

Termination and Continuity of Greedy Growing for Tree-Structured Vector Quantizers

Andrew B. Nobel, *Member, IEEE*, and Richard A. Olshen, *Member, IEEE*

Abstract—Tree-structured vector quantizers (TSVQ) provide a computationally efficient, variable-rate method of compressing vector-valued data. In applications, the problem of designing a TSVQ from empirical training data is critical. Greedy growing algorithms are a common and effective approach to the design problem. They are recursive procedures that produce a TSVQ one node at a time by optimizing a simple splitting criterion at each step. While unsupervised greedy growing algorithms are well-understood from an experimental point of view, there has been little theory to support their use, or to examine their behavior on large training sets. In this paper we present a rigorous analysis of a greedy growing algorithm proposed by Riskin and Gray, and Balakrishnan. The first part of the paper is a description of the algorithm and an examination of its asymptotic behavior as it applies to a fixed, absolutely continuous distribution. The second part of the paper establishes the structural consistency of the algorithm with respect to a convergent sequence of distributions. As an application, we obtain results concerning the large-sample empirical behavior of the algorithm when it is applied to stationary ergodic training vectors.

Index Terms—Greedy algorithms, tree-structured vector quantizers, data compression, structural consistency.

I. INTRODUCTION

TREE-STRUCTURED vector quantizers (TSVQ's) provide computationally efficient means of compressing multivariate data that arise in a variety of applications, including medical imaging and speech recognition. While lacking the optimality properties of full-search techniques, TSVQ's are easier to implement, and they possess progressive transmission properties that make them very attractive in practice. TSVQ's give rise to variable-rate codes that frequently outperform fixed-rate, full-search techniques with the same average number of bits per sample.

Greedy growing algorithms [15], [4], [25], [26], [1] are a widely used and effective method of producing TSVQ's from empirical data. Greedy algorithms produce a labeled binary tree, one node at a time, by recursively optimizing a simple splitting criterion at each step. Greedy algorithms do

not employ lookahead; recent work [27] indicates that in many cases lookahead is not very helpful.

In this paper we consider a particular greedy growing algorithm, originally proposed by Riskin [25], Riskin and Gray [26], and Balakrishnan [1]. The algorithm, henceforth referred to as *the greedy growing algorithm*, seeks to optimize a performance/complexity tradeoff at each stage of its operation. The performance of a TSVQ is judged by its distortion, and its complexity by the expected depth of the tree. While our results apply to the case in which the complexity of a tree is measured by its size (e.g., number of nodes), other notions of complexity are not being addressed here.

Although the greedy growing algorithm is well understood from an experimental standpoint, there has been little theory (cf. [1]) to support its use, or to examine its behavior on large training sets. This paper presents a rigorous analysis of the algorithm. Central to the approach taken here is the fact that the algorithm can be applied to any probability distribution on \mathbb{R}^k . The first part of the paper is a study of the algorithm as it applies to a fixed distribution having a density with bounded support. Termination of the algorithm is established when it operates with a rate-based stopping criterion, and its nonterminating behavior is examined.

The second part of the paper is an argument for the continuity of the algorithm as it applies to a fixed, convergent sequence of distributions. As an application we deduce uniform termination of the algorithm under rate-constrained operation, and establish the large-sample structural consistency of the algorithm when it is applied to the empirical distributions of stationary ergodic training vectors.

A. Greedy Growing and Medical Image Compression

Some radiographic technologies, such as computerized tomography (CT) and magnetic resonance imaging (MR) are, by nature, digital. Other modalities are becoming digital as reliable storage, retrieval, and transmission of images become increasingly important. At present the amount of data produced by such techniques (2 Mbytes for X-ray images, 1/2 Mbyte for a CT image) threatens even the most modern approaches to archiving and retrieval, and consequently compression of digital image data is imperative.

Lossless compression techniques seldom yield compression ratios greater than 4:1 in practice. On the other hand, it has been shown recently [7], [6] that in applications of CT to the detection of lung lesions and mediastinal adenopathy, one can achieve compression ratios of 10:1 without loss of clinical accuracy using lossy TSVQ produced from algorithms like

Manuscript received December 20, 1993; revised March 13, 1995. This work was supported in part by the National Science Foundation under Grants DMS-9101528 and MIP-9016974, and by the National Institutes of Health under Grants CA49697 and CA55325. This work was completed while A. Nobel was Beckman Institute Fellow at the Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign.

A. B. Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 USA.

R. A. Olshen is with the Division of Biostatistics, Stanford University, Stanford, CA 94305 USA.

Publisher Item Identifier S 0018-9448(96)00437-3.

those studied here. Application to the measurement of vessels from MR chest scans [30] has shown that there is no significant reduction in accuracy when images compressed up to 16:1 are used.

The application of TSVQ to the lossy coding of medical images is straightforward. Each image in a set of "training" images is divided into disjoint (typically rectangular) blocks containing k pixels. The intensities of pixels within the block can be viewed as a vector in k -dimensional Euclidean space \mathbb{R}^k . Thus the training images yield a sequence of vectors, each representing a region of pixel intensities. The greedy algorithm is applied to the empirical distribution of these vectors, in conjunction with some stopping criterion, to produce a labeled tree.

In some cases it is not the pixel vectors themselves to which the greedy algorithm is applied. Frequently a predictor, usually of Wiener-Hopf type, is applied to previously encoded pixel blocks in order to predict the next coded pixel block. The residuals from this fit are then quantized. If the decoder has the predictor and knows the order in which pixel blocks have been coded, the image can be reconstructed from the residuals. Readers will see readily that our mathematics applies in a straightforward way when the greedy algorithm is applied to the cited residuals.

In many applications the greedy algorithm is used to produce a large initial tree whose terminal nodes are then successively pruned off in an optimal way to minimize the increase in distortion per decrease in bit rate [5]. If the initial tree is grown for a fixed number of steps, application of our work to the pruned subtrees of the large tree is straightforward. If the initial tree is grown from an absolutely continuous distribution with a bit-rate constraint, our results imply that the tree is of finite depth, and they apply readily to the finitely many optimally pruned subtrees.

B. Relation to Previous Work

In [20] and [22] Pollard established the asymptotic consistency of design methods for full-search vector quantizers. He showed that empirically optimal full-search vector quantizers converge to the optimal quantizer Q^* with the same number of codewords when Q^* is unique. Adopting a more analytic approach, Sabin and Gray [29] established the asymptotic consistency of the generalized Lloyd algorithm for fixed-rate, full-search quantizers. In both cases the authors considered stationary, ergodic training vectors, as we do here. In [21] Pollard established a central limit theorem for the codewords of empirically optimal quantizers designed from independent training vectors.

An important aspect of our analysis is the problem posed by nonuniqueness of the greedy algorithm: the algorithm need not have a unique output when it is applied to a fixed distribution. Sabin and Gray [29] addressed nonuniqueness of the Lloyd algorithm in a direct way. Although it is similar in spirit, the analysis of variable-rate, tree-structured schemes involves additional complications.

In several respects the results presented here improve upon previous work concerning the asymptotic properties of tree-

structured statistical methods. Gordon and Olshen [10] and Breiman *et al.* [2] obtained results on the Bayes risk consistency of tree-structured classification schemes. Gordon and Olshen [10]–[12], Breiman *et al.* [2], and Butler *et al.* [3] have established sufficient conditions for the consistency of algorithms producing tree-structured schemes for regression estimation and survival analysis. In each of these cases, however, the algorithms to which the cited papers apply are supervised: in order to ensure that the terminal regions contain a minimum number of points, and that their diameters tend to 0, some stages of the algorithm ignore what would be dictated by the data and relevant optimization criteria. The analysis given here applies to an unsupervised version of the greedy growing algorithm.

Lugosi and Nobel [14], and Nobel [17], have established weak sufficient conditions for the L_2 -consistency of unsupervised classification and regression schemes based on data-dependent histograms. Their results apply to trees produced by the greedy growing algorithm [19]. LeBlanc and Crowley [13] have studied application of an unsupervised tree-structured algorithm to the empirical distributions of data in a survival analysis context.

As described below, application of the greedy algorithm is based on recursive selection of optimal partitions with respect to a simple cost criterion. As finding such partitions can be computationally prohibitive, most implementations of the algorithm select successive partitions using a two-means version of the *Lloyd algorithm* at each node. The Lloyd algorithm will eventually find a local minimum of the cost criterion, but it may not find an empirically optimal partition.

C. Summary of Results

The paper has two parts. The first, consisting of Sections II–V, provides an analysis of the greedy growing algorithm when it is applied to a fixed absolutely continuous distribution with bounded support. Definitions of tree-structured vector quantizer, distortion, rate, centroid, and optimal splitting are given in the next section. Section III gives a precise description of the greedy growing algorithm, including stopping criteria and the possibility of nonterminating output. Theorem 1 of Section IV states that under nonterminating operation the greedy growing algorithm produces a sequence of TSVQ whose expected depth tends to infinity, and whose distortion tends to zero. As an immediate corollary we establish termination of the algorithm with a rate-based stopping criterion. Section V presents a counterexample showing that the assumption of bounded support in Theorem 1 cannot be weakened in general.

The second part of the paper concerns three problems: i) uniform termination of the algorithm; ii) structural consistency of the algorithm with respect to a convergent sequence of distributions; and iii) the large sample empirical performance of the algorithm. Section VI is devoted to the nonuniqueness of greedy growing, and to the statement and discussion of Theorems 2, 3, and 4. Theorem 2 shows that for each distribution P there is a *uniform bound* on the depth of every tree produced with a fixed, rate-based stopping criterion. Theorems

3 and 4 concern the structural consistency of the algorithm. Let distributions P_1, P_2, \dots converge in a suitable fashion to an absolutely continuous distribution P with bounded support, and let B be any finite rate. Theorem 3 shows that when n is sufficiently large every tree produced from P_n under the constraint B is matched to a tree produced from the limiting distribution P under the constraint B . Matching entails structural isomorphism and the closeness of codewords. Theorem 4 addresses the special case in which P_n is the empirical distribution of a stationary ergodic training sequence.

Proofs of Theorems 2–4 are given in Section VIII. Section VII contains a number of definitions and technical preliminaries. The Appendix contains the proofs of several technical results that are stated in the text.

II. PRELIMINARIES

A. Vector Quantizers

A *vector quantizer* is a mapping $Q : \mathbb{R}^k \rightarrow \mathcal{C}$, where \mathbb{R}^k is k -dimensional Euclidean space, and $\mathcal{C} = \{c_1, \dots, c_N\}$ is a finite collection of vectors in \mathbb{R}^k known as the *codebook* of Q . Thus Q assigns each vector $x \in \mathbb{R}^k$ to a representative $c_i \in \mathcal{C}$, and in this way Q induces a partition of \mathbb{R}^k with cells $A_i = \{x : Q(x) = c_i\}$, $i = 1, \dots, N$. In statistical terminology, Q is a multivariate clustering scheme.

B. TSVQ and Associated Regions

Let T be a binary tree with a single root node. The *depth* of a node $v \in T$ is the length of the path leading from the root to v . The root node itself has depth zero, its children have depth one, and so on. The terminal nodes (leaves) of T will be denoted by \tilde{T} . A binary tree T' is said to be a *subtree* of T , written $T' \leq T$, if T' and T share the same root and every node of T' is a node of T . If $T' \leq T$ and $T' \neq T$ then T' is said to be a *proper subtree* of T , written $T' < T$.

A tree-structured vector quantizer is described by a binary tree T whose nodes are labeled with distinct vectors in \mathbb{R}^k . Let Q_T be the quantizer corresponding to T . The representative $Q_T(x)$ of a vector x is determined by a sequence of binary comparisons that trace a path through T , beginning at the root node: at each internal node v , x moves to that child of v whose label is nearest to x in Euclidean distance; the representative $Q_T(x)$ is the vector labeling the terminal node where the path ends. Thus vectors labeling the terminal nodes of T form the codebook of Q_T .

