Contents

Vol. 17, No. 2, 2008

ISSN 1066-5307, Mathematical Methods of Statistics, 2008, Vol. 17, No. 2, pp. 113–122. © Allerton Press, Inc., 2008.

Adaptation on the Space of Finite Signed Measures

E. Giné[1] and R. Nickl[1*]

[1]*Dept. of Math., University of Connecticut, USA*

Received February 05, 2008

Abstract—Given an i.i.d. sample from a probability measure P on $\mathbb{R}$, an estimator is constructed that efficiently estimates P in the bounded-Lipschitz metric for weak convergence of probability measures, and, at the same time, estimates the density of P – if it exists (but without assuming it does) – at the best possible rate of convergence in total variation loss (that is, in L^1-loss for densities).

Key words: kernel density estimator, exponential inequality, adaptive estimation, total variation loss, bounded Lipschitz metric, L^1-loss.

2000 Mathematics Subject Classification: primary 62G07; secondary 60F05.

DOI: 10.3103/S1066530708020014

1. INTRODUCTION

Viewing the set of all probability measures on $\mathbb{R}$ as a subset of the Banach space $M(\mathbb{R})$ of finite signed Borel measures on $\mathbb{R}$, one has two 'natural' topologies: the 'strong' norm topology given by the norm

$$\|\mu\|_{TV} := |\mu|(\mathbb{R}), \tag{1}$$

where $|\mu|$ is the usual total variation measure of $\mu \in M(\mathbb{R})$; and the usual topology of weak convergence, where

$$\mu_n \to \mu \text{ weakly} \qquad \Longleftrightarrow \qquad \int_{\mathbb{R}} f d(\mu_n - \mu) \to 0 \quad \forall f \in \mathsf{C}(\mathbb{R}).$$

The topology of weak convergence can be metrized on bounded subsets of $M(\mathbb{R})$, so in particular on the set of all probability measures on $\mathbb{R}$, and a commonly used metric is the bounded Lipschitz metric given by

$$\beta(\mu, \nu) = \sup_{f \in \mathcal{F}_{BL}} \left| \int_{\mathbb{R}} f\, d(\mu - \nu) \right| = \|\mu - \nu\|_{\mathcal{F}_{BL}}, \tag{2}$$

for $\mu, \nu \in M(\mathbb{R})$ and where

$$\mathcal{F}_{BL} = \left\{ f : \mathbb{R} \to \mathbb{R} : \|f\|_{BL} := \|f\|_\infty + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} \leq 1 \right\} \tag{3}$$

is the unit ball in the space of bounded Lipschitz functions.

Let $X_1, \ldots, X_n$ be independent real-valued random variables each having law P, and denote by $P_n = n^{-1} \sum_{j=1}^n \delta_{X_j}$ the usual empirical measure. We assume throughout that the X_j's, $j = 1, \ldots, n$, are the coordinate projections of the infinite product probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P^{\mathbb{N}})$, and we set $\Pr := P^{\mathbb{N}}$. Given the sample, the statistical goal is to estimate P, and the Banach space $M(\mathbb{R})$ suggests two natural loss functions to evaluate the performance of an estimator, namely, $\|\cdot\|_{TV}$ and β. For each

*E-mail: nickl@math.uconn.edu

given loss functions, optimal estimators exist: on the one hand, $\|\mu\|_{TV} = \|m\|_1$ for absolutely continuous $\mu \in M(\mathbb{R})$ with Lebesgue-density m, so estimation of (absolutely continuous) P in $\|\cdot\|_{TV}$-loss reduces to density estimation in L^1-loss, which is a well treated subject in nonparametric statistics, cf., e.g., Devroye and Lugosi [5]. Here the usual phenomenon occurs that the best possible rate of convergence for estimating the density p_0 of P depends on the smoothness properties of p_0, and this rate is always slower than $1/\sqrt{n}$ if no finite-dimensional model is assumed. On the other hand, estimation of P in bounded-Lipschitz loss β was considered in Giné and Zinn [10]. There it was shown that the empirical process over the bounded Lipschitz ball $\mathcal{F}_{BL}$ satisfies the uniform central limit theorem if P has a moment of order larger than one, and that a marginally weaker condition is necessary for the CLT to hold. This implies, in particular, that the empirical measure P_n estimates P efficiently w.r.t. the metric β (for this notion of efficiency, see, e.g., van der Vaart and Wellner [18], p. 420) and has convergence rate $\beta(P_n, P) = O_P(n^{-1/2})$. Note however that P_n is not consistent in $\|\cdot\|_{TV}$-loss, since $\|P_n - P\|_{TV} = 2$ for every n and absolutely continuous P. So the question arises whether optimality in both loss functions can be achieved by a single estimator, and we will answer this question in the affirmative in this note.

Adaptive density estimation in the i.i.d. density model on the real line in L^1-loss has been treated in the literature before (see Remark 2 below), but to the best of our knowledge, all these results achieve the minimax rate of convergence only within a logarithmic factor. Our results show that *optimal* rate-adaptive estimators (without a logarithmic penalty) can be constructed in the i.i.d. density model. More generally, Theorem 1 below shows that optimally rate-adaptive estimators possessing the plug-in property of Bickel and Ritov [1] exist. The results of the present article also have applications to semiparametric higher order efficiency problems, similar to those studied in Golubev and Levit [11] and Dalalyan, Golubev and Tsybakov [3].

Some of the methods and ideas of the present article are inspired by recent results in Giné and Nickl [9], who considered the conceptually related problem of optimal estimation of a distribution function and its density in the supnorm.

