Probab. Th. Rel. Fields 77, 521–536 (1988)

Probability Theory and Related Fields

© Springer-Verlag 1988

The Kernel Estimate is Relatively Stable

Luc Devroye*

School of Computer Science, McGill University, 805 Sherbrooke Street West, Montreal, Canada H3A 2K6

Summary. Consider the Parzen-Rosenblatt kernel estimate $f_n = (1/n) \sum_{i=1}^{n} K_h(x - X_i)$, where $h > 0$ is a constant, K is an absolutely integrable function with integral one, $K_h(x) = (1/h^d) K(x/h)$, and $X_1, \ldots, X_n$ are iid random variables with common density f on R^d. We show that for all $\varepsilon > 0$,

$$\sup_{h>0, f} P(|\int |f_n - f| - E\int |f_n - f|| > \varepsilon) \leqq 2e^{-\frac{n\varepsilon^2}{32 \int |K|^2}}.$$

We also establish that f_n is relatively stable, i.e.

$$\frac{\int |f_n - f|}{E\int |f_n - f|} \to 1 \quad \text{in probability as } n \to \infty,$$

whenever $\liminf \sqrt{n} E \int |f_n - f| = \infty$. We also study what happens when h is allowed to depend upon the data sequence.

1. Introduction

We consider the standard problem of estimating a density f on R^d from an iid sample $X_1, \ldots, X_n$ drawn from f. A *density estimate* f_n, a Borel measurable function of $x, X_1, \ldots, X_n$, is *relatively stable* when

$$\frac{J_n}{E(J_n)} \to 1 \quad \text{in probability as } n \to \infty,$$

where $J_n = \int |f_n - f|$ is the L_1 error. It is *strongly relatively stable* when the convergence is in the almost sure sense.

* This work was sponsored by NSERC Grant A3456 and FCAC Grant EQ-1678

The notion of relative stability is important in comparative studies of density estimates. Comparing relatively stable density estimates on the basis of $E(J_n)$ is fair since the actual error J_n is with high probability close to its mean. The situation is more complicated for example when $J_n/E(J_n)$ tends to a nondegenerate limit law; the conservative elements among us could be tempted to choose a density estimate with a larger asymptotic mean but a smaller asymptotic variance. Dilemmas of this sort do not occur for relatively stable density estimates. Of course, one should keep in mind that the actual value of $E(J_n)$ is more important than anything else- just consider that any data-independent estimate is relatively stable.

Another important point is that simulations of the performance (L_1 error) of relatively stable density estimates are very cheap since J_n (the computed performance) is with high probability close to $E(J_n)$. In other words, it is not necessary to average over several simulation runs. As we will see below, J_n is already an average of sorts because of the global integral in its definition.

The literature on minimax lower and upper bounds for the L_1 error deals almost exclusively with $E(J_n)$, and not with other quantities such as the p-th quantiles of J_n. In view of the relative stability of most nonparametric estimates, it is less important to develop minimax theories based upon quantities other than $E(J_n)$, except in special circumstances. One such situation is when the classes of densities considered in the minimax theory are very small ("parametric"), so that specially designed estimates ("parametric estimates") are better suited.

Most parametric density estimates are not relatively stable. Take for example the class of densities $f=pg+(1-p)h$ where g, h are known disjoint densities ($\int gh=0$), and p is the unknown mixture parameter. If p is estimated from the data by the obvious frequency estimate p_n, and $f_n=p_n g+(1-p_n)h$, then $J_n=2\,|p-p_n|$, and thus, by the central limit theorem,

$$\frac{J_n}{2\sqrt{\frac{1}{n}p(1-p)}} \to |N| \quad \text{in distribution}$$

where N is a normal random variable. It is clear that $J_n/E(J_n)\to|N|/E(|N|)$ in distribution as $n\to\infty$. Therefore, the estimate is not relatively stable for any density in the given class.

In contrast, popular nonparametric density estimates such as the kernel and histogram estimates are relatively stable for all densities. This is due to the local nature of these estimates: densities are estimated locally by considering a limited number of close data points. Locally, the error's standard deviation can be of the same order of magnitude as the error's mean. Yet, because the L_1 criterion sums a lot of many "nearly independent" local errors, the variation in the local errors averages out, rendering the estimates relatively stable. Thus, if we had picked a local criterion such as $|f_n-f|/E(|f_n-f|)$ to define relative stability, then relative stability would effectively force the bias term to dominate the variational term, i.e., $E(|f_n-f|)\sim|E(f_n)-f|$. For nonparametric density estimates, one can usually achieve this by taking the smoothing parameter large enough. Yet, this is a suboptimal strategy because the smallest asymptotic errors are obtained by balancing the bias and variation terms. Thus, "local relative stability" and "locally optimal rate of convergence" are conflicting notions.

Relative stability is a concept first studied by Abou-Jaoude (1977). He encountered some problems in the decomposition of J_n into a bias component and a variational component, due to the nature of the L_1 criterion. Interestingly, even though $E(f_n)$ always is relatively stable, and $\int |f_n - E(f_n)|/E(\int |f_n - E(f_n)|)$ can be shown to converge to 1 in probability for all f for the kernel and histogram estimates (Abou-Jaoude, 1977; Devroye and Gyorfi, 1985, pp. 23–33), these facts do not in general imply relative stability. The two components should thus not be separated.

