View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

JOURNAL OF MULTIVARIATE ANALYSIS 15, 147–163 (1984)

Almost Surely Consistent Nonparametric Regression from Recursive Partitioning Schemes

LOUIS GORDON*

University of Southern California

AND

RICHARD A. OLSHEN†

University of California, San Diego

Communicated by M. Rosenblatt

Presented here are results on almost sure convergence of estimators of regression functions subject to certain moment restrictions. Two somewhat different notions of almost sure convergence are studied: unconditional and conditional given a training sample. The estimators are local means derived from certain recursive partitioning schemes. © 1984 Academic Press, Inc.

1. INTRODUCTION AND SUMMARY

This paper is concerned with the almost sure convergence of within-box means to conditional expectations. These "boxes" are derived from the recursive partitioning of a "feature space," the Euclidean range of the explanatory variables in a regression problem. The within-box means and the partitioning are derived from a training sample, which we imagine accrues sequentially. The cited partitions we study here are required to be nested as the training sample grows. A class of partitioning algorithms which leads to almost surely consistent estimators is presented in Section 5; we refer the interested reader to the book [3] and to references of our previous papers [12, 13] for a much more extensive list of recursive partitioning derived estimators in classification and regression. The reader is urged to have [13] at hand when reading this paper.

Received November 1981; revised September 1983.

* The bulk of Louis Gordon's work on this paper was completed while he was on the staff of the Energy Information Administration, U. S. Department of Energy.

† Research supported in part by National Science Foundation Grant MCS 79-06228 to the University of California, San Diego.

AMS 1970 subject classifications: Primary, 62G05; Secondary, 62E20, 62G30, 68G10.

Key words and phrases: Recursive partioning schemes, tree-structured methods, almost sure convergence.

0047-259X/84 $3.00

Copyright © 1984 by Academic Press, Inc.
All rights of reproduction in any form reserved.

Suppose that (X, Y), $(X_1, Y_1),\ldots, (X_n, Y_n)$ are independent and identically distributed (iid) random vectors on the probability space $(\Omega, \mathscr{F}, P)$, with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^1$. If $E\{Y\}$ exists, then $h(x) = E\{Y \mid X = x\}$ is the regression of Y on X. The assumption is that X and our training sample $\{(X_i, Y_i)_{i=1}^n\}$ are given, and that h is to be estimated. Our estimators will be written $\hat{h}_n = \hat{h}_n(X) = \hat{h}_n(X, (X_1, Y_1),\ldots, (X_n, Y_n))$ and are nearly always simple averages of those Y_i's, $1 \leqslant i \leqslant n$, for which X_i lies in the same "box" of a partition of $\mathbb{R}^d$ as X does.

Informally, a box is a polyhedron with at most a preassigned number of faces. The notion of just what a box is has been discussed in [12, 13, 3]; what we need here is nearly the same as what is in [3]. The precise definition and other necessary preliminaries are discussed in Section 2. Note that a recursive partitioning scheme has associated with it a binary tree. See [3] for details of the association. Note also that the rules of Section 5 and many of those of [3] are invariant to strictly monotonic transformations of the X coordinate axes.

Section 3 is devoted to a proof that if (i) Y satisfies a certain moment restriction; (ii) the cited partitions are nested as n grows without bound; (iii) the "diameter" of a randomly chosen box (in the partition based on $(X_1, Y_1),\ldots, (X_n, Y_n)$) tends to 0 in probability; and (iv) each box contains at least $k = k(n)$ X's from among $X_1,\ldots, X_n$ (for suitable $k(n)$); then $\hat{h}_n(X)$ tends to $h(X)$ almost surely. The proof is accomplished by a sequence of lemmas, which invoke the martingale convergence theorem and adaptation of the arguments of [12, 13, 3] along with certain technical considerations which were not necessary in the cited work. We are not aware of any result like that of Section 3 for other nonparametric estimators of possibly unbounded functions. Devroye [6] has given a result like that of Section 3 for certain kernel and nearest neighbor estimators and bounded Y.

The more conventional notion of almost sure convergence in regression is that of Section 4. There it is shown that if $\hat{h}_n(X)$ tends almost surely to $h(X)$ and Y satisfies certain moment conditions, then $E\{|\hat{h}_n(X) - h(X)|^p \mid (X_1, Y_1),\ldots, (X_n, Y_n)\}$ tends almost surely to 0 for prescribed values of p. Devroye [6] has given a result like that of Section 4 for certain kernel and nearest neighbor estimators in the case when Y is bounded. Geman [10] and Geman and Hwang [11] have given another argument which applies when $d = 1$, $p = 2$, $Y \in L^2$, and the estimators of $h(X)$ are given by the method of sieves. Their arguments generalize to $d > 1$ and more general scenarios.

Nearly all results of the present paper, in contrast to those of [12, 13, 3], require that the partitions $Q^{(n)}$ of $\mathbb{R}^d$ be nested as n increases. A major problem which led to this is related to problems in the differentiation of integrals and is discussed in Section 6. However, there are other considerations which dictate interest in nested $Q^{(n)}$'s.

As data accrue sequentially and a recursive partitioning algorithm produces nested $Q^{(n)}$'s, a given $\hat{h}_n$—that is, its corresponding tree (see [3, 13])—can be used to facilitate preprocessing necessary to the determination of subsequent partitions.

2. Preliminaries

This section introduces notation and terminology beyond what was introduced in Section 1. Here we discuss the fundamental concepts of box, basic box, and partition. The presentation closely follows those of [3, 13] and is included for the sake of completeness. Finally, we provide a short digression on the Blackwell–Dubins lemma which is used in Section 4.

Let $d_1 \geqslant d+1$ be a fixed positive integer. A *box* is a set $B \subset \mathbb{R}^d$ which is the solution set to a system of at most d_1 inequalities, each inequality being of the form $b_1x_1 + \cdots + b_dx_d \leqslant c$ or $b_1x_1 + \cdots + b_dx_d < c$, where c and $b_1, \ldots, b_d$ are real numbers with at least one $b_i \neq 0$. If for each linear inequality defining B, exactly one b_i, $1 \leqslant i \leqslant d$, is not 0, then B is a *basic box*.