2. ADAPTATION ON THE SPACE OF FINITE SIGNED MEASURES

2.1. Basic Setup

We start with some basic notation. For an arbitrary (non-empty) set M, $\ell^\infty(M)$ will denote the Banach space of bounded real-valued functions H on M normed by $\|H\|_M := \sup_{m \in M} |H(m)|$, but $\|H\|_\infty$ is used for $\sup_{x \in \mathbb{R}} |H(x)|$. For Borel-measurable functions $h \colon \mathbb{R} \to \mathbb{R}$ and Borel measures μ on $\mathbb{R}$, we set $\mu h := \int_{\mathbb{R}} h\, d\mu$, and we denote by $\mathcal{L}^p(\mathbb{R}, \mu)$ the usual Lebesgue-spaces of Borel-measurable functions from $\mathbb{R}$ to $\mathbb{R}$. If $d\mu(x) = dx$ is Lebesgue measure, we set shorthand $\mathcal{L}^p(\mathbb{R}) := \mathcal{L}^p(\mathbb{R}, \mu)$, and, for $1 \le p < \infty$, we abbreviate the norm by $\|\cdot\|_p$. The convolution $f * g(x)$ of two measurable functions f, g on $\mathbb{R}$ is defined by $\int_{\mathbb{R}} g(x-y) f(y)\, dy$ if the integral converges. Similarly, if μ is any finite signed measure and f is a measurable function, convolution is defined as $\mu * f(x) = \int_{\mathbb{R}} f(x-y)\, d\mu(y)$ if the integral exists. We refer to p. 237 in de la Peña and Giné [4] for the following definitions: the empirical process indexed by $\mathcal{F} \subseteq \mathcal{L}^2(\mathbb{R}, P)$ is given by $f \mapsto \sqrt{n}(P_n - P)f = n^{-1/2} \sum_{j=1}^n (f(X_j) - Pf)$. Convergence in law of random elements in $\ell^\infty(\mathcal{F})$ is defined in the usual way, and will be denoted by the symbol $\rightsquigarrow_{\ell^\infty(\mathcal{F})}$. The class $\mathcal{F}$ is said to be *P-Donsker* if $\sqrt{n}(P_n - P) \rightsquigarrow_{\ell^\infty(\mathcal{F})} G_P$, where G_P is the Brownian bridge indexed by $\mathcal{F}$ (that is, a centered Gaussian process with covariance $EG_P(f)G_P(g) = P[(f - Pf)(g - Pg)]$) and if G_P is sample-bounded and sample-continuous w.r.t. the covariance metric. We also introduce the following function spaces, where we restrict ourselves, for simplicity, to *integer* $t > 0$: we denote by $\mathcal{W}_1^t(\mathbb{R})$ the space of integrable functions f whose derivatives $D^\alpha f$ up to order t exist, and $D^\alpha f \in \mathcal{L}^1(\mathbb{R})$ for all $0 \le \alpha \le t$.

We will consider the usual smoothed empirical process (kernel density estimator): if $X_1, \dots, X_n$ are i.i.d. on the real line, then

$$p_n^K(h, x) = P_n * K_h(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right), \qquad x \in \mathbb{R}, \tag{4}$$

where the kernel $K\colon \mathbb{R}\to\mathbb{R}$ is a symmetric, integrable function that integrates to 1, $K_h(x) := h^{-1}K(x/h)$, and $h := h_n \searrow 0$, $h_n > 0$. The kernel K is of order $r > 0$ if

$$\int_{\mathbb{R}} y^j K(y)\,dy = 0 \quad \text{for } j = 1,\dots,r-1, \quad \text{and} \quad \int_{\mathbb{R}} |y|^r |K(y)|\,dy < \infty.$$

We will denote by $P_n^K(h)$ the random measure defined by $P_n^K(h)(A) = \int_A p_n^K(h,x)\,dx$ for every Borel set $A \subseteq \mathbb{R}$. The dependence of h on n will be assumed without displaying.

2.2. The Main Theorem

For the construction of the estimator, we will have to know a bound on some moment of P, that is, we consider the model

$$\mathcal{P}(\gamma,H) = \left\{P \text{ a Borel probability measure on } \mathbb{R}\colon \int_{\mathbb{R}} (1+|x|)^{2\gamma}\,dP(x) \le H\right\}$$

for some $H < \infty$, $\gamma > 1/2$. See Remark 3 for further discussion. Note that, if P is known to be supported in a bounded interval $[a,b]$, the constant H can be easily calculated as a function of a and b only, and the following results then hold for all probability measures on $[a,b]$. To construct our estimator, we will use the kernel density estimator $p_n^K(h)$ from (4). The crucial problem is to find a good data-driven bandwidth $\hat{h}_n$, that optimally adapts to the unknown smoothness of the density of P. Here we will use a modification of Lepski's method (see Lepski [14]) and refinements given, among others, in Lepski and Spokoiny [15]). Define the grid

$$\mathcal{H} := \left\{h_k = \rho^{-k}\colon k \in \mathbb{N}\cup\{0\},\ \rho^{-k} > n^{-1}(\log n)^2\right\}, \tag{5}$$

where $\rho > 1$ is arbitrary. The number of elements in this grid is of order $\log n$ and we denote by $h_{\min}$ the last (i.e., smallest) element in the grid. We construct $\hat{h}_n$ as follows: first, we check whether

$$\beta\bigl(P_n^K(h_{\min}), P_n\bigr) \le \frac{1}{\sqrt{n}\log n}$$

holds. If this is not satisfied, we set $\hat{h}_n = 0$. Otherwise, we proceed to check whether

$$\|p_n^K(h_{\min}^+) - p_n^K(h_{\min})\|_1 \le \sqrt{\frac{M}{nh_{\min}}} \qquad \text{and} \qquad \beta\bigl(P_n^K(h_{\min}^+), P_n\bigr) \le \frac{1}{\sqrt{n}\log n}$$

simultaneously hold, where $h_{\min}^+$ is the last but one element in the grid $\mathcal{H}$ and where $M = 17L^2$ with

$$L := L(\gamma,H,K) = \left[\frac{2H}{2\gamma-1}\int_{\mathbb{R}} K^2(u)(1+|u|)^{2\gamma}\,du\right]^{1/2}.$$