We will show that for the kernel estimate, consistency implies relative stability. It suffices to note that everything that follows remains valid for the histogram estimate as well. In this paper, the *kernel estimate* is defined as follows:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

where $h > 0$ is a *smoothing factor*, K is an absolutely integrable function called the *kernel*, and $K_h(x) = (1/h^d) K(x/h)$ (Parzen, 1962; Rosenblatt, 1956). We consider smoothing factors h that are functions of n only. Relative stability in the L_2 sense (replace J_n by $\int (f_n - f)^2$ in the definition of relative stability) has been established by Hall (1982) under some regularity conditions on h, f and K, when $d = 1$. Later, Hall (1984) refined this result and obtained the asymptotic law of $\int (f_n - f)^2$ when f has two uniformly continuous derivatives on R^d. Unfortunately, for a variety of reasons, L_1 relative stability cannot be obtained from Hall's results. For example, the relative sizes of the L_2 error and the L_1 error are not related (see Devroye and Gyorfi, 1985), Even if we could show that one follows from the other, we would still need to impose regularity conditions such as $\int f^2 < \infty$.

2. Main Results

We consider only L_1 kernels, i.e. kernels K with $\int |K| < \infty$. Recall that $\int K = 1$, but that K does not have to be nonnegative. The main result from which most other results will be derived is given in Theorem 1:

Theorem 1. *Consider a kernel estimate with L_1 kernel K. For all $\varepsilon > 0$,*

$$\sup_{h > 0, f} P(|J_n - E(J_n)| > \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{32 \int^2 |K|}}.$$

We emphasize that Theorem 1 is valid for *all densities* on R^d. Also, the inequality is uniform over all h and all densities f. It is true that we pay a price for the uniformity; for particular cases, better inequalities are obtainable. Yet, it is the uniformity that will allow us to establish the relative stability of automatic kernel estimates. When considered as a function of K, the bound is smallest for nonnegative K.

It is worth mentioning that the inequality remains valid if J_n is replaced by $\int|f_n-g|$, where g is an arbitrary integrable function (e.g., $g\equiv 0$ is allowed), and the kernel K, while absolutely integrable, violates the condition $\int K=1$.

We will state the main corollaries of Theorem 1 as theorems. Taking $\varepsilon=u/\sqrt{n}$ for some constant u, we obtain

Theorem 2. *Consider a kernel estimate with L_1 kernel K. For $u>0$,*

$$\sup_{h>0,f} P(\sqrt{n}\,|J_n-E(J_n)|>u)\leqq 2e^{-\frac{u^2}{32\int^2|K|}}.$$

In particular, the kernel estimate is relatively stable whenever $\liminf\limits_{n\to\infty}\sqrt{n}\,E(J_n)=\infty$.

If we take $\varepsilon=c\sqrt{\log(n)/n}$ for some constant $c>\sqrt{32}\int|K|$, we can conclude

Theorem 3. *Consider the kernel estimate with L_1 kernel K. For all f and all sequences of smoothing factors $h=h_n$,*

$$\limsup_{n\to\infty}\sqrt{\frac{n}{\log n}}\,|J_n-E(J_n)|\leqq\sqrt{32}\int|K| \quad \textit{almost surely.}$$

In particular, the kernel estimate is strongly relatively stable whenever $\liminf\limits_{n\to\infty}\sqrt{\frac{n}{\log n}}\,E(J_n)=\infty$.

Theorem 3 is probably suboptimal because strong convergence is derived via the Borel-Cantelli lemma.

It is interesting to apply Theorems 1–3 to several special cases. Consider first the univariate case ($d=1$) studied from other points of view by Devroye and Gyorfi (1985). For symmetric nonnegative kernels, and all f, we have

$$\liminf_{n\to\infty}\inf_h n^{\frac{2}{5}}E(J_n)\geqq\gamma>0$$

for some universal constant γ which is at least equal to 0.8 (Devroye and Gyorfi, 1985, p. 79). Combining this with Theorems 1–3 yields

Theorem 4. *Assume that K is a nonnegative symmetric kernel, and that f is an arbitrary density on R^1. Then the kernel estimate is strongly relatively stable, regardless of how $h=h_n$ is chosen as a function of n. For fixed $u>0$,*

$$\limsup_{n\to\infty}P\left(\frac{|J_n-E(J_n)|}{E(J_n)}>\frac{u}{\gamma n^{1/10}}\right)\leqq 2e^{-\frac{u^2}{32\int^2|K|}}.$$

For all f and all sequences of smoothing factors $h=h_n$,

$$\limsup_{n\to\infty}\frac{n^{1/10}}{\sqrt{\log n}}\,\frac{|J_n-E(J_n)|}{E(J_n)}\leqq\frac{\sqrt{32}\int|K|}{\gamma} \quad \textit{almost surely.}$$

Also,

$$\frac{|J_n-E(J_n)|}{E(J_n)}=O(n^{-\frac{1}{10}}) \quad \textit{in probability as } n\to\infty.$$

The simple formula $E(|X|)=\int_0^\infty P(|X|>t)\,dt$ can be used to show that

$$E(|J_n-E(J_n)|)\leqq\frac{C}{\sqrt{n}}$$

where we can take

$$C=\int_0^\infty 2e^{-\frac{t^2}{32\int^{2}|K|}}\,dt=\sqrt{32\pi}\int|K|.$$

Thus, $|J_n-E(J_n)|$ decreases to zero roughly as $1/\sqrt{n}$, which is usually much faster than the rate with which $E(J_n)$ tends to zero, i.e., $n^{-2/5}$ or slower for any nonnegative kernel K and $d=1$.

