The Annals of Statistics
1994, Vol. 22, No. 1, 185–210

BIAS REDUCTION IN KERNEL DENSITY ESTIMATION BY SMOOTHED EMPIRICAL TRANSFORMATIONS

BY DAVID RUPPERT[1] AND DAREN B. H. CLINE[2]

Cornell University and Texas A&M University

A modification of kernel density estimation is proposed. The first step is ordinary kernel estimation of the density and its cdf. In the second step the data are transformed, using this estimated cdf, to an approximate uniform (or normal or other target) distribution. The density and cdf of the transformed data are then estimated by the kernel method and, by change of variable, converted to new estimates of the density and the cdf of the original data. This process is repeated for a total of k steps for some integer k greater than 1.

If the target density is uniform, then the order of the bias is reduced, provided that the density of the observed data is sufficiently smooth. By proper choice of bandwidth, rates of squared-error convergence equal to those of higher-order kernels are attainable. More precisely, k repetitions of the process are equivalent, in terms of rate of convergence, to a $2k$-th-order kernel. This transformation-kernel estimate is always a bona fide density and appears to be more effective at small sample sizes than higher-order kernel estimators, at least for densities with interesting features such as multiple modes. The main theoretical achievement of this paper is the rigorous establishment of rates of convergence under multiple iteration.

Simulations using a uniform target distribution suggest that the possibility of improvement over ordinary kernel estimation is of practical significance for samples sizes as low as 100 and can become appreciable for sample sizes around 400.

1. Introduction. Suppose we have an independent sample $X_1, \ldots, X_n$ from a density f_X. Many methods have been proposed for the estimation of f_X. One that has been extensively studied and that is appealing for its simplicity is the kernel density estimator (KDE):

$$\hat{f}_X(x; h) = n^{-1} \sum_{i=1}^{n} K_h\{x - X_i\}.$$

Here $K_h\{x\} = h^{-1}K\{h^{-1}x\}$, where K is a symmetric "kernel" function that integrates to 1. See Silverman (1986) for an excellent overview of kernel density estimation.

Received March 1991; revised January 1993.

[1] Partially supported by the Army Research Office through the Mathematical Science Institute at Cornell University and by NSF Grant DMS-90-02791.

[2] Partially supported by the Army Research Office through the Mathematical Science Institute at Cornell University and by NSF Grant DMS-90-01011.

AMS 1991 *subject classifications.* 62G07, 62G20.

Key words and phrases. Bandwidth selection, bias reduction, boundary kernels, empirical processes, improved rates of convergence, transformation to a target distribution, variable bandwidths.

The bias at x has a formal asymptotic expansion of the form

$$\sum_{j=1}^{\infty} h^{2j} f_X^{(2j)}(x) \int u^{2j} K(u)\, du/(2j)!.$$

One approach to bias reduction is to choose K so that $\int u^{2j} K(u)\, du = 0$, for $j = 1, \ldots, k-1$ for some positive integer k [Parzen (1962), Bartlett (1963), Singh (1977, 1979)]. A kernel with this property and $\int u^{2k} K(u)\, du \neq 0$ is said to be of order $2k$. If K is nonnegative, which guarantees that f_X is a bona fide density, then K can be of order at most 2.

Despite their potential bias reduction, the so-called higher-order kernels of order 4 or more have several drawbacks. They may give a negative density estimate, although this problem is easily fixed [Gajek (1986)]. More seriously, exact integrated mean squared error (IMSE) results of Marron and Wand (1992) suggest that higher-order kernels only improve significantly over the usual second-order nonnegative kernels for densities without interesting features or for enormous samples. Our empirical results in Section 3 corroborate the results of Marron and Wand (1992) and show that for the densities that we investigate Gajek's modification has very little effect on the IMSE of a higher-order kernel estimate.

A second approach to bias reduction was found by Abramson (1982) and is called the adaptive kernel estimator by Silverman (1986). Abramson's estimator is discussed in more detail in Section 3. Also see Jones (1990), Hall and Marron (1988) and Hall (1990).

A different approach to bias reduction is taken here. First the data are transformed to $Y_i = g(X_i)$, where g is a smooth, monotonic function; g will be chosen so that the density of Y_i,

$$f_Y(y;g) = f_X(g^{-1}(y)) \left\{ \frac{d}{dy} g^{-1}(y) \right\},$$

has

$$\left| f_Y^{(2j)}(y;g) \right| = \left| \frac{d^{2j+1}}{dy^{2j+1}} F_X\left(g^{-1}(y)\right) \right|$$

"small", at least for $j = 1, \ldots, k$ for some positive k. With g chosen in this manner, $f_Y(y;g)$ can be accurately estimated by an ordinary KDE. The KDE

$$\hat{f}_Y(y;g,h) = n^{-1} \sum_{i=1}^{n} K_h\{y - Y_i\} \tag{1.1}$$

can then be "back-transformed" by change of variables to an estimator of f_X:

$$\hat{f}_X(x;g,h) = \hat{f}_Y(g(x);g,h) g'(x). \tag{1.2}$$

We will call $\hat{f}_X(x;g,h)$ a transformation-kernel density estimator (TKDE).

For this strategy to be effective, the function g must depend on the X_i's. In this paper, $g = G \circ \hat{F}_X$, where $\hat{F}_X$ is a *smooth* estimate of the cdf F_X of $X_1, \ldots, X_n$, and G is the inverse cdf of some target distribution, for example, the uniform or normal distribution.

A uniform target distribution (where G is the identity function) is particularly interesting, since its density has all derivatives equal to 0 so that bias is asymptotically negligible. For example, suppose K has support $[-1, 1]$. If g were exactly F_X so that f_Y were exactly uniform[0, 1], then $\hat{f}_Y(y;g,h)$ would be unbiased for y in $[h, 1-h]$ and therefore $\hat{f}_X(x;g,h)$ would be unbiased for x in $[g^{-1}(h), g^{-1}(1-h)]$. Moreover, the bias near the boundaries could be eliminated by the use of a so-called boundary kernel (see Section 3).

Of course, bias is not completely eliminated when g is only an estimate of F_X, but in Section 2 an asymptotic study of the TKDE is made. It is shown that if a uniform target distribution is used, then the order of the bias can be reduced by the transformation. For example, suppose that g is $\hat{F}_X(x;h_1)$, the indefinite integral of the ordinary KDE $\hat{f}_X(x;h_1)$. If $h_i = c_i n^{-1/9}$ and $c_i > 0$, for $i = 1, 2$, then the squared error of $\hat{f}_X(x;\hat{F}_X(\cdot;h_1),h_2)$ is of order $O_P(n^{-8/9})$ as $n \to \infty$ rather than $O_P(n^{-4/5})$ as for an ordinary KDE—see Section 2. Moreover, this process can be iterated, letting g be $\hat{F}_X(x;\hat{F}_X(\cdot;h_1),h_2)$, the indefinite integral of $\hat{f}_X(x;\hat{F}_X(\cdot;h_1),h_2)$, and so on. By Theorem 2.1, if h_i is of order $n^{-1/(4t+1)}$, for $i = 1, \ldots, t$, then the squared error of the t-step TKDE is $O_P(n^{-(4t)/(4t+1)})$. Here the 1-step TKDE is the ordinary KDE, the 2-step TKDE is $\hat{f}_X(x;\hat{F}_X(\cdot;h_1),h_2)$, and so on.

Thus, the uniform target TKDE can achieve the rates of squared-error convergence of the higher-order kernels [see Singh (1979)]. Like the higher-order kernels, TKDE's will, of course, require the existence of higher-order derivatives of f_X, since these derivatives are necessary for the existence of estimators achieving such faster rates uniformly over classes of densities; see Stone (1980) for details. The empirical work in Section 3 shows that for some densities the TKDE produces noticeably different estimates than do higher-order kernels. Densities with sharp or multiple peaks are estimated more accurately by the TKDE than by higher-order kernels. The normal density is somewhat more accurately estimated by a higher-order kernel estimate, although the TKDE is quite acceptable at the normal density and is superior to a second-order kernel estimate there.

All empirical work in this paper uses the uniform target. However, we anticipate that other target distributions may prove to be useful. We have had some success with the normal target, although the results are too preliminary to report here. In our theoretical work, we assume an arbitrary smooth target G whenever this assumption is no more difficult to work with than to assume a uniform target.

The use of transformations in kernel density estimation has been proposed by Devroye and Györfi (1985), Silverman (1986) and Wand, Marron and Ruppert (1991). Wand, Marron and Ruppert consider only parametric fami-

lies of g's, which does not lead to improved rates of convergence. Devroye and Györfi do not recommend nonparametric transformations because of the difficulties in establishing consistency and rates of convergence of the resulting estimators. The present paper appears to be the first to investigate nonparametric transformations.

Rudemo (1991) has suggested a parametric TKDE with uniform target. His proposal is to fit a parametric model to the data and to transform by the model cdf evaluated at the maximum likelihood estimator. Of course, it is only assumed that the parametric model provides a rough fit to f_X—if f_X were known to belong to the parametric family, then one would stop at the MLE. Rudemo's interesting proposal seems worth pursuing, but, like other parametric TKDE's, it will not achieve rates of squared-error convergence faster than $O(n^{-4/5})$ unless the density actually is a member of the parametric family.

2. Asymptotics. The following assumptions will be used throughout.

(A1) $X_1, \ldots, X_n$ are iid from the density f_X, $f_X(x) > 0$, $t \geq 1$ is an integer and f_X has $2t$ bounded derivatives in a neighborhood of x. Define $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$.

(A2) K is a symmetric kernel with support $[-1, 1]$ and with $2t + 2$ continuous derivatives.

(A3) G is a $(2t + 1)$-times continuously differentiable function on $[0, 1]$ and $G'(F_X(x)) > 0$.

(A4) For $j = 1, \ldots, t$, $h_j \downarrow 0$ and $\widehat{f}_j$ and $\widehat{F}_j$ are defined as follows:

$$\widehat{f}_1(x) = \widehat{f}_X(x; h_1) = n^{-1} \sum_{i=1}^{n} K_{h_1}\{x - X_i\},$$

$$\widehat{F}_1(x) = \int_{-\infty}^{x} \widehat{f}_1(u)\,du;$$

if $t \geq 2$, then, for $j = 2, \ldots, t$,

$$\begin{aligned}\widehat{f}_j(x) &= \widehat{f}_X(x; G \circ \widehat{F}_{j-1}, h_j) \\ &= n^{-1} \sum_{i=1}^{n} K_{h_j}\left\{G[\widehat{F}_{j-1}(x)] - G[\widehat{F}_{j-1}(X_i)]\right\} G'[\widehat{F}_{j-1}(x)]\widehat{f}_{j-1}(x)\end{aligned}$$

and

$$\widehat{F}_j(x) = \int_{-\infty}^{x} \widehat{f}_j(u)\,du.$$

Let $\widehat{f}_j^{(l)}(x)$ be the l-th derivative of $\widehat{f}_j$.

