_n for designing estimators $\hat{\theta}_n = \theta_n(\Theta)$ satisfying the following properties:

1) For some constant C independent of n and $R_n = \inf_{T_n} R(T_n)$, $R(\hat{\theta}_n) \leqq C\, R_n$.

2) Assuming that the true parameter space is Θ but we use a model Θ_n such that $d^2(\Theta, \Theta_n) \leqq C\, R_n$, is it true that when $\hat{\theta}_n = \theta_n(\Theta_n)$

$$\sup_{\theta \in \Theta} \mathbb{E}_{\theta,n}[d^2(\theta, \hat{\theta}_n)] \leqq K(C) R_n$$

for some fixed function K. In this case we shall say that the procedure θ_n is "stable". A study of the minimax risk is given in Birgé (1983), mainly in the case of i.i.d. variables (density estimation) and Hellinger distance. It is proved there that, under some regularity assumptions (uniformly bounded likelihood ratios say), the minimax risk is determined (up to multiplicative constants) by the metric entropy of the parameter space. Using the same framework, it is proved in Birgé (1984) that in general (apart from some pathological parameter spaces) stability should occur for some well-chosen estimators. Therefore this paper will be mainly devoted to the relationship between the metric entropy structure of the parameter space and the performances of M.C.E. We shall conclude that as soon as the entropy function is regular enough and not too large M.C.E. will have the optimal rate of convergence, but when it becomes too large only suboptimal rates can be expected. Moreover, irregular behaviour of the entropy functions can lead to serious troubles for this estimator which implies "unstability". All these only concern unrestricted M.C.E. i.e. those estimators for which the minimization is carried out over the whole parameter space. In a subsequent work we shall investigate the properties of some restricted M.C.E. and show that they behave in a much better way.

We do not know much about systematic studies of M.C.E. properties. Actually M.C.E. are some sort of generalization of Huber's M-estimators and they apparently appear for the first time in Huber (1967) and, up to our knowledge, the name M.C.E. was coined by Pfanzagl (1969). A study of consistency is to be found in Reiss (1978) but we believe that the first systematic approach to the nonparametric case is in Severini and Wong (1987).

While there exists an enormous amount of literature concerning nonparametric curve estimation (density or regression estimation) and parametric maximum likelihood or least squares estimators as well, the number of papers devoted to maximum likelihood density estimation or nonparametric least squares estimation appears to be more limited. Some general results are to be found in Reiss (1984) and several papers have been devoted to the Grenander estimator (M.L.E. for monotone densities) by Grenander (1956 and 1980), Groeneboom (1985), Prakasa Rao (1983), Barlow et al. (1972), Birgé (1989). There are probably other references that we are not aware of but some recent and noticeable results are to be found in

Severini and Wong (1987), Van de Geer (1990a) and Nemirovskii et al. (1984) and (1985) which mainly motivated the present work. Van de Geer (1990a, b) gives fairly general and optimal results in the regression framework for L.S.E. or M.$\mathbb{L}$1.N.E. under some metric assumptions. Those results are restricted to cases where the dimension is not too large, i.e. when some entropy condition (*) is satisfied. Under such a condition the rates of convergence of the estimators are shown to be optimal but nothing is said about the case when (*) is not true.

The preprint by Severini and Wong (1987) was, to our knowledge, the first attempt to give a general theory of rates of convergence for M.C.E. related to dimension. Unfortunately, the results do not always lead to the optimal rates of convergence and, in particular, do not allow to recover the classical $n^{-1/2}$ rate in the parametric case. It should be noted here that the assumptions concerning the bounds on their parameter space are weaker than ours and therefore the results not directly comparable. The authors also make use, in the i.i.d. case, of the uniform metric, which is not the adequate one for these problems (see Birgé 1983 for example).

As to Nemirovskii et al. (1984) and (1985), they give a fairly general account of the regression situation, not only for L.S.E. or M.$\mathbb{L}$1.N.E. estimators but also for other minimization criteria, with very precise asymptotic results, but only for classical Sobolev-type classes of functions and not under purely metric assumptions. These classes happen to belong to the entropy domain which was considered by Van de Geer (1990a) and always satisfy condition (*).

In view of these different attempts, we shall present a unified treatment for all those situations, including the more general regression framework of Nemirovskii et al. (1984) and (1985) and various other examples. The presentation is inspired by Van de Geer (1990a) and we shall get the optimal rates of convergence under entropy restrictions similar to condition (*) of Van de Geer (1990a). We shall also prove rates of convergence when (*) is not satisfied, but in this case they happen to be suboptimal. Since this could very well be a weakness of our proof rather than a weakness of the estimators themselves, we shall exhibit some counterexamples leading to lower bounds for the rates of convergence of these estimators. They show that, without additional assumptions, nothing better can be expected and that the estimators are actually suboptimal in some circumstances.

To be more specific let us denote by S_α the class of functions s on $[0, 1]$ such that for x, y in $[0, 1]$ $f(x) \geqq 1/10$ and $|f(x) - f(y)| \leqq |x - y|^\alpha$. The entropy structure of such spaces is well-known, see for example Kolmogorov and Tikhomirov (1961) or Lorentz (1966), and the rate of convergence of optimal estimators, for such classes or similar classes with analogous entropy properties should be $n^{-\frac{\alpha}{2\alpha+1}}$ for various models including density or regression estimation. We shall prove that, when $\alpha > 1/2$, the rate of convergence of M.C.E. is actually the optimal one but, for $\alpha < 1/2$, one can only guarantee a convergence at a rate $n^{-\alpha/2}$ for M.C.E.

Indeed, there is a split at some stage, corresponding to the entropy structure of 1/2-Hölderian functions on $[0, 1]$. The same split occurs in the proof of functional central limit theorems, for analogous reasons. Our counterexamples below show that, under this type (purely metric) of assumptions on the parameter space, our results are optimal whatever the entropy structure. Of course, α-Hölderian densities are rather wild when $\alpha < 1/2$ but the same type of split carries out to the multidimensional case and the class of densities with uniformly bounded second derivatives on $[0, 1]^k$ has an entropy structure which does not satisfy (*) as soon as

$k \geqq 4$. Our counterexamples only deal with the one-dimensional case in order to keep the length of the proofs decent but it is clear that one could extend them and show that M.L.E. would not reach the right rate of convergence when estimation is carried over some particular spaces of densities on $[0, 1]^5$ with bounded second derivatives. A modification of our construction shows that no stability is to be expected from M.C.E. Those estimators can have a very poor behaviour in case of a small (smaller than the expected error) misspecification of the parameter space. The proof of the counterexample connected with M.L.E. is more complicated because of the introduction of a convex parameter space. The first version, involving the simpler non-convex space that we use in the regression framework led to an interrogation: could convexity improve the performance of M.L.E.? This could very well be the case and was not clear for us until we derived the new counterexample. For sake of simplicity we did not modify the regression case.

In what follows, we shall set up a framework which is close to the one initiated by Severini and Wong (1987), although our assumptions and proofs will clearly be different in order to get the optimal rates of convergence. Section 2 will be devoted to the presentation of this framework, the corresponding assumptions and the main theorem which gives the precise relationship between the entropy structure of the parameter space and the rate of convergence of M.C.E. In Sect. 3, we will apply these results to various situations, including M.L.E. and regression estimation, showing how they can be fitted to the general framework and discussing the assumptions. We already mentioned that the cases of L.S.E. and M.𝕃1.N.E. have been partially treated by Van de Geer (1990a). At the time of the writing of this paper we received a preprint by the same author dealing with M.L.E. using Hellinger distance. Some similarities between the two papers are therefore unavoidable, although the techniques and results appear to be different. The entropy conditions in Van de Geer (1990b) are much more restrictive than ours but the assumptions on the uniform bounds are weaker, which is not surprising in a paper which only considers M.L.E. The generality we want to deal with implies some additional restrictions in order to make the whole machinery work. It is clear that in each particular case, some ad hoc procedures could be used to simplify the assumptions, but this is not the purpose of this paper. Finally, in Sect. 4, we shall develop, for both M.L.E. and L.S.E., lower bounds arguments which prove that the results of the main theorem are optimal, under the type of assumptions we work with. They also show that M.C.E. are definitely unstable.

2 The abstract model

Our framework follows more or less the classical nonparametric regression model, although it is intended to describe various other models too, as will become obvious later. We assume the existence of $2n$ independent variables $X_1, \ldots, X_n$ and $W_1, \ldots, W_n$ with values on measurable spaces $\mathscr{X}$ and $\mathscr{W}$ respectively. We are given some abstract parameter space S and some measurable transformation f, indexed by S from $\mathscr{X} \times \mathscr{W}$ to another measurable space $\mathscr{Z}$. We only observe the variables $Z_i = f(X_i, W_i, s)$ where the parameter s is unknown and the distribution of the X_i's and W_i's depends on s. Let us denote by P_s^i the joint distribution of (X_i, W_i, Z_i) when s obtains, by $\mathbb{E}_s$ the expectation with respect to probability $\bar{P}_s = n^{-1} \sum_{i=1}^n P_s^i$ and by P_n the empirical distribution of the Z_i's.

Our purpose is to estimate the unknown s on the basis of the observations. The estimators that we shall consider in this paper are defined by the mean of a "contrast function".

Definition 1. A real function γ on $\mathscr{X} \times S$ is a *contrast function* if for all s in S,

$$\inf_{t \in S} \mathbb{E}_s[\gamma(Z, t)] = \mathbb{E}_s[\gamma(Z, s)] .$$

Since s is a minimizer of $\mathbb{E}_s[\gamma(Z, t)]$, a natural candidate for an estimator of s would be a minimizer of the expectation of $\gamma(Z, t)$ with respect to the empirical distribution of the Z_i's which leads to the following definition.

Definition 2. Given some contrast function γ, a *minimum contrast estimator* or M.C.E. will be any minimizer $\hat{s}(Z_1, \ldots, Z_n)$ of the function $t \to \frac{1}{n} \sum_{i=1}^{n} \gamma(Z, t) = \gamma_n(t)$ (*empirical contrast*).

Remark. Actually, $\hat{s}$ may very-well not exist at all, but then we could define ε-minimum contrast estimators (or ε-M.C.E.) in the usual way: an ε-M.C.E. will be any $\hat{s}$ such that

$$\gamma_n(\hat{s}) \leqq \inf_t \gamma_n(t) + \varepsilon .$$

Due to the very definition of our estimators, it is clear that their performances are naturally measured in terms of the "loss function ℓ" on $S \times S$:

$$\ell(s, t) = \int (\gamma(z, t) - \gamma(z, s)) d\bar{P}_s = \mathbb{E}_s[\gamma(Z, t) - \gamma(Z, s)] .$$

The assumptions that we need in order to control the regularity of γ and the "size", when properly defined, of the parameter space S may be rather restrictive when applied to γ itself. To allow more flexibility we shall introduce an auxillary function $\tilde{\gamma}$ defined on the same space $\mathscr{X} \times S$ to which our assumptions apply and another loss function $d(s, t)$ which, in the applications, will be equivalent to $\ell^{1/2}$.

In what follows, the true value s of the parameter is fixed and all auxilliary quantities might depend on it.

(A1) There exists two positive functions M and $\mathscr{W}$ and Δ on $\mathscr{X} \times S^2$ and a positive constant B such that for all $x \in \mathscr{X}$, $w \in \mathscr{W}$, t, $u \in S$ and $z = f(x, w, s)$ we have

$$|\tilde{\gamma}(z, t) - \tilde{\gamma}(z, u)| \leqq M(w) \Delta(x, t, u) \leqq BM(w) \ \bar{P}_s \text{ a.s.}$$

Remark. Usually S will be a set of functions acting on $\mathscr{X}$ and we must think of $\Delta(x, t, u)$ as $|t(x) - u(x)|$ or similar quantities.

(A2) There exists positive constants α and Γ such that for $1 \leqq i \leqq n$

$$\int \exp[\alpha M(w)] dP_s^i \leqq \Gamma .$$

(A3) For any $\delta > 0$, let us denote by $\mathscr{S}_\delta$ a covering of S of minimum cardinality $\exp[H(\delta, S)]$ such that for all $S_\delta \in \mathscr{S}_\delta$

$$\mathbb{E}_s\left[\left(\sup_{t, u \in S_\delta} \Delta^2(X, t, u)\right)^*\right] \leqq \delta^2 ,$$

then Card $\mathscr{S}_\delta < +\infty$. Here Y^* denotes the measurable envelope of Y.

We define the pseudo-distance d on S by

$$d^2(t,u) = \mathbb{E}_s[\Delta^2(X,t,u)]$$

and let v_n be the normalized centered empirical process defined by

$$v_n = \sqrt{n}(P_n - \bar{P}_s),$$

which means that for any function $g(z)$ on $\mathscr{X}$ we have

$$v_n(g) = n^{-1/2} \sum_{i=1}^{n} [g(Z_i) - \mathbb{E}_s(g(Z))].$$

(A4) There exists a constant $C \geqq B\alpha/\Gamma$ such that for any t in S whenever $\gamma_n(t) \leqq \gamma_n(s) + \varepsilon$, then

$$C^{-1}d^2(s,t) \leqq -n^{-1/2} v_n[\tilde{\gamma}(z,t) - \tilde{\gamma}(z,s)] + \varepsilon.$$

In many cases $\tilde{\gamma}$ and γ are related in a very simple way and then (A4) will be a consequence of

(A′4) There exists a function Ψ defined on S such that

$$\tilde{\gamma}(z,t) = \gamma(z,t) + \Psi(t)$$

and a constant $C \geqq B\alpha/\Gamma$ such that for any t in S $d^2(s,t) \leqq C\ell(s,t)$.

Indeed, due to the centering in v_n we see that

$$v_n[\tilde{\gamma}(z,t) - \tilde{\gamma}(z,s)] = v_n[\gamma(z,t) - \gamma(z,s)] = n^{1/2}[\gamma_n(t) - \gamma_n(s) - \ell(s,t)].$$

Remarks. 1) All the assumptions only involve differences between functions $\tilde{\gamma}$ and we can therefore always assume without loss of generality that $\tilde{\gamma}(z,s) = 0, \forall z \in \mathscr{X}$. This fact will be used systematically in the proofs.

2) In (A2) α and Γ are not uniquely defined but in view of Theorem 1 below it is clear that we should choose the ratio $\sqrt{\Gamma}/\alpha$ as small as possible. When $M(w) = M$ is constant, the best choice is $\alpha = 2/M$ leading to $\sqrt{\Gamma}/\alpha = Me/2$.

We can clearly define numbers $H(\delta, S')$ for any subset S' of S as in (A3) assuming S' to be the parameter space. These numbers, defined for $\delta > 0$ and S' any subset of S will play the role of entropy numbers and will determine the rate of convergence of M.C.E. as the following theorem shows:

Theorem 1. *Let us denote by s the true parameter and assume that* (A1) *to* (A4) *hold. Let* $B_s(\sigma) = \{t \in S : d(s,t) \leqq \sigma\}$. *Let* $a = 1 \vee \left(\dfrac{\alpha B}{64\Gamma C}\right)$ *and let* $\tilde{H}$ *be a function from* $]0, +\infty[\times]0,1]$ *to* $[1, +\infty[$ *such that*:

(i) $\qquad H(u, B_s(\sigma)) \leqq \tilde{H}(u,\sigma)$, *for any* $(u,\sigma) \in]0, +\infty[\times]0,1]$,

setting

$$\varphi(\sigma) = \int_{\frac{a\sigma^2}{128\Gamma C}}^{2a\sigma} \tilde{H}^{1/2}(u, 2\sigma)\,du,$$

(ii) *the function* $\sigma \to \varphi(\sigma)/\sigma$ *is non increasing and continuous.*

Then, there is an absolute constant K such that, if n is large enough, the equation

$$\frac{2KC\sqrt{\Gamma}}{\alpha}\varphi(\sigma) = \sqrt{n}\sigma^2, \tag{2.1}$$

has a unique solution σ^* *and the following bound holds for any positive* λ:
$\mathbb{P}^*[\exists t$ *such that* $\gamma_n(t) \leqq \gamma_n(s) + \varepsilon$ *and* $d^2(s,t) > \max(2C\varepsilon, \lambda\sigma^{*2})] \leqq 8.1\exp(-\lambda/2)$,
where $\mathbb{P}^*$ *denotes the outer measure associated with* $\mathbb{P} = \otimes_{i=1}^n P_s^i$.

