Recursive Partitioning to Reduce Distortion

Andrew B. Nobel *

November 1, 1996

Abstract

Adaptive partitioning of a multidimensional feature space plays a fundamental role in the design of data-compression schemes. Most partition-based design methods operate in an iterative fashion, seeking to reduce distortion at each stage of their operation by implementing a linear split of a selected cell. The operation and eventual outcome of such methods is easily described in terms of binary tree-structured vector quantizers.

This paper considers a class of simple growing procedures for tree-structured vector quantizers. Of primary interest is the asymptotic distortion of quantizers produced by the unsupervised implementation of the procedures. It is shown that application of the procedures to a convergent sequence of distributions with a suitable limit yields quantizers whose distortion tends to zero. Analogous results are established for tree-structured vector quantizers produced from stationary ergodic training data.

The analysis is applicable to procedures employing both axis-parallel and oblique splitting, and a variety of distortion measures. The results of the paper apply directly to unsupervised procedures that may be efficiently implemented on a digital computer.

Appears in IEEE Transactions on Information Theory,
vol. 43, no. 4, pp. 1122-1133, 1997

*Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@stat.unc.edu. This work was completed while he was on leave as a Beckman Institute Fellow at the Beckman Institute for Advanced Science and Technology, University of Illinois U-C, and was supported in part by National Science Foundation Grant DMS-9501926.

1 Introduction

Adaptive partitioning of a multidimensional feature space plays a fundamental role in the empirical design of data compression schemes. Most partition-based design schemes operate in an iterative fashion, seeking to reduce distortion at each stage of their operation through selective refinement of a given partition. At each iteration, a single cell of the current partition is divided into two new cells by a halfspace. Both the selection of the cell and the choice of the splitting halfspace are based on the achievable reduction of average distortion with respect to a given distribution.

The operation and eventual output of such recursive partitioning schemes can be described in terms of binary tree-structured vector quantizers (TSVQs). Tree-structured vector quantizers provide a computationally efficient means of compressing multidimensional data arising in a variety of applications, including medical imaging and speech recognition. When candidate splits are selected using the two-means algorithm, recursive partitioning methods are formally equivalent to greedy growing algorithms [17, 6, 30, 31, 2]. Though greedy growing is the basic tool by which tree-structured vector quantizers are produced from finite data sets, there has been little analysis to support the unsupervised use of such algorithms, or to examine their behavior on large training sets.

In this paper a class of simple recursive partitioning procedures for producing tree-structured vector quantizers is studied. Of primary interest is the asymptotic distortion of quantizers produced by the unsupervised implementation of these procedures.

1.1 Overview

A tree-structured partition is described by a binary tree and a function that assigns a vector to each node of the tree. Each leaf of the tree corresponds to a cell of the partition. A tree-structured vector quantizer is defined by assigning a vector representative to each leaf of the tree. The recursive growing procedures studied here produce tree-structured vector quantizers, and their associated partitions, one node at a time, in a stepwise optimal fashion.

The input to the growing procedure consists of a distribution P on $\mathbb{R}^d$, and an integer k indicating the number of iterations to perform. At each iteration the procedure carries out the following steps:

a. It produces a good candidate split of each leaf/cell using a local splitting rule;

b. It selects the leaf/cell having the best candidate split;

c. It implements that split by adding two children to the selected leaf, and assigning a suitable vector to each child.

In steps (a) and (b) the assessment of 'goodness' is based on the reduction in distortion that is achievable by a given split. In practice P is the empirical distribution of a finite training set.

Implementation of the recursive growing procedure depends on the distribution P, the iteration number k, and the local splitting rule, each of which is fixed at the outset of its operation. The procedure is adaptive, in that it does not require any prior knowledge of its input distribution: P need not be drawn from a parametric family, nor is it necessary that P satisfy regularity or smoothness conditions. Like other adaptive methods (e.g. classification and regression trees [3]), this flexibility makes the output of the procedure sensitive to its input: changing a single element of a training sequence may result in a markedly different tree.

The sensitivity of adaptive methods makes their analysis problematic. For the procedures considered here, this is compounded by the fact that what happens at a given iteration of the procedure depends critically on what took place at each previous iteration. Analysis of adaptive procedures is commonly undertaken under some form of supervision. Supervision involves external oversight of the procedure, usually through post-hoc modifications, which insure that it behaves in a desired fashion. Under supervision a procedure will not always act in accordance with its input and applicable optimization criteria: it may be forced to take some alternative action that is dictated by theoretical rather than data-analytic considerations.

The growing procedures considered here are entirely unsupervised. No modifications are made for the sake of the analysis.

1.2 Relation to Previous Work

Tree-structured vector quantizers were introduced by Buzo *et al.* [5] in the context of speech coding. Greedy growing algorithms for the design of TSVQ have been suggested by a number of authors. Makhoul et al. [17] proposed splitting the node making the greatest contribution to the overall distortion. The reduction of distortion criteria considered here was first proposed by Chou [6], and applied to greedy growing of balanced (fixed rate) trees. Riskin and Gray [31], and Balakrishnan [2], considered splitting criterion based on the complexity-normalized reduction of distortion. For further details concerning the design and application of TSVQ, we refer the interested reader to the comprehensive book

of Gersho and Gray [9].

The analysis of design methods for vector quantizers of a fixed dimension has focused primarily on nearest-neighbor quantizers with a fixed number of codewords. Pollard [25, 27] showed that empirically optimal nearest-neighbor quantizers converge to the optimal quantizer Q^* for the underlying distribution of the training vectors when Q^* is unique. Adopting a more analytic approach, Sabin and Gray [32] established the asymptotic consistency of the generalized Lloyd (k-means) algorithm. Nobel and Olshen [23] studied the structural consistency of tree-structured quantizers produced by a greedy algorithm similar to that considered here (see the discussion below). In each of these papers the authors considered stationary, ergodic training vectors. In [26] Pollard established a central limit theorem for the codewords of empirically optimal quantizers designed from independent training vectors.

Recursive partitioning of a multidimensional feature space has a long history in statistics, dating from the work of Morgan and Sonquist [18] (see also Sonquist [33]), Anderson [1], Patrick and Fisher [24], and others. More recently, a number of authors have undertaken a systematic analysis of histogram classification and regression schemes based on data-dependent partitions. Gordon and Olshen [10] and later Brieman et al. [3], found sufficient conditions for the consistency of classification rules based on recursive partitioning of a Euclidean observation space. Gordon and Olshen [11, 12], Breiman et al. [3], and Butler et al. [4] established similar results in the context of regression and survival analysis. In each case, the algorithms to which the cited papers apply require supervision to insure that the diameters of the underlying partitions tend to zero.

LeBlanc and Crowley [13] studied application of an unsupervised tree-structured algorithm to the empirical distributions of data in a survival analysis context. Improving the earlier results of Breiman et al. [3], Lugosi and Nobel [16] and Nobel [21] established weak sufficient conditions for the consistency of unsupervised histogram classification and regression schemes based on data-dependent partitions with non-rectangular cells. Their results apply to trees produced by the recursive growing procedures discussed here.

Nobel and Olshen [23] analyzed a greedy growing algorithm for TSVQ that was proposed in [30, 2]. The algorithm employs a splitting criterion that is equal to the reduction of distortion divided by the increase in bit rate (expected depth of the tree). Termination of the algorithm is rate-based, rather than iteration-based as it is here. The analysis in [23] is concerned primarily with termination of the algorithm, and with the structural consistency of trees produced from a convergent sequence of distributions. The results of this paper show that trees produced by repeated application of the complexity normalized splitting

criterion to a fixed, compactly supported distribution have distortion tending to zero (see [23] for more details).

1.3 Summary

Precise definitions of tree structured partitions and tree structured vector quantizers are given in the next section. Section 3 contains a description of local splitting rules and a precise definition of the recursive growing procedure. Section 4 is devoted to the presentation and discussion of the principle results of the paper. Analysis of the recursive growing procedure begins in Section 5, where several important properties of the distortion-based splitting criterion are established. Sections 6 and 7 examine the asymptotic distortion of quantizers produced from a convergent sequence of distributions. It is shown in Theorem 1 that the empirical distortion of trees produced from a convergent sequence of distributions will tend to zero if the limiting distribution has finite second moment and the size of the trees tends to infinity. Theorem 3 shows that the same is true of distortions measured with respect to the limiting distribution, if that distribution is absolutely continuous. Section 8 considers application of the recursive growing algorithm to the empirical distributions of stationary ergodic training vectors. The proofs of several technical results are given in the appendix.