Our first result shows that using the uniform target distribution achieves the same rate of convergence as with a higher-order kernel. See Section 5 for

proofs. In (2.2) and elsewhere, the convergence of the TKDE and its derivatives is established at each x uniformly over a shrinking neighborhood of x. Although such a result may appear only slightly more general than pointwise convergence, it should be clear from the proofs why this extra generality was sought; to study the j-th iterate at x, one needs convergence of the $(j-1)$-th iterate uniformly in a neighborhood of x.

THEOREM 2.1. *Let G be the identity function on $[0,1]$. Define $h_0 = n^{-1/(4t+1)}$. Suppose that $t \geq 2$ and, for $j = 1, \ldots, t$,*

$$h_j/h_0 \to c_j > 0 \quad \text{as } n \to \infty. \tag{2.1}$$

Then, for each $M > 0$, $1 \leq j \leq t$ and $0 \leq l \leq 2t$, we have

$$\sup_{|x'-x| \leq Mh_0} \left|\widehat{f}_j^{(l)}(x') - f^{(l)}(x')\right| = O_P\left(h_j^{\min(2j,\,2t-l)}\right). \tag{2.2}$$

The following special case is worth noting.

COROLLARY 2.1. *Under the assumptions of Theorem 2.1, we have, for $t \geq j \geq t - l/2$,*

$$\sup_{|x'-x| \leq Mh_0} \left|\widehat{f}_j^{(l)}(x') - f^{(l)}(x')\right| = O_P\left(n^{-(2t-l)/(4t+1)}\right). \tag{2.3}$$

REMARK 2.1. Theorem 2.1 covers only the case where the maximum number of bounded derivatives of f_X in a neighborhood of x is even. Suppose f_X has only $2t-1$ bounded derivatives in a neighborhood of x and (2.1) holds for $h_0 = n^{-1/(4t-1)}$. Then instead of (2.2) one can obtain

$$\sup_{|x'-x| \leq Mh_0} \left|\widehat{f}_j^{(l)}(x') - f^{(l)}(x')\right| = O_P\left(h_j^{\min(2j,\,2t-l-1)}\right),$$

for $1 \leq j \leq t$ and $0 \leq l \leq 2t-1$, and instead of (2.3) one can obtain, for $t \geq j \geq t - l/2 - 1/2$,

$$\sup_{|x'-x| \leq Mh_0} \left|\widehat{f}_j^{(l)}(x') - \widehat{f}^{(l)}(x')\right| = O_P\left(n^{-(2t-1-l)/(4t-1)}\right). \tag{2.4}$$

Together (2.3) and (2.4) show that rates for squared-error convergence given by Singh (1979) for higher-order kernel estimators are achievable by a TKDE. These rates are optimal within the class of all density estimators, both for f_X and for its derivatives, by results in Stone [(1980), Section 3].

Now consider use of a normal or other nonuniform target distribution. The following result shows that if h_1 converges to 0 more slowly than h_2, then $\widehat{f}_2$ behaves asymptotically as if one used the exact, rather than the estimated, transformation to the target.

THEOREM 2.2. *Let $t = 2$ in* (A1)–(A4). *Let G^{-1} be a cdf so that the cdf of Y is $F_Y = G^{-1}$. Let h_1 and h_2 satisfy*

$$h_1 = c_1 n^{-1/5}(\log n), \quad \text{for some } c_1 > 0 \quad \text{and} \quad n^{1/5}h_2 \to c_2 > 0.$$

Let $T = G \circ F_X$. Then, for any fixed $M > 0$ and letting $h_0 = n^{-1/5}$,

$$\sup_{|x'-x| \le Mh_0} \left|\hat{f}_2(x') - \hat{f}_X(x'; T, h_2)\right| = o_P\left(n^{-2/5}\right). \tag{2.5}$$

Let $\kappa_1 = \int u^2 K(u)\,du$ and $\kappa_2 = \int K^2(u)\,du$. Then

$$\begin{aligned} &n^{2/5}\left(\hat{f}_2(x) - f_X(x)\right) \\ &\quad \to_D N\left(\frac{c_2^2}{2}\kappa_1 f_Y^{(2)}(T(x))T'(x),\, c_2^{-1}\kappa_2 f_Y(T(x))[T'(x)]^2\right). \end{aligned} \tag{2.6}$$

With the bandwidths given as in Theorem 2.1, the asymptotic distribution of $\hat{f}_t(x)$ appears quite complicated. In contrast, Theorem 2.2 shows the asymptotics to be rather simple for nonuniform targets. As the proof of Theorem 2.2 reveals, the reason for the simple asymptotics when using a nonuniform target is that higher rates of convergence are not being achieved. However, the bias of the KDE at the Gaussian distribution can be reduced by using a Gaussian kernel and correcting for variance-inflation; see Jones (1991). Using Jones' estimator after an estimated transformation to normality seems promising, but is beyond the scope of this paper. On the other hand, when using the normal target, the following theorem shows that if one is willing to sacrifice a little on the rate of convergence, then one can get a simple asymptotic distribution. This could be useful for obtaining confidence intervals. Note also that (2.6) forms the basis for selecting the constant c_2 and that c_1 plays no role in the asymptotic distribution.

THEOREM 2.3. *Let G and h_0 be as in Theorem* 2.1, *suppose that $t \ge 2$, that* (2.1) *holds for $j = 1, \ldots, t-1$ and that*

$$h_t/h_{t-1} \to 0. \tag{2.7}$$

Then

$$(nh_t)^{1/2}\left(\hat{f}_t(x) - \hat{f}_X(x; F_X, h_t)\right) = o_P(1), \tag{2.8}$$

so that

$$(nh_t)^{1/2}\left(\hat{f}_t(x) - f_X(x)\right) \to_D N\left(0, \kappa_2(f_X(x))^2\right), \tag{2.9}$$

where κ_2 is as in Theorem 2.2.

3. Examples. We implemented a discretized TKDE by linearly prebinning the data to 160 bins [see Jones (1989) for a definition of linear prebinning]. Since the bin width is much smaller than the bandwidths used, prebinning has little effect on the IMSE of the estimate [Jones (1989)]. In fact, we were unable to see any effects of prebinning when comparing prebinned and nonbinned estimates visually.

In every iteration, the bandwidth was $h = H \times IQR$, where H is a "bandwidth factor" that is fixed across iterations and IQR is the interquartile range of the transformed data, the initial transformation being the identity function. Some initial experimentation showed no advantage in terms of accuracy of the estimate in allowing H to vary between iterations—we have not yet pursued the idea suggested by Theorem 2.3 of letting the final H be smaller than the others for inferential purposes. The Gaussian kernel ϕ was used for the first iteration. To handle the boundaries of the uniform target distribution, in subsequent iterations we replaced the Gaussian kernel by the boundary kernel

$$K(Y_i; y, h) = \frac{1}{h}\left\{\phi\left(\frac{Y_i - y}{h}\right) + \phi\left(\frac{Y_i + y}{h}\right) + \phi\left(\frac{2 - y - Y_i}{h}\right)\right\}, \quad \text{for } 0 \le y, Y_i \le 1.$$

A referee questioned the use of a noncompactly supported kernel because of the boundary effects after transformation. While this might be a problem with larger bandwidths, it did not seem to be a problem here. In particular, we also tried the compacted supported triweight kernel, scaled to have variance equal to 1, and found virtually no difference between the triweight and Gaussian kernels.

The second row of plots in Figure 1 shows the first, second and sixth iterates of the TKDE with $H = 0.23$ applied to five random samples of size $n = 400$ from the normal mixture, $\frac{2}{3}N(0, 1) + \frac{1}{3}N(1, (0.2)^2)$. On this plot we show the integrated squared bias (IBIAS), integrated variance (IVAR) and IMSE estimated from 500 Monte Carlo samples with integration of the squared bias, variance and mean squared error performed over the interval $(-4, 4)$. The value of H was chosen to minimize the estimated IMSE of the sixth iteration. The true density is plotted as a dotted curve. The first iterate is the ordinary KDE and has a substantial bias near the peak at 1. This bias is gradually removed by iteration, although the estimates become somewhat more variable with iteration. Keeping $H = 0.23$, we computed 12 iterations and found that the iterates did not converge but instead became increasingly variable. Of course, this nonconvergence does not contradict the theoretical results of Section 2.