One could wonder what happens for general (i.e., not necessarily symmetric or nonnegative) L_1 kernels K. Theorem 5.16 of Devroye and Gyorfi (1985, p. 136) states that when $h\to 0$, $\sqrt{n}E(J_n)\to\infty$ as $n\to\infty$, for all f. It is known that $h\to 0$ is necessary for the convergence to 0 of $E(J_n)$ of even a single f when K is nonnegative, but that this is not necessarily so when K is allowed to take negative values. Combining this with Theorem 2 shows

Theorem 5. *For any L_1 kernel K with $\int K=1$, any density f on R^1, and any sequence of smoothing factors $h\to 0$, the kernel estimate is relatively stable. In particular, for any nonnegative kernel and any density f on R^1, consistency (i.e., $E(J_n)\to 0$) implies relative stability.*

Our theorems would not allow one to obtain relative stability for *all* L_1 kernels. It is known (Devroye and Gyorfi 1985, p. 144) that for some classes of densities, $\limsup\sqrt{n}\,E(J_n)<\infty$ provided that h converges to an appropriate positive constant, and K is a kernel whose characteristic function is one in an open neighborhood of the origin. The latter condition, combined with $\int K=1$, $\int|K|<\infty$, forces K to take both positive and negative values. Since $E(J_n)$ is of the same order of magnitude ($n^{-1/2}$) as $|J_n-E(J_n)|$, relative stability is not obtainable from our theorems in these cases. It even seems that the estimates with constant h are not relatively stable for these classes of densities.

Consider $d=1$ again. Let s be an even positive integer. For general class s kernels, i.e., kernels K that are symmetric, square integrable, and satisfy $\int(1+|x|^s)|K(x)|\,dx<\infty$, $\int K=1$, $\int x^i K(x)\,dx=0$ for $0<i<s$, and $\int x^s K(x)\neq 0$, it is known that

$$\liminf_{n\to\infty}\inf_h n^{\frac{s}{2s+1}}E\int|f_n-f|\geqq c$$

where $c>0$ is a constant depending upon K only (Devroye, 1987). The second statement of Theorem 5 remains valid for class s kernels K and all densities

f on the real line. Also, $|J_n - E(J_n)|/E(J_n) = O(n^{-1/(4s+2)})$ in probability, and, in analogy with Theorem 4,

$$\limsup_{n\to\infty} \frac{n^{\frac{1}{4s+2}}}{\sqrt{\log n}} \frac{|J_n - E(J_n)|}{E(J_n)} \leqq \frac{\sqrt{32}\int |K|}{c} \quad \text{almost surely.}$$

3. Some Lemmas

Lemma 1. *Let X be any random variable with finite mean, and let a be an arbitrary real number. Then*

$$||X-a| - E(|X-a|)| \leqq |X - E(X))| + E(|X - E(X)|).$$

Proof of Lemma 1. Observe that

$$|X-a| - |X-E(X)| \leqq |a - E(X)| \leqq E(|X-a|),$$

and that

$$\begin{aligned} &E(|X-a|) - E(|X-E(X)|) \leqq E(|a-E(X)|) \\ &= |a - E(X)| \leqq |a - X| + |X - E(X)|. \quad \square \end{aligned}$$

Consider next an increasing sequence $\Sigma_0, \ldots, \Sigma_n$ of sub-σ-fields of a basic probability space, where Σ_0 is trivial. A sequence of random variables Z_i, $1 \leqq i \leqq n$, is called a martingale difference sequence if each Z_i is Σ_i-measurable, and if $E(Z_i|\Sigma_{i-1}) = 0$ for each i. We have for every Σ_n-measurable random variable Y,

$$Y - E(Y) = \sum_{i=1}^{n} Z_i$$

where

$$Z_i = E(Y|\Sigma_i) - E(Y|\Sigma_{i-1}),$$

so that the Z_i's form a martingale difference sequence. Most inequalities for sums of iid zero mean random variables are also applicable, with minor modifications, to martingale difference sequences (see e.g., Steiger (1969), Millar (1969), Freedman (1975), Burkholder (1973), Azuma (1967), Chow (1966) and Stout (1974, Sect. 4.2)). One that is particularly useful for us is the following exponential inequality due to Azuma (1967) (see also Stout (1974, pp. 238–239)):

Lemma 2.

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > \varepsilon\right) \leqq 2e^{-\frac{\varepsilon^2}{2\int_{i=1}^{n} \|Z_i\|_\infty^2}}$$

where $\|Z_i\|_\infty$ *is the essential supremum norm of* Z_i.

4. Proof of Theorem 1

Let us formally define $\Sigma_i=\sigma(X_1, \ldots, X_i)$, and

$$Y\equiv\int|f_n-f|-E(\int|f_n-f|).$$

To be able to apply Lemma 2, we need an upper bound on every

$$|Z_i|=|E(\int|f_n-f||\Sigma_i)-E(\int|f_n-f||\Sigma_{i-1})|,$$

since obviously $Y=E(\int|f_n-f||\Sigma_n)-E(\int|f_n-f||\Sigma_0)$. We note that with the notation $W_{i,k}=\sum_{i\leqq j\leqq k}(K_h(x-X_j)-f)/n$,

$$\begin{aligned}|\mathrm{Z}_i|&\leqq\int|E(|f_n-f||\Sigma_i)-E(|f_n-f||\Sigma_{i-1})|\\&=\int|E(|W_{1,i-1}+W_{i,i}+W_{i+1,n}||\Sigma_i)-E(|W_{1,i-1}+W_{i,i}+W_{i+1,n}||\Sigma_{i-1})|\\&\leqq\int\sup_a|E(|a+W_{i,i}|)-|a+W_{i,i}||\\&\leqq\int|W_{i,i}-E(W_{i,i})|+\int E(|W_{i,i}-E(W_{i,i})|)\qquad\text{(Lemma 1)}\\&\leqq\frac{4\int|K|}{n}.\end{aligned}$$

Thus, by Lemma 2,

$$P(|\int|f_n-f|-E(\int|f_n-f|)|>\varepsilon)\leqq2e^{-\frac{\varepsilon^2}{2n(4\int|K|/n)^2}}=2e^{-\frac{n\varepsilon^2}{32\int^2|K|}}.\qquad\square$$

5. Data-Based Smoothing Factors

An *automatic kernel estimate* is a kernel estimate in which the smoothing factor H is a Borel measurable function of the data $X_1, \ldots, X_n$. For particular examples, we refer to chapter 6 of Devroye and Gyorfi (1985). Here we establish the relative stability of most automatic kernel estimates. For the automatic kernel estimate with smoothing factor H, we will write J_n or J_{nH} for the error, and f_n or f_{nH} for the estimate, so as to make the dependence upon H explicit.

