Consistency of Data-driven Histogram Methods for Density Estimation and Classification

Gábor Lugosi * Andrew Nobel †

submitted: April 14, 1993 *revised:* April 15, 1995

Abstract

We present general sufficient conditions for the almost sure L_1-consistency of histogram density estimates based on data-dependent partitions. Analogous conditions guarantee the almost-sure risk consistency of histogram classification schemes based on data-dependent partitions. Multivariate data is considered throughout.

In each case, the desired consistency requires shrinking cells, subexponential growth of a combinatorial complexity measure, and sub-linear growth of the number of cells. It is *not* required that the cells of every partition be rectangles with sides paralles to the coordinate axis, or that each cell contain a minimum number of points. No assumptions are made concerning the common distribution of the training vectors.

We apply the results to establish the consistency of several known partitioning estimates, including the k_n-spacing density estimate, classifiers based on statistically equivalent blocks, and classifiers based on multivariate clustering schemes.

Appears in Ann. Stat. 1996, v.26, pp.687-706

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

Key Words: Partitioning rules, histogram density estimation, histogram classification, statistically equivalent blocks, Vapnik-Chervonenkis inequality.

*Gábor Lugosi is with the Dept. of Mathematics, Faculty of Elect. Engineering, Technical University of Budapest, Hungary. Email: lugosi@vma.bme.hu . His research was supported in part by the National Science Foundation under Grant No. NCR-91-57770

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

1 Introduction

A natural method of estimating local properties of data in nonparametric statistics is to partition the space of observations into cells, and then compute statistics locally within each cell. This leads to histogram estimates of an unknown density, and to partition-based classification rules. The simplest histogram methods partition the space into congruent intervals or cubes whose size and position depends on the number of available data points, but not on the data itself. These methods provide estimates that are consistent, regardless of the underlying distribution of the data. Abou-Jaoude (1976a), (1976c) gave necessary and sufficient conditions under which a sequence of regular partitions gives rise to L_1-consistent estimates for every density (see also Devroye and Györfi (1985)). A similar result for classification and regression estimates based on cubic partitions was obtained by Devroye and Györfi (1983). The weak (in-probability) consistency of these schemes can also be deduced from the general result of Stone (1977).

Statistical practice suggests that histogram estimators based on data-dependent partitions will provide better performance than those based on a fixed sequence of partitions. Theoretical evidence for this superiority was given by Stone (1985). 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. In one dimensional problems statistically equivalent blocks reduce to *k-spacing estimates* (Mahalanobis (1961), Parthasarathy and Bhattacharya (1961), Van Ryzin (1973)), where the k-th, $2k$-th, ... order statistics determine the partition of the real line.

Many other data-dependent partitioning schemes have been introduced in the literature (cf. Devroye (1988)). In many cases the partition is described by a binary tree, each of whose leaves corresponds to a cell of the partition. The tree structure makes computation of the corresponding classification rule or density estimate fast, and provides a ready interpretation of the estimate. The consistency of tree-structured classification and regression was investigated by Gordon and Olshen (1978), (1980), (1984) in a general framework, and was extended by Breiman, Friedman, Olshen and Stone (1984).

Existing conditions for the consistency of histogram classification and density estimation using data-dependent partitions require significant restrictions. The conditions of Breiman *et al.* (1984) for consistent classification require that each cell of every partition belongs to a fixed Vapnik-Chervonenkis class of sets, and that every cell must contain at least k_n points,

where $k_n/\log n \to \infty$ as the sample size n tends to infinity. Chen and Zhao (1987), and Zhao, Krishnaiah, and Chen (1990) restrict their attention to density estimates based rectangular partitions.

This paper presents general sufficient conditions for the almost-sure L_1 consistency of histogram classification and density estimates that are based on data-dependent partitions. Analogous conditions for the consistency of histogram regression estimates are addressed in Nobel (1994).

In the next section two combinatorial properties of partition families are defined, and a Vapnik-Chervonenkis type large deviation inequality is established. In Section 3, common features of the estimates investigated in the paper are defined. Sections 4 and 5 are devoted to the consistency results for density estimation and classification, respectively.

Our results establish consistency under significantly weaker conditions than those imposed by Breiman *et al.* (1984) and Zhao, Krishnaiah, and Chen (1990), and are readily applicable to a number of existing partitioning schemes. In Section 6 the results are applied to establish the consistency k_n-spacing density estimates, classifiers based on statistically equivalent blocks, and classifiers based on clustering of the data.

2 A Vapnik-Chervonenkis Inequality for Partitions

Let $\mathbb{R}^d$ denote d-dimensional Euclidean space. An ordered sequence $(x_1, \ldots, x_n) \in \mathbb{R}^{n \cdot d}$ will be denoted by x_1^n. By a partition of $\mathbb{R}^d$ we mean a finite collection $\pi = \{A_1, \ldots A_r\}$ of Borel-measurable subsets of $\mathbb{R}^d$, referred to as cells, with the property that (i) $\cup_{j=1}^r A_j = \mathbb{R}^d$ and (ii) $A_i \cap A_j = \emptyset$ if $i \neq j$. Let $|\pi|$ denote the number of cells in π.

Let $\mathcal{A}$ be a (possibly infinite) family of partitions of $\mathbb{R}^d$. The *maximal cell count* of $\mathcal{A}$ is given by

$$m(\mathcal{A}) = \sup_{\pi \in \mathcal{A}} |\pi| \, .$$

The complexity of $\mathcal{A}$ will be measured by a combinatorial quantity similar to the growth function for classes of sets that was proposed by Vapnik and Chervonenkis (1971). Fix n points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\}$. Let $\Delta(\mathcal{A}, x_1^n)$ be the number of distinct partitions

$$\{A_1 \cap B, \ldots, A_r \cap B\} \tag{1}$$

of the finite set B that are induced by partitions $\{A_1, \ldots, A_r\} \in \mathcal{A}$. Note that the order of

appearance of the individual sets in (1) is not important. It is easy to see that $\Delta(\mathcal{A}, x_1^n) \leq m(\mathcal{A})^n$. Define the *growth function* of $\mathcal{A}$ as follows:

$$\Delta_n^*(\mathcal{A}) = \max_{x_1^n \in \mathbb{R}^{n \cdot d}} \Delta(\mathcal{A}, x_1^n) \tag{2}$$

is the largest number of distinct partitions of any n point subset of $\mathbb{R}^d$ that can be induced by the partitions in $\mathcal{A}$.

Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $X_i \sim \mu$ and let μ_n denote the *empirical distribution* of $X_1, \ldots, X_n$. We wish to establish a large deviations inequality for random variables of the form

$$\sup_{\pi \in \mathcal{A}} \sum_{A \in \pi} |\mu_n(A) - \mu(A)|, \tag{3}$$

where $\mathcal{A}$ is an appropriate family of partitions. Our analysis relies on the well-known inequality of Vapnik and Chervonenkis (1971). Consider a class $\mathcal{C}$ of subsets of $\mathbb{R}^d$. The *shatter coefficient* $S_n(\mathcal{C})$ is defined to be the maximum cardinality of the collection $\{B \cap C : C \in \mathcal{C}\}$, as B ranges over subsets of $\mathbb{R}^d$ containing n points. Vapnik and Chervonenkis (1971) showed that for each $n \geq 1$ and each $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{A \in \mathcal{C}} |\mu_n(A) - \mu(A)| > \epsilon\right\} \leq 4\, S_{2n}(\mathcal{C})\, e^{-n\epsilon^2/8}. \tag{4}$$

Remark: In order to insure measurability of the supremum in (3), it is necessary to impose regularity conditions on uncountable collections of partitions. Suppose that $m(\mathcal{A}) = r < \infty$. Let Ω consist of all measurable functions $f : \mathbb{R}^d \to \{1, \ldots, r\}$. Each function in Ω corresponds to a measurable partition of $\mathbb{R}^d$ having at most r cells, and each partition in $\mathcal{A}$ corresponds to a finite collection of functions in Ω. Let $\Omega' \subseteq \Omega$ be the collection of all such functions associated with partitions in $\mathcal{A}$. It is assumed that each family $\mathcal{A}$ considered here gives rise to a collection Ω' that contains a countable sub-collection Ω_0 with the property that every function in Ω' is the pointwise limit of a sequence of functions in Ω_0. It is easy to show (c.f. Pollard (1984), pp.38-39) that the supremum in (3) is measurable when $\mathcal{A}$ has this property.

The following lemma presents a Vapnik-Chervonenkis inequality for partition families. A similar inequality, for families of rectangular partitions, was established by Zhao, Krishnaiah, and Chen (1990).

Lemma 1 *Let $\mathcal{A}$ be any collection of partitions of $\mathbb{R}^d$. For each $n \geq 1$ and every $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{\pi\in\mathcal{A}} \sum_{A\in\pi} |\mu_n(A) - \mu(A)| > \epsilon\right\} \leq 4\,\Delta^*_{2n}(\mathcal{A})\,2^{m(\mathcal{A})}\,e^{-n\epsilon^2/32}. \tag{5}$$

Remark: A longer, but more general, proof can be found in Lugosi and Nobel (1993). The argument below was suggested by Andrew Barron.

Proof of Lemma 1: For each partition $\pi = \{A_1, \ldots, A_r\} \in \mathcal{A}$ let $\mathcal{B}(\pi)$ be the collection of all 2^r sets that can be expressed as the union of cells of π. Let

$$\mathcal{B}(\mathcal{A}) = \{A \in \mathcal{B}(\pi) : \pi \in \mathcal{A}\}$$

be the collection of all such unions, as π ranges through $\mathcal{A}$. Fix π for the moment and define

$$\tilde{A} = \bigcup_{A\in\pi:\mu_n(A)\geq\mu(A)} A\,.$$

Then clearly

$$\begin{aligned}\sum_{A\in\pi} |\mu_n(A) - \mu(A)| &= 2\left(\mu_n(\tilde{A}) - \mu(\tilde{A})\right) \\ &\leq 2 \sup_{A\in\mathcal{B}(\pi)} |\mu_n(A) - \mu(A)|\,.\end{aligned}$$

Consequently,

$$\begin{aligned}\sup_{\pi\in\mathcal{A}} \sum_{A\in\pi} |\mu_n(A) - \mu(A)| &\leq 2 \sup_{\pi\in\mathcal{A}} \sup_{A\in\mathcal{B}(\pi)} |\mu_n(A) - \mu(A)| \\ &= 2 \sup_{A\in\mathcal{B}(\mathcal{A})} |\mu_n(A) - \mu(A)|. \end{aligned} \tag{6}$$

A straightforward argument shows that $S_{2n}(\mathcal{B}(\mathcal{A})) \leq 2^{m(\mathcal{A})}\Delta^*_{2n}(\mathcal{A})$. In conjunction with (4) and (6) it then follows that

$$\begin{aligned}\mathbb{P}\left\{\sup_{\pi\in\mathcal{A}} \sum_{A\in\pi} |\mu_n(A) - \mu(A)| > \epsilon\right\} &\leq \mathbb{P}\left\{\sup_{A\in\mathcal{B}(\mathcal{A})} |\mu_n(A) - \mu(A)| > \frac{\epsilon}{2}\right\} \\ &\leq 4\,\Delta^*_{2n}(\mathcal{A})\,2^{m(\mathcal{A})}\,e^{-n\epsilon^2/32},\end{aligned}$$

as desired. □

The results of Sections 4 and 5 rely on the following corollary of Lemma 1, whose proof is an easy application of the Borel-Cantelli Lemma.

Corollary 1 *Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $X_i \sim \mu$, and let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of partition families. If as n tends to infinity*

(a) $n^{-1}m(\mathcal{A}_n) \to 0$ *and*

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

then

$$\sup_{\pi \in \mathcal{A}_n} \sum_{A \in \pi} |\mu_n(A) - \mu(A)| \to 0 \tag{7}$$

with probability one.

3 Data-driven Partitioning Schemes

The density and classification estimates studied below have several common features. In each case an estimate is produced in two stages from a *training set* T_n that consists of n i.i.d. random variables $Z_1, \ldots, Z_n$ taking values in a set $\mathcal{X}$. For density estimation $\mathcal{X} = \mathbb{R}^d$, while for classification $\mathcal{X} = \mathbb{R}^d \times \{1, \ldots, M\}$. Using T_n a partition $\pi_n = \pi_n(Z_1, \ldots, Z_n)$ is produced according to a prescribed rule. The partition π_n is then used in conjunction with T_n to produce a density estimate as in Section 4, or a classification rule as in Section 5. In either case, the training set is "used twice" and it is this feature of data-dependent histogram methods that distinguish them from fixed histogram methods.

An *n-sample partitioning rule* for $\mathbb{R}^d$ is a function π_n that associates every n-tuple $(z_1, \ldots z_n) \in \mathcal{X}^n$ with a measurable partition of $\mathbb{R}^d$. Applying the rule π_n to $Z_1, \ldots, Z_n$ produces a random partition $\pi_n(Z_1, \ldots, Z_n)$. A *partitioning scheme* for $\mathbb{R}^d$ is a sequence of partitioning rules

$$\Pi = \{\pi_1, \pi_2, \ldots\}$$

Associated with every rule π_n there is a fixed, non-random family of partitions

$$\mathcal{A}_n = \{\pi_n(z_1, \ldots, z_n) : z_1, \ldots, z_n \in \mathcal{X}\}.$$

Thus every partitioning scheme Π is associated with a sequence $\{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$ of partition families. In what follows the random partitions $\pi_n(Z_1, \ldots, Z_n)$ will be denoted simply by π_n. With this convention in mind, for every $x \in \mathbb{R}^d$ let $\pi_n[x]$ be the unique cell of π_n that contains the point x.