For example, if P and K are supported in $[0,1]$ we have $H = 4$ (with $\gamma = 1$) and may choose $L = 4\sqrt{2}\|K\|_2$. If the latter does not occur, we set $\hat{h}_n = h_{\min}$, and otherwise, we define $\hat{h}_n$ as

$$\hat{h}_n = \max\left\{h \in \mathcal{H}\colon \|\,p_n^K(h) - p_n^K(g)\|_1 \le \sqrt{M/ng} \quad \forall g < h,\ g \in \mathcal{H} \right.$$
$$\left. \text{and} \quad \beta(P_n^K(h), P_n) \le \frac{1}{\sqrt{n}\log n}\right\}.$$

The estimator is $P_n^K(\hat{h}_n) =: P_n^K(\hat{h}_n,\gamma,H)$ with the convention that $P_n^K(0) := P_n$. The following theorem shows that this estimator is asymptotically optimal both in β and in $\|\cdot\|_{TV}$-loss, see the remark following the theorem for details. In what follows, we say that a sequence of events A_n is *eventual* if $\lim_m \Pr(\cap_{n\ge m} A_n) = 1$.

Theorem 1. *Let $X_1, \ldots, X_n$ be i.i.d. on $\mathbb{R}$ with common law $P \in \mathcal{P}(\gamma, H)$ for some $H < \infty, \gamma > 1/2$. Let $P_n^K(\hat{h}_n)$ be defined as above, where K is a kernel function of order $T+1$, $0 \le T < \infty$ integer, such that $\int_{\mathbb{R}}[(1+|x|)^\gamma K(x)]^2\,dx < \infty$. Then*

$$\sqrt{n}\big(P_n^K(\hat{h}_n) - P\big) \rightsquigarrow_{\ell^\infty(\mathcal{F}_{BL})} G_P, \tag{6}$$

so in particular

$$\beta(P_n^K(\hat{h}_n), P) = O_P\left(\frac{1}{\sqrt{n}}\right).$$

If P possesses a Lebesgue-density p_0, then {the Lebesgue density $p_n^K(\hat{h}_n)$ of $P_n^K(\hat{h}_n)$ exists} is eventual, and

$$\|p_n^K(\hat{h}_n) - p_0\|_1 = o_P(1). \tag{7}$$

If, in addition, $p_0 \in \mathcal{W}_1^t(\mathbb{R})$ for some $0 < t \le T$, we have

$$\|p_n^K(\hat{h}_n) - p_0\|_1 = O_P\big(n^{-\frac{t}{2t+1}}\big). \tag{8}$$

Remark 1. (Modification of Lepski's method.) Our modification of Lepski's [14] method, which follows Theorem 2 in Giné and Nickl [9], basically consists in applying the usual method, but confined to estimators that are contained in a $\|\cdot\|_{\mathcal{F}_{BL}}$-ball of size $o(1/\sqrt{n})$ around the empirical measure P_n.

Remark 2. (Minimax Rates, Related Results.) The minimax rate of convergence in L^1-loss over balls of densities in $\mathcal{W}_1^t(\mathbb{R})$ is $n^{-t/(2t+1)}$ (e.g., Chapter 15 in Devroye and Lugosi [5]), which is achieved by the estimator in the above theorem. Inspection of the proof shows that (8) holds uniformly over sets of the form $\{p \in \mathcal{W}_1^t(\mathbb{R})\colon \sum_{0\le\alpha\le t}\|D^\alpha p\|_1 \le D\}$, and it can be shown that (7) holds uniformly over precompact subsets of $\mathcal{L}^1(\mathbb{R})$. Also, the convergence in law in (6) is uniform over the class of probability measures $\mathcal{P}(\gamma, H)$, since $\mathcal{F}_{BL}$ is a $\mathcal{P}(\gamma, H)$-uniform Donsker class (cf. Corollary 5 and Remark 2 in Nickl and Pötscher [16]). The question whether (8) in Theorem 1 can be obtained for adaptive density estimators on $\mathbb{R}$ has been treated in several places in the literature. For example, Donoho *et al.* [6], Kerkyacharian *et al.* [13] (for compactly supported densities) and Juditsky and Lambert-Lacroix [12] (for densities on $\mathbb{R}$) treated adaptation in general L^p-loss, $1 \le p < \infty$, for compactly supported densities, by wavelet-based estimators, but they had to pay a logarithmic penalty in the rate of convergence.

Remark 3. (Moment Conditions.) Efficient estimation of P in the metric β (that is, in the Banach space $\ell^\infty(\mathcal{F}_{BL})$, for this notion of efficiency cf. van der Vaart and Wellner [18], p. 420) is only possible if a tight Brownian bridge process over $\mathcal{F}_{BL}$ exists, hence, by the Giné and Zinn [10] result discussed in the Introduction, the moment condition on P imposed in Theorem 1 cannot be relaxed.

