Histogram Regression Estimation Using Data-dependent Partitions

Andrew Nobel [*]

September 1995

Abstract

We establish general sufficient conditions for the L_2-consistency of multivariate histogram regression estimates based on data-dependent partitions. These same conditions insure the consistency of partitioning regression estimates based on local polynomial fits, and, with an additional regularity assumption, the consistency of histogram estimates for conditional medians.

Our conditions require shrinking cells, subexponential growth of a combinatorial complexity measure, and sub-linear growth of restricted cell counts. It is not assumed that the cells of every partition be rectangles with sides parallel to the coordinate axis, or that each cell contain a minimum number of points. Response variables are assumed to be bounded throughout.

Our results may be applied to a variety of partitioning schemes. We establish the consistency of histogram regression estimates based on cubic partitions with data-dependent offsets, k-thresholding in one dimension, and empirically optimal nearest neighbor clustering schemes. In addition, it is shown that empirically optimal regression trees are consistent when the size of the trees grows with the number of samples at an appropriate rate.

Appears in Annals of Statistics, vol.24, pp.1084-1105

[*] Andrew Nobel is with the Dept. of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@stat.unc.edu . This work was completed while he was a Beckman Institute Fellow at the Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign.

AMS 1991 subject classifications: Primary 62G07; Secondary 62H30

Key words and phrases. Regression estimation, histogram estimates, regression trees, multivariate clustering, Vapnik-Chervonenkis theory, data-dependent partitions.

1 Introduction

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X, Y) \in \mathbb{R}^d \times [-K, K]$ be independent and identically distributed random vectors defined on a common probability space. We may view the random vector $X \in \mathbb{R}^d$ as a collection of measurements that are related in a stochastic fashion to a response variable $Y \in [-K, K]$, whose value is of interest. The joint distribution of (X, Y) is assumed to be unknown. We wish to estimate the regression function

$$r(x) = E(Y|X = x) \in [-K, K],$$

based on a training set of the form

$$T_n = (X_1, Y_1), \ldots, (X_n, Y_n). \tag{1}$$

Formally, a regression estimate is any function $\hat{f}_n(\cdot, T_n) : \mathbb{R}^d \to \mathbb{R}$ that depends on the training set. In most of what follows the dependence of $\hat{f}_n$ on T_n will be supressed. A sequence of estimates $\{\hat{f}_n\}$ is said to be *strongly L_2-consistent* if

$$\int |\hat{f}_n(x) - r(x)|^2 dP(x) \to 0 \ \text{ w.p.1},$$

where P denotes the distribution of the random vector X. When the expected value of the integral tends to zero, the estimates $\hat{f}_n$ are said to be *weakly* L_2 consistent.

When parametric models are not available, a natural means of estimating a multivariate regression function is to partition the observation space $\mathbb{R}^d$ into cells, and then form estimates locally, within each cell, based on the response variables. The simplest example of this are histogram estimates based on data-independent partitions that consist of infinitely many congruent, rectangular cells. Typically, the size and location of the cells depends only on the cardinality of T_n, not on the geometrical or numerical properties of its constituent vectors. If the common dimensions of the rectangles shrink at an appropriate rate, results of Stone (1977) establish the weak consistency of the associated regression estimates, regardless of the underlying distribution of the data. Devroye and Györfi (1985) established the strong consistency of regression estimates based on data-independent cubic partitions, and they derived exponential bounds for the L_1-error of the estimates.

While regression estimates based on data-independent partitions are easy to implement, statistical practice suggests that estimates based on suitably chosen *data-dependent* partitions will provide better small-sample performance. For example, if the measurement vectors $X_1, \ldots, X_n$

fall into two distinct clusters, a data-dependent partition could allocate the majority of its cells within these clusters. By considering the response variables Y_i a data-dependent partition can separate large and small values into separate cells, or allocate more cells to regions in which the behavior of the response variables is erratic. The flexibility of data-dependent partitions is likely to be most beneficial when the dimension d of the measurement variables is large and the sample size n is moderate.

In what follows we restrict our attention to measurable partitions having at most a countable number of cells. An *n-sample partitioning rule* is a deterministic mapping $\psi_n(\cdot)$ that associates each n-length sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ with a partition π of $\mathbb{R}^d$. When it is applied to a training set T_n, the rule ψ_n yields a random partition $\psi_n(T_n)$. The dependence of $\psi_n(T_n)$ on T_n will be suppressed when no confusion will arise. With this convention in mind, let $\psi_n[x]$ be the unique cell of $\psi_n(T_n)$ containing the vector x. Given a partitioning rule ψ_n and a training set T_n, we define the histogram regression estimate

$$\hat{r}_n(x) = \frac{\sum_{i=1}^n Y_i I\{X_i \in \psi_n[x]\}}{\sum_{i=1}^n I\{X_i \in \psi_n[x]\}} \tag{2}$$

for each $x \in \mathbb{R}^d$, with the convention that $\hat{r}_n(x) = 0$ when both the numerator and the denominator are zero. Note that $\hat{r}_n$ is piecewise constant on the cells of ψ_n, and that $|\hat{r}_n| \leq K$. Below we present general sufficient conditions for the consistency of such estimates, and then verify these conditions in a number of specific examples.

1.1 Discussion of Related Work

The simplest data-dependent partitioning methods are based on *statistically equivalent blocks* (Anderson (1966), Patrick and Fisher (1967)), in which each cell contains the same number of points. When $d = 1$ statistically equivalent blocks reduce to k-spacing estimates (Parthasarathy and Bhattacharya (1961)), where the k-th, $2k$-th,... order statistics of $X_1, \ldots, X_n$ determine the partition of the real line. For a discussion of these and other related partitioning rules in the context of pattern recognition, we refer to the survey paper of Devroye (1988).

Theoretical evidence for the superiority of data-dependent histogram methods is suggested by Stone (1985). Stone (1977) gave necessary and sufficient conditions for the weak L_2 consistency of regression estimates based on data-dependent local averages. His conditions apply to histogram estimates based on partitioning rules $\psi_n(X_1, \ldots, X_n)$ that do not make use of the response variables $Y_1, \ldots, Y_n$.

In many cases of practical interest, histogram regression estimates are described by binary trees. Regression trees are produced in an iterative fashion by recursive partitioning schemes that seek at each step to minimize an empirical criterion function. The consistency of regression trees produced by means of supervised axis-parallel splitting was established by Gordon and Olshen (1984, 1980). Gordon and Olshen (1978) established similar results for classification trees.

Sufficient conditions for the weak L_2-consistency of the estimates (2) considered here can be found in the book of Breiman, Friedman, Olshen, and Stone (1984). Their result may be summarized as follows.

Theorem A (Breiman *et al.***)** *A sequence $\psi_1, \psi_2, \ldots$ of partitioning rules gives rise to weakly L_2-consistent estimates $\hat{r}_n$ if*

a. Each cell of ψ_n contains at least $\alpha_n \log n$ points, where $\alpha_n \to \infty$;

b. There is a collection of sets $\mathcal{C}$ having finite Vapnik Chervonenkis dimension such that $\psi_n[x] \in \mathcal{C}$ for every x, every n, and every T_n.

c. $P\{x : \operatorname{diam}(\psi_n[x]) > \gamma\} \to 0$ with probability one for every $\gamma > 0$.

The collection $\mathcal{C}$ may consists of all d-dimensional rectangles with sides parallel to the coordinate axes, or more generally all polytopes having at most $M < \infty$ faces. Under slightly stronger conditions, Gordon and Olshen (1984) established the strong L_2-consistency of the estimates $\hat{r}_n$, and showed that $\hat{r}_n \to r$ almost surely when the partitions $\{\psi_n\}$ are nested. Chaudhuri, Huang, Loh, and Yao (1994) considered regression estimates that are constructed by fitting a polynomial of fixed degree within each cell of a data-dependent partition, based on a local least squares criterion. They gave sufficient conditions under which, with probability tending to one, such functions give uniformly good estimates of the regression function, and a prespecified number of its derivatives, on any fixed compact set. Their conditions are similar to those of Theorem A above, with the exception of an additional regularity assumption that insures the invertibility of the matrices giving the optimal coefficients within each cell. We note here that the work cited above applies to unbounded response variables Y_i satisfying various moment restrictions.

Conditions (a) and (b) restrict the applicability of Theorem A and related results. Condition (a) is difficult to verify when ψ_n is defined through minimization of an empirical criterion function, as in the tree-structured methods mentioned above. The required minimization does not generally insure that each cell contains a minimum number of points. Condition (b) requires that the cells of $\psi_n(T_n)$ be of fixed complexity, regardless of the sample size, while in practice analysis of the data may warrant increasing the cell complexity as n tends to infinity. In applying Theorem A to a particular method under study, satisfaction of (a), (b), and (c) is frequently accomplished by altering the method through supervisory oversight.