In this way every labeled tree T corresponds to a hierarchical partitioning scheme for \mathbb{R}^k . Each node $v \in T$ can be associated with a region $V \subseteq \mathbb{R}^k$ in a recursive fashion. If v is the root node, let $V = \mathbb{R}^k$. Otherwise, v has a parent u whose associated region U has been previously defined. Suppose that vectors a and b label the node v and its sibling v' , respectively. Then

$$V = \{x \in U : \|x - a\| \leq \|x - b\|\}$$

while the sibling v' of v has an associated region

$$V' = \{x \in U : \|x - b\| \leq \|x - a\|\}.$$

Here $\|\cdot\|$ denotes the ordinary Euclidean norm on \mathbb{R}^k . The region V consists of all those vectors x whose path from the root node takes them through v ; V' is characterized similarly. If v lies at depth k in T , it is easy to see that V is a closed, convex polytope with at most k faces. The regions associated with terminal nodes $v \in \tilde{T}$ are called *terminal regions*, and will also be denoted by \tilde{T} .

It is important to note that regions associated with distinct nodes may overlap on their boundaries. In applying the quantizer, a tie-breaking scheme must be used to encode vectors that are equidistant from two labels. While the terminal regions of T do not form a partition of \mathbb{R}^k , the presence of ties does not affect the analysis for absolutely continuous distributions, as the regions of overlap have probability zero.

C. Distortion and Rate of TSVQ

The performance of a quantizer that assigns regions of \mathbb{R}^k to one or more representatives can be evaluated in terms of a *distortion measure* $\rho : \mathbb{R}^k \times \mathbb{R}^k \rightarrow [0, \infty)$ where $\rho(a, b)$ is the distortion incurred when a is represented by b . We assume that $\rho(a, b) = \phi(\|a - b\|)$, where $\phi : \mathbb{R} \rightarrow [0, \infty)$ is a univariate function having the following properties:

- 1) $\rho_1: \phi(0) = 0$;
- 2) $\rho_2: \phi$ has a continuous derivative ϕ' that is positive almost everywhere.

It follows that $\rho(a, a) = 0$ for every a , and that ρ is a continuous function on the product $\mathbb{R}^k \times \mathbb{R}^k$. Note that the r th power distortion $\rho(a, b) = \|a - b\|^r$ satisfies our conditions when $r \geq 1$.

When a tree-structured quantizer Q_T is applied to a random vector $X \in \mathbb{R}^k$ having distribution P , its performance will be measured in terms of the *expected distortion*,

$$D(T, P) = E\rho(X, Q_T(X)) = \int \rho(x, Q_T(x)) dP(x).$$

The complexity of a tree-structured vector quantizer Q_T will be measured in terms of its rate. The *rate* $R(T, P)$ of T with respect to P is just the expected depth of the terminal nodes of T

$$R(T, P) = \sum_{v \in \tilde{T}} \text{depth}(V) \cdot P(V).$$

Equivalently, $R(T, P)$ is the expected number of comparisons needed to quantize a random vector X that is distributed according to P . In many cases, $R(T, P)$ is less than the number of comparisons needed to perform the table lookup of an optimally chosen full search quantizer with the same expected distortion. When no ambiguity will arise, the distortion $D(T, P)$ and the rate $R(T, P)$ will be written as $D(T)$ and $R(T)$, omitting reference to the distribution P .

D. Properties of Centroids

Fix a distribution P on \mathbb{R}^k and consider a region $V \subseteq \mathbb{R}^k$ with $P(V) > 0$. If every vector in V is assigned to a single representative c , the associated average distortion is given by

the integral

$$\int_V \rho(x, c) dP(x). \quad (1)$$

Any vector $c \in \mathbb{R}^k$ that minimizes this integral is said to be a *centroid* for V with respect to P . Define

$$\mathcal{C}_1(V, P) = \left\{ c \in \mathbb{R}^k : \int_V \rho(x, c) dP = \inf_{a \in \mathbb{R}^k} \int_V \rho(x, a) dP \right\}$$

to be the collection of all centroids for V with respect to P . Note that $\mathcal{C}_1(V, P)$ can contain more than one vector.

If every vector in $x \in V$ may be assigned to one of two vectors a and b , then the expected distortion is minimized by assigning x to a when $\rho(x, a) \leq \rho(x, b)$, and assigning x to b otherwise. Using the notation $\alpha \wedge \gamma = \min(\alpha, \gamma)$, this assignment yields an overall distortion

$$\int_V \rho(x, a) \wedge \rho(x, b) dP(x) \quad (2)$$

and splits V into two regions

$$\begin{aligned} V_{ab} &= \{x \in V : \|x - a\| \leq \|x - b\|\} \\ &= \{x \in V : \rho(x, a) \leq \rho(x, b)\} \\ V_{ba} &= \{x \in V : \|x - b\| \leq \|x - a\|\} \\ &= \{x \in V : \rho(x, b) \leq \rho(x, a)\}. \end{aligned}$$

The definition of ρ ensures the second equality in each case. The boundary between V_{ab} and V_{ba} is the hyperplane $S = \{x : \|x - a\| = \|x - b\|\}$, which represents the perpendicular bisector of the line connecting a and b . The pair (a, b) is said to be a *centroid pair* for V with respect to P if it minimizes the distortion integral (2). Define

$$\begin{aligned} \mathcal{C}_2(V, P) &= \left\{ (a, b) \in \mathbb{R}^k \times \mathbb{R}^k : \int_V \rho(x, a) \wedge \rho(x, b) dP \right. \\ &= \left. \inf_{c, d \in \mathbb{R}^k} \int_V \rho(x, c) \wedge \rho(x, d) dP \right\} \end{aligned}$$

to be the collection of all centroid pairs for V with respect to P . The regions V_{ab} , V_{ba} associated with a centroid pair (a, b) are said to be a *best split* for V .

The following lemmas establish properties of centroids and centroid pairs that will be needed in the remainder of the paper. If V is a subset of \mathbb{R}^k , then V° denotes the interior of V .

Lemma 1: Let P be a distribution on \mathbb{R}^k that is absolutely continuous with respect to Lebesgue measure. If $V \subseteq \mathbb{R}^k$ is convex and $V^\circ = \emptyset$, then $P(V) = 0$.

Proof: If V is convex and $V^\circ = \emptyset$, then it follows from standard facts about convex sets [28] that V is contained in an affine subspace S of \mathbb{R}^k having dimension not more than $k - 1$. As the Lebesgue measure $\lambda(S) = 0$ for any such set, $P(S) = 0$, and consequently $P(V) = 0$. \square

Definition [Convex Support]: Let P have a density f with respect to Lebesgue measure. The *convex support* of P , denoted S_P , is the closed convex hull of $\{x : f(x) > 0\}$. If P has bounded support, then S_P is compact.

Lemma 2: Let P be a distribution on \mathbb{R}^k with a density and bounded support. If $V \subseteq \mathbb{R}^k$ has $P(V) > 0$, then

- a) $\mathcal{C}_1(V, P) \subseteq S_P$.
- b) $\mathcal{C}_1(V, P) \neq \emptyset$
- c) $\mathcal{C}_2(V, P) \neq \emptyset$.

Moreover, for any centroid pair $(a, b) \in \mathcal{C}_2(V, P)$,

- d) $P(V_{ab}), P(V_{ba}) > 0$
- e) $a \in \mathcal{C}_1(V_{ab}, P)$ and $b \in \mathcal{C}_1(V_{ba}, P)$.

Proof a): Note that the distortion integral (1) is unchanged if we replace V by $V \cap S_P$. Thus we may assume that $V \subseteq S_P$. A straightforward argument using the monotonicity of ϕ shows that any centroid for V must lie in the closed convex hull V^* of V . As $V \subseteq S_P$, $V^* \subseteq S_P$, and the result follows.

Proof b), c): The existence of a centroid for V with respect to P follows from the compactness of S_P and the continuity of the integral (1) with respect to the representative c . The argument for part c) is similar.

Proof d): Suppose to the contrary that $P(V_{ba}) = 0$. As P has a density, there is a vector c such that $P(V_{ac}), P(V_{ca}) > 0$. By an easy argument,

$$\begin{aligned} \int_V \rho(a, x) \wedge \rho(b, x) dP &= \int_V \rho(a, x) dP \\ &> \int_V \rho(a, x) \wedge \rho(c, x) dP. \end{aligned}$$

This contradicts the fact that (a, b) is a centroid pair for V with respect to P .

Proof e): These relations follow immediately from Lloyd's necessary conditions for optimality (c.f. [9]). \square

III. THE GREEDY GROWING ALGORITHM

A. The Design Problem for TSVQ

Let X_1, X_2, \dots be stationary ergodic random vectors, with each X_i distributed according to some unknown distribution P . The basic *design problem* for us is like that of [26] and [14]. We successively balance improved performance and increased complexity by splitting a node so as to maximally decrease distortion while minimally increasing the average depth of the resulting binary tree. In implementations of the algorithm, the decoder knows the codewords (centroids) associated with each terminal node as well as the rule by which an image being coded was scanned. Decoding is by simple table lookup. Therefore, for each pixel block, what is archived or transmitted is only its path from root to terminal node. Thus the complexity of a tree may be judged in terms of its expected depth. The design problem may be stated as follows:

For a fixed rate B , use the training set X_1, \dots, X_n to produce a labeled tree T such that $R(T, P) \leq B$, and $D(T, P)$ is as small as possible.

Note that both the rate R and the distortion D are expressed in terms of the *unknown* distribution P . In practice one may, as an approximation, replace P by the empirical distribution \hat{P}_n of the training set. Unfortunately, even the simpler problem of minimizing $D(Q, \hat{P}_n)$ subject to the constraint $R(T, \hat{P}_n) \leq B$ is not computationally feasible. The greedy growing algorithm

addresses the design of TSVQ by finding a computationally efficient, approximate solution to this latter problem.

B. Description of the Algorithm

The greedy growing algorithm [25], [26], [1] is applied to a fixed probability distribution P on \mathbb{R}^k . In experimental situations P is the empirical distribution of a set of training vectors. The algorithm produces a nested sequence of labeled binary trees. At each stage of its operation the greedy algorithm splits a terminal node/region of the tree T produced at the previous stage. A candidate node/region is selected in order to maximize the decrease in distortion per increase in rate. Termination of the sequence depends both on the distribution P and on the stopping criterion used by the algorithm.

Fix a distribution P and a labeled binary tree T . Let v be a terminal node of T with an associated region V . Passing from an optimal representative $c \in \mathcal{C}_1(V, P)$ to an optimal pair $(a, b) \in \mathcal{C}_2(V, P)$ yields a decrease in distortion

$$\Delta D^*(V) = \int_V \rho(x, c) dP(x) - \int_V \rho(x, a) \wedge \rho(x, b) dP(x).$$

Accommodating two representatives for V entails splitting the node v and labeling its children with the vectors a and b . In the new tree one must perform an additional comparison for every vector $x \in V$. Thus the corresponding increase in the expected depth of T is just the probability of V under P :

$$\Delta R(V) = P(V).$$

Thus splitting a terminal node $v \in T$ entails a reduction in distortion, measured by $\Delta D^*(V)$, and an increase in rate, measured by $\Delta R(V) = P(V)$. The proof of the next lemma is similar to that of Lemma 2 d).

Lemma 3: If P has a density and $V \subseteq \mathbb{R}^k$ satisfies $P(V) > 0$, then there is a split of V for which $\Delta D^*(V) > 0$. \square

At each stage of its operation the greedy growing algorithm seeks to maximize the reduction of distortion per increase in bit rate. In the absence of an external stopping criterion, application of the algorithm to a fixed distribution P can be described formally as follows.