For comparison, we plotted kernel density estimates (KDE) in the top row and Abramson's (1982) adaptive kernel density estimator (AKDE) in the bottom row for the *same* five samples. In the top row, KORD is the order of the kernel. The kernels for the KDE's were the Gaussian-based kernels of Deheuvels (1977) [see also Wand and Schucany (1990)], so KORD = 2 means the Gaussian kernel. The KDE's were converted to bona fide densities by

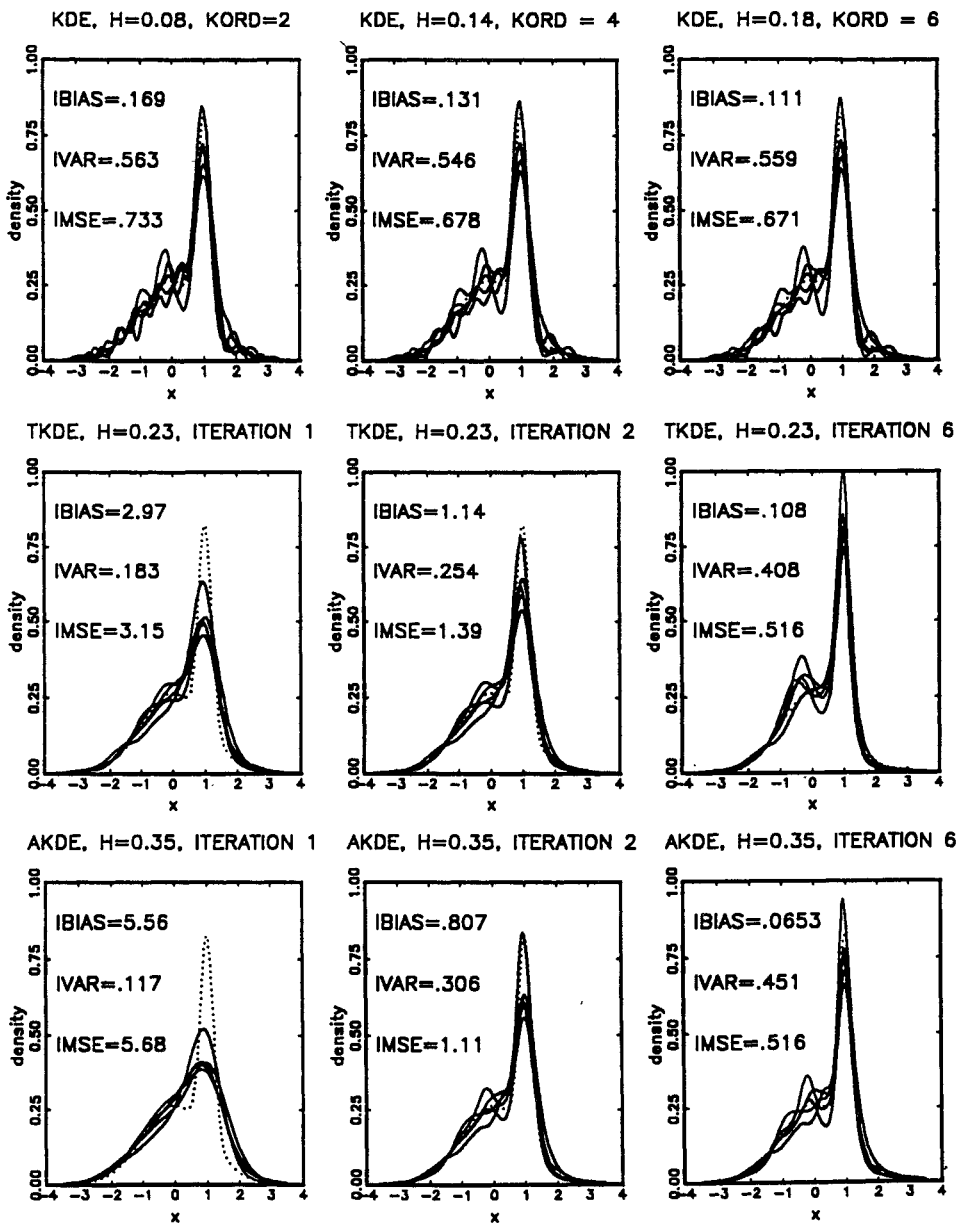

FIG. 1. *KDE, TKDE and AKDE. All estimates are calculated at the same five samples of* $n = 400$ *observations from the normal mixture,* $\frac{2}{3}N(0,1) + \frac{1}{3}N(1,(0.2)^2)$: *fixed bandwidth factor H; (dotted curve) true density; (solid curves) estimates. IBIAS, IVAR and IMSE based on* 500 *Monte Carlo repetitions. KORD is the kernel order. The KDE with KORD* ≥ 4 *uses Gajek's modification.*

Gajek's (1986) algorithm. The values of H used in each plot of the top row minimize the estimated IMSE for the estimator in that plot.

We used the implementation of Abramson's estimate proposed by Silverman [(1986), page 101]—that estimator is

$$\hat{f}_A(x) = (nh\lambda_i)^{-1} \sum_{i=1}^{n} K\left\{\frac{x - X_i}{h\lambda_i}\right\}, \quad \text{where } \lambda_i = \left\{\frac{\tilde{f}(X_i)}{gm}\right\}^{-1/2},$$

$\tilde{f}$ is a pilot estimate of f_X, and gm is the geometric mean of $f_X(X_1), \ldots, f_X(X_n)$. The first iteration was the ordinary KDE. In subsequent iterations, the pilot estimate was the estimate from the previous iteration. $H = 0.35$ minimizes IMSE for iteration 6. The overall impression from Figure 1 is that the TKDE and ADKE perform quite similarly for this density and sample size. The TKDE is somewhat more successful than the AKDE at showing the bimodal structure—this is true not only at the five estimates illustrated here but at larger numbers of samples that we have examined. The higher-order kernels do reduce bias compared to the KDE, but higher-order kernel estimators have larger IMSE's and have a far "wigglier" appearance than the TKDE's and AKDE's. We feel that the IMSE of the higher-order estimates is acceptable but their visual appearance is not. Of course, the higher-order kernel estimates can be made as smooth as the TKDE and AKDE's by increasing H, but that leads to a notable underestimation of the peaks and a sizable increase in IMSE. The original higher-order kernel estimates were never far below zero, and Gajek's algorithm had very little effect on the appearance of the higher-order kernel estimator and decreased their integrated bias and variance by less than 0.1%.

In an interesting paper, Hall and Marron (1988) show that the second iterate of the AKDE attains the rate of $O_P(n^{-8/9})$ but further iteration does not improve this rate. We found that the IMSE of the AKDE decreases at least till the third or fourth iterate, but does stabilize after that. $H = 0.33$ minimizes the estimated IMSE of the fourth iterate.

Figure 2 compares the kernel estimators for $n = 400$ standard normal observations. Again, the comparison is among TKDE, AKDE and KDE's of order 2, 4 and 6. In each plot, estimates of IBIAS, IVAR and IMSE based on 500 Monte Carlo repetitions are shown, and H in each plot was chosen to minimize the IMSE estimates. This is a case where the higher-order kernels perform somewhat better than the TKDE and AKDE, while the latter two estimators are rather similar.

In the two Monte Carlo experiments exhibited here and in a third experiment with lognormal data that we will not report, bias reduction (either by a higher-order kernel, the TKDE or the AKDE) does lead to reduction in IMSE compared to the KDE with a second-order kernel. None of the three methods of bias reduction dominates the other two. The AKDE and TKDE are noticeably better than higher-order kernel estimators at multimodal data, especially if the peaks are sharp. For lognormal data, none of the estimators

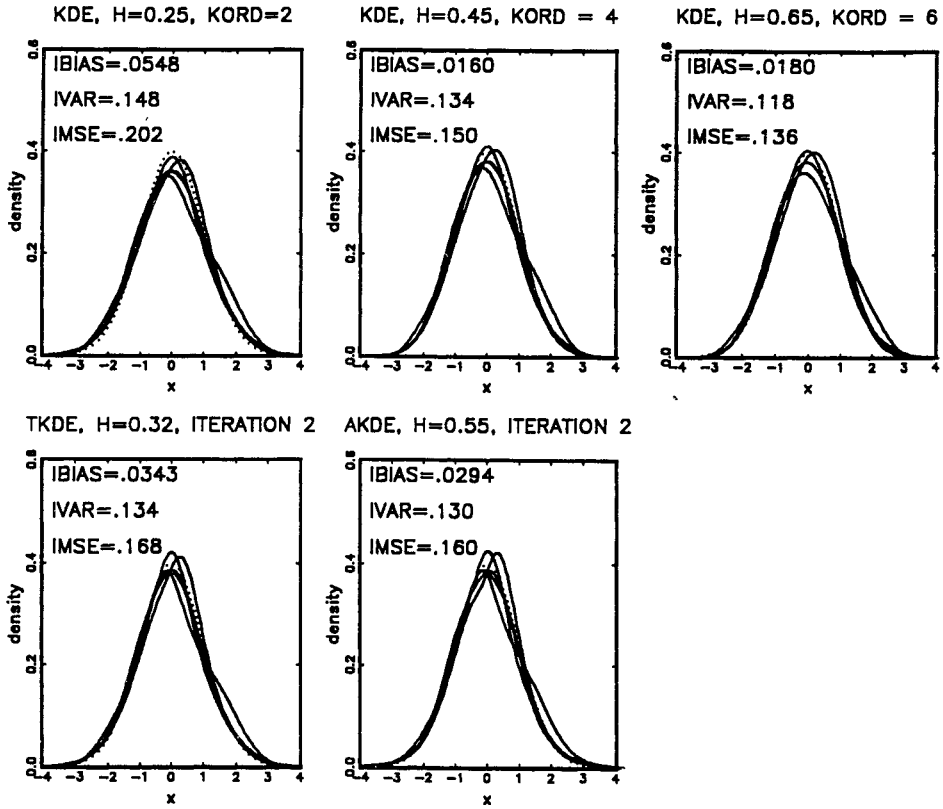

FIG. 2. *KDE, TKDE and AKDE. All estimates are calculated at the same five samples of* $n = 400$ *observations from the standard normal distribution: fixed bandwidth factor* H*; (dotted curve) true density; (solid curves) estimates. IBIAS, IVAR and IMSE based on* 500 *Monte Carlo repetitions. The KDE with KORD* ≥ 4 *uses Gajek's modification.*

discussed here compares well with the parametric TKDE of Wand, Marron and Ruppert (1991).