In [12, 13] we have emphasized the invariance of recursive partitioning derived rules for classification and regression to strictly monotonic transformations for the coordinate axes of $\mathbb{R}^d$. With the notions of box and $\hat{h}_n(X)$ given in this paper, for the cited invariance to hold not only must all boxes be basic boxes, but also each of the hyperplanes which determine the boundary of a box must contain at least one X_i, $i \leqslant n$, which belongs to the training sample. The general definition of box allows for linear combination splits, as defined in Section 5.2 of [3].

We reserve Q as a generic symbol for a finite partition of $\mathbb{R}^d$, all of whose component subsets are boxes B. For $x \in \mathbb{R}^d$, we denote by $B(x)$ the unique box in Q containing x. If a sequence of partitions is discussed, the index is superscripted, and the same indexing is carried to boxes. For example, $Q^{(n)}$ denotes an element is a sequence of partitions, and $B^{(n)}(x)$ is that box in $Q^{(n)}$ containing x.

As was mentioned in Section 1, we think of (X_1, Y_1), $(X_2 Y_2)$,..., as being observed in sequence. At each stage n of sampling we are given $Q^{(n)}$, a partition of $\mathbb{R}^d$ which depends measurably on $(X_1, Y_1) \cdots (X_n, Y_n)$. We further assume throughout that the partitions observed at the various stages of sampling are nested; that is, $Q^{(n+1)}$ is a refinement of $Q^{(n)}$ (which is not to preclude that $Q^{(n+1)} = Q^{(n)}$).

Write F for the common distribution of X and the X_i's and let $\operatorname{supp} X$ denote the support of F. The *diameter* $D_n(x)$ of the box of $Q^{(n)}$ containing x is defined as

$$D_n(x) = \min(D'_n(x), 1),$$

where

$$D_n'(x) = \sup\{\|z - y\| : y, z \in B^{(n)}(x) \cap \operatorname{supp} X\}$$

and $\|z - y\|$ is the Euclidean distance between z and y. Because we assume that the $Q^{(n)}$ are nested, $D_n(X)$ is monotone nonincreasing in n.

We shall have recourse in Section 4 to a version of the dominated convergence theorem for conditional expectations. This lemma is variously known as the Blackwell–Dubins lemma, for its appearance in [2], or as Hunt's lemma, for example, in Chung and Walsh [5] who refer to Hunt [14, p. 47]. The lemma, by either name, says that if Y_n tends almost surely to Y, $\sup |Y_n|$ is integrable, and $\mathscr{F}_n$ is either an increasing or decreasing sequence of σ-fields, then

$$E\{Y_n \mid \mathscr{F}_n\} \text{ tends almost surely to } E\{Y \mid \mathscr{F}\},$$

where $\mathscr{F}$ is either $\bigcap \mathscr{F}_n$ or $\bigvee \mathscr{F}_n$ as the σ-fields $\mathscr{F}_n$ are either decreasing or increasing.

During the course of our work we introduce various constants: $c_1, c_2, \ldots$. The values of these numbers are not material to our arguments.

3. Unconditional Almost Sure Convergence

In this section we state and prove our main result. In particular, Theorem 3.6 gives conditions sufficient to guarantee almost sure consistency for a class of nonparametric estimators of h. Proof of the theorem is accomplished by a sequence of lemmas.

There is some measure-theoretic delicacy involved in arguing with what appear in our notation to be conditional expectations given adaptively determined σ-fields of subsets of $\mathbb{R}^d$. Thus, we require the following notation, in which $\sigma\{\cdot\}$ is the σ-field of subsets of $\mathscr{B}$ determined by the random variables described inside the brackets, and (as usual) $\mathscr{C} \vee \mathscr{D}$ is the smallest σ-field which contains the σ-fields $\mathscr{C}$ and $\mathscr{D}$.

$$\mathscr{S}_n = \sigma\{(X_1, Y_1), \ldots, (X_n, Y_n)\}. \tag{3.1}$$

$$\mathscr{F}_n = \sigma\{(X_1, Y_1), \ldots, (X_n, Y_n), I_B(X) : B \in Q^{(n)}\}, \tag{3.2}$$

where $I_B(X)$ is 1 if $X(\omega) \in B$

and 0 otherwise.

If B is a box, then (3.3)

$$\mu(B) = E\{Y I_B(X)\}$$

and

$$\hat{\mu}_n(B) = \frac{1}{n} \sum_{i=1}^{n} Y_i I_B(X_i).$$

For $x \in \mathbb{R}^d$, if $P(X \in B^{(n)}(x)) > 0$ then $h_n(x) = \mu(B^{(n)}(x))/P(X \in B^{(n)}(x))$; otherwise $h_n(x) = 0$. (3.4)

For $x \in \mathbb{R}^d$ and $B^{(n)}(x) = B$, $\hat{h}_n(x) = [\hat{\mu}_n(B)/\hat{F}_n(B)]\, I_{\{\hat{F}_n(B) \geq k(n)/n\}}$, (3.5)

where 0/0 is taken to be 0, and $\hat{F}_n(B)$ is the empirical probability of B based on $X_1, X_2, \ldots, X_n$. In (3.5) $k(n)$ is a nondecreasing sequence of positive integers which tends to ∞. The ideas of this section are summarized in the next result.

3.6. THEOREM. *Assume that*

$E\{|Y|^p\} < \infty$ *for some* $p > 1$; (3.7)

$D_n(X)$ *tends to* 0 *in probability as* n *tends to* ∞; (3.8)

for $n = 1, 2, \ldots$, *the partition* $Q^{(n+1)}$ *refines the partition* $Q^{(n)}$ *of* $\mathbb{R}^d$*—it may happen that* $Q^{(n+1)} = Q^{(n)}$; (3.9)

$n^{1/p} \log n/k(n)$ *tends to* 0 *as* n *tends to* ∞; (3.10)

$I_{\{\hat{F}_n(B^{(n)}(X)) < k(n)/n\}}$ *tends to* 0 *almost surely as* n *tends to* ∞. (3.11)

Then $\hat{h}_n(X)$ *tends to* $h(X)$ *almost surely as* n *tends to* ∞.