We will proceed as in Theorems 1–5. The kernels we will consider here are *smooth*, i.e., there exists a constant C such that

$$\sup_{1\leqq h\leqq c}\int|K-K_h|\leqq C(c^d-1)$$

for all $c>1$. C is called the *smoothness constant*. Smoothness of a kernel implies that small changes in h induce proportionally small changes on f_n with regard to the L_1 distance. For example, let K be nonnegative, nonincreasing along rays, and $\int|K|<\infty$. (A kernel K is said to be *nonincreasing along rays* if $K(ux)\leqq K(x)$ for all $x\in R^d$ and all $u\geqq1$.) For such kernels, we can take $C=2$ (see Devroye and Gyorfi (1985, p. 187)). When $d=1$ and K is absolutely continuous, we can take $C=\int|x||K'(x)|dx+\int|K|$ (Devroye, 1987). Similarly, when K is Lipschitz on the real line, K is smooth.

Theorem 6. *Consider an automatic kernel estimate on R^d based upon a smooth kernel K with smoothness-constant C. Then, for all $\varepsilon > 0$, and all $0 < a < b < \infty$,*

$$P(|J_n - E(J_n)| > \varepsilon) \leqq P(H \notin [a,b]) + \left(2 + \frac{8C + d\varepsilon}{\varepsilon} \log\left(\frac{b}{a}\right)\right) 2e^{-\frac{n\varepsilon^2}{512\int^{2}|K|}}$$

provided that

$$P(H \notin [a,b]) \leqq \frac{\varepsilon}{8(1+\int|K|)}$$

and

$$4\left(1+\int|K|\right)\left(2 + \frac{8C + d\varepsilon}{\varepsilon} \log\left(\frac{b}{a}\right)\right) e^{-\frac{n\varepsilon^2}{2048\int^{2}|K|}} \leqq \frac{\varepsilon}{8}.$$

Theorem 6 is valid for *all densities* on R^d, and will be our starting point. For example, we have

Theorem 7. *Consider an automatic kernel estimate on R^d based upon a smooth kernel K. Assume that the smoothing factor H is such that $P(H \notin [a_n, b_n]) = o(\sqrt{\log n/n})$ for some sequences of positive numbers $a_n \leqq b_n$ having the property that $b_n/a_n = e^{O(n^k)}$ for some finite k. Then,*

$$\sqrt{n}\,|J_n - E(J_n)| = O(\sqrt{\log(n)}) \quad \textit{in probability as } n \to \infty.$$

If in addition $\sum_n P(H \notin [a_n, b_n]) < \infty$, then

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log n}}\,|J_n - E(J_n)| \leqq \sqrt{2048(k+2)}\int|K| \quad \textit{almost surely.}$$

The ratio of the boundaries of the intervals $[a_n, b_n]$ can diverge at any exponential rate. We could take $[e^{-n}, e^n]$ for example. In this light, the conditions on the rate of decrease of $P(H \notin [a_n, b_n])$ are not restrictive at all. Theorems 6 and 7 allow us to say things about $|J_n - E(J_n)|$ provided that we know how widely H varies for the application at hand. However, the Theorems are useless as stated when, say, $H = 0$ with probability 1/2. Granted, such estimates are totally useless, but it is important to make this point nevertheless. Fortunately, when H is extremely small, both J_n and its expected value approach $1 + \int|K|$, and similarly when H tends to infinity for n fixed. Thus, with good bounds on just how close J_n is to $1 + \int|K|$ in these limit cases, it is possible to insure that when $H \notin [a,b]$, $|J_n - E(J_n)|$ is smaller than a given small value. The rate with which J_n tends to $1 + \int|K|$ in the extreme cases depends upon the peakedness and the smoothness of f and K. It should come as no surprise that for further results, just such conditions on f are needed. This is the price paid for letting H swing widely. In Theorem 9, we state the results that can be obtained by just such a technique.

Before we state Theorem 9, it is helpful to recall the following:

Theorem 8 (Devroye, 1987, Theorem S 1). *Consider a kernel estimate on R_1 based upon a smooth absolutely integrable class s kernel. Then*

$$\lim_{n\to\infty} \frac{E(\inf_h J_{nh})}{\inf_h E(J_{nh})} = 1.$$

Theorem 8 is useful for bounding the denominator of $|J_n - E(J_n)|/E(J_n)$ of automatic kernel estimates. For example, it is known that for class s kernels and $d=1$, and all f,

$$\inf_h E(J_{nh}) \geqq (c+o(1))\, n^{-s/(2s+1)},$$

where c is a positive constant depending upon K only (for nonnegative kernels, see Devroye and Gyorfi, 1985; in general, see Devroye, 1987). The fact that the infimum in the inequality is over all $h>0$ does not imply that the same lower bound is valid for the expected L_1 error of all automatic kernel estimates. Indeed, we always have

$$\inf_h E(J_{nh}) \geqq E(\inf_h J_{nh}).$$

However, Theorem 8 tells us that indeed

$$E(\inf_h J_{nh}) \geqq (c+o(1))\, n^{-s/(2s+1)}.$$

Theorem 9. *Consider any automatic kernel estimate on R^1 based upon an absolutely integrable Lipschitz (hence, smooth) kernel K with compact support. Assume that $f \in L \log_+ L$ (i.e. $\int f \log(1+f) < \infty$), and that $\int \log(1+|x|) f(x)\, dx < \infty$. Then*