The TKDE and AKDE were computed for the Buffalo snowfall data, winters 1910/11 to 1972/73 [Silverman (1986)]. The first (dotted curve) and sixth (solid curve) iterates are plotted in Figure 3. We also included the KDE with KORD = 2 (dotted curve) and KORD = 8 (solid curve). The upper plots are the TKDE at $H = 0.20$ and 0.30, the middle plots are the AKDE at $H = 0.40$ and 0.50 and the bottom plots are of the KDE at $H = 0.35$ and 0.45. Here and in the following "incomes" example, the values of H were chosen subjectively to make the smoothness of the TKDE, AKDE, and KDE comparable. After discretization, $IQR \doteq 34.4$ for the untransformed data, so, for example, $H = 0.20$ corresponds to $h = 6.88$. An interesting but unresolved question is whether these data come from a unimodal or trimodal density. The rightmost plots in Figure 3 suggest

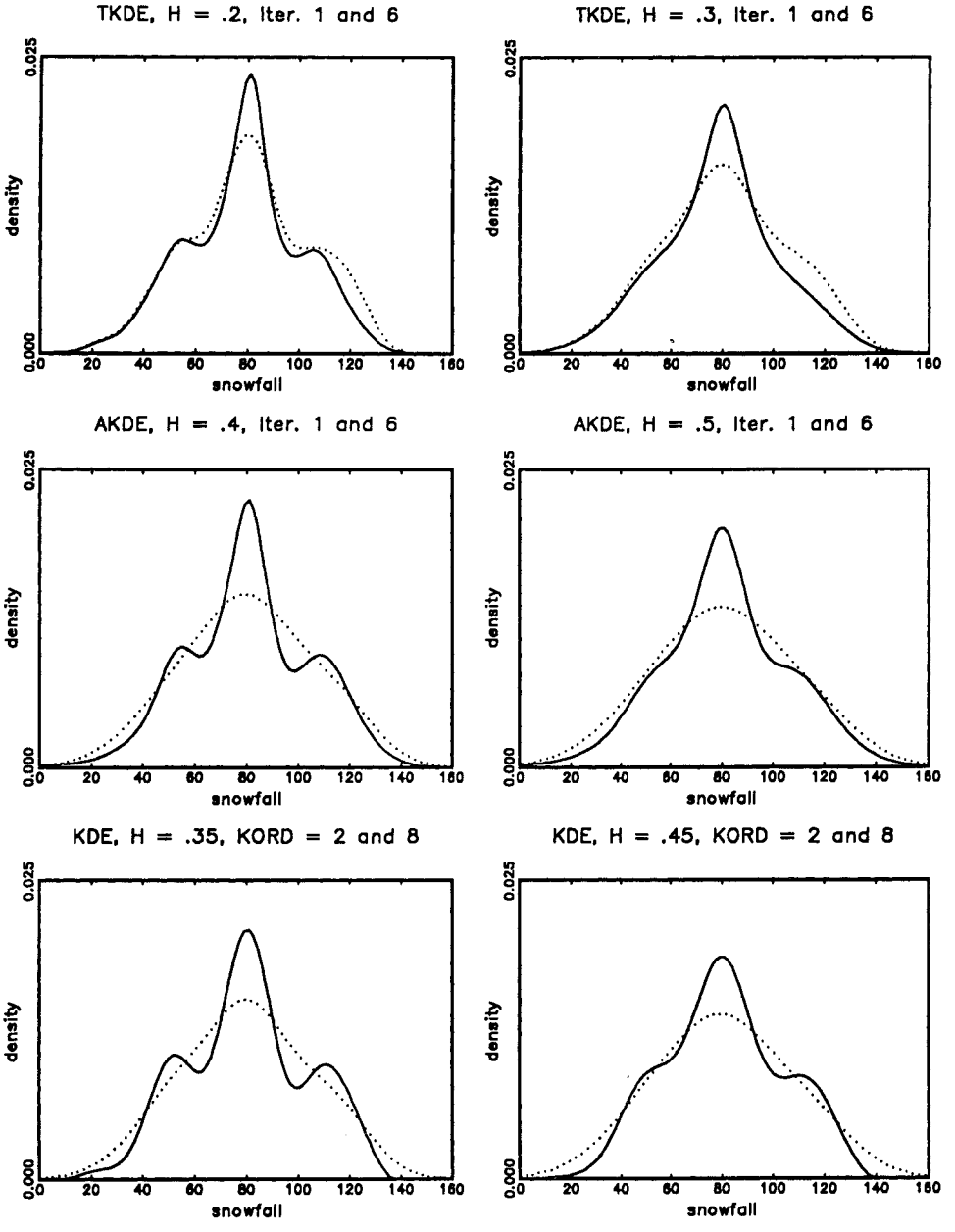

FIG. 3. *Estimates for the Buffalo snowfall data: (dotted curve) first iterate of TKDE or AKDE or second-order kernel KDE; (solid curve) sixth iterate of TKDE or AKDE or eighth-order kernel KDE with Gajek's modification.*

a compromise position, that the density is unimodal but with a sharp spike at the center and "humps" in the shoulders. The TKDE plots in Figure 3 are similar to those of the parametric TKDE in Ruppert and Wand (1992). The three estimators are different, and having all three does give us more insight into the shape of this distribution than having only one.

In Figure 4, we plot the TKDE, AKDE and KDE for the "incomes data" discussed in Wand, Marron and Ruppert (1991). The data are the incomes of slightly more than 7000 British subjects for the year 1975 and have been standardized to have a mean of 1. The IQR is 0.74 for the untransformed data, so $H = 0.11$ corresponds to $h = 0.0814$. As in Figure 3, the dotted curve is the ordinary KDE and the solid curve is the sixth iterate of TKDE or AKDE or KORD = 8. The TKDE indicates two well-defined peaks. The AKDE and KDE with KORD = 8 has a somewhat shorter left peak and a rougher right peak. Compared to the KDE with second-order kernel, the TKDE and AKDE—as well as the parametric TKDE of Wand, Marron and Ruppert (1991)—can show the bimodal structure in the data without having extensive random variation in the right tail.

4. Further comments. The bandwidths in Theorem 2.1–2.3 are deterministic sequences. It would be possible to extend these results to data-based bandwidths. In fact, the empirical process theory used in Section 5 would be ideal for doing this—see Pollard's (1984) study of random bandwidths for the KDE. To emphasize the main ideas, we have avoided the additional technicalities of random bandwidths.

The TKDE is similar to the proposals of Abramson (1984) for transforming locally to a density that is linear. Abramson does not consider rates of convergence or the possibility of iteration. He uses a generalized nearest neighbor estimator, to use the terminology of Silverman (1986). If we transformed to a linear density function, rather than a constant density function, then bias reduction would still be achieved in the interior, but boundary bias would necessitate a boundary kernel not everywhere nonnegative [Rice (1984)].

For extremely skewed or heavy-tailed densities, the poor behavior of the preliminary KDE may seriously degrade the performance of the TKDE. In such situations we recommend using an initial parametric transformation as discussed in Wand, Marron and Ruppert (1991) and in Ruppert and Wand (1992). Similarly, if the density has compact support, then the boundary bias of the initial KDE will persist in later iterations. This problem can be avoided by using a parametric TKDE developed by Marron and Ruppert (1995) to get the initial estimator. When tested on compactly supported densities with sharply peaked modes, the combination of initial boundary transformation and then the TKDE developed in this paper has proved very successful. An advantage of the nonparametric TKDE presented here is that it fits into a family of parametric and nonparametric transformations that can estimate a wide variety of estimators. It is doubtful if any single estimator such as a higher-order KDE will have nearly the flexibility of this family.

Why are the AKDE and TKDE more similar to each other than either is

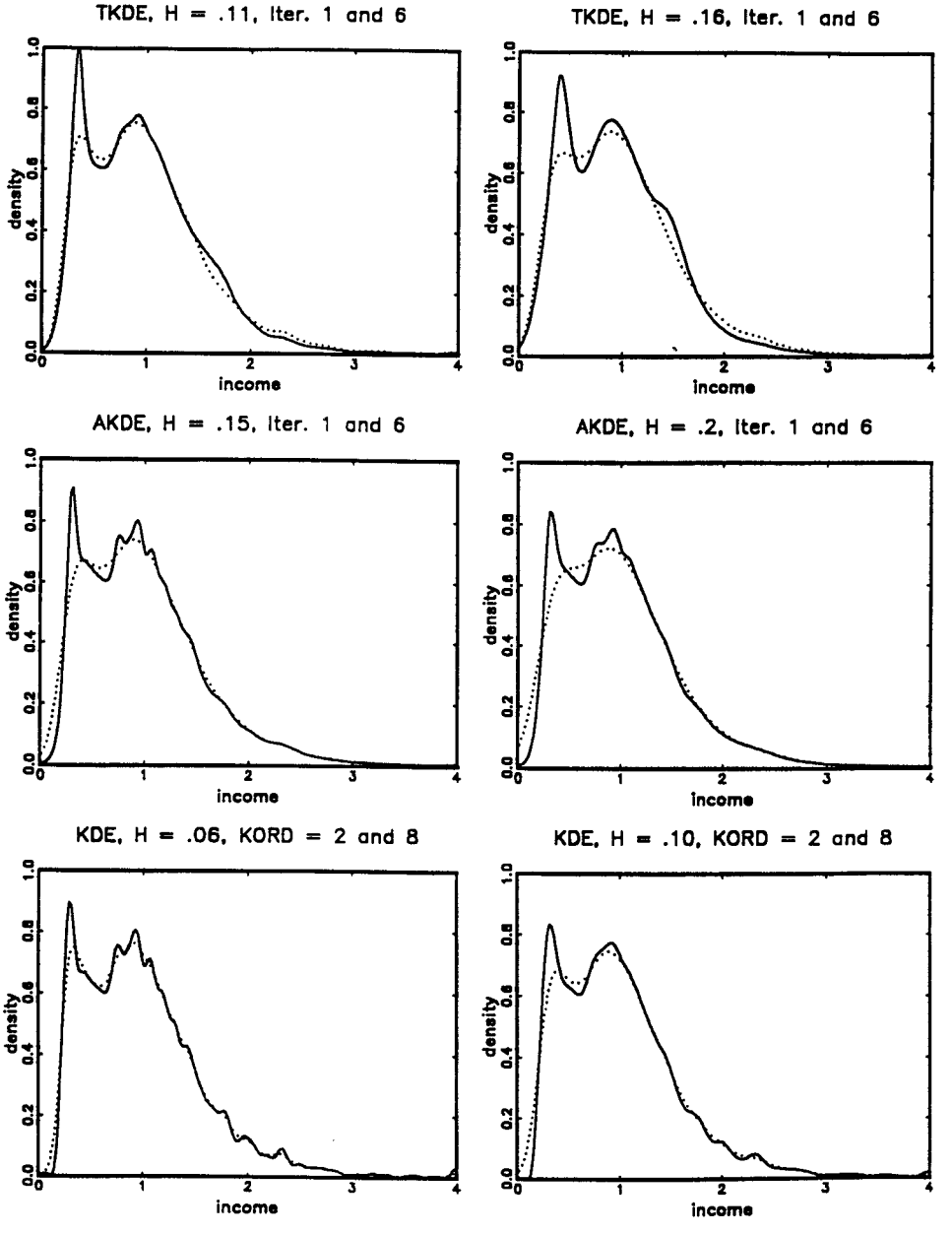

FIG. 4. *Estimates for the incomes data: (dotted curve) first iterate of TKDE or AKDE or second-order kernel KDE; (solid curve) sixth iterate of TKDE or AKDE or eighth-order kernel KDE with Gajek's modification.*

to the higher kernel estimates? The KDE with kernel of any order applies a constant amount of smoothing—as measured by the bandwidth—across the range of x. The AKDE smooths more where f_X is small than where the density is large. If as a heuristic we use the approximation $x \doteq X_i$, then the effective bandwidth of the AKDE at x is $h(f_X(x))^{-1/2}$. By the heuristic argument in Wand, Marron and Ruppert (1991), the TKDE effectively has a local bandwidth of size $h(f_X(x))^{-1}$. Of course, these heuristics should not be pushed too far. Applied to bias, they would erroneously suggest that the AKDE and TKDE *cannot* reduce the order of the bias. However, they do give some insight into the similarity between the TKDE and AKDE.