Theorem 3.6 is the consequence of a sequence of lemmas which is our concern for most of the remainder of this section. Throughout, we assume without loss of generality that Y is nonnegative. We also assume throughout this section that (3.7) through (3.11) are satisfied.

3.12. LEMMA. $E\{Y \mid \mathcal{F}_n\} = h_n(X)$.

Proof. For each box B let $A(B) = E\{YI_B(X) \mid \mathcal{S}_n\}/E\{I_B(X) \mid \mathcal{S}_n\}$. It follows from the independence of (X, Y) and the training sample and from Fubini's theorem that $h_n(X) = \sum_{B \in Q^{(n)}} A(B)\, I_B(X)$. Since $h_n(X)$ is $\mathcal{F}_n$

measurable, in order to prove the lemma it suffices to show that $E\{I_S I_B(X)\, h_n(X)\} = E\{YI_B(X) T_S\}$ for $B \in Q^{(n)}$ and $S \in \mathscr{S}_n$. But

$$\begin{aligned} E\{I_S I_B(X)\, h_n(X)\} &= E\{I_S I_B(X)\, A(B)\} \\ &= E\{I_S A(B)\, E\{I_B(X) \mid \mathscr{S}_n\}\} \\ &= E\{I_S E\{YI_B(X) \mid \mathscr{S}_n\}\} \\ &= E\{E\{YI_S I_B(X) \mid \mathscr{S}_n\}\} = E\{YI_B(X) I_S\}. \end{aligned}$$

3.13. LEMMA. *If $D_n(X)$ tends in probability to 0, then $h_n(X) - h(X)$ tends almost surely to 0.*

Proof. Because the partitions $Q^{(n)}$ are nested, the random variables $D_n(X)$ are nonincreasing, and it follows that (3.8) is equivalent to the assumption that $D_n(X)$ tends almost surely to 0. Because the σ-fields $\mathscr{F}_n$ are nondecreasing, the random variables $h_n(X) = E\{Y \mid \mathscr{F}_n\}$ are an expectation bounded martingale. A special case of the martingale convergence theorem implies that $h_n(X)$ tends almost surely to $E\{Y \mid \mathscr{F}_\infty\}$, where $\mathscr{F}_\infty = \bigvee \mathscr{F}_n$. Denote by $\mathscr{F}_n^*$ and $\mathscr{F}_\infty^*$ the completions of $\mathscr{F}_n$ and $\mathscr{F}_\infty$, respectively; and observe that (almost surely) $h_n(X) = E\{Y \mid \mathscr{F}_n^*\}$, and $h_n(X)$ tends almost surely to $E\{Y \mid \mathscr{F}_\infty^*\}$. In the next two paragraphs it will be argued that $E\{Y \mid \mathscr{F}_\infty^*\} = E\{Y \mid X\}$ almost surely.

Let $K = \operatorname{supp} X$ and V be an arbitrary open subset of $\mathbb{R}^d$. Then

$$I_V(X) = \overline{\lim} \sum_{\substack{B \in Q^{(n)} \\ B \cap K \subset V}} I_B(X)$$

almost surely because $D_n(X)$ tends to 0 almost surely. So $\sigma(X)$ is generated by a (countable) family of $\mathscr{F}_\infty^*$ sets; we conclude that X is $\mathscr{F}_\infty^*$ measurable and $h(X) = E\{Y \mid X\}$ is $\mathscr{F}_\infty^*$ measurable.

We now verify the conditional expectation property for indicators of sets which generate $\mathscr{F}_\infty^*$. To that end let $S \in \mathscr{S}_n$ and $B \in Q^{(n)}$.

$$\begin{aligned} E\{I_S I_B(X)\, Y\} &= E\{I_S I_B(X)\, E\{Y \mid \sigma(X) \vee \mathscr{S}_n\}\} \\ &= E\{I_S I_B(X)\, E\{Y \mid X\}\}; \end{aligned} \tag{3.14}$$

the last equality is a consequence of the independence of (X, Y) and $\mathscr{S}_n$. From a standard monotone class argument and (3.14) one concludes that for all $T \in \mathscr{F}_\infty^*$, $E\{YI_T\} = E\{I_T E\{Y \mid X\}\}$. Because $\sigma\{X\} \subset \mathscr{F}_\infty^*$, $E\{Y \mid X\} = E\{Y \mid \mathscr{F}_\infty^*\}$ almost surely. And since $E\{Y \mid \mathscr{F}_n^*\}$ tends almost surely to $E\{Y \mid \mathscr{F}_\infty^*\}$, it follows that $h_n(X)$ tends almost surely to $E\{Y \mid X\} = h(X)$.

The following theorem is a corollary of Theorem 12.2 of [3]. In the theorem (in a slight abuse of notation) we write $P([X \in B] \cap [Y \in I])$ (for

boxes $B \subset \mathbb{R}^d$ and intervals of real numbers I) as $F(B, I)$. Similarly, $\hat{F}_n(\cdot, \cdot)$ refers in an obvious way to empirical probabilities based on $(X_1, Y_1),\ldots,$ (X_n, Y_n).

3.15. THEOREM. [3] *Given $\varepsilon > 0$ and $p' > 0$, there exists a constant $c_1 = c_1(p', \varepsilon, d)$ such that for all $n \geqslant N(p', \varepsilon, d)$ sufficiently large there is a set in $\mathscr{B}$ of probability at least $1 - n^{-(2p'+1)}$ on which simultaneously for all intervals I and all boxes $B \in Q^{(n)}$.*

$$|\hat{F}_n(B, I) - F(B, I)| \leqslant \varepsilon \hat{F}(B, I) + c_1 \frac{\log n}{n} \tag{3.16}$$

and

$$|\hat{F}_n(B, I) - F(B, I)| \leqslant \varepsilon F(B, I) + c_1 \frac{\log n}{n}.$$

In what follows, for each positive γ, $\hat{h}_{n,\gamma}$ is defined as $\hat{h}_n$ except that Y_i which appears in the definition of $\hat{\mu}_n$ is replaced by $\max(-\gamma, \min(Y_i, \gamma))$. The reader will note that $\hat{h}_{n,\gamma}$ was called $\hat{h}_n$ in [13].