Let A be any subset of $\mathbb{R}^d$. The *diameter* of A is the maximum Euclidean distance between any two points of A:

$$\text{diam}(A) = \sup_{x,y \in A} \|x - y\|.$$

For each $\gamma > 0$ let A^γ be the set of points in $\mathbb{R}^d$ that are within distance γ of some point in A,

$$A^\gamma = \left\{ x : \inf_{y \in A} \|x - y\| < \gamma \right\}.$$

4 Histogram Density Estimation

In this section we investigate the consistency of histogram density estimates based on data-dependent partitions. Let μ be a probability distribution on $\mathbb{R}^d$ having density f, so that

$$\mu(A) = \int_A f(x)dx.$$

for every Borel subset A of $\mathbb{R}^d$. Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$, each distributed according to μ, and let μ_n be the empirical distribution of $X_1, \ldots, X_n$. Fix a partitioning scheme $\Pi = \{\pi_1, \pi_2, \ldots\}$ for $\mathbb{R}^d$. Applying the n'th rule in Π to $X_1, \ldots, X_n$ produces a partition $\pi_n = \pi_n(X_1^n)$ of $\mathbb{R}^d$. The partition π_n, in turn, gives rise to a natural histogram estimate of f as follows. For each vector $x \in \mathbb{R}^d$ let

$$f_n(x) = \begin{cases} \mu_n(\pi_n[x])/\lambda(\pi_n[x]) & \text{if } \lambda(\pi_n[x]) < \infty \\ 0 & \text{otherwise .} \end{cases} \tag{8}$$

Here λ denotes the Lebesgue measure on $\mathbb{R}^d$. Note that f_n is itself a function of the training set $X_1, \ldots, X_n$, and that f_n is piecewise constant on the cells of π_n. The sequence of estimates $\{f_n\}$ is said to be *strongly L_1-consistent* if

$$\int |f(x) - f_n(x)|dx \to 0. \tag{9}$$

with probability one as $n \to \infty$. The strong distribution-free consistency of kernel and non-data dependent histogram estimates has been thoroughly studied by Devroye and Györfi (1985).

Remark: While the estimates f_n are always non-negative, they need not integrate to one, indeed $\int f_n(x)dx(x)$ is just the fraction of those points $X_1, \ldots, X_n$ lying in cells $A \in \pi_n$ for which $\lambda(A)$ is finite. The consistency of the normalized estimates is addressed in Corollary 2 below.

Proposition 1 *Let f be a density function on $\mathbb{R}^d$, and for some $\epsilon < 1/2$ let $g \geq 0$ satisfy*

$$\int |f - g|dx < \epsilon\,.$$

If $\hat{g}(x) = g(x)/\int g(y)dy$ is the normalized density corresponding to g, then

$$\int |f - \hat{g}|dx < \frac{8\epsilon}{3}.$$

Proof: In this proof all integrals are understood with respect to Lebesgue measure. Since $|\int g - \int f| \le \int |g - f| < \epsilon$, it follows that $1 - \epsilon \le \int g \le 1 + \epsilon$. Therefore,

$$\begin{aligned}\int \left| f - \frac{g}{\int g} \right| &\le \int \left| f - \frac{f}{\int g} \right| + \int \left| \frac{f}{\int g} - \frac{g}{\int g} \right| \\ &= \int f \left| 1 - \frac{1}{\int g} \right| + \frac{1}{\int g} \int |f - g| \\ &< 1 - \frac{1}{1+\epsilon} + \frac{\epsilon}{1-\epsilon} \le \frac{8\epsilon}{3}\,.\end{aligned}$$

□

The following theorem extends previous work of Zhao, Krishnaiah, and Chen (1990) who found sufficient conditions for the strong L_1 consistency of histogram density estimates based on infinite, data-dependent rectangular partitions. Our result differs from theirs in two respects. First, we place no restriction on the geometry of the partitions outside of the growth condition *(b)* below. Secondly, the condition *(c)* weakens their requirement that for λ-almost every x the cells containing x have diameter tending to zero.

Theorem 1 *Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ whose common distribution μ has a density f. Let $\Pi = \{\pi_1, \pi_2, \ldots\}$ be a fixed partitioning scheme for $\mathbb{R}^d$, and let $\mathcal{A}_n$ be the collection of partitions associated with the rule π_n. If as n tends to infinity,*

(a) $n^{-1} m(\mathcal{A}_n) \to 0$,

(b) $n^{-1} \log \Delta_n^*(\mathcal{A}_n) \to 0$, *and*

(c) $\mu\{x : \mathrm{diam}(\pi_n[x]) > \gamma\} \to 0$ *with probability one for every* $\gamma > 0$,

then the density estimates f_n are strongly consistent in L_1:

$$\int |f(x) - f_n(x)| dx \to 0$$

with probability one.

Proof: Fix a number $\epsilon \in (0, 1/2)$. It follows from Proposition 1 and standard arguments that there is a continuous density g on $\mathbb{R}^d$ such that $\{x : g(x) > 0\}$ is bounded and $\int |f - g| dx < \epsilon$. Let ν be the measure corresponding to g and set $S_\nu = \{x : g(x) > 0\}$.

Fix n and let $\pi_n = \pi_n(X_1^n)$ be the random partition produced from $X_1, \ldots, X_n$. Let f_n be as in (8) above and define the auxiliary functions

$$\tilde{f}_n(x) = \begin{cases} \mu(\pi_n[x])/\lambda(\pi_n[x]) & \text{if } \lambda(\pi_n[x]) < \infty \\ 0 \quad \text{otherwise} \end{cases}$$

and

$$\tilde{g}_n(x) = \begin{cases} \nu(\pi_n[x])/\lambda(\pi_n[x]) & \text{if } \lambda(\pi_n[x]) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that

$$|f - f_n| \leq |f - g| + |\tilde{g}_n - \tilde{f}_n| + |g - \tilde{g}_n| + |\tilde{f}_n - f_n|, \tag{10}$$

so the L_1 error of f_n is bounded above by the sum of the integrals of each term on the right-hand side above. By design, $\int |f - g| dx < \epsilon$, and it is easy to see that

$$\int |\tilde{g}_n - \tilde{f}_n| dx \leq \sum_{A \in \pi_n} |\nu(A) - \mu(A)| \leq \int |f - g| dx < \epsilon$$

as well.