The results of this paper are based in part on combinatorial properties of partition families, and related exponential inequalities. In this respect we follow the work of Vapnik and Chervonenkis (1971, 1981). Application of these ideas to histogram estimation originated with Breiman *et al.* (1984) and in the later work of Zhao, Krishnaiah, and Chen (1991), who considered histogram density estimates based on data-dependent partitions with rectangular cells. Adopting an approach similar to that taken here, Lugosi and Nobel (1995) established general sufficient conditions for the consistency of histogram density estimates and classification rules based on finite, data dependent partitions.

1.2 Statement of Main Result

Let Π be a family of partitions of $\mathbb{R}^d$. For each set $V \subseteq \mathbb{R}^d$ define the restricted cell count

$$m(\Pi : V) = \max_{\pi \in \Pi} |\{A \in \pi : A \cap V \neq \emptyset\}| ,$$

where $|\{A \in \pi : A \cap V \neq \emptyset\}|$ measures the number of cells of π that intersect V. The unrestricted cell count is defined by $m(\Pi) = m(\Pi : \mathbb{R}^d)$. The complexity of Π is measured in terms of a combinatorial quantity proposed by Lugosi and Nobel (1995), which is similar to the growth function for classes of sets introduced by Vapnik and Chervonenkis (1971). Let $C = \{x_1, \ldots, x_n\}$ contain n vectors in $\mathbb{R}^d$. Every element $\pi = \{A_j\} \in \Pi$ induces a partition $\{A_j \cap C\}$ of the finite set C. Let $\Delta(x_1^n, \Pi)$ be the number of distinct partitions of C that are induced by the elements of Π (the order of appearance of the individual sets is not important). The partitioning number

$$\Delta_n^*(\Pi) = \max\{\Delta(x_1^n, \Pi) \,:\, x_1, \ldots, x_n \in \mathbb{R}^d\} \tag{3}$$

measures the maximum number of different partitions of any n point set that can be induced by members of Π.

Example 1: Let $\mathcal{U}_k$ be the family of all partitions of $\mathbb{R}$ into k non-empty intervals. Then

$m(\mathcal{U}_k) = k$, and for any sequence of numbers $x_1 < x_2 < \ldots < x_n$ an easy combinatorial argument shows that $\Delta(x_1^n, \mathcal{U}_k) = \binom{n+k-1}{n}$. It follows that $\Delta_n^*(\mathcal{U}_k) = \binom{n+k-1}{n}$.

Example 2: The *nearest-neighbor* partition of k vectors $c_1, \ldots, c_k \in \mathbb{R}^d$ has k cells $A_1, \ldots, A_k$. The cell A_i contains those vectors $x \in \mathbb{R}^d$ that are closer to c_i than any other c_j, with ties broken in favor of the vector having least index. Let $\mathcal{V}_k$ be the family of nearest-neighbor partitions of k vectors in $\mathbb{R}^d$. Then $m(\mathcal{V}_k) = k$. Moreover, it is readily verified that every cell of a partition $\pi \in \mathcal{V}_k$ is the intersection of $(k-1)$ halfspaces. It is well-known (c.f. Cover (1965)) that halfspaces in $\mathbb{R}^d$ can intersect n vectors $x_1, \ldots, x_n$ in at most n^d different ways. Thus each cell of a partition in $\mathcal{V}_k$ can intersect $x_1, \ldots, x_n$ in at most $n^{(k-1)d}$ different ways. As every partition of $\mathcal{V}_k$ contains k cells, it follows that $\Delta_n^*(\mathcal{V}_k) \leq n^{k^2 d}$.

Example 3: A *tree-structured partition* is described by a pair (T, τ), where T is a binary tree and $\tau : T \to \mathbb{R}^d$ is a node function that assigns a *test vector* in $\mathbb{R}^d$ to every $t \in T$. Every vector $x \in \mathbb{R}^d$ is associated, through a sequence of binary comparisons, with a descending path in T: beginning at the root, and at each subsequent internal node of T, x moves to that child of its current node whose test vector is nearest to x in Euclidean distance. In case of ties, x moves to the left child of its current node. For each $t \in T$ let U_t contain those vectors x whose path includes t. Then $U_t = \mathbb{R}^d$ when t is the root node of T, and if t is an internal node then U_t is split between its children by the hyperplane that forms the perpendicular bisector of their test vectors. If t is at distance k from the root, then U_t is a polytope having at most k faces. The partition generated by (T, τ) is comprised of the sets U_t associated with the leaves (terminal nodes) of T.

If at each internal node of T the comparison between the test vectors labeling its children is based on a single coordinate of x, then each cell of the resulting partition is a d-dimensional rectangle. Tree-structured partitions of this sort, based on axis-parallel splits, are the basis for the regression trees considered by Breiman *et al.* (1984).

Let $\mathcal{T}_k$ contain all the tree-structured partitions generated by binary trees T having k leaves. Clearly $m(\mathcal{T}_k) = k$. Consider an internal node t whose children are assigned test vectors u and v. Finding the test vector closest to x is equivalent to testing the membership of x in a closed halfspace that is bounded by the perpendicular bisector of u and v. A binary tree T with k leaves has $k-1$ internal nodes, and therefore each partition (T, τ) is based on at most $k-1$ intersecting halfspaces. As each of these halfspaces can dichotomize n points in at most n^d ways, their intersection can partition n points in at most $n^{(k-1)d}$ ways. Thus $\Delta_n^*(\mathcal{T}_k) \leq n^{(k-1)d}$.

Our principle result is stated below. Analogous conditions for the consistency of histogram classification and density estimates were established by Lugosi and Nobel (1995). Note that the range of a partitioning rule ψ_n is a deterministic family of partitions on $\mathbb{R}^d$.

Theorem 1 *Let $\psi_1, \psi_2, \ldots$ be fixed partitioning rules, and let Π_n be the range of ψ_n. Let the histogram regression estimate $\hat{r}_n$ be defined using ψ_n and T_n as in (2). Suppose that as n tends to infinity,*

(a) $n^{-1} m(\Pi_n : V) \to 0$ for every compact set $V \subseteq \mathbb{R}^d$,

(b) $n^{-1} \log \Delta_n^*(\Pi_n) \to 0$,

(c) For every $\gamma > 0$ and $\delta \in (0,1)$,

$$\inf_{S : P(S) \geq 1-\delta} P\{x : \operatorname{diam}(\psi_n[x] \cap S) > \gamma\} \to 0 \quad \textit{wp1}. \tag{4}$$

Then $\int |\hat{r}_n - r|^2 dP \to 0$ with probability one.

Thus a sequence of partitioning rules gives rise to consistent histogram regression estimates if the cells of the selected partitions shrink, and if the rules take values in a sequence of partition families whose restricted cell counts grow sub-linearly, and whose partitioning numbers grow sub-exponentially. Note that no assumptions are made on the distribution of the pair $(X, Y) \in \mathbb{R}^d \times [-K, K]$. The conclusions of the theorem remain valid when conditions (a) and (b) are replaced by their natural almost-sure equivalents.

Condition (a) of Theorem 1 is significantly weaker than the corresponding condition of Theorem A, which requires that each cell contain a minimum number of points. Condition (b) allows the partitions $\psi_n(T_n)$ to become more complex as the sample size n increases. In particular, the cells of each partition need not be rectangles, or polytopes with a fixed number of sides. Conditions (a) and (b) of Theorem A can be shown to imply condition (b) of Theorem 1.

In many cases of interest, the combinatorial conditions of Theorem 1 can be incorporated, without need of supervision, into the design of partitioning rules. In this case the principal task of analysis becomes verification of the shrinking cell condition (c).

1.3 Outline

A Vapnik-Chervonenkis inequality for partition families, and several of its consequences, are established in the next section. The proof of Theorem 1 is presented in Section 3. In Section 4 it is shown that the conditions of Theorem 1 insure the consistency of partitioning regression estimates based on local polynomial fits of fixed degree. Sufficient conditions for the consistency of histogram conditional quantile estimates are presented in Section 5.

The conditions of Theorem 1 are applicable to a variety of partitioning rules. Cubic partitions with data-dependent offsets are considered in Section 6. Section 7 is devoted to regression estimates in one dimension based on k-thresholding, a generalization of ordinary k-spacing in which the partitioning rule depends on the response variables Y_i.