It is unlikely that a uniformly best method of density estimation exists. What is needed in practice is a variety of estimation methods and an understanding of their differences. The various TKDE's provide a very flexible family of estimators. The nonparametric TKDE presented here seems ideal for densities with sharp peaks. The parametric TKDE's in Wand, Marron and Ruppert (1991), Ruppert and Wand (1992), and Marron and Ruppert (1995) are designed for positive right-skewed densities, roughly symmetric heavy-tailed densities and compactly supported densities, respectively. One can of course use a parametric TKDE as the initial estimator for the nonparametric TKDE, which increases the flexibility of the TKDE approach.

Müller and Zhou (1991) discuss another approach to kernel estimation, local bandwidths, that we have not considered here because this methodology does not improve the rate of convergence of a kernel estimator. Nonetheless, local bandwidths could lead to improvements in the IMSE and the ability to estimate peaks. It would be interesting to compare local bandwidth kernel estimation with the TKDE and AKDE, but that comparison is beyond the scope of this paper.

5. Proofs. The main purpose of this section is to prove Theorems 2.1–2.3. First we need some introductory lemmas. Recall that (A1)–(A4) are assumed throughout. We will use the following notation: if $\{U_n\}$ and $\{W_n\}$ are sequences of random variables, then we write $U_n = O_P(W_n)$ if, for each $\varepsilon > 0$, there exist M and N depending on ε such that $\mathrm{P}\{|U_n| \le M|W_n|\} > 1 - \varepsilon$, for all $n \ge N$.

LEMMA 5.1. *Fix $D > 0$ and suppose that $h = h_n$ is a sequence such that $h > 0$, $h \to 0$ and $nh \to \infty$. For $l = 1, \ldots, 2t$ define, for $s \in [0,1]$,*

$$X_n^l(s) = \sqrt{nh^{(2l+1)}}\left\{\widehat{f}_1^{(l)}\left(x + Dh\left(s - \tfrac{1}{2}\right); h\right) - \mathrm{E}\widehat{f}_1^{(l)}\left(x + Dh\left(s - \tfrac{1}{2}\right); h\right)\right\},$$

and let X^l be a stationary, mean-zero Gaussian process on $[0,1]$ with

$$\mathrm{Cov}\left(X^l(s), X^l(s')\right) = f_X(x)\int K^{(l)}(u + D(s - s'))K^{(l)}(u)\,du.$$

Then X_n^l converges weakly to X^l in $C[0,1]$, as $n \to \infty$, and

$$\sup_{s\in[0,1]} \left|X_n^{(l)}(s)\right| = O_P(1) \quad \text{as } n \to \infty. \tag{5.1}$$

PROOF. (5.1) will follow from the weak convergence since

$$\sup_{s\in[0,1]} \left|X^{(l)}(s)\right| < \infty \quad \text{with probability 1,}$$

by standard results on the continuity of stationary Gaussian processes [Leadbetter, Lindgren and Rootzén (1983)]. Convergence of the finite-dimensional distributions is a standard calculation. To show that X_n^l is tight, set

$$Y_{n,i}(s) = \frac{1}{\sqrt{nh}} K^{(l)}\left(\frac{x - X_i}{h} - D\left(s + \frac{1}{2}\right)\right).$$

Then

$$\begin{aligned}
&\mathbb{E}\left(X_n^l(s) - X_n^l(s')\right)^2 \\
&\quad \le n\mathbb{E}\left(Y_{n,i}(s) - Y_{n,i}(s')\right)^2 \\
&\quad = h^{-1}\int \left\{K^{(l)}\left(\frac{x-y}{h} + D\left(s - \frac{1}{2}\right)\right) - K^{(l)}\left(\frac{x-y}{h} + D\left(s' - \frac{1}{2}\right)\right)\right\}^2 f(y)\,du \\
&\quad = \int \left\{K^{(l)}\left(u + D(s-s')\right) - K^{(l)}(u)\right\}^2 f\left(x - hu - Dh\left(s' - \frac{1}{2}\right)\right) du \\
&\quad = \int \left\{K^{(l+1)}\left(u + \theta(s-s')\right) D(s-s')\right\}^2 f\left(x - hu - Dh\left(s' - \frac{1}{2}\right)\right) du \\
&\quad = D^2(s-s')^2 \int \left\{K^{(l+1)}\left(u + \theta(s-s')\right)\right\}^2 f\left(x - hu - Dh\left(s' - \frac{1}{2}\right)\right) du \\
&\quad \le M(s-s')^2,
\end{aligned}$$

where θ depends on u and is bounded by D. From this it is easy to see that

$$\mathbb{E}\left(X_n^l(s) - X_n^l(s')\right)^2 \le M(s-s')^2$$

uniformly in n, s, s'. Tightness follows from Theorem 12.3 and (12.51) of Billingsley (1968) with $\gamma = 2$ and $\alpha = 2$. □

LEMMA 5.2. *Suppose that* $h = cn^{-1/(4t+1)}$ *for some* $c > 0$, $\Delta = h(\log n)^{1/2}$ *and* $C > 0$. *Define*

$$\begin{aligned}
\mathcal{G} = \Big\{ & g\text{: } g \text{ is a } (2t+1)\text{-th degree polynomial, } g(x) = 0, \\
& \left|g^{(m)}(x) - (G \circ F_X)^{(m)}(x)\right| \le \Delta \quad \text{for } m = 1, \ldots, 2t \\
& \text{and } \left|g^{(2t+1)}(x) - (G \circ F_X)^{(2t+1)}(x)\right| \le C \Big\}.
\end{aligned} \tag{5.2}$$

[*Recall from* (A1) *that* x *is in the interior of* $\operatorname{supp} X$. $\mathcal{G}$ *depends on* x *but this will not be made explicit in the notation.*] *Let* $\widehat{f}_X(\cdot; g, h)$ *be defined by* (1.2). *Then, for any fixed* $\eta > 0$, $M > 0$ *and nonnegative integer* $l \le 2t$,

$$\sup_{g\in\mathcal{G}} \sup_{|x'-x|\le Mh} \left|\widehat{f}_Y^{(l)}\left(g(x'); g, h\right) - \mathbb{E}\widehat{f}_Y^{(l)}\left(g(x'); g, h\right)\right| = O_P\left(h^{2t-l}\right). \tag{5.3}$$

PROOF. Since $\hat{f}_Y(g(x');g,h)$ is unchanged if we replace $g(\cdot)$ by $g(\cdot)-g(x)$, there is no loss in generality when we assume that $g(x)=0$ for all g in $\mathcal{G}$.

Fix $M>0$ and let $l\in\{0,1,\ldots,2t\}$. Since $G\circ F_X$ has a positive derivative at x and $\Delta\to 0$, there exist $N>0$ and $M_1>0$ such that, for all $n\geq N$, for all x' and x'' in $[x-Mh, x+Mh]$ and for all g in $\mathcal{G}$, we have

$$|g(x')-g(x'')|\geq M_1^{-1}|x'-x''|. \tag{5.4}$$

Since supp $K=[-1,1]$, by (5.4)

$$\begin{aligned}&\hat{f}_Y^{(l)}(g(x');g,h)\\ &\quad=\frac{1}{nh^{1+l}}\sum_{i=1}^{n}K^{(l)}\{h^{-1}[g(X_i)-g(x')]\}I\{|X_i-x'|\leq M_1h\},\end{aligned} \tag{5.5}$$

for all $n\geq N$, g in $\mathcal{G}$ and x' such that $|x-x'|\leq Mh$. Define

$$K_h\{\,\cdot\,;x',l\}=h^{-1}K^{(l)}\{h^{-1}(\cdot-x')\}$$

and

$$k(\cdot;x',g,l)=K^{(l)}\left(\frac{g(\cdot)-g(x')}{h}\right)-K^{(l)}\left(\frac{G\circ F_X(\cdot)-G\circ F_X(x')}{h}\right).$$

(Note that the dependence on n and h is suppressed here as elsewhere.) Consider the class of functions

$$\mathcal{H}^{(l)}=\{k\{\,\cdot\,;x',g,l\}\colon g\in\mathcal{G}\quad\text{and}\quad x'\in[x-Mh,x+Mh]\}.$$

We will show that the $L_1(F_X)$ covering numbers [Pollard (1984), page 25] of $\mathcal{H}^{(l)}$ satisfy

$$N_1(\varepsilon,\mathcal{H}^{(l)})\leq M_2\varepsilon^{-(2t+2)}, \tag{5.6}$$

for some M_2 and all $\varepsilon>0$. M_1, M_2 and the the constants $M_3,\ldots,M_{20}$ introduced later in the proofs are independent of n, l, h and ε. If $g\in\mathcal{G}$, then we can write

$$g(x')=\sum_{m=1}^{2t+1}\alpha_m/m!(x'-x)^m$$

where $|\alpha_m-(G\circ F_X)^{(m)}(x)|\leq\Delta$ for $m=1,\ldots,2t$, and $|\alpha_m-(G\circ F_X)^{(m)}(x)|\leq C$ for $m=2t+1$. For each $\varepsilon>0$, consider $\mathcal{G}_\varepsilon$, the set of all g^* in $\mathcal{G}$ whose coefficients α_m^* are of the form

$$\alpha_m^*=(G\circ F_X)^{(m)}(x)+j\Delta_\varepsilon\quad\text{for } j\text{ an integer and } |j|\leq[\varepsilon^{-1}]+1\text{ for } m=1,\ldots,2t,$$

and

$$\alpha_{2t+1}^*=(G\circ F_X)^{(2t+1)}(x)+jC\varepsilon,\quad\text{for } j\text{ an integer and } |j|\leq[\varepsilon^{-1}]+1.$$