Comments. Concerning the existence and the choice of a function $\tilde{H}$ fulfilling (i) and (ii), it is worth noticing that it is always possible to define $\tilde{H}(u,\sigma) = 1 \vee H(u,S)$. In the standard finite dimensional parametric case however, we shall see in Sect. 3 that $H(u, B_s(\sigma))$ is typically bounded by $A\,\mathrm{Log}(\sigma/u)$ and therefore $\tilde{H}(u,\sigma) = (A\,\mathrm{Log}(\sigma/u)) \vee 1$ will be a more judicious choice, allowing to ensure a $n^{-1/2}$ rate of convergence which is precisely what one can expect under these circumstances. It is also worth to notice the unpleasant effect of the bound B which prevents us to work with an unbounded distance. Even in the simplest case of finite dimension and $\tilde{H}(u,\sigma)$ as defined above, for large B, $\varphi(\sigma)$ will essentially be of order $a\sigma$ leading to σ^* of order $B/\sqrt{n}$. For unbounded sets, there is a need for a preliminary estimator in order to restrict oneself to a fixed ball of moderate size.

Theorem 1 is a consequence of Assumption (A4) and of the following result which might prove interesting by itself:

Theorem 2. *Assume that Assumptions* (A1), (A2) *and* (A3) *hold. Let* $K = 1920$ *and* a, $\tilde{H}$ *and* φ *be defined as in Theorem* 1 *for some* $C > 0$. *Let* σ^{*2} *be defined by eq.* (2.1). *Then, for any non-negative integer* L *such that* $2^{2L}\sigma^{*2} \leqq BC\Gamma/\alpha$, *the following inequality holds*:

$$P^*\left(\sup_{t\in S}\frac{|\nu_n(\tilde{\gamma}(\cdot,t)-\tilde{\gamma}(\cdot,s))|}{d^2(s,t)\vee 2^{2L}\sigma^{*2}} \geqq \sqrt{n}/(2C)\right) \leqq 8.1\exp(-2^{2L+1}n\sigma^{*2}/b),$$

with $b = 9a^2K^2C^2\Gamma/\alpha^2 \leqq n\sigma^{*2}$.

The proof of these theorems will be defered to the Appendix. An immediate corollary is as follows:

Corollary 1. *Let* $\hat{s}(Z_1,\ldots,Z_n)$ *be an* ε-M.C.E. *Under* (A1) *to* (A4) *we have for any* $p > 0$ *and some absolute constant* C_p:

$$\mathbb{E}^*[d^p(s,\hat{s})] \leqq (2C\varepsilon)^{p/2} + C_p\sigma^{*p}.$$

Proof of Corollary 1

$$\begin{aligned}\mathbb{E}^*[d^p(s,\hat{s})] &= \int_0^{+\infty} \mathbb{P}^*[d^2(s,\hat{s}) > t^{2/p}]dt\\ &\leqq \int_{2\varepsilon C\sigma^{*-2}}^{+\infty} \mathbb{P}^*[d^2(s,\hat{s}) > \lambda\sigma^{*2}]\lambda^{p/2-1}\frac{p}{2}\sigma^{*p}d\lambda + (2C\varepsilon)^{p/2}\\ &\leqq (2C\varepsilon)^{p/2} + 4.05p\sigma^{*p}\int_0^{+\infty}\lambda^{p/2-1}e^{-\lambda/2}d\lambda\,. \quad \square\end{aligned}$$

Remark. For practical purposes, ε should be small compared to σ^{*2} in order not to affect the risk of the estimators. Unfortunately, we do not usually know the constants involved in the model. Nevertheless, it seems that a choice of the type $\varepsilon = n^{-2}$ would be safe in most cases.

Some comments concerning assumption (A3) are necessary in order to relate this assumption to previous works on rates of convergence and mainly to Le Cam (1973), Birgé (1983) and Van de Geer (1990a) where those rates are related to the entropy structure of the parameter space. In our framework, another notion will be useful, which was introduced by Dudley (1978). In order to keep the presentation simple, we shall restrict ourselves to the following case:

Definition 3. Let $(\Theta, \|\cdot\|)$ be some subset of an $\mathbb{L}^p$-space of real functions, with $1 \leqq p \leqq \infty$. For any subset S of θ we shall denote by $H(\delta, S) = \operatorname{Log} N(\delta, S)$ its δ-entropy with bracketing where $N(\delta, S)$ denotes the smallest number of pairs (θ_i^-, θ_i^+), $1 \leqq i \leqq N(\delta, S)$ such that

$$\text{(i)} \qquad \theta_i^- \leqq \theta_i^+ \text{ and } \|\theta_i^+ - \theta_i^-\| \leqq \delta$$

$$\text{(ii)} \qquad S \subset \bigcup\nolimits_{i=0}^{N(\delta,S)} \{\theta \in \Theta | \theta_i^- \leqq \theta \leqq \theta_i^+\} .$$

It is clear that entropy with bracketing is larger than ordinary entropy and that the two numbers will be the same for $p = \infty$. The notion can also be extended to Hellinger distance which is actually an $\mathbb{L}^2$-norm on the square roots of the densities.

In most cases, $\Delta(x, t, u)$ will be some function of the difference $|t(x) - u(x)|$ (or eventually $|\sqrt{t(x)} - \sqrt{s(x)}|$) and d will be an $\mathbb{L}^p$-norm. If we can control entropy with bracketing for this norm, it will clearly be easy to check (A3), since we shall then have when $\theta_i^+ \geqq t \geqq \theta_i^-$ and $\theta_i^+ \geqq u \geqq \theta_i^-$, $\Delta(x, t, u) \leqq \Delta(x, \theta_i^+, \theta_i^-)$. Therefore, in examples, checking (A3) will amount to compute entropy with bracketing, and mainly for $\mathbb{L}^2$-norm. This is usually not much more complicated than computing ordinary entropy and the results are quite similar. When we have entropy results concerning uniform (sup-norm) metric, they immediately extend to entropy with bracketing and we shall not give details here, refering to Kolmogorov and Tihomirov (1961) or Lorentz (1966) for entropy computations. $\mathbb{L}^2$-entropies of Sobolev and other spaces are studied by Birman and Solomjak (1967) and can be extended to $\mathbb{L}^2$-entropy with bracketing as shown in Birgé and Massart (1993). Applications to estimation can be found in Birgé (1983) and (1986) (with further computations) and Van de Geer (1990a).

We shall content ourselves here to consider one very classical situation which has already been extensively studied (Birgé 1983 and 1986, Van de Geer 1990a, Stone 1982, Ibragimov and Has'minskii 1983, etc.) of parameter spaces of the following type. We shall denote by $S(m, \alpha, A)$ the set of functions s on some given compact interval of the real line such that $s^{(m)}$ exists and

$$|s^{(m)}(x) - s^{(m)}(y)| \leqq A|x - y|^\alpha .$$

For such a space $H(\delta, S)$ will typically be of order $\delta^{-\frac{1}{m+\alpha}}$ (for $\mathbb{L}^p$-norm, the situation being more complicated with Hellinger as shown in Birgé 1986) and optimal rates of convergence δ_n are given by the solution of the equation

$$\text{(2.2)} \qquad n\delta_n^2 = H(\delta_n, S)$$

leading to δ_n of order $n^{-\frac{m+\alpha}{2(m+\alpha)+1}}$, this being true both for density estimation and regression curve estimation (with gaussian errors, say). It is easy to check that, for $H(x, S)$ of order $x^{-1/t}$, $t > 0$ the solution σ_n of Eq. (2.1) if of order $n^{-\frac{t}{2t+1}}$ when the integral is convergent at zero, i.e. when $t > 1/2$ and $n^{-t/2}$ when $t < 1/2$. Therefore M.C.E. estimators will be optimal when $t > 1/2$ and more generally when the integral involved is equivalent to $\sigma_n H^{1/2}(\sigma_n, S)$ since then solving (2.1) basically amounts to solving (2.2). When this is not true, only suboptimal rates can be derived from Theorem 1 and this is the case for $t < 1/2$. Clearly this does not prove that M.C.E. are suboptimal but only that Theorem 1 leads to such rates. But we shall see, in Sect. 4, that, for some subspace of $S(0, \alpha, A)$ the actual rate is not better than $n^{-\alpha/2}$ for $\alpha < 1/2$ which shows that the theorem cannot be improved without additional assumptions. We actually do not know what is the true rate of convergence of M.L.E. for density estimation or L.S.E. for regression curve estimation with parameter space $S(0, \alpha, A)$ with $0 < \alpha < 1/2$.

3 Illustrations

We shall give here several examples of classical models for which the preceding theory works.

A) M.L.E. *for density estimation*

We observe here n i.i.d. variables $X_1, \ldots, X_n$ and we can take $W_i = 0, Z_i = X_i$. The parameter s is the density of the distribution of the X_i's with respect to some measure μ and it is known to belong to some function space S. For simplicity we shall always identify the value of the parameter and the corresponding probability. The M.L.E. is a M.C.E. with contrast function $\gamma(z, s) = -\log(s(z))$. The corresponding natural loss function is given by

$$\ell(s, t) = \mathbb{E}_s[\gamma(Z, t) - \gamma(Z, s)] = \int \log \frac{s(x)}{t(x)} s(x) d\mu(x) = K(s, t)$$

where K denotes the Kullback–Leibler information number. Unfortunately, this choice of a loss function leads to serious problems because it is unbounded. The following example shows that, even if $\sup_{t \in S} \|t/s\|_\infty < \infty$, M.L.E. could very well be divergent.

Define S to be a parametric set of densities f_θ on $[0, 2]$ with $0 \leqq \theta \leqq \frac{1}{2}$ and f_θ defined as follows

$$f_\theta(x) = 0 \text{ for } 0 \leqq x \leqq \theta\,,$$

$$f_\theta(x) = x - \theta \text{ for } \theta \leqq x \leqq 2\theta\,,$$

$$f_\theta(x) = \frac{1}{2}\left[1 + \frac{\theta^2}{2(1-\theta)^2}\right](x - 2\theta) + \theta \text{ for } 2\theta \leqq x \leqq 2\,.$$

Assume that f_0 is the true density. Clearly for $\theta > 0$, $K(f_0, f_\theta) = +\infty$. On the other hand, if the smallest observation $X_{(1)}$ is larger than 2θ, then

$$\sum_{i=1}^{n} \log f_\theta(X_i) > \sum_{i=1}^{n} \log f_0(X_i)\,.$$

The simplest solution to the problem is to assume that the family of densities is uniformly bounded from above and from below by some constants c^{-1} and c, say, and to use uniform distance to control the fluctuations of γ. Indeed we shall then have

$$|\gamma(x, t) - \gamma(x, u)| \leqq c|t(x) - u(x)|\,.$$

Taking $\tilde{\gamma} = \gamma$, $M = B = c$ and $\Delta(x, t, u) = |t(x) - u(x)|$ will clearly solve the problem when S is compact in $\mathbb{L}^\infty$. The entropy numbers which will appear will correspond to this metric structure. (A′4) will always be satisfied in this context (see the relations between Kullback information and $\mathbb{L}^2$-distances in Birgé 1983). We shall not develop on this, because we would like to avoid such assumptions and also because we know from Birgé (1983) that rates of convergence for estimators should be related to the metric structure with Hellinger metric rather than sup-norm. Indeed, when all likelihood ratios are bounded by some fixed constant A, which is clearly the case in the above setting, we know (Birgé 1983, Lemma 4.4) that for some constant $C(A) > 4$:

$$2h^2(s, t) \leqq K(s, t) \leqq C(A)h^2(s, t) \tag{3.1}$$

where h denotes Hellinger distance on densities:

$$h^2(s, t) = \frac{1}{2} \int (\sqrt{s(x)} - \sqrt{t(x)})^2 \, d\mu(x) \, .$$

This definition extends to non-negative elements of $\mathbb{L}^1(\mu)$.

This implies that, up to a multiplicative constant, $K(s, t)$ could be replaced by $h^2(s, t)$ which leads us to adopt a slightly different approach and consider the problem of maximum likelihood estimation with loss function h^2. With this loss function, it is not necessary to assume that $\| s/t \|_\infty$ is bounded but some control on the ratios t/s is clearly needed. Actually, the counterexample of Bahadur (1958) shows that, even with a compact parameter space, M.L.E. can diverge when likelihood ratios are unbounded. Our assumptions will be as follows

$$A = \sup_{t \in S} \| t/s \|_\infty < +\infty \ \bar{P}_s \text{ a.s.} \tag{Aa}$$

(Ab) For any $\delta > 0$, we can find a minimal number $N(\delta, S)$ of pairs of non-negative functions in $\mathbb{L}^1(\mu)$ (t_i^+, t_i^-), $1 \leqq i \leqq N(\delta, S)$ with $h(t_i^+, t_i^-) \leqq \delta$ and $S = \bigcup_{i=1}^{N(\delta,S)} S_\delta^i$ with $S_\delta^i = \{t \in S | t_i^- \leqq t \leqq t_i^+\}$. $N(\delta, S)$ is related to the notion of entropy with bracketing as given in Definition 3.

To put the problem in our basic framework, we choose

$$\tilde{\gamma}(z, t) = -\log \left[\frac{1}{2} (t(z) + s(z)) \right] . \tag{3.2}$$

The following lemma will be useful to check (A4):

Lemma 1. *For any pair of probabilities P and Q*

$$h(P, Q) \geqq h\left(P, \frac{P+Q}{2}\right) \geqq bh(P, Q) \, , \tag{3.3}$$

$$\textit{with } b = \left(\frac{3 - 2\sqrt{2}}{2} \right)^{1/2} \approx 0.293.$$

Proof. The left hand side inequality easily follows from convexity arguments since $h^2(P, Q) = 1 - \int \sqrt{dPdQ}$. The reverse inequality is more computational. Let us define the function $t(x)$ on $[-1, +\infty[$ by $\sqrt{1+x} = 1 + \frac{x}{2} - \frac{t(x)}{8} x^2$. $t(x)$ is a continuous decreasing function with $t(-1) = 4$, $t(0) = 1$ and $t(+\infty) = 0$. Assume that $Q \ll P$ and put $u = \frac{dQ}{dP} - 1$, then

$$\rho(P, Q) = \int \sqrt{1 + u}\, dP = 1 - \frac{1}{8} \int t(u) u^2 dP$$

since $\int u dP = 0$. The same argument applies to $\rho\left(P, \frac{P+Q}{2}\right)$ with u replaced by $\frac{1}{2}\frac{dQ + dP}{dP} - 1 = \frac{u}{2}$ and we get

$$\rho\left(P, \frac{P+Q}{2}\right) = 1 - \frac{1}{32} \int t\left(\frac{u}{2}\right) u^2 dP\,.$$

It follows that

$$h^2(P, Q)/h^2\left(P, \frac{P+Q}{2}\right) = 4 \int t(u) u^2 dP \Big/ \int t\left(\frac{u}{2}\right) u^2 dP\,.$$

One can check by elementary calculus and some numerical computations that the ratio $t(x)/t(x/2)$ is bounded by $t(-1)/t(-1/2)$ which leads to the constant in the inequality. The general case follows if we replace Q by $(1 - \varepsilon)Q + \varepsilon P$ and let ε go to zero. □

Remark. The constant cannot be improved as shown by the following example: take $\frac{dQ}{dP} = 0$ with probability $\frac{\varepsilon}{1 + \varepsilon}$ and $\frac{dQ}{dP} = 1 + \varepsilon$ with probability $\frac{1}{1 + \varepsilon}$. Then the ratio is:

$$4 \frac{t(-1)\frac{\varepsilon}{1+\varepsilon} + t(\varepsilon)\frac{\varepsilon^2}{1+\varepsilon}}{t(-1/2)\frac{\varepsilon}{1+\varepsilon} + t(\varepsilon/2)\frac{\varepsilon^2}{1+\varepsilon}} \xrightarrow[\varepsilon \to 0]{} \frac{16}{t(-1/2)}\,.$$

$t(-1/2) = 8(3 - 2\sqrt{2})$ and therefore the ratio $h^2(P, Q)/h^2\left(P, \frac{P+Q}{2}\right)$ is $\frac{2}{3} - 2\sqrt{2}$.