- 1) **[Initialize]:** Set the iteration count $r = 0$ and let T_0 consist of a single node labeled by a centroid $c \in \mathcal{C}_1(P, \mathbb{R}^k)$.
- 2) **[Iteration]:** If $\Delta D^*(V) = 0$ for every terminal region $V \in \tilde{T}_r$, then stop. Otherwise, find a region $V \in \tilde{T}_r$ for which

$$\frac{\Delta D^*(V)}{\Delta R(V)} = \max_{V' \in \tilde{T}_r} \frac{\Delta D^*(V')}{\Delta R(V')}. \quad (3)$$

Form T_{r+1} by splitting the terminal node v associated with V and labeling its children with the components of a centroid pair $(a, b) \in \mathcal{C}_2(P, V)$ that achieves $\Delta D^*(V)$. Increment r and return to the beginning of step 2).

The algorithm produces a nested sequence $T_0 \leq T_1 \leq \dots$ of labeled binary trees. If P is absolutely continuous, an inductive application of Lemmas 2 and 3 shows that the algorithm does not terminate. If P is the empirical distribution of a training

set of size n , the algorithm terminates when, after at most n steps, every terminal region contains a single atom.

In practice the algorithm employs a *stopping criterion* that is meant to ensure its termination. An *iteration-based* stopping criterion stipulates that the algorithm should run for a fixed (finite) number of steps, or until $\Delta D^*(V) = 0$ for every terminal region. In this case, the algorithm is guaranteed to produce a finite sequence of trees. As dictated by the design problem, it is common in practice to impose a *rate-based* stopping criterion B . In this case the algorithm terminates when splitting would produce a tree with $R(T, P) > B$, or when $\Delta D^* = 0$ for every terminal region. Termination of the algorithm with a rate-based stopping criterion is not immediate: there exist infinite labeled trees with finite expected depth. This problem is addressed in the next two sections.

IV. PERFORMANCE ON A FIXED DISTRIBUTION

Here we consider the behavior of the greedy algorithm when it is applied to a fixed distribution P on \mathbb{R}^k . Throughout this section, and in the remainder of the paper, we make two critical assumptions regarding P :

- A1). There exists a bounded set $A \subset \mathbb{R}^k$ such that $P(A) = 1$.
- A2). P has a density with respect to Lebesgue measure.

The principal results of this section are summarized in the following theorem, whose proof appears following several preliminary lemmas.

Theorem 1: Let $T_0 < T_1 < T_2, \dots$ be a nonterminating sequence of labeled trees produced by applying the greedy growing algorithm to a distribution P that satisfies A1) and A2). Then

- a) $R(T_r) \rightarrow \infty$ as $r \rightarrow \infty$.
- b) $D(T_r) \rightarrow 0$ as $r \rightarrow \infty$.

As an immediate corollary of Theorem 1 it is evident that greedy growing with a rate-based stopping criterion will always produce a finite tree.

Corollary 1: If the greedy growing algorithm is applied to a distribution P with a rate-based stopping criterion $B < \infty$, it will produce only a finite number of distinct trees before terminating. \square

Definition: A binary tree T contains a balanced subtree of depth k if, for each $j = 0, \dots, k$, T contains all 2^j possible nodes at distance j from the root.

Lemma 4: Let $T_0 < T_1 < T_2 < \dots$ and P be as in Theorem 1. For every $k \geq 1$ there is an integer $r(k, P)$ such that T_r contains a balanced subtree of depth k whenever $r \geq r(k, P)$.

Proof: Assume to the contrary that there is an integer $k \geq 1$ such that none of T_1, T_2, \dots contains a balanced subtree of depth k . Then each tree T_r must have a terminal node whose depth is at most $k - 1$. As T_1, T_2, \dots are nested, there exists an integer r_0 and a *fixed* node u such that $u \in \tilde{T}_r$ for every $r \geq r_0$. The terminal region U associated with u has $P(U) > 0$ by Lemma 2 d). As this probability is not concentrated at a single point

$$\eta \triangleq \Delta D^*(U) / \Delta R(U) > 0.$$

For each $r \geq r_0$ the algorithm selects and splits a node $v \in \bar{T}_r$ that is different from u , and therefore $\Delta D^*(v)/\Delta R(v) \geq \eta$. Consequently, there is a sequence of nodes v_1, v_2, \dots such that for each k :

- i) $v_k \in T_r$ when r is sufficiently large
- ii) v_{k+1} is a descendant of v_k
- iii) $\Delta D^*(v_k)/\Delta R(v_k) \geq \eta$.

Let a_k and b_k be the labels assigned to the children of v_k . By Lemma 2 a) each of the sequences $\{a_k\}$ and $\{b_k\}$ is contained in the compact set S_P . Thus there exist vectors $a^*, b^* \in S_P$ and integers k_1, k_2, \dots such that $a_{k_j} \rightarrow a^*$ and $b_{k_j} \rightarrow b^*$ as j tends to infinity. The nodes $\{v_{k_j}\}$ continue to possess properties i), ii), and iii). In what follows $\{v_{k_j}\}$ will be indexed as $\{v_m\}$ for simplicity.

Suppose that $a^* = b^* \triangleq c^*$. For each index m in the sequence above, the vectors a_m and b_m that label the children of node v_m split its associated region V_m into two subregions

$$V_m^a = \{x \in V_m : \rho(x, a_m) \leq \rho(x, b_m)\}$$

and

$$V_m^b = \{x \in V_m : \rho(x, b_m) \leq \rho(x, a_m)\}.$$

Letting $\alpha \vee \beta = \max(\alpha, \beta)$, it follows from the definition of ΔD^* that

$$\begin{aligned} \Delta D^*(V_m) &\leq \int_{V_m} \rho(c^*, x) dP(x) - \int_{V_m^a} \rho(a_m, x) dP(x) \\ &\quad - \int_{V_m^b} \rho(b_m, x) dP(x) \\ &\leq \int_{V_m^a} |\rho(c^*, x) - \rho(a_m, x)| dP(x) \\ &\quad + \int_{V_m^b} |\rho(c^*, x) - \rho(b_m, x)| dP(x) \\ &\leq P(V_m) \cdot \sup \{ |\rho(c^*, x) - \rho(a_m, x)| \\ &\quad \vee |\rho(c^*, x) - \rho(b_m, x)| : x \in S_P \}. \end{aligned}$$

As m tends to infinity $a_m, b_m \rightarrow c^*$, and as $\rho(\cdot, \cdot)$ is uniformly continuous on $S_P \times S_P$ the inequality above shows that $\Delta D^*(V_m)/P(V_m) \rightarrow 0$. However, this contradicts property iii) of $\{v_m\}$, and we conclude that $a^* \neq b^*$.

Let $\delta = \|a^* - b^*\| > 0$ and select an integer m_0 so that $\|a_m - a^*\| < \delta/8$ and $\|b_m - b^*\| < \delta/8$ whenever $m \geq m_0$. Fix an index $m > m_0$. It is easy to see that $\|a_m - b_m\| \geq 3\delta/4$, while $\|a_m - a_{m_0}\| < \delta/4$ and $\|b_m - b_{m_0}\| < \delta/4$. Therefore, a_m is contained in the interior of $V_{m_0}^a$, while b_m is contained in the interior of $V_{m_0}^b$. But v_m is a proper descendent of v_{m_0} , and therefore the vectors a_m and b_m labeling the children of v_m must lie within one of $V_{m_0}^a$ or $V_{m_0}^b$, but not both. We again arrive at a contradiction, exhausting the possibilities for a^* and b^* , and conclude that the assertion of the lemma is valid. \square

Proof of Theorem 1: The proof of part a) follows immediately from Lemma 4, as any tree T_r containing a balanced subtree of depth k has rate $R(T_r) \geq k$.

As for part b), let trees $S_0 \leq S_1 \leq \dots$ be produced by a greedy growing algorithm that aims to maximize the reduction in distortion at each stage of its operation. Formally, such a procedure is described by the algorithm of Section III when

(3) is replaced by

$$\Delta D^*(V) = \max_{V' \in \bar{T}_r} \Delta D^*(V').$$

It is shown in [18] that $D(S_k) \rightarrow 0$ as $k \rightarrow \infty$. Moreover, it is clear that $S_k \leq T_r$ when T_r contains a balanced subtree of depth k . As splitting nodes (in any order) always reduces the distortion of a tree, Lemma 4 shows that for each k

$$\lim_{r \rightarrow \infty} D(T_r) = \inf_r D(T_r) \leq D(S_k).$$

The result follows by letting k tend to infinity. \square

V. A COUNTEREXAMPLE

Theorem 1 is proved under the assumption that the distribution P is supported on a bounded subset of \mathbb{R}^k . The following example shows that this condition is necessary in some cases, and that it cannot be replaced by a weaker condition involving finite moments of P or even the existence of a moment generating function in a neighborhood of 0.

Set the dimension $k = 1$, and let $\rho(x, y) = |x - y|^2$ be the squared error distortion on \mathbb{R} . Take P to be the one-sided exponential distribution with density

$$f(x) = \begin{cases} \exp(-x), & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$$

Note that P has a finite moment generating function on $(-1, \infty)$. Let U be a subset of \mathbb{R} with $P(U) > 0$, and suppose that every point of U is represented by the centroid

$$c = P(U)^{-1} \int_U x f(x) dx.$$

Suppose that U is split into two regions U_1 and U_2 , which are represented by their respective centroids

$$c_1 = P(U_1)^{-1} \int_{U_1} x f(x) dx$$

and

$$c_2 = P(U_2)^{-1} \int_{U_2} x f(x) dx. \quad (4)$$

Then the relative decrease in the integrated squared error over U is given by the equation [2, sec. 9.3]

$$\frac{\Delta D(U_1, U_2)}{P(U)} = \frac{P(U_1)P(U_2)}{P^2(U)} |c_1 - c_2|^2. \quad (5)$$

Let the greedy growing algorithm be applied to P with no stopping criterion. For simplicity, the terminal regions of the resulting trees will be described as subsets of $[0, \infty)$, the support of P . At the first stage of its operation, the algorithm finds a best split $[0, a]$, $[a, \infty)$ of the region $V_0 = [0, \infty)$ associated with T_0 . Numerical evaluation using (5) shows that

$$a \approx 1.5936 \text{ and } \eta_0 = \frac{\Delta D^*(V_0)}{\Delta R(V_0)} \approx 0.6476.$$

The algorithm splits $[0, \infty)$ producing a tree T_1 with terminal regions $V_1 = [0, a]$ and $V_2 = [a, \infty)$.

An easy calculation shows that (4) and (5) hold for every region $U \subseteq V_2 = [a, \infty)$ if f is replaced by its suitably normalized restriction to V_2

$$\tilde{f}(x) = \frac{f(x)}{\int_a^\infty f(y)dy}, \quad x \in [a, \infty).$$

But $\tilde{f}(x)$ is just a shifted exponential density concentrated on $[a, \infty)$. It follows that the best split of $U = [a, \infty)$ is just $[a, 2a]$, $[2a, \infty)$. Moreover, the value $\Delta D^*(V_2)/\Delta R(V_2)$ of this split is simply $\eta_0 \approx 0.6476$, as before. As for V_1 , a numerical evaluation shows that

$$\eta_1 = \frac{\Delta D^*(V_1)}{\Delta R(V_1)} \approx 0.1385.$$

As $\eta_0 > \eta_1$, the algorithm splits the infinite interval $V_2 = [a, \infty)$, producing a tree T_2 with terminal regions $W_1 = [0, a]$, $W_2 = [a, 2a]$, and $W_3 = [2a, \infty)$. The renormalization property above shows that

$$\frac{\Delta D^*(W_1)}{\Delta R(W_1)} = \frac{\Delta D^*(W_2)}{\Delta R(W_2)} = \eta_1 \quad \text{while} \quad \frac{\Delta D^*(W_3)}{\Delta R(W_3)} = \eta_0.$$

Consequently, the greedy algorithm splits W_3 into $[2a, 3a]$ and $[3a, \infty)$, producing a new tree T_3 .