Multivariate clustering schemes provide natural partitioning rules for histogram regression. It is shown in Section 8 that estimates based on empirically optimal nearest-neighbor clustering schemes are consistent when the number of cluster centers grows with the size of the training set at an appropriate rate.

In Section 9 it is shown that empirically optimal regression trees are consistent when the size k_n of the tree grows as $o(n/\log n)$.

2 Preliminary Results

A non-empty collection $\mathcal{F}$ of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ is said to be uniformly bounded with envelope L if $|f(x)| \leq L$ for every $x \in \mathbb{R}^d$ and every $f \in \mathcal{F}$. For each $\epsilon > 0$ and each sequence of vectors $x_1, \ldots, x_n$, the covering number $N(x_1^n, \epsilon, \mathcal{F})$ is the size of the smallest collection $\mathcal{G}$ such that

$$\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)| < \epsilon \tag{5}$$

for every $f \in \mathcal{F}$. Any collection $\mathcal{G}$ satisfying (5) is said to be an ϵ-cover for $\mathcal{F}$ on x_1^n. If no finite ϵ-cover exists then $N(x_1^n, \epsilon, \mathcal{F}) = \infty$. A class $\mathcal{F}$ will be called *nice* if it is uniformly bounded, and there is a function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$N(x_1^n, \epsilon, \mathcal{F}) \leq \phi(\epsilon) \tag{6}$$

for every $\epsilon > 0$, every n, and every sequence $x_1, \ldots, x_n \in \mathbb{R}^d$. When (6) holds, ϕ will be said to majorize the covering numbers of $\mathcal{F}$.

Definition: Given a class of real-valued functions $\mathcal{G}$ on $\mathbb{R}^d$ and a partition family Π, define

$$\mathcal{G} \circ \Pi = \left\{ f = \sum_{A_j \in \pi} g_j I_{A_j} : \pi = \{A_j\} \in \Pi,\, g_j \in \mathcal{G} \right\}.$$

Each function f in $\mathcal{G} \circ \Pi$ is obtained by applying a different function $g \in \mathcal{G}$ within each cell of a selected partition $\pi \in \Pi$.

Consider a sequence $X_1, X_2, \ldots \in \mathbb{R}^d$ of independent random vectors having a common distribution P. Let $\hat{P}_n$ be the empirical distribution of $X_1, \ldots, X_n$. For every bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$ define

$$Pf = \int f(x) dP(x) ,$$

and

$$\hat{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) .$$

Our results rely on exponential inequalities concerning the uniform deviation of P and $\hat{P}_n$ over suitable classes of functions. Given a uniformly bounded class of functions $\mathcal{F}$ on $\mathbb{R}^d$, define

$$\Lambda_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} |\hat{P}_n f - Pf| . \tag{7}$$

If $\mathcal{H}$ is a uniformly bounded class of functions $h : \mathbb{R}^d \times [-K, K] \to \mathbb{R}$ let

$$\tilde{\Lambda}_n(\mathcal{H}) = \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i) - Eh(X, Y) \right| .$$

Remark: To insure measurability of the suprema $\Lambda_n(\mathcal{F})$, we assume that every class $\mathcal{F}$ under consideration contains a countable subclass $\mathcal{F}_0$ with the property that every function in $\mathcal{F}$ is the pointwise limit of a sequence of functions in $\mathcal{F}_0$. This condition may be extended to partition families by viewing each partition π as a mapping from $\mathbb{R}^d$ into the set of natural numbers.

Proposition 1 *Let $\mathcal{G}$ be a class of functions on $\mathbb{R}^d$ whose covering numbers are majorized by $\phi(\cdot)$, and let Π be any partition family with $m(\Pi) < \infty$. For each sequence $x_1, \ldots, x_n \in \mathbb{R}^d$ and every $\epsilon > 0$,*

$$N(x_1^n, \epsilon, \mathcal{G} \circ \Pi) \ \le \ \Delta(x_1^n, \Pi)\, \phi(\epsilon)^{m(\Pi)} \ \le \ \Delta_n^*(\Pi)\, \phi(\epsilon)^{m(\Pi)} .$$

Proof: Fix $x_1, \ldots, x_n \in \mathbb{R}^d$ and $\epsilon > 0$. Call two elements $\pi, \pi' \in \Pi$ equivalent if they induce the same partition of $x_1, \ldots, x_n$. If $f \in \mathcal{G} \circ \Pi$ then there is a partition $\pi = \{A_j\} \in \Pi$ and functions $g_j \in \mathcal{G}$ such that

$$f = \sum_{A_j \in \pi} g_j I_{A_j} . \tag{8}$$

For each j let $\mathcal{F}_j$ be an ϵ-cover for $\mathcal{G}$ on $A'_j = \{x_1, \ldots, x_n\} \cap A_j$ such that $|\mathcal{F}_j| \le \phi(\epsilon)$. To each function g_j appearing in (8) there corresponds an approximating function $f_j \in \mathcal{F}_j$ such that

$$\frac{1}{n_j} \sum_{x_i \in A'_j} |g_j(x_i) - f_j(x_i)| < \epsilon,$$

where $n_j = |A'_j|$. If we define $f' = \sum_{A_j \in \pi} f_j I_{A_j}$, then it is easy to see that

$$\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f'(x_i)| < \epsilon.$$

For suitably chosen $f_j \in \mathcal{F}_j$, every function $\tilde{f} \in \mathcal{G} \circ \Pi$ defined in terms of a partition equivalent to π can be approximated by a similar estimate f'. Thus the collection of all such functions $\tilde{f}$ can be covered on x_1^n by no more than $\Pi_{j=1}^{|\pi|} |\mathcal{F}_j| \le \phi(\epsilon)^{|\pi|}$ approximating functions. As the partitions in Π fall into at most $\Delta(x_1^n, \Pi)$ equivalence classes, the result follows.

An application of the basic Vapnik Chervonenkis inequality for classes of functions (cf. Vapnik and Chervonenkis (1981) or Pollard (1984)) gives the following bound.

Lemma 1 *Let $\mathcal{G}$ be a class of functions with envelope L whose covering numbers are majorized by $\phi(\cdot)$. If Π is a partition family for which $m(\Pi) < \infty$, then for every $t > 0$,*

$$\mathbb{P}\left\{\Lambda_n(\mathcal{G} \circ \Pi) > t\right\} \;\le\; \Delta_n^*(\Pi)\, \phi(t)^{m(\Pi)} \exp\left[\frac{-nt^2}{32L^2}\right].$$

□

Proposition 2 *For every uniformly bounded class $\mathcal{G}$, and every partition family Π,*

$$\sup_{g \in \mathcal{G}} \sup_{\pi \in \Pi} \sum_{A \in \pi} |\hat{P}_n(gI_A) - P(gI_A)| \;\le\; \Lambda_n(\mathcal{G}' \circ \Pi)\,,$$

where $\mathcal{G}' = \mathcal{G} \cup (-\mathcal{G})$ contains every function $g \in \mathcal{G}$ and its additive inverse.

Proof: Fix a function $g \in \mathcal{G}$ and a partition $\pi \in \Pi$. Given $X_1 = x_1, \ldots, X_n = x_n$ define

$$\pi_1 = \{A \in \pi : \hat{P}_n(gI_A) \ge P(gI_A)\} \;\text{ and }\; \pi_2 = \{A \in \pi : \hat{P}_n(gI_A) < P(gI_A)\}\,.$$

Then it is easy to see that

$$\begin{aligned} &\sum_{A \in \pi} |\hat{P}_n(gI_A) - P(gI_A)| \\ &\quad = \left| \sum_{A \in \pi_1} (\hat{P}_n(gI_A) - P(gI_A)) - \sum_{A \in \pi_2} (\hat{P}_n(gI_A) - P(gI_A)) \right| \\ &\quad = |\hat{P}_n f - Pf|\,, \end{aligned}$$

where $f = \sum_{A \in \pi_1} g I_A + \sum_{A \in \pi_2} (-g) I_A$ is an element of $\mathcal{G}' \circ \Pi$. Consequently,

$$\sum_{A \in \pi} |\hat{P}_n(gI_A) - P(gI_A)| \leq \Lambda_n(\mathcal{G}' \circ \Pi)\,,$$

and the result follows as $g \in \mathcal{G}$ and $\pi \in \Pi$ were arbitrary.

The following inequality extends Lemma 1 of Lugosi and Nobel (1995). Though it will not be needed in what follows, it may be of independent interest.