It is clear, since $|\log(a) - \log(b)| \leqq \sqrt{2}|a - b|$ for $a, b \geqq 1/\sqrt{2}$ that

$$|\tilde{\gamma}(x, t) - \tilde{\gamma}(x, u)| = 2\sqrt{2}\Delta(x, t, u) = 2\left|\log\sqrt{\frac{t + s}{2s}} - \log\sqrt{\frac{u + s}{2s}}\right|$$

$$\leqq 2\sqrt{2}\left|\sqrt{\frac{t + s}{2s}} - \sqrt{\frac{u + s}{2s}}\right| \leqq \frac{2}{\sqrt{s}}|\sqrt{t} - \sqrt{u}|$$

Therefore (A1) is satisfied with $M = 2\sqrt{2}$ and $B = \frac{1}{2\sqrt{2}} \log\left(\frac{1 + A}{2}\right)$. We also get $\mathbb{E}[\Delta^2(X, t, u)] \leqq h^2(t, u)$ which allows us to take $d = h$, to see that (A3) is a consequence of (Ab) and get $H(u, B_s(\sigma)) = \log N(u, B_s(\sigma))$. In order to check (A4) we note that the concavity of the logarithm implies that when $\gamma_n(t) \leqq \gamma_n(s) + \varepsilon$, then

$$\sum_{i=1}^{n} \log\left(\frac{t(Z_i) + s(Z_i)}{2}\right) \geqq \sum_{i=1}^{n} \log(s(Z_i)) - n\varepsilon/2$$

and therefore

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{\gamma}(Z_i, t) \leqq \frac{1}{n}\sum_{i=1}^{n} \tilde{\gamma}(Z_i, s) + \varepsilon/2 .$$

Using (3.1) and (3.3) we get

$$n^{-1/2}\nu_n[\tilde{\gamma}(\cdot, t) - \tilde{\gamma}(\cdot, s)] \leqq \varepsilon/2 - \mathbb{E}_s[\log(2s/(t+s))] = \varepsilon/2 - K(s, (s+t)/2)$$
$$\leqq \varepsilon/2 - 2h^2(s, (s+t)/2) \leqq \varepsilon/2 - 2b^2h^2(s, t) .$$

All assumptions are satisfied and Theorem 1 allows us to control $h(s, \hat{s})$ in terms of $N(u, B_s(\sigma))$ which is entropy with bracketing for Hellinger distance. We can also note that the effect of A is moderate since B is of order of $\log(A)$.

Remark on the parametric case. If the model is parametric, i.e. $S = \{s_\theta,\ \theta \in \Theta\}$ where Θ is a bounded set in $\mathbb{R}^k$, and is regular enough in the following sense: (Ac) for some positive α

$$D'\|\theta - \theta'\| \leqq h^\alpha(s_\theta, s_{\theta'}) \leqq D\|\theta - \theta'\|$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^k$ and D, D' are positive constants, then we can use the local entropy with bracketing with respect to Hellinger distance. This local entropy is defined to be $\log N(\delta, B_\sigma(S, h))$, where $B_\sigma(S, h)$ is the ball with radius σ with respect to the Hellinger distance h. Condition (Ac) implies (see Le Cam 1973) that $N(\delta, B_\sigma(S, h))$ is of order of 1 $\vee$ $(\delta/\sigma)^{-k/\alpha}$. Using the same arguments as above we can apply Theorem 1 with $\tilde{H}(u, \sigma) = (A'\log(\sigma/u)) \vee 1$. Therefore, if conditions (Aa) and (Ac) hold, we find that the M.L.E. $\hat{\theta}$ of θ converges to the true value of the parameter with rate $n^{-1/2}$.

B) Another contrast function for density estimation

Another simple, although not currently used, contrast function in the i.i.d case, could be, assuming that S is some compact set in $\mathbb{L}^2(\mu)$ where μ is a probability measure and $\|\cdot\|$ denotes $\mathbb{L}^2$ norm

$$\gamma(z, t) = \|t\|^2 - 2t(z) .$$

Then $\mathbb{E}_s[\gamma(z,\ t)] = \|t - s\|^2 - \|s\|^2$ which is clearly minimal for $s = t$. If we assume that S is bounded in sup-norm by some constant A and that it has finite entropy with bracketing in $\mathbb{L}^2(\mu)$, it is easy to check assumptions (A1) to (A4) with $\tilde{\gamma} = \gamma$. Clearly, the entropy property implies that $\sup_{t \in S}\|t\| = A'$ and therefore

$$|\gamma(z, t) - \gamma(z, u)| \leqq 2A'|\,\|t\| - \|u\|\,| + 2|t(z) - u(z)| .$$

We can take $M = 2$, $B = (A + A'^2)$ and $\Delta(z, t, u)$ to be one half of this upper bound. Then $d(s, t)$ is equivalent to $\mathbb{L}^2$-norm. More precisely

$$\|t - s\| \leqq d(t, u) \leqq (1 + A')\|t - u\| .$$

Since $\ell(s, t) = \|t - s\|^2$, (A4) is satisfied. Finally, (A3) only requires that S has finite entropy with bracketing in $\mathbb{L}^2(\mu)$ which is our assumption.

Remark. Another contrast function which could be used in the same context would be

$$\gamma(z, t) = -t(z)/\|t\| ,$$

but it does not lead to very attractive values of d and ℓ and we shall not develop this example here.

C) Regression models.

In this case, our observations are $Z_i = (X_i, Y_i)$, $1 \leqq i \leqq n$ with

$$Y_i = s(X_i) + W_i\,,$$

s belonging to S and the underlying variables (X_i, W_i) being independent with respective distributions $R_i \otimes Q_i$. When R_i is a Dirac measure, we get the fixed design model. Usually the W_i's will be i.i.d. with distribution Q but this is not necessary.

We shall build the estimator with the use of a function F which has the following properties:

(Ca) F is convex, non linear (but possibly piecewise linear) and for some version F' of the derivative

$$\mathbb{E}[F'(W_i)] = 0 \quad \text{for } 1 \leqq i \leqq n\,.$$

In this case, our contrast function is defined by

$$\gamma(z, t) = F(y - t(x))\,.$$

Then

$$\mathbb{E}_s[\gamma(Z, t)] = n^{-1} \sum_{i=1}^{n} \mathbb{E}[F(s(X_i) - t(X_i) + W_i)]\,.$$

The convexity of F implies that

$$F(s(X_i) - t(X_i) + W_i) \geqq F(W_i) + (s(X_i) - t(X_i))F'(W_i)$$

and the centering of the variables $F'(W_i)$ shows that γ is a contrast function with the corresponding loss function

$$\ell(s, t) = n^{-1} \sum_{i=1}^{n} \mathbb{E}[F(s(X_i) - t(X_i) + W_i) - F(W_i)]\,.$$

We shall have to put some additional restrictions on the model in order to check our assumptions:

(Cb) There exists some constant A such that for $1 \leqq i \leqq n$ and all t, u, in S,

$$|t(X_i) - u(X_i)| \leqq A \quad R_i \text{ a.s.}$$

(Cc) The function F cannot grow too fast in the sense that for some constants $a, b, w_0 > 0$

$$|F'(w - a)| \leqq b|F'(w)| \quad \text{for } w \leqq -w_0$$

$$|F'(w + a)| \leqq b|F'(w)| \quad \text{for } w \geqq w_0\,.$$

(Cd) For some constants $\alpha', \Gamma' > 0$.

$$\mathbb{E}[\exp(\alpha'|F'(W_i)|)] \leqq \Gamma' \quad 1 \leqq i \leqq n\,.$$

(Ce) If G is defined by

$$G(w, h) = F(w + h) - F(w) - hF'(w)$$

then for some constants c_0, C_0, $h_0 \geqq 0$ and $1 \leqq i \leqq n$

$$c_0 \leqq h^{-2}\mathbb{E}[G(W_i, h)] \leqq C_0 \quad \text{if } |h| \leqq h_0 . \tag{3.4}$$

Remark. The convexity of F implies that G is non-negative and non-decreasing with respect to h for $h > 0$, non-increasing for $h < 0$.

Proposition 1. *Assumptions* (Ca)–(Ce) *imply* (A1), (A2) *and* (A4) *with* $\tilde{\gamma} = \gamma$.

Proof. We shall first define M, B and Δ:

$$\begin{aligned}|\gamma(Z_i, t) - \gamma(Z_i, u)| &= |F(s(X_i) - t(X_i) + W_i) - F(s(X_i) - u(X_i) + W_i)| \\ &\leqq |t(X_i) - u(X_i)| \max(|F'(W_i - \Delta)|, |F'(W_i + \Delta)|)\end{aligned}$$

by (Cb) and the convexity of F. We can therefore define $\Delta(x, t, u) = |t(x) - u(x)|$ and $M(w)$ as the second factor. (A1) is satisfied with $B = \Delta$ and in order to check (A2), it is enough, in view of (Cd) to prove that $M(W_i) \leqq D_1 |F'(W_i)| + D_2$. We shall content ourselves to prove that

$$|F'(w + \Delta)| \leqq D_1 |F'(w)| + D_2 \quad \forall w \in \mathscr{W} ,$$

the other inequality being proved in the same way. The inequality is clearly true for $w \geqq w_0$ by repeated applications of (Cc). It is also true when $F'(w + \Delta) \leqq 0$ by monotonicity of F'. Finally, for $F'(w + \Delta) > 0$ and $w < w_0$ we can choose $D_2 = F'(w_0 + \Delta)$.

With our definition of Δ, d is associated to some $\mathbb{L}^2(\mu)$ norm with μ defined as

$$\mu = n^{-1} \sum_{i=1}^{n} R_i .$$

ℓ can be expressed as a function of G in the following way

$$\ell(s, t) = n^{-1} \sum_{i=1}^{n} \mathbb{E}[G(W_i, s(X_i) - t(X_i))] .$$

(A'4) will follow if we can show that, for $|h| \leqq \Delta$ and some positive constant c

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[G(W_i, h)] \geqq ch^2 .$$

This is true by (Ce) if $|h| \leqq h_0$ and for $\Delta \geqq |h| > h_0$, the properties of G imply that

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}[G(W_i, h)] \geqq c_0 h_0^2 \geqq c_0 h_0^2 \Delta^{-2} h^2 . \quad \square$$

Remarks. (i) In order to check (A'4), we only used one half of inequality (3.4). The other half will be useful to get lower bounds results on rates of convergence.
(ii) Clearly, apart from (Cb), all assumptions are related to the structure of the "errors" W_i and not the "explanatory variables" X_i. The results will therefore be valid for any choice of the X_i's.

In order to illustrate those assumptions, we shall consider the popular choices $F(w) = w^2$ (least squares estimators) and $F(w) = |w|$ (minimum $\mathbb{L}^1$ regression estimators).

The case of $F(w) = |w|$ is the most complicated because F' is not continuous, but it requires weaker assumptions on the errors. In this case $F'(x) = \operatorname{sgn}(x)$ where we define

$$\operatorname{sgn}(x) = 1 \quad \text{for } x \geqq 0; \quad \operatorname{sgn}(x) = -1 \text{ for } x < 0 \,.$$

One can then write

$$|x + h| = |x| + h\operatorname{sgn}(x) + 2|x + h|\mathbb{1}_{\{-1\}}(\operatorname{sgn}(hx))\mathbb{1}_{]|x|,+\infty]}(|h|)$$

and therefore

$$G(x, h) = 2|x + h|\mathbb{1}_{\{-1\}}(\operatorname{sgn}(hx))\mathbb{1}_{]|x|,+\infty[}(|h|) \,.$$

It is clear, under such circumstances that we only have to assume that the W_i's have a median zero and that (Ce) is satisfied. For $h > 0$ we get

$$\mathbb{E}[G(W_i, h)] = 2\int_{-h}^{0} (x + h)dQ_i(x)$$

from which we derive

$$hQ_i\left[-\frac{h}{2}, 0\right] \leqq \mathbb{E}[G(W_i, h)] \leqq 2hQ_i([-h, 0]) \,.$$

An analogous result holds for $h < 0$ and (Ce) will be satisfied if we assume that the distributions Q_i have densities with respect to Lebesgue measure which are uniformly bounded away from zero and from infinity in some vicinity of zero.

If $F(w) = w^2$, the assumptions are clearly satisfied if the W_i''s are centered with exponential moments which is the assumption used by Stone (1982) and slightly better than Van de Geer (1990a) who uses subgaussian tails but far from what is expected in such a situation, as shown in Nemirovskii et al. (1985). It is actually possible to substantially improve our results in this case and to weaken the moment condition (A2) when the approximating nets used for entropy computations are derived from finite dimensional linear approximations of S. Since the details are not obvious and the existence of such approximations suggests the use of some better type of estimator, namely restricted M.C.E. which will be dealt with in a subsequent work, we defer this extension to this context. Another desirable extension is to unbounded parameter sets. In the case of density estimation, the need for positivity and bounded integral plus the entropy restrictions will usually lead to bounded parameter sets. A trivial example would be the set S of densities on $[0, 1]$ satisfying the Lipschitz condition

$$|t(x) - t(y)| \leqq |x - y| \quad \forall x, y \in [0, 1] \,.$$

For regression functions, the situation is different and a more natural model would be a regression function u of the form $\theta + t(x)$ with $\theta \in \mathbb{R}$ and $t \in S$. More generally we shall assume the following model

$$Y_i = \sum_{j=1}^{q} \theta_j \varphi_j(X_i) + s(X_i) + W_i, \quad s \in S, \theta = (\theta_1, \ldots, \theta_q) \in \mathbb{R}^q$$

where the φ_j's are bounded functions. We also assume that the differences $|t(x) - u(x)|$ for t, u in S are bounded by A which is assumption (Cb) restricted to S and that the two norms defined by

$$|||\theta||| = \sup_j \theta_j; \quad \|\theta\|^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{j=1}^{q} \theta_j \varphi_j(X_i)\right)^2\right]$$

are equivalent and especially that

$$|||\theta|||^2 \leqq D \|\theta\|^2 . \tag{3.5}$$

Without loss of generality we can also assume that

$$|\varphi_j(x)| \leqq 1 \quad \forall x \in \mathscr{X}, \quad j = 1, \ldots, q . \tag{3.6}$$

Let us denote by $(\hat{\theta}, \hat{s})$ the L.S.E. in $\mathbb{R}^q \times S$ (assuming that it exists and $\varepsilon = 0$ for simplicity), and put

$$T_\theta = \sum_{i=1}^n \left[\sum_{j=1}^q \theta_j \varphi_j(X_i) \right]^2 - n\|\theta\|^2; \quad T = \sup_{\theta \in \mathbb{R}^q} (|T_\theta| / |||\theta|||^2) .$$

In this setting d is the distance in $\mathbb{L}^2(\mu)$. Let us assume without loss of generality that $\theta = 0$. Then from the definition of the L.S.E. we know that

$$\sum_{i=1}^n \left[\sum_{j=1}^q \hat{\theta}_j \varphi_j(X_i) + \hat{s}(X_i) - s(X_i) - W_i \right]^2 \leqq \sum_{i=1}^n W_i^2$$

from which we easily derive using (Cb) that

$$\left[\sum_{i=1}^n \left[\sum_{j=1}^q \hat{\theta}_j \varphi_j(X_i) \right]^2 \right]^{1/2} \leqq A\sqrt{n} + 2\left(\sum_{i=1}^n W_i^2 \right)^{1/2}$$

and by (3.5)

$$n\|\hat{\theta}\|^2 \leqq T|||\hat{\theta}|||^2 + 2A^2 n + 4 \sum_{i=1}^n W_i^2 \leqq TD\|\hat{\theta}\|^2 + 2A^2 n + 4 \sum_{i=1}^n W_i^2 . \tag{3.7}$$

Let k be chosen and assume that $T \leqq n/(2D)$ which is always true in the case of fixed design points since then $T = 0$. If $\sum_{i=1}^n W_i^2 \leqq knA^2/2$, then by (3.5)

$$|||\hat{\theta}|||^2 \leqq 4DA^2(1 + k) .$$

And also in any case

$$d((s, 0), (\hat{s}, \hat{\theta})) \leqq \|\hat{\theta}\| + A .$$

Let us consider the event $A = \{|||\hat{\theta}|||^2 \leqq 4DA^2(1 + k)\}$ and S' the parameter set $\{(\theta, s), s \in S, |||\theta|||^2 \leqq 4DA^2(1 + k)\}$. Then if A occurs the L.S.E. in S' is the global L.S.E. and we can apply our theory to S' in order to compute H and σ^* relative to S' and get bounds of the type $\mathbb{E}[d^2((s, 0), (\hat{s}, \hat{\theta})\mathbb{1}_A] \leqq C_1 \sigma^{*2}$. We also have when $T \leqq n/(2D)$

$$d^2((s, 0), (\hat{s}, \hat{\theta})) \leqq 10A^2 + \frac{16}{n} \sum_{i=1}^n W_i^2 .$$

Taking into account the fact that any moment of the variables W_i's can be bounded in terms of α' and Γ' independently of i and n, it comes from Marcinkiewicz–Zygmund's inequality and Markov's inequality that, for and any $p \geqq 2$ and $x \geqq \frac{2}{n} \sum_{i=1}^n \mathbb{E}(W_i^2)$:

$$\mathbb{P}\left[\sum_{i=1}^n W_i^2 \geqq nx \right] \leqq C(p, \alpha', \Gamma')(\sqrt{n}x)^{-p}$$

yielding, via Hölder's inequality:

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i^2 \mathbb{1}_{[\sum_{i=1}^{n} W_i^2 \geqq nx]}\right] \leqq [C(p, \alpha', \Gamma')]^{1-1/p}\left[\sum_{i=1}^{n} \| W_i \|_p\right](\sqrt{n}x)^{-p+1} .$$

Noting that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(W_i^2)$ is also bounded independently of n and choosing $p = 3$ say, we finally get that for large enough k:

$$\mathbb{E}[d^2((s, 0), (\hat{s}, \hat{\theta}))\mathbb{1}_{\Lambda^c}\mathbb{1}_{[T \leqq n/2D]}] \leqq C_2/n .$$

Since the parameter set contains an Euclidean space, σ^* cannot be of order smaller than $n^{-1/2}$ and finally we see that

$$\mathbb{E}[d^2((s, 0), (\hat{s}, \hat{\theta}))\mathbb{1}_{[T \leqq n/2D]}] \leqq C_3 \sigma^{*2} .$$

In the case of fixed design points, $T = 0$ and we are done. In general, since the X_i's are available one can check whether $T \leqq n/2D$ is true or not. In any case we get by Bernstein's inequality (see Appendix 2 and formula (6.4)) using the bound on the $|\varphi_j|$'s

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} (\varphi_i(X_i)\varphi_k(X_i) - \mathbb{E}(\varphi_j(X_i)\varphi_k(X_i))_j\right| \geqq \lambda\sqrt{n}\right] \leqq 2\exp\left[\frac{-\lambda^2}{8 + 4\lambda/3\sqrt{n}}\right].$$

Now if those differences are smaller than $\lambda\sqrt{n}$ for all j, k we get

$$T_\theta \leqq \sum_{j,k=1}^{q} |\theta_j||\theta_k|\lambda\sqrt{n} \leqq \lambda\sqrt{n}(q|\!|\!|\theta|\!|\!|)^2$$

and finally

$$\mathbb{P}[|T| \geqq \lambda\sqrt{n}q^2] \leqq 2q^2 \exp\left[\frac{-\lambda^2}{8 + 4\lambda/3\sqrt{n}}\right].$$

If we choose $\lambda = \dfrac{\sqrt{n}}{2Dq^2}$ we see that

$$\mathbb{P}[|T| \geqq n/2D] \leqq 2q^2 \exp(-C_5 n) .$$

This gives our extension to unbounded parameter spaces. A particular case could be the space of functions of bounded variation, with a variation bounded by a given fixed constant V on $[0, 1]$. Any such function can be written as $\theta + s$ with $\theta \in \mathbb{R}$ and $\sup_x|s(x)| \leqq V/2$. This would fit perfectly in our framework but leads us to the second question: how can we control the $\mathbb{L}^2$-entropy with bracketing of spaces of functions of bounded variation or other classical function spaces? A lot is known on ordinary $\mathbb{L}^2$-entropies. The necessary extensions, using classical methods of Birman and Solomjak (1967) are provided by Birgé and Massart (1993) and we shall not insist on this matter. It allows to deal with the classical Sobolev spaces on $[0, 1]^m$, say, provided that the regularity is good enough compared to m. Those multidimensional extensions provide new examples for which the integral defining $\varphi(\sigma)$ in Theorem 1 is not convergent at 0. This is the case in one dimension for α-Hölderian densities with $\alpha < 1/2$ but when the dimension increases, the split occurs for smoother functions. Roughly speaking, if we work with spaces of functions of smoothness q (not necessarily an integer) the function $\tilde{H}(u, \sigma)$ is of order $u^{-m/q}$. As soon as $m > 2q$ we see that $\varphi(\sigma)$ is of order $\sigma^{2-m/q}$ which leads to a σ^* of order $n^{-q/2m}$ instead of the optimal rate which should be $n^{-\frac{q}{2q+m}}$. Multidimensional examples show that the split occurs for differentiable functions for $m = 3$. Let us also notice that in the case of fixed design points, a better choice for $\tilde{\gamma}$ would be

$$\tilde{\gamma}(z, t) = (y - t(x))^2 - 2s(x)t(x) - t^2(x) .$$

In this case

$$\tilde{\gamma}(z, t) - \tilde{\gamma}(z, u) = 2w(u(x) - t(x))$$

and assumption (A.1) becomes much simpler with $M(w) = 2w$. This does not change anything else since then

$$\nu_n[\tilde{\gamma}(z, t) - \tilde{\gamma}(z, u)] = \nu_n[\gamma(z, t) - \gamma(z, u)] .$$

D) Ill-posed problems

The framework is similar to the one used in regression, with independent variables X_i, W_i, $1 \leqq i \leqq n$ and observations $Z_i = (X_i, Y_i)$ but now the X_i's and the elements of S belong to the same $\mathbb{L}^2$-space with respect to some measure μ with a scalar product denoted by $\langle \cdot, \cdot \rangle$. The model is given by

$$Y_i = \langle X_i, s \rangle + W_i, \quad 1 \leqslant i \leqslant n .$$

Usually, the X_i's will be deterministic but they could be random as well. We shall again deal with least squares estimation and therefore our assumptions on the W_i's will be the same as in the regression model. They should be centered variables with exponential moments. The contrast function will clearly be

$$\gamma(z, t) = (y - \langle x, t \rangle)^2 = \tilde{\gamma}(z, t)$$

leading to the loss function

$$\ell(s, t) = n^{-1} \sum_{i=1}^{n} \mathbb{E}[\langle X_i, s - t \rangle^2] .$$

We also have

$$\begin{aligned} |\gamma(z, t) - \gamma(z, u)| &= |2y - \langle x, t + u \rangle| |\langle x, u - t \rangle| \\ &= |2w + \langle x, 2s - t - u \rangle| |\langle x, u - t \rangle| \\ &\leqq (2|w| + \|x\| \, \|2s - t - u\|) |\langle x, u - t \rangle| . \end{aligned}$$

We shall assume the following

(Da) For $1 \leqq i \leqq n$, $\mathbb{E}(W_i) = 0$ and $\mathbb{E}[\exp(\alpha' |W_i|)] \leqq \Gamma'$.

(Db) $\sup_{t \in S} \|t\| \leqq A/2$; $\sup_i \|X_i\| \leqq A'$ a.s.

We can therefore choose $\Delta(x, t, u) \leqq |\langle x, t - u \rangle|$, $B = AA'$ and $M(w) = 2(|w| + AA')$. This leads to the distance $d^2(s, t) = \ell(s, t)$. For the entropy counts, it is actually enough to control the entropy with respect to the $\mathbb{L}^2$-norm $\|t - u\|$. Suppose that t and u belong to the same $\mathbb{L}^2$ ball of radius ε, then

$$\Delta^2(x, t, u) \leqq \|t - u\|^2 \leqq (2\varepsilon)^2$$

and Assumption (A3) will trivially be satisfied with $\delta = 2\varepsilon$. This is a case where only usual $\mathbb{L}^2$-entropy is needed rather than entropy with bracketing.

E) Mixture models and deconvolution

We observe here n i.i.d. variables $X_1, \ldots, X_n$ and we can take $W_i = 0$, $Z_i = X_i$. The distribution of the X_i's is assumed to have a density with respect to some measure μ which is known to be of the following form:

$$f_s(z) = \int f(\theta, z)\, ds(\theta)$$

where the parameter s belongs to some subspace S of distribution functions on $\mathbb{R}$ and $f(\cdot, \cdot)$ is a known non-negative mesurable function such that:
(Ea) f_s is a density for all $s \in S$.
(Eb) $\sup_t \| f_t \|_\infty = A < +\infty$.
(Ec) There exists non negative constants C_1, C_2, j, m, β_i and points z_i such that

$$|f_t(z) - f_u(z)| \leqq C_1 \sum_{i=1}^{m} \beta_i |t(z + z_i) - u(z + z_i)| + C_2 \| t^{(j)} - u^{(j)} \|_p .$$

We consider the contrast function

$$\gamma(z, t) = \| f_t \|_2^2 - 2f_t(z) .$$

With this choice we exactly have as in B):

$$\ell(s, t) = \| f_s - f_t \|_2^2 .$$

Defining $\tilde{\gamma}(z, t) = -2f_t(z)$, it follows from (Eb) that $\tilde{\gamma}$ satisfies condition (A1) with $\Delta(z, t, u) = |f_t(z) - f_u(z)|$, $B = A$ and $M = 2$. Now conditions (A2) and (A'4) are clearly satisfied. In order to verify (A3) we need some entropy condition. We shall here assume that 1) $\mathbb{L}^2$-entropy with bracketing and 2) $\mathbb{L}^p$-entropy are bounded (note that condition (i) is unnecessary when $C_i = 0$). We can now give three examples where this framework applies.

Mixture models. (i) $f(\theta, z)$ is bounded and differentiable with respect to θ, $\left| \frac{\partial}{\partial \theta} f(\cdot, \cdot) \right|$ is bounded by some constant b and all elements of S have their support in the same compact interval, $[0, 1]$ say. Then integrating by parts we get

$$f_t(z) - f_u(z) = -\int \frac{\partial}{\partial \theta} f(\theta, z)(t(\theta) - u(\theta))\, d\theta$$

and (Ec) is satisfied with $C_1 = 0$, $C_2 = b$, $j = 0$ and $p = 1$. It follows from Birgé and Massart (1993) that $H(\delta, S) \leqq K/\delta$. Therefore Theorem 1 implies that a M.C.E. converges to the true value of the parameter with rate $n^{-1/3}$ (with respect to the pseudo-metric $\ell^{1/2}$).

Deconvolution. (ii) $f_s(z) = \int k(z - \theta)\, ds(\theta)$, where k is a bounded density of the form:

$$k(z) = \sum_{i=1}^{m} \beta_i \mathbb{1}_{[z_i, +\infty[} + \int_{-\infty}^{z} k'(u)\, du, \quad \beta_i \geqq 0 .$$

Moreover k' belongs to $\mathbb{L}^2$ and the elements of S are continuous with support in $[0, 1]$. Integration by parts leads to

$$f_t(z) - f_u(z) = \sum_{i=1}^{m} \beta_i (t(z + z_i) - u(z + z_i)) + \int k'(z - \theta)(t(\theta) - u(\theta))\, d\theta$$

which implies by Cauchy–Schwarz that (Ec) holds with $C_1 = 1$, $C_2 = \|k'\|_2$, $j = 0$ and $p = 2$. By Birgé and Massart (1993), $\mathbb{L}^2$ entropy with bracketing of bounded decreasing functions is of order $1/\delta$ which implies rates of order $n^{-1/3}$.

We can readily see that this rate cannot be improved in general. Let us consider the set of distribution functions s such that $\int s = 1/\beta$ where β is a fixed constant, $\beta > 1$, as the parameter set S. Let μ be the Lebesgue measure on $[0, 1]$ and $f(\theta, z) = \beta \mathbb{1}_{\theta \leqq z}$, then we simply have:

$$f_s = \beta s$$

and we know from Birgé (1989) for instance that, in this situation, the $\mathbb{L}^1$-minimax risk (and therefore the $\mathbb{L}^2$-minimax risk) is bounded from below by $Kn^{-1/3}$.

(iii) The framework is the same but k instead of k' belongs to $\mathbb{L}^2$ and the elements of S have densities with respect to Lebesgue measure. Then, by Cauchy–Schwarz

$$|f_t(z) - f_u(z)| \leqq \|k\|_2 \|t' - u'\|_2 ,$$

hence (Ec) holds with $C_1 = 0$, $C_2 = \|k\|_2$, $j = 1$ and $p = 2$. The rates of convergence will be derived in the usual way from the $\mathbb{L}^2$ entropy properties of the densities of the elements of S.

Remarks. (i) This framework can be extended to the case of θ belonging to $\mathbb{R}^k$. (ii) If the likelihood ratios f_t/f_u are uniformly bounded, the same theory can be developed with the contrast function $\gamma(z, t) = -\log f_t(z)$ leading to maximum likelihood deconvolution. Unfortunately, the loss function will then be $K(f_s, f_t)$ or equivalently $\| f_s - f_t \|_2^2$ which can be much smaller than $\|s' - t'\|_2^2$.

4 Suboptimality of minimum contrast estimators

As we already mentioned, the rates of convergence given by Theorem 1 in the case of a "large" parameter set for which the integral $\int_0^1 H^{1/2}(x, S)\,dx$ is divergent at zero are not the optimal ones, as defined by the general theory in Birgé (1983). Therefore the question naturally arises of knowing whether the proof of Theorem 1 is suboptimal or the estimator itself is. The answer will be given in this section where we shall exhibit sets S for which the "bad" rate announced by Theorem 1 is actually the true one, for two different models: density estimation and regression. This does not mean that M.C.E. are always suboptimal when the entropy condition does not hold. We do not know anything about it, but if M.C.E. are optimal when the entropy condition does not hold, it is clear that some additional circumstance occurs, which is not a consequence of the assumptions of Theorem 1. Under the assumptions we use, it is impossible to derive better rates of convergence and those could very well be the right ones in some particular situations. In particular, even on a convex set of densities with uniformly bounded likelihood ratios, M.L.E. can be suboptimal.

The main reason for this trouble is the following. The solution δ_n of Eq. (2.2) only involves $H(x, S)$ for $x \geqq \delta_n$. Even if $H(x, S)$ is explosive for x smaller than δ_n, this will not change anything to the optimal rate for n observations. On the contrary, the integral in Theorem 1 depends on values of $H(x, S)$ for x smaller than σ_n. As a consequence, M.C.E. are definitely unstable and will be sensitive to perturbations of S which only involves $H(x, S)$ for $x < \delta_n$. This is not the case for optimal estimators as shown in Birgé (1984).

A) Construction of the parameter spaces

In order to define a convenient parameter space, let us start with the basic functions

$$\bar{f}(x) = x\mathbb{1}_{[0;\,1/4[}(x) + 1/4\,\mathbb{1}_{[1/4;\,3/4]}(x) + (1-x)\mathbb{1}_{]3/4;\,1]}(x)\,,$$

$$v(x) = x\mathbb{1}_{[0;\,1/2]}(x) + (1-x)\mathbb{1}_{]1/2;\,1]}(x)\,.$$

We shall then define on $[-1/2; 1/2]$ and $[-\eta;\eta]$ respectively with $\lambda, \theta, \eta > 0$

$$f_\lambda(x) = \lambda(\bar{f}(2x) - \bar{f}(-2x)); \quad v_{\theta,\eta} = \theta(v(x/\eta) - v(-x/\eta))\,,$$

which clearly satisfy

$$\|f_\lambda\|_\infty = \lambda/4; \quad \|v_{\theta,\eta}\|_\infty = \theta/2; \quad \int f_\lambda(x)\,dx = \int v_{\theta,\eta}(x) = 0$$

$$\int f_\lambda^2(x)\,dx = \lambda^2/24; \quad \int v_{\theta,\eta}^2(x)\,dx = \eta\theta^2/2\,.$$

We shall say that the set $\mathbf{y} = \{y_1, \ldots, y_k\}$ satisfies the property $A(\eta)$ if all intervals $]y_i - \eta; y_i + \eta[$, $1 \leqq i \leqq k$ are disjointed and included into the set $J = [-\frac{3}{8}; -\frac{1}{8}] \cup [\frac{1}{8}; \frac{3}{8}]$. Here k is an arbitrary positive integer.

Assuming that the following relationships hold with constants c, $0 < c \leqq 1$ and α, $0 < \alpha \leqq 1$.

$$\theta = c\lambda^2; \quad \eta^\alpha = 2^{\alpha-1}\theta; \quad 0 \leqq \lambda \leqq 1/4\,,$$

which implies that $\theta \leqq \lambda/4$ and $\eta \leqq 1/8$, we can define for $\mathbf{y} \in A(\eta)$,

$$w_{\lambda,\mathbf{y}}(x) = \sum_{i=1}^{k} v_{\theta,\eta}(y_i - x)$$

and notice that whatever λ, λ', $\int f_\lambda(x) w_{\lambda',\mathbf{y}} = 0$. We shall define the parameter spaces Θ_α and $\bar{\Theta}_\alpha$. Θ_α is the set of all densities of the form $1 + f_\lambda(x) + w_{\lambda,\mathbf{y}}(x)$ where $0 \leqq \lambda \leqq 1/4$ and $\mathbf{y}$ is any vector of any dimension satisfying $A(\eta)$ with η as above. $\bar{\Theta}_\alpha$ will be the convex hull of Θ_α, i.e. the set of g's of the form

$$\begin{aligned} g &= 1 + \sum_{l=1}^{L} \mu_l(f_{\lambda_l}(x) + w_{\lambda_l,\mathbf{y}_l}(x)),\ \mu_l \geqq 0,\ \sum_{l=1}^{L} \mu_l = 1 \\ &= 1 + f_{\bar\lambda}(x) + \sum_{l=1}^{L} \mu_l w_{\lambda_l,\mathbf{y}_l}(x) \end{aligned}$$

with $\bar\lambda = \sum_{l=1}^{L} \mu_l \lambda_l$.