2 Basic Definitions

2.1 Vector Quantizers

A vector quantizer is a map $Q : \mathbb{R}^d \to \mathcal{C}$ where $\mathcal{C} = \{c_1, \ldots, c_m\} \subseteq \mathbb{R}^d$ is a finite set of representative vectors known as the *codebook* of Q. In statistical terminology, Q is a multivariate clustering scheme, and the vectors in $\mathcal{C}$ are its cluster centers. Every quantizer Q gives rise to a partition of $\mathbb{R}^d$ having cells $A_i = \{x : Q(x) = c_i\}$ for $i = 1, \ldots, m$. The cell containing x is defined by

$$Q[x] = \{x' : Q(x) = Q(x')\}\,.$$

When Q is applied to a random vector $X \in \mathbb{R}^d$ its performance will be judged in terms of the *distortion*

$$D(Q) = E\|Q(X) - X\|^2 = \int \|Q(x) - x\|^2 dP(x)\,, \tag{1}$$

where $\|\cdot\|$ is the ordinary Euclidean norm on $\mathbb{R}^d$.

2.2 Half-spaces and Polytopes

A closed halfspace $H \subseteq \mathbb{R}^d$ is any set of the form $H = \{x : x \cdot \omega \geq a\}$, where $\omega \in \mathbb{R}^d$ is a fixed weight vector, and a is a real-valued threshold. An open half-space is the complement of a closed halfspace. Let $\mathcal{H}$ denote the collection of all open and closed halfspaces in $\mathbb{R}^d$. A halfspace H is said to be *axis-parallel* if one component of its weight vector has absolute value one, and the rest are zero, e.g. $H = \{x : x_i \geq a\}$. Let $\mathcal{H}_0$ denote the collection of all axis-parallel halfspaces.

In what follows, "polytope" refers to any finite intersection of half-spaces. Thus a polytope may be bounded or unbounded, closed, open, or neither. Let $\mathcal{U}$ denote the collection of all d-dimensional polytopes.

2.3 Tree-structured Partitions

Let T be a finite binary tree. The *depth* $d(t)$ of a node $t \in T$ is the length of the shortest path from t to the root node of T. The root node itself has depth zero, its children have depth one, their children have depth two, 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' have the same root node, 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 $T' < T$. For every tree T and every integer $r \geq 0$ let T^r be the truncated subtree of T containing only those nodes $t \in T$ for which $d(t) \leq r$.

A *tree-structured partition* is described by a pair (T, τ), where T is a binary tree and $\tau : T \to \mathbb{R}^d$ is a node function that assigns a *test vector* in $\mathbb{R}^d$ to every $t \in T$. Every vector $x \in \mathbb{R}^d$ is associated with a descending path in T through a sequence of binary comparisons: beginning at the root, and at each subsequent internal node of T, x moves to that child of its current node whose test vector is nearest to x in Euclidean distance. In case of ties, x moves to the *left* child of its current node. The *cell* U_t of a node $t \in T$ is the set of vectors x whose path contains t. Thus the cell of the root node is $\mathbb{R}^d$, and the cell of an internal node is split between its children by the hyperplane that forms the perpendicular bisector of their test vectors. The cell of a node t with depth $d(t) = k$ is a polytope having at most k faces.

More formally, let $t_0, t_1, \ldots, t_k$ be a descending path in T from the root node t_0 to another node $t = t_k$. For $j = 1, \ldots, k$ let $u_j = \tau(t_j)$, and let u'_j be the test vector assigned

to the sibling of t_j. Then

$$U_t = \bigcap_{j \in A} \{x : \|x - u_j\| \leq \|x - u'_j\|\} \cap \bigcap_{j \in B} \{x : \|x - u_j\| < \|x - u'_j\|\},$$

where A contains those indices j for which t_j is the left sibling of its parent, and B contains those indices for which it is the right sibling of its parent. The cells $\{U_t : t \in \tilde{T}\}$ associated with the terminal nodes of T form a partition of $\mathbb{R}^d$. This collection will be referred to as the partition defined by (T, τ).

2.4 Tree-structured Vector Quantizers

A *tree-structured vector quantizer* (TSVQ) is described by a triple (T, τ, Rep), where (T, τ) is a tree-structured partition and $\text{Rep} : T \to \mathbb{R}^d$ is a node function that assigns a vector representative to each $t \in T$. $\text{Rep}(\cdot)$ defines a vector quantizer by assigning a vector representative to each cell of the partition defined by (T, τ), formally,

$$T(x) = \sum_{t \in \tilde{T}} \text{Rep}(t) I\{x \in U_t\}. \tag{2}$$

In what follows T will be used to denote a binary tree, a triple (T, τ, Rep) describing a tree-structured vector quantizer, and the associated mapping given by (2). In each case its intended meaning will be clear from the context. When the quantizer T defined in (2) is applied to a random variable X, its distortion is given by

$$D(T) = E\|X - T(X)\|^2 = \int \|x - T(x)\|^2 dP(x).$$

Given a triple (T, τ, Rep) and a subtree $T' \leq T$, there is a unique quantizer (T', τ', Rep') that is determined by restricting the domains of τ and Rep to T'. Following the notational convention above, this will be abbreviated by writing $T' \leq T$ without further comment.

3 The Recursive Growing Procedure

This section gives a complete description of the recursive growing procedure, which is presented in 3.3 below. The next two subsections are devoted to reduction of distortion and local splitting rules for the procedure.

3.1 Splitting and Reduction of Distortion

Here "splitting" refers to the partitioning of a convex polytope $U \in \mathcal{U}$ by a half-space $H \in \mathcal{H}$ into two constituent polytopes $U \cap H$ and $U \cap H^c$ that are themselves elements of $\mathcal{U}$. Fix a

distribution P on $\mathbb{R}^d$ having finite second moment. If every vector belonging to a polytope $U \in \mathcal{U}$ is assigned to a single representative u, the resulting distortion is given by

$$D(U,u) = \int_U \|x-u\|^2 dP(x)\,.$$

The optimal representative for U is its *centroid* (or center of mass) with respect to P, namely

$$c = \mathrm{c}(U,P) = \frac{1}{P(U)} \int_U x\, dP \;\in\; \mathbb{R}^d\,.$$

The distortion achieved by the c is denoted by

$$D^*(U) = D(U,c) = \inf_{u \in \mathbb{R}^d} D(U,u)\,.$$

If U is split by a halfspace H, and an optimal representative is assigned to each of the resulting polytopes, then the resulting reduction of distortion is given by

$$\Delta D(U:H) = D^*(U) - D^*(U \cap H) - D^*(U \cap H^c)\,. \tag{3}$$

It is shown below that $\Delta D(U:H) \geq 0$ for every $U \in \mathcal{U}$ and every $H \in \mathcal{H}$. The larger the value of $\Delta D(U:H)$, the more effective is splitting with H as a means reducing distortion.

Remark: Both $D^*(\cdot)$ and $\Delta D(\cdot)$ depend on an underlying distribution P on $\mathbb{R}^d$. When a sequence $P_1, P_2, \ldots$ of such distributions and a fixed reference distribution P are under consideration, quantities evaluated with respect to P_n will be subscripted by n, e.g. $D_n^*(\cdot)$ and $\Delta D_n(\cdot)$, while those evaluated with respect to P will be written without subscripts, as above.

3.2 Local Splitting Rules

Every halfspace split of a polytope reduces distortion, but some halfspaces are better than others. The selection of a splitting halfspace for a given region under a known distribution is carried out by a *local splitting rule*. Let $\mathcal{P}$ denote the set of all distributions on $\mathbb{R}^d$ having finite second moment. Formally, a local splitting rule is a function $\psi : \mathcal{U} \times \mathcal{P} \to \mathcal{H}$ that selects a *closed* halfspace to split a polytope U under a distribution P. In applications of the recursive growing procedure, U is a cell of a tree-structured partition, P is the empirical distribution of a finite training sequence $X_1, \ldots, X_n$, and ψ is an algorithm that seeks to find a splitting halfspace for which the reduction of distortion is large. See Murthy, Kasif, and Salzberg [19] for an account of splitting rules in the context of decision trees.

In the simplest case ψ selects a halfspace in order to maximize the reduction of distortion over *all* possible candidates, i.e.

$$\psi_1(U, P) = \arg\max_{H \in \mathcal{H}} \Delta D(U : H) .$$

While ψ_1 is optimal, it is not possible in practice to find the best split of a region U in an efficient fashion when the training sequence is large and $d > 1$. As a more efficient alternative, one may select the best candidate from a subset $\mathcal{H}' \subseteq \mathcal{H}$ for which a search is computationally feasible, i.e.

$$\psi_2(U, P) = \arg\max_{H \in \mathcal{H}'} \Delta D(U : H) .$$

The most natural choice of $\mathcal{H}'$ is the collection $\mathcal{H}_0$ of axis-parallel halfspaces, in which case ψ_2 is implemented as follows: (1) for $k = 1, \ldots, d$ sort the training vectors $X_1, \ldots, X_n$ according to the values of their k'th coordinate; (2) search for the best halfspace perpendicular to each coordinate; (3) select from among the d candidate halfspaces that one giving the greatest overall reduction of distortion. These steps require $O(dn \log n)$ operations in the worst case.