Lemma 2 *Let $\mathcal{G}$ be a class of functions with envelope L whose covering numbers are majorized by $\phi(\cdot)$. If Π is a partition family with $m(\Pi) < \infty$, then for every $t > 0$,*

$$\mathrm{IP}\left\{ \sup_{g \in \mathcal{G}} \sup_{\pi \in \Pi} \sum_{A \in \pi} |\hat{P}_n(gI_A) - P(gI_A)| > t \right\} \;\leq\; 2^{m(\Pi)}\, \Delta_n^*(\Pi)\, \phi(t)^{m(\Pi)} \,.\, \exp\left[\frac{-nt^2}{32L^2} \right]$$

Proof: As $N(x_1^n, t, \mathcal{G}') \;\leq\; 2N(x_1^n, t, \mathcal{G}) \;\leq\; 2\phi(t)$, the stated inequality follows immediately from Lemma 1 and Proposition 2. □

Lemma 3 *Let $\mathcal{G}$ be a nice class of functions, and let $\Pi_1, \Pi_2, \ldots$ be a sequence of partition families. If*

$$n^{-1} m(\Pi_n : V) \to 0 \tag{9}$$

for every compact set $V \subset \mathrm{IR}^d$, and

$$n^{-1} \log \Delta_n^*(\Pi_n) \to 0\,, \tag{10}$$

then

$$\Lambda_n(\mathcal{G} \circ \Pi_n) \to 0 \tag{11}$$

with probability one.

Proof: If the unrestricted cell counts are such that $n^{-1} m(\Pi_n) \to 0$, then (11) is an immediate consequence of Lemma 1 and the Borel-Cantelli Lemma. In the more general case, fix $\epsilon > 0$ and let $L < \infty$ be an envelope for $\mathcal{G}$. Select a compact set $V \subseteq \mathrm{IR}^d$ such that $L \cdot P(V^c) \leq \epsilon$, and define the class of functions $\mathcal{G}_V = \{gI_V : g \in \mathcal{G}\}$, which is readily seen to be nice. For each partition $\pi = \{A_j\} \in \Pi_n$ define its restriction $\pi' = \{A_j \cap V\} \cup \{V^c\}$, and let

$$\Pi_n' = \{\pi' : \pi \in \Pi_n\}\ .$$

It is easily verified that $\Delta_n^*(\Pi_n') \leq \Delta_n^*(\Pi_n)$, and that $m(\Pi_n') \leq m(\Pi_n : V) + 2$. By virtue of the unrestricted case above, $\Lambda_n(\mathcal{G}_V, \Pi_n') \to 0$ with probability one. Now note that

$$|\Lambda_n(\mathcal{G}, \Pi_n) - \Lambda_n(\mathcal{G}_V, \Pi_n)| \leq L\hat{P}_n(V^c) + LP(V^c)\,,$$

and as $\Lambda_n(\mathcal{G}_V, \Pi_n) = \Lambda_n(\mathcal{G}_V, \Pi_n')$ for each n,

$$\begin{aligned}\limsup_{n\to\infty} \Lambda_n(\mathcal{G}, \Pi_n) &\leq \limsup_{n\to\infty} \Lambda_n(\mathcal{G}_V, \Pi_n) + 2\epsilon \\ &= \limsup_{n\to\infty} \Lambda_n(\mathcal{G}_V, \Pi_n') + 2\epsilon \\ &= 2\epsilon\end{aligned}$$

with probability one. Since $\epsilon > 0$ was arbitrary, the result follows. □

Lemma 4 *Let $\mathcal{G}$ be a nice class of functions with envelope L, and let $\Pi_1, \Pi_2, \ldots$ be partition families satisfying (9) and (10). Let $\{\hat{f}_n\}$ and $\{\hat{g}_n\}$ be regression estimates such that for each n and each training sequence T_n,*

a. $\hat{f}_n(\cdot, T_n)$ and $\hat{g}_n(\cdot, T_n)$ lie in $\mathcal{G} \circ \Pi_n$,

b. $\sum_{i=1}^n (\hat{f}_n(X_i) - Y_i)^2 \leq \sum_{i=1}^n (\hat{g}_n(X_i) - Y_i)^2$.

Then $\{\hat{f}_n\}$ is strongly consistent if $\{\hat{g}_n\}$ is strongly consistent.

Proof: It is well known, and easy to verify, that for every bounded function $f : \mathbb{R}^d \to \mathbb{R}$,

$$P|f - r|^2 = E|f(X) - Y|^2 - E|r(X) - Y|^2\,. \tag{12}$$

For each element $\pi = \{A_j\} \in \Pi_n$ define an associated partition $\tilde{\pi} = \{A_j \times [-K, K]\}$ of $\mathbb{R}^d \times [-K, K]$, and let $\tilde{\Pi}_n = \{\tilde{\pi} : \pi \in \Pi_n\}$. It is readily verified that

$$\Delta_n^*(\tilde{\Pi}_n) = \Delta_n^*(\Pi_n) \;\text{ and }\; m(\tilde{\Pi}_n : V \times [-K, K]) = m(\Pi_n : V) \tag{13}$$

for every compact subset V of $\mathbb{R}^d$. Let $\mathcal{H}$ contain all those functions $h : \mathbb{R}^d \times [-K, K] \to \mathbb{R}$ of the form $h(x, y) = (g(x) - y)^2$, where $g \in \mathcal{G}$. Then $\mathcal{H}$ is nice, and it follows from Lemma 3 and the equations (13) that

$$\sup_{f \in \mathcal{G}\circ\Pi_n} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 - E(f(X) - Y)^2 \right| = \tilde{\Lambda}_n(\mathcal{H} \circ \tilde{\Pi}_n) \to 0 \tag{14}$$

with probability one. Now consider the estimates $\{\hat{f}_n\}$ and $\{\hat{f}_n\}$. By virtue of (14) and the assumptions above,

$$\begin{aligned}
P|\hat{f}_n - r|^2 &= E|\hat{f}_n(X) - Y|^2 - E|r(X) - Y|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i) - Y_i)^2 - E|r(X) - Y|^2 + \tilde{\Lambda}_n(\mathcal{H} \circ \tilde{\Pi}_n) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n(X_i) - Y_i)^2 - E|r(X) - Y|^2 + \tilde{\Lambda}_n(\mathcal{H} \circ \tilde{\Pi}_n) \\
&\leq E|\hat{g}_n(X) - Y|^2 - E|r(X) - Y|^2 + 2\tilde{\Lambda}_n(\mathcal{H} \circ \tilde{\Pi}_n) \\
&= P|\hat{g}_n - r|^2 + 2\tilde{\Lambda}_n(\mathcal{H} \circ \tilde{\Pi}_n)\,.
\end{aligned}$$

If the sequence $\{\hat{g}_n\}$ is strongly consistent, then the last term above tends to zero with probability one. □

3 Proof of Theorem 1

Definition: A sequence $Z_1, Z_2, \ldots$ of random variables defined on the same probability space as (X, Y) is said to be of order $o^*(1)$, written $Z_n = o^*(1)$, if as n tends to infinity $Z_n \to 0$ with probability one.

Proof of Theorem 1: Define an auxilliary function

$$\tilde{r}_n(x) = \frac{\sum_{i=1}^{n} r(X_i) I\{X_i \in \psi_n[x]\}}{\sum_{i=1}^{n} I\{X_i \in \psi_n[x]\}}\,. \tag{15}$$

Note that $|\tilde{r}_n| \leq K$, and that $\tilde{r}_n$ is piecewise constant on the cells of ψ_n. By an obvious upper bound,

$$P|r - \hat{r}_n|^2 \leq 2P|\tilde{r}_n - \hat{r}_n|^2 + 2P|r - \tilde{r}_n|^2\,. \tag{16}$$

The first term above measures the variance of $\hat{r}_n$, while the second term measures its bias. We show in turn that each is of order $o^*(1)$.