Proposition 2. *Denoting* $f(x) = g(x) - 1$, *we have for all* g *in* $\bar{\Theta}_\alpha$

(i) $\|f\|_\infty \leqq 3\bar\lambda/8 \leqq 3/32$; $|f_{\bar\lambda}(x)| \leqq 2|f(x)| \leqq 3|f_{\bar\lambda}(x)|$;

(ii) $\|f\|_2^2 \geqq \bar\lambda^2/24$;

(iii) $h^2(1, g) \geqq \bar\lambda^2/210$;

(iv) $|g(x) - g(y)| \leqq |x - y|^\alpha$ *for any* x, y *in* $[-1/2, 1/2]$.

Proof. (i) One has $|f_\lambda(x) + w_{\lambda,\mathbf{y}}(x)| \leqq 3\lambda/8$ since $\theta \leqq \lambda/4$ and the first inequality follows by convexity. The second uses the same arguments since $|w_{\lambda,\mathbf{y}}(x)| \leqq \frac{1}{2}|f_\lambda(x)|$.

(ii) Follows from orthogonality between f_λ and $w_{\lambda',\mathbf{y}}(x)$, whatever λ, λ' and $\mathbf{y}$ are and (iii) from Lemma 4.1 in Birgé (1983) and our upper bound on g.

(iv) The assumption $A(\eta)$ implies that f_λ is constant on the supports of the $v_{\theta,\eta}(y_i - \cdot)$ which are themselves disjointed. Considering that the largest monotonous parts of f_λ and $v_{\theta,\eta}$ have respective lengths 1/4 and η and the maximal slopes are 2λ and θ/η we have to check that

$$2\lambda\delta + \theta\eta^{-1}\varepsilon \leqq (\varepsilon + \delta)^\alpha \quad \text{for } 0 \leqq \delta \leqq 1/4; \quad 0 \leqq \varepsilon \leqq \eta .$$

Since $(\varepsilon + \delta)^\alpha \geqq 2^{\alpha-1}(\varepsilon^\alpha + \delta^\alpha)$ by concavity and $\varepsilon \leqq \varepsilon^\alpha \eta^{1-\alpha}$, it is enough to check that $2\lambda\delta + \theta\eta^{-\alpha}\varepsilon^\alpha \leqq 2^{\alpha-1}(\varepsilon^\alpha + \delta^\alpha)$ which clearly follows from our upper bound on λ. The result follows by convexity. □

Then it follows from our proposition that $\bar{\Theta}_\alpha$ is a convex set of α-Hölderian densities. We shall use $\bar{\Theta}_\alpha$ as a parameter space to derive lower bounds on rates of convergence of M.C.E. and demonstrate the limitations of those estimators. Since $\bar{\Theta}_\alpha$ is a subset of the set of α-Hölderian functions, from Birgé (1983) we know that, at least in the i.i.d. case, rates of convergence of optimal estimators should be at least $n^{\frac{-\alpha}{2\alpha+1}}$, which could be readily checked with suitable histograms. Similar results also hold in the regression framework. Computations below will show that M.C.E. cannot reach this rate for $\alpha < 1/2$.

B) Maximum likelihood estimators

Let us now consider what happens when we use M.L.E. on $\bar{\Theta}_\alpha$ with $c = 1$. Assume the observations $X_1, \ldots, X_n$ are i.i.d. and uniform on $[-\frac{1}{2}; \frac{1}{2}]$ and $\hat{g}_n$ is the M.L.E.

Theorem 3. *There exists some positive constant ε such that*

$$\liminf_n P[h(1; \hat{g}_n) \geqq \varepsilon(n \log(n))^{-\alpha/2}] \geqq 1/2 .$$

An immediate consequence of this theorem is the fact that, as soon as $\alpha < 1/2$, the rate of convergence of M.L.E. is not better than $(n \log(n))^{-\alpha/2}$ which is suboptimal as compared to $n^{\frac{-\alpha}{2\alpha+1}}$. The rate is actually bounded by $n^{-\alpha/2}$, as follows from Sect. 3. The $\log(n)$ gap is likely to be due to our proof and we believe that the actual rate is $n^{-\alpha/2}$.

The proof of the theorem relies in a crucial way on the following technical lemma (see the proof in the Appendix) about spacings of uniform random variables.

Lemma 2. *Let $X_1, \ldots, X_n$ be n independant uniform variables on $[-1/2; 1/2]$ and N_n the cardinal of the maximal subset S of $\{X_1; \cdots; X_n\}$ such that S satisfies $A\left(\frac{0.2}{n+1}\right)$ and for any pair X_i, X_j with $X_i \in S$ and $X_j \notin S$, $|X_i - X_j| \geqq \frac{0.2}{n+1}$. Then*

$$P\left[N_n \geqq \frac{n}{8}\right] \xrightarrow[n \to +\infty]{} 1 .$$

Proof of Theorem 3. Let $g(x) = 1 + f(x)$ be an element of $\bar{\Theta}_\alpha$. We already noticed that $\|f\|_\infty \leqq 3/32$. Now for $|x| \leqq 3/32$ we can write ($x^{-2}[\log(1 + x) - x]$ being increasing) with $a = 0.47$, $b = 0.534$

$$x - bx^2 \leqq \log(1 + x) \leqq x - ax^2 .$$

This implies that the likelihood function can be written

$$\sum_{j=1}^{n} \log g(X_j) = \sum_{j=1}^{n} f(X_j) - \sum_{j=1}^{n} c_j f^2(X_j); \quad a \leqq c_j \leqq b\,.$$

It follows from Proposition 2(i) that

$$\frac{1}{4} f_{\bar\lambda}^2(X_j) \leqq f^2(X_j) \leqq \frac{9}{4} f_{\bar\lambda}^2(X_j)$$

then

$$\sum_{j=1}^{n} \log g(X_j) = \bar\lambda \sum_{j=1}^{n} f_1(X_j) + \sum_{l=1}^{L} \mu_l \sum_{j=1}^{n} w_{\lambda_l, y_l}(X_j) - M_n \bar\lambda^2 \sum_{j=1}^{n} f_1^2(X_j)$$

with $a/4 \leqq M_n \leqq 9b/4$ and finally

$$\sum_{j=1}^{n} \log g(X_j) = \frac{\bar\lambda\sqrt{n}}{2\sqrt{6}} S_n - \frac{\bar\lambda^2 n}{24} M_n T_n + \sum_{l=1}^{L} \mu_l W_{n,l}$$

where

$$S_n \xrightarrow{\mathscr{L}} N(0,1); \quad 0.11 \leqq M_n \leqq 1.202; \quad T_n \xrightarrow{\text{a.s.}} 1; \quad W_{n,l} = \sum_{j=1}^{n} w_{\lambda_l, y_l}(X_j)\,.$$

Claim. For $\varepsilon > 0$, *there exists* K_ε *such that uniformly over* $\bar\Theta_\alpha$ *with probability larger than* $1 - \varepsilon$

$$\sum_{l=1}^{L} W_{n,l} \leqq K_\varepsilon n \bar\lambda (\log(n)/n)^{\alpha/2}$$

and therefore for large n and some K with great probability

$$\sum_{j=1}^{n} \log g(X_j) \leqq K n \bar\lambda (\log(n)/n)^{\alpha/2}.$$

On the other hand, Lemma 2 proves the existence of a subset of size N_n of the X_j's, which we shall denote by $\{X_{j_1}, \dots, X_{jN_n}\}$ such that if $3\eta/2 \leqq \frac{0.2}{n+1}$ *the density*

$$\bar g(x) = 1 + f_\lambda(x) + \sum_{i=1}^{N_n} v_{\theta,\eta}\left(X_{j_i} + \frac{\eta}{2} - x\right)$$

belongs to Θ_α and asymptotically $N_n \geqq \frac{n}{8}$ with a probability close to 1. Then, with large probability

$$\sum_{j=1}^{n} \log \bar g(X_j) \geqq -\lambda\sqrt{n} - \frac{n\lambda^2}{19.9} + \frac{n\lambda^2}{16}\,.$$

Choosing $\eta = \frac{1}{8n}$ in the definition of $\bar g$, we find that the corresponding λ is given by $\lambda^2 = (4n)^{-\alpha}/2$. The log-likelihood for $\bar g$ is therefore of order $n^{1-\alpha}$ which implies that the M.L.E. is obtained for values of $\bar\lambda$ larger than $\delta(n\log(n))^{-\alpha/2}$ for some $\delta > 0$, hence the result.

There remains to prove our claim. Let us first consider the following situation of n i.i.d. uniform variables on $[a, b]$ with common c.d.f. F and empirical distribution F_n and let $Z_n = \sqrt{n}(F_n - F)$ be the normalized empirical process. Let also u be a fixed absolutely continuous function defined on $[0, 1]$ and satisfying $u(0) = u(1) = 0$ and $\int_0^1 u(x)\,dx = 0$. We shall consider the following set $\mathscr{W}_\alpha$ of functions defined on $[a, b]$ for $0 < \alpha \leqq 1$:

$$\mathscr{W}_\alpha = \left\{ w(x) = \lambda^2 \sum_{i=1}^{k} u((y_i - x)/\eta),\ 0 < \lambda \leqq 1/4, \eta^\alpha = 2^{1-\alpha}\lambda^2 \right\}$$

where $\{y_i\}_{0 \leqq i \leqq k}$ is a sequence satisfying

$$y_0 = 0; \quad y_{i+1} \geqq y_i + \eta; \quad y_k \leqq b; \quad k \leqq (b-a)/\eta\ .$$

Under such conditions, for each $\varepsilon > 0$ there exists some K_ε such that

$$\mathbb{P}\left[\sup_{w \in \mathscr{W}_\alpha} \frac{1}{n\lambda} \sum_{j=1}^{n} w(X_j) \geqq K_\varepsilon (n^{-1/2} \log(n))^\alpha \right] \leqq \varepsilon\ . \tag{4.1}$$

Indeed, since $\mathbb{E}[w(X)] = 0$ by assumption, we have to control $n^{1/2} \sup_w \lambda^{-1} Z_n(w)$. But clearly, integrating by parts we get

$$\begin{aligned} Z_n(w) &= \lambda^2 \sum_{i=1}^{k} \int_{y_i - \eta}^{y_i} u((y_i - x)/\eta)\, dZ_n(x) \\ &= \frac{\lambda^2}{\eta} \sum_{i=1}^{k} \int_{y_i - \eta}^{y_i} u'((y_i - x)/\eta) Z_n(x)\, dx \\ &= \lambda^2 \sum_{i=1}^{k} \int_{y_i - \eta}^{y_i} \eta^{-1} u'((y_i - x)/\eta)[Z_n(x) - Z_n(y_i)]\, dx \\ &\leqq \lambda^2 k \sup_{|x-y| \leqq \eta} |Z_n(x) - Z_n(y)| \int_0^1 |u'(x)|\, dx \\ &\leqq \lambda^2 (b-a)/\eta \, \|u'\|_1 \sup_{|x-y| \leqq \eta} |Z_n(x) - Z_n(y)|\ . \end{aligned}$$

Now the inequality by Mason, Shorack and Wellner (see Shorack and Wellner 1986, p. 545) shows that, for $\eta > 0$, $t > 0$

$$\mathbb{P}\left[\sup_{|x-y| \leqq \eta} |Z_n(x) - Z_n(y)| \geqq t \right] \leqq \frac{K_1}{\eta} \exp\left[-\frac{K_2 t^2}{\eta + t/\sqrt{n}} \right],$$

where the K_i's denote different constants. Therefore, summing those probabilities for values of η of the form $2^j/n, j \geqq 0$ with $t = K_3(\eta/2\log(n))^{1/2}$ we get for K_3 large enough

$$\mathbb{P}\left[\exists \eta \geqq \frac{1}{n}, \sup_{|x-y| \leqq \eta} |Z_n(x) - Z_n(y)| \geqq K_3 (\eta \log(n))^{1/2} \right] \leqq \varepsilon\ .$$

From this we deduce that with probability larger than $1 - \varepsilon$, uniformly for $\lambda \geqq (\log(n)/n)^{\alpha/2}$

$$\sqrt{n} Z_n(w) \leqq K_4 \lambda^{2 - 1/\alpha} (n \log(n))^{1/2} \leqq n\lambda K_4 (\log(n)/n)^{\alpha/2}\ .$$

On the other hand since $\sum_{j=1}^{n} w(X_j) \leqq n\|u\|_\infty \lambda^2$, for $\lambda \leqq (\log(n)/n)^{\alpha/2}$

$$\sqrt{n}\, Z_n(w) \leqq n\lambda \|u\|_\infty (\log(n)/n)^{\alpha/2} ,$$

which proves (4.1) from which our claim follows. □

Remark. As a by-product of our proof, we get a lower bound for convergence of empirical processes. Let G_α be the set of functions g defined on $[-1/2; 1/2]$ and such that

$$|g(x) - g(y)| \leqq |x-y|^\alpha ; \ \|g\|_\infty \leqq 1 .$$

Let P be the uniform probability on $[-1/2; 1/2]$ and P_n the empirical measure derived from n i.i.d. variables with distribution P. Now, all g's of the form $g(x) = \sum_{i=1}^{k} v_{\theta,\eta}(y_i - x)$ with $\eta^\alpha = 2^{1-\alpha}\theta$ and the y_i's satisfying $A(\eta)$ are in G_α and centered for P. But there exists some

$$\bar{g}(x) = \sum_{i=1}^{N_n} v_{\theta,\eta}\left(X_{ji} + \frac{\eta}{2} - x \right)$$

defined as above such that with probability tending to 1, $P_n \bar{g} \geqq \frac{1}{32}(4n)^{-\alpha}$. This implies that the classical $n^{-1/2}$ rate of convergence in the T.C.L. cannot be true over the class G_α for $\alpha < \frac{1}{2}$, a result which goes back to Bakhvalov. Another consequence is the fact that the ratio $\sup_{g \in G_\alpha} \dfrac{|P_g - P_n g|}{\|g\|_2}$ will be bounded away from zero, for any α in $]0, 1]$.

C) *Regression*

Let us now consider a regression framework of the type $Y_j = g(X_j) + \varepsilon_j$ as described in Sect. 3, where g belongs to Θ_α, the ε_j's are i.i.d. and independent of the X_j's which are themselves independent and uniformly distributed over $[-1/2, 1/2]$. Given some convex function F satisfying

$$\mathbb{E}(F'(\varepsilon_j)) = 0,\ \mathbb{E}(|F'(\varepsilon_j)|) = \sigma',\ \operatorname{var}(F'(\varepsilon_j)) = \sigma^2 ,$$

we shall define our estimator $\hat{g}_n$ to be the minimizer over Θ_α of the quantity $\sum_{j=1}^{n} F(Y_j - g(X_j))$ or equivalently of $\sum_{j=1}^{n} [F(Y_j - g(X_j)) - F(\varepsilon_j)]$. As before we assume that the true value of g is 1, therefore the quantity of interest can be written with $g = 1 + f$ as

$$\begin{aligned} R_n(g) &= \sum_{j=1}^{n} [F(Y_j - g(X_j)) - F(\varepsilon_j)] = \sum_{j=1}^{n} [F(\varepsilon_j - f(X_j)) - F(\varepsilon_j)] \\ &= -\sum_{j=1}^{n} f(X_j) F'(\varepsilon_j) + \sum_{j=1}^{n} G(\varepsilon_j, -f(X_j)) \\ &= -\lambda \sum_{j=1}^{n} f_1(X_j) F'(\varepsilon_j) - \sum_{j=1}^{n} w_{\lambda,y}(X_j) F'(\varepsilon_j) + \sum_{j=1}^{n} G(\varepsilon_j, -f(X_j)) . \end{aligned}$$

Using the properties of G and the same arguments as before we get

$$R_n(g) = \frac{\lambda\sqrt{n}\sigma}{2\sqrt{6}} S_n + n\lambda^2 M_n T_n + nM'_n V_n$$

where

$$S_n \xrightarrow{\mathscr{L}} N(0,1); \quad 0 \leqq M_n \leqq \frac{9}{4} C_0 \| f_1 \|_\infty^2 \leqq \frac{9C_0}{64};$$

$$T_n \xrightarrow{\text{a.s.}} 1; \quad V_n \xrightarrow{\text{a.s.}} 1; \quad |M'_n| \leqq \sigma'\theta/2$$

and then for large n with large probability

$$R_n(g) \geqq -\lambda\sqrt{n\sigma} - n\sigma'\theta/2 .$$

On the other hand, using again Lemma 2 we can define $\bar{g}$ belonging to Θ_α if $3\eta/2 \leqq \frac{0.2}{n+1}$ by

$$\bar{g}(x) = 1 + f_\lambda(x) + \sum_{i=1}^{N_n} v_{\theta,\eta}\left(X_{j_i} + \frac{\eta}{2}\frac{F'(\varepsilon_{j_i})}{|F'(\varepsilon_{j_i})|} - x\right)$$

where $P[N_n \geqq \frac{n}{8}] \xrightarrow[n \to +\infty]{} 1$. Therefore, asymptotically with a large probability we get

$$R_n(\bar{g}) \leqq \lambda\sqrt{n\sigma} + \frac{nC_0}{7}\lambda^2 - \theta/2 \sum_{i=1}^{N_n} |F'(\varepsilon_{j_i})| .$$

Since the ε_j's and X_j's are independent, the last term obeys the law of large numbers which implies that with large probability

$$R_n(\bar{g}) \leqq \lambda\sqrt{n\sigma} + \frac{nC_0}{7}\lambda^2 - \frac{n\sigma'}{16}\theta .$$

If we define Θ_α by $\theta = \frac{3C_0}{\hat{\sigma}}\lambda^2$, we can just conclude as before and get:

Theorem 4 *In the above regression framework, the rate of convergence of* M.C.E. *over Θ_α will satisfy for some positive ε*

$$\liminf_n P[h(1, \hat{g}_n) \geqq \varepsilon n^{-\alpha/2}] \geqq 1/2 .$$

Remark. With a slight modification of the proof we can extend the result to the case where the sequence $\{X_j\}_{j \geqq 1}$ satisfies Lemma 2, possibly with different values of the constants. This will be the case if, for example, the X_j's are deterministic and equispaced or i.i.d. with a density with respect to Lebesgue measure, which is bounded away from zero and infinity on a subinterval of J.