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log n}}\, |J_{nH} - E(J_{nH})| \leqq \sqrt{10240} \int |K| \qquad \textit{almost surely.}$$

If additionally K is a class s kernel, then the estimate is strongly relatively stable, and in fact,

$$\limsup_{n\to\infty} \frac{n^{\frac{1}{4s+2}}}{\log n} \frac{|J_{nH} - E(J_{nH})|}{E(J_{nH})} \leqq \frac{1}{c} \sqrt{10240} \int |K| \qquad \textit{almost surely,}$$

where c is the positive constant depending upon K only that was introduced immediately following Theorem 5.

Theorem 9 is the only theorem in this paper that does not apply to all densities on the real line. We have explained above why. Extremely small or extremely large values of H occurring with too high a probability cause the statements of Theorem 9 to fail for some densities with enormous infinite tails or very steep peaks. Let us describe the borderline cases. The peak condition ($f \in L \log_+ L$) is violated when $f(x) \sim 1/(x \log^2(1/x))$ as $x \downarrow 0$, but is not when f is monotone $\downarrow$ on $[0, \infty)$, and $f(x) \sim 1/(x \log^2(1/x) \log^{1+\varepsilon} \log(1/x))$ as $x \downarrow 0$, where

$\varepsilon>0$ is arbitrary. The moment condition on f is violated when $f(x)\sim 1/(x\log^2(x))$ as $x\uparrow\infty$, but for monotone densities on the positive halfline, it is satisfied when $f(x)\sim 1/(x\log^2(x)\log^{1+\varepsilon}\log(1/x))$ as $x\uparrow\infty$, where $\varepsilon>0$.

In Theorem 9, as in the previous Theorems, no attempt was made to optimize the constants in the inequality. Also, the conditions on f and K are such that the constants in the asymptotic upper bounds of Theorem are not affected, except perhaps via the factor $\int|K|$. This means that there is some room for improvement. By fine-tuning the arguments, the conditions on f and K can be relaxed to the point that the upper bounds become functionals involving both functions. In view of the fact that K can be chosen by the user, and that the conditions on f are truly weak in the present form, we will not pursue this matter any further.

6. Proofs of Theorems 6–9

Proof of Theorem 6. Let $S=\{h_0, h_1, \ldots, h_m\}$ be a collection of h-values defined as follows:

$$h_i=a\left(1+\frac{d\varepsilon}{8C}\right)^{\frac{i}{d}},\qquad i=0,1,2,\ldots,m.$$

The number m is obtained from the condition that h_m is the first number at least equal to b. Thus, using $\log(1+u)\geqq u/(1+u)$, valid for $u\geqq 0$, we have

$$m=\left[\left(d\log\left(\frac{b}{a}\right)\right)\Big/\log\left(1+\frac{d\varepsilon}{8C}\right)\right]\leqq 1+\frac{8C+d\varepsilon}{\varepsilon}\log\left(\frac{b}{a}\right).$$

Let V be determined from H by the following rule: $V=h_i$ if $h_i\leqq H<h_{i+1}$ for some $0\leqq i<m$. If $H<a$, then $V=a$. If $H>b$, then $V=b$. In S, replace h_m by $\min(h_m, b)$. Let D be the event $[H\in[a,b]]$ and let D^c be its complement. Note that

$$\begin{aligned}|J_{nH}-J_{nV}|&\leqq|J_{nH}-J_{nV}|I_D+2(1+\textstyle\int|K|)I_{D^c}\\&\leqq C\left(\left(1+\frac{d\varepsilon}{8C}\right)^{1/d}-1\right)+2(1+\textstyle\int||K|)I_{D^c}\\&\leqq C\left(\frac{d\varepsilon}{8Cd}\right)+2(1+\textstyle\int|K|)I_{D^c}=\frac{\varepsilon}{8}+2(1+\textstyle\int|K|)I_{D^c}\end{aligned}$$

and

$$|E(J_{nH})-E(J_{nV})|\leqq\frac{\varepsilon}{8}+2(1+\textstyle\int|K|)P(D^c).$$

Thus

$$\begin{aligned}&P(|J_{nH}-E(J_{nH})|>\varepsilon)\\&\quad\leqq P\left(|J_{nV}-E(J_{nV})|+\frac{\varepsilon}{4}+2(1+\textstyle\int|K|)I_{D^c}+2(1+\textstyle\int|K|)P(D^c)>\varepsilon\right)\\&\quad\leqq P\left(|J_{nV}-E(J_{nV})|>\frac{\varepsilon}{2}\right)+P(D^c)\end{aligned}$$

when $P(D^c) \leqq \varepsilon/(8(1+\int|K|))$. Next,

$$\begin{aligned}|J_{nV}-E(J_{nV})| &= |\sum_i J_{nh_i} I_{[V=h_i]} - E(\sum_i J_{nh_i} I_{[V=h_i]})|\\ &= |\sum_i (J_{nh_i}-E(J_{nh_i}))\, I_{[V=h_i]} - E(\sum_i (J_{nh_i}-E(J_{nh_i}))\, I_{[V=h_i]})|\\ &\leqq \max_i |J_{nh_i}-E(J_{nh_i})| + E(\max_i |J_{nh_i}-E(J_{nh_i})|)\end{aligned}$$