Consider the first term in (16). By an easy application of Proposition 2 and Lemma 3,

$$\sup_{\pi \in \Pi_n} \sum_{A \in \pi} |\hat{P}_n(A) - P(A)| = o^*(1). \tag{17}$$

For each cell $A \in \psi_n(T_n)$ let $n_A = n\hat{P}_n(A)$ be the number of vectors X_i in A. By virtue of (17) and the definition of $\tilde{r}_n$,

$$P|\hat{r}_n - \tilde{r}_n|^2 \leq 2K \cdot \sum_{A \in \psi_n} \left| \frac{1}{n_A} \sum_{X_i \in A} (Y_i - r(X_i)) \right| P(A)$$

$$
\begin{aligned}
&\leq \; 2K \cdot \sum_{A \in \psi_n} \left| \frac{1}{n_A} \sum_{X_i \in A} (Y_i - r(X_i)) \right| \hat{P}_n(A) + o^*(1) \\
&= \; 2K \cdot \sum_{A \in \psi_n} \left| \frac{1}{n} \sum_{X_i \in A} (Y_i - r(X_i)) \right| + o^*(1) \\
&\leq \; 2K \cdot \sup_{\pi \in \Pi_n} \sum_{A \in \pi} \left| \frac{1}{n} \sum_{X_i \in A} (Y_i - r(X_i)) \right| + o^*(1) \,. \qquad (18)
\end{aligned}
$$

Let Q denote the joint distribution of (X, Y) and let $\hat{Q}_n$ be the empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. For each element $\pi = \{A_j\} \in \Pi_n$ define an associated partition $\tilde{\pi} = \{A_j \times [-K, K]\}$ of the set $\mathbb{R}^d \times [-K, K]$, and let $\tilde{\Pi}_n = \{\tilde{\pi} : \pi \in \Pi_n\}$. The function $h(x, y) = y - r(x)$, defined for $(x, y) \in \mathbb{R}^d \times [-K, K]$, is bounded and satisfies $\int_{A \times \mathbb{R}} h(x, y) dQ = 0$ for every measurable subset A of $\mathbb{R}^d$. Rewriting (18) in terms of the distribution Q, the family $\tilde{\Pi}_n$, and h gives

$$
\begin{aligned}
P|\hat{r}_n - \tilde{r}_n|^2 \;\; &\leq \;\; 2K \cdot \sup_{\tilde{\pi} \in \tilde{\Pi}_n} \sum_{B \in \tilde{\pi}} |\hat{Q}_n(hI_B) - Q(hI_B)| + o^*(1) \\
&\leq \;\; 2K \cdot \tilde{\Lambda}_n(\{h, -h\} \circ \tilde{\Pi}_n) + o^*(1) \,,
\end{aligned}
$$

where the second inequality follows from Proposition 2. By virtue of (13) and the conditions of the theorem the first term above is of order $o^*(1)$. Thus $P|\hat{r}_n - \tilde{r}_n| = o^*(1)$.

Consider now the second term of (16). Fix $\epsilon > 0$ and let $g : \mathbb{R}^d \to [-K, K]$ be a uniformly continuous function such that $P|r - g|^2 < \epsilon$. Let $\gamma > 0$ be chosen so that $|g(x_1) - g(x_2)| < \epsilon^{1/2}$ whenever $\|x_1 - x_2\| < \gamma$. Fix $\delta > 0$ so small that $8K^2\delta \leq \epsilon$, and let S be any set satisfying $P(S) \geq 1 - \delta$. Define $\bar{g}_n$ by averaging g over those X_i within the cells of $\psi_n \cap S$,

$$
\bar{g}_n(x) = \frac{\sum_{i=1}^n g(X_i) I\{X_i \in \psi_n[x] \cap S\}}{\sum_{i=1}^n I\{X_i \in \psi_n[x] \cap S\}} \,.
$$

Both $\tilde{r}_n$ and $\bar{g}_n$ are constant on the cells of ψ_n. In particular, the definition of $\tilde{r}_n$ insures that

$$
\sum_{X_i \in A} (r(X_i) - \tilde{r}_n(X_i))^2 \; \leq \; \sum_{X_i \in A} (r(X_i) - \bar{g}_n(X_i))^2
$$

for each cell $A \in \psi_n$, so that for each training set T_n,

$$
\hat{P}_n|r - \tilde{r}_n|^2 \; \leq \; \hat{P}_n|r - \bar{g}_n|^2 \,. \qquad (19)
$$

Let $\mathcal{G}$ contain all the functions of the form $g(x) = (r(x) - a)^2$, where $a \in [-K, K]$. Then $\mathcal{G}$ is nice, and it is evident that $|r - \tilde{r}_n|^2$ and $|r - \tilde{g}_n|^2$ are contained in $\mathcal{G} \circ \Pi_n$ for each n.

Applying Lemma 3 to the sequence $\{\mathcal{G} \circ \Pi_n\}$, one deduces that

$$\hat{P}_n|r - \tilde{r}_n|^2 - P|r - \tilde{r}_n|^2 = o^*(1) \text{ and } \hat{P}_n|r - \tilde{g}_n|^2 - P|r - \bar{g}_n|^2 = o^*(1). \tag{20}$$

It follows from (19), (20), and the choice of S that

$$\begin{aligned} P|r - \tilde{r}_n|^2 &\leq \hat{P}_n|r - \tilde{r}_n|^2 + o^*(1) \\ &\leq \hat{P}_n|r - \bar{g}_n|^2 + o^*(1) \\ &\leq P|r - \bar{g}_n|^2 + o^*(1) \\ &\leq 2P|r - g|^2 + 2P|g - \bar{g}_n|^2 + o^*(1) \\ &\leq 2P|g - \bar{g}_n|^2 + 2\epsilon + o^*(1) \\ &= 2P|g - \bar{g}_n|^2 I_S + 2P|g - \bar{g}_n|^2 I_{S^c} + 2\epsilon + o^*(1) \\ &\leq 2P|g - \bar{g}_n|^2 I_S + 3\epsilon + o^*(1)\,. \end{aligned}$$

If $x \in S$ and $\operatorname{diam}(\psi_n[x] \cap S) < \gamma$ then $|g(x) - \bar{g}_n(x)|^2 < \epsilon$. Therefore

$$P|r - \tilde{r}_n|^2 \leq 8K^2 \cdot P\{x : \operatorname{diam}(\psi_n[x] \cap S) \geq \gamma\} + 4\epsilon + o^*(1)\,.$$

As the right-hand side depends on S only through its probability,

$$P|r - \tilde{r}_n|^2 \leq \inf_{S : P(S) \geq 1-\delta} 8K^2 \cdot P\{\operatorname{diam}(\psi_n[x] \cap S) > \gamma\} + 4\epsilon + o^*(1)\,.$$

Condition (c) of the theorem guarantees that $P|r - \tilde{r}_n|^2 \leq 4\epsilon + o^*(1)$, and the result follows as $\epsilon > 0$ was arbitrary. □

4 Local Fitting with Truncated Polynomials

The histogram estimate $\hat{r}_n$ is piecewise constant on the cells of $\psi_n(T_n)$ and has discontinuities along the boundaries of these cells. If the regression function $r(\cdot)$ is known to be smooth, or if the sample size is large, more sophisticated local estimates may be appropriate.

The results of Theorem 1 can be extended in a natural way to regression estimates based on local polynomial fits. Fitting a suitably truncated polynomial within each cell of ψ_n gives estimates with better approximation capabilities, and potentially better performance. Discontinuities along cell boundaries remain, but may be corrected if necessary by employing weighted averages of estimates from different cells, as in Chaudhuri *et al.* (1994)

For each vector $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$ and each sequence $\alpha = (\alpha_1, \ldots, \alpha_d)$ of non-negative integers, let $u^\alpha = u_1^{\alpha_1} \cdots u_d^{\alpha_d}$ and $|\alpha| = \alpha_1 + \ldots + \alpha_d$. Fix an integer $k \geq 0$ and let

$$\mathcal{G} = \left\{ g(x) = \sum_{|\alpha| \leq k} c(\alpha) x^\alpha \ : \ c(\alpha) \in \mathbb{R} \right\}$$

to be the class of k'th order multivariate polynomials on $\mathbb{R}^d$. Set $L = Kb$ for some $b \geq 1$ and define the class

$$\tilde{\mathcal{G}} = \{ g(\cdot) \wedge l \vee (-l) \ : \ g \in \mathcal{G} \, , \, 0 \leq l \leq L \}$$

of truncated polynomials which has envelope L. As $\mathcal{G}$ is a finite dimensional vector space of real-valued functions, it follows from Lemma 28 of Pollard (1984, Chapter 2) that $\mathcal{G}$ is a VC-graph class, and an easy argument shows that the same is true for $\tilde{\mathcal{G}}$. Lemma 25 of Pollard (1984, Chapter 2) shows that $\mathcal{G}$ is nice.