3.17. LEMMA. *Let γ_n be an arbitrary sequence of positive constants, and assume that $p > 1$. Given $\frac{1}{2} > \varepsilon > 0$, there exists a constant $c_2 = c_2(p, \varepsilon, d)$ for which, with probability at least $1 - n^{-(2p+1)}$,*

$$\begin{aligned} |\hat{h}_{n,\gamma_n}(X) - h_n(X)| < {} & E\{(Y - \gamma_n)_+ \mid \mathscr{F}_n\} + c_2 \gamma_n \log n/k(n) \\ & + h_n(X)[I_{\{\hat{F}_n(B^{(n)}(X)) < k(n)/n\}} + 5\varepsilon + c_1 \log n/k(n)]. \end{aligned} \tag{3.18}$$

Proof. As we did in [13], we write

$$\begin{aligned} |\hat{h}_{n,\gamma_n}(X) - h_n(X)| \leqslant {} & E\{(Y - \gamma_n)_+ \mid \mathscr{F}_n\} \\ & + \frac{1}{\hat{F}_n(B^{(n)}(X))} \int_0^{\gamma_n} |\hat{P}_n(y, X) - P_n(y, X)|\, dy\, I_{\{\hat{F}_n(B^{(n)}(X)) \geqslant k(n)/n\}} \\ & + \int_0^{\gamma_n} P_n(y, X) \left| \frac{1}{\hat{F}_n(B^{(n)}(X))} - \frac{1}{F(B^{(n)}(X))} \right| dy\, I_{\{\hat{F}_n(B^{(n)}(X)) \geqslant k(n)/n\}} \\ & + h_n(X) I_{\{\hat{F}_n(B^{(n)}(X)) < k(n)/n\}} \\ \stackrel{\text{def}}{=} {} & \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}, \end{aligned}$$

where $P_n(y, x) = F(B^{(n)}(x), (y, \infty))$, $\hat{P}_n(y, x) = \hat{F}_n(B^{(n)}(x), (y, \infty))$, and $\hat{F}_n(B^{(n)}(X), \mathbb{R}^1)$ is abbreviated to $\hat{F}_n(B^{(n)}(X))$.

Terms I and IV are repeated verbatim in the statement of the lemma. There remains the task of bounding terms II and III.

Because ε, p, and d are fixed, we may choose c_1 as in Theorem 3.15. From that theorem it follows that for $n = n(p, \varepsilon, d)$ sufficiently large, with probability at least $1 - n^{-(2p+1)}$

$$\begin{aligned} \text{II} &\leqslant \frac{2/(1-\varepsilon)}{F(B^{(n)}(X))} \int_0^{\gamma_n} \varepsilon P_n(y, X)\, dy + c_1 \frac{\gamma_n \log n}{k(n)} \\ &\leqslant 4\varepsilon \int_0^{\gamma_n} \frac{P_n(y, X)}{F(B^{(n)}(X))}\, dy + c_1 \frac{\gamma_n \log n}{k(n)} \\ &\leqslant 4\varepsilon\, h_n(X) + c_1 \frac{\gamma_n \log n}{k(n)}. \end{aligned}$$

An application of Theorem 3.15 to III shows that for $n = n(p, \varepsilon, d)$ sufficiently large, on the same set of probability $1 - n^{-(2p+1)}$ as appeared in the argument for II,

$$\begin{aligned} \text{III} &\leqslant \int_0^{\gamma_n} \frac{P_n(y, X)}{F(B^{(n)}(X))} \left(\varepsilon + c_1 \frac{\log n}{k(n)} \right) dy \\ &\leqslant (\varepsilon + c_1(\log n/k(n)))\, h_n(X). \end{aligned}$$

3.19. LEMMA. *Let $\gamma_n = n^{1/p}$; then as n tends to ∞ $\hat{h}_{n,\gamma_n}(X) - h_n(X)$ tends to 0 almost surely.*

Proof. Choose and fix ε for which $0 < \varepsilon < \frac{1}{2}$. From Lemma 3.17 and the Borel–Cantelli lemma it follows that almost surely, (3.18) holds for all but finitely many n. We now show that each of the three summands of (3.18) tends to 0 almost surely.

Since the σ-fields $\mathcal{F}_n$ are monotone nondecreasing, $(Y - \gamma_n)_+$ tends to 0 almost surely, and $(Y - \gamma_n)_+ \leqslant Y$, the Blackwell–Dubins lemma implies that $E\{(Y - \gamma_n)_+ \mid \mathcal{F}_n\}$ tends to 0 almost surely.

That the second term of (3.18) tends to 0 follows from (3.10).

Because $h_n(X) = E\{Y \mid \mathcal{F}_n\}$, Kolmogorov's inequality for martingales implies that $\sup_n h_n(X) < \infty$ almost surely. Hence, in view of (3.11), we may make the third summand of (3.18) arbitrarily small on a set of arbitrarily large probability by a judicious choice of ε in Lemma 3.17.

3.20. LEMMA. *Write $\gamma_n = n^{1/p}$; then $\hat{h}_n(X) - \hat{h}_{n,\gamma_n}(X)$ tends to 0 almost surely.*

Proof. If $\hat{F}_n(B^{(n)}(X)) < k(n)/n$ then the cited difference is 0, so suppose that $\hat{F}_n(B^{(n)}(X)) \geqslant k(n)/n$. In that case

$$|\hat{h}_n(X) - \hat{h}_{n,\gamma_n}| = \int_{\gamma_n}^{\infty} \frac{\hat{P}_n(y, X)}{\hat{F}_n(B^{(n)}(X))}\, dy,$$

which is not more than the

$$\sup_{j \leqslant n} (Y_j - \gamma_n)_+ .$$

But it follows from a result of Robbins quoted in [1] that $\sup_{j \geqslant n}(Y_j - \gamma_n)_+$ is almost surely eventually 0.