Continuing in this fashion, it can be seen that the algorithm produces a nonterminating sequence of trees $T_0 < T_1 < \dots$ in such a way that T_{r+1} is obtained by splitting the unbounded terminal region $[ra, \infty)$ of T_r into $[ra, (r+1)a]$ and $[(r+1)a, \infty)$. By organizing the terminal regions of T_r in increasing order, from left to right, one obtains an unbalanced tree each of whose leaves emanate from a single, long spine. In particular

$$\begin{aligned} R(T_r) &= \sum_{k=0}^r k \cdot P[ka, (k+1)a] + r \cdot P[ra, \infty) \\ &= \sum_{k=0}^{r-1} e^{-ak} \end{aligned}$$

and consequently

$$\lim_{r \rightarrow \infty} R(T_r) = \frac{1}{1 - e^{-a}} \approx 1.255$$

is finite. Moreover, as the algorithm never splits a bounded terminal region of any tree T_r , it is clear that

$$\lim_{r \rightarrow \infty} D(Q_{T_r}) > 0.$$

To summarize, for the one-sided exponential distribution, neither of the conclusions of Theorem 1 is valid.

VI. ENSEMBLE PROPERTIES OF GREEDY GROWING

Previous sections considered the behavior of the greedy growing algorithm when it is applied to a fixed, absolutely continuous distribution P with bounded support. This second part of the paper considers three problems: i) uniform termination of the algorithm; ii) structural consistency of the algorithm with respect to a convergent sequence of distributions; and iii) the large sample empirical performance of the algorithm.

An important aspect of each problem is the nonuniqueness of greedy growing, which necessitates viewing the output of the algorithm as an ensemble or collection of trees. Nonuniqueness is described in the next subsection. Subsequent subsections are devoted to the statement and discussion of Theorems 2, 3, and 4, which address problems i), ii), and iii), respectively.

A. Nonuniqueness

Study of the greedy growing algorithm is complicated by the fact that the algorithm is not guaranteed to produce a unique sequence of trees from a given distribution. Nonuniqueness arises from ties that may occur at various stages of the algorithm's operation. A tie *between* nodes occurs when two or more terminal nodes maximize the ratio $\Delta D^*/\Delta R$. In this case, the algorithm is allowed to split any one of these nodes and then continue. A tie *within* a node occurs when a terminal node maximizing $\Delta D^*/\Delta R$ has more than one best split. In this case, the algorithm can select any of the best splits and then continue. Ties are the result of symmetries in the underlying distribution P . For instance, if P is multivariate Gaussian with unit covariance and $\rho(x, y) = \|x - y\|^2$, there are infinitely many best splits of the root node \mathbb{R}^k , and for each of these the two resulting regions yield the same value of $\Delta D^*/\Delta R$.

Nonuniqueness can be addressed by assuming that the greedy algorithm produces a unique tree from the underlying distribution P . However, as the example above indicates, common distributions violate this assumption, and there are no natural conditions that ensure it is satisfied. Tie-breaking schemes provide another approach to nonuniqueness, but it appears that no natural tie-breaking scheme exists. Here nonuniqueness of the algorithm is addressed directly, by describing its behavior in terms of an *ensemble* of possible outcomes.

Definition: Fix a distribution P on \mathbb{R}^k and a rate $B < \infty$. Let $\mathcal{S}(P, B)$ denote the set of all possible labeled trees T produced when the greedy growing algorithm is applied to P with any rate-based stopping criterion $B' \leq B$.

B. Uniform Termination

Let P be an absolutely continuous distribution with bounded support. If B is finite Theorem 1 shows that every $T \in \mathcal{S}(P, B)$ has finite depth. One can show more generally that there is a *uniform* bound on the depth of *every* tree in $\mathcal{S}(P, B)$.

Theorem 2: There is an integer $M = M(P, B)$ such that every tree in $\mathcal{S}(P, B)$ is produced within M iterations of the greedy growing algorithm. Equivalently, there is an integer $K = K(P, B)$ such that the maximum depth of every tree in $\mathcal{S}(P, B)$ is at most K .

C. Structural Consistency

TSVQ are typically produced from experimental data. Consider a sequence of n training vectors drawn from a stationary ergodic process whose underlying distribution P is unknown. In most applications, the greedy growing algorithm is applied to the empirical distribution of the training vectors. Alternatively, one might use the vectors to form a more sophisticated estimate of P and then apply the algorithm to this estimate.

In either case, the input to the algorithm will vary with n , and it will not in general be absolutely continuous. It is reasonable to assume, however, that as the number of training vectors grows, the estimates will improve, approaching P in the limit as $n \rightarrow \infty$. If the unknown distribution P is smooth and has bounded support, one may ask whether trees produced from the estimates resemble trees produced from P . Though nonuniqueness precludes a direct answer in terms of an individual sequence of trees, Theorems 3 and 4 below answer this question in the affirmative.

Definition [Convergence of Distributions]: Let P_1, P_2, \dots , P be probability distributions on \mathbb{R}^k . The sequence $\{P_n\}$ is said to converge to P , written $P_n \rightarrow P$, if

- i) $P_n(S_P) = 1$ for each n .
- ii) $\int f dP_n \rightarrow \int f dP$ for every measurable function f that is bounded on Λ .

More generally, S_P may be replaced by any compact, convex set having P -probability one. In what follows P will be absolutely continuous. No such assumptions are made on the individual P_i .

Theorem 3: Let P_1, P_2, \dots converge to an absolutely continuous distribution P with compact support. For every $\epsilon, \delta > 0$ and every $B < \infty$ there is a finite integer N having the following property: for each tree

$$T \in \bigcup_{n=N}^{\infty} \mathcal{S}(P_n, B)$$

there is a tree

$$T' \in \mathcal{S}(P, B)$$

that is close to T in the sense that

- i) T and T' are isomorphic as directed graphs.
- ii) Vectors labeling corresponding nodes of T and T' are at most ϵ apart.
- iii) $P\{x : \|Q_T(x) - Q_{T'}(x)\| > \epsilon\} < \delta$.

The isomorphism condition i) says that there is a bijection $\theta : T \rightarrow T'$ which preserves the ancestral relation, e.g., $\theta(u)$ is an ancestor of $\theta(v)$ if and only if u is an ancestor of v . A stronger version of Theorem 3, involving a notion of structural isomorphism based on admissible sequences, is proved in Section VIII.

One immediate consequence of Theorem 3 is worth mentioning. Let $|T|$ denote the number of nodes in T . By Theorem 2 there exists an integer K such that $|T| \leq K$ for every $T \in \mathcal{S}(P, B)$. It follows from part i) of the theorem that when n is sufficiently large, every tree $T \in \mathcal{S}(P_n, B)$ has at most K nodes.

D. Large Sample Performance

Let $X_1, X_2, \dots \in \mathbb{R}^k$ be a stationary ergodic sequence of training vectors with $X_i \sim P$, and let \hat{P}_n be the empirical distribution of X_1, \dots, X_n . The ergodic theorem shows that for every measurable function f that is bounded on S_P the integrals $\int f d\hat{P}_n \rightarrow \int f dP$ with probability one. The proof of Theorem 3 relies on two applications of Theorem 1, which appears in the Appendix. Each application of Theorem

1 involves the integrals of countably many approximating functions. Consequently, the almost everywhere convergence guaranteed by the ergodic theorem is enough to show that, for the purposes of our analysis, $\hat{P}_n \rightarrow P$ with probability one. In this way Theorem 3 may be applied to analyze the large sample performance of the greedy growing algorithm.

Theorem 4: For almost every sample sequence of the process $\{X_i\}$, every $\epsilon, \delta > 0$, and every $B < \infty$, there is a finite integer N (depending on the sample sequence) having the following property: for each tree

$$T \in \bigcup_{n=N}^{\infty} \mathcal{S}(\hat{P}_n, B)$$

there is a tree

$$T' \in \mathcal{S}(P, B)$$

that is close to T in the sense that

- i) T and T' are isomorphic as directed graphs.
- ii) Vectors labeling corresponding nodes of T and T' are at most ϵ apart.
- iii) $P\{x : \|Q_T(x) - Q_{T'}(x)\| > \epsilon\} < \delta$. □

VII. TECHNICAL PRELIMINARIES

This section contains a number of preliminary definitions and lemmas that are central to the proofs of Theorems 2-4 in Section VI. Studying the ensemble properties of the greedy growing algorithm requires an analytical means of describing trees and the process by which they are constructed. In the next two subsections it is shown that each tree produced by the algorithm can be represented by a sequence of vectors, called a trajectory, that encodes the stepwise production of the tree from a single root node. Convergence of sets, tree isomorphisms, and several related lemmas are presented in Subsection VII-C, while Subsection VII-D studies the topological properties of trajectories.

A. Admissible Sequences

An admissible sequence is a concatenation of centroid-pairs that is in direct correspondence with a labeled binary tree. Admissible sequences and their corresponding trees are defined inductively with respect to a fixed distribution P on \mathbb{R}^k .

A vector-pair $\mathbf{c} = (a, b)$ in $\mathbb{R}^k \times \mathbb{R}^k$ is admissible if $\mathbf{c} = \mathbf{0}$ or if $\mathbf{c} \in \mathcal{C}_2(\mathbb{R}^k, P)$. If $\mathbf{c} = \mathbf{0}$ let $T(\mathbf{c})$ consist of a single node labeled with any vector in $\mathcal{C}_1(\mathbb{R}^k, P)$. If $\mathbf{c} = (a, b) \neq \mathbf{0}$, let $T(\mathbf{c})$ consist of a root node with two children. The root node is labeled by any vector in $\mathcal{C}_1(\mathbb{R}^k, P)$, and its children are labeled by a and b in either order.

Suppose now that $r > 1$ and that $(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$ is admissible. If $\mathbf{c}_{r-1} = \mathbf{0}$, then $(\mathbf{c}_1, \dots, \mathbf{c}_r)$ is admissible if and only if $\mathbf{c}_r = \mathbf{0}$; in this case $T(\mathbf{c}_1, \dots, \mathbf{c}_r) = T(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$. If $\mathbf{c}_{r-1} \neq \mathbf{0}$, then $(\mathbf{c}_1, \dots, \mathbf{c}_r)$ is admissible if $\mathbf{c}_r = \mathbf{0}$ or if there is a terminal region $V \in \hat{T}(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$ such that $\mathbf{c}_r \in \mathcal{C}_2(V, P)$ is a centroid pair for V with respect to P . In the former case $T(\mathbf{c}_1, \dots, \mathbf{c}_r) = T(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$, while in the latter $T(\mathbf{c}_1, \dots, \mathbf{c}_r)$ is obtained from $T(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$

by splitting the terminal node $v \in \tilde{T}(\mathbf{c}_1, \dots, \mathbf{c}_{r-1})$ associated with V , and labeling its children by the components of \mathbf{c}_r in either order.

Definition [Admissible Sequence]: A singly infinite sequence of vector-pairs $\mathbf{s} \in \prod_{i=1}^{\infty} \mathbb{R}^k \times \mathbb{R}^k$ is admissible if its initial segments $(\mathbf{s}(1), \dots, \mathbf{s}(r))$ are admissible for every $r \geq 1$. Here $\mathbf{s}(r)$ denotes the r th vector pair in \mathbf{s} . Let $\mathcal{A}(P)$ denote the set of all infinite sequences that are admissible with respect to P . For a fixed sequence $\mathbf{s} \in \mathcal{A}(P)$ we make the following definitions:

- $|\mathbf{s}| = \max \{r : \mathbf{s}(r) \neq \mathbf{0}\}$ is called the *length* of \mathbf{s} .
- \mathbf{s} is *terminating* if $|\mathbf{s}| < \infty$.
- \mathbf{s} is *nonterminating* if $|\mathbf{s}| = \infty$.
- Let $T_0(\mathbf{s})$ consist of a single root node labeled by any vector in $C_1(\mathbb{R}^k, P)$.
- For $r \geq 1$ let $T_r(\mathbf{s})$ be the labeled binary tree corresponding to $(\mathbf{s}(1), \dots, \mathbf{s}(r))$.
- If \mathbf{s} is terminating, define $T(\mathbf{s}) = T_{|\mathbf{s}|}(\mathbf{s})$.
- $\tilde{T}_r(\mathbf{s})$ denotes the terminal regions/nodes of $T_r(\mathbf{s})$.
- $\tilde{T}(\mathbf{s})$ denotes the terminal regions/nodes of $T(\mathbf{s})$.