Given a partitioning rule ψ_n and a training set T_n, we construct a piecewise polynomial regression estimate by fitting a suitable function $g \in \tilde{\mathcal{G}}$ within each cell of $\psi_n(T_n)$. For each cell $A \in \psi_n(T_n)$ let

$$g_A = \arg\min_{g \in \mathcal{G}} \sum_{X_i \in A} (g(X_i) - Y_i)^2 \tag{21}$$

be the best k'th order polynomial fit to those pairs (X_i, Y_i) for which $X_i \in A$, and set

$$l_A = b \cdot \max\{ |Y_i| \ : \ X_i \in A \} \, .$$

If the range $[-K, K]$ of the response variables Y_i is known, then define the estimate

$$\hat{f}_n(x) = \sum_{A \in \psi_n} g_A(x) \wedge K \vee (-K) \cdot I\{x \in A\} \, ; \tag{22}$$

otherwise define

$$\hat{f}_n(x) = \sum_{A \in \psi_n} g_A(x) \wedge l_A \vee (-l_A) \cdot I\{x \in A\} \, . \tag{23}$$

In either case, $\hat{f}_n(x) \in \tilde{\mathcal{G}} \circ \Pi_n$. Moreover the optimality of g_A and the choice of trunction level insures that

$$\sum_{i=1}^n (\hat{f}_n - Y_i)^2 \leq \sum_{i=1}^n (\hat{r}_n - Y_i)^2$$

for every training set T_n. The consistency of the estimates $\{\hat{f}_n\}$ follows immediately from Theorem 1 and Lemma 4. Note that truncation eliminates the need for regularity assumptions concerning the local least-squares fit within each cell, assumptions that are difficult to verify in practice.

Theorem 2 *Let $\psi_1, \psi_2, \ldots$ be a sequence of partitioning rules satisfying conditions* (a)*,* (b)*, and* (c) *of Theorem 1. If regression estimates $\hat{f}_n$ are defined using truncated local polynomial fits as in (22) or (23), then $P|\hat{f}_n - r|^2 \to 0$ with probability one.*

When it is desireable to do so, we may center the polynomial fit within each cell, replacing $g(X_i)$ in (21) by $g(X_i - x_0)$, where x_0 is the average of those vectors lying in A, or some other centrally located point of A.

5 Estimation of Conditional Medians

For each $x \in \mathbb{R}^d$ and $a \in [-K, K]$ the conditional cumulative distribution function of Y given $X = x$ is given by

$$F(a, x) = \mathbb{P}\{Y \leq a | X = x\}\,.$$

The conditional median of Y given $X = x$ is defined by

$$m(x) = \inf\{a : F(a, x) \geq 1/2\}\,.$$

Here we study histogram estimates of $m(\cdot)$ that are based on local order statistics of the response variables.

Given a partitioning rule ψ_n and a training set $T_n = (X_1, Y_1), \ldots, (X_n, Y_n)$, define for each x a histogram estimate

$$\hat{m}_n(x) = \inf\{a : \hat{F}_n(a, x) \geq 1/2\}$$

of the conditional median $m(x)$, in terms of the corresponding estimate

$$\hat{F}_n(a, x) = \frac{\sum_{i=1}^n I\{Y_i \leq a\} I\{X_i \in \psi_n[x]\}}{\sum_{i=1}^n I\{X_i \in \psi_n[x]\}}$$

of the conditional cumulative distribution function. Then $\hat{m}_n(x)$ is just the $\lfloor (k+1)/2 \rfloor$'th order statistic among the k numbers $\{Y_i : X_i \in \psi_n[x]\}$. In particular, if k is odd then $\hat{m}_n(x)$ is the ordinary median of $\{Y_i : X_i \in \psi_n[x]\}$.

Theorem 3 *Let $\psi_1, \psi_2, \ldots$ be partitioning rules satisfying conditions (a), (b), and (c) of Theorem 1. If for P-almost every $x \in \mathbb{R}^d$*

$$\mathbb{P}\{m(x) < Y \leq m(x) + \epsilon \,|\, X = x\} > 0 \;\; \textit{for every } \epsilon > 0\,, \tag{24}$$

then $P|\hat{m}_n - m| \to 0$ with probability one.

Proof: Fix $\epsilon > 0$ and note that $P|\hat{m}_n - m| \leq 2\epsilon + 2KP(A_n) + 2KP(B_n)$, where the sets

$$A_n = \{x : m(x) \geq \hat{m}_n(x) + 2\epsilon\} \text{ and } B_n = \{x : \hat{m}_n(x) \geq m(x) + 2\epsilon\}.$$

depend on the training set T_n through $\hat{m}_n$. It is enough to show that $P(A_n)$ and $P(B_n)$ are of order $o^*(1)$.

Fix a number $a \in [-K, K]$ and define the indicator random variable $Y_i' = I\{Y_i \leq a\}$ for each $i \geq 1$. The regression function of Y_i' is $F(a, x)$, and its histogram estimate is just $\hat{F}_n(a, x)$. Although ψ_n is a function of T_n rather than $(X_1, Y_1'), \ldots, (X_n, Y_n')$, the analysis of Theorem 1 still applies: for each $a \in [-K, K]$,

$$\int |\hat{F}_n(a, x) - F(a, x)| \, dP(x) \to 0 \tag{25}$$

with probability one.

Select numbers $-(K + \epsilon) = a_0 < a_1 < \ldots < a_l < a_{l+1} = (K + \epsilon)$ such that $a_{i+1} - a_i < \epsilon$ for $i = 0, \ldots, l$. For each $x \in \mathbb{R}^d$ let $j(x)$ be the unique integer in $\{1, \ldots, l - 1\}$ such that $m(x) \in [a_{j(x)}, a_{j(x)+1})$, and define $h(x) = 1/2 - F(a_{j(x)-1}, x)$. The definition of $m(x)$ insures that $P\{h > 0\} = 1$. If $x \in A_n$ then $\hat{m}_n(x) \leq m(x) - 2\epsilon \leq a_{j(x)-1}$, and therefore

$$\hat{F}_n(a_{j(x)-1}, x) - F(a_{j(x)-1}, x) \geq 1/2 - F(a_{j(x)-1}, x) = h(x). \tag{26}$$

For each $\delta > 0$ define $V_\delta = \{x : h(x) \geq \delta\}$. By virtue of (26),

$$\begin{aligned} \int \max_{0 \leq j \leq l} |\hat{F}_n(a_j, x) - F(a_j, x)| dP(x) &\geq \int_{A_n} h(x) dP(x) \\ &\geq \delta P(A_n \cap V_\delta) \\ &\geq \delta[P(A_n) - P(V_\delta^c)], \end{aligned}$$

and it follows from (25) that for every $\delta > 0$,

$$\limsup_{n \to \infty} P(A_n) \leq P(V_\delta^c).$$

with probability one. As $\delta \to 0$, the probability $P(V_\delta^c) \to P\{x : h(x) = 0\} = 0$, so that $P(A_n) = o^*(1)$. A similar analysis may be carried out for the events $\{B_n\}$ using the regularity condition (24), which insures that $F(a, x)$ is increasing in each neighborhood of $a = m(x)$. □

Remark: An obvious modification of the preceding argument establishes the consistency of histogram estimates for the conditional quantiles $m_\alpha(x)$, with $\alpha \in (0, 1)$. One need only require that (24) hold at P-almost every point $m_\alpha(x)$.

6 Cubic Partitions with Data-dependent Offsets

A cubic partition has congruent, rectangular cells whose dimensions are specified in advance of the data. Adding an offset vector to each cell of a cubic partition shifts the cells in the direction of the vector, while maintaining their axis-parallel orientation. An application of Theorem 1 yields conditions for the consistency of regression estimates based on cubic partitions with data-dependent offsets.

Fix a sequence $\{h_{jn}\}$ of positive numbers for each $j = 1, \ldots, d$. For every $n \geq 1$ define the infinite cubic partition

$$\theta_n = \{ \ [(r_1 - 1)h_{1n} \, , \, r_1 h_{1n}) \times \cdots \times [(r_d - 1)h_{dn} \, , \, r_d h_{dn}) \ : \ r_1, \ldots, r_d \in \mathbb{Z} \, \} \ .$$

having congruent, rectangular cells with edge lengths $h_{1n}, \ldots, h_{dn}$. Let $\Pi_n = \{\theta_n + c : c \in \mathbb{R}^d\}$ contain all the shifts of θ_n by an arbitrary offset c.

Suppose that $d = 1$, in which case each cell of θ_n is an interval of length h_n. Fix $x_1, \ldots, x_n \in \mathbb{R}$. If $\pi \in \Pi_n$ then π and $\pi + h_n$ are identical, so in assessing the partitioning number $\Delta(x_1^n, \Pi_n)$ it is enough to consider shifts $c \in [0, h_n)$ of length at most h_n. As c increases from 0 to h_n, the partition induced by $\theta_n + c$ changes only if the boundary of a cell crosses one of the points $x_1, \ldots, x_n$. This happens exactly n times, once for each x_i, and it follows that $\Delta(x_1^n, \Pi_n) \leq n$.