where we used the fact that $\sum_i I_{[V=h_i]}=1$. Thus,

$$\begin{aligned}&P(|J_{nH}-E(J_{nH})|>\varepsilon)\\ &\quad\leqq P\left(\max_{0\leqq i\leqq m} |J_{nh_i}-E(J_{nh_i})| + E(\max_{0\leqq i\leqq m} |J_{nh_i}-E(J_{nh_i})|) > \frac{\varepsilon}{2}\right) + P(D^c)\\ &\quad\leqq (m+1)\sup_{h>0} P\left(|J_{nh}-E(J_n h)| > \frac{\varepsilon}{4}\right) + P(D^c)\end{aligned}$$

when $P(D^c) \leqq \varepsilon/(8(1+\int|K|))$ and $E(\max_i |J_{nh_i}-E(J_{nh_i})|) \leqq \varepsilon/4$. The last condition is satisfied when $2(1+\int|K|)(m+1)\sup_{h>0} P(|J_{nh}-E(J_{nh})|>\varepsilon/8) \leqq \varepsilon/8$. Now apply Theorem 1. □

Proof of Theorem 7. Theorem 7 follows directly from Theorem 6, first by taking $\varepsilon = c\sqrt{\log n/n}$ for some large constant c, and then by setting c equal to $\sqrt{2048(k+2+\delta)}\int|K|$ for some small $\delta>0$, and applying the Borel-Cantelli lemma. □

Proof of Theorem 9. k and q are positive constants to be picked further on. Define the interval $A=[a,b] \overset{\Delta}{=} [e^{-n^k}, e^{n^q}]$, and relate the random variable H^* to H in the following manner:

$$H^* = \begin{cases} e^{n^q}, & \text{if } H > e^{n^q}\\ H, & \text{if } e^{n^q} \geqq H \geqq e^{-n^k}.\\ e^{-n^k}, & \text{if } e^{-n^k} > H\end{cases}$$

By Theorem 7,

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log n}}\, |J_{nH^*}-E(J_{nH^*})| \leqq \sqrt{2048(k+q+2)}\int|K| \qquad \text{almost surely.}$$

(We will later see that we can take k and q slightly larger than $\frac{5}{2}$ and $\frac{1}{2}$ respectively.) Also,

$$|J_{nH}-J_{nH^*}| \leqq 2\sup_{h\notin A} |1+\int|K|-J_{nh}|.$$

By the triangle inequality, the fact that $|J_{nh}| \leqq 1 + \int |K|$, and an application of Theorem 8, Theorem 9 is proved if we can show that

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log n}} \sup_{h\notin A} |1+\int|K|-J_{nh}| = 0 \quad \text{almost surely}$$

and

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log n}}\, E\Big(\sup_{h\notin A} |1+\int|K|-J_{nh}|\Big) = 0.$$

We will achieve this by showing that for every θ, $\zeta > 0$, we have for all n large enough,

$$P\left(\sup_{h\notin A} |1+\int|K|-J_{nh}| > \theta \sqrt{\frac{\log n}{n}}\right) \leqq \frac{2}{n^{1+\xi}},$$

and applying the Borel-Cantelli lemma. This will be done in turn by considering the suprema over $h<a$ and $h>b$ separately.

Assume that K vanishes off $[-s, s]$. Take $a = \frac{s\delta}{2n}$ where $\delta > 0$ is a constant to be picked further on. Let N be the number of X_i's for which $[X_i - 2a, X_i + 2a]$ has at least one X_j with $j \neq i$, and let D be the union of the sets $[X_i - a, X_i + a]$ for those X_i *not* counted in N. Note that $1 \geqq \int_D |f_{nh}|/\int|K| \geqq \frac{n-N}{n}$, uniformly over $h \leqq a$. We have for arbitrary $\varepsilon > 0$,

$$P(\sup_{h<a} |\int|f_{nh}-f| - (1+\int|K|)| \geqq \varepsilon) \leqq P\left(\sup_{h<a} \int_D |f_{nh}| \leqq \int|K| - \frac{\varepsilon}{3}\right)$$

if $\delta < \rho(f, \varepsilon)$ where $\rho \triangleq s^{-1}\int\psi(f)\exp\left(-\frac{12\int\psi(f)}{\varepsilon}\right)$, and $\psi(u) \triangleq u\log(1+u)$. To show this, we note that ψ is convex, and that for $u \geqq 0$, $\psi^{-1}(u) \leqq 2u/\log(1+u)$. Let λ be Lebesgue measure. For any set B with $\lambda(B) \leqq \int\psi(f)$,

$$\begin{aligned}\int_B f &\leqq \lambda(B)\,\psi^{-1}\left(\frac{\int\psi(f)}{\lambda(B)}\right) \quad \text{(Jensen's inequality)}\\ &\leqq \frac{2\int\psi(f)}{\log(1+\int\psi(f)/\lambda(B))}.\end{aligned}$$

Take δ so small that uniformly over all sets B with $\lambda(B) < s\delta$, $\int_B f < \varepsilon/6$. By the inequality given above, this can be done if $\delta \leqq s^{-1}\int\psi(f)\Big/\left(\exp\left(\frac{12\int\psi(f)}{\varepsilon}\right) - 1\right)$. Note in passing that $\lambda(D) \leqq s\delta$. It suffices to show that $\sup_{B:\lambda(B)<s\delta} \int_B |f_{nh}| \geqq \int|K| - \varepsilon/3$ implies that $\int|f_{nh}-f| \geqq 1 + \int|K| - \varepsilon$. We have $\int_B |f_{nh}-f| \geqq \int|K|$

$-\varepsilon/3-\int_B f$, $\int_{B^c}|f_{nh}-f|\geqq\int_{B^c}f-\int_{B^c}|f_{nh}|$, which is at least $1-\int_B f-\varepsilon/3$. Summing this and noting that $\int_B f\leqq\varepsilon/6$ shows that $\int|f_{nh}-f|\geqq 1+\int|K|-\varepsilon$.