Proof of Theorem 3.6. Write $\gamma_n = n^{1/p}$ and

$$\hat{h}_n(X) - h(X) = (\hat{h}_n(X) - \hat{h}_{n,\gamma}(X)) + (\hat{h}_{n,\gamma}(X) - h_n(X)) + (h_n(X) - h(X))$$

$$\stackrel{\text{def}}{=} \text{V} + \text{VI} + \text{VII}.$$

Term V tends almost surely to 0 as a consequency of Lemma 3.20, whereas terms VI and VII likewise tend almost surely to 0 in view, respectively, of Lemma 3.19 and Lemma 3.13.

4. Conditional Almost Sure Convergence

We mentioned in Section 1 that the notion of almost sure convergence as studied in the previous section differs from the notions of convergence as presented in [6, 10, 11]. In this section it is shown that, given suitable moment conditions, the notion of convergence established for nested recursive partitioning estimators in Section 3 implies the integrated form of almost sure convergence studied in the cited references. In particular, we show that if $E\{|Y|^p\} < \infty$ for some $p > 2$, then $E\{\sup_n |\hat{h}_n(X)|^r\} < \infty$ for $r < p/2$. This, together with the Blackwell–Dubins lemma, establishes the result which follows.

4.1. Theorem. *Assume that*

$$E\{|Y|^p\} < \infty \text{ for some } p > 2r \geqslant 2; \tag{4.2.}$$

$$D_n(X) \text{ tends to } 0 \text{ in probability as } n \text{ tends to } \infty; \tag{4.3}$$

$$\text{For all } n = 1, 2, \ldots, \text{ the partition } Q^{(n+1)} \text{ of } U \text{ refines (but may be identical to) the partition } Q^{(n)}; \tag{4.4}$$

$$n^{1/r} \log n/k(n) \text{ tends to } 0 \text{ as } n \text{ tends to } \infty; \text{ and} \tag{4.5}$$

$$I_{\{\hat{F}_n(B^{(n)}(X)) < k(n)/n\}} \text{ tends to } 0 \text{ almost surely as } n \text{ tends to } \infty. \tag{4.6}$$

It then follows that

$$E\{|\hat{h}_n(X) - h(X)|^r \mid \mathscr{S}_n\} \text{ tends to } 0 \text{ almost surely as } n \text{ tends to } \infty.$$

We prove Theorem 4.1 by means of Lemmas 4.7, 4.8, and 4.9. Throughout we assume that (4.2) through (4.6) are in force and that $Y \geqslant 0$; also, take q to be a constant for which $1 < q < p - r$.

4.7. LEMMA. *From the stated assumptions it follows that*

$$E\{(\sup_n h_n(X))^p\} < \infty.$$

Proof. We have from Lemma 3.12 that $h_n(X) = E\{Y \mid \mathscr{F}_n\}$, and so $h_n(X)$ is a martingale. From the well-known inequality of Doob [7, p. 317],

$$E\{\sup_n h_n^p(X)\} \leqslant \left(\frac{p}{p-1}\right)^p E\{Y^p\}.$$

4.8. LEMMA. *Let* $\gamma_n = n^{1/q}$. *Then*

$$E\{\sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)|^p\} < \infty;$$

that is, $\sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)|$ *belongs to* $L^p = L^p(\Omega, \mathscr{B}, P)$.

Proof. Choose and fix ε for which $0 < \varepsilon < \frac{1}{2}$. Let I_n^* be the indicator of the event that (for n) the inequality (3.18) is violated. By noticing that

$$\sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)|$$

$$\leqslant \sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)| I_n^* + \sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)| (1 - I_n^*)$$

and that

$$\sup_n \gamma_n I_n^* \leqslant \sum_{n=1}^{\infty} \gamma_n I_n^*,$$

we have that

$$\begin{aligned}\sup_n |\hat{h}_{n,\gamma_n}(X) - h_n(X)| &\leqslant \sup_n h_n(X) I_n^* + \sum_{n=1}^{\infty} I_n^* \gamma_n \\ &\quad + \sup_n h_n(X) \\ &\quad + \sup_n c_2 \gamma_n \log n/k(n) \\ &\quad + \sup_n h_n(X)[1 + 5\varepsilon + c_1 \log n/k(n)] \\ &\stackrel{\text{def}}{=} \text{VIII} + \text{IX} + \text{X} + \text{XI} + \text{XII}.\end{aligned}$$

Terms VIII, X, and XII are in L^p because of Lemma 4.7. In view of Lemma 3.17, compute thus with IX and the L^p norm:

$$\left\| \sum_{n=1}^{\infty} I_n^* \gamma_n \right\|_p \leqslant \sum_{n=1}^{\infty} \| I_n^* \gamma_n \|_p$$
$$\leqslant \sum_{n=1}^{\infty} n^{1/q} n^{-[2+(1/p)]} < \infty,$$

and so IX is bounded in L^p. The boundedness of XI is a consequence of (4.5).

4.9. LEMMA. $E\{\sup_n |\hat{h}_n(X) - \hat{h}_{n,\gamma_n}(X)|^r\} < \infty$.

Proof. $\hat{h}_{n,\gamma_n}(X) - \hat{h}_n(X)$ can be written (see (3.16) of [13])

$$\int_{\gamma_n}^{\infty} \{1 - \hat{F}_n(y \mid B^{(n)}(X))\}\, dy\, I_{\{\hat{F}_n(B^{(n)}(X)) \geqslant k(n)/n\}},$$

which is not more than the

$$\max_{j \leqslant n} (Y_j - \gamma_n)_+ . \tag{4.10}$$

As was noted in Section 3, (4.10) is almost surely eventually 0. Because the γ_n are increasing, there almost surely exists some $n_0 = n_0(\omega)$ for which

$$\max_n \max_{j \leqslant n} (Y_j - \gamma_n)_+ = (Y_{n_0} - \gamma_{n_0})_+ .$$

Hence

$$\sup_n |\hat{h}_{n,\gamma_n}(X) - \hat{h}_n(X)| \leqslant \max_n (Y_n - \gamma_n)_+ ,$$

and also

$$E\{\sup_n |\hat{h}_{n,\gamma_n}(X) - \hat{h}_n(X)|^r\} \leqslant \sum_{n=1}^{\infty} \int_0^{\infty} r y^{r-1} (1 - F(y + \gamma_n))\, dy$$
$$\leqslant \sum_{n=1}^{\infty} r E(Y^p) \int_0^{\infty} y^{r-1} (y + \gamma_n)^{-p}\, dy, \tag{4.11}$$

where we have made use of the Markov inequality in (4.11). The expression (4.11) is not more than

$$\sum_{n=1}^{\infty} c_3 n^{-(p-r)/q},$$

which is finite since $p - r > q$.