Suppose now that $d > 1$. For each $j = 1, \ldots, d$ let the family Π_n^j be generated by shifts of the partition

$$\theta_n^j = \{ \ \mathbb{R} \times \cdots \times [(r_j - 1)h_{jn} \, , \, r_j h_{jn}) \times \cdots \times \mathbb{R} \ : \ r_j \in \mathbb{Z} \, \} \ ,$$

each of whose cells is a slab perpendicular to the j'th coordinate axis. By varying the components of c one at a time, it can be shown that for each sequence $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\Delta(x_1^n, \Pi_n) \leq \Delta_n^*(\Pi_n^1) \cdots \Delta_n^*(\Pi_n^d) \, .$$

It follows from the case $d = 1$ above that $\Delta_n^*(\Pi_n^j) \leq n$ for each j, and therefore the partitioning numbers $\Delta_n^*(\Pi_n) \leq n^d$ satisfy condition (b) of Theorem 1.

The other conditions of Theorem 1 lead to restrictions on the constants h_{jn} that are identical to those required for fixed cubic partitions. The shrinking cell condition (c) is satisfied if and only if

$$\lim_{n \to \infty} h_{jn} \to 0 \quad \text{for } j = 1, \ldots, d \, . \tag{27}$$

As for condition (a), assume without loss of generality that V is a compact set of the form $[-t,t]\times\cdots\times[-t,t]$, with $t>0$. Each cell of $\pi\in\Pi_n$ has volume $\prod_{j=1}^d h_{jn}$, and if (27) holds, then

$$m(\Pi_n : V) \leq \frac{(2t+1)^d}{\prod_{j=1}^d h_{jn}}$$

when n is sufficiently large. Thus the restricted covering numbers grow sublinearly if

$$n\cdot\prod_{j=1}^d h_{jn}\to\infty \quad \text{as} \quad n\to\infty\,. \tag{28}$$

The conclusion of the following theorem holds regardless of the joint distribution of $(X,Y)\in \mathbb{R}^d\times[-K,K]$.

Theorem 4 *For each n, let the regression estimate $\hat{r}_n$ be based on a shifted version of the partition θ_n. If (27) and (28) hold, then $P|r-\hat{r}_n|^2\to 0$ with probability one.*

7 K-thresholding with Variable Weights

In this section we consider a Y-dependent variant of the k-spacing regression estimate of Parthasarathy and Battacharya (1961). Let $d=1$, so that $X_1,X_2,\ldots$ are real-valued, and assume that the distribution P of X_i has a density with respect to Lebesgue measure. Let $F:\mathbb{R}\to[0,\infty)$ be a non-negative weight function that is bounded away from zero and infinity on every compact set. The partition $\psi_n(T_n)$ is found by ordering $X_1,\ldots,X_n$, and then grouping them into intervals based on the weights $F(Y_i)$ of the corresponding response variables. In scanning from left to right, a new interval is begun when the running sum of the weights $F(Y_i)$ associated with X_i in the current interval exceeds a preassigned threshold.

Let $k_n>0$ be a preassigned threshold value. Let ρ be the unique permutation on $\{1,\ldots,n\}$ such that $X_{\rho(1)}<X_{\rho(2)}<\ldots<X_{\rho(n)}$. (Such a permutation exists with probability one as P has a density.) Set $m_1=1$ and recursively define successive threshold times

$$m_{r+1} = 1+\min\left\{m\geq m_r : \sum_{k=m_r}^{m} F(Y_{\rho(k)})\geq k_n\right\}$$

until $\sum_{k=m_s}^n F(Y_{\rho(k)})<k_n$. Let $\psi_n(T_n)$ be a partition of $\mathbb{R}$ into intervals $\{A_1,\ldots,A_s\}$ such that $X_{\rho(m_j)},\ldots,X_{\rho(m_{j+1}-1)}\in A_j$ for $j=1,\ldots,s-1$, and $X_{\rho(m_s)},\ldots,X_{\rho(n)}\in A_s$. If the weight function F is identically 1, then $\psi_n(T_n)$ is the ordinary k-spacing partition of $\mathbb{R}$.

For suitable sequences of threshold values, Theorem 1 shows that histogram regression estimates based on ψ_n are consistent. The proof, which makes use of Example 1 above, is similar to that given in Lugosi and Nobel (1995) for k-spacing density estimates, and is therefore ommitted.

Theorem 5 *Let a regression estimate $\hat{r}_n$ be formed by averaging response variables within the cells of $\psi_n(T_n)$. If $k_n \to \infty$ and $k_n/n \to 0$ then $P|\hat{r}_n - r|^2 \to 0$ with probability one.* □

Remark: In higher dimensions, similar results can be obtained for a Y-dependent version of the partitioning scheme proposed by Gessaman (1970). The Y-independent case is discussed in Lugosi and Nobel (1995).

8 Nearest-neighbor Clustering

Clustering of multivariate data is a widely used method of statistical analysis. Clustering schemes typically partition the training set by minimizing an empirical error criterion. In this section we establish the consistency of regression estimates based on nearest-neighbor clustering of the (unlabeled) measurement vectors X_i. Let $\|\cdot\|$ be the usual Euclidean norm on $\mathbb{R}^d$.

A clustering scheme is a function $C : \mathbb{R}^d \to \mathcal{C}$ that associates every vector $x \in \mathbb{R}^d$ with one of a finite number of cluster centers $\mathcal{C} = \{c_1, \ldots c_m\} \subseteq \mathbb{R}^d$. Each clustering scheme has a corresponding partition $\pi = \{A_1, \ldots, A_m\}$ with cells $A_j = \{x : Q(x) = c_j\}$. A clustering scheme is said to be *nearest neighbor* if for each $x \in \mathbb{R}^d$,

$$C(x) = \arg\min_{c_j \in \mathcal{C}} \|x - c_j\|,$$

with ties broken in favor of the center c_j having the least index. In this case the partition of C is the nearest-neighbor partition of its cluster centers. See Hartigan (1975) or Gersho and Gray (1992) for more details concerning multivariate clustering and its applications.

The risk of a clustering scheme C is commonly measured by $R(C) = \int \|x - C(x)\|^2 dP$, the expected squared distance between a random vector $X \sim P$ and its corresponding cluster center. The *empirical risk* of C is given by

$$R_n(C) = \int \|x - C(x)\|^2 d\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \|X_i - C(X_i)\|^2 . \tag{29}$$

From a training set $T_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ and a clustering scheme C one may produce a histogram regression estimate by averaging the response variables Y_i within the cells of C. We consider estimates based on empirically optimal nearest neighbor clustering schemes.

Definition: For each $x \in \mathbb{R}^d$ and every $\delta > 0$ let $B(x, \delta) = \{v : \|u - v\| < \delta\}$. The *support* set of P is defined by

$$S_P = \{x : P(B(x, \delta)) > 0 \quad \text{for every} \quad \delta > 0\}.$$

It is easy to see that S_P is a closed set with $P(S_P) = 1$.

Proposition 3 *For each $n \geq 1$ let $C_n : \mathbb{R}^d \to \mathcal{C}_n$ minimize the empirical risk $R_n(C)$ over all nearest neighbor clustering schemes having k_n cluster centers. If $E\|X\|^2 < \infty$ and $k_n \to \infty$, then for every compact set $V \subseteq \mathbb{R}^d$,*

$$\max_{u \in S_P \cap V} \min_{c \in \mathcal{C}_n} \|u - c\| \to 0$$

with probability one.

Proof: By the monotonicity of $\|\cdot\|^2$, the cluster centers of an empirically optimal scheme must lie within the closed convex hull of $X_1, \ldots, X_n$. As this set is compact, the continuity of $\|\cdot\|^2$ insures that C_n exists.

Let $u_1, u_2, \ldots$ be dense in $\mathbb{R}^d$, with $u_1 = 0$. For each n let C'_n be the nearest neighbor clustering scheme with centers $\{u_1, \ldots, u_n\}$. For each compact set $V \subseteq \mathbb{R}^d$,

$$R_n(C_n) \;\leq\; R_n(C'_n) \;\leq\; \max_{x \in V} \min_{1 \leq j \leq k_n} \|x - a_j\|^2 \;+\; \int_{V^c} \|x\|^2 d\hat{P}_n$$

and therefore

$$\limsup_{n \to \infty} R_n(C_n) \leq \int_{V^c} \|x\|^2 dP \,.$$

with probability one. As V was arbitrary, $R_n(C_n) \to 0$ with probability one.