There exists $M_3 > 0$ such that, for every $g \in \mathcal{G}$, there exists $g^* \in \mathcal{G}_\varepsilon$ such that

$$\sup_{|x'-x| \le (M+3M_1)h} |g(x') - g^*(x')| \le M_3\varepsilon. \tag{5.7}$$

There exists M_4 such that, for all $\varepsilon > 0$ and for all n,

$$\operatorname{card}[\mathcal{G}_\varepsilon] \le M_4\varepsilon^{-(2t+1)}. \tag{5.8}$$

Suppose g and g^* satisfy (5.7), $|x'-x| \le Mh$ and $|x'-x''| \le \varepsilon$. Then there exist M_5, M_6, M_7 such that

$$\begin{aligned}
&\int \left| K^{(l)}\left(\frac{g(u)-g(x')}{h}\right) - K^{(l)}\left(\frac{g^*(u)-g^*(x'')}{h}\right)\right| du \\
&\quad \le M_5 I_{\{2M_1h \le |x'-x''| \le \varepsilon\}} \int \left(I_{\{|u-x'| \le M_1h\}} + I_{\{|u-x''| \le M_1h\}}\right) du \\
&\quad + \frac{M_6}{h} I_{\{|x'-x''| \le 2M_1h\}} \int_{-3M_1h}^{3M_1h} \left(|g(x'+v) - g^*(x'+v)|\right. \\
&\qquad\qquad \left. + |g(x') - g^*(x')| + |g^*(x') - g^*(x'')|\right) dv \\
&\quad \le M_7\varepsilon.
\end{aligned} \tag{5.9}$$

Applying (5.9) twice (once with g and g^*, as it stands, and once with $g = g^* = G \circ F_X$), there is M_8 such that

$$\mathbb{E}\,|k(X;x',g,l) - k(X;x'',g^*,l)| \le M_8\varepsilon.$$

If $\mathcal{Y}(\varepsilon) = [x - Mh, x + Mh] \cap \{(h\varepsilon j) : j \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of integers, then, for some $M_9 > 0$,

$$\operatorname{card}[\mathcal{Y}(\varepsilon)] \le M_9\varepsilon^{-1} \tag{5.10}$$

for all $\varepsilon > 0$ and for all n. By (5.7) and (5.9), we can cover $\mathcal{H}^{(l)}$ by $L_1(F_X)$ balls with centers in the set

$$\mathcal{H}_\varepsilon^{(l)} = \{k(\cdot;x',g,l) : g \in \mathcal{G}\varepsilon \text{ and } x' \in \mathcal{Y}(\varepsilon)\}$$

and $L_1(F_X)$ radii equal to $M_{10}\varepsilon$, for some M_{10}. By (5.8) and (5.10), card $\mathcal{H}_\varepsilon^{(l)} \le M_{11}\varepsilon^{-(2t+2)}$, for some M_{11}. Therefore, (5.6) holds.

By (5.2) for $g \in \mathcal{G}$, $|x'-x''| \le M_1h$, $|x'-x| \le (M+M_1)h$ and $|x''-x| \le (M+M_1)h$, there exist M_{12}, M_{13} and M_{14} so that

$$\begin{aligned}
|g(x'') - g(x') - G \circ F_X(x'') + G \circ F_X(x')| &\le \int_{x'}^{x''} |g'(u) - (G \circ F_X)'(u)|\, du \\
&\le 2M_1h\left(M_{12}\Delta + M_{13}h^{2t}\right) \le M_{14}h\Delta.
\end{aligned}$$

By (5.5) there are M_{15} and M_{16} such that

$$\begin{aligned}(5.11)\qquad &\mathbb{E}k^2(X_1;x',g,l)\\ &\le \frac{M_{15}}{h^2}\mathbb{E}\Big\{I_{\{|X_1-x'|\le M_1h\}}\big(g(X_1)-g(x')-G\circ F_X(X_1)\\ &\qquad +G\circ F_X(x')\big)^2\Big\}\le M_{16}h\Delta^2.\end{aligned}$$

Since card $\mathcal{H}_\varepsilon^{(l)} \le M_{11}\varepsilon^{-(2t+2)}$ and by (5.11) one can apply Theorem 37 of Pollard [(1984), page 34] with $\delta_n = \sqrt{M_{16}}h^{1/2}\Delta$ and, with $\eta > 0$ fixed, $\alpha_n = \big((\log n)^{(1+\eta)}/(n\delta_n^2)\big)^{1/2}$ so that

$$\delta_n^2\alpha_n = \delta_n\left((\log n)^{1+\eta}/n\right)^{1/2} = O\left(h^{2t+2}(\log n)^{1+\eta/2}\right).$$

To simplify notation, let

$$\widehat{f}_{Y,*}^{(l)}(y;g,h) = \widehat{f}_Y^{(l)}(y;g,h) - \mathbb{E}\widehat{f}_Y^{(l)}(y;g,h).$$

One gets from the application of Pollard's theorem that

$$\begin{aligned}&\sup_{|x'-x|\le Mh}\sup_{g\in\mathcal{G}}\left|\widehat{f}_{Y,*}^{(l)}(g(x');g,h) - \widehat{f}_{Y,*}^{(l)}\left(G(F_X(x'));\,G\circ F_X,h\right)\right|\\ (5.12)\qquad &= \sup_{|x'-x|\le Mh}\sup_{g\in\mathcal{G}}\left|h^{-l-1}\sum_{i=1}^n\left\{k(X_i;x',g,l)-\mathbb{E}k(X_i;x',g,l)\right\}\right|\\ &= o\left(h^{-l-1}\delta_n^2\alpha_n\right) = o\left(h^{2t-l}\right),\end{aligned}$$

almost surely, as $n\to\infty$. By Lemma 5.1

$$\begin{aligned}(5.13)\qquad &\sup_{|x'-x|\le Mh}\left|\widehat{f}_{Y,*}^{(l)}\left(G\left(F_X(x')\right);G\circ F_X,h\right)\right|\\ &= O_P\left(h^{-1/2-l}n^{-1/2}\right) = O_P\left(h^{2t-l}\right).\end{aligned}$$

Hence, by (5.12) and (5.13),

$$\sup_{|x'-x|\le Mh}\sup_{g\in\mathcal{G}}\left|\widehat{f}_{Y,*}^{(l)}(g(x');g,h)\right| = O_P\left(h^{2t-l}\right),$$

which proves (5.3). □

LEMMA 5.3. *Suppose that $t \ge 2$. Let $c_j' = 2c_j$, let $A_t = M$ and, for $j = t-1,\ldots,1$, let $A_j = A_{j+1} + c_{j+1}'M_1$, where M_1 satisfies* (5.4). *Suppose* (2.1) *holds and note that $h_j \le c_j'h_0$ for n large, $j = 1,\ldots,t$. Suppose, for some $j < t$, that* (2.2) *holds with $M = A_j$. Then, for every $l \le 2t$,*

$$\sup_{|x'-x|\le A_{j+1}h_0}\left|\widehat{f}_Y^{(l)}\left(\widehat{F}_j(x');\widehat{F}_j,h_{j+1}\right) - f_Y^{(l)}\left(\widehat{F}_j(x');\widehat{F}_j\right)\right| = O_P\left(h_{j+1}^{\min(2(j+1),\,2t-l)}\right).$$

PROOF. Let $\nu > 0$. Define $\widetilde{F}_j(x') = \sum_{m=1}^{2t+1} \widehat{F}_j^{(m)}(x)(x'-x)^m/m!$. Then, by (2.2) with $M = A_j$,

$$\sup_{|x'-x| \le A_j h_0} \left| \widetilde{F}_j^{(l)}(x') - F_X^{(l)}(x') \right| = O_P\left(h_0^{\min(2j,\, 2t-l+1)}\right), \tag{5.14}$$

for $1 \le l \le 2t+1$. [The summation in the definition of $\widetilde{F}_j$ starts at 1, not 0, so that $\widetilde{F}_j$ is in $\mathcal{G}$. This does not affect the validity of (5.14) for $l \ge 1$.] Thus with G equal to the identity function there are an N and a C such that for all $n \ge N$ the probability that $\widetilde{F}_j \in \mathcal{G}$ is at least $1-\nu$. To establish O_P convergence rates, it suffices to limit our discussion to the event $\widetilde{F}_j \in \mathcal{G}$ since ν may be arbitrarily small.

By Lemma 5.2, therefore,

$$\begin{aligned} &\sup_{|x'-x| \le A_j h_0} \left| \widehat{f}_Y^{(l)}\left(g(x');g,h_{j+1}\right) - \mathbb{E}\widehat{f}_Y^{(l)}\left(g(x');g,h_{j+1}\right) \right|_{g=\widehat{F}_j} \\ &\qquad = O_P\left(h_{j+1}^{2t-1}\right), \end{aligned} \tag{5.15}$$

for all $l \le 2t$.

There are $Q_{m,l}(x';g)$, polynomials in $g', \ldots, g^{(l+1-m)}$, such that

$$f_Y^{(l)}\left(g(x');g\right) = \left(g'(x')\right)^{-l-1} \sum_{m=0}^{l} Q_{m,l}(x';g) f_X^{(m)}(x').$$

Therefore, for $0 \le l \le 2t$, there is M_{17} such that, for $g \in \mathcal{G}$,

$$\begin{aligned} &\sup_{|x'-x| \le A_{j+1}h_0} \left| f_Y^{(l)}\left(g(x');g\right) - f_Y^{(l)}\left(F_X(x');F_X\right) \right| \\ &\quad \le \sup_{|x'-x| \le A_{j+1}h_0} \sum_{m=0}^{l} \left| f_X^{(m)}(x') \right| \Big| \left(g'(x')\right)^{-l-1} Q_{m,l}(x';g) \\ &\qquad\qquad - \left(f_X(x')\right)^{-l-1} Q_{m,l}\left(x';F_X\right) \Big| \\ &\quad \le M_{17} \sup_{|x'-x| \le A_{j+1}h_0} \sum_{m=1}^{l+1} \left| g^{(m)}(x') - F_X^{(m)}(x') \right|. \end{aligned} \tag{5.16}$$

In particular, using (5.14) in (5.16),

$$\begin{aligned} &\sup_{|x'-x| \le A_{j+1}h_0} \left| \widehat{f}_Y^{(l)}\left(\widehat{F}_j(x');\widetilde{F}_j,h_{j+1}\right) - f_Y^{(l)}\left(F_X(x');F_X\right) \right| \\ &\qquad = O_P\left(h_{j+1}^{\min(2j,\, 2t-l)}\right). \end{aligned} \tag{5.17}$$