Proof of Theorem 4.1. Lemmas 4.7, 4.8, and 4.9 imply that

$$E\{\sup_n |\hat{h}_n(X) - h(X)|^r\} < \infty.$$

From Theorem 3.6 it follows that $\hat{h}_n(X) - h(X)$ tends to 0 almost surely. The Blackwell–Dubins lemma implies that

$$E\{|\hat{h}_n(X) - h(X)|^r \mid \mathcal{S}_n\} \text{ tends to 0 almost surely.}$$

The reader may be interested to note that we have paid the price of hypothesis (4.2) in order to preserve the affine invariance of $\hat{h}_n$. For example, the conclusion of Theorem 4.1 holds with $r = p$ and $k(n) = n^\theta$ for the truncated estimators of [13] provided $2p\theta > p + 2$. The cited price is confined to Lemma 4.9 in the present paper. We believe that a conclusion like that mentioned for the estimators of [13] holds for suitably chosen affinely invariant truncated estimators.

5. Implementing and Almost Surely Consistent Splitting Rule

We now sketch an adaptive splitting rule to which apply the arguments of Theorems 3.6 and 4.1. Therefore, the corresponding sequence of estimators converges almost surely in the two described senses to $E\{Y \mid X\}$. In order to simplify exposition of the rule we assume throughout this section that F has continuous marginal distributions on the d coordinate axes of X. An additional simplifying assumption in force is that all boxes are basic boxes. The algorithm can be implemented so that $\hat{h}_n(X)$ is invariant to strictly monotonic transformations of the coordinate axes of X.

The nesting of partitions, which enables us to prove the two cited theorems, entails certain difficulties in the definition of the splitting rule. For as n grows, a terminal box B which once had more than the desired $k = k(n)$ members of the training sample may have fewer members if sufficiently few observations among $X_{n+1}, X_{n+2}, \ldots$, belong to B. In order to demonstrate that (3.11) (which is the same thing as (4.6)) holds, we employ a large deviation result which ensures that for F almost every x, and the splitting rule to be sketched, the possible lapses in box content are almost surely temporary. The tree which corresponds to our splitting rule will grow fitfully, one terminal node at a time. In view of the foregoing discussion and the results of [12, 13, 3], we base our splitting rule on two guiding principles: split only those boxes which are sufficiently full; monitor and, if necessary, enforce the occurrence of quantile splits (called quantile cuts in [12, 13]. For a given box B, we say that a jth *p-quantile split* has been achieved if the box is refined by a split perpendicular to coordinate axis j into daughter boxes B',

B'' so that $\max(\hat{F}_n(B'), \hat{F}_n(B'')) \leqslant p\hat{F}_n(B)$. Necessarily, p is at least $1/2$. When, as is assumed here, the coordinates of X have continuous distributions, jth p-quantile splits can always be implemented. It is possible but technically difficult to prescribe algorithms which apply when the distribution of X is completely general.

The splitting rule requires an increasing sequence $k(n)$, and a quantile splitting target q, $\frac{1}{2} < q < \frac{3}{4}$; $k(n) = 3k'(n)$, where $k'(n) = c \log n$ and the constant c will be specified. We begin our prescription with the definition of $\hat{h}_1$: set $\hat{h}_1 \equiv Y_1$. In what follows we describe how to split and how to update $\hat{h}_{n-1}$ for $n = 2, 3, \ldots$. Throughout, $|T|$ is the cardinality of the set T. If $x \in B$ and $B \in Q^{(m)}$, we write $h_m(B)$ for the common value of $h_m(x)$ on B.

Cycle through $B \in Q^{(n-1)}$ for which $X_n \notin B$. If $|\{i: i \leqslant n, X_i \in B\}| < k(n)$, set $\hat{h}_n(B) = 0$. Otherwise, set $\hat{h}_n(B) = \hat{h}_{n-1}(B)$.]5.1)

Examine $B^{(n-1)}(X_n)$. If $|\{i: i \leqslant n, X_i \in B^{(n-1)}(X_n)\}| < k(n)$, set $\hat{h}_n(B) = 0$. If $k(n) \leqslant |\{i: i \leqslant n, X_i \in B^{(n-1)}(X_n)\}| < 8k(n)$, average $\{Y_i: i \leqslant n, X_i \in B^{(n-1)}(X_n)\}$. Set $\hat{h}_n(X_n)$ equal to that average, and stop. (5.2)

Otherwise, $|\{i: i \leqslant n, X_i \in B^{(n-1)}(X_n)\}| \geqslant 8k(n)$, and $B^{(n-1)}(X_n)$ is split according to specifications which follow. All splits are made according to the risk reduction splitting rule (see [3]) subject to the constraint that every split is a q-quantile split. Require a split on axis j if (5.3)

among the previous $4d$ ancestor nodes of $B^{(n-1)}(X_n)$ none involves a split of axis j, and (5.3a)

j is the smallest index of an axis satisfying (5.3a). (See [3] for details of the relationship between $\hat{h}_n$ and a binary tree.) (5.3b)

Compute within box averages of the Y_i, $i \leqslant n$, for the new boxes determined in (5.3). For each new box B set $\hat{h}_n(B)$ equal to its corresponding average, and stop. (5.4)

With the splitting process defined, $D_n(X)$ tends to 0 in probability in view of Lemma 4.1 and Theorems 3.13 and 3.12 of [13]. Therefore, (3.8) and (4.3) are satisfied. We now use the Borel–Cantelli lemma to demonstrate that almost surely, $\hat{F}_n(B^{(n)}(X))$ is sufficiently large for all but finitely many n so that (3.11) and (4.6) are satisfied.