If \mathbf{s} is nonterminating, then $\mathbf{s}(r) \neq \mathbf{0}$ for every r . If \mathbf{s} is terminating, then $T(\mathbf{s})$ is the limit of $\{T_r(\mathbf{s})\}_{r=1}^{\infty}$. Convergence of admissible sequences is defined on a component basis: if $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}$ are elements of $\prod_{i=1}^{\infty} \mathbb{R}^k \times \mathbb{R}^k$, then $\mathbf{s}_n \rightarrow \mathbf{s}$ if and only if $\mathbf{s}_n(r) \rightarrow \mathbf{s}(r)$ as vectors in \mathbb{R}^{2k} for every $r \geq 1$.

B. Trajectories

When the greedy growing algorithm is applied to a fixed distribution P on \mathbb{R}^k , it produces a nested sequence $T_0 \leq T_1, \dots$ of labeled binary trees. This sequence describes a *trajectory* of the algorithm's operation. If the algorithm fails to terminate, its trajectory is infinite. This is the case, for instance, when P is absolutely continuous and the algorithm is not subject to a stopping criterion. When the algorithm produces finitely many trees $T_0 < \dots < T_r$, its trajectory can be extended by setting $T_{r+1} = T_{r+2} = \dots = T_r$.

Every trajectory of the greedy growing algorithm can be represented by an admissible sequence of centroid pairs $\mathbf{s} \in \prod_{i=1}^{\infty} \mathbb{R}^k \times \mathbb{R}^k$. Recall that T_0 consists of a single node labeled with a centroid $c \in C_1(\mathbb{R}^k, P)$. Consider any subsequent stage r of the algorithm's operation. If $T_r = T_{r-1}$ then the algorithm has terminated and we set $\mathbf{s}(r) = \mathbf{0}$. Otherwise, T_r was produced from T_{r-1} by splitting any terminal region $V \in \tilde{T}_{r-1}$ that maximized the ratio $\Delta D^*/\Delta R$. Let $\mathbf{s}(r) \in C_2(V, P)$ be the centroid pair labeling the children of V .

By performing this translation for each $r \geq 1$, one may represent every trajectory $\{T_i\}$ of the greedy growing algorithm by an infinite sequence \mathbf{s} that is unique up to a reordering of its component pairs. Moreover, it is clear that $\{T_i\}$ can be completely recovered from \mathbf{s} , so that the two descriptions are interchangeable. Thus we may, without ambiguity, refer to the sequence \mathbf{s} as a trajectory of the greedy growing algorithm. Trajectories of a different sort, based on iterations of a continuous map, were employed by Sabin and Gray [29] in their analysis of the empirical behavior of the Lloyd algorithm.

The definitions and results of the previous subsection carry over to trajectories with the help of the following elementary proposition.

Proposition 1: A sequence \mathbf{s} is a trajectory if and only if i) \mathbf{s} is admissible, and ii) for every $r \geq 1$ if $\mathbf{s}(r) \neq \mathbf{0}$, then $T_r(\mathbf{s})$ is produced by splitting a node $v \in \tilde{T}_{r-1}(\mathbf{s})$ for which

$$\frac{\Delta D^*(v)}{\Delta R(v)} = \max \frac{\Delta D^*(u)}{\Delta R(u)}$$

where the maximum is over $u \in \tilde{T}_{r-1}(\mathbf{s})$. \square

Definition [Trajectories]: Fix a distribution P on \mathbb{R}^k . For every $B < \infty$ let $\mathcal{T}(P, B)$ consist of all the trajectory sequences produced when the greedy growing algorithm is applied to P with any rate-based stopping criterion $B' \leq B$. Elements of $\mathcal{T}(P, B)$ will be referred to as rate-constrained trajectories.

Remark: The collection $\mathcal{T}(P, B)$ is analogous to the ensemble of trees $\mathcal{S}(P, B)$ defined in the previous section. However, there are important differences: a trajectory provides a detailed description of the tree that explicitly encodes the way in which the tree was produced by the algorithm. Every trajectory in $\mathcal{T}(P, B)$ corresponds to a tree in $\mathcal{S}(P, B)$, though a given tree may be the product of numerous trajectories. All subsequent analysis will address individual trajectories and collections $\mathcal{T}(P, B)$ defined as above.

C. Isomorphism and Convergence of Sets

Two trees produced by greedy growing are structurally isomorphic if they were produced by the algorithm in the same manner, and they are isomorphic as directed graphs.

Definition [Tree Isomorphism]: Let $T = T(\mathbf{c}_1, \dots, \mathbf{c}_r)$ and $T' = T(\mathbf{c}'_1, \dots, \mathbf{c}'_r)$ be described by finite admissible sequences. Then T and T' are *structurally isomorphic*, written $T \cong T'$, if there is a bijection $\theta : T \rightarrow T'$ such that

- If v_1 is a descendent of v_2 , then $\theta(v_1)$ is a descendent of $\theta(v_2)$.
- If siblings v_1 and v_2 are labeled by the first and second components of \mathbf{c}_j , then $\theta(v_1)$ and $\theta(v_2)$ are labeled by the first and second components of \mathbf{c}'_j .

Note that $(\mathbf{c}_1, \dots, \mathbf{c}_r)$ and $(\mathbf{c}'_1, \dots, \mathbf{c}'_r)$ may be admissible with respect to different distributions: the isomorphism relation does not capture the probabilistic structure of labeled trees.

Definition [Convergence of Sets]: Let A_1, A_2, A_3, \dots be subsets of \mathbb{R}^k . The *liminf* of the sequence $\{A_n\}$ is defined by

$$\underline{\lim} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} A_m.$$

Similarly, the *limsup* of $\{A_n\}$ is defined by

$$\overline{\lim} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m.$$

Recall that $\underline{\lim} A_n \subseteq \overline{\lim} A_n$. The sequence $\{A_n\}$ is said to *converge* to A , written $A_n \rightarrow A$, if

$$A^\circ \subseteq \underline{\lim} A_n \subseteq \overline{\lim} A_n \subseteq \bar{A}.$$

Here A° denotes the interior of A , and \bar{A} denotes the closure of A . Note that the limit of a given sequence may not be unique.

The next four lemmas are used to establish the closure properties of the next section. They relate set convergence, tree isomorphism, and centroids. Proofs can be found in Appendices II and IV.

Lemma 5: Let $\{\mathbf{c}_{1n}, \dots, \mathbf{c}_{mn}\}_{n=1}^\infty$ be admissible sequences converging to an admissible sequence $(\mathbf{c}_1, \dots, \mathbf{c}_m)$. Let $T_n = T(\mathbf{c}_{1n}, \dots, \mathbf{c}_{mn})$ and $T = T(\mathbf{c}_1, \dots, \mathbf{c}_m)$, and suppose that for each n there is an isomorphism $\theta_n : T \rightarrow T_n$. Fix a node $v \in T$ with region V , and let V_n be the region associated with $\theta_n(v) \in T_n$. Then $V_n \rightarrow V$. \square

Remark: Note that each sequence $(\mathbf{c}_{1n}, \dots, \mathbf{c}_{mn})$ appearing in the statement of Lemma 5 may be admissible with respect to a different distribution.

Let P be an absolutely continuous distribution with compact support. The following lemma is a consequence of the fact that if V is convex with $P(V) > 0$ then $C_1(V, P) \subseteq V^\circ$ (see Appendix I). Recall that $S(x, \delta) = \{y : |x - y| \leq \delta\}$.

Lemma 6: Let V_1, V_2, \dots be convex sets converging to V with $P(V) > 0$. If $(a, b) \in C_2(V, P)$ then there exist a number $\delta > 0$ and an integer N such that $S(a, \delta), S(b, \delta) \subseteq V_n^\circ$ for every $n \geq N$. \square

Let P_1, P_2, \dots converge to P . The next two lemmas establish the stability of centroids and continuity of the incremental performance measures $\Delta D^*(\cdot)$ and $\Delta R^*(\cdot)$.

Lemma 7: Let A_1, A_2, \dots be sets converging to A with $P(A) > 0$. If $\mathbf{c}_n \in C_2(A_n, P_n)$ for $n \geq 1$ and $\mathbf{c}_n \rightarrow \mathbf{c}$, then $\mathbf{c} \in C_2(A, P)$. \square

Lemma 8: Let V_1, V_2, \dots be polytopes having at most L faces. If $V_n \rightarrow V$ with $P(V) > 0$, then $\Delta D^*(V_n, P_n) \rightarrow \Delta D^*(V)$ and $\Delta R^*(V_n, P_n) \rightarrow \Delta R^*(V)$. \square

D. Closure Properties and Compactness

Topological properties of trajectories form the basic analytical tools used to analyze the asymptotic behavior of the greedy growing algorithm. The following lemma shows that a limit of P_n -admissible sequences is admissible with respect to P .

Lemma 9: Let P_1, P_2, \dots converge to an absolutely continuous distribution P with compact support. If $\mathbf{s}_n \in \mathcal{A}(P_n)$ for $n \geq 1$ and $\mathbf{s}_n \rightarrow \mathbf{s}$, then the following statements are valid.

1) For every $r \geq 1$ the sequence $\mathbf{s}(1), \dots, \mathbf{s}(r)$ is admissible, with a corresponding tree $T_r(\mathbf{s})$. In particular

- i) If $\mathbf{s}(r) = \mathbf{0}$, then $\mathbf{s}(r+1) = \mathbf{s}(r+2) = \dots = \mathbf{0}$;
- ii) If $\mathbf{s}(r) \neq \mathbf{0}$, then there is a region $V \in \tilde{T}_{r-1}(\mathbf{s})$ such that $\mathbf{s}(r) \in C_2(V, P)$.

2) There is an integer N_r such that $T_r(\mathbf{s}_n) \cong T_r(\mathbf{s})$ whenever $n \geq N_r$.

It follows from 1) that the infinite sequence \mathbf{s} is an element of $\mathcal{A}(P)$.

Proof: The result is established by induction on r . The proof is broken into four steps. Recall that $T_0(\mathbf{s})$ consists of a single node with an associated region \mathbb{R}^k .

Step 1 [$r = 1$ and $\mathbf{s}(1) = \mathbf{0}$]: If $\mathbf{s}_n(1)$ is nonzero infinitely often then there are integers $n_1 < n_2 < \dots$ such that $\mathbf{s}_{n_k}(1) \in C_2(\mathbb{R}^k, P_{n_k})$ for each k . By assumption, $\mathbf{s}_{n_k}(1) \rightarrow \mathbf{s}(1)$, so

Lemma 7 implies that $\mathbf{s}(1) = \mathbf{0} \in C_2(\mathbb{R}^k, P)$. However, $\mathbf{0}$ is not an admissible centroid pair, and consequently there must be an integer N_1 such that $\mathbf{s}_n(1) = \mathbf{0}$ for every $n \geq N_1$. As each sequence \mathbf{s}_n is admissible,

$$\mathbf{s}_n(1) = \mathbf{s}_n(2) = \dots = \mathbf{0} \text{ for every } n \geq N_1 \quad (6)$$

and therefore $\mathbf{s}(1) = \mathbf{s}(2) = \dots = \mathbf{0}$. Thus 1) is satisfied. In addition, (6) shows that when $n \geq N_1$ both $T_1(\mathbf{s})$ and $T_1(\mathbf{s}_n)$ are isomorphic to a tree consisting of a single root node. This establishes 2).

Step 2 [$r = 1$ and $\mathbf{s}(1) \neq \mathbf{0}$]: There is an integer N_1 such that $\mathbf{s}_n(1) \neq \mathbf{0}$ whenever $n \geq N_1$. As above, Lemma 7 ensures that $\mathbf{s}(1) \in C_2(\mathbb{R}^k, P)$. For $n \geq N_1$ the trees $T_1(\mathbf{s})$ and $T_1(\mathbf{s}_n)$ consist of a root node with two children. By accounting for the order in which the terminal nodes of $T_1(\mathbf{s}_n)$ are labeled, it is easy to define an isomorphism $\theta_n : T_1(\mathbf{s}) \rightarrow T_1(\mathbf{s}_n)$ for each $n \geq N_1$.