Iterative methods, such as the Generalized Lloyd (2-means) algorithm [14], successively improve the performance of any given halfspace. When iterative methods fail to yield substantial improvements in performance, local perturbation and random search may be used to find a new initial halfspace to which the method can be applied, thereby avoiding local minima. In this case the overall best half-space encountered is taken to be the output of the rule.

Definition: A local splitting rule ψ will be called *admissible* if, for every $U \in \mathcal{U}$ and every distribution $P \in P$,

$$\Delta D(U : \psi(U, P)) \geq \sup_{H \in \mathcal{H}_0} \Delta D(U : H) .$$

Thus admissible local splitting rules reduce the distortion of every polytope by at least as much as the best axis-parallel halfspace. Initiating iterative methods and random searches with the best $H \in \mathcal{H}_0$ insures that these methods are admissible, at moderate computational cost. These and other admissible methods can be readily implemented on a digital computer.

3.3 Description of the Growing Procedure

The recursive growing procedure takes as input a distribution P on $\mathbb{R}^d$ having finite second moment, and an integer k specifying the maximum number of iterations it will perform. During its execution the procedure produces a nested sequence of TSVQ. The initial tree

consists of a single root node. Subsequent trees are obtained by splitting a single terminal node of the tree produced at the previous step. In the basic iterative step of the algorithm a terminal region of the current tree is selected and split by a halfspace: children are appended to the corresponding node of the tree and the node functions $\tau(\cdot)$ and $\mathrm{Rep}(\cdot)$ are extended to the new tree in an optimal way. Selection of the terminal region and halfspace is based on the reduction of distortion ΔD and the local splitting rule ψ. Termination of the algorithm occurs after k iterations, or when no improvement of the current tree is possible.

Fix in Advance: A local splitting rule ψ.

Inputs: (1) A distribution P on $\mathbb{R}^d$ with finite second moment, and (2) an iteration count $k \geq 0$.

Initialize: Set $j = 0$. Form an initial tree T_0 consisting of a single root node t_0 for which (a) $\tau(t_0) = \mathbf{0}$ and (b) $\mathrm{Rep}(t_0) = \mathrm{c}(\mathbb{R}^d, P)$.

Iterate:

1. For each $t \in \tilde{T}_j$ obtain a half-space $H_t = \psi(U_t, P)$.

2. If $\Delta D(U_t, H_t) = 0$ for every $t \in \tilde{T}_j$ then output T_j and terminate.

3. Otherwise, select $t^* = \arg\max_{t \in \tilde{T}} \Delta D(U_t : H_t)$.

4. Produce T_{j+1} by adding left and right children t_1 and t_2, respectively, to t^*.

5. Augment the node functions $\tau(\cdot)$ and $\mathrm{Rep}(\cdot)$ as follows:

- **a.** Select $\tau(t_1)$ and $\tau(t_2)$ so that $H = \{x : \|x - \tau(t_1)\| \leq \|x - \tau(t_2)\|\}$.
- **b.** Define $\mathrm{Rep}(t_i) = \mathrm{c}(U_{t_i}, P)$ for $i = 1, 2$.

6. Increment $j := j + 1$. If $j = k$ then output T_k and stop.

Remarks:

a. The procedure treats bounded and unbounded cells in the same way. As P has finite second moment, centroids $\mathrm{c}(U, P)$ exist for every $U \in \mathcal{U}$.

b. Recall that $H_t = \psi(U_t, P)$ is closed by definition. The selection of $\tau(t_1)$ and $\tau(t_2)$ insures that ties are broken in favor of the left daughter node, but any two vectors having the boundary of H as their perpendicular bisector will do.

3.4 Non-uniqueness

The recursive growing procedure is not guaranteed to produce a unique sequence of trees from a fixed input. Non-uniqueness arises from ties that may occur during the procedure's operation. A tie *between* nodes occurs when two or more terminal nodes maximize ΔD, in which case the algorithm may split any one of these nodes and then continue. In the same way, a local splitting rule seeking to maximize $\Delta D(U_t : H)$ over $H \in \mathcal{H}'$ at some leaf t may need to choose between several equivalent alternatives (a tie *within* a node). Ties typically result from symmetries in the input distribution. Non-uniqueness of the algorithm is addressed by describing its behavior in terms of an *ensemble* of possible outcomes.

Definition: Fix a local splitting rule ψ. Let $\mathrm{Alg}(P, k)$ be the collection of possible TSVQ produced by applying the recursive procedure to P for k steps using the rule ψ. Let

$$\mathrm{Alg}(P) = \bigcup_{k=0}^{\infty} \mathrm{Alg}(P, k)$$

be the collection of all TSVQ that can be produced by the procedure from P in any finite number of steps.

4 Overview of Principal Results

Analysis of an unsupervised procedure is invariably complicated by the fact that one cannot impose, for the sake of technical convenience, additional structure or constraints on its operation. There does not appear to be a direct connection between the iterative reduction of local distortion, and vanishing global distortion. When successive trees are produced from different distributions, T_n need not be a subtree of T_{n+1}, and indeed the two may have markedly different structures. We have endeavored to make minimal assumptions on the limiting, or fixed, distribution P. Implicitly at least, the analysis must contend with unbounded cells, cells having large aspect ratios, and irregular support sets. Assuming that P is compactly supported, or that P has a continuous density bounded away from zero on a convex set, would lead to simpler proofs.

4.1 Convergent Distributions

In practice, the recursive growing procedure is applied to the empirical distribution $\hat{P}_n$ of a sequence $X_1, \ldots, X_n$ of random vectors. The X_i are are obtained through experimentation or simulation, and are usually assumed to be stationary and ergodic. Consider a tree

$T_n \in \text{Alg}(\hat{P}_n, k_n)$ that is produced from $\hat{P}_n$ in k_n steps. If n is large then $\hat{P}_n$ should approximate the common distribution P of the vectors X_i. If k_n is also large, the procedure should insure that $D(T_n) \approx D_n(T_n) \approx 0$. We first investigate the behavior of the procedure with respect to a fixed, convergent sequence of distributions, and then specialize to the empirical distributions of a stationary ergodic process.

Definition: Let $P_1, P_2, \ldots$ and P be probability distributions on $\mathbb{R}^d$. The sequence P_n converges to P, written $P_n \to P$, if

$$\int f dP_n \to \int f dP$$

for every function $f : \mathbb{R}^d \to \mathbb{R}$ that is integrable with respect to P.

Fix an admissible local splitting rule ψ and suppose that $P_1, P_2, \ldots$ converge to a distribution P. Let $T_n \in \text{Alg}(P_n, k_n)$ be produced from P_n in k_n steps. The quantity $D_n(T_n)$ measures the distortion of T_n with respect to the distribution P_n from which it was produced. Adopting terminology that is standard when $P_n = \hat{P}_n$ is the empirical distribution of a finite training set, $D_n(T_n)$ will be referred to as the *empirical distortion* of T_n.

Theorem 1 *Let ψ be an admissible splitting rule, and suppose that $P_1, P_2, \ldots$ converge to a distribution P with finite second moment. If $T_n \in \text{Alg}(P_n, k_n)$ and $k_n \to \infty$, then $D_n(T_n) \to 0$.*

If the recursive growing procedure is applied to a fixed distribution P without a termination criterion (e.g. $k = \infty$), it will produce a (possibly infinite) sequence of tree-structured vector quantizers. Each iteration of the algorithm reduces distortion so the sequence of distortions has a limit. Setting $P_n = P$ in the theorem above shows that the limit is zero when ψ is admissible.

Theorem 2 *Let ψ be an admissible local splitting rule, and let P be any distribution with finite second moment. If the recursive growing procedure is applied repeatedly to P, it will produce a sequence $T_0 \leq T_1 \leq \ldots$ for which $D(T_n) \to 0$.*

In practice, the empirical distortion of a tree T_n is typically of less interest than its distortion under the (unknown) limiting distribution P. Our main result concerns the asymptotic behavior of T_n with respect to P. Recall that $T[x] = \{x' : T(x') = T(x)\}$ is the

cell-function of T, and that the diameter of a set $U \subseteq \mathbb{R}^d$ is the maximum distance between any two points of U, $\mathrm{diam}(U) = \sup_{u,v \in U} \|u - v\|$.