2.3. Proof of Theorem 1

(I) First note that the class of functions $\mathcal{F}$ is P-Donsker for every probability measure P satisfying $\int_{\mathbb{R}} |x|^{2\gamma}\,dP(x) < \infty$ for some $\gamma > 1/2$, see, e.g., Theorem 2 in Giné and Zinn [10]. Then (6) follows from

$$\|P_n^K(\hat{h}_n) - P_n\|_{\mathcal{F}_{BL}} = o(1/\sqrt{n}).$$

(II) For the case, where P possesses a density p_0, we need the following. Using Minkowski's inequality for integrals we have

$$\big(E\|p_n^K(h) - Ep_n^K(h)\|_1^2\big)^{1/2} \le \int_{\mathbb{R}} \left(E\left|\frac{1}{n}\sum_{j=1}^n K_h(x - X_j) - K_h * p_0(x)\right|^2\right)^{1/2} dx$$

$$\leq \frac{1}{\sqrt{nh}} \int_{\mathbb{R}} \left((K^2)_h * p_0(x)\right)^{1/2} dx.$$

Adapting the proof of Lemma 1 in Giné and Mason [7] to obtain explicit constants, we have

$$\int_{\mathbb{R}} \left((K^2)_h * p_0(x)\right)^{1/2} dx \leq \left(\frac{2H}{2\gamma - 1} \int_{\mathbb{R}} K^2(u)(1+|u|)^{2\gamma}\, du\right)^{1/2} := L,$$

and hence we have that

$$E\|p_n^K(h) - Ep_n^K(h)\|_1^2 \leq L^2 \frac{1}{nh} := L^2\sigma^2(h,n). \tag{9}$$

For the bias, assuming $p_0 \in \mathcal{W}_1^t(\mathbb{R})$, we have for some constant $0 < L' < \infty$ and some $0 < \zeta < 1$

$$\begin{aligned}\|Ep_n^K(h) - p_0\|_1 &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} K(u)[p_0(x-uh) - p_0(x)]\, du \right| dx \\ &\leq \frac{h^t}{t!} \int_{\mathbb{R}} |K(u)||u|^t \int_{\mathbb{R}} |D^t p_0(x - uh\zeta)|\, dx\, du = L'h^t := B(h, p_0) \end{aligned}\tag{10}$$

since $D^t p_0 \in \mathcal{L}^1(\mathbb{R})$. If it is only known that p_0 exists we still have

$$\|Ep_n^K(h) - p_0\|_1 = \|K_h * p_0 - p_0\|_1 = o(1),$$

cf., e.g., Theorem 9.1 in Devroye and Lugosi [5].

Proof of (7) *and* (8). By Lemma 1 below with $\lambda = 1/\log n$ and $h = h_{\min}$ we obtain that $\{\hat{h}_n \geq h_{\min}\}$ is eventual, and hence the density $p_n^K(\hat{h}_n)$ exists eventually. Expectations in the rest of the proof are taken over the event $\{p_n^K(\hat{h}_n)$ exists$\}$.

Define h_p by the balance equation

$$h_p = \max\left\{h \in \mathcal{H}\colon B(h, p_0) \leq \frac{\sqrt{M}}{4}\sigma(h,n)\right\}.$$

It is easily verified that $h_p \simeq n^{-1/(2t+1)}$ if $p_0 \in \mathcal{W}_1^t(\mathbb{R})$ for some $0 < t \leq T$, cf. (II). If p_0 exists but is not contained in $\mathcal{W}_1^t(\mathbb{R})$ for some $t > 0$, we set $h_p = h_{\min}$. Then we define $\tilde{\sigma}(h_p, n)$ as $\sigma(h_p, n)$ if $t > 0$ and set $\tilde{\sigma}(h_p, n) = \max\left(\sigma(h_p, n), (4/\sqrt{M})B(h_p, p_0)\right)$ otherwise, so that

$$B(h_p, p_0) \leq (\sqrt{M}/4)\tilde{\sigma}(h_p, n)$$

always holds. Clearly $\sigma(h_p, n) = O(\tilde{\sigma}(h_p, n))$ and we note that for $t > 0$

$$\tilde{\sigma}(h_p, n) = \sigma(h_p, n) \simeq n^{-t/2t+1)} \tag{11}$$

is the rate of convergence required in (8), but $\tilde{\sigma}(h_p, n) \to 0$ as soon as P has a density.

We will consider the cases $\{\hat{h}_n \geq h_p\}$ and $\{\hat{h}_n < h_p\}$ separately. First, by definition of $\hat{h}_n$, h_p and (9) we have

$$\begin{aligned}E\|p_n^K(\hat{h}_n) - p_0\|_1 I_{\{\hat{h}_n \geq h_p\}} &\leq E\Big(\|p_n^K(\hat{h}_n) - p_n^K(h_p)\|_1 + \|p_n^K(h_p) - Ep_n^K(h_p)\|_1 + B(h_p, p_0)\Big) I_{\{\hat{h}_n \geq h_p\}} \\ &\leq \sqrt{M}\sigma(h_p, n) + L\sigma(h_p, n) + \frac{\sqrt{M}}{4}\tilde{\sigma}(h_p, n) = O(\tilde{\sigma}(h_p, n)).\end{aligned}$$