The previous facts can now be combined to conclude that for δ small enough,

$$P(\sup_{h<a}|\int|f_{nh}-f|-(1+\int|K|)|\geqq\varepsilon)\leqq P\left(\frac{N}{n}>\frac{\varepsilon}{3\int|K|}\right)+P\left(\frac{EN}{n}>\frac{\varepsilon}{3\int|K|}\right).$$

By Markov's inequality, this can be made smaller than a given small constant ε_1 if

$$\frac{EN}{n}\leqq\min(1,\varepsilon_1)\frac{\varepsilon}{3\int|K|}\overset{\Delta}{=}\varepsilon_2.$$

Note that we will take $\varepsilon=\theta\sqrt{\log n/n}$, and $\varepsilon_1=1/n^{1+\xi}$. Thus, for all n large enough, $\varepsilon_2=(\theta\sqrt{\log n})/(3\int|K|n^{\zeta+3/2})$. But

$$\begin{aligned}EN/n&\leqq\int f(x)\min\left(1,n\int_{x-s\delta/n}^{x+s\delta/n}f(y)\,dy\right)dx\\&\leqq 2s\delta\int f(x)\frac{n}{2s\delta}\int_{x-s\delta/n}^{x+s\delta/n}f(y)\,dy\,dx\\&\leqq 2s\delta\int f(x)\psi^{-1}\left(\frac{\int\psi(f)}{2s\delta/n}\right)\qquad\text{(Jensen's inequality)}\\&\leqq\frac{2n\int\psi(f)}{\log(1+n\int\psi(f)/(2s\delta))}.\end{aligned}$$

Thus, we need to choose δ such that

$$\delta\leqq\frac{n}{2s}\int\psi(f)\bigg/\left(\exp\left(\frac{2n\int\psi(f)}{\varepsilon_2}\right)-1\right).$$

We can now choose k because $a=\exp(-n^k)$ and $a=s\delta/(2n)$. One can verify that for n large enough,

$$\delta=\frac{2n}{s}e^{n^k}\leqq\frac{n}{2s}\int\psi(f)\bigg/\left(\exp\left(\frac{2n\int\psi(f)}{\varepsilon_2}\right)-1\right)$$

when $k=\frac{5}{2}+\zeta+\eta$ for arbitrary small $\eta>0$. We conclude that for all n large enough,

$$P(\sup_{h<a}|\int|f_{nh}-f|-(1+\int|K|)|\geqq\theta\sqrt{\frac{\log n}{n}}\right)\leqq\frac{1}{n^{1+\zeta}}.$$

We finally consider the case $h>b$. Let ω be the modulus of continuity of K defined by $\omega(u)=\sup_x\sup_{|y|\leqq u}|K(x)-K(x+y)|$. By our assumptions on K, $\omega(u)$

$\leqq lu$ for some constant $l>0$ and all $u>0$. Let t and $T>t$ be positive numbers chosen in such a way that $\int_{t\leqq|x|\leqq T}|K|\geqq\int|K|-\frac{\varepsilon}{4}$ and $\sup_z\int_{z-t}^{z+t}|K|\leqq\frac{\varepsilon}{4}$. Also, T should be so large that $\int_{|x|\geqq T} f<\varepsilon/(6\int|K|)$. This fixes t and T once and for all. Let N be the number of X_i's with $|X_i|\geqq T$. We have the following inequality:

$$\int_{th\leqq|x|\leqq Th}|f_{nh}-K_h|\leqq(T-t)\,\omega\left(\frac{T}{b}\right)+\frac{N}{n}\int|K|.$$

This can best be seen by noting that

$$\begin{aligned}|f_{nh}-K_h|&=\left|\frac{1}{n}\sum_{i=1}^{n}(K_h(x-X_i)-K_h(x))\right|\\&\leqq\frac{1}{n}\sum_{i:|X_i|\leqq T}\sup_{|y|\leqq T}|K_h(x-y)-K_h(x)|+\frac{1}{n}\sum_{i:|X_i|>T}|K_h(x-X_i)|\\&\leqq\frac{1}{h}\,\omega\left(\frac{T}{h}\right)+\frac{1}{n}\sum_{i:|X_i|>T}|K_h(x-X_i)|.\end{aligned}$$

Now, integrating over the given interval and noting that $h\geqq b$ yields the result. For $h\geqq b$,

$$\begin{aligned}\int|f_{nh}-f|\geqq&\int_{th\leqq|x|\leqq Th}|f_{nh}|-\int_{th\leqq|x|\leqq Th}f+\int_{|x|\leqq th}f-\int_{|x|\leqq th}|f_{nh}|\\\geqq&\int_{th\leqq|x|\leqq Th}|K_h|-\int_{th\leqq|x|\leqq Th}|f_{nh}-K_h|-\int_{th\leqq|x|\leqq Th}f\\&+\int_{|x|\leqq th}f-\frac{1}{n}\sum_{i=1}^{n}\int_{|x|\leqq th}|K_h(x-X_i)|\\\geqq&\int|K|-\frac{\varepsilon}{4}-(T-t)\,\omega\left(\frac{T}{h}\right)-\int|K|\frac{N}{n}+1-2\int_{th\leqq|x|}f-\sup_z\int_{z-t}^{z+t}|K|\\\geqq&\int|K|-\frac{\varepsilon}{4}-(T-t)\,\omega\left(\frac{T}{b}\right)-\int|K|\frac{N}{n}+1-2\int_{tb\leqq|x|}f-\frac{\varepsilon}{4}\\\geqq&\int|K|+1-\frac{2\varepsilon}{3}-\int|K|\frac{N}{n}\end{aligned}$$

if b is so large that $(T-t)\,\omega(T/b)+2\int_{th\leqq|x|}f\leqq\varepsilon/6$. Thus,

$$P(\sup_{h\geqq b}\int|f_{nh}-f|\leqq1+\int|K|-\varepsilon)\leqq P\left(\frac{N}{n}\geqq\frac{\varepsilon}{3\int|K|}\right).$$