Theorem 3 *Let ψ be an admissible splitting rule, and suppose that $P_1, P_2, \ldots$ converge to an absolutely continuous distribution P with finite second moment. If $T_n \in \mathrm{Alg}(P_n, k_n)$ and $k_n \to \infty$, then*

a. $D(T_n) \to 0$.

b. $P\{x : \mathrm{diam}(T_n[x]) > \epsilon\} \to 0$ *for every* $\epsilon > 0$.

Note that $\mathrm{diam}(T_n[x])$ accounts for all the points in the cell, not just those that lie in the support of P. Theorem 3 may be applied to the empirical distributions of a stationary ergodic process. Let $X_1, X_2, \ldots$ be a stationary ergodic sequence of random vectors in $\mathbb{R}^d$ and let $\hat{P}_n$ be the empirical distribution of $X_1, \ldots, X_n$.

Theorem 4 *Let ψ be an admissible splitting rule, and suppose that the distribution P of X_1 is absolutely continuous with finite second moment. If $T_n \in \mathrm{Alg}(\hat{P}_n, k_n)$ and $k_n \to \infty$, then with probability one,*

a. $D(T_n) \to 0$.

b. $P\{x : \mathrm{diam}(T_n[x]) > \epsilon\} \to 0$ *for every* $\epsilon > 0$.

4.2 General Distortion Measures

Implementation of the recursive growing procedure is governed by the local splitting rule ψ and the reduction of distortion ΔD. The use of squared Euclidean distance as a distortion measure is not critical to the analysis. Consideration of more general distortion measures is possible, at the expense of some technical complications in the proofs. Let $\rho : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a distortion measure having the following properties:

a. $\rho(u, v) \geq 0$ and $\rho(u, v) = 0$ if and only if $u = v$.

b. $\rho(u, v)$ is jointly continuous in both variables.

c. If $\|x - u\| > \|x - v\|$ then $\rho(x, u) > \rho(x, v)$.

d. For every compact set $\Lambda \subset \mathbb{R}^d$, and every sequence of vectors $v_n \to \infty$,

$$\min_{x \in \Lambda} \rho(x, v_n) \to \infty.$$

Following the definitions for squared-error distortion given in the previous section, for each polytope $U \in \mathcal{U}$ and each $v \in \mathbb{R}^d$ let

$$D_\rho(U, v) = \int_U \rho(x, v) dP(x) \,,$$

be the average distortion incurred when every element of U is assigned to a single representative v, and let

$$D^*_\rho(U) = \inf_{v \in \mathbb{R}^d} D(U, v)$$

be the optimal single-representative distortion for U with respect to ρ. Any vector v achieving the infimum above is referred to as a centroid of U with respect to P. All the results of the previous section apply to a recursive growing procedure that is defined in terms of the reduction

$$\Delta D_\rho(U : H) = D^*_\rho(U) - D^*_\rho(U \cap H) - D^*_\rho(U \cap H^c) \,,$$

provided that the moment condition is replaced by the assumption that the (fixed or limiting) distribution P satisfies

$$\int \max_{v \in \Lambda} \rho(x, v) dP(x) < \infty$$

for every compact set $\Lambda \subseteq \mathbb{R}^d$. In this more general setting one may consider r'th power distortion measures $\rho(u, v) = \|u - v\|^r$ with $r \geq 1$, and input-weighted distortion measures of the form $\rho(u, v) = \sigma(u)\|u - v\|^r$.

5 Preliminary Results

In what follows it will be assumed that the moment condition

$$M(P) \stackrel{\triangle}{=} \int \|x\|^2 dP < \infty$$

holds for every distribution P under consideration. This insures that $D^*(U)$ and $\Delta D(U : H)$ are well-defined for every polytope U and every halfspace H. In order to simplify notation, define

$$\Delta D(U : \psi) = \Delta D(U : \psi(U, P))$$

whenever $U \in \mathcal{U}$ and P is the same distribution that is used to evaluate ΔD.

Several basic properties of the ΔD splitting criterion are summarized in the following propositions. Propositions 1 and 2 appear in [6]: proofs are given below for the sake of completeness. See [7] for more details concerning monotone tree functionals in the context of pruning.

Proposition 1 *For every polytope $U \in \mathcal{U}$ and every half-space $H \in \mathcal{H}$, the quantity $\Delta D(U : H)$ is non-negative.*

Proof: If $u = \mathrm{c}(U, P)$, then by the definition of $D^*(\cdot)$,

$$\begin{aligned} D^*(U) &= \int_{U \cap H} \|x - u\|^2 dP(x) + \int_{U \cap H^c} \|x - u\|^2 dP(x) \\ &\geq D^*(U \cap H) + D^*(U \cap H^c)\,, \end{aligned}$$

so that $\Delta D(U : H) \geq 0$.

Proposition 2 *If $T \in \mathrm{Alg}(P)$ and $S \leq T$ then $D(S) \geq D(T)$.*

Proof: If $S \neq T$ then there exists a node $s \in \tilde{S}$ whose children s_1 and s_2 are nodes of T. Let S' be the subtree of T having nodes $S \cup \{s_1, s_2\}$. Consider the hyperplane $H = \psi(P, U_s) \in \mathcal{H}$ such that $U_{s_1} = U_s \cap H$ and $U_{s_2} = U_s \cap H^c$. Then

$$D(S) - D(S') = D^*(U_s) - D^*(U_{s_1}) - D^*(U_{s_2}) = \Delta D(U_s : H) \geq 0$$

If $S' \neq T$ repeat the argument. At each step distortion is reduced, and the result follows. □

Proposition 3 *For each local splitting rule ψ and each tree-structured partition (T, τ), $\sum_{t \in T} \Delta D(U_t : \psi) \leq M(P)$.*

Proof: By expanding each term in the sum using (3) it is evident that

$$\begin{aligned} \sum_{t \in T} \Delta D(U_t : \psi) &= D^*(\mathbb{R}^d) - \sum_{t \in \tilde{T}} \Delta D(U_t : \psi) \\ &\leq D^*(\mathbb{R}^d)\,. \end{aligned}$$

The optimality of $D^*(\cdot)$ insures that the last term is less than $M(P)$. □

The following lemmas summarize two geometric properties of distortion that will play a key role in the results that follow.

Lemma 1 *If ψ is an admissible splitting rule, then for every set $U \subseteq \mathbb{R}^d$, every distribution P, and every number $\beta > 0$,*

$$\Delta D(U : \psi) \geq \frac{\beta^2}{2d^2} \cdot P\{x \in U : \|x - \mathrm{c}(U, P)\| > \beta\}\,. \tag{4}$$

Proof: Fix a distribution P, a polytope $U \in \mathcal{U}$ with $P(U) > 0$, and a number $\beta > 0$. As ψ is permissible, it suffices to exhibit a halfspace $H \in \mathcal{H}_0$ such that $\Delta D(U : H)$ satisfies the stated bound. For fixed $a > 0$ and $k = 1, \ldots, d$ define closed axis-parallel halfspaces

$$H_k^+(a) = \{x : x_k \geq a\} \quad \text{and} \quad H_k^-(a) = \{x : x_k \leq -a\}\,,$$

and let $B(x, \lambda)$ be the Euclidean ball of radius λ centered at the vector x. A straightforward argument shows that for every $\beta > 0$,

$$B(0, \beta)^c \subseteq \bigcup_{k=1}^{d} [H_k^+(d^{-1/2}\beta) \cup H_k^-(d^{-1/2}\beta)]\,.$$

Let $c = \mathrm{c}(U, P)$ and define halfspaces $H_{2j-1} = H_j^-(d^{-1/2}\beta) + c$ and $H_{2j} = H_j^+(d^{-1/2}\beta) + c$ for $j = 1, \ldots, d$. Then

$$B(c, \beta)^c \subseteq \bigcup_{i=1}^{2d} H_j\,,$$

and by the union bound

$$P(U \cap B(c, \beta)^c) \leq \sum_{j=1}^{2d} P(U \cap H_j)\,.$$

Therefore there is a half-space H among $H_1, \ldots, H_{2d}$ for which

$$P(U \cap H) \geq \frac{P(U \cap B(c, \beta)^c)}{2d}\,. \tag{5}$$

We will show that $\Delta D(U : H)$ satisfies the lower bound (4) above.