In the case $\{\hat{h}_n < h_p\}$ we have the following: if $h_p = h_{\min}$, then $\{\hat{h}_n < h_p\}$ cannot occur, so (7) is proved if $t = 0$ and will follow from (8) in case $t > 0$, which we assume for the rest of the proof. [Note that then $\tilde{\sigma}(h_p, n) = \sigma(h_p, n)$.] Since

$$E\|p_n^K(\hat{h}_n) - p_0\|_1 I_{\{\hat{h}_n < h_p\}} \leq \sum_{h \in \mathcal{H}\colon h < h_p} E\Big[\big(\|p_n^K(h) - Ep_n^K(h)\|_1 + \|Ep_n^K(h) - p_0\|_1\big) I_{\{\hat{h}_n = h\}}\Big]$$

$$\leq \sum_{h\in\mathcal{H}:\, h<h_p} \big(E\|p_n^K(h) - Ep_n^K(h)\|_1^2\big)^{1/2}\big(EI_{\{\hat{h}_n=h\}}\big)^{1/2} + \frac{\sqrt{M}}{4}\sigma(h_p,n),$$

by (9), it remains to show that

$$\sum_{h\in\mathcal{H}:\, h<h_p} \sigma(h,n)\cdot\sqrt{\Pr(\hat{h}_n = h)} = O(\sigma(h_p,n)) \tag{12}$$

is satisfied. Pick any $h \in \mathcal{H}$ so that $h < h_p$, denote by h^+ the previous element in the grid (i.e., $h^+ = \rho h$) and observe that

$$\sqrt{\Pr(\hat{h}_n = h)} \leq \Bigg(\sum_{g\in\mathcal{H}:\, g\leq h} \Pr\big(\|\, p_n^K(h^+) - p_n^K(g)\|_1 > \sqrt{M}\sigma(g,n)\big)\Bigg)^{1/2}$$
$$+ \left(\Pr\left(\sqrt{n}\|P_n^K(h^+) - P_n\|_{\mathcal{F}_{BL}} > \frac{1}{\log n}\right)\right)^{1/2} =: A + B. \tag{13}$$

First, by definition of the grid and (9) we have

$$\sum_{h\in\mathcal{H}:\, h<h_p} \sigma(h,n)\cdot B \leq d(\log n)\sigma(h_{\min},n)\sqrt{\exp\left\{-L\min\left(\frac{1}{(h_p\log n)^2}, \frac{\sqrt{n}}{h_p\log n}\right)\right\}}$$
$$= o\big(\sigma(h_p,n)\big) \tag{14}$$

for n large, where we have applied Lemma 1 below with $\lambda = 1/\log n$ and $h = h^+ \leq h_p$.

For the term including A we first observe that

$$\|p_n^K(h^+) - p_n^K(g)\|_1 \leq \|\, p_n^K(h^+) - Ep_n^K(h^+)\|_1 + \|p_n^K(g) - Ep_n^K(g)\|_1 + B(h^+,p_0) + B(g,p_0),$$

where $B(h^+,p_0) + B(g,p_0) \leq (\sqrt{M}/2)\sigma(g,n)$, since $g < h^+ \leq h_p$. Consequently,

$$\Pr\Big(\|p_n^K(h^+) - p_n^K(g)\|_1 > \sqrt{M}\sigma(g,n)\Big) \leq \Pr\Big(\|p_n^K(h^+) - Ep_n^K(h^+)\|_1 > (1/4)\sqrt{M}\sigma(h^+,n)\Big)$$
$$+ \Pr\Big(\|p_n^K(g) - Ep_n^K(g)\|_1 > (1/4)\sqrt{M}\sigma(g,n)\Big).$$

Now Lemma 2 below gives

$$\sum_{g\in\mathcal{H}:\, g\leq h} \Pr\Big(\|p_n^K(h^+) - p_n^K(g)\|_1 > \frac{1}{4}\sqrt{M}\sigma(g,n)\Big) \leq L''\log n\exp\Big\{-\frac{1}{L'h}\Big\}$$

and then

$$\sum_{h\in\mathcal{H}:\, h<h_p} \sigma(h,n)\cdot A = O\Bigg((\log n)^{3/2}\sigma(h_{\min},n)\sqrt{\exp\Big\{-\frac{1}{L'h_p}\Big\}}\Bigg) = o\big(\sigma(h_p,n)\big).$$

Now this, (13), and (14) verify (12), which completes the proof, given Lemmas 1 and 2.

The following two exponential inequalities were used in the proof.

Lemma 1. *Suppose that P satisfies $H := H(\gamma) = \int_{\mathbb{R}} |x|^{2\gamma}\, dP(x) < \infty$ for some $\gamma > 1/2$. Set $t = 0$ in what follows, or assume that P has a density p_0 with respect to Lebesgue measure such that $p_0 \in \mathcal{W}_1^t(\mathbb{R})$ for some $t > 0$. Let $h := h_n \to 0$ as $n \to \infty$ satisfy $h \geq (\log n/n)$, and let K be a kernel of order $t+1$. Define $\gamma' = \gamma$ if $\gamma \neq 1$, and $\gamma' = 1 - \delta$ for some arbitrary $0 < \delta < 1/2$ otherwise, and then define $\kappa = \min(1,\gamma')$. Then there exist finite positive constants $L := L(K)$ and $\Lambda_0 := \Lambda_0(K, H, \int_{\mathbb{R}} |D^t p_0(y)|\, dy)$ such that for all $\lambda \geq \Lambda_0 \max\big(\min(h^{1-1/2\kappa}, \sqrt{n}h), \sqrt{n}h^{t+1}\big)$ and $n \in \mathbb{N}$,*

$$\Pr\big(\sqrt{n}\|P_n^K(h) - P_n\|_{\mathcal{F}_{BL}} > \lambda\big) \leq 2\exp\left\{-L\min\left(\frac{\lambda^2}{h^2}, \frac{\sqrt{n}\lambda}{h}\right)\right\}.$$