Step 3 [$r > 1$ and $\mathbf{s}(r) = \mathbf{0}$]: Assume that i) and ii) hold for the sequence $(\mathbf{s}(1), \dots, \mathbf{s}(r-1))$. It is easy to see that 1) and 2) will hold for $(\mathbf{s}(1), \dots, \mathbf{s}(r))$ if there is an integer $N_r \geq N_{r-1}$ with the property that $\mathbf{s}_n(r) = \mathbf{0}$ whenever $n \geq N_r$. If $\mathbf{s}_n(r)$ is nonzero, then by definition $\mathbf{s}_n(r)$ splits a terminal node $u_n \in \tilde{T}_{r-1}(\mathbf{s}_n)$ with an associated region U_n , so that $\mathbf{s}_n(r) \in C_2(U_n, P_n)$. If $n \geq N_{r-1}$ then the induction hypothesis ensures that there is an isomorphism $\theta_n : T_{r-1}(\mathbf{s}_n) \rightarrow T_{r-1}(\mathbf{s})$, and it is easy to see that $\theta_n(u_n)$ is a terminal node of $T_{r-1}(\mathbf{s})$. Therefore, if $\mathbf{s}_n(r)$ is nonzero for infinitely many n , there must exist a fixed terminal node $v \in \tilde{T}_{r-1}(\mathbf{s})$ and a sequence $\{n_k\}$ such that $\theta_{n_k}(u_{n_k}) = v$ for every k . As $\mathbf{s}_{n_k} \rightarrow \mathbf{s}$, Lemma 5 implies that U_{n_k} converges to the region V associated with v . The inductive hypothesis then insures that $P(V) > 0$, and consequently $\mathbf{s}(r) = \mathbf{0} \in C_2(V, P)$ by Lemma 7. However $\mathbf{0}$ is not an admissible centroid pair, and consequently $\mathbf{s}_n(r)$ can be nonzero for only finitely many n . Equivalently, $\mathbf{s}_n(r) = \mathbf{0}$ for every n greater than some integer $N_r \geq N_{r-1}$.

Step 4 [$r > 1$ and $\mathbf{s}(r) \neq \mathbf{0}$]: There is an integer $N \geq N_{r-1}$ such that $\mathbf{s}_n(r)$ is nonzero for every $n \geq N$. The argument of the previous paragraph shows that $\mathbf{s}(r) \in C_2(V, P)$, where V is associated with a terminal node $v \in \tilde{T}_{r-1}(\mathbf{s})$. This establishes (a). For $n \geq N_{r-1}$ let $V_n \in \tilde{T}_{r-1}(\mathbf{s}_n)$ be the region associated with $\theta_n^{-1}(v)$. Then $V_n \rightarrow V$ by Lemma 5. If $\mathbf{s}_n(r) = (a_n, b_n)$ and $\mathbf{s}(r) = (a, b)$, then $(a_n, b_n) \rightarrow (a, b)$; Lemma 6 shows that $a_n, b_n \in V_n^\circ$ for n larger than some integer $N_r \geq N_{r-1}$. Therefore, $\mathbf{s}_n(r)$ splits the region V_n (equivalently the node $\theta_n^{-1}(v)$) whenever $n \geq N_r$. For each such n define a mapping $\theta_n^* : T_r(\mathbf{s}_n) \rightarrow T_r(\mathbf{s})$ as follows. If $u \in T_r(\mathbf{s}_n) \cap T_{r-1}(\mathbf{s}_n)$, set $\theta_n^*(u) = \theta_n(u)$. Otherwise, u is a child of $\theta_n^{-1}(v)$: if u is labeled by a_n , let $\theta_n^*(u)$ be the child of v labeled by a ; if u is labeled by b_n , let $\theta_n^*(u)$ be the child of v labeled by b . It is easy to see that θ_n^* is an isomorphism. This establishes 2) and completes the proof of the lemma. \square

The next result, analogous to Lemma 9, states that a limit of P_n -trajectories is a trajectory with respect to P . This closure property will be used frequently in the next section.

Lemma 10: If $\mathbf{s}_n \in T(P_n, B)$ for each $n \geq 1$ and $\mathbf{s}_n \rightarrow \mathbf{s}$, then $\mathbf{s} \in T(P, B)$.

Proof: As $\mathcal{T}(P_n, B) \subseteq \mathcal{A}(P_n)$ for each n , Lemma 9 shows that $\mathbf{s} \in \mathcal{A}(P)$. Now fix $r \geq 1$ and suppose that $\mathbf{s}(r) \neq \mathbf{0}$. The terminal nodes v_1, \dots, v_m of $T_{r-1}(\mathbf{s})$ have regions V_1, \dots, V_m . For every $n \geq N_{r-1}$ there is an isomorphism $\theta_n : T_{r-1}(\mathbf{s}) \rightarrow T_{r-1}(\mathbf{s}_n)$. Let $V_{jn} \in \tilde{T}_{r-1}(\mathbf{s}_n)$ be the region associated with $\theta_n(v_j)$. For each j , $V_{jn} \rightarrow V_j$ by Proposition 5. Moreover, if $l \in [1, m]$ is such that $\mathbf{s}(r) \in \mathcal{C}_2(V_l, P)$, then $\mathbf{s}_n(r) \in \mathcal{C}_2(V_{nl}, P_n)$ for every $n \geq N_r$, and consequently

$$\frac{\Delta D^*(V_{nl}, P_n)}{\Delta R(V_{nl}, P_n)} = \max_{1 \leq j \leq m} \frac{\Delta D^*(V_{nj}, P_n)}{\Delta R(V_{nj}, P_n)}$$

when n is sufficiently large. Letting n tend to infinity and applying Lemma 8 one finds that

$$\frac{\Delta D^*(V_l)}{\Delta R(V_l)} = \max_{1 \leq j \leq m} \frac{\Delta D^*(V_j)}{\Delta R(V_j)}.$$

It follows from Proposition 1 that \mathbf{s} is a trajectory of the greedy growing algorithm.

It remains to show that \mathbf{s} is a terminating sequence whose corresponding tree has expected depth at most B . It is easy to see that for each integer r

$$R(T_r(\mathbf{s})) = \lim_{n \rightarrow \infty} R(T_r(\mathbf{s}_n), P_n) \leq B$$

where the inequality follows because $\mathbf{s}_n \in \mathcal{T}(P_n, B)$ for every n . Theorem 1 shows that \mathbf{s} is terminating, so that $T(\mathbf{s}) = T_{|\mathbf{s}|}(\mathbf{s})$ is well-defined. It then follows from the inequality above that $R(T(\mathbf{s})) \leq B$. \square

If P has bounded support then the closed set $S_P \subseteq \mathbb{R}^k$ is compact. Let $\Gamma_P = \prod_{i=1}^{\infty} S_P \times S_P$. Then $\mathcal{A}(P) \subseteq \Gamma_P$, and if $P_n \rightarrow P$ then $\mathcal{A}(P_n) \subseteq \Gamma_P$ for each n . A straightforward diagonalization argument (or Tychonoff's theorem) shows that Γ_P is compact when it is endowed with the usual product topology. Convergence in the product topology is equivalent to the component-wise convergence defined in the previous section. Thus setting $P_n = P$ for each n , it follows from Lemma 10 that $\mathcal{T}(P, B)$ is a closed subset of Γ_P . As a closed subset of a compact set is compact, the following proposition is immediate.

Proposition 2: For every $B < \infty$ the set $\mathcal{T}(P, B)$ of rate-constrained trajectories is compact. \square

VIII. PROOFS OF PRINCIPAL RESULTS

A. Uniform Termination

Proof of Theorem 2: Fix a trajectory $\mathbf{s} \in \mathcal{T}(P, B)$. Corollary 1 of Theorem 1 shows that \mathbf{s} is terminating so that $T(\mathbf{s})$ is well-defined. Let

$$h(\mathbf{s}) = \min_{V \in \tilde{T}(\mathbf{s})} P(V)$$

be the smallest probability of any terminal region in $\tilde{T}(\mathbf{s})$. Then $h(\mathbf{s}) > 0$ as $P(V) > 0$ for each $V \in \tilde{T}(\mathbf{s})$. Define $\gamma = \inf_{\mathbf{s} \in \mathcal{T}(P, B)} h(\mathbf{s})$.

Consider the set $\mathcal{T}(P, B)$ with the topology it inherits as a subspace of Γ_P . If trajectories $\mathbf{s}_1, \mathbf{s}_2, \dots \in \mathcal{T}(P, B)$ converge to a trajectory $\mathbf{s} \in \mathcal{T}(P, B)$, it can be seen from Lemma 9

and the proof of Lemma 8 that $h(\mathbf{s}_n) \rightarrow h(\mathbf{s})$. Thus h is continuous. The compactness of $\mathcal{T}(P, B)$ then ensures that $\gamma = h(\mathbf{s}^*)$ for some $\mathbf{s}^* \in \mathcal{T}(P, B)$ and it follows that $\gamma > 0$. If $|\tilde{T}(\mathbf{s})|$ denotes the number of terminal nodes of $T(\mathbf{s})$, then

$$1 = \sum_{V \in \tilde{T}(\mathbf{s})} P(V) \geq |\tilde{T}(\mathbf{s})| h(\mathbf{s}) \geq |\tilde{T}(\mathbf{s})| \gamma$$

so that $|T(\mathbf{s})| < 2|\tilde{T}(\mathbf{s})| \leq 2/\gamma < \infty$ for every $\mathbf{s} \in \mathcal{T}(P, B)$. This yields the desired bound. \square

Corollary 2: There are constants $L < \infty$ and $\gamma > 0$ such that every terminal region of every tree $T \in \mathcal{T}(P, B)$ is a polytope with probability greater than γ , having at most L faces. If \mathcal{U} is the collection of all such regions, \mathcal{U} has finite Vapnik–Chervonenkis dimension [31]. \square

B. Structural Consistency

The following result, stated in terms of trajectory sequences, strengthens Theorem 3 by replacing graph-theoretic isomorphism with the stronger notion of structural isomorphism introduced above.

Theorem 5: Let P_1, P_2, \dots converge to an absolutely continuous distribution P with compact support. For every $\epsilon, \delta > 0$ and every $B < \infty$ there is a finite integer N having the following property: for every trajectory

$$\mathbf{s} \in \bigcup_{n=N}^{\infty} \mathcal{T}(P_n, B)$$

there is a trajectory

$$\mathbf{s}' \in \mathcal{T}(P, B)$$

that is close to \mathbf{s} in the sense that

- i) $T(\mathbf{s})$ and $T(\mathbf{s}')$ are structurally isomorphic.
- ii) $\|\mathbf{s}(i) - \mathbf{s}'(i)\| \leq \epsilon$ for each $i \geq 1$.
- iii) $P\{x : \|Q_{T(\mathbf{s})}(x) - Q_{T(\mathbf{s}')} (x)\| > \epsilon\} < \delta$.

Proof of Theorem 5: Suppose that the result fails to hold. Then there exist numbers $0 < \epsilon, \delta, B < \infty$, and an infinite sequence of trajectories

$$\{\mathbf{s}_{n_k} \in \mathcal{T}(P_{n_k}, B)\} \subseteq \Gamma_P \quad (7)$$

such that for each k no trajectory $\mathbf{s} \in \mathcal{T}(P, B)$ is close to \mathbf{s}_{n_k} in the sense of i), ii), and iii). It suffices then to show that every sequence of the form (7) contains a further subsequence along which each of i), ii), and iii) hold.