Let v^* be the vector in H closest to c. Then the inequality $(c - v^*)^t(x - v^*) \leq 0$ holds for every $x \in H$ (cf. [15][p. 50]), so that

$$\|x - c\|^2 - \|x - v^*\|^2 \geq \|c - v^*\|^2 = \frac{\beta^2}{d}\,.$$

If c_1 is the centroid of $V = U \cap H$ and c_2 is the centroid of $W = U \cap H^c$, then

$$\begin{aligned}
\Delta D(U : H) &= D(U, c) - D(V, c_1) - D(W, c_2) \\
&= (D(V, c) - D(V, c_1)) + (D(W, c) - D(W, c_2)) \\
&\geq D(V, c) - D(V, c_1) \\
&\geq D(V, c) - D(V, v^*) \\
&= \int_V (\|x - c\|^2 - \|x - v^*\|^2) dP \\
&\geq \frac{\beta^2}{d} P(V) \\
&\geq \frac{\beta^2}{2d^2} P(U \cap B(c, \beta)^c)\,.
\end{aligned}$$

The first two inequalities follow from the choice of c_1 and c_2, while the last follows from (5). □

Lemma 2 *Let $P_1, P_2, \ldots$ converge to P with $M(P) < \infty$. For every $\delta > 0$ there exist constants $\beta, \gamma > 0$, depending only on δ and P, such that for every tree $T \in \mathrm{Alg}(P_n)$ with $D_n(T) \geq \delta$,*

$$P_n\{x : \|x - T(x)\| \geq \beta\} > \gamma$$

when n is sufficiently large.

Proof: Suppose that there is a number $K < \infty$ such that $P_n(B(0, K)) = 1$ for each n. Then $\|T(x)\| \leq K$ for each $T \in \mathrm{Alg}(P_n)$ and each $x \in \mathbb{R}^d$. If some $T \in \mathrm{Alg}(P_n)$ satisfies $D_n(T) \geq \delta > 0$ then as $\|x - T(x)\|^2 \leq 4K^2$,

$$\begin{aligned} \delta &\leq \int \|x - T(x)\|^2 dP_n \\ &\leq \frac{\delta}{2} + 4K^2 P_n \left\{ x : \|x - T(x)\| \geq \sqrt{\frac{\delta}{2}} \right\} \end{aligned}$$

and consequently

$$P_n \left\{ x : \|x - T(x)\| \geq \sqrt{\frac{\delta}{2}} \right\} > \frac{\delta}{4K^2}$$

which gives the desired bound. The more general case, in which the distributions P_n do not share a bounded support set, is considered in Appendix A below. □

6 Convergence of Empirical Distortions

When it is applied to a fixed distribution, the recursive growing procedure creates a nested sequence of tree-structured vector quantizers. Every element $T \in \mathrm{Alg}(P, k)$ has an associated *trajectory* of the form

$$S_0 < S_1 < \ldots < S_{k'} \,. \tag{6}$$

The initial tree S_0 consists of a single root node, the final tree $S_{k'} = T$, and for each $j = 0, 1, \ldots, k' - 1$ the tree S_{j+1} is produced from S_j by one iteration of the growing procedure. The trajectory of T explicitly describes its production from the root node under the distribution P. When $k' = k$ the trajectory (6) is said to be *complete*. If $k' < k$ then the algorithm terminated prior to its k'th iteration, and in this case $\Delta D(U_t, \psi(P, U_t)) = 0$ for every $t \in \tilde{T}$. The analytical properties of such trajectories were studied in [23]. Trajectories

of a different sort, based on iterations of a continuous map, were studied by Sabin and Gray [32] in their analysis of the empirical behavior of the Lloyd algorithm.

Lemma 3 *If $T \in \mathrm{Alg}(P,k)$ has a complete trajectory of the form $S_0 < S_1 < \ldots < S_k = T$ then*

$$\min_{1\le j\le k} \sum_{s\in\tilde{S}_{j-1}} \Delta D(U_s:\psi) \;\le\; \frac{M(P)}{w(k)}$$

where $w(k) = \sum_{j=1}^{k} j^{-1}$.

Proof: For $j = 1,\ldots,k$ let $s_{j-1} \in \tilde{S}_{j-1}$ be the terminal node that is split to form S_j. Since S_{j-1} has exactly j such nodes, the greedy node selection protocol insures that

$$j^{-1} \sum_{s\in\tilde{S}_{j-1}} \Delta D(U_s:\psi) \;\le\; \Delta D(U_{s_{j-1}}:\psi)\,.$$

Therefore,

$$w(k)\cdot \min_{1\le j\le k} \sum_{s\in\tilde{S}_{j-1}} \Delta D(U_s:\psi) \;\le\; \sum_{j=1}^{k} \Delta D(U_{s_{j-1}}:\psi) \;\le\; \sum_{t\in T} \Delta D(U_t:\psi)\,,$$

which is less than $M(P)$ by Proposition 3. □

Proof of Theorem 1: If T_n has an incomplete trajectory then for each $t \in \tilde{T}_n$, $\Delta D_n(U_t : \psi) = 0$. It follows from Lemma 1 that $P_n(U_t)$ is zero, or is concentrated at the single point $\mathrm{c}(U_t, P_n)$. In either case, $D_n^*(U_t) = 0$ for each $t \in \tilde{T}_n$, and it follows that $D_n(T_n) = 0$.

Now fix $\delta > 0$ and consider a tree $T_n \in \mathrm{Alg}(P_n, k_n)$ for which $D_n(T_n) > \delta$. The trajectory of T_n is necessarily complete, and by the preceding Lemma there exists $S_n \le T_n$ such that $S_n \in \mathrm{Alg}(P_n)$ and

$$\sum_{s\in\tilde{S}_n} \Delta D_n(U_s:\psi) \;\le\; \frac{M(P_n)}{w(k_n)}\,. \tag{7}$$

It follows from Proposition 2 that $D_n(S_n) > \delta$, and in view of Lemma 2, there exist constants $\beta, \gamma > 0$, depending only on δ and P, such that

$$P_n\{x : \|x - S_n(x)\| \ge \beta\} > \gamma$$

when n is large. As the terminal regions of S_n form a partition of $\mathbb{R}^d$, the admissibility of ψ and Lemma 1 imply that

$$\sum_{s\in\tilde{S}_n} \Delta D_n(U_s:\psi) \;\ge\; \frac{\beta^2}{2d^2} P_n\{x : \|x - S_n(x)\| \ge \beta\} \;\ge\; \eta\,, \tag{8}$$

where $\eta = \gamma\beta^2/2d^2 > 0$ is independent of n. As $M(P_n) \to M(P) < \infty$ and $w(k_n)$ grows without bound as n tends to infinity, (7) and (8) imply that $D_n(T_n) > \delta$ for at most finitely many values of n. As $\delta > 0$ was arbitrary, the proof is complete. □

7 Asymptotic Distortion for Convergent Distributions

This section is concerned with the distortion of trees produced from a convergent sequence of distributions $P_n \to P$. Consider trees $T_n \in \text{Alg}(P_n, k_n)$. If $k_n \to \infty$ then the empirical distortion $D_n(T_n) \to 0$ by Theorem 1. It is natural to assume that $D_n(T_n)$ is close to $D(T_n)$ when n is large because $P_n \approx P$, but a rigorous proof must address two critical problems. The first problem is that the trees T_n grow larger and more complicated as k_n tends to infinity. The second problem is that the non-uniqueness of $\text{Alg}(\cdot)$ and the variation of P_n with n make T_n a 'moving target'. In particular, T_n need not be a subtree of T_{n+1}.

Lemma 4 shows that for a fixed margin of error $\epsilon > 0$, we may focus our attention on truncated versions T_n^r of T_n. Here $r = r(\epsilon)$ is a finite integer that tends to infinity as $\epsilon \to 0$. Proposition 4 shows that when r is fixed the difference $D_n(T_n^r) - D(T_n^r) \to 0$ if the representative vectors of each tree T_n^r, $n \geq 1$, are contained in a fixed compact set. The existence of such a compact set is established in Lemma 5.