Proof. We start with a remark on measurability, which will also be needed in the application of Talagrand's inequality below: since f and $K_h * f$ are continuous functions, $(P_n^K(h) - P_n)f$ is a random variable for each $f \in \mathcal{F}_{BL}$. Furthermore, there is a countable $\mathcal{F}_0 \subseteq \mathcal{F}_{BL}$ such that

$$\sup_{f \in \mathcal{F}_0} |(P_n - P)(K_h * f - f)| = \sup_{f \in \mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| \tag{15}$$

except perhaps on a set of zero probability. To see this, let $\mathcal{F}_{BL}(l)$ be the unit ball of the space of bounded Lipschitz functions on $[-l, l]$, which is relatively compact for the sup-norm (by Ascoli's theorem), and let $\mathcal{F}_l$ be a countable (sup-norm) dense subset of $\mathcal{F}_{BL}(l)$. Extend each $f \in \mathcal{F}_l$ as $f(x) = f(-l)$ for $x < -l$ and $f(x) = f(l)$ for $x > l$, and still denote, with some abuse of notation, this set of extensions as $\mathcal{F}_l$. Then $\mathcal{F}_0 := \cup_{l=1}^{\infty} \mathcal{F}_l$ is a countable subset of $\mathcal{F}_{BL}$, and, using tightness of K and P, it is easy to see that (15) holds for all ω such that $|X_j(\omega)| < \infty$, $j \in \mathbb{N}$, and for all n.

The proof of the lemma follows Theorem 1 in Giné and Nickl [9], but requires substantial technical modifications. We use the decomposition

$$P_n * K_h - P_n = P_n * K_h - P * K_h - P_n + P + P * K_h - P,$$

so that

$$\|P_n^K(h) - P_n\|_{\mathcal{F}_{BL}} \leq \sup_{f \in \mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| + \|P * K_h - P\|_{\mathcal{F}_{BL}}. \tag{16}$$

For the "bias term" we have, as in Lemma 4 in Giné and Nickl [8], for given $f \in \mathcal{F}_{BL}$ with $\bar{f}(x) = f(-x)$, that

$$(P * K_h - P)f = \int_{\mathbb{R}} K(t)[P * \bar{f}(ht) - P * \bar{f}(0)]\, dt. \tag{17}$$

For every $0 < \alpha \leq t$, we have $D^\alpha(p_0 * \bar{f}) = D^\alpha p_0 * \bar{f}$, see, e.g., Lemma 5b in Giné and Nickl [8], and, with the convention that $D^0 p_0 = P$, we obtain

$$\|D^\alpha p_0 * \bar{f}\|_\infty \leq \|D^\alpha p_0\|_{TV} \|f\|_\infty < \infty,$$

where $\|D^\alpha p_0\|_{TV}$ denotes the total variation norm (see (1) above) of the measure $D^\alpha p_0(y)dy$, which is equal to the L^1-norm of $D^\alpha p_0$ for $\alpha > 0$. Summarizing, the function $P * \bar{f}$ possesses bounded derivatives up to order t. Furthermore, since $D^t p_0(y)dy$ gives rise to a finite signed measure, and since $f \in \mathcal{F}_{BL}$, we obtain (interpreting $D^0 p_0(y)\, dy$ as $dP(y)$)

$$\begin{aligned} |r|^{-1}|D^t p_0 * \bar{f}(x + r) - D^t p_0 * \bar{f}(x)| &= |r|^{-1}\left| \int_{\mathbb{R}} [f(r + y - x) - f(y - x)] D^t p_0(y)\, dy \right| \\ &\leq \int_{\mathbb{R}} |D^t p_0(y)| dy < \infty \end{aligned}$$

and hence $P * \bar{f}$ has bounded derivatives up to order t and the tth derivative (in case $t = 0$ the function $P * \bar{f}$ itself) is a bounded Lipschitz function. Now this, (17) and the fact that the kernel is of order $t + 1$ give, by straightforward Taylor expansions,

$$\|P * K_h - P\|_{\mathcal{F}_{BL}} \leq Ch^{t+1}$$

for some constant C depending only on $\int_{\mathbb{R}} |D^t p_0(y)|\,dy$ and K. This and (16) imply, by assumption on λ, that

$$\Pr\left(\sqrt{n}\|P_n^K(h) - P_n\|_{\mathcal{F}_{BL}} > \lambda\right) \le \Pr\left(\sqrt{n} \sup_{f\in\mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| > \lambda - C\sqrt{n}h^{t+1}\right)$$

$$\le \Pr\left(n \sup_{f\in\mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| > \frac{\sqrt{n}\lambda}{2}\right). \tag{18}$$

We will apply Talagrand's inequality to the class

$$\tilde{\mathcal{F}}_{BL} = \left\{K_h * f - f - P(K_h * f - f) \colon f \in \mathcal{F}_{BL}\right\}$$

to bound the last probability, but first we need some preliminary facts:

(a) We have

$$\sup_{f\in\mathcal{F}_{BL}} \|K_h * f - f\|_{2,P} \le \sup_{f\in\mathcal{F}_{BL}} \|K_h * f - f\|_\infty \le h \int_{\mathbb{R}} |K(u)||u|du := \sigma, \tag{19}$$

since

$$|K_h * f(x) - f(x)| = \left|\int_{\mathbb{R}} K(u)[f(x - uh) - f(x)]\,du\right| \le h \int_{\mathbb{R}} |K(u)||u|\,du.$$

(b) Clearly, (19) implies that the envelope U of $\tilde{\mathcal{F}}_{BL}$ can be taken to be of order $C'h$ for $C' = 2\int_{\mathbb{R}} |K(u)||u|\,du$.