For $l \le 2t-2, g \in \mathcal{G}$ and $|x-x'| \le A_{j+1}h_0$,

$$\begin{aligned} &\mathbb{E}\widehat{f}_Y^{(l)}\left(g(x');g,h_{j+1}\right) - f_Y^{(l)}\left(g(x');g\right) \\ &\qquad = \frac{h_{j+1}^2}{2} \int w^2 K(w) f_Y^{(l+2)}\left(g(x') + h_{j+1}\theta_2 w;g\right)\, dw, \end{aligned} \tag{5.18}$$

where $|\theta_2| \le 1$ and depends on w. Let M_1 be as in (5.4). For $|x' - x| \le A_{j+1}h_0$, the mean value theorem gives $g(x') + h_{j+1}\theta_2 w = g(x'')$, where

$$|x'' - x| \le (A_{j+1} + c'_{j+1}M_1)h_0 = A_j h_0.$$

For $l > 0$, of course, $f_Y^{(l)}(F_X(x'); F_X) = 0$. Applying (5.16) to $f_Y^{(l+2)}$ in (5.18), we have, for some M_{18},

$$\begin{aligned} &\left|\mathbb{E}\hat{f}_Y^{(l)}\left(g(x'); g, h_{j+1}\right) - f_Y^{(l)}\left(g(x'); g\right)\right|_{g=\tilde{F}_j} \\ &\quad \le M_{18}h_{j+1}^2 \sup_{|x'-x| \le A_j h_0} \sum_{m=1}^{l+3} \left|\tilde{F}_j^{(m)}(x') - F_X^{(m)}(x')\right| \\ &\quad = O_P\left(h_{j+1}^2 h_{j+1}^{\min(2j,\, 2t-l-2)}\right) = O_P\left(h_{j+1}^{\min(2(j+1),\, 2t-l)}\right). \end{aligned} \tag{5.19}$$

Similar mean value arguments establish (5.19) for $l = 2t - 1, 2t$ with rates $O_P(h_{j+1})$ and $O_P(1)$, respectively. By (5.15) and (5.19),

$$\begin{aligned} &\sup_{|x'-x| \le A_{j+1}h_0} \left|\hat{f}_Y^{(l)}\left(\tilde{F}_j(x'); \tilde{F}_j, h_{j+1}\right) - f_Y^{(l)}\left(\tilde{F}_j(x'); \tilde{F}_j\right)\right| \\ &\quad = O_P\left(h_{j+1}^{\min(2(j+1),\, 2t-l)}\right). \end{aligned} \tag{5.20}$$

Since

$$\sup_{|x'-x| \le A_j h_0} \left|\hat{F}_j(x') - \tilde{F}_j(x')\right| \le 2 \sup_{|x'-x| \le A_j h_0} \left|\hat{F}_j^{(2t+1)}(x')\right| (A_j h_0)^{2t+1} = O_P\left(h_{j+1}^{2t+1}\right),$$

then, for some M_{19},

$$\begin{aligned} &\sup_{|x'-x| \le A_{j+1}h_0} \left|\hat{f}_Y^{(l)}(\hat{F}_j(x'); \hat{F}_j, h_{j+1}) - \tilde{f}_Y^{(l)}(\tilde{F}_j(x'); \tilde{F}_j, h_{j+1})\right| \\ &\quad \le \sup_{|x'-x| \le A_j h_0} \frac{1}{nh_{j+1}^{l+1}} \sum_{i=1}^{n} \left| K^{(l)}\left(\frac{\hat{F}_j(X_i) - \hat{F}_j(x')}{h_{j+1}}\right) \right. \\ &\qquad\qquad \left. - K^{(l)}\left(\frac{\tilde{F}_j(X_i) - \tilde{F}_j(x')}{h_{j+1}}\right) \right| I_{\{|X_i - x'| \le M_1 h_{j+1}\}} \\ &\quad \le \sup_{|x'-x| \le A_j h_0} \frac{M_{19}}{nh_{j+1}^{l+2}} \sum_{i=1}^{n} \left(\left|\hat{F}_j(X_i) - \tilde{F}_j(X_i)\right| \right. \\ &\qquad\qquad \left. + \left|\hat{F}_j(x') - \tilde{F}_j(x')\right| \right) I_{\{|X_i - x'| \le M_1 h_{j+1}\}} \\ &\quad \le 2M_{19} \sup_{|x'-x| \le A_j h_0} \left|\hat{F}_j(x') - \tilde{F}_j(x')\right| \sup_{|x'-x| \le A_j h_0} \frac{1}{nh_{j+1}^{l+2}} \sum_{i=1}^{n} I_{\{|X_i - x'| \le M_1 h_{j+1}\}} \\ &\quad = O_P\left(h_{j+1}^{2t-l}\right). \end{aligned} \tag{5.21}$$

Just as in (5.16), there is M_{20} such that, for $0 \le l \le 2t$,

$$\begin{aligned}
&\sup_{|x'-x| \le A_{j+1}h_0} \left| f_Y^{(l)}\left(\widetilde{F}_j(x);\widetilde{F}_j\right) - f_Y^{(l)}\left(\widehat{F}_j(x');\widehat{F}_j\right)\right| \\
&\quad \le \sup_{|x'-x| \le A_{j+1}h_0} \sum_{m=0}^{l} \left|f_X^{(m)}(x')\right| \left| \left(\widetilde{F}_j'(x')\right)^{-l-1} Q_{m,l}\left(x';\widetilde{F}_j\right) \right. \\
&\qquad\qquad \left. - \left(\widehat{F}_j'(x')\right)^{-l-1} Q_{m,l}\left(x';\widehat{F}_j\right)\right| \qquad (5.22) \\
&\quad \le M_{20} \sup_{|x'-x| \le A_{j+1}h_0} \sum_{m=1}^{l+1} \left|\widetilde{F}_j^{(m)}(x') - \widetilde{F}_j^{(m)}(x')\right| \\
&\quad = O_P\left(h_{j+1}^{2t-l}\right).
\end{aligned}$$

The result thus follows from (5.20)–(5.22). □

REMARK 5.1. What is interesting here (and surprising) is that the target uniform density is not approached at the new, faster rate, at least not until the next iteration; see (5.17). What *is* approached at the faster rate is the density from which a back-transformation using $\widehat{F}_X$ gives f_X; after all, this is the back-transformation used in the estimate.

PROOF OF THEOREM 2.1. Fix $M > 0$. Let A_j be as in Lemma 5.3. We will prove (2.2) with M replaced by $A_j > A_t = M$. The proof will be by induction on j. For the case $j = 1$, note that $\widehat{f}_1$ is an ordinary KDE and, by assumption (A1), f_X has at least $2t$ bounded derivatives in a neighborhood of x, so for $l \le 2t-2$,

$$\sup_{|x'-x| \le A_j h_0} \left|\widehat{f}_1^{(l)}(x') - f_X^{(l)}(x')\right| = O_P\left(h_1^2 + n^{-1/2}h_1^{-l-1/2}\right) = O_P\left(h_1^2\right), \qquad (5.23)$$

by Lemma 5.1 and the standard result that the asymptotic bias of $\widehat{f}_1^{(l)}(x')$ is proportional to $h_1^2 f_X^{l+2}(x')$ uniformly over $|x'-x| \le A_1 h_0$. Thus (2.2) holds for $l \le 2t-2$. For $l = 2t-1$ and $l = 2t$, $\left|\widehat{f}_1^{(l)}(x') - \mathbb{E}\widehat{f}_1^{(l)}(x')\right| = O_P(h_1) = O_P\left(h_1^{2t-l}\right)$ and $O_P(1)$, respectively. Therefore, (2.2) holds for $l = 2t-1$ and $2t$ as well.

Now suppose $t > 1$ and (2.2) holds for $l \in \{0, \ldots, 2t\}$ and some integer j less than t. We will prove that (2.2) holds for $l \in \{0, \ldots, 2t\}$ and $j+1$, which will complete the induction.

There exist $P_{m,l}(x', g)$, polynomials in $g', \ldots, g^{(l+1-m)}$, such that

$$\widehat{f}_X^{(l)}(x';g,h) = \sum_{m=0}^{l} \widehat{f}_Y^{(m)}\left(g(x');g,h\right) P_{m,l}(x',g)$$

and

$$f_X^{(l)}(x') = \sum_{m=0}^{l} f_Y^{(m)}\left(g(x');g\right) P_{m,l}(x',g).$$

Taking $g = \widehat{F}_j$, therefore, and using Lemma 5.3,

$$\sup_{|x'-x| \le A_{j+1}h_0} \left|\widehat{f}_X^{(l)}\left(x';\widehat{F}_j,h_{j+1}\right) - f_X^{(l)}(x')\right|$$

$$\le \sup_{|x'-x| \le A_{j+1}h_0} \sum_{m=0}^{l} \left|\widehat{f}_Y^{(m)}\left(\widehat{F}_j(x');\widehat{F}_j,h_{j+1}\right) - f_Y^{(m)}\left(\widehat{F}_j(x');\widehat{F}_j\right)\right| \left|P_{m,l}\left(x',\widehat{F}_j\right)\right|$$

$$= O_P\left(h_{j+1}^{\min(2(j+1),\,2t-l)}\right). \qquad \square$$

REMARK 5.2. If f_X has $(2t-1)$ bounded derivatives in a neighborhood of x, then (5.23) will hold for $l = 2t-2$ with h_1^2 on the RHS replaced by h_1. This is the first step toward proving the result in Remark 2.1, whose proof is otherwise quite similar to that of Theorem 2.1.