Unstability of the risk of M.C.E. estimators. The phenomenon is well-known, at least for the M.L.E., and we shall not go into great details but just give some illustrative example derived from our previous construction.

Let Θ_L be the set of Lipschitz densities on $[-1/2, 1/2]$ which means that $|g(x) - g(y)| \leqq |x - y|$ for g in Θ_L . In this case the optimal rate of convergence is $n^{-1/3}$ and it follows from Birgé (1984) that the rate will not change if we replace for each n, Θ_L by Θ_n such that any element of Θ_n is at a distance of Θ_L smaller than $n^{-1/3}$. This is due to the fact that for n observations the risk is, in this case, only determined by the metric structure relative to balls of radius $n^{-1/3}$ or larger.

Modifying the structure for smaller balls will only slightly affect the minimax risk as shown in Birgé (1984). Unfortunately, our proofs of convergence of M.C.E. do involve smaller balls in a crucial way through a chaining argument. The following illustration will show that this is not only a defect of the proof but a real drawback of the estimator.

For each n, we define a subset Θ_n of our set Θ_1 by $\theta = \lambda^2, \eta = 2c_n^2 n^{-5/3}, c_n \leqq 1$. Clearly, any element g of Θ_n is within $(n\eta\theta^2/2)^{1/2} \leqq \frac{c_n}{8} n^{-1/3}$ in $\mathbb{L}^2$-norm, (and similarly in Hellinger distance) of a Lipschitz density $1 + f_\lambda$. Now the arguments used in the proof of Theorem 3 show that with large probability when n goes to infinity we shall have uniformly over Θ_n

$$\sum_{j=1}^{n} \log g(X_j) \leqq \lambda\sqrt{n} + \frac{n\theta}{2} = \lambda\sqrt{n} + \frac{n\lambda^2}{2}$$

and for some $\bar{g}$

$$\sum_{j=1}^{n} \log \bar{g}(X_j) \geqq -\lambda\sqrt{n} - \frac{n\lambda^2}{19.9} + \frac{n\theta}{16} = -\lambda\sqrt{n} + \frac{n\lambda^2}{82} .$$

The M.L.E. will occur for large values of λ although the distance between Θ_n and Lipschitz densities can be made arbitrarily small by a proper choice of c_n. Of course, this implies that perturbations $v_{\theta,\eta}$ will be rather wild. If we assume some smoothness on them, various rates of convergence may occur but in the worst case above, the M.L.E. will not even be consistent.

5 Appendix 1

The purpose of this appendix is to provide a proof of Lemma 2. The following large deviation inequality for uniform spacings will be helpful.

Lemma 3 *Let $U_{(0)} = 0, U_{(1)}, \ldots, U_{(n)}, U_{(n+1)} = 1$ be the order statistics corresponding to n* i.i.d. *variables, uniformly distributed on $[0, 1]$ and $V_i = U_{(i+1)} - U_{(i)}$, $0 \leqq i \leqq n$ be the corresponding spacings. Then, setting $J = [a/(n+1), 1]$, we have for any positive ε, δ*

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_J(V_{i-1})\mathbb{1}_J(V_i) < bn\right) \leqq \exp(-(n+1)(\delta - \log(1+\delta)))$$

$$+ 2\exp(-(n-1)\varepsilon^2) \tag{5.1}$$

with $b = \exp(-2a(1+\delta)) - \varepsilon$.

Proof. Let us first recall some classical facts about large deviation theory.

- If S is binomial $\mathscr{B}(n, p)$ and $0 \leqq \varepsilon$, then

$$\mathbb{P}(S - np \geqq n\varepsilon) \leqq \exp(-2n\varepsilon^2) \tag{5.2}$$

(see Massart 1990, for a more precise inequality).

- If $W_0, W_1, \ldots, W_n$ are i.i.d. exponential r.v.'s with parameter 1, by the Cramér–Chernoff inequality, we get for $\varepsilon < 0$

$$\mathbb{P}\left(\sum_{i=0}^{n} (W_i - 1) > (n+1)\varepsilon\right) \leqq (1+\varepsilon)^{n+1}\exp(-(n+1)\varepsilon) . \tag{5.3}$$

It is well known that the joint distribution of the V_i's is the same as the joint distribution of the $(W_i/\sum_{j=0}^{n} W_j)$'s, so that the left-hand side of (5.1) is equal to

$$\mathbb{P}_0 = \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}_I(\tilde{W}_i) < bn\right)$$

with $\tilde{W}_i = W_i \wedge W_{i-1}$ and $I = [a\bar{W}, 1]$, where

$$\bar{W} = \sum_{j=0}^{n} W_j/(n+1)\,.$$

Now, assuming that $n = 2m + 1$ is odd, it is clear that the event $\{\sum_{i=1}^{n} \mathbb{1}_I(\tilde{W}_i) \geqq bn\}$ is included in the intersection of the events

$$\{\bar{W} \leqq t\}, \left\{\sum_{i=1}^{m+1} \mathbb{1}_{[at,1]}(\tilde{W}_{2i-1}) \geqq b(m+1)\right\}, \quad \left\{\sum_{i=1}^{m} \mathbb{1}_{[at,1]}(\tilde{W}_{2i}) \geqq bm\right\}.$$

Therefore, an upper bound for $\mathbb{P}_0$ is

$$\mathbb{P}(\bar{W} \geqq t) + \mathbb{P}\left(\sum_{i=1}^{m+1} \mathbb{1}_{[0,at[}(\tilde{W}_{2i-1}) \geqq (1-b)(m+1)\right)$$
$$+ \mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_{[0,at[}(\tilde{W}_{2i}) \geqq (1-b)m\right).$$

Assuming $t = 1 + \delta$, we can bound the first term using (5.3) by

$$\exp(-(n+1)(\delta - \log(1+\delta)))\,.$$

For the second term, we use (5.2) after convenient centering by $p = \mathbb{P}(\tilde{W}_j < at) = 1 - e^{-2at}$, since the corresponding $\tilde{W}_j$'s are independent. Assuming that $\varepsilon = 1 - b - p = e^{-2at} - b$, we get the bound $\exp(-2(m+1)\varepsilon^2)$. The last term can be bounded analogously by $\exp(-2m\varepsilon^2)$. The conclusion follows, the case of $n = 2m$ being similar. □

Proof of Lemma 2. Assuming that M_n observations fall in $[-3/8, -1/8]$, let us order them and denote them $Y_{(1)}, \ldots, Y_{(M_n)}$. With $Y_{(0)} = -3/8$ and $Y_{(M_n+1)} = -1/8$, the variables $U_{(i)} = 4(Y_{(i)} + 3/8)$ satisfy the assumptions of Lemma 3 with M_n replacing n. Clearly $Y_{(i)}$ will belong to S if $V_{i-1} \wedge V_i \geqq 4(0.4/(n+1))$. Let N'_n be the number of those $Y_{(i)}$'s, it is enough, using the same arguments for the interval $[1/8, 3/8]$ to prove that $\mathbb{P}(N'_n \geqq n/16) \to 1$ as $n \to +\infty$. Now for $M_n + 1 \leqq \beta(n+1)$, we have

$$N'_n \geqq \sum_{i=1}^{M_n} \mathbb{1}_J(V_{i-1})\mathbb{1}_J(V_i); \quad J = [1.6\beta/(M_n+1), 1]$$

and if $M_n \geqq \alpha n$, $(M_n/(16\alpha)) \geqq n/16$. This allows us to conclude that

$$\mathbb{P}(N'_n \geqq n/16) \geqq \mathbb{P}\left(\sum_{i=1}^{M_n} \mathbb{1}_J(V_{i-1})\mathbb{1}_J(V_i) \geqq M_n/(16\alpha), M_n \in \mathcal{J}_n\right)$$
$$\geqq \mathbb{E}\left(\mathbb{P}\left(\sum_{i=1}^{M_n} \mathbb{1}_J(V_{i-1})\mathbb{1}_J(V_i) \geqq M_n/(16\alpha) \mid M_n\right)\mathbb{1}_{\mathcal{J}_n}(M_n)\right)$$
$$\geqq \min_{m \in \mathcal{J}_n} \mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_J(V_{i-1})\mathbb{1}_J(V_i) \geqq m/(16\alpha)\right) P(M_n \in \mathcal{J}_n)$$

where $\mathscr{I}_n\{m \in \mathbb{N} : m \geqq \alpha n$ and $m + 1 \leqq \beta(n + 1)\}$. Now if $\alpha < 1/4 < \beta$, the last factor converges to 1. The first one can be bounded from below using Lemma 3 with $a = 1.6\beta$ and $b = 1/(16\alpha)$ which clearly allows ε, δ to be positive for convenient choices of α and β. □

6 Appendix 2

The purpose of this Appendix is to provide a proof of Theorems 1 and 2. This proof is based on a control of the modulus of continuity of the empirical process $v_n = \sqrt{n}(P_n - \bar{P}_s)$ indexed by the family $\{\tilde{\gamma}(\cdot, t) - \tilde{\gamma}(\cdot, s), t \in S\}$. Without loss of generality we assume throughout the proofs that $\tilde{\gamma}(\cdot, s) = 0$.

Proposition 3 *Suppose that assumptions* (A1), (A2) *and* (A3) *hold and* $B = 1$. *Then, for any positive* a, σ *and* λ *such that* $\sigma \leqq 1 \leqq a$ *and*

$$(6.1) \qquad (K\sqrt{\Gamma}/\alpha) \int_{\lambda\alpha/(64\Gamma\sqrt{n})}^{a\sigma} \max(1, H^{1/2}(u, B_s(\sigma)))du < \lambda$$

$$(6.2) \qquad \lambda \leqq 32(a\Gamma/\alpha)\sigma^2\sqrt{n}$$

the following bound is available

$$(6.3) \qquad \mathbb{P}^*\left(\sup_{t \in B_s(\sigma)} |v_n(\tilde{\gamma}(\cdot, t) - \tilde{\gamma}(\cdot, s))| \geqq \lambda\right) \leqq 8\exp(-4\alpha^2\lambda^2/(K^2a^2\Gamma\sigma^2))$$

where $K = 1920$.

In order to prove Proposition 3 we shall use a chaining argument with adaptive truncatures. This technique was intiated by Bass (1985) in the context of set-indexed partial sum processes and then used by Ossiander (1988) and next by Andersen et al. (1989) in order to prove uniform central limit theorems for function-indexed empirical processes.

Since we shall use Bernstein's inequality repeatedly we prefer to recall the precise statement of this inequality. To do this we need to give some notation and definition.

Definition 4 Let ρ and c be positive constants. We define $\mathscr{H}(\rho, c)$ to be the set of random variables X such that $\mathbb{E}|X|^m \leqq (m!/2)\rho c^{m-2}$, for any integer $m \geqq 2$.

We can now state Bernstein's inequality (see Shorack and Wellner 1986, p. 855).

Lemma 4 *Let* $X_1, \ldots, X_n$ *be independent random variables which are centered at expectations. Let* $S_n = X_1 + \cdots + X_n$ *and assume that* X_k *belongs to* $\mathscr{H}(\rho_k, c)$ *for any* $1 \leqq k \leqq n$. *Then, for any positive* λ

$$\mathbb{P}(|S_n| \geqq \lambda\sqrt{n}) \leqq 2\exp(-\lambda^2/2(v + (c\lambda/\sqrt{n})))$$

where $v = n^{-1}\sum_{k=1}^n \rho_k$.

Remark. It is very easy to verify that if X belongs to $\mathscr{H}(\rho, c)$ then $X - \mathbb{E}X$ belongs to $\mathscr{H}(4\rho, 2c)$. Thus, if the variables X's have the same properties as in Lemma 4 but are no longer assumed to be centered, the following exponential inequality is still available:

$$(6.4) \qquad \mathbb{P}(|S_n - \mathbb{E}S_n| \geqq \lambda\sqrt{n}) \leqq 2\exp(-\lambda^2/4(2v + (c\lambda/\sqrt{n}))) .$$

We are now in position to prove Proposition 3.

Proof of Proposition 3 Let us keep in mind that $\Delta(x, t, u) \leqq 1$ since $B = 1$ and fix the value of the constant $K = 1920$. Now using a regularization argument, it is enough to prove (6.3) if we assume the following condition to be fulfilled instead of (6.1)

$$(6.5) \qquad (K\sqrt{\Gamma}/\alpha) \int_{\lambda\alpha/(64\Gamma\sqrt{n})}^{a\sigma} H^{1/2}(u)du \leqq \lambda$$

where H is a continuous and strictly decreasing function such that $\max(1, H(\delta, B_s(\sigma))) \leqq H(\delta)$ for any positive δ and $H(\delta) \uparrow +\infty$ as $\delta \downarrow 0$. Since H is decreasing (6.5) and (6.2) imply that $\lambda \geqq (K\sigma\sqrt{\Gamma}/\alpha)H^{1/2}(a\sigma)(a - a\sigma/2)$ which means (since $\sigma \leqq 1$) that $\lambda^2 \geqq (\Gamma/4)(aK\sigma/\alpha)^2 H(a\sigma)$ and thus $H(a\sigma) \leqq 4\alpha^2\lambda^2/(\Gamma K^2 a^2 \sigma^2)$. So, since H is continuous and strictly decreasing with $H(\delta) \uparrow +\infty$ as $\delta \downarrow 0$ we may define $\delta_0 = H^{-1}(4\alpha^2\lambda^2/(\Gamma K^2 a^2\sigma^2))$ and we have

$$(6.6) \qquad \delta_0 \leqq a\sigma .$$

Next we define $\delta_i = 2^{-i}\delta_0$ for any nonnegative integer i. Since s is fixed we shall omit to mention explicitly (in the notations) the dependence with respect to s of the various quantities we have to deal with. So, for any $t \in S$, let $f_t = \tilde{\gamma}(\cdot, t)$ (we recall that $\tilde{\gamma}(\cdot, s) = 0$). Now, since (A3) holds for $B_s(\sigma)$, for each nonnegative integer i, we may choose a covering $\mathscr{S}_i = \{S_{1,i}, \ldots, S_{J_i,i}\}$ of $B_s(\sigma)$ with $J_i \leqq \exp H(\delta_i)$ and

$$\mathbb{E}_s\left[\left(\sup_{t,u \in S_{j,i}} \Delta^2(X, t, u)\right)^*\right] \leqq \delta_i^2, \quad \text{for } 1 \leqq j \leqq J_i .$$

For any positive δ we set $\mathbb{H}(\delta) = \sum_{\delta_k \geqq \delta} H(\delta)$. This function $\mathbb{H}$ will be useful because the mapping $(\pi_0, \ldots, \pi_k)$ below ranges in a finite set with cardinality bounded by $\exp(\mathbb{H}(\delta_k))$. For each pair (j, i) with $i \geqq 0$, $1 \leqq j \leqq J_i$ choose some point $s_{j,i}$ in $S_{j,i}$ and define the mappings π_i from S to $\{1, \ldots, J_i\}$ in such a way that t belongs to $S_{j,i}$ whenever $\pi_i(t) = j$. Setting $t_i = s_{\pi_i(t),i}$ and $T_i = S_{\pi_i(t),i}$ we then define

$$\Pi_k f_t = \max_{i \leqq k}\left(f_{t_i} - M\left(\sup_{u \in T_i} \Delta(\cdot, u, t_i)\right)^*\right)$$

and

$$\Delta_k(f_t) = \min_{i \leqq k}\left(\sup_{u \in T_i} \Delta(\cdot, u, t_i)\right)^* .$$