(c) We will establish the expectation bound

$$nE \sup_{f\in\mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| \le C'' \min(\sqrt{n}h^{1-1/2\kappa}, nh) \tag{20}$$

for C'' some finite positive constant depending only on H. That this expression is dominated by $C''nh$ follows immediately from (b). Note that the set $\cup_{h>0}\{K_h * f - f \colon f \in \mathcal{F}_{BL}\}$ is contained in the class of functions $3\|K\|_1 \cdot \mathcal{F}_{BL}$ in view of $\|K_h * f - f\|_{BL} \le \|K_h * f\|_{BL} + 1$, $\|K_h * f\|_\infty \le \|K\|_1$, and

$$|r|^{-1}|K_h * f(x + r) - K_h * f(x)| = |r|^{-1}\left|\int_{\mathbb{R}} K_h(y)[f(x + r - y) - f(x - y)]\,dy\right|$$

$$\le \int_{\mathbb{R}} |K_h(y)|dy = \|K\|_1.$$

Then, the bracketing metric entropy $\log N_{[]}\left(\varepsilon, 3\|K\|_1 \cdot \mathcal{F}_{BL}, \|\cdot\|_{2,P}\right)$ can be shown to be dominated by a constant depending only on H times $\varepsilon^{-1/\kappa}$, see Theorem 1.2 (with $\beta = 0$, $s = d = 1$, $p = q = \infty$, $\mu = P$) and Remark 2 in Nickl and Pötscher [16]. Now, the bracketing-expectation bound for empirical processes contained in the third inequality in Theorem 2.14.2 in van der Vaart and Wellner [18] yields (20) in view of (b).

We now apply Talagrand's inequality, see (21) below, with $x = L\min\left(\frac{\lambda^2}{h^2}, \frac{\sqrt{n}\lambda}{h}\right)$ for suitable L and with σ and U as in (a) and (b), to the expression (18). We need to check the following three bounds.

(I) First we have, for n large enough

$$nE \sup_{f\in\mathcal{F}} |(P_n - P)(K_h * f - f)| \le C'' \min(\sqrt{n}h^{1-1/2\kappa}, nh) \le \frac{\sqrt{n}\lambda}{6}$$

by (20) and the assumption on λ.

(II) Note that $V \le n\sigma^2 + C'C''h\min(\sqrt{n}h^{1-1/2\kappa}, nh) \le C'''nh^2$ and then, for L small enough,

$$\sqrt{2Vx} \le 2\sqrt{LC'''}\sqrt{nh^2\frac{\lambda^2}{h^2}} \le \frac{\sqrt{n}\lambda}{6}.$$

(III) Furthermore

$$\frac{Ux}{3} \le LC'h\frac{\sqrt{n}\lambda}{3h} \le \frac{\sqrt{n}\lambda}{6}.$$

Summarizing, the sum of the terms in (I)–(III) is smaller than $\sqrt{n}\lambda/2$ if L is chosen suitably small, and we obtain from (21) for the given choice of x that

$$\Pr\left\{n \sup_{f\in\mathcal{F}_{BL}} |(P_n - P)(K_h * f - f)| > \frac{\sqrt{n}\lambda}{2}\right) \le 2\exp\{-x\},$$

which completes the proof of the lemma. □

Lemma 2. *We have*

$$\Pr\left(\|p_n^K(g) - Ep_n^K(g)\|_1 > (1/4)\sqrt{M}\sigma(g,n)\right) \le 2\exp\left(-\frac{1}{L'g}\right)$$

for every $g \le h_p$, $g \in \mathcal{H}$, $n \in \mathbb{N}$, *and some constant* $0 < L' < \infty$.

Proof. We will apply Talagrand's inequality to $K_g(\cdot - X) - K_g * p_0$ which is a $\mathcal{L}^1(\mathbb{R})$-valued random variable (since the mapping $x \mapsto f(\cdot - x)$ is continuous from $\mathbb{R}$ to $\mathcal{L}^1(\mathbb{R})$ for integrable f). First we note that, since the unit ball B of $\mathcal{L}^\infty(\mathbb{R})$ is compact and metrizable, hence separable, for the weak* topology induced by $\mathcal{L}^1(\mathbb{R})$, there is a countable subset B_0 of B such that $\|H\|_1 = \sup_{f\in B_0}\left|\int_{\mathbb{R}} H(t)f(t)\,dt\right|$ for all $H \in \mathcal{L}^1(\mathbb{R})$. Since $K_g, P * K_g$ are in $\mathcal{L}^1(\mathbb{R})$, we have

$$\|p_n^K(g) - Ep_n^K(g)\|_1 = \|P_n - P\|_{\mathcal{K}}$$

for

$$\mathcal{K} = \left\{x \mapsto \int_{\mathbb{R}} f(t)K_g(t-x)\,dt - \int_{\mathbb{R}} f(t)K_g * p_0(t)\,dt\colon f \in B_0\right\},$$

so that we can apply Talagrand's inequality with the countable class $\mathcal{K}$.