PROOF OF THEOREM 2.2. Let $\widehat{T} = G \circ \widehat{F}_1$ and

$$\widetilde{T}(x') = \sum_{m=1}^{3} \widehat{T}^{(m)}(x)(x'-x)^m/m!.$$

Using (5.12) in the proof of Lemma 5.2 with $t = 1$, one can show that, for any $M > 0$ and any $\eta > 0$,

$$\sup_{|x'-x| \le Mh_0} \left|\widehat{f}_{Y,*}\left(\widetilde{T}(x);\widetilde{T},h_2\right) - \widehat{f}_{Y,*}\left(T(x);T,h_2\right)\right| \tag{5.24}$$

$$= o_P\left(h_2^3(\log n)^{(1+\eta)/2}\right) = o_P\left(n^{-2/5}\right).$$

Using Lemma 5.1 and the fact that $n^{1/5}h_1 \to \infty$, one can show that, for any $M > 0$,

$$\sup_{|x'-x| \le Mh_0} \left|\widehat{f}_1^{(l)}(x') - f_X^{(l)}(x')\right| \tag{5.25}$$

$$= O_P\left(n^{-1/2}h_1^{-(1/2+l)}\right) = o_P(1), \quad \text{for } l = 0, 1 \text{ and } 2,$$

and

$$\sup_{|x'-x| \le Mh_0} \left|\widehat{f}_1^{(3)}(x') - f_X^{(3)}(x')\right| = O_P\left(n^{-1/2}h_1^{-7/2}\right). \tag{5.26}$$

Next, by calculations similar to those leading to (5.21) and using (5.26) and the fact that $h_2/h_1 \to 0$,

$$\sup_{|x'-x| \le Mh_0} \left|\widehat{f}_Y\left(\widehat{T}(x');\widehat{T},h_2\right) - \widehat{f}_Y\left(\widetilde{T}(x');\widetilde{T},h_2\right)\right|$$

$$= O_P\left(\sup_{|x'-x| \le (M+M_1)h_0} \left|\widehat{T}^{(3)}(x')\right| h_2^2\right) \tag{5.27}$$

$$= O_P\left(h_2^3 n^{-1/2}h_1^{-7/2}\right) = o_P\left(n^{-1/2}h_1^{-1/2}\right) = o_P\left(n^{-2/5}\right).$$

By (5.25) and a standard approximation of bias to order h_2^2, one can show that

$$\sup_{|x'-x|\le Mh_0}\left|\left(\mathbb{E}\hat{f}_Y\left(g(x');g,h_2\right)\right)\Big|_{g=\tilde{T}}-\mathbb{E}\hat{f}_Y\left(T(x');T,h_2\right)\right| = o_P\left(h_2^2\right)=o_P\left(n^{-2/5}\right). \tag{5.28}$$

By (5.24), (5.27) and (5.28),

$$\sup_{|x'-x|\le Mh_0}\left|\hat{f}_Y\left(\hat{T}(x');\hat{T},h_2\right)-\hat{f}_Y\left(T(x');T,h_2\right)\right| = o_P\left(n^{-2/5}\right). \tag{5.29}$$

Next,

$$\begin{aligned}&\sup_{|x'-x|\le Mh_0}\left|\hat{f}_Y\left(\hat{T}(x');\hat{T},h_2\right)\hat{T}'(x')-\hat{f}_Y\left(T(x');T,h_2\right)T'(x')\right|\\ &\le \sup_{|x'-x|\le Mh_0}\left|\hat{f}_Y\left(\hat{T}(x');\hat{T},h_2\right)\left(\hat{T}'(x')-T'(x')\right)\right|\\ &\quad+\sup_{|x'-x|\le Mh_0}\left|\left(\hat{f}_Y\left(\hat{T}(x');\hat{T},h_2\right)-\hat{f}_Y\left(T(x');T,h_2\right)\right)T'(x)\right|\\ &=o_P\left(n^{-2/5}\right).\end{aligned} \tag{5.30}$$

This proves (2.5). By standard results,

$$n^{(2/5)}\left(\hat{f}_Y(y;T,h_2)-f_Y(y;T)\right)\to_{\mathcal{D}} N\left(\frac{c_2^2}{2}\kappa_1 f_Y''(y;T),c_2^{-1}\kappa_2 f_Y(y;T)\right).$$

This and (2.5) imply (2.6) since $\hat{f}_X(x;T,h_2)=\hat{f}_Y(y;T,h_2)T'(x)$. □

PROOF OF THEOREM 2.3. This proof is similar to that of Theorem 2.1. In the latter, we already verified by induction that (2.2) holds for $j=1,\dots,t-1$. Using (2.7), we can show that

$$\left|\left(\mathbb{E}\hat{f}_Y\left(g(x);g,h_t\right)\right)\Big|_{g=\tilde{F}_{t-1}}-f_Y\left(\tilde{F}_{t-1}(x);\tilde{F}_{t-1}\right)\right| = o_P\left(h_0^{2t}\right)=o_P\left(n^{-2t/(4t+1)}\right). \tag{5.31}$$

Also, since $F_X(X_i)$ is exactly uniform(0,1) and K has compact support, for all large n,

$$\mathbb{E}\hat{f}_Y\left(F_X(x);F_X,h_t\right)=f_Y\left(F_X(x);F_X\right). \tag{5.32}$$

By (5.31) and (5.32) and since $f_Y(F_X(x);F_X)F_X'(x) = f_X(x) = f_Y\left(\widetilde{F}(x);\widetilde{F}\right)\widetilde{F}'(x)$,

$$\begin{aligned}
&\left|\left(\mathbb{E}\widehat{f}_Y\left(g(x);g,h_t\right)\right)\Big|_{g=\widetilde{F}_{t-1}}\widetilde{F}_{t-1}'(x) - \mathbb{E}\widehat{f}_Y\left(F_X(x);F_X,h_t\right)F_X'(x)\right| \\
&\quad\le \left|\left(\mathbb{E}\widehat{F}_Y\left(g(x);g,h_t\right)\right)\Big|_{g=\widetilde{F}_{t-1}}\widetilde{F}_{t-1}'(x) - f_Y\left(\widetilde{F}_{t-1}(x);\widetilde{F}_{t-1}\right)\widetilde{F}_{t-1}'(x)\right| \\
&\qquad + \left|\mathbb{E}\widehat{f}_Y(F_X(x);F_X,h_t)F_X'(x) - f_Y(F_X(x);F_X)F_X'(x)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.33}$$

By the same type of argument that established (5.12),

$$\begin{aligned}
&\left|\widehat{f}_{Y,*}\left(g(x);g,h_t\right)\Big|_{g=\widetilde{F}_{t-1}} - \widehat{f}_{Y,*}\left(F_X(x);F_X,h_t\right)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.34}$$

Also,

$$\begin{aligned}
&\left|\widehat{f}_{Y,*}\left(g(x);g,h_t\right)\Big|_{g=\widetilde{F}_{t-1}}\left(F'(x) - F_X'(x)\right)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.35}$$

By (5.34) and (5.35),

$$\begin{aligned}
&\left|\widehat{f}_{Y,*}\left(g(x);g,h_t\right)\Big|_{g=\widetilde{F}_{t-1}}\widetilde{F}'(x) - \widehat{f}_{Y,*}\left(F_X(x);F_X,h_t\right)F_X'(x)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.36}$$

By (5.33) and (5.36),

$$\begin{aligned}
&\left|\widehat{f}_Y\left(F_X(x);F_X,h_t\right)F_X'(x) - \widehat{f}_Y\left(\widetilde{F}_{t-1}(x);\widetilde{F}_{t-1},h_t\right)\widetilde{F}_{t-1}'(x)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.37}$$

Finally, by an argument similar to that establishing (5.21),

$$\begin{aligned}
&\left|\widehat{f}_Y\left(\widehat{F}_{t-1}(x);\widehat{F}_{t-1},h_t\right)\widehat{F}_{t-1}'(x) - \widehat{f}_Y\left(\widetilde{F}_{t-1}(x);\widetilde{F}_{t-1}h_t\right)\widetilde{F}_{t-1}'(x)\right| \\
&\quad = o_P\left(n^{-2t/(4t+1)}\right).
\end{aligned} \tag{5.38}$$

(5.37) and (5.38) prove (2.8). Then (2.9) follows easily from (2.8). □

Acknowledgments. We thank Steve Marron for pointing out the connection between our work and Abramson (1984) and for supplying GAUSS

subroutines for linear binning and interpolation. We thank an Associate Editor and reviewers for their thoughtful comments.

REFERENCES

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimation—a square root law. *Ann. Statist.* **10** 1217–1223.

ABRAMSON, I. S. (1984). Adaptive density flattening—a metric distortion principle for combating bias in nearest neighbor methods. *Ann. Statist.* **12** 880–886.

BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25** 245–254.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

DEHEUVELS, P. (1977). Estimation non-paramétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl.* **25** 5–42.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

GAJEK, L. (1986). On improving density estimators which are not bona fide functions. *Ann. Statist.* **14** 1612–1618.

HÄRDLE, W., MARRON, J. S. and WAND, M. P. (1990). Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B* **52** 223–232.

HALL, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77** 529–535.

HALL, P. and MARRON, J. S. (1988). Variable window width kernel estimates of probability densities. *Probab. Theory Related Fields* **80** 37–49.

JONES, M. C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.* **84** 733–741.

JONES, M. C. (1990). Variable kernel density estimates and variable kernel density estimates. *Austral. J. Statist.* **32** 361–371. [Correction (1991) **33** 119.]

JONES, M. C. (1991). On correcting for variance inflation in kernel density estimation. *Comput. Statist. Data Anal.* **11** 3–15.

LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

MARRON, J. S. and RUPPERT, D. (1995). Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser. B*. To appear.

MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

MÜLLER, H.-G. and ZHOU, H. (1991). Comment on "Transformations in density estimation" by M. P. Wand, J. S. Marron and D. Ruppert. *J. Amer. Statist. Assoc.* **86** 356–358.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, Berlin.

RICE, J. (1984). Boundary modification for kernel regression. *Comm. Statist. A—Theory Methods* **13** 893–900.

RUDEMO, M. (1991). Comment on "Transformations in density estimation" by M. P. Wand, J. S. Marron and D. Ruppert. *J. Amer. Statist. Assoc.* **86** 353–354.

RUPPERT, D. and WAND, M. P. (1992). Correcting for kurtosis in density estimation. *Austral. J. Statist.* **34** 19–29.

SHEATHER, S. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.

SILVERMAN, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SINGH, R. S. (1977). Estimation of derivatives of a density. *Ann. Statist.* **5** 400–404.

SINGH, R. S. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika* **66** 177–180.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

WAND, M. P., MARRON, J. S. and RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86** 343–361.

WAND, M. P. and SCHUCANY, W. R. (1990). Gaussian-based kernels for curve estimation and window width selection. *Canad. J. Statist.* **18** 197–204.

SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
CORNELL UNIVERSITY
ETC BUILDING
ITHACA, NEW YORK 14853-3801

DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843