By Lemma 10 and the compactness of Γ_P there is a subsequence $\{\mathbf{s}_{m_k}\}$ of $\{\mathbf{s}_{n_k}\}$ that converges to a fixed trajectory $\mathbf{s}^* \in \mathcal{T}(P, B)$. Theorem 1 insures that \mathbf{s}^* is terminating: let $T' = T(\mathbf{s}^*)$. When k is large Lemma 9 shows that $T(\mathbf{s}_{m_k}) \cong T'$ and that the corresponding labels of $T(\mathbf{s}_{m_k})$ and T' are at most ϵ apart. Thus i) and ii) hold along $\{\mathbf{s}_{m_k}\}$.

Consider a fixed terminal region V of T' and let V_{m_k} be the corresponding region of $T(\mathbf{s}_{m_k})$. Then $V_{m_k} \rightarrow V$ by Lemma 5, and as $\{\mathbf{s}_{m_k}\} \rightarrow \mathbf{s}^*$ the centroid of V_{m_k} converges to that of V as well. It follows that iii) holds when k is sufficiently large. \square

APPENDIX I
INTERIOR LEMMA

Lemma 11: Let P be a distribution on \mathbb{R}^k that is absolutely continuous with respect to Lebesgue measure. If V is a bounded convex set with $P(V) > 0$, then $\mathcal{C}_1(V, P) \subseteq V^\circ$.

Remark: Note that the result is immediate in the case of squared error distortion $\rho(u, v) = \|u - v\|^2$. Condition $\rho 2$ of Section II-C is not the most general under which the conclusion is true. Conditions $\rho 1$ and $\rho 2$ cover every distortion function known by us to have been implemented in the application of TSVQ to compressing medical images.

Definition: Let A be a subset of \mathbb{R}^k . The *boundary* of A is defined to be $\partial A = \bar{A} \setminus A^\circ$.

Proof: For each $c \in \mathbb{R}^k$ define

$$H(c) = \int_V \phi(\|c - x\|) f(x) dx.$$

Let $c^* = \operatorname{argmin} H(c)$. The special case in which $k = 1$ turns out to be the general one, and without loss of generality we can take V to be the unit interval. It is enough to show that $H'(0) < 0$, for then $c^* \neq 0$, and by a change of variable $c^* \neq 1$ so that $c^* \in V^\circ$. For $0 < h < 1$

$$\begin{aligned} \frac{H(h) - H(0)}{h} &= \int_0^1 \frac{\phi(|h-x|) - \phi(|x|)}{h} f(x) dx \\ &= \int_0^1 \frac{\phi(|h-x|) - \phi(|x|)}{|h-x| - |x|} \\ &\quad \cdot \frac{|h-x| - |x|}{h} f(x) dx \\ &= o(1) + \int_{2h}^1 \frac{\phi(|h-x|) - \phi(|x|)}{|h-x| - |x|} (-1) f(x) dx. \end{aligned}$$

By the dominated convergence theorem and $\rho 2$, this last term tends to

$$- \int_0^1 \phi'(x) f(x) dx < 0$$

as h tends to 0.

Suppose now that $k > 1$, and let c^* be a candidate point in ∂V where the minimum of $H(c)$ is attained. For $\alpha \geq 0$ define $\psi(\alpha) = \phi(\alpha^{1/2})$, so that

$$H(c) = \int_V \psi(\|c - x\|) f(x) dx.$$

We coordinatize \mathbb{R}^k in a particular way, namely, with the origin at c^* , and the positive first coordinate in the direction of a normal w whose initial segment is interior to V . This is always possible. The remaining coordinates are chosen to be orthogonal to w . For $x \in V$ and any c in the one-dimensional subspace $W = \{\alpha w\}$ of \mathbb{R}^k spanned by w , write $x - c = x - x_w + (x_w - c)$, where x_w is the projection of x onto W . The two summands are orthogonal, so

$$\psi(\|c - x\|) = \psi(\|x - x_w\|^2 + \|x_w - c\|^2).$$

As c varies in W , the first summand in the argument of ψ is constant, and the second is a univariate squared distance. Write

$f(x) = f(y|x_w)f(x_w)$, where $y = x - x_w$, and express H as an iterated integral

$$H(c^*) = \int \left[\int \psi(\|x - x_w\|^2 + \|x_w - c^*\|^2) f(y|x_w) dy \right] f(x_w) dx_w.$$

Write

$$\begin{aligned} \psi(\|x - x_w\|^2 + \|x_w - c^*\|^2) &= \psi(\cdot) \\ &= [\psi(\cdot) - \psi(\|x - x_w\|^2)] \\ &\quad + \psi(\|x - x_w\|^2) \end{aligned}$$

and rewrite $H(c^*)$ as the corresponding sum of two integrals. Now compute the directional derivative interior to V along W . The derivative of the second term is 0, and that of the first term is negative by the previous result for $k = 1$. This contradicts the assumption that $c^* \in \mathcal{C}_1(V, P)$ and completes the proof. \square

APPENDIX II
CONVERGENCE OF SETS

Proposition 3: Let $A_n \rightarrow A$ and $B_n \rightarrow B$. If both A and B are closed, then $A_n \cap B_n \rightarrow A \cap B$.

Proof: Recall that $A_n \rightarrow A$ if

$$A^\circ \subseteq \liminf A_n \subseteq \overline{\lim} A_n \subseteq \bar{A}.$$

Using the basic properties of the liminf and the limsup

$$\begin{aligned} (A \cap B)^\circ &= A^\circ \cap B^\circ \subseteq (\liminf A_n) \cap (\liminf B_n) \\ &= \liminf (A_n \cap B_n) \subseteq \overline{\lim} (A_n \cap B_n) \subseteq (\overline{\lim} A_n) \\ &\quad \cap (\overline{\lim} B_n) \subseteq \bar{A} \cap \bar{B} \\ &= \overline{A \cap B}. \end{aligned}$$

where the last inequality follows since A and B are closed. \square

Proof of Lemma 5: The theorem is proved by induction on the depth of v in T . If v is the root node of T , then $\theta_n(v)$ is the root node of T_n . Therefore, $V_n = V = \mathbb{R}^k$ for each n , and $V_n \rightarrow V$. If $\operatorname{depth}(v) > 0$, then v has a parent u , and $\theta_n(u)$ is the parent of $\theta_n(v)$. If regions U and U_n are associated with u and $\theta_n(u)$, respectively, then $U_n \rightarrow U$ by the induction hypothesis.

Suppose that $\mathbf{c}_j = (a, b)$ labels the children of u , and that v is labeled by a . Then $\mathbf{c}_{jn} = (a_n, b_n)$ labels the children of $\theta_n(u)$, and $\theta_n(v)$ is labeled by a_n . As a consequence

$$V = U \cap H_{ab} \quad \text{and} \quad V_n = U_n \cap H_{a_n b_n}$$

where

$$H_{uv} = \{x : \|x - u\| \leq \|x - v\|\}.$$

As $(a_n, b_n) \rightarrow (a, b)$ by assumption, it is easy to see that $H_{a_n b_n} \rightarrow H_{ab}$. Since U and H_{ab} are closed, $V_n \rightarrow V$ by Proposition 3. \square

Proposition 4: Let A_1, A_2, \dots be convex sets converging to A . If $x \in A^\circ$, then there is a number $\delta > 0$ and an integer n_0 such that $S(x, \delta) \subseteq A_n^\circ$ for every $n \geq n_0$. Here

$$S(x, \delta) = \{y : \|x - y\| \leq \delta\}.$$

Proof: If $x \in A^\circ$ there is a number $\epsilon > 0$ such that $S(x, \epsilon) \subseteq A^\circ$. Within $S(x, \epsilon)$ there are $2k$ points x_1, \dots, x_{2k} whose convex hull D has a nonempty interior that contains x . Choose $\delta > 0$ so that $S(x, \delta) \subseteq D^\circ$.

For $n \geq 1$ let $C_n = \bigcap_{m \geq n} A_m$. As $C_n \nearrow \lim A_n \supseteq A^\circ$, there is an integer n_0 such that $x_1, \dots, x_{2k} \in C_n$ whenever $n \geq n_0$. As each set C_n is convex, $S(x, \delta) \subseteq D \subseteq C_n$ for every $n \geq n_0$, and the result follows. \square

Proof of Lemma 6: The result follows from Lemma 11 and Proposition 4.

Definition: Let A and B be subsets of \mathbb{R}^k . The *symmetric difference* of A and B is $A \Delta B = (A \cap B^c) \cup (B \cap A^c)$.

Proposition 5: Let A_1, A_2, \dots be convex sets converging to A , and let P be a probability distribution. If $P(\partial A) = 0$ then $P(A_n \Delta A) \rightarrow 0$.

Proof: Define $D_n = (\bigcup_{m \geq n} A_m)$. Note that $D_n \searrow \lim A_n \subseteq \bar{A}$ so that $D_n \cap \bar{A}^c \searrow \emptyset$. As $P(\partial A) = 0$, the inequality $P(A_n \cap A^c) = P(A_n \cap \bar{A}^c) \leq P(D_n \cap \bar{A}^c)$ holds for every n . It follows that $P(A_n \cap A^c) \rightarrow 0$ as $n \rightarrow \infty$. By a similar argument, $P(A \cap A_n^c) \rightarrow 0$ as $n \rightarrow \infty$. \square

APPENDIX III

BRACKETING CLASSES AND UNIFORM CONVERGENCE

If $P_n \rightarrow P$ then $\int f dP_n \rightarrow \int f dP$ for every fixed, measurable function f that is bounded on S_P . In our analysis it is frequently the case that both the function and the distribution will change with n . To address this problem we establish the *uniform convergence* of $\int f dP_n$ to $\int f dP$ over an infinite class of functions that will contain the ‘‘moving target’’ f_n for sufficiently large values of n .

Definition: Let P be a probability distribution on \mathbb{R}^k . A class \mathcal{F} of measurable functions $f : \mathbb{R}^k \rightarrow \mathbb{R}$ is said to satisfy a *bracketing condition* with respect to a distribution P if i) there exists a number $K < \infty$ such that $|f(x)| \leq K$ for every $f \in \mathcal{F}$ and every $x \in S_P$, and ii) for every $\epsilon > 0$ there is a finite set of functions $\mathcal{G}_\epsilon = \{g_1, \dots, g_r\}$ such that every $f \in \mathcal{F}$ has approximating functions $\underline{g}, \bar{g} \in \mathcal{G}_\epsilon$ satisfying

- $\underline{g} \leq f \leq \bar{g}$.
- $\int (\bar{g} - \underline{g}) dP \leq \epsilon$.

The bracketing class \mathcal{G}_ϵ need not be contained in \mathcal{F} . Note that condition i) is frequently replaced by integrability with respect to P .

Lemma 12: Let \mathcal{H} be the collection of indicator functions of closed half-spaces in \mathbb{R}^k . If P has a density, then \mathcal{H} is bracketing with respect to P .

Proof: Fix $\epsilon > 0$ and let S be a closed sphere of diameter M , centered at the origin, for which $P(S^c) \leq \epsilon/2$. As P has a density, there is a number $\gamma > 0$ such that $P(A) \leq \epsilon/2$ whenever A has Lebesgue measure $\lambda(A) \leq M^{k-1}\gamma$. Partition S into a finite number of sets $\{B_1, \dots, B_L\}$ such that the diameter of each cell B_i is less than $\gamma/2$. For each $H \in \mathcal{H}$ let $\partial^* H$ be the union of those cells B_i that intersect ∂H .

By an elementary argument, $\lambda(\partial^* H \cap S) < M^{k-1}\gamma$, so that $P(\partial^* H \cap S) < \epsilon/2$.

Let \mathcal{G}_ϵ contain the indicator functions of those sets that can be obtained by taking unions of sets among S^c, B_1, \dots, B_n . An easy argument shows that \mathcal{G}_ϵ is a bracketing class for \mathcal{H} . \square

Lemma 13: Let \mathcal{F}_1 and \mathcal{F}_2 be bracketing with respect to P . The product

$$\mathcal{F} = \mathcal{F}_1 \cdot \mathcal{F}_2 = \{f_1 \cdot f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$$

is bracketing with respect to P . \square

Definition: For $L \geq 1$ let \mathcal{P}_L be the collection of all closed polytopes $V \subseteq \mathbb{R}^k$ having at most L faces. Define $\mathcal{H}_L = \{I_V : V \in \mathcal{P}_L\}$ to be the corresponding class of indicator functions.