What we demonstrate is that

$$\sum_{n=1}^{\infty} P\left(\bigcup_{B \in Q^{(n)}: \hat{F}_n(B) > k(n)/n} \bigcup_{s=1}^{\infty} [\hat{F}_{n+s}(B) \leqslant k'(n+s) \Big/ n+s] \right) \tag{5.5}$$

converges. The argument which is presented here is essentially due to Ken Alexander—his argument extends an earlier one of ours. Denote by E_n the events whose probabilities are being summed in (5.5). Fix an $\varepsilon > 0$ for which

$$(1-\varepsilon)^2 > 2/3. \tag{5.6}$$

Let $c_1 = c_1(1, \varepsilon, d)$ as in Theorem 3.15, and pick $c > c_1$ large enough that

$$3c - 3c\varepsilon - c_1 \geqslant 2c/(1-\varepsilon); \tag{5.7}$$

(5.6) implies that there is a c which satisfies (5.7). Clearly,

$$P(E_n) \leqslant P\left(E_n \cap \bigcap_{B \in Q^{(n)}} \left[|\hat{F}_n(B) - F(B)| \leqslant \varepsilon \hat{F}_n(B) + c_1 \frac{\log n}{n}\right]\right)$$
$$+ P\left(\bigcup_{B \in Q^{(n)}} |\hat{F}_n(B) - F(B)| > \varepsilon \hat{F}_n(B) + c_1 \frac{\log n}{n}\right). \tag{5.8}$$

It follows from Theorem 3.15 that for n sufficiently large the second probability in (5.8) is at most n^{-3}. On the event whose probability is the first term on the right hand side of (5.8), $\hat{F}_n(B) > k(n)/n$ for each B, so

$$F(B) > (3c - 3c\varepsilon - c_1)\frac{\log n}{n} \geqslant \frac{2c}{1-\varepsilon}\frac{\log n}{n}.$$

Also, on the cited event for some s and $B \in Q^{(n)}$, $\hat{F}_{n+s}(B) \leqslant k'(n+s)/(n+s) = c \log(n+s)/(n+s)$. Since for $n \geqslant 3$, $2c \log n/(1-\varepsilon)n \geqslant (c + c_1)\log(n+s)/(1-\varepsilon)(n+s)$, for s and B as described

$$\hat{F}_{n+s}(B) \leqslant (1-\varepsilon)F(B) - c_1 \frac{\log(n+s)}{n+s} \qquad \text{for} \quad n \geqslant 3. \tag{5.9}$$

The first probability on the right hand side of (5.8) has therefore been shown to be bounded, for $n \geqslant 3$, by

$$\sum_{s=1}^{\infty} P\left(\bigcup_{B \in Q^{(n)}} \left[\hat{F}_{n+s}(B) \leqslant (1-\varepsilon)F(B) - c_1 \frac{\log(n+s)}{n+s}\right]\right). \tag{5.10}$$

Theorem 3.15 implies that for n sufficiently large the sum (5.10) is at most

$$\sum_{s=1}^{\infty} (n+s)^{-3} = O(n^{-2}).$$

Therefore, $\sum_{n=1}^{\infty} P(E_n) < \infty$, and so (3.11) and (4.6) are satisfied for $k(n)$ as given. The argument applies also to show that (3.11) and (4.6) also hold for the algorithm (5.1)–(5.4) and $k(n) = 3k'(n) = o(n)$ for any nondecreasing

sequence of positive integers for which $\log n = o(k(n))$. In view of Theorem 3.6, for such $k(n)$ and p satisfying (3.7) and (3.10), $\hat{h}_n(X)$ tends to $h(X)$ almost surely as n tends to ∞, while

$$E\{|\hat{h}_n(X) - h(X)|^r \mid \mathscr{S}_n\} \text{ tends to 0 almost surely}$$

as n tends to ∞ for p and r which satisfy (4.2) and (4.5).

For example, if $n^{2/3} \log n/k(n)$ tends to 0, then the unconditional almost sure convergence of Theorem 3.6 holds if $E\{|Y|^{3/2}\} < \infty$, and the conditional almost sure convergence of Theorem 4.1 holds for $r = 3/2$ if $E\{|Y|^{3+\delta}\} < \infty$ for any $\delta > 0$.

6. The Non-nested Case Appears to be Difficult

The principal difficulty in proving a theorem like Theorem 3.6 in case the $Q^{(n)}$'s are not nested lies in the putative convergence of $h_n(X)$ to $h(X)$. If a general recursive partitioning rule guarantees that $D_n(X)$ tends almost surely to 0, then it follows from [3], Fubini's theorem, and what Garsia [9] calls Banach's principle that for all (X, Y) with $Y \in L^1$, $h_n(X)$ tends almost surely to $h(X)$ provided that (in an obvious notation)

$$\sup_n E\{Y' \mid Q^{(n)}\}(X) < \infty \qquad \text{a.e.}$$

for each $Y' \in L^1$. In other words, if $D_n(X)$ tends to 0 almost surely, and the cited suprema are finite, then the $Q^{(n)}$'s differentiate $h(X)$ for each integrable Y almost surely. From the viewpoint of Shilov and Gurevich [16], if the coordinatewise marginal distributions of X are continuous and $Q^{(n)}$ is comprised of basic boxes, then when $D_n(X)$ tends to zero almost surely, $h_n(X)$ tends almost surely to $h(X)$ if the $Q^{(n)}$'s and the distribution of X form a Vitali system.

In the remainder of the paper we assume that $\operatorname{supp} X \subset U$, the unit cube in $\mathbb{R}^d$. Arguments of R. Dudley [8] show that either (i) $D_n(X)$ tends to 0 almost surely implies that the $Q^{(n)}$'s differentiate $h(X)$ for each integrable Y almost surely, or (ii) there are an X, a sequence $Q^{(n)}$ of partitions of U for which $D_n(X)$ tends to 0 almost surely, and an open set $\mathscr{O} \subset U$ for which the $Q^{(n)}$'s do not differentiate $I_{\mathscr{O}}$ almost surely. We sketch Dudley's arguments.