It follows from those definitions that

$$(6.7) \qquad \Pi_k f_t \leqq f_t \leqq \Pi_k f_t + 2M\Delta_k(f_t)$$

and

$$(6.8) \qquad \mathbb{E}_s(\Delta_k^2(f_t)) \leqq \delta_k^2 .$$

We finally set $\mathscr{F} = \{f_t, t \in B_s(\sigma)\}$. From now on we shall write "f" instead of "f_t" for short and "$\Delta_k f$" instead of "$\Delta_k(f)$". We need to define a few more parameters. Let $A = 48\sqrt{\Gamma}/\alpha$, and for any integer k

$$(6.9) \qquad \eta_k = A\delta_k \mathbb{H}(\delta_k)^{1/2} ,$$

$$(6.10) \qquad a_k = (2\Gamma/\alpha)\sqrt{n}\delta_k^2/\eta_{k+1} .$$

Let $N = \min\{k \geqq 0 : \delta_k \leqq (\alpha/(16\Gamma))\lambda/\sqrt{n}\}$ and for any $f \in \mathscr{F}$,

$$\mu(f) = (\min\{k \geqq 0 : \Delta_k f > a_k\}) \wedge N .$$

When $N \geqq 1$, the following decomposition is available:

$$f = \Pi_0 f + (f - \Pi_N f) + \sum_{k=0}^{N-1} (\Pi_{k+1} f - \Pi_k f) . \tag{6.11}$$

Now, summating by parts we get

$$\sum_{k=0}^{N-1} (\Pi_{k+1} f - \Pi_k f)\mathbb{1}_{\mu(f) < k+1} = \sum_{k=0}^{N-1} (f - \Pi_k f)\mathbb{1}_{\mu(f) = k} + (\Pi_N f - f)\mathbb{1}_{\mu(f) < N} .$$

Then, plugging this identity in Eq. (6.11), we finally get the decomposition that we shall use below, when $N \geqq 1$

$$f = \Pi_0 f + \sum_{k=0}^{N} (f - \Pi_k f)\mathbb{1}_{\mu(f) = k} + \sum_{k=0}^{N-1} (\Pi_{k+1} f - \Pi_k f)\mathbb{1}_{\mu(f) \geqq k+1} ; \tag{6.12.a}$$

when $N = 0$, we shall simply use

$$f = \Pi_0 f + (f - \Pi_0 f) . \tag{6.12.b}$$

Proof of Proposition 3 when $N \geqq 1$. The following inequality derives straightforwardly from (6.12.a)

$$\mathbb{P}^*\left(\sup_{f \in \mathscr{F}} |\nu_n(f)| \geqq \lambda\right) \leqq \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3 ,$$

where

$$\mathbb{P}_1 = \mathbb{P}\left(\sup_{f \in \mathscr{F}} |\nu_n(\Pi_0 f)| \geqq \lambda/8\right) ,$$

$$\mathbb{P}_2 = \mathbb{P}^*\left(\sup_{f \in \mathscr{F}} \sum_{k=0}^{N} |\nu_n((f - \Pi_k f)\mathbb{1}_{\mu(f) = k})| \geqq 3\lambda/8 + 4\sum_{k=1}^{N} \eta_k\right) ,$$

$$\mathbb{P}_3 = \sum_{k=0}^{N-1} \mathbb{P}\left(\sup_{f \in \mathscr{F}} |\nu_n((\Pi_{k+1} f - \Pi_k f)\mathbb{1}_{\mu(f) \geqq k+1})| > \eta_{k+1}\right)$$

provided that the following inequality holds:

$$\sum_{k=1}^{N} \eta_k \leqq \lambda/10 . \tag{6.13}$$

In order to control the probabilities $\mathbb{P}_i$, $i = 1, 2, 3$ it is useful to notice some elementary relations based on the fact that a_k and Δ_k are non-increasing

$$\{\mu(f) = k\} \subset \{a_k < \Delta_k f \leqq a_{k-1}\} , \text{ for } 1 \leqq k \leqq N - 1 , \tag{6.14}$$

$$\{\mu(f) = 0\} = \{a_0 < \Delta_0 f\}, \{\mu(f) = N\} \subset \{\Delta_N f \leqq a_{N-1}\} .$$

Control of $\mathbb{P}_1$. We simply note that $|(\Pi_0 f)(z_i)| \leqq M(w_i)((\Delta_0 f)(x_i) + \Delta(x_i, s, t_0))$. Now x_i and w_i are independent, $P^i(M^m) \leqq \Gamma m!/\alpha^m$ by (A2) and $\Delta \leqq 1$ by (A1), hence

$$(\Pi_0 f)(z_i) \in \mathscr{H}((4/\alpha^2)\Gamma(P^i(\Delta_0^2 f + \Delta^2(s, t_0)), 2/\alpha) .$$

Therefore, bound (6.4) yields

$$\mathbb{P}_1 \leqq 2\exp(H(\delta_0))\exp(-2^{-6}\alpha^2\lambda^2/(32\Gamma(\sigma^2 + \delta_0^2) + \alpha\lambda/\sqrt{n}))\,,$$

thus, using (6.2), (6.6) and the fact that $a \geqq 1$, we get

$$\mathbb{P}_1 \leqq 2\exp(H(\delta_0))\exp(-2^{-13}\alpha^2\lambda^2/(\Gamma\sigma^2 a^2))$$

which finally gives, using the definition of δ_0 and the crude bound $K^2 > 2^{16}$,

$$\mathbb{P}_1 \leqq 2\exp(-H(\delta_0))\,.$$

Control of $\mathbb{P}_2$. We note that $|g| \leqq h$ implies

$$|\nu_n(g)| \leqq |\nu_n(h)| + 2\sqrt{n}\bar{P}(h) \tag{6.15}$$

therefore, bound (6.7) yields

$$|\nu_n(f - \Pi_k f)\mathbb{1}_{(\mu(f)=k)}| \leqq 2|\nu_n(M(\Delta_k f)\mathbb{1}_{(\mu(f)=k)})| + 4\sqrt{n}\bar{P}(M(\Delta_k f)\mathbb{1}_{(\mu(f)=k)})\,;$$

thus, since M and $(\Delta_k f)\mathbb{1}_{(\mu(f)=k)}$ are independent and $P^i(M) \leqq \Gamma/\alpha$, we get

$$|\nu_n(f - \Pi_k f)\mathbb{1}_{(\mu(f)=k)}| \leqq 2R_k(f) + 4\sqrt{n}(\Gamma/\alpha)E_k(f) \tag{6.16}$$

where

$$R_k(f) = |\nu_n(M(\Delta_k f)\mathbb{1}_{(\mu(f)=k)})|$$

and

$$E_k(f) = \bar{P}((\Delta_k f)\mathbb{1}_{(\mu(f)=k)})\,.$$

Control of E_k. For each integer k such that $0 \leqq k \leqq N - 1$, we use (6.14), (6.8) and (6.10) to obtain the bound

$$E_k(f) \leqq \bar{P}((\Delta_k f)\mathbb{1}_{(\Delta_k f > a_k)}) \leqq \delta_k^2 a_k \leqq (\alpha/(2\Gamma))\eta_{k+1}/\sqrt{n}$$

and for $k = N$, we simply use the Cauchy–Schwarz inequality and the definition of δ_N to get

$$E_N(f) \leqq \bar{P}(\Delta_N(f)) \leqq \delta_N \leqq (\alpha/(16\Gamma))\lambda/\sqrt{n}\,.$$

Control of R_k. For each integer k such that $1 \leqq k \leqq N$, we note that (6.14) implies for any integer $m \geqq 2$, because of (A2)

$$\begin{aligned}\mathbb{E}((M(\Delta_k f)\mathbb{1}_{(\mu(f)=k)})(z_i))^m) &\leqq \mathbb{E}(M^m(w_i)(\Delta_k^m f)(x_i)\mathbb{1}_{((\Delta_k f)(x_i) \leqq a_{k-1})}) \\ &\leqq P^i(M^m)P^i(\Delta_k^2 f)a_{k-1}^{m-2} \leqq (\Gamma/\alpha^2)P^i(\Delta_k^2 f)(a_{k-1}/\alpha)^{m-2}m!\,.\end{aligned}$$

Therefore $(M(\Delta_k f)\mathbb{1}_{(\mu(f)=k)})(z_i) \in \mathscr{H}((2/\alpha^2)\Gamma P^i(\Delta_k^2 f), a_{k-1}/\alpha)$. Thus using bound (6.4) and (6.8), we get

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_k(f) \geqq \eta_k\right) \leqq 2\exp\mathbb{H}(\delta_k)\exp\left(\frac{\alpha^2\eta_k^2}{4(4\Gamma\delta_k^2 + \alpha\eta_k a_{k-1}/\sqrt{n})}\right).$$

Therefore, using (6.10), our bound for R_k becomes

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_k(f) \geqq \eta_k\right) \leqq 2\exp(\mathbb{H}(\delta_k) - \eta_k^2\alpha^2/(48\Gamma\delta_k^2))$$

and (6.9) yields

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_k(f) \geqq \eta_k\right) \leqq 2\exp(-\mathbb{H}(\delta_k)(A^2\alpha^2/(48\Gamma)-1)).$$

Finally, it follows from the monotonicity of the function H and the definition of the constant A that

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_k(f) \geqq \eta_k\right) \leqq 2\exp(-(k+1)H(\delta_0)).$$

When $k=0$, we simply use $\Delta_0 f \leqq 1$ and next proceed in the same way as above to verify that the variable $(M(\Delta_0 f)\mathbb{1}_{(\mu(f)=0)})(z_i)$ belongs to $\mathscr{H}((2/\alpha^2)\Gamma P^i(\Delta_0^2 f), 1/\alpha)$. Now the inequalities (6.4) and (6.8) yield

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_0(f) \geqq \lambda/16\right) \leqq 2\exp\left(H(\delta_0) - \frac{2^{-8}\alpha^2\lambda^2}{(16\Gamma\delta_0^2 + \alpha\lambda/(4\sqrt{n}))}\right).$$

Using (6.2) and (6.6), we get

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_0(f) \geqq \lambda/16\right) \leqq 2\exp\left(H(\delta_0) - \frac{2^{-12}\alpha^2\lambda^2}{\Gamma\sigma^2(a^2 + a/2))}\right).$$

We conclude exactly as we did for the control of $\mathbb{P}_1$ that

$$\mathbb{P}\left(\sup_{f\in\mathscr{F}} R_0(f) \geqq \lambda/16\right) \leqq 2\exp(-H(\delta_0)).$$

Collecting the above estimates, we get from (6.16) that

$$\begin{aligned}\mathbb{P}_2 &\leqq 2\exp(-H(\delta_0))\left(\sum_{k=0}^{N}\exp(-kH(\delta_0))\right)\\ &\leqq 2\exp(-H(\delta_0))(1-\exp(-H(\delta_0)))^{-1}.\end{aligned}$$

Taking into account that $\mathbb{P}_2 \leqq 1$, this bound finally becomes

$$\mathbb{P}_2 \leqq 3\exp(-H(\delta_0)).$$

Control of $\mathbb{P}_3$. We note that for any integer k such that $0 \leqq k \leqq N-1$, the inclusion $\{\mu(f) \geqq k+1\} \subset \{\Delta_k f \leqq a_k\}$ holds. Then we can verify in the same way as for the control of $\mathbb{P}_2$ above that $((\Pi_{k+1}f - \Pi_k f)\mathbb{1}_{\mu(f)\geqq k+1})(z_i)$ belongs to $\mathscr{H}((16/\alpha^2)\Gamma P^i(\Delta_k^2(f) + \Delta_{k+1}^2(f)), 4a_k/\alpha)$. So, applying bound (6.4) again gives

$$\mathbb{P}_3 \leqq 2\sum_{k=0}^{N-1}\exp(2\mathbb{H}(\delta_{k+1}))\exp\left(-\frac{\alpha^2\eta_{k+1}^2}{4(160\Gamma\delta_{k+1}^2 + 4\alpha a_k\eta_{k+1}/\sqrt{n})}\right).$$

Thus, using (6.10) we get

$$\mathbb{P}_3 \leqq 2\sum_{k=0}^{N-1}\exp(2\mathbb{H}(\delta_{k+1}) - \alpha^2\eta_{k+1}^2/(768\Gamma\delta_{k+1}^2)),$$

and then by (6.9)

$$\mathbb{P}_3 \leqq 2\sum_{k=0}^{N-1}\exp(-\mathbb{H}(\delta_{k+1})((\alpha^2A^2/(768\Gamma)) - 2)).$$

Now taking into account the definition of A, we conclude by the same way as we did for the control of $\mathbb{P}_2$ and finally get

$$\mathbb{P}_3 \leqq 2 \sum_{k=0}^{N-1} \exp(-\mathbb{H}(\delta_{k+1})) \leqq 3\exp(-H(\delta_0)) .$$

End of the proof of Proposition 3 when $N \geqq 1$. Collecting the above bounds for the probabilities $\mathbb{P}_i, i = 1, 2, 3$ we get

$$\mathbb{P}^*\left(\sup_{f\in\mathscr{F}}|\nu_n(f)| \geqq \lambda\right) \leqq 8\exp(-H(\delta_0))$$

provided that condition (6.13) is fulfilled. It only remains to show that (6.5) implies condition (6.13). This will be a straightforward consequence of the following claim.

Claim 1

$$\sum_{k=1}^{N} \delta_k(\mathbb{H}(\delta_k))^{1/2} \leqq 4 \int_{\delta_{N-1}}^{\delta_0} H^{1/2}(u)du .$$

Proof of Claim 1

$$\sum_{k=1}^{N} \delta_k\left(\sum_{j\leqq k} H(\delta_j)\right)^{1/2} \leqq \sum_{k=1}^{N} \delta_k\left(\sum_{j\leqq k} H^{1/2}(\delta_j)\right).$$

Recalling that $\delta_j = 2^{-j}\delta_0$, we have

$$\sum_{k=1}^{N} \delta_k\left(\sum_{j\leqq k} H(\delta_j)\right)^{1/2} \leqq \sum_{j=0}^{N} H(\delta_j)^{1/2}\left(\sum_{k\geqq j} \delta_k\right) \leqq 2 \sum_{j=0}^{N} \delta_j H(\delta_j)^{1/2} ;$$

thus, since H is non increasing

$$\sum_{k=1}^{N} \delta_k\left(\sum_{j\leqq k} H(\delta_j)\right)^{1/2} \leqq 4\left(\sum_{j=0}^{N} \int_{\delta_{j+1}}^{\delta_j} H^{1/2}(u)du\right)$$

which proves Claim 1.

Now, using (6.9), we can deduce from Claim 1 that

$$\sum_{k=1}^{N} \eta_k \leqq 4A \int_{\delta_{N-1}}^{\delta_0} H^{1/2}(u)\,du .$$

So, noticing that since $N \geqq 1, \delta_{N-1} > \alpha\lambda/(16\Gamma\sqrt{n})$ and using (6.6)

$$\sum_{k=1}^{N} \eta_k \leqq 4A \int_{\alpha\lambda/(64\Gamma\sqrt{n})}^{a\sigma} H^{1/2}(u)du$$

which clearly means (taking into account the definitions of the constants A and K) that (6.5) implies condition (6.13). Hence the proof of Proposition 3 is complete when $N \geqq 1$.