The last term in (10) involves the difference between μ_n and μ on cells of the random partition π_n. By considering the worst-case behavior over the range of $\pi_n(\cdot)$, we obtain an upper bound to which the results of Section 2 apply:

$$\begin{aligned} \int |\tilde{f}_n - f_n| dx &\leq \sum_{A \in \pi_n} |\mu_n(A) - \mu(A)| \\ &\leq \sup_{\pi \in \mathcal{A}_n} \sum_{A \in \pi} |\mu_n(A) - \mu(A)|, \end{aligned}$$

and it follows from Corollary 1 of Lemma 1 that

$$\lim_{n \to \infty} \int |\tilde{f}_n - f_n| dx = 0$$

with probability one.

It remains to consider the third term in (10). Let $\delta > 0$ be so small that $\delta \lambda(S_\nu^1) \leq \epsilon$, where S_ν^1 denotes the 1-blowup of S_ν. Let $\gamma \in (0, 1)$ be such that for every set $A \subseteq \mathbb{R}^d$ having diameter less than γ,

$$\sup_{x,y \in A} |g(x) - g(y)| < \delta.$$

Let π_n^* be the collection of cells $A \in \pi_n$ for which $\lambda(A)$ is finite. Then

$$\begin{aligned} \int |g(x) - \tilde{g}_n(x)| dx &= \sum_{A \in \pi_n^*} \int_A \left| g(x) - \frac{\nu(\pi_n[x])}{\lambda(\pi_n[x])} \right| dx + \sum_{A \notin \pi_n^*} \int_A g(x) dx \\ &\leq \sum_{A \in \pi_n^*} \int_A \left| g(x) - \frac{\nu(\pi_n[x])}{\lambda(\pi_n[x])} \right| dx + \nu\{x : \text{diam}(\pi_n[x]) \geq \gamma\}. \end{aligned} \tag{11}$$

An application of Fubini's Theorem shows that if $A \in \pi_n^*$, then

$$\begin{aligned}\int_A \left| g(x) - \frac{\nu(\pi_n[x])}{\lambda(\pi_n[x])} \right| dx &= \lambda(A)^{-1} \int_A |g(x)\lambda(A) - \nu(A)|\, dx \\ &= \lambda(A)^{-1} \int_A \left| g(x) \int_A dy - \int_A g(y) dy \right| dx \\ &\leq \lambda(A)^{-1} \int_{A\times A} |g(x) - g(y)| dx dy\,. \end{aligned} \tag{12}$$

If $A \cap S_\nu = \emptyset$ then

$$\int_{A\times A} |g(x) - g(y)| dx dy = 0. \tag{13}$$

Suppose that $A \cap S_\nu \neq \emptyset$. If $\mathrm{diam}(A) < \gamma$ then $A \subseteq S_\nu^\gamma$ and it follows that

$$\int_{A\times A} |g(x) - g(y)| dx dy \leq \delta\lambda^2(A) = \delta\lambda^2(A \cap S_\nu^\gamma)\,. \tag{14}$$

On the other hand, if $\mathrm{diam}(A) \geq \gamma$ then

$$\int_{A\times A} |g(x) - g(y)| dx dy \leq 2 \int_{A\times A} g(x) dx dy = 2\nu(A)\lambda(A)\,. \tag{15}$$

Combining (11) - (14) shows that

$$\begin{aligned}\int |g(x) - \tilde{g}_n(x)| dx &\leq 3\nu(\{x : \mathrm{diam}(\pi_n[x]) \geq \gamma\}) + \delta\lambda(S_\nu^\gamma) \\ &\leq 3\mu(\{x : \mathrm{diam}(\pi_n[x]) \geq \gamma\}) + \frac{3}{2}\epsilon + \delta\lambda(S_\nu^\gamma)\,, \end{aligned}$$

where the second inequality follows from the fact that for every Borel set $A \subseteq \mathbb{R}^d$,

$$|\nu(A) - \mu(A)| \leq \frac{1}{2}\int |f - g| dx < \frac{1}{2}\epsilon\,.$$

Letting $n \to \infty$ and making use of assumption (c) in the statement of the theorem,

$$\limsup_{n\to\infty} \int |g(x) - \tilde{g}_n(x)| dx \leq \frac{3}{2}\epsilon + \delta\lambda(S_\nu^\gamma) \leq \frac{5}{2}\epsilon$$

with probability one. The result may now be established by letting ϵ tend to zero. □

The consistency of the estimates $\{f_n\}$ extends immediately to their normalized versions using Proposition 1.

Corollary 2 *Under the assumptions of Theorem 1 the L_1-error of the normalized partitioning density estimates converges to zero with probability one.* □

5 Histogram Classification

In the classification problem, a measurement vector $X \in \mathbb{R}^d$ is associated in a stochastic fashion with a class label Y taking on finitely many values. Let $(X,Y),(X_1,Y_1),(X_2,Y_2),\ldots$ be independent and identically distributed with $X \in \mathbb{R}^d$ and $Y \in \{1,\ldots,M\}$. Each measurable decision rule $g : \mathbb{R}^d \to \{1,\ldots,M\}$ has an associated error probability, or risk,

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

The decision rule minimizing $L(\cdot)$ is given by

$$g^*(x) = \arg\max_{k=1,\ldots,M} P_k(x),$$

where $P_k(x) = \mathbb{P}\{Y = k | X = x\}$ is the *a posteriori* probability of the k-th class given that $X = x$. Define $L^* = L(g^*)$.

Let g_n be a decision rule that is based on the training set $T_n = (X_1,Y_1),\ldots,(X_n,Y_n)$. The error probability of g_n is a random variable given by

$$L(g_n) = \mathbb{P}\{g_n(X) \neq Y | T_n\}.$$

A sequence $\{g_n\}$ of data-dependent decision rules is said to be strongly risk consistent if $L(g_n) \to L^*$ with probability one as n tends to infinity.

Let $\Pi = \{\pi_1,\pi_2,\ldots\}$ be a fixed partitioning scheme for $\mathbb{R}^d$. The partitioning rule π_n assigns a measurable partition of $\mathbb{R}^d$ to each sequence $(x_1,y_1),\ldots,(x_n,y_n)$ of labeled vectors. Of interest here are decision rules that are defined by forming a class-majority votes within the cells of $\pi_n(T_n)$. Supressing the dependence of $\pi_n(T_n)$ on T_n, define

$$g_n(x) = k \quad \text{if} \quad \sum_{X_i \in \pi_n[x]} I\{Y_i = k\} \geq \sum_{X_i \in \pi_n[x]} I\{Y_i = l\} \quad \text{for} \quad l = 1,\ldots,M, \tag{16}$$

where $I\{C\}$ denotes the indicator of an event C. Ties are broken in favor of the class having the smallest index. We emphasize that the partition π_n can depend on the vectors X_i, and on their labels Y_i as well.

The weak consistency of histogram classification rules whose partitions depend only on the vectors X_i may be established using the general result of Stone's (1977). The strong universal consistency of histogram classification rules based on data independent cubic partitions was shown by Devroye and Györfi (1983). Gordon and Olshen (1978), (1980), and (1984) established universal consistency for classification and regression schemes based on data-dependent,

rectangular partitioning of $\mathbb{R}^d$. The most general existing conditions for the risk consistency of the classification rules studied here can be found in the book of Breiman, Friedman, Olshen and Stone (1984). These conditions are discussed further in Section 6.