Lemma 4 *Let $\{P_n\}_{n=1}^{\infty}$ be as in Theorem 1 and let $T_n \in \text{Alg}(P_n, k_n)$ for $n \geq 1$. If $k_n \to \infty$ then*

$$\lim_{r \to \infty} \left[\limsup_{n \to \infty} D_n(T_n^r) \right] = 0 \tag{9}$$

Proof: It is enough to show that $D_n(T_n^{r_n}) \to 0$ for every increasing sequence of integers $r_n \to \infty$. Fix such a sequence and recall that the production of each tree $T_n \in \text{Alg}(P_n, k_n)$ is described by a finite trajectory

$$\text{root} = S_{0,n} \leq S_{1,n} \leq \ldots \leq S_{k_n',n} = T_n \, .$$

where $k_n' \leq k_n$. Let l_n be the last stage at which every vertex of $S_{l,n}$ has depth at most r_n,

$$l_n = \max\{l : d(s) \leq r_n \text{ for every } s \in S_{l,n}\} \, ,$$

and define the corresponding tree

$$S_n = S_{l_n,n} \in \text{Alg}(P_n, l_n) \, .$$

Partition the integers into sets

$$\mathcal{N}_1 = \{n : r_n \geq k_n'\} \;\text{ and }\; \mathcal{N}_2 = \{n : k_n' > r_n\} \, .$$

If $n \in \mathcal{N}_1$ then $l_n = k_n'$ so that $S_n = T_n$. Therefore $D_n(S_n) = D_n(T_n) \to 0$ if $n \in \mathcal{N}_1$ tends to infinity. If $n \in \mathcal{N}_2$ it is easy to see that $l_n \geq r_n$. As $r_n \to \infty$ it follows from Theorem 1 that

$D_n(S_n) \to 0$ if $n \in \mathcal{N}_2$ tends to infinity. Thus $D_n(S_n) \to 0$ and since $D_n(T_n^{r_n}) \leq D_n(S_n)$ it follows that $D_n(T_n^{r_n}) \to 0$. $\square$

Definition: For each integer $r \geq 0$ and every number $K \in (0, \infty)$ let $\mathcal{T}(r, K)$ be the family of all tree-structured vector quantizers T such that

a. $d(t) \leq r$ for every $t \in T$, and

b. $\|T(x)\| \leq K$ for every $x \in \mathbb{R}^d$

Thus $\mathcal{T}(r, K)$ contains all those TSVQ whose leaves have depth at most r, and whose codewords lie in a sphere of radius K about the origin.

The proofs of Lemma 5 and Proposition 4 can be found in Appendices B and C, respectively.

Lemma 5 *Let $P_1, P_2, \ldots$ converge to an absolutely continuous distribution P with $M(P) < \infty$, and let $T_n \in \mathrm{Alg}(P_n)$ for $n \geq 1$. For each $r \geq 0$ there exists a number $K(r) < \infty$ such that*

$$\| \mathrm{c}(U_t, P_n)\| \leq K(r)$$

for each $n \geq 1$ and each $t \in T_n^r$.

Proposition 4 *If $P_n \to P$ with $M(P) < \infty$ then for each fixed $r \geq 0$ and each number $K \in (0, \infty)$,*

$$|D(T) - D_n(T)| \to 0 \tag{10}$$

uniformly over T in $\mathcal{T}(r, K)$.

In order to establish Theorem 3 it is first necessary to establish an asymptotic connection between the distortion of a vector quantizer and the size of its cells. Recall that the cell-function of a quantizer Q is defined by $Q[x] = \{x' : Q(x') = Q(x)\}$ for each $x \in \mathbb{R}^d$. The following result is due to Nobel [22].

Theorem A *Let P be an absolutely continuous distribution on $\mathbb{R}^d$ and let $Q_1, Q_2, \ldots$ be vector quantizers having convex cells. If $D(Q_n) \to 0$, then*

$$P\{x : \mathrm{diam}(Q_n[x]) > \epsilon\} \to 0$$

for every $\epsilon > 0$.

Proof of Theorem 3: An application of Lemma 5 shows that for every $r \geq 0$ and every $n \geq 1$,

$$\begin{aligned} D(T_n^r) &\leq D_n(T_n^r) + |D(T_n^r) - D_n(T_n^r)| \\ &\leq D_n(T_n^r) + \sup_{T \in \mathcal{T}(r,K(r))} |D(T) - D_n(T)| \,. \end{aligned}$$

It follows from (10) that

$$\limsup_{n\to\infty} D(T_n^r) \leq \limsup_{n\to\infty} D_n(T_n^r) \tag{11}$$

for every $r \geq 0$, and in view of (9),

$$\lim_{r\to\infty} \left[\limsup_{n\to\infty} D(T_n^r) \right] = 0 \,.$$

Thus there exist trees $S_1, S_2, \ldots$ such that $S_n \leq T_n$ for each n, and $D(S_n) \to 0$. Theorem A shows that $\mathrm{diam}(S_n[x]) \to 0$ in P-probability. But $S_n \leq T_n$ implies that $\mathrm{diam}(S_n[x]) \geq \mathrm{diam}(T_n[x])$ for each $x \in \mathbb{R}^d$, and it follows that $\mathrm{diam}(T_n[x]) \to 0$ in P-probability.

Fix $\delta > 0$ and let U_n be the union of those cells $T_n[x]$ such that $\mathrm{diam}(T_n[x]) > \delta$. Integrating over each cell in turn, the optimality of $T(x)$ guarantees that

$$\begin{aligned} D(T_n) &\leq \delta^2 + \int_{U_n} \|x - T_n(x)\|^2 dP \\ &\leq \delta^2 + \int_{U_n} \|x\|^2 dP \end{aligned}$$

Since $P(U_n) \to 0$ and P has finite second moment,

$$\limsup_{n\to\infty} D(T_n) \leq \delta^2,$$

and as $\delta > 0$ was arbitrary, the proof is complete. □

Remark: The proof of Theorem 3 applies to suitably pruned subtrees of T_n. Let Prune$(\cdot)$ be any pruning scheme and define $S_n = \mathrm{Prune}(T_n)$ to be the pruned subtree of T_n. If for each $r \geq 1$ there exists $N(r) < \infty$ such that $S_n^r = T_n^r$ for each $n \geq N(r)$, then $D(S_n) \to 0$.

8 Large Sample Performance

In practice, recursive partitioning schemes are commonly applied to distributions derived from finite training sets, which are obtained from experiments or simulations. If the random vectors comprising the training set share a common distribution P, it is natural to ask whether large trees produced from large training sets will yield effective compression with respect to P.

It is assumed in Theorem 3 that $\int f dP_n \to \int f dP$ for every function f that is integrable with respect to P. However, a careful inspection of the proof shows that it is enough to require the convergence of integrals for every function in a countable family $\mathcal{F}$, which we now describe. Let $\mathbb{Q}$ denote the set of rational numbers. Let $\mathcal{B}_r$ be the collection of balls $B(u, a)$ and $\mathcal{H}_r$ the collection of halfspaces $H = \{x : u \cdot x \geq a\}$ with $u \in \mathbb{Q}^d$ and $a \in \mathbb{Q}$. Let $\mathcal{U}_r$ be the collection of all finite intersections of halfspaces $H \in \mathcal{H}_r$, and define $\mathcal{F}$ to be all those functions of the form

$$f(x) = \max_{v \in B_1} \|x - v\|^2 \cdot I_{B_2}(x) \quad \text{with} \quad B_1, B_2 \in \mathcal{B}_r$$

and

$$f(x) = \|x - z\|^2 \cdot I_U(x) \quad \text{with} \quad z \in \mathbb{Q}^d,\ U \in \mathcal{U}_r\,.$$

Then $\mathcal{F}$ is countable, and it can be shown that the convergence of P_n to P in Theorem 3 can be replaced by the assumption that $\int f dP_n \to \int f dP$ for every $f \in \mathcal{F}$.

Let $X_1, X_2, \ldots \in \mathbb{R}^d$ be a stationary ergodic sequence of random vectors with $X_i \sim P$, and let $\hat{P}_n$ denote the empirical distribution of $X_1, \ldots, X_n$. As $\mathcal{F}$ is countable, the ergodic theorem insures that, with probability one, $\int f d\hat{P}_n \to \int f dP$ for every $f \in \mathcal{F}$. Thus Theorem 4 follows immediately from the strengthened version of Theorem 3.

Acknowledgments

The author wishes to thank an anonymous referee who suggested a simpler means of establishing Theorem 1. Their argument now appears Lemma 3 and the proof of Theorem 1 that follows it.

Appendix A: Proof of Lemma 2

Proof of Lemma 2: We consider the general case in which the distributions P_n do not share a bounded support set. Given $\delta > 0$ set $\epsilon = \delta/5$ and define constants $\alpha_1 \leq \alpha_2 \leq \alpha_3$ as follows. Let α_1 be so large that

$$\int_{B(0,\alpha_1)^c} \|x\|^2 dP < \epsilon\,. \tag{12}$$

Choose $\alpha_2 > \alpha_1$ in order to make

$$K \;=\; \inf\{\|u - v\|^2 : \|u\| \leq \alpha_1,\ \|v\| > \alpha_2\} \;\geq\; \frac{\alpha_1^2}{\epsilon} \int \|x\|^2 dP\,, \tag{13}$$

and select $\alpha_3 > \alpha_2$ so that

$$\int_{B(0,\alpha_3)^c} \max\{\|x-v\|^2 : \|v\| \le \alpha_2\} dP < \epsilon\,. \tag{14}$$

To simplify the notation let $\Lambda_j = B(0,\alpha_j)$ for $j = 1,2,3$.