To do this, observe that $\mathcal{K}$ is uniformly bounded by $2\|K\|_1 := U$, since

$$\sup_{f,x}\left|\int_{\mathbb{R}} f(t)K_g(t-x)\,dt\right| \le \|K\|_1.$$

Similarly, we have

$$\sup_f E\Big(\int_{\mathbb{R}} f(t)K_g(t-x)\,dt\Big)^2 \le \|K\|_1^2 := \sigma^2.$$

Also we have as in (9)

$$E\|n(P_n - P)\|_{\mathcal{K}} = E\left\|\sum_{j=1}^n \big(K_h(\cdot - X_j) - EK_h(\cdot - X)\big)\right\|_1 \le L\sqrt{\frac{n}{h}},$$

where L is specified before (9).

Now Talagrand's inequality, see (21), gives with $x = 1/(L'g)$ that

$$\Pr\left(n\|p_n^K(g) - Ep_n^K(g)\|_1 > L\sqrt{\frac{n}{g}} + \sqrt{\left(2n\|K\|_1^2 + 4\|K\|_1 L\sqrt{\frac{n}{g}}\right)\frac{1}{L'g}} + \frac{2\|K\|_1}{3L'g}\right) \le 2e^{-\frac{1}{L'g}}.$$

But this inequality implies the lemma, since

$$\sqrt{\frac{n}{g}}\left[L + \frac{\sqrt{2}\|K\|_1}{\sqrt{L'}} + \frac{2\sqrt{L\|K\|_1}}{\sqrt{L'}(ng)^{1/4}} + \frac{2\|K\|_1}{3L'\sqrt{ng}}\right] \le \frac{\sqrt{M}}{4}\sqrt{\frac{n}{g}}$$

by suitable choice of L' and recalling $M = 17L^2$. □

2.4. Appendix: Talagrand's Inequality

Let $X_1, \ldots, X_n$ be i.i.d. with law P on $\mathbb{R}$, and let $\mathcal{F}$ be a P-centered (i.e., $\int f dP = 0$ for all $f \in \mathcal{F}$) countable class of real-valued functions on $\mathbb{R}$, uniformly bounded by the constant U. Let σ be any positive number such that $\sigma^2 \ge \sup_{f\in\mathcal{F}} E(f^2(X))$, and set $V := n\sigma^2 + 2UE\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}}$. Then, Bousquet's [2] version of Talagrand's inequality (Talagrand [17]), with constants, is as follows (see Theorem 7.3 in Bousquet [2]): for every $x \ge 0$

$$\Pr\left\{\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}} \ge E\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}} + \sqrt{2Vx} + \frac{Ux}{3}\right\} \le 2e^{-x}. \tag{21}$$

REFERENCES

1. J. P. Bickel and Y. Ritov, "Nonparametric Estimators which Can Be 'Plugged-In'," Ann. Statist. **31**, 1033–1053 (2003).
2. O. Bousquet, "Concentration Inequalities for Sub-Additive Functions Using the Entropy Method", in: *Progress in Probability*, Vol. 56: *Stochastic Inequalities And Applications*, Ed. by E. Giné, C. Houdré, and D. Nualart (Birkhäuser, Boston, 2003), pp. 213–247.
3. A. S. Dalalyan, G. K. Golubev, and A. B. Tsybakov, "Penalized Maximum Likelihood and Semiparametric Second-Order Efficiency", Ann. Statist. **34**, 169–201 (2006).
4. V. de la Peña and E. Giné, *Decoupling. From Dependence to Independence* (Springer, New York, 1999).
5. L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation* (Springer, New York, 2001).
6. D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density Estimation by Wavelet Thresholding", Ann. Statist. **24**, 508–539 (1996).
7. E. Giné and D. M. Mason, "On Local U-Statistic Processes and the Estimation of Densities of Functions of Several Sample Variables", Ann. Statist. **35**, 1105–1145 (2007).
8. E. Giné and R. Nickl, "Uniform Central Limit Theorems for Kernel Density Estimators", Probab. Theory Related Fields, 2008 (in press).
9. E. Giné and R. Nickl, "An Exponential Inequality for the Distribution Function of the Kernel Density Estimator, with Applications to Adaptive Estimation", Probab. Theory Related Fields, 2008 (in press).
10. E. Giné and J. Zinn, "Empirical Processes Indexed by Lipschitz Functions", Ann. Probab. **14**, 1329–1338 (1986).
11. G. K. Golubev and B. Y. Levit, "Distribution Function Estimation: Adaptive Smoothing", Math. Methods Statist. **5**, 383–403 (1996).
12. A. Juditsky and S. Lambert-Lacroix, "On Minimax Density Estimation on $\mathbb{R}$", Bernoulli **10**, 187–220 (2004).
13. G. Kerkyacharian, D. Picard, and K. Tribouley, "L^p Adaptive Density Estimation", Bernoulli **2**, 229–247 (1996).
14. O. V. Lepski, "Asymptotically Minimax Adaptive Estimation. I. Upper Bounds. Optimally Adaptive Estimates", Theory Probab. Appl. **36**, 682–697 (1991).
15. O. V. Lepski and V. G. Spokoiny, "Optimal Pointwise Adaptive Methods in Nonparametric Estimation", Ann. Statist. **25**, 2512–2546 (1997).
16. R. Nickl and B. M. Pötscher, "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type", J. Theoret. Probab. **20**, 177–199 (2007).
17. M. Talagrand, "New Concentration Inequalities in Product Spaces", Invent. Math. **126**, 505–563 (1996).
18. A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes* (Springer, New York, 1996).