Here we establish the strong risk consistency of the rules $\{g_n\}$ for a wide class of partitioning schemes Π. The next theorem is analogous to Theorem 1 for density estimation.

Theorem 2 *For each n let $\mathcal{A}_n$ be the collection of partitions associated with the n-sample partitioning rule π_n. Let μ be the distribution of X. If as n tends to infinity*

(a) $n^{-1}m(\mathcal{A}_n) \to 0$,

(b) $n^{-1}\log \Delta_n^*(\mathcal{A}_n) \to 0$, *and*

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

$$\inf_{S:\mu(S)\geq 1-\delta} \mu\{x : \mathrm{diam}(\pi_n[x] \cap S) > \gamma\} \to 0 \quad \textit{with probability one,}$$

then the classification rules $\{g_n\}$ defined in (16) are risk consistent:

$$L(g_n) \to L^*$$

with probability one.

Theorem 2 implies the distribution free consistency of partitioning schemes for which condition (c) is satisfied for every distribution of (X,Y). An example of such a scheme will be given in Section 6. The proof of Theorem 2 relies on the following elementary inequality (c.f. Devroye and Györfi (1985)).

Lemma A *Let $\beta_1(x), \ldots, \beta_M(x)$ be real-valued functions on $\mathbb{R}^d$, and define the decision rule*

$$h(x) = \arg\max_{1\leq k\leq M} \beta_k(x).$$

Then

$$L(h) - L^* \leq \sum_{k=1}^{M} \int |P_k(x) - \beta_k(x)|\, \mu(dx)\,.$$

Proof of Theorem 2: Observe that the classification rule g_n defined in (16) can be rewritten in the form:

$$g_n(x) = \arg\max_{1\leq k\leq M} \left\{ \frac{n^{-1}\sum_{i=1}^n I\{X_i \in \pi_n[x], Y = k\}}{\mu(\pi_n[x])} \right\}.$$

For $k = 1, \ldots, M$ define

$$P_{k,n}(x) = \frac{n^{-1} \sum_{i=1}^{n} I\{X_i \in \pi_n[x], Y = k\}}{\mu(\pi_n[x])},$$

and note that by Lemma A, it is enough to show that

$$\int |P_k(x) - P_{k,n}(x)| \, \mu(dx) \to 0 \text{ a.s.}$$

for each k. Fix $k \in \{1, \ldots, M\}$ and define

$$m(x) = P_k(x) \quad \text{and} \quad m_n(x) = P_{k,n}(x).$$

Fix $\epsilon > 0$ and let $r : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support such that

$$\int |m(x) - r(x)| \mu(dx) < \epsilon.$$

Define the auxiliary functions

$$\tilde{m}_n(x) = \frac{E\left(I\{Y = k\} I\{X \in \pi_n[x]\} | \, T_n \right)}{\mu(\pi_n[x])}$$

and

$$\tilde{r}_n(x) = \frac{E\left(r(X) I\{X \in \pi_n[x]\} | \, T_n \right)}{\mu(\pi_n[x])},$$

and note that both are piecewise-constant on the cells of the partition π_n. We begin with the following upper bound:

$$\begin{aligned} & |m(x) - m_n(x)| \\ & \quad \leq \quad |m(x) - r(x)| + |r(x) - \tilde{r}_n(x)| + |\tilde{r}_n(x) - \tilde{m}_n(x)| + |\tilde{m}_n(x) - m_n(x)|. \end{aligned} \tag{17}$$

The integral of the first term on the right hand side of (17) is smaller than ϵ by the definition of $r(x)$. As for the third term,

$$\begin{aligned} \int |\tilde{r}_n(x) - \tilde{m}_n(x)| \mu(dx) \quad &= \quad \sum_{A \in \pi_n} \left| \int_A m(x) \mu(dx) - \int_A r(x) \mu(dx) \right| \\ &\leq \quad \int |m(x) - r(x)| \mu(dx) < \epsilon. \end{aligned}$$

Now let η be the distribution of $(X, I\{Y = k\})$ on $\mathbb{R}^d \times \{0, 1\}$, and let η_n be the empirical measure of $(X_1, I\{Y_1 = k\}), \ldots, (X_n, I\{Y_n = k\})$. For each partition $\pi = \{A_1, \ldots, A_r\} \in \mathcal{A}_n$, define a partition $\tilde{\pi}$ of $\mathbb{R}^d \times \{0, 1\}$ via

$$\tilde{\pi} = \{A_1 \times \{0\}, \ldots A_r \times \{0\}\} \cup \{A_1 \times \{1\}, \ldots A_r \times \{1\}\},$$

and let $\mathcal{B}_n = \{\tilde{\pi} : \pi \in \mathcal{A}_n\}$. Then

$$\begin{aligned}
\int |\tilde{m}_n(x) - m_n(x)|\mu(dx) &= \sum_{A\in\pi_n} \left| \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = k\}I\{X_i \in A\} - E\left(I\{Y = k\}I\{X \in A\}|\, T_n\right)\right| \\
&= \sum_{A\in\pi_n} |\eta_n(A\times\{1\}) - \eta(A\times\{1\})| \\
&\leq \sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} |\eta_n(A\times\{1\}) - \eta(A\times\{1\})| \\
&= \sup_{\tilde{\pi}\in\mathcal{B}_n} \sum_{B_j\in\tilde{\pi}} |\eta_n(B_j) - \eta(B)|\,.
\end{aligned}$$

It is easy to see that $m(\mathcal{B}_n) = 2m(\mathcal{A}_n)$ and $\Delta_n^*(\mathcal{B}_n) = \Delta_n^*(\mathcal{A}_n)$. In conjunction with Corollary 1 of Lemma 1, conditions (a) and (b) above imply that

$$\int |\tilde{m}_n(x) - m_n(x)|\mu(dx) \to 0 \ \text{ a.s.}$$

It remains to consider the second term on the right-hand side of (17). An application of Fubini's theorem gives the following bound:

$$\begin{aligned}
\int |r(x) - \tilde{r}_n(x)|\mu(dx) &= \sum_{A:\mu(A)\neq 0} \int_A \left| r(x) - \frac{E\left(r(X)I\{X\in A\}|\, T_n\right)}{\mu(A)}\right| \mu(dx) \\
&= \sum_{A:\mu(A)\neq 0} \frac{1}{\mu(A)} \int_A |r(x)\mu(A) - E\left(r(X)I\{X\in A\}|\, T_n\right)|\, \mu(dx) \\
&= \sum_{A:\mu(A)\neq 0} \frac{1}{\mu(A)} \int_A \left| r(x)\int_A \mu(dy) - \int_A r(y)\mu(dy)\right| \mu(dx) \\
&\leq \sum_{A:\mu(A)\neq 0} \frac{1}{\mu(A)} \int_A \int_A |r(x) - r(y)|\mu(dx)\mu(dy)\,.
\end{aligned}$$