Proof of Proposition 3 when $N = 0$. We first recall that $N = 0$ means

$$\delta_0 \leqq \alpha\lambda/(16\Gamma\sqrt{n}) . \tag{6.17}$$

Next, it follows from (6.12b) that

$$\mathbb{P}^*\left(\sup_{f\in\mathscr{F}}|\nu_n(f)| \geqq \lambda\right) \leqq \mathbb{P}_1 + \mathbb{P}_2 \tag{6.18}$$

where

$$\mathbb{P}_1 = \mathbb{P}\left(\sup_{f\in\mathscr{F}}|\nu_n(\Pi_0 f)| \geqq \lambda/8\right), \quad \mathbb{P}_2 = \mathbb{P}\left(\sup_{f\in\mathscr{F}}|\nu_n(f - \Pi_0 f)| \geqq 3\lambda/4\right).$$

The control of $\mathbb{P}_1$ does not differ from the one that we have already performed when $N \geqq 1$ and gives

$$\mathbb{P}_1 \leqq 2\exp(-H(\delta_0)).$$

In order to bound $\mathbb{P}_2$, we use (6.15) which leads to

$$\mathbb{P}_2 \leqq \exp(H(\delta_0)) \sup_{f\in\mathscr{F}} \mathbb{P}(2|\nu_n(M\Delta_0 f)| \geqq 3\lambda/4 - 4\sqrt{n}\bar{P}(M\Delta_0 f)).$$

Now $P^i(M) \leqq \Gamma/\alpha$ and, using (6.17), we get

$$\bar{P}(M\Delta_0 f) \leqq \Gamma\delta_0/\alpha \leqq \lambda/(16\sqrt{n}).$$

Hence

$$\mathbb{P}_2 \leqq \exp(H(\delta_0)) \sup_{f\in\mathscr{F}} \mathbb{P}(|\nu_n(M\Delta_0 f)| \geqq \lambda/4).$$

But it is easy to verify that $M\Delta_0 f(z_i)$ belongs to $\mathscr{H}(2\Gamma\delta_0^2/\alpha^2, 1/\alpha)$ hence using bound (6.4) we get

$$\mathbb{P}_2 \leqq 2\exp(H(\delta_0))\exp(-2^{-5}\alpha^2\lambda^2/(8\Gamma\delta_0^2 + \alpha\lambda/(2\sqrt{n})))$$

and then, by (6.2) and (6.6)

$$\mathbb{P}_2 \leqq 2\exp(H(\delta_0))\exp(-2^{-8}\alpha^2\lambda^2/(3\Gamma\sigma^2 a^2))$$

which implies (using the definition of δ_0 and the crude bound $K^2 > 3\cdot 2^{11}$)

$$\mathbb{P}_2 \leqq 2\exp(-H(\delta_0)).$$

Collecting the above estimates, we finally get from (6.18)

$$\mathbb{P}^*\left(\sup_{f\in\mathscr{F}}|\nu_n(f)| \geqq \lambda\right) \leqq 4\exp(-H(\delta_0)).$$

Hence the proof of Proposition 3 has been completed. □

Proof of Theorem 2. Let us first assume that $B = 1$. For any integer j, let $S_j = \{t \in S : 2^{2j}\sigma^{*2} \leqq d^2(s,t) < 2^{2(j+1)}\sigma^{*2} \wedge 1\}$. Then, since $S = (\cup_{j\geqq L} S_j) \cup B_s(2^L\sigma^*)$, we have

$$\mathbb{P}^*\left(\sup_{t\in S}\frac{|\nu_n(\tilde{\gamma}(\cdot,t))|}{d^2(s,t) \vee 2^{2L}\sigma^{*2}} \geqq \sqrt{n}/(2C)\right) \leqq \mathbb{P}_1 + \mathbb{P}_2$$

where

$$\mathbb{P}_1 = \sum_{j\geqq L} \mathbb{P}^*\left(\sup_{t\in S_j}|\nu_n(\tilde{\gamma}(\cdot,t))| \geqq \sqrt{n}\,2^{2j}\sigma^{*2}/(2C)\right)$$

and

$$\mathbb{P}_2 = \mathbb{P}^*\left(\sup_{t \in B_s(2^L\sigma^*)} |\nu_n(\tilde{\gamma}(\cdot, t))| \geqq \sqrt{n}2^{2L}\sigma^{*2}/(2C) \right).$$

Control of $\mathbb{P}_1$. To bound $\mathbb{P}_1$ we can assume that $2^{2L}\sigma^{*2} \leqq 1$ (otherwise $\mathbb{P}_1 = 0$). Now, for any integer j such that $S_j \neq \emptyset$ we apply Proposition 3 with $\lambda = \sqrt{n}\,2^{2j}\sigma^{*2}/(2C)$ and $\sigma^2 = 2^{2(j+1)}\sigma^{*2} \wedge 1$. We may do this only if conditions (6.1) and (6.2) are fulfilled. These conditions will be a fortiori fulfilled if the following inequalities hold

$$(K\sqrt{\Gamma}/\alpha)\varphi(2^j\sigma^*) \leqq \sqrt{n}\,2^{2j}\sigma^{*2}/(2C) \tag{6.19}$$

and

$$2^{2j}\sigma^{*2} \leqq (64a\Gamma C/\alpha)(2^{2(j+1)}\sigma^{*2} \wedge 1)\,. \tag{6.20}$$

Noting that $S_j \neq \emptyset$ implies that $(2^{2(j+1)}\sigma^{*2}) \wedge 1 \geqq 2^{2j}\sigma^{*2}$ we see that (6.20) is fulfilled whenever $64a\Gamma C/\alpha \geqq 1$ which is precisely ensured by our choice of a. On the other hand, since the function $\varphi(\sigma)/\sigma^2$ is decreasing, inequality (6.19) is implied by Eq. (2.1). So, it comes from (6.3) that

$$\mathbb{P}_1 \leqq 8 \sum_{j \geqq L} \exp\left(-\frac{\alpha^2 2^{2j} n\sigma^{*2}}{4\Gamma K^2 a^2 C^2} \right).$$

Since $\tilde{H} \geqq 1 \geqq \sigma^*$, (2.1) implies that $n\sigma^{*2} \geqq b$. Thus,

$$\mathbb{P}_1 \leqq 8 \sum_{j \geqq L} \exp(-9 \cdot 2^{2j} n\sigma^{*2}/(4b)) \leqq 8 \sum_{k \geqq 0} \exp(-9 \cdot 2^{2L} 2^{2k} n\sigma^{*2}/(4b))$$

which finally implies since $2^{2k} \geqq 1 + 2k$ and $n\sigma^{*2} \geqq b$

$$\mathbb{P}_1 \leqq 8.09 \exp(-9 \cdot 2^{2L} n\sigma^{*2}/(4b))\,.$$

Control of $\mathbb{P}_2$. To bound $\mathbb{P}_2$ we use Proposition 3 again with $\lambda = \sqrt{n}2^{2L}\sigma^{*2}/(2C)$ and $\sigma^2 = 2^{2L}\sigma^{*2} \wedge 1$. Conditions (6.1) and (6.2) will be fulfilled if

$$(K\sqrt{\Gamma}/\alpha) \int_{\alpha\sigma^{*2}/(128\Gamma C)}^{a2^L\sigma^*} \tilde{H}^{1/2}(u, L\sigma^*)du \leqq \sqrt{n}\,2^{2L}\sigma^{*2}/(2C) \tag{6.21}$$

and

$$2^{2L}\sigma^{*2} \leqq (64a\Gamma C/\alpha)(2^{2L}\sigma^{*2} \wedge 1)\,. \tag{6.22}$$

Clearly (6.19) with $j = L$ implies (6.21). On the other hand $2^{2L}\sigma^{*2} \leqq \Gamma C/\alpha$ and $a = 1 \vee (\alpha/(64\Gamma C))$ so that (6.22) is fulfilled. Thus we may apply inequality (6.3), which gives

$$\mathbb{P}_2 \leqq 8\exp(-9K^2 2^{2L} n\sigma^{*2}/(4b))\,.$$

Now, collecting the above estimates for $\mathbb{P}_1$ and $\mathbb{P}_2$ and taking into account the facts that $n\sigma^{*2} \geqq b$ and $K = 1920$ straightforward calculations yield (6.18), thus completing the proof of Theorem 2 when $B = 1$. Let us now consider the general case of $B \neq 1$. We can always go back to our special case setting $\Delta'(x, t, u) = B^{-1}\Delta(x, t, u)$ and $M'(w) = BM(w)$. Then we can derive assumptions (A1), (A2) and (A3) for Δ' and M' changing the constants to $B' = 1$, $\alpha' = B^{-1}\alpha$, $\Gamma' = \Gamma$ and $H(\delta, S)$ into $\underline{H}'(\delta, S) = \underline{H}(B\delta, S)$ with a new distance $d'(s, t) = B^{-1}d(s, t)$. This implies that $\tilde{H}'(u, \sigma) = \tilde{H}(Bu, B\sigma)$. Defining $C' = B^{-2}C$ and $a' = 1 \vee \left(\dfrac{\alpha'}{64\Gamma' C'} \right) = a$ we can apply the proof for the case $B = 1$ to the setting with the primes everywhere and get the conclusion (6.18) provided that

$2^{2L}\sigma^{*\prime 2} \leqq C'\Gamma'/\alpha'$ where $\sigma^{*\prime}$ is the solution of

$$\sqrt{n}\sigma^{*\prime 2} = \frac{2KC'\sqrt{\Gamma'}}{\alpha'} \int_{\frac{\alpha'\sigma^{*\prime 2}}{128\Gamma'C'}}^{2a'\sigma^{*\prime}} \tilde{H}'^{1/2}(u, 2\sigma^{*\prime})du\ .$$

Going back to the original setting without primes and making a change of variable inside the integral leads to the conclusion that $B\sigma^{*\prime}$ is the solution σ^* of Eq. (2.1). This allows to translate (6.18) which is true for the prime values into the original setting with $\sigma^{*\prime} = B^{-1}\sigma^*$. It is easy to see that removing the primes does not affect the probability. The exponent should therefore remain unchanged which leads, since $b'^2 = 9a'^2K^2C'^2\Gamma'/\alpha'^2 \leqq n\sigma^{*\prime 2}$ to

$$\sigma^{*2}/b^2 = \frac{\sigma^{*\prime}\alpha'^2}{9a'^2K^2C'^2\Gamma'}\ .$$

This gives the final value of b together with $b^2 \leqq n\sigma^{*2}$. The conclusion follows since $2^{2L}\sigma^{*\prime 2} \leqq C'\Gamma'/\alpha'$ is equivalent to $2^{2L}\sigma^{*2} \leqq BC\Gamma/\alpha$. □

We are now in position to prove Theorem 1.

Proof of Theorem 1 Since $d(s, t) \leqq B$ and $C \geqq B\alpha/\Gamma$ we can always assume that

$$\lambda\sigma^{*2} \leqq BC\Gamma/\alpha\ . \tag{6.23}$$

Let Ω_λ be the event

$$|\nu_n(\tilde{\gamma}(\cdot, t))| \leqq \sqrt{n}(d^2(s, t) \vee (\lambda\sigma^{*2}))/(2C), \text{ for any } t \in S\ .$$

Now, let $T_\varepsilon(\omega) = \{t \in S : \gamma_n(t) \leqq \gamma_n(s) + \varepsilon\}$, then for any $\omega \in \Omega_\lambda$ and any $t \in T_\varepsilon(\omega)$, we have because of (A4)

$$d^2(s, t)/C \leqq -n^{-1/2}\nu_n(\tilde{\gamma}(\cdot, t)) + \varepsilon \leqq (d^2(s, t) \vee (\lambda\sigma^{*2}))/(2C) + \varepsilon$$

which means that for any $\omega \in \Omega_\lambda$ and any $t \in T_\varepsilon(\omega)$, we have

$$d^2(s, t) \leqq (2C\varepsilon) \vee (\lambda\sigma^{*2})\ .$$

Next we note that since we may assume that $8 \cdot 1 \exp(-\lambda/2) \leqq 1$, we may a fortiori assume that $\lambda \geqq 1$. So, we can define L to be the integer such that $2^{2L} \leqq \lambda < 2^{2(L+1)}$. Because of (6.23), we have $2^{2L}\sigma^{*2} \leqq BC\Gamma/\alpha$ thus, we can use Theorem 2 and get

$$\mathbb{P}^*(\Omega_\lambda^c) \leqq 8.1 \exp(-\lambda/2)$$

implying Theorem 1. □

References

Andersen, N.T., Giné, E., Ossiander, M., Zinn, J.: The central limit theorem and the law of iterated logarithm for empirical processes under local conditions. Probab. Theory Relat. Fields **77**, 271–305 (1988)

Bahadur, R.R.: Examples of inconsistency of maximum likelihood estimates. Sankhyã, Ser. A **20**, 207–210 (1958)

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D.: Statistical inference under order restrictions. New York: Wiley 1972

Bass, R.F.: Law of the iterated logarithm for set-indexed partial sum processes with finite variance. Z. Warscheinlichkeitstheor. Verw. Geb. **70**, 591–608 (1985)

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

Brigé, L.: Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. Ann. Inst. Henri Poincaré, Sect. B **20**, 201–223 (1984)

Birgé, L.: On estimating a density using Hellinger distance and some other strange facts. Probab. Theory Relat. Fields **71**, 271–291 (1986)

Birgé, L.: The Grenander estimator: a nonasymptotic approach. Ann. Stat. **17**, 1532–1549 (1989)

Birgé, L., Massart, P.: The $\mathbb{L}^2$ entropy with bracketing. Manuscript (1993)

Birman, M.S., Solomjak, M.Z.: Piecewise-polynomial approximation of functions of the classes W_p. Mat. Sb. **73**, 295–317 (1967)

Donoho, D.L.: Statistical estimation and optimal recovery. Technical report. University of California, Berkeley (1989)

Dudley, R.M.: A course on empirical processes. In: Ecole d'Eté de Probabilités de Saint-Flour XII-1982. Berlin Heidelberg New York: Springer 1984

Efroimovitch, S.Yu., Pinsker, M.S.: Estimation of square-integrable density on the basis of a sequence of observations. Probl. Inf. Transm. **17**, 182–196 (1981)

Grenander, U.: On the theory of mortality measurement. II. Skand. Aktuarietidskr. **39**, 125–153 (1956)

Grenander, U.: Abstract inference. New-York: Wiley 1980

Groeneboom, P.: Estimating a monotone density. In: LeCam, L.M., Olshen, R.A. (eds.)Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. 2, pp. 539–555. Monterey, CA: Wadsworth 1985

Huber, P.J.: The behavior of maximum likelihood estimates under non-standard conditions. In: Proc. 5th Berkeley Symp. Math. Stat. Probab., vol. 1, pp. 221–233. Berkeley, CA: University of California Press 1967

Ibragimov, I.A., Has'minskii, R.Z.: On estimate of the density function. Zap. Nauchn. Semin. Leningr. Otd. Mat. Inst. Steklova **98**, 61–85 (1980)

Ibragimov, I.A., Has'minskii, R.Z.: Estimation of distribution density. J. Sov. Math. **21**, 40–57 (1983)

Kolmogorov, A.N., Tikhomirov, V.M.: ε-entropy and ε-capacity of sets in function spaces. Transl., II. Ser., Am. Math. Soc. **17**, 277–364 (1961)

Le Cam, L.M.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**, 38–53 (1973)

Le Cam, L.M.: Asymptotic methods in statistical decision theory. Berlin Heidelberg New-York: Springer 1986

Lorentz, G.G.: Approximation of functions. New-York: Holt, Rinehart, Winston 1966

Massart, P.: The tight constant in the D.K.W. inequality. Ann. Probab. **18**, 1269–1283 (1990)

Nemirovskii, A.S. Polyak, B.T., Tsybakov, A.B.: Signal processing by the nonparametric maximum-likelihood method. Probl. Inf. Transm. **20**, 177–192 (1984)

Nemirovskii, A.S., Polyak, B.T., Tsybakov, A.B.: Rate of convergence of nonparametric estimates of maximum-likelihood type. Probl. Inf. Transm. **21**, 258–272 (1985)

Ossiander, M.: A central limit theorem under metric entropy with L_2 bracketing. Ann. Probab. **15**, 897–919 (1987)

Pfanzagl, J.: On the measurability and consistency of minimum contrast estimates. Metrika **14**, 249–272 (1969)

Pinsker, M.S.: Optimal filtration of square-integrable signals in gaussian noise. Probl. Inf. Transm. **16**, 120–133 (1980)

Prakasa Rao, B.L.S.: Estimation of a unimodal density. Sankhya, Ser. A **31**, 23–36 (1983)

Reiss, R.-D.: Consistency of minimum contrast estimators in non-standard cases. Metrika **25**, 129–142 (1978)

Reiss, R.-D.: Sharp rates of convergence of maximum likelihood estimators in nonparametric models. Z. Wahrscheinlichkeitstheor. Verw. Geb. **65**, 473–482 (1984)

Severini, T.A., Wong, W.H.: Convergence rates of maximum likelihood and related estimates in general parameter spaces. (Preprint 1987)

Shorack, G.R., Wellner, J.A.: Empirical processes with applications to statistics. New-York: Wiley 1986

Stone, C.J.: Optimal rates of convergence for nonparametric regression. Ann. Stat. **10**, 1040–1053 (1982)

Van de Geer, S.: Estimating a regression function. Ann. Stat. **18**, 907–924 (1990a)

Van de Geer, S.: Hellinger-consistency of certain nonparametric maximum likelihood estimates. n° 614, Utrecht University (Preprint 1990b)