Lemma 14: If P has a density then \mathcal{H}_L is bracketing with respect to P for each $L \geq 1$.

Proof: Fix L of all polytopes $V \in \mathcal{P}_L$. It is easy to see that \mathcal{H}_L is just the L -fold product $\mathcal{H} \cdot \dots \cdot \mathcal{H}$. The result follows from Lemma 12 and Lemma 13. \square

The next theorem is an immediate consequence of a well-known result concerning uniform laws of large numbers [8], [23].

Theorem A: Let \mathcal{F} satisfy a bracketing condition with respect to the distribution P . If $P_n \rightarrow P$, then

$$\sup_{f \in \mathcal{F}} \left| \int f dP - \int f dP_n \right| \rightarrow 0$$

as n tends to infinity. \square

APPENDIX IV

CLOSURE OF CENTROIDS

The uniformity results of the previous section are applied here to establish Lemma 8 and Lemma 7.

Definition: Fix a distribution P on \mathbb{R}^k . For every pair $\mathbf{c} = (a, b) \in \mathbb{R}^k \times \mathbb{R}^k$ and every subset $V \subseteq \mathbb{R}^k$, let

$$\psi(\mathbf{c}, V, P) = \int_V \rho(x, a) \wedge \rho(x, b) dP(x)$$

be the least distortion achievable in assigning vectors $x \in V$ to one of the representatives a or b . Note that

$$\mathcal{C}_2(V, P) = \left\{ \mathbf{c}^* : \psi(\mathbf{c}^*, V, P) = \inf_{\mathbf{c} \in \mathbb{R}^k} \psi(\mathbf{c}, V, P) \right\}.$$

In what follows P_1, P_2, \dots are distributions on \mathbb{R}^k converging to an absolutely continuous distribution P having bounded support.

Lemma 15: If a sequence $A_1, A_2, \dots \in \mathcal{P}_L$ converges to a set A with $P(\partial A) = 0$, then

$$\sup_{\mathbf{c}} |\psi(\mathbf{c}, A_n, P_n) - \psi(\mathbf{c}, A, P)| \rightarrow 0$$

the supremum being over $\mathbf{c} \in S_P \times S_P$.

Proof: The problem may be broken into two parts: for each n and each pair \mathbf{c} .

$$\begin{aligned} & |\psi(\mathbf{c}, A_n, P_n) - \psi(\mathbf{c}, A, P)| \leq \\ & |\psi(\mathbf{c}, A_n, P) - \psi(\mathbf{c}, A, P)| + \\ & |\psi(\mathbf{c}, A_n, P) - \psi(\mathbf{c}, A_n, P_n)|. \end{aligned}$$

As $\rho(\cdot, \cdot)$ is continuous on the compact set $S_P \times S_P$, the constant

$$M = \sup_{x, y \in S_P} \rho(x, y)$$

is finite. For each pair $\mathbf{c} = (a, b) \in S_P \times S_P$

$$\begin{aligned} & |\psi(\mathbf{c}, A_n, P) - \psi(\mathbf{c}, A, P)| \\ & \leq \int \rho(a, x) \wedge \rho(b, x) |I_{A_n}(x) - I_A(x)| dP(x) \leq \\ & MP(A_n \Delta A). \end{aligned}$$

Since this last inequality is independent of \mathbf{c} , Lemma 5 shows that as $n \rightarrow \infty$

$$\sup_{\mathbf{c}} |\psi(\mathbf{c}, A_n, P) - \psi(\mathbf{c}, A, P)| \rightarrow 0. \quad (8)$$

Now define a class $\mathcal{G} = \{\rho(a, \cdot) \wedge \rho(b, \cdot) : a, b \in S_P\}$. The continuity of ρ and the compactness of S_P ensure that \mathcal{G} is bracketing with respect to P [24], [16]. Let

$$\mathcal{F} = \mathcal{G} \cdot \mathcal{P}_L = \{\rho(\cdot, a) \wedge \rho(\cdot, b) I_V(\cdot) : a, b \in S_P, V \in \mathcal{P}_L\}.$$

By Lemmas 13 and 14 the class \mathcal{F} is bracketing with respect to P . Moreover, it is clear that if $f(x) = \rho(x, a) \wedge \rho(x, b) I_V(x) \in \mathcal{F}$ and $\mathbf{c} = (a, b)$ then

$$\psi(\mathbf{c}, V, P_n) = \int f dP_n \quad \text{and} \quad \psi(\mathbf{c}, V, P) = \int f dP.$$

By applying Theorem 1 to \mathcal{F} , it follows that

$$\sup_{\mathbf{c}, V} |\psi(\mathbf{c}, V, P) - \psi(\mathbf{c}, V, P_n)| \rightarrow 0$$

the supremum being over $\mathbf{c} \in S_P \times S_P$ and $V \in \mathcal{P}_L$. Combining this last relation with (8) completes the proof. \square

Proof of Lemma 8: A straightforward application of Lemma 15 shows that $\Delta D^*(V_n, P_n) \rightarrow \Delta D^*(V)$. As for ΔR , note that

$$\begin{aligned} |\Delta R(V_n, P_n) - \Delta R(V, P)| &= |P_n(V_n) - P(V)| \\ &\leq |P_n(V_n) - P(V_n)| \\ &\quad + |P(V_n) - P(V)|. \end{aligned}$$

Combining Lemma 14 and Theorem 1 establishes that

$$\sup_{V \in \mathcal{P}_L} |P_n(V) - P(V)| \rightarrow 0$$

and it follows easily that $|P_n(V_n) - P(V_n)| \rightarrow 0$ as $n \rightarrow \infty$. In addition, the difference $|P(V_n) - P(V)| \rightarrow 0$ by Proposition 5. \square

Proof of Lemma 7: For each $A \subseteq \mathbb{R}^k$ and each distribution H let $\psi^*(A, H)$ be the minimum of $\psi(\mathbf{c}, A, H)$ over $\mathbf{c} \in S_P \times S_P$. For each n

$$\begin{aligned} |\psi(\mathbf{c}, A, P) - \psi^*(A, P)| &\leq |\psi(\mathbf{c}, A, P) - \psi(\mathbf{c}_n, A, P)| \\ &\quad + |\psi(\mathbf{c}_n, A, P) - \psi(\mathbf{c}_n, A_n, P_n)| \\ &\quad + |\psi(\mathbf{c}_n, A_n, P_n) - \psi^*(A_n, P_n)| \\ &\quad + |\psi^*(A_n, P_n) - \psi^*(A, P)|. \end{aligned}$$

As n tends to infinity, the second and fourth terms on the right-hand side tend to zero by Lemma 15. The third term is zero for each n as $\mathbf{c}_n \in C_2(A_n, P_n)$, and the first term tends to zero because $\psi(\cdot, A, P)$ is continuous. Therefore $\psi(\mathbf{c}, A, P) = \psi^*(A, P)$, and consequently $\mathbf{c} \in C_2(A, P)$. \square

ACKNOWLEDGMENT

The authors wish to thank R. Gray and his research group at Stanford University for their encouragement and suggestions. Our understanding of Lemma 11 was helped considerably by interactions with D. Burkholder. We also acknowledge T. Linder and one of the reviewers, whose careful proofreading led to a number of corrections and improvements.

REFERENCES

- [1] M. Balakrishnan, "Variable rate structured vector quantization and applications to multiresolution image coding," Rensselaer Poly. Inst., Troy, NY, 1991.
- [2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," Wadsworth Int., Belmont, CA, 1984.
- [3] J. H. Butler, E. A. Gilpin, L. Gordon, and R. A. Olshen, "Tree-structured survival analysis. II," Div. of Biostatistics, Stanford University, Rep. 133, Sept., 1989.
- [4] P. A. Chou, "Applications of information theory to pattern recognition and the design of decision trees and trellises," Stanford Univ., Inform. Syst. Lab., 1988.
- [5] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 299-315, 1989.
- [6] P. C. Cosman, H. C. Davidson, C. J. Bergin, C. J. Tseng, L. E. Moses, E. A. Riskin, R. A. Olshen, and R. M. Gray, "Thoracic CT images: Effect of lossy image compression on diagnostic accuracy," *Radiology*, vol. 190, pp. 517-524, 1994.
- [7] P. C. Cosman, C. Tseng, R. M. Gray, R. A. Olshen, L. E. Moses, H. C. Davidson, C. J. Bergin, and E. A. Riskin, "Tree-structured vector quantization of CT chest scans: image quality and diagnostic accuracy," *IEEE Trans. Med. Imag.*, vol. 12, pp. 727-739, 1993.
- [8] J. DeHardt, "Generalizations of the Glivenko-Cantelli theorem," *Ann. Math. Stat.*, vol. 42, pp. 2050-2055, 1971.
- [9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.
- [10] L. Gordon and R. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Statist.*, vol. 6, pp. 515-533, 1978.
- [11] ———, "Consistent nonparametric regression from recursive partitioning schemes," *J. Multivariate Anal.*, vol. 10, pp. 611-627, 1980.
- [12] ———, "Almost sure consistent nonparametric regression from recursive partitioning schemes," *J. Multivariate Anal.*, vol. 15, pp. 147-163, 1984.
- [13] M. LeBlanc and J. Crowley, "Survival trees by goodness of split," *J. Amer. Statist. Assoc.*, vol. 88, pp. 457-467, 1993.
- [14] G. Lugosi and A. B. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," Beckman Inst., Univ. of Illinois, Urbana-Champaign, Rep. UIUC-BI-93-01; to appear in *Ann. Stat.*, 1996.
- [15] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, pp. 1551-1558, Nov. 1985.
- [16] A. B. Nobel, "On uniform laws of averages," Ph.D. dissertation, Stanford Univ., Inform. Syst. Lab., 1992.
- [17] ———, "Histogram regression estimation using data-dependent partitions," Beckman Inst., Univ. of Illinois, Urbana-Champaign, Rep. UIUC-BI-94-01, 1993; to appear in *Ann. Stat.*, 1996.

- [18] ———, "Recursive partitioning to reduce distortion," Beckman Inst., Univ. of Illinois, Urbana-Champaign, Rep. UIUC-BI-95-01, 1995.
- [19] ———, "Vector quantization and consistent histogram estimation," 1995, in preparation
- [20] D. Pollard, "Strong consistency of k-means clustering," *Ann. Statist.*, vol. 9, pp. 135–140, 1981.
- [21] ———, "A central limit theorem for k-means clustering," *Ann. Probab.*, vol. 10, pp. 919–926, 1982.
- [22] ———, "Quantization and the method of k -means," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 199–205, 1982.
- [23] ———, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.
- [24] R. R. Rao, "Relations between weak and uniform convergence of measures with applications," *Ann. Math. Stat.*, vol. 33, pp. 659–680, 1962.
- [25] E. A. Riskin, "Variable rate vector quantization of images," Rep., Stanford Univ., Inform. Syst. Lab., 1990.
- [26] E. A. Riskin and R. M. Gray, "A greedy tree growing algorithm for the design of variable rate vector quantizers," *IEEE Trans. Signal Processing*, vol. 39, pp. 2500–2507, 1991.
- [27] ———, "Lookahead in growing tree-structured vector quantizers," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing* (Toronto, Ont., Canada, May, 1991), pp. 2289–2292.
- [28] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
- [29] M. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. 32, 1986.
- [30] C. Tseng, S. M. Perlmuter, P. C. Cosman, K. C. P. Li, C. J. Bergin, R. A. Olshen, and R. M. Gray, "Effect of tree-structured vector quantization on the accuracy of vessel measurements in MR chest scans," 1993, submitted to *IEEE Trans. Med. Imag.*
- [31] V.N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.