Fix $\delta \in (0,1)$ and let $\gamma > 0$ be chosen so that if $A \subseteq \mathbb{R}^d$ satisfies $\mathrm{diam}(A) < \gamma$ then $|r(x) - r(y)| < \delta$ for every $x, y \in A$. Let $K < \infty$ be a uniform upper bound on $|r|$. Let $S \subset \mathbb{R}^d$ be such that $\mu(S) \geq 1 - \delta$. If $\mathrm{diam}(A \cap S) \geq \gamma$ then

$$\frac{1}{\mu(A)} \int_A \int_A |r(x) - r(y)|\mu(dx)\mu(dy) \leq 2K\mu(A).$$

If, on the other hand, $\mathrm{diam}(A \cap S) < \gamma$ then

$$\begin{aligned}
&\frac{1}{\mu(A)} \int_A \int_A |r(x) - r(y)|\mu(dx)\mu(dy) \\
&\leq \frac{1}{\mu(A)} \left(\int_{A\cap S}\int_{A\cap S} |r(x) - r(y)|\mu(dx)\mu(dy) + 2\int_A \int_{A\setminus S} |r(x) - r(y)|\mu(dx)\mu(dy)\right) \\
&\leq \frac{1}{\mu(A)} \left(\delta\mu^2(A) + 4K\mu(A)\mu(A\setminus S)\right) \\
&= \delta\mu(A) + 4K\mu(A\setminus S).
\end{aligned}$$

Summing over the cells $A \in \pi_n$, and noting that $\mu(S^c) < \delta$, these bounds show that

$$\int |r(x) - \tilde{r}_n(x)|\mu(dx) \quad \leq \quad 2K\mu\{x : \mathrm{diam}(\pi_n[x] \cap S) \geq \gamma\} + (4K+1)\delta\,.$$

Take the infimum of both sides above over $S \subset \mathbb{R}^d$ with $\mu(S) \geq 1 - \delta$ and then let n tend to infinity. By condition (c) of the theorem,

$$\limsup_{n\to\infty} \int |r(x) - \tilde{r}_n(x)|\mu(dx) \leq \delta(4K+1) \quad \text{a.s.}$$

In summary, we have shown that

$$\limsup_{n\to\infty} \int |m(x) - m_n(x)|\mu(dx) \leq 2\epsilon + \delta(4K+1) \ \text{ a.s..}$$

As ϵ and δ were arbitrary, the proof is complete. □

Remark: The similarity between the conditions of Theorem 1 and Theorem 2 is apparent. Condition (c) of Theorem 2 is weaker than condition (c) of Theorem 1, however, as one can see by taking $S = \mathbb{R}^d$ in the argument above. Consistent density estimation requires more stringent conditions on the diameter of the partition-cells than does consistent classification.

6 Applications

6.1 Relation to a previous result

Breiman *et al.* considered classification rules based on tree-structured partitions. Tree-structured partitions are produced recursively: beginning with a single cell containing all of $\mathbb{R}^d$, refinements are made in an iterative fashion by splitting a selected cell of the current partition with a hyperplane that is based on the data. If the rule $\pi_n(\cdot)$ makes k such splits, then the resulting partition contains $k+1$ cells, each of which is a convex polytope. Breiman *et al.* (1984) establish the consistency of classification rules g_n defined as in (16) under three conditions:

a. For every n and every training sequence T_n, each cell of $\pi_n(T_n)$ is a polytope having at most B faces, where B is fixed.

b. Each cell of π_n contains at least k_n of the vectors $X_1, \ldots, X_n$, where $k_n / \log n \to \infty$.

c. A "shrinking cell" condition that implies condition (c) of Theorem 2.

Using Theorem 2 it can be shown that conditions (b) and (c) alone suffice to insure the consistency of classification rules based on tree-structured partitioning schemes.

Theorem 3 *Let $\Pi = \{\pi_1, \pi_2, \ldots\}$ be a sequence of tree-structured partitioning rules for $\mathbb{R}^d$. Suppose that for every training sequence T_n, each cell of the partition $\pi_n(T_n)$ contains at least k_n of $X_1, \ldots, X_n$, where*

$$\frac{k_n}{\log n} \to \infty \,. \tag{18}$$

If the shrinking cell condition (c) of Theorem 2 is satisfied, then the classification rules $\{g_n\}$ based on Π are risk consistent.

Proof: Let $\mathcal{A}_n$ denote the collection of all possible partitions produced by the rule $\pi_n(\cdot)$. Each partition $\pi_n(T_n)$ contains at most n/k_n cells, so that

$$\frac{m(\mathcal{A}_n)}{n} \leq \frac{1}{k_n} \to 0 \,.$$

The recursive nature of the partitioning rule insures that each partition $\pi_n(T_n)$ is based on at most $m(\mathcal{A}_n) = n/k_n$ hyperplane splits. Each such split can dichotomize $n \geq 2$ points in $\mathbb{R}^d$ in at most n^d different ways (*cf.* Cover (1965)). It follows that the number of different ways n vectors can be partitioned by $\pi \in \mathcal{A}_n$ is bounded by

$$\Delta_n^*(\mathcal{A}_n) \leq \left(n^d\right)^{n/k_n} \,,$$

and consequently

$$\frac{1}{n} \log \Delta_n^*(\mathcal{A}_n) \leq \frac{d \log n}{k_n} \to 0 \,.$$

Thus conditions (a) and (b) of Theorem 2 are satisfied, and the proof is complete. □

6.2 k-spacing density estimates

Consider the k_n-spacing estimate of a univariate density. Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables whose distribution μ has a density f on $\mathbb{R}$. Let $X^{(1)} < X^{(2)} < \ldots < X^{(n)}$ be the order statistics obtained by a suitable permutation of $X_1, \ldots, X_n$. (This permutation exists with probability one as μ has a density.) The rule π_n partitions the real line into intervals such that each interval, with the possible exception of the rightmost, contains k_n points. Let $m = \left\lceil \frac{n}{k_n} \right\rceil$. Then

$$\pi_n(X_1^n) = \{A_1, \ldots, A_m\},$$

where

$$X^{(k_n(j-1)+1)}, \ldots, X^{(k_n j)} \in A_j \,,$$

for each $j = 1, \ldots, m-1$, and

$$X^{(k_n(m-1)+1)}, \ldots, X^{(n)} \in A_m \,.$$

Theorem 4 applies to any partition having these properties: the endpoints of the individual cells are not important. The density estimate f_n is defined by

$$f_n(x) = \begin{cases} k_n/\lambda(\pi_n[x]) & \text{if } x \in \cup_{j=1}^{m-1} A_j \\ (n - k_n(m-1))/\lambda(\pi_n[x]) & \text{if } x \in A_m \\ 0 & \text{otherwise.} \end{cases}$$

Abou-Jaoude (1976b) established the strong L_1-consistency of this estimate when the density f of μ is Riemann-integrable. An application of Theorem 1 gives the best possible result.