Note that $EN/n = \int\limits_{|x| \geqq T} f < \varepsilon/(6\int|K|)$ by our choice of T. Thus,

$$P\left(\frac{N}{n} \geqq \frac{\varepsilon}{3\int|K|}\right) \leqq P\left(\frac{N-EN}{n} \geqq \frac{\varepsilon}{6\int|K|}\right) \leqq e^{-\frac{n\varepsilon}{24\int|K|}}.$$

The last inequality is obtained by an application of Bennett's inequality (Bennett, 1962), which in a form convenient for us states that if X is binomial (n, p), then $P(X-EX>nu) \leqq \exp(-nu^2/(2p(1-p)+2u))$ for $u>0$. This is used with $p \leqq u = \varepsilon/(6\int|K|)$. We now replace ε by $\theta\sqrt{\log n/n}$, and note that, provided that we can choose t, T and b as required for such an ε,

$$P\left(\sup_{h \geqq b} \left|\int|f_{nh}-f| - (1+\int|K|)\right| > \theta\sqrt{\frac{\log n}{n}}\right) \leqq e^{-\frac{\theta\sqrt{n\log n}}{24\int|K|}}.$$

It remains to be verified that we can choose t, T and b. By the Lipschitz condition, we see that $|K| \leqq \sqrt{l\int|K|} \overset{\Delta}{=} K^*$, so that we can pick $t = \varepsilon/(8K^*)$. Next, T is picked by

$$T = \exp\left(\frac{6\int|K|\int\log(1+|x|)f(x)\,dx}{\varepsilon}\right).$$

This insures that

$$\int\limits_{|x| \geqq T} f \leqq \frac{\int\log(1+|x|)f(x)\,dx}{\log(1+T)} \leqq \frac{\varepsilon}{6\int|K|}.$$

The condition $\int\limits_{t \leqq |x| \leqq T} |K| \geqq \int|K| - \frac{\varepsilon}{4}$ is obviously satisfied for n large enough, by the choice of t, and the fact that $T \to \infty$. Finally, the condition involving b is satisfied if $lT^2/b \leqq \varepsilon/12$ and $\int\limits_{tb \leqq |x|} f \leqq \varepsilon/24$. Using the same logarithmic bounding technique again, we see that it suffices to take

$$b \geqq \max\left(\frac{12lT^2}{\varepsilon}, \frac{1}{t}\exp\left(\frac{24\int\log(1+|x|)f(x)\,dx}{\varepsilon}\right)\right).$$

Since $b = \exp(n^q)$, we see that such a b can be found provided that $q > 1/2$. □

References

Abou-Jaoude, S.: La convergence L_1 et L_∞ de certains estimateurs d'une densite de probabilite. These de Doctorat d'Etat, University Paris VI, France, 1977

Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J., Ser. II. **37**, 357–367 (1967)

Bennett, G.: Probability inequalities for the sum of independent random variables. J. Am. Stat. Assoc. **57**, 33–45 (1962)

Burkholder, D.L.: Distribution function inequalities for martingales. Ann. Probab. **1**, 19–42 (1973)

Chow, Y.S.: Some convergence theorems for independent random variables. Ann. Math. Stat. **37**, 1482–1493 (1966)

Devroye, L.: The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. Ann. Stat. **11**, 896–904 (1983)

Devroye, L., Gyorfi, L.: Nonparametric density estimation. The L_1 View, New York: Wiley 1985

Devroye, L.: Asymptotic performance bounds for the kernel estimate. Technical Report, School of Computer Science, McGill University, Montreal, Canada, 1987

Devroye, L.: An application of the Efron-Stein inequality in density estimation. Ann. Stat. **15**, 1317–1320 (1987)

Devroye, L.: An L1 asymptotically optimal kernel estimate. Technical Report, School of Computer Science, McGill University, Montreal, Canada, 1987

Freedman, D.A.: On tail probabilities for martingales. Ann. Probab. **3**, 100–118 (1975)

Hall, P.: Limit theorems for stochastic measures of the accuracy of density estimators. Stochastic Processes Appl. **13**, 11–25 (1982)

Hall, P.: Central limit theorem for integrated square error of multivariate nonparametric density estimators. J. Multivariate Anal. **14**, 1–16 (1984)

Millar, P.W.: Martingales with independent increments. Ann. Math. Stat. **40**, 1033–1041 (1969)

Parzen, E.: On the estimation of a probability density function and the mode. Ann. Math. Stat. **33**, 1065–1076 (1962)

Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **27**, 832–837 (1956)

Steiger, W.L.: A best possible Kolmogoroff-type inequality for martingales and a characteristic property. Ann. Math. Stat. **40**, 764–769 (1969)

Stout, W.F.: Almost sure convergence. New York: Academic Press 1974

Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. Cambridge: Cambridge Universit) Press 1927

Received September 2, 1986; received in revised form November 13, 1987