Let T be any tree in $\mathrm{Alg}(P_n)$. When n is sufficiently large the contribution to $D_n(T)$ from vectors x such that $\|x\| > \alpha_3$ or $\|T(x)\| > \alpha_2$ is uniformly bounded from above. Indeed, by virtue of (14) and the fact that $P_n \to P$,

$$\begin{aligned}
&\int \|x - T(x)\|^2 I\{x \in \Lambda_3^c \ \vee\ T(x) \in \Lambda_2^c\} dP_n \\
&\quad\le \int \|x - T(x)\|^2 I\{T(x) \in \Lambda_2^c\} dP_n + \int \|x - T(x)\|^2 I\{x \in \Lambda_3^c \ \wedge\ T(x) \in \Lambda_2\} dP_n \\
&\quad\le \int \|x - T(x)\|^2 I\{T(x) \in \Lambda_2^c\} dP_n + \int_{\Lambda_3^c} \max\{\|x-v\|^2 : v \in \Lambda_2\} dP_n \\
&\quad\le \int \|x - T(x)\|^2 I\{T(x) \in \Lambda_2^c\} dP_n + \epsilon + o(1)\,, \qquad (15)
\end{aligned}$$

where $o(1) \to 0$ as $n \to \infty$ and is independent of T. The last integral in (15) may be bounded as follows:

$$\begin{aligned}
&\int \|x - T(x)\|^2 I\{T(x) \in \Lambda_2^c\} dP_n \\
&\quad\le \int \|x\|^2 I\{T(x) \in \Lambda_2^c\} dP_n \\
&\quad\le \int \|x\|^2 I\{x \in \Lambda_1 \wedge T(x) \in \Lambda_2^c\} dP_n + \int \|x\|^2 I\{x \in \Lambda_1^c\} dP_n \\
&\quad\le \alpha_1^2 P_n\{x : \|x - T(x)\| \ge K\} + \int_{\Lambda_1^c} \|x\|^2 dP_n \qquad (16) \\
&\quad\le \frac{\alpha_1^2 D_n(T)}{K} + \epsilon + o(1) \qquad (17) \\
&\quad\le \frac{\alpha_1^2 \int \|x\|^2 dP_n}{K} + \epsilon + o(1) \\
&\quad\le 2\epsilon + o(1)\,.
\end{aligned}$$

The first inequality follows from the optimality property of centroids by integrating over each cell $T[x]$ in turn, and noting that $T \in \mathrm{Alg}(P_n)$ implies $T(x) = \mathrm{c}(T[x], P_n)$ for every x. Inequality (16) follows from the definition of K in (13), while (17) is a consequence of Markov's inequality and (12). The last inequality follows from (13).

Each of the terms $o(1)$ appearing in (15) and (17) is independent of T, and therefore, when n is sufficiently large,

$$\int \|x - T(x)\|^2 I\{x \in \Lambda_3^c \vee T(x) \in \Lambda_2^c\} dP_n \le 4\epsilon$$

for each tree $T \in \mathrm{Alg}(P_n)$. In particular, if $D_n(T) > \delta\,(= 5\epsilon)$ then

$$\int \|x - T(x)\|^2 I\{x \in \Lambda_3 \wedge T(x) \in \Lambda_2\} dP_n > \epsilon\,.$$

Since $\|x - T(x)\|^2$ is bounded on the event $\{x \in \Lambda_3 \wedge T(x) \in \Lambda_2\}$, the argument for the special case considered above applies, and the result follows. □

Appendix B: Proof of Lemma 5

We require two preliminary results regarding the asymptotic behavior of the reduction $\Delta D_n(U_n : H_n)$ when $P_n \to P$. Proposition 5 concerns non-negligible sets and Lemma 6 concerns lopsided splits.

Proposition 5 *Let $P_n \to P$ and let $U_1, U_2, \ldots \subseteq \mathbb{R}^d$ be such that $P_n(U_n) \geq \alpha > 0$. Then there exists a number $M < \infty$ such that*

$$\|\,\mathrm{c}(U_n, P_n)\| \leq M$$

for each n, and if ψ is an admissible local splitting rule, then

$$\Delta D_n(U_n : \psi) > \epsilon$$

for some fixed $\epsilon > 0$ and n sufficiently large.

Proof: Select a bounded set $\Lambda \subset \mathbb{R}^d$ such that $P(\Lambda) > 1 - \alpha/2$. Setting $z_n = \mathrm{c}(U_n, P_n)$ it is clear that

$$\begin{aligned}
\int \|x\|^2 dP_n &\geq \int_{U_n} \|x\|^2 dP_n \\
&\geq \int_{U_n} \|x - z_n\|^2 dP_n \\
&\geq \int_{U_n \cap \Lambda} \|x - z_n\|^2 dP_n \\
&\geq \min_{x \in \Lambda} \|x - z_n\|^2 P_n(U_n \cap \Lambda)\,.
\end{aligned}$$

By letting $n \to \infty$ and then rearranging terms one finds that

$$\limsup_{n \to \infty} \left[\min_{x \in \Lambda} \|x - z_n\|^2 \right] \leq \frac{2}{\alpha} \int \|x\|^2 dP \;<\; \infty\,.$$

The desired bound on $\|z_n\|$ is an immediate consequence of this last inequality because Λ was assumed to be bounded.

As for the second inequality above, the conditions on P_n and U_n hold along every subsequence of $\{1, 2, \ldots\}$, so that by passing to a suitable subsequence and renumbering if necessary, it is enough to show that

$$\limsup_{n\to\infty} \Delta D_n(U_n : \psi) > 0\,.$$

The argument above shows that the vectors $\{z_n = \mathrm{c}(U_n, P_n)\}$ are contained in a fixed compact set. Extract a convergent subsequence $\{z_m\}$ with limit z^*. As P has a density there is a number $\beta > 0$ such that $P(B(z^*, 2\beta)) \le \alpha/3$. By Proposition 1 the inequality

$$\Delta D_n(U_n : \psi) \ge \frac{\beta^2}{2d^2} \cdot P_n(U_n \cap B(z_n, \beta)^c)$$

holds for each n, and as $P_n \to P$, the choice of β insures that

$$\Delta D_n(U_n : \psi) \ge \frac{\beta^2 \alpha}{4d^2} > 0$$

when n is sufficiently large. □

Lemma 6 *Let $P_n \to P$ and let $U_1, U_2, \ldots \subseteq \mathbb{R}^d$ be such that $P_n(U_n) \ge \alpha > 0$. If $H_1, H_2, \ldots$ are half-spaces for which $P_n(U_n \cap H_n^c) \to 0$ then $\Delta D_n(U_n : H_n) \to 0$.*

Proof: Set $V_n = U_n \cap H_n$ and $W_n = U_n \cap H_n^c$ for each $n \ge 1$. By assumption $P_n(W_n) \to 0$, so that

$$\liminf_{n\to\infty} P_n(V_n) = \liminf_{n\to\infty} P_n(U_n) \ge \alpha\,.$$

By virtue of Proposition 5 there is a fixed compact set $\Lambda \subseteq \mathbb{R}^d$ containing both $\mathrm{c}(U_n, P_n)$ and $\mathrm{c}(V_n, P_n)$ for each n. If A is any subset of $\mathbb{R}^d$ then

$$\begin{aligned}
\Delta D_n(U_n : H_n) &= D_n^*(U_n) - D_n^*(V_n) - D_n^*(W_n) \\
&\le D_n^*(U_n) - D_n^*(V_n) \\
&= \inf_{v\in\Lambda} \int_{U_n} \|x - v\|^2 dP_n - \inf_{v\in\Lambda} \int_{V_n} \|x - v\|^2 dP_n \\
&\le \sup_{v\in\Lambda} \left| \int_{U_n} \|x - v\|^2 dP_n - \int_{V_n} \|x - v\|^2 dP_n \right| \\
&\le \sup_{v\in\Lambda} \int_{W_n} \|x - v\|^2 dP_n \\
&\le \int_{W_n} \sup_{v\in\Lambda} \|x - v\|^2 dP_n \\
&\le P_n(W_n) \cdot \sup\{\|u - v\|^2 : u \in A, v \in \Lambda\} + \int_{A^c} \sup_{v\in\Lambda} \|x - v\|^2 dP_n\,.
\end{aligned}$$

Thus if A is bounded,

$$\limsup_{n\to\infty} \Delta D_n(U_n : H_n) \le \int_{A^c} \sup_{v\in\Lambda} \|x - v\|^2 dP\,.$$

Suitable choice of A makes the right hand side arbitrarily small, and the result follows. □