Theorem 4 *Let f_n be the k_n-spacing estimate given above. Then*

$$\lim_{n\to\infty} \int |f(x) - f_n(x)| dx = 0 \ \textit{a.s.}$$

if $k_n \to \infty$ and $k_n/n \to 0$ as n tends to infinity.

Remark: Abou-Jaoude (1976b) showed that the conditions on the block size k_n are necessary for universal consistency, so the conditions above are optimal.

Proof of Theorem 4: Let $\mathcal{A}_n$ contain all the partitions of $\mathbb{R}$ into $m = \left\lceil \frac{n}{k_n} \right\rceil$ intervals. Then $m(\mathcal{A}_n) \le n/k_n + 1$, so that condition (a) of Theorem 1 is satisfied. The partitioning number $\Delta_n^*(\mathcal{A}_n)$ is equal to the number of ways n fixed points can be partitioned by m intervals, so that

$$\Delta_n^*(\mathcal{A}_n) = \binom{n+m}{n} .$$

Let h be the binary entropy function, defined by $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in (0,1)$. Note that h is increasing on $(0, 1/2]$, h is symmetric about $1/2$, and that $h(x) \to 0$ as $x \to 0$. It is well known (*c.f.* Csiszár and Körner (1981)) that $\log \binom{s}{t} \le sh(t/s)$, and consequently

$$\begin{aligned} \log \Delta_n^*(\mathcal{A}_n) &\le (n+m) h\left(\frac{m}{n+m}\right) \\ &\le 2nh(1/k_n) \,. \end{aligned}$$

As $k_n \to \infty$, the last inequality implies that

$$\frac{1}{n} \log \Delta_n^*(\mathcal{A}_n) \to 0 \,,$$

which establishes condition *(b)* of Theorem 1.

Now fix $\gamma, \epsilon > 0$ and let B be so large that $\mu([-B,B]^c) < \epsilon$. Then

$$\mu\{x : \mathrm{diam}(\pi_n[x]) > \gamma\} \leq \epsilon + \mu(\, \{x : \mathrm{diam}(\pi_n[x]) > \gamma\} \cap [-B,B]\,) \,.$$

There are at most $2B/\gamma$ disjoint intervals of length greater than γ in $[-B,B]$, and consequently

$$\begin{aligned} \mu(\, \{x : \mathrm{diam}(\pi_n[x]) > \gamma\} \cap [-B,B]\,) &\leq \frac{2B}{\gamma} \max_{A \in \pi_n} \mu(A) \\ &\leq \frac{2B}{\gamma} \left(\max_{A \in \pi_n} \mu_n(A) + \max_{A \in \pi_n} |\mu(A) - \mu_n(A)| \right) \\ &\leq \frac{2B}{\gamma} \left(\frac{k_n}{n} + \sup |\mu(A) - \mu_n(A)| \right) , \end{aligned}$$

where in the last inequality the supremum is taken over all intervals in $\mathbb{R}$. The first term in the parenthesis tends to zero by assumption, while the second term tends to zero with probability one by an obvious extension of the classical Glivenko-Cantelli theorem. In summary, for any γ, $\epsilon > 0$,

$$\limsup_{n \to \infty} \mu\{x : \mathrm{diam}(\pi_n[x]) > \gamma\} \leq \epsilon \quad \text{a.s.}$$

so that condition (c) of Theorem 1 is satisfied. This completes the proof. □

6.3 Classification using statistically equivalent blocks

Classification rules based on statistically equivalent blocks are analogous to the k-spacing density estimate studied above. If the observations X_i are real-valued, then the partition for the statistically equivalent blocks classification rule agrees with the partition used by the k-spacing density estimate. Note that partitions of this sort are well-defined only if data points do not coincide.

For multivariate data the k-spacing partitioning scheme can be generalized in several ways. Consider a training sequence $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, \ldots, M\}$ such that $d \geq 2$ and the distribution of X_i has non-atomic marginals. We consider a partitioning method proposed by Gessaman (1970). Let $m_n = \left\lceil \left(\frac{n}{k_n}\right)^{1/d} \right\rceil$. Now project the vectors $X_1, \ldots, X_n$ onto the first coordinate axis. Based on these projections, partition the data into m_n sets using hyperplanes perpendicular to the first coordinate axis, in such a way that each set (with the possible

exception of the last) contains an equal number of points. This produces m_n cylindrical sets. In the same way, partition each of these cylindrical sets along the second axis into m_n boxes, such that each box contains the same number of data points. Continuing in a similar fashion along the remaining coordinate axes produces m_n^d rectangular cells, each of which (with the possible exception of those on the boundary) contains k_n points. The corresponding classification rule g_n is defined as in Section 5, by taking a majority vote among those labels Y_i whose corresponding vectors X_i lies within a given cell. The consistency of this classification rule can be established by an argument similar to that given for the k_n-spacing density estimate above. It is sufficient to verify that the conditions of Theorem 2 are satisfied. The only minor difference is in the computation of Δ_n^*, which in this case is upper bounded by $\binom{n+m}{n}^d$. The following theorem summarizes the result.

Theorem 5 *Assume that the distribution μ of X has non-atomic marginals. Then the classification rule based on Gessaman's partitioning scheme is consistent if $k_n \to \infty$ and $k_n/n \to 0$ as n tends to infinity.* □

To consider distributions with possibly atomic marginals the partitioning algorithm must be modified, since for large n every such atom will have more than k_n data points with the same corresponding component. Such a modification is possible, but it is not discussed here.

Remark: Consistency of Gessaman's classification scheme can also be derived from the results of Gordon and Olshen (1978) under the additional condition $k_n/\sqrt{n} \to \infty$. Results in Breiman *et al.* (1984) can be used to improve this condition to $k_n/\log n \to \infty$. Theorem 5 guarantees consistency under the optimal condition $k_n \to \infty$.

6.4 Clustering-based partitioning schemes

Clustering is a widely used methods of statistical data analysis. Clustering schemes divide the data into a finitely many disjoint groups by minimizing an empirical error measure, such as the average squared distance from the cluster centers. In this section we outline the application of our results to classification rules and density estimates based on nearest-neighbor clustering of the (unlabeled) measurement vectors X_i.