First note that standard approximations which use the regularity of Borel measures on U imply that (ii) holds if we find X, $Q^{(n)}$'s, and a bounded Borel f on U for which $D_n(X)$ tends to 0 almost surely and the $Q^{(n)}$'s do not differentiate f almost surely. Suppose now that (i) fails. Then there are an integrable Y, X, and $Q^{(n)}$'s with $D_n(X)$ tending almost surely to 0 for which the $Q^{(n)}$'s do not differentiate $h(X)$ almost surely. Write $Y = Y^+ - Y^-$ to

conclude that without loss we may assume $Y \geqslant 0$. But if differentiation fails for Y, it also fails for $Y+1$. Thus, without loss, suppose that $Y \geqslant 1$ and that

$$\sum_{\substack{B \in Q^{(n)} \\ F(B)>0}} \frac{\mu(B)}{F(B)} I_B(X)$$

does not tend almost surely to $E\{Y \mid X\}$. Write the displayed sum as

$$\sum_{\substack{B \in Q^{(n)} \\ F(B)>0}} \frac{E\{Y\}\, \nu(B)}{E\{Y^{-1} Y I_B\}} I_B(X), \tag{6.1}$$

where $E\{Y\}\, \nu(B) \stackrel{\text{def}}{=} \mu(B)$. The expression (6.1) is (almost surely) the reciprocal of the conditional expectation of Y^{-1} when X has distribution ν, and Y^{-1} is clearly bounded. That is, when X has distribution ν and the $Q^{(n)}$'s are as with Y, then still $D_n(X)$ tends to 0 almost surely, and yet the $Q^{(n)}$'s do not differentiate the bounded function Y^{-1}. Thus, the cited dichotomy is demonstrated. We can actually say a bit more, as the following arguments demonstrate.

Suppose that $\mathscr{O}$ is any open subset of U. If $D_n(X)$ tends to 0 almost surely, then almost surely $E\{I_{\mathscr{O}} \mid Q^{(n)}\}(X)$ tends to 1 on $\mathscr{O}$. Also, $E\{I_{\mathscr{O}} \mid Q^{(n)}\}(X)$ tends to $I_{\mathscr{O}}$ in L^1 [13]. So Fatou's lemma implies

$$E\{\underline{\lim}\, E\{I_{\mathscr{O}} \mid Q^{(n)}\}(X)\} \leqslant \underline{\lim}\, E\{E\{I_{\mathscr{O}} \mid Q^{(n)}\}(X) = E\{I_{\mathscr{O}}\} = P(\mathscr{O}).$$

The foregoing observations guarantee that

$$\underline{\lim}\, E\{I_{\mathscr{O}} \mid Q^{(n)}\}(x) = I_{\mathscr{O}}(x) \qquad F\text{-almost everywhere,}$$

and therefore that (ii) obtains only if

$$\overline{\lim}\, E\{I_{\mathscr{O}} \mid Q^{(n)}\}(x) > 0 \tag{6.2}$$

on a subset of $\mathscr{O}^c$ of positive probability. One point of Theorem 3.6 is that with nested partitions (6.2) occurs at most on a set of F measure 0.

When X has a uniform distribution, famous results of Jessen, Marcinkiewicz, and Zygmund [15] and of Busemann and Feller [4] bear upon the almost sure convergence of $h_n(X)$ to $h(X)$. Thus, if X is uniform on U, and $D_n(X)$ tends to 0 almost surely, then the $Q^{(n)}$'s almost surely differentiate each Borel f on U for which $\int |f| (\log(1 + + |f|)^{d-1}\, dF$ is finite, that is, each $f \in L(\log L)^{d-1}$. (The condition is automatic if $d = 1$.) It would be enough to have f merely integrable if we could the ratios of sides of boxes which comprise the $Q^{(n)}$'s away from 0 and ∞ (which requirement would be inconsistent with the described invariance of our estimators to strictly monotonic transformations of the coordinates of X.)

REFERENCES

[1] BARNDORFF-NIELSEN, O. (1963). On the limit behavior of extreme order statistics. *Ann. Math. Statist.* **34**, 992–1002.

[2] BLACKWELL, D., AND DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33**, 882–886.

[3] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. (1984). *Classification and Regression Trees.* Wadsworth International, Belmont, CA.

[4] BUSEMANN, H., AND FELLER, W. (1934). Zur differentiation der Lebesgueschen Integrale. *Fund. Math.* **22**, 226–256.

[5] CHUNG, K. L., AND WALSH, J. B. (1969). To reverse a Markov process. *Acta Math.* **123**, 225–251.

[6] DEVROYE, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9**, 1310–1319.

[7] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[8] DUDLEY, R. M. (1980). Private communication.

[9] GARSIA, A. M. (1970). *Topics in Almost Everywhere Convergence.* Markham, Chicago.

[10] GEMAN, S. (1981). Sieves for nonparametric estimation of densities and regressions. In *Reports in Pattern Analysis* No. 99. Division of Applied Mathematics, Brown University, Providence.

[11] GEMAN, D., AND HWANG, C. R. (1982). Nonparametric maximum likelihood estimation by the mthod of sieves. *Ann. Statist.* **10** 401–414.

[12] GORDON, L., AND OLSHEN, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Ann. Statist.* **6** 515–533.

[13] GORDON, L., AND OLSHEN, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **10** 611–627.

[14] HUNT, G. A. (1966). *Martingales et Processus de Markov.* Dunod, Paris.

[15] JESSEN, B., MARCINKIEWICZ, J., AND ZYGMUND, A. (1935). Note on the differentiability of multiple integrals. *Fund. Math.* **25** 217–234.

[16] SHILOV, G. E., AND GUREVICH, B. L. (1977). *Integral, Measure, and Derivative: A Unified Approach.* Dover, New York.