Proof of Lemma 5: In view of Proposition 5 it suffices to show that for each $r \geq 0$ there exists $\delta > 0$ and $N < \infty$, both depending on r, such that

$$P_n(U_t) > \delta \tag{18}$$

for each $n \geq N$ and each $t \in T_n^r$. Recall that T_n^0 consists of a single root node t_0 with $U_{t_0} = \mathbb{R}^d$. If the desired property fails to hold, then let $r_0 \geq 1$ be the least integer r such that

$$\liminf_{n\to\infty} \min\{P_n(U_t) : t \in T_n^r\} = 0\,.$$

Find a sequence of nodes $\{s_n\}$, with $s_n \in T_n^{r_0-1}$, such that the cell U_n of s_n is split by a halfspace, $H_n = \psi(P_n, U_n)$ and $\liminf P_n(U_n \cap H_n^c) = 0$. As r_0 is minimal, $\liminf P_n(U_n) > 0$, and it follows from Proposition 5 that

$$\liminf_{n\to\infty} \Delta D_n(U_n : \psi) > 0\,.$$

On the other hand, Lemma 6 shows that

$$\liminf_{n\to\infty} \Delta D_n(U_n : \psi) = 0\,.$$

In this way we arrive at a contradiction and conclude that (18) must hold for every $r \geq 0$. This completes the proof. □

Appendix C: Proof of Proposition 4

Definition: Let P be a probability distribution on $\mathbb{R}^d$. A class $\mathcal{F}$ of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *bracketing* with respect to P if

a. There exists a P-integrable function F such that $|f(x)| \leq F(x)$ for every $f \in \mathcal{F}$ and every $x \in \mathbb{R}^d$,

b. For every $\epsilon > 0$ there is a finite set of functions $\mathcal{G}_\epsilon = \{g_1, ..., g_r\}$ such that every $f \in \mathcal{F}$ has bracketing functions $\underline{g}$, $\overline{g} \in \mathcal{G}_\epsilon$ with the property that

   (a) $\underline{g} \leq f \leq \overline{g}$

   (b) $\int(\overline{g} - \underline{g})dP \leq \epsilon$.

Note that the bracketing class $\mathcal{G}_\epsilon$ need not be contained in $\mathcal{F}$. The function F in condition (i) is called an *envelope* for $\mathcal{F}$. If $\mathcal{F}$ has a constant envelope $F \equiv K < \infty$, then it is said to be uniformly bounded.

As an easy consequence of the definition, it can be seen that a product of bracketing classes is again a bracketing class.

Lemma 7 *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be bracketing with respect to P. If $\mathcal{F}_1$ is uniformly bounded then 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.* □

Proposition 6 *Let Λ be a compact subset of $\mathbb{R}^d$. If $\int \|x\|^2 dP < \infty$ then the class*

$$\mathcal{G} = \{g_z(x) = \|x - z\|^2 : z \in \Lambda\}$$

is bracketing with respect to P.

Proof: The function $F(x) = \sup_{u \in \Lambda} \|x - u\|^2$ is a P-integrable envelope for $\mathcal{F}$. The finite approximation condition follows from the continuity of $h(u, v) = \|u - v\|^2$ and the compactness of Λ. (See [29] or [20] for more details.) □

A short proof of the following result can be found in [23]. See also the example in Chapter 2 of Pollard [28].

Proposition 7 *Let $\mathcal{H}$ be the set of indicator functions of open and closed half-spaces in $\mathbb{R}^d$. If P has a density then $\mathcal{H}$ is bracketing with respect to P.*

The following theorem establishes a uniformity property of bracketing classes with respect to a convergent sequence of distributions. For more details and a proof, see [8, 28].

Theorem B *If $P_n \to P$ and $\mathcal{F}$ is bracketing with respect to P then*

$$\sup_{f \in \mathcal{F}} \left| \int f dP - \int f dP_n \right| \to 0 \quad \text{as } n \to \infty\,. \qquad \square$$

Application of Theorem B and the propositions concerning bracketing classes above establishes the uniform convergence of $D_n(T)$ to $D(T)$ over $\mathcal{T}(r, M)$ when r and M are fixed.

Proof of Proposition 4: Fix r and M and let $\mathcal{H}^r$ be the r-fold product of $\mathcal{H}$ with itself. Then $\mathcal{H}^r$ is bracketing with respect to P by Lemma 7. Consider a tree $T \in \mathcal{T}(r, M)$. Each

terminal node t of T has a representative vector $z_t \in B(0, M)$ and an associated polytope U_t having at most r faces. In particular, the indicator function of U_t is contained in $\mathcal{H}^r$. As T has at most 2^r terminal nodes,

$$\begin{aligned} & |D(T) - D_n(T)| \\ & \quad \leq \sum_{t \in \tilde{T}} \left| \int_{U_t} \|x - z_t\|^2 dP - \int_{U_t} \|x - z_t\|^2 dP_n \right| \\ & \quad \leq 2^r \cdot \sup_{f \in \mathcal{F}} \left| \int f dP - \int f dP_n \right| , \end{aligned}$$

where

$$\mathcal{F} = \{ f_z(x) = \|x - z\|^2 \cdot h(x) : z \in B(0, M) \, , \, h \in \mathcal{H}^r \} \, .$$

Note that the upper bound above does not depend on the choice of $T \in \mathcal{T}(r, M)$. Application of Lemma 7 to $\mathcal{H}^r$ and the class $\mathcal{G}$ of Proposition 6 shows that $\mathcal{F}$ is bracketing with respect to P. The result now follows from Theorem B. □

References

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

[2] M. Balakrishnan. *Variable Rate Structured Vector Quantization and Applications to Multiresolution Image Coding.* PhD thesis, Rensselaer Polytechnic Institute, Troy, NY, 1991.

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

[4] J.H. Butler, E.A. Gilpin, L. Gordon, and R.A. Olshen. Tree-structured survival analysis, II. Technical Report 133, Stanford University, Division of Biostatistics, September 1989.

[5] A. Buzo, A.H. Gray, R.M. Gray, and J.D. Markel. Speech coding based upon vector quantization. *IEEE Trans. Acoust. Speech Signal Process.*, 28:562–574, 1980.

[6] P.A. Chou. *Applications of Information Theory to Pattern Recognition and the Design of Decision Trees and Trellises.* PhD thesis, Stanford University, Information Systems Lab, 1988.

[7] P.A. Chou, T. Lookabaugh, and R.M. Gray. Optimal pruning with applications to tree-structered source coding and modeling. *IEEE Trans. Inform. Theory*, 37:31–42, 1989.

[8] J. DeHardt. Generalizations of the Glivenko-Cantelli theorem. *Ann. Math. Stat.*, 42:2050–2055, 1971.

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

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

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

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

[13] M. LeBlanc and J. Crowley. Survival trees by goodness of split. *J. Amer. Statist. Assoc.*, 88:457–467, 1993.

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

[15] D.G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, NY, 1969.

[16] G. Lugosi and A.B. Nobel. Consistency of data-driven histogram methods for density estimation and classification. 1996. To appear in Annals of Statistics.

[17] J. Makhoul, S. Roucos, and H. Gish. Vector quantization in speech coding. *Proc. IEEE*, 73:1551–1558, November 1985.

[18] J.N. Morgan and J.A. Sonquist. Problems in the analysis of survey data and a proposal. *J. Amer. Statist. Assoc.*, 58:415–435, 1963.

[19] S.K. Murthy, S. Kasif, and S. Salzberg. A system for induction of obliques decision trees. *J. Artificial Intelligence Research*, 2:1–32, 1994.

[20] A.B. Nobel. *On Uniform Laws of Averages*. PhD thesis, Stanford University, Information Systems Lab, 1992.

[21] A.B. Nobel. Histogram regression estimation using data-dependent partitions. 1996. To appear in Annals of Statistics.

[22] A.B. Nobel. Vanishing distortion and shrinking cells. *IEEE Trans. Inform. Theory*, 42:1303–1305, 1996.

[23] A.B. Nobel and R.A. Olshen. Termination and continuity of greedy growing for tree structured vector quantizers. *IEEE Trans. Inform. Theory*, 42:1–15, 1996.

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

[25] D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9:135–140, 1981.

[26] D. Pollard. A central limit theorem for k-means clustering. *Ann. Probab.*, 10:919–926, 1982.

[27] D. Pollard. Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, 28(2):199–205, 1982.

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

[29] R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *Ann. Math. Stat.*, 33:659–680, 1962.

[30] E.A. Riskin. *Variable Rate Vector Quantization of Images*. PhD thesis, Stanford University, Information Systems Lab, 1990.

[31] E.A. Riskin and R.M. Gray. A greedy growing algorithm for the design of variable rate vector quantizers. *IEEE Trans. Signal Proc.*, 39:2500–2507, 1991.

[32] M. Sabin and R.M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Trans. Inform. Theory*, 32, 1986.

[33] J. Sonquist. Multivariate model building: The validation of a search strategy. Technical report, Institute for Social Research, University of Michigan, Ann Arbor, 1970.