A *clustering scheme* is a function $C : \mathbb{R}^d \to \mathcal{C}$, where $\mathcal{C} = \{c_1, \ldots c_m\} \subseteq \mathbb{R}^d$ is a finite set of vectors known as *cluster centers*. Every clustering scheme C is associated with a partition $\pi = \{A_1, \ldots, A_m\}$ of $\mathbb{R}^d$ having cells $A_j = \{x : Q(x) = c_j\}$. A clustering scheme $C(\cdot)$ 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 π of C is just the nearest-neighbor partition of the vectors $\{c_1, \ldots, c_m\}$. See Hartigan (1975) or Gersho and Gray (1992) for more details concerning multivariate clustering and its applications.

Let $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathbb{R}^d \times \{1, \ldots, M\}$ be i.i.d. and suppose that the distribution μ of X_1 has bounded support. The *risk* of a clustering scheme C is defined to be $R(C) = \int \|x - C(x)\|^2 d\mu(x)$, and the *empirical risk* of C with respect to $X_1, \ldots, X_n$ is given by

$$\hat{R}_n(C) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - C(X_i)\|^2 . \tag{19}$$

(Here $\|\cdot\|$ denotes the usual Euclidean norm.) From a training set $T_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ and a clustering scheme C_n one may produce a classification rule g_n be taking class-majority votes within the cells of C_n. Suitable choice of C_n insures that g_n is risk consistent.

Theorem 6 *Assume that the distribution μ of X_i has bounded support. Let C_n minimize the empirical risk $R_n(C)$ over all nearest neighbor clustering schemes C with k_n cluster centers. Let the classification rule g_n be defined within the cells of C_n by a majority vote as in (16). If $k_n \to \infty$ and $n^{-1} k_n^2 \log n / n \to 0$, then $L(g_n) \to L^*$ with probability one.*

Proof: Let $\mathcal{V}_k$ be the family of all nearest-neighbor partitions of k vectors in $\mathbb{R}^d$. Then $m(\mathcal{V}_k) = k$, and every cell of a partition $\pi \in \mathcal{V}_k$ is bounded by $(k-1)$ hyperplanes representing points that are equidistant from two vectors. It is well-known (c.f. Cover (1965)) that n vectors $x_1, \ldots, x_n$ in $\mathbb{R}^d$ can be split by hyperplanes in at most n^d different ways. Therefore the cells of partitions in $\mathcal{V}_k$ can intersect $x_1, \ldots, x_n$ in at most $n^{(k-1)d}$ different ways. Each partition contains at most k cells, so that $\Delta_n^*(\mathcal{V}_k) \le n^{k^2 d}$, and consequently

$$\frac{1}{n} \log \Delta_n^*(\mathcal{V}_{k_n}) \le \frac{d k_n^2 \log n}{n} \to 0$$

by assumption. Thus condition (b) of Theorem 2 is satisfied.

It remains to establish the shrinking cell condition of Theorem 2. Fix $\gamma, \delta > 0$ and let $c_1, \ldots, c_{k_n}$ be the cluster centers of the scheme C_n that minimizes (19). Define

$$S_n = \bigcup_{j=1}^{k_n} B(c_j, \gamma/2) \cap A_j,$$

where A_j is the cell of c_j and $B(x, \alpha)$ denotes the open ball of radius α around the vector x. It is evident that

$$\mu\{x : \text{diam}(\pi_n[x] \cap S_n) > \gamma\} = 0,$$

so it suffices to show that $\mu(S_n) \to 1$ with probability one. Using a large-deviation inequality of Linder, Lugosi, and Zeger (1993) for the empirical squared error of nearest-neighbor clustering schemes, it can be shown that

$$R(C_n) \to 0 \tag{20}$$

with probability one. (Here we have made use of the fact that the X_i are bounded.) By the Markov inequality,

$$1 - \mu(S_n) \leq \left(\frac{2}{\gamma}\right)^2 R(C_n)$$

for each n, and it follows that $\mu(S_n) \to 1$ as desired. □

Suppose now that $X_1, X_2, \ldots \in \mathbb{R}^d$ are i.i.d. and that the distribution μ of X_1 has a density with bounded support. Let π_n be the partition associated with the nearest-neighbor clustering scheme C_n minimizing (19). It follows from a general result of Nobel (1995) that if $R(C_n) \to 0$ then $\text{diam}(\pi_n[X]) \to 0$ in probability. Thus (20) insures that the shrinking cell condition of Theorem 1 is satisfied, and we obtain the following analogue of Theorem 6 above.

Theorem 7 *Let C_n minimize the empirical risk $R_n(C)$ over all nearest neighbor clustering schemes C with k_n cluster centers. Let the density estimate f_n be defined within the cells of C_n as in (8). If $k_n \to \infty$ and $n^{-1}k_n^2 \log n/n \to 0$, then $\int |f_n - f| dx \to 0$ with probability one.*

References

[1] S. Abou-Jaoude. Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. *Annales de l'Institut Henri Poincaré*, 12:213–231, 1976.

[2] S. Abou-Jaoude. La convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité. *Annales de l'Institut Henri Poincaré*, 12:299–317, 1976.

[3] S. Abou-Jaoude. Sur une condition nécessaire et suffisante de L_1-converence presque complète de l'estimateur de la partition fixe pour une densité. *Comptes Rendus de l'Académie des Sciences de Paris*, 283:1107–1110, 1976.

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

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

[6] X. R. Chen and L. C. Zhao. Almost sure L_1-norm convergence for data-based histogram density estimates. *Journal of Multivariate Analysis*, 21:179–188, 1987.

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

[8] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

[9] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.

[10] L. Devroye and L. Györfi. Distribution-free exponential bound 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, 1983. Akadémiai Kiadó.

[11] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York, 1985.

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

[13] M. P. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *Annals of Mathematical Statistics*, 41:1344–1346, 1970.

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

[15] L. Gordon and R. A. Olshen. Asymptotically efficient solutions to the classification problem. *Annals of Statistics*, 6:515–533, 1978.

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

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

[18] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, empirical quantizer design, and universal lossy source coding. *IEEE Transactions on Information Theory*, 40:1728–1740, 1994.

[19] 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.

[20] P. C. Mahalanobis. A method of fractile graphical analysis. *Sankhya Series A*, 23:41–64, 1961.

[21] A. Nobel. Histogram regression estimation using data-dependent partitions. Technical report UIUC-BI-94-01, Beckman Institute, University of Illinois, Urbana-Champaign, 1994.

[22] A. Nobel. Recursive partitioning to reduce distortion. Technical report UIUC-BI-95-01, Beckman Institute, University of Illinois, Urbana-Champaign, 1995.

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

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

[25] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[26] C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 8:1348–1360, 1977.

[27] 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, Vol. II*, pages 513–520, Belmont, CA., 1985. Wadsworth.

[28] J. Van Ryzin. A histogram method of density estimation. *Communications in Statistics*, 2:493–506, 1973.

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

[30] L. C. Zhao, P. R. Krishnaiah, and X. R. Chen. Almost sure L_r-norm convergence for data-based histogram estimates. *Theory of Probability and its Applications*, 35:396–403, 1990.