By applying the strong law of large numbers to a countable, dense subset of S_P, it may be established that

$$\mathbb{P}\left\{\liminf_{n \to \infty} P_n(B(u, \delta)) > 0 \;\; \text{for every} \;\; u \in S_P, \, \delta > 0\right\} = 1. \tag{30}$$

Suppose for the moment that there is a compact set $V \subseteq \mathbb{R}^d$ and a number $\delta > 0$ such that

$$\mathbb{P}\left\{\limsup_{n \to \infty} \left[\max_{u \in S_P \cap V} \min_{c \in \mathcal{C}_n} \|u - c\|\right] > \delta\right\} > 0\,. \tag{31}$$

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots$ be a sequence of codebooks corresponding to a sample point in this event. As $S_P \cap V$ is compact, there exists a sequence of vectors $\{u_{n_k}\}$ such that $u_{n_k} \to u^* \in S_P$ and $\mathcal{C}_{n_k} \subseteq B^c(u_{n_k}, \delta)$ for every k. Thus when k is sufficiently large,

$$R_{n_k}(C_{n_k}) \geq \frac{\delta}{3} \hat{P}_{n_k}(B(u^*, \delta/3)) \, ,$$

and it follows from (30) and (31) that

$$\mathrm{IP}\left\{\liminf_{n\to\infty} R_n(C_n) > 0\right\} > 0.$$

But this contradicts the fact that $R(C_n) \to 0$ with probability, and we conclude that (31) cannot hold. □

Theorem 6 *Let C_n minimize the empirical risk $R_n(C)$ over all nearest neighbor clustering schemes with k_n cluster centers. Define $\hat{r}_n$ by averaging the response variables Y_i within the cells of C_n. If $E\|X\|^2$ is finite, $k_n \to \infty$, and $n^{-1}k_n^2 \log n \to 0$, then $P|r - \hat{r}_n|^2 \to 0$ with probability one.*

Proof: The partition associated with C_n lies in the family $\mathcal{V}_{k_n}$ containing all nearest-neighbor partitions of k_n vectors in IR^d. Example 3 of Section 1.2 shows that $m(\mathcal{V}_{k_n}) = k_n$ and $\Delta_n^*(\mathcal{V}_{k_n}) \leq n^{k_n^2 d}$. Our assumptions on k_n guarantee that

$$n^{-1} m(\mathcal{V}_{k_n}) \to 0 \quad \text{and} \quad n^{-1} \log \Delta_n^*(\mathcal{V}_{k_n}) \to 0 \, .$$

The shrinking cell condition (c) of Theorem 1 follows easily from Proposition 3 above. □

9 Empirically Optimal Regression Trees

Tree-structured partitions were defined in Example 3 of Section 2. A regression tree is a function $f : \mathrm{IR}^d \to \mathrm{IR}$ defined by assigning a number to each cell of a tree-structured partition (T, τ). Alternatively, one may define a regression tree by augmenting the pair (T, τ) with an additional node function $\alpha : T \to \mathrm{IR}$, and setting $f(x) = \alpha(t)$ for every vector x whose path through T ends at the leaf node t. In the notation of Section 2, a k-node regression tree is a function $f \in \mathcal{G} \circ \mathcal{T}_k$, where $\mathcal{G}$ is the the class of constant functions taking values in $[-K, K]$.

The regression function $r(\cdot)$ minimizes the predictive risk $J(f) = E|f(X) - Y|^2$ over all functions $f : \mathrm{IR}^d \to \mathrm{IR}$. In practice, when the training set T_n is given but the distribution of

(X, Y) is unknown, it is common to seek an estimate $\hat{f}_n$ that minimizes the empirical risk

$$J_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

over a suitable class of regression estimates. Empirically optimal regression trees yield consistent estimates of $r(\cdot)$ if the size of the trees grows with n at a controlled rate.

Theorem 7 *Let $\hat{f}_n$ minimize the empirical risk $J_n(f)$ over all the k_n-node regression trees $f \in \mathcal{G} \circ \mathcal{T}_{k_n}$. If $k_n \to \infty$ and $k_n = o(n/\log n)$, then $P|\hat{f}_n - r|^2 \to 0$ with probability one.*

Proof: As k_n grows without bound, there exists a fixed, tree-structured partitions $\pi_n \in \mathcal{T}_{k_n}$ such that for every compact set $V \subseteq \mathrm{IR}^d$,

$$\max_{A \in \pi_n} \operatorname{diam}(A \cap V) \to 0\,. \tag{32}$$

Let $\hat{g}_n$ be the histogram regression estimate produced from T_n by averaging the response variables Y_i within the cells of π_n.

Under the assumption that $k_n = o(n/\log n)$, the bounds derived in Example 3 of Section 1.2 show that $n^{-1} m(\mathcal{T}_{k_n}) \to 0$ and $n^{-1} \Delta_n^*(\mathcal{T}_{k_n}) \to 0$. It then follows from (32) and Theorem 1 that $\{\hat{g}_n\}$ is strongly consistent. Since for each n $\hat{f}_n$ and $\hat{g}_n$ are contained in $\mathcal{G} \circ \mathcal{T}_{k_n}$, and $\hat{f}_n$ is empirically optimal, the consistency of the estimates $\hat{f}_n$ follows from Lemma 4. □

We remark that the partitions π_n above may be chosen to have rectangular cells. Thus the conclusion of the theorem holds for the empirically optimal regression trees employing axis-parallel splits. It should also be noted that the estimates $\hat{f}_n$ need *not* have shrinking cells in the sense of (4). When n is large, the empirically optimal tree will not divide regions of the measurement space on which the regression function is constant.

Acknowledgements

The author has benefitted from numerous conversations with Richard Olshen. The helpful comments of Gábor Lugosi and Amir Dembo contributed to the content and organization of the paper.

References

[1] T.W. Anderson. Some nonparametric multivariate procedures based on statistically equivalent blocks. In P.R. Krishnaiah, editor, *Multivariate Analysis*, pages 5–27. Academic Press, 1966.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth International, Belmont, CA., 1984.

[3] P. Chaudhuri, M.-C. Huang, W.-Y. Loh, and R. Yao. Piecewise polynomial regression trees. *Statistica Sinica*, 4:143–167, 1994.

[4] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on Electronic Computers*, 14:326–334, 1965.

[5] L. Devroye. Automatic pattern recognition: a study of the probability of error. *IEEE Trans. on Pattern Anal. and Mach. Intelligence*, 10(4):530–543, July 1988.

[6] L. Devroye and L. Györfi. Distribution-free exponential bounds on the l_1 error of partitioning estimates of a regression function. In F. Konecny, J. Mogyoródi, and W. Wertz, editors, *Proc. of the Fourth Pannonian Symposium on Mathematical Statistics*, pages 67–76, Budapest, Hungary, 1985. Akadémiai Kiadó.

[7] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, Boston, 1992.

[8] M.P. Gessaman. A consistent nonparametric multivariate density esimator based on statistically equivalent blocks. *Ann. Math. Stat.*, 41:1344–1346, 1970.

[9] L. Gordon and R. Olshen. Asymptotically efficient solutions to the classification problem. *Ann. Statist.*, 6:515–533, 1978.

[10] L. Gordon and R. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.

[11] L. Gordon and R. Olshen. Almost sure consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15:147–163, 1984.

[12] J.A. Hartigan. *Clustering Algorithms.* John Wiley, New York, 1975.

[13] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans. Comm.*, 28:84–95, 1980.

[14] G. Lugosi and A.B. Nobel. Consistency of data-driven histogram methods for density estimation and classification. Technical Report UIUC-BI-93-01, Beckman Institute, University of Illinois, Urbana-Champaign, 1993. Accepted for publication in Ann. Stat.

[15] K.R. Parthasarathy and P.K. Bhattacharya. Some limit theorems in regression theory. *Sankhya Series A*, 23:91–102, 1961.

[16] E.A. Patrick and F.P. Fisher. Introduction to the performance of distribution-free conditional risk learning systems. Technical Report TR-EE-67-12, Purdue University, Lafayette, Indiana, 1967.

[17] D. Pollard. *Convergence of Stochastic Processes.* Springer, New York, 1984.

[18] C.J. Stone. Consistent nonparametric regression. *Ann. Stat.*, 8:1348–1360, 1977.

[19] C.J. Stone. An asymptotically optimal histogram selection rule. In L. Le Cam and R.A. Olshen, editors, *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer*, volume II, pages 513–520. Wadsworth, 1985.

[20] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[21] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Probab. Appl.*, 26:532–553, 1981.

[22] L.C. Zhao, P.R. Krishnaiah, and X.R. Chen. Almost sure L_r-norm convergence for data-based histogram density estimates. *Theory Probab. Appl.*, 35:396–403, 1990.