Quantization and the Method of k-Means

DAVID POLLARD

Abstract—**Asymptotic results from the statistical theory of k-means clustering are applied to problems of vector quantization. The behavior of quantizers constructed from long training sequences of data is analyzed by relating it to the consistency problem for k-means.**

I. Introduction

THE THEORY developed in the statistical literature for the method of k-means can be applied to the study of optimal k-level vector quantizers. In this paper, I describe some of this theory, including a consistency theorem (Section II) and a central limit theorem (Section IV) for k-means cluster centers. These results help to explain the behavior of optimal vector quantizers constructed from long stretches of ergodic training sequences. I also offer a new proof (Section III) for the consistency theorem, based on an identification of the optimal quantizer with the measure that minimizes a Vasershtein-like distance between an empirical measure and a collection of discrete measures corresponding to k-level quantizers.

By a k-level quantizer I mean a map q from some Euclidean space $\mathbb{R}^d$ into a subset $\{a_1, a_2, \cdots, a_k\}$ of itself. Such a map can be used to convert a d-dimensional input signal X into an output $q(X)$ that can take on at most k different values. An optimal k-level quantizer for a probability distribution P on $\mathbb{R}^d$ minimizes the distortion, as measured by the mean-squared error $\mathbb{P}\,|X - q(X)|^2$, for a random vector X with distribution P. (Instead of the traditional symbol $\mathbb{E}$, I use $\mathbb{P}$ to denote expectations as well as probabilities. A similar convention applies for expectations with respect to the probability measure P.) Of course, this makes sense only if the expected squared Euclidean distance $P\,|x|^2$ is finite, or, equivalently, the L^2 norm $\|X\| = (\mathbb{P}\,|X|^2)^{1/2}$ of the corresponding random vector X is finite. Such a constraint will remain on P throughout this paper.

Searching for an optimal k-level quantizer for P is equivalent to the k-means problem: find a set $A = \{a_1, a_2, \cdots, a_k\}$ of cluster centers to minimize the within cluster sum of squares

$$W(a_1, a_2, \cdots, a_k; P) = P\Big(\min_i |X - a_i|^2\Big).$$

The corresponding quantizer maps each x into its nearest center a_i. This formulation makes the dependence on P more explicit. Changing P would lead to a new optimal A; but, intuitively, changing P by only a small amount should not affect the locations of the optimal centers too greatly. This is the idea underlying the consistency theorem for k-means clustering.

Define the empirical measure P_n^ω by placing mass n^{-1} at each of the first n members of a long training sequence of observations on P. The superscript ω is to emphasize that P_n^ω is a random measure. The empirical measure can be used as an estimator of P; the optimal quantizer for P_n^ω provides a good approximation to the optimal quantizer for P.

1 Consistency Theorem: Let P_n^ω be the empirical measure constructed from a stationary, ergodic sequence of observations on a distribution P for which $\mathrm{P}\,|x|^2 < \infty$. Let $A_n^\omega = \{a_{n1}^\omega, \cdots, a_{nk}^\omega\}$ be any set of k cluster centers that minimizes $W(\cdot; P_n^\omega)$. If the set $A = \{a_1, \cdots, a_k\}$ that minimizes $W(\cdot, P)$ is unique (up to relabeling of its members) then, under an appropriate labeling of the members of A_n^ω, the sequence $\{a_{ni}^\omega\}$ converges to a_i for each i and almost every ω.

Notice the awkwardness caused by the need to match up the centers correctly. Especially in the multidimensional case, this can greatly complicate the notation and the flow of the proofs. The approach to be introduced in Section III avoids all these complications.

A full proof of the consistency theorem for a special case appears in Section II together with an outline of the extra complications encountered in extending to the general case. The development is based on [10], which generalized the one-dimensional results of [6]. I also discuss in that section the possibility of replacing minimum mean-squared error as the criterion of optimality.

Section III treats the consistency theorem from a different point of view. To each optimal quantizer q for a distribution P there corresponds a discrete probability measure Pq^{-1}, its image measure. I give a simple characterization for such image measures. This correspondence between optimal quantizers and discrete measures has nice continuity properties: if P' lies close to P, then the image measures are also close to one another. Applied to P and the empirical measure P_n^ω, this gives a result equivalent to the consistency theorem.

All this theory sidesteps one problem: results on the asymptotic behavior of the global minimum for $W(\cdot; P_n^\omega)$ do not apply directly to the practicable quantization algorithms, which search out local optima. Confirmation of the

Manuscript received August 3, 1981; revised November 3, 1981. This work was supported in part by the National Science Foundation, Grant MCS-8102725, and in part by the Air Force Office of Scientific Research, Grant F49620-79-C-0164.

The author is with the Statistics Department, Yale University, New Haven, CT 06520.

0018-9448/82/0300-0199$00.75 ©1982 IEEE

suspicion that all local minima get swept in towards a global minimum as n increases would therefore be of great theoretical comfort. This is a gap that needs to be filled.

II. Consistency for k-Means Clustering

For the proof of the consistency theorem, the 2-means problem for a distribution concentrated on a bounded interval of the real line presents the fewest difficulties. So, to begin with, consider sampling from the uniform distribution on $[0, 1]$. The problem is to show that the optimal pair of centers constructed from a stationary, ergodic sequence $\xi_1, \xi_2, \cdots$ of observations on that distribution converges to the pair of centers found by minimizing

$$W(a, b) = W(a, b; \text{uniform distribution}) = \int_0^1 \min\bigl(|x-a|^2, |x-b|^2\bigr)\, dx.$$

By symmetry I may assume that $a \le b$. Break the range of integration into the intervals $[0, 1/2(a+b))$ and $[1/2(a+b), 1]$, then integrate out the two quadratics

$$W(a, b) = a^3/3 + (1-b)^3/3 + (b-a)^3/12. \quad (1)$$

This function takes on its minimum value of $1/48$ at $a = 1/4, b = 3/4$. Moreover, given any $\epsilon > 0$ there exists a $\delta > 0$ (which you can actually calculate) such that $W(a, b) > \epsilon + 1/48$ whenever $0 \le a \le b \le 1$ and $\max(|a - 1/4|, |b - 3/4|) > \delta$. Put another way, a must be within a distance δ of $1/4$ and b within a distance δ of $3/4$ whenever

$$W(a, b) \le \epsilon + 1/48. \quad (2)$$

To prove that the optimal empirical cluster centers a_n^ω and b_n^ω lie within δ of $1/4$ and $3/4$ I have only to check that they satisfy this inequality.

The values of a_n^ω and b_n^ω are found by minimizing, subject to the constraint $0 \le a_n^\omega \le b_n^\omega \le 1$, the function

$$W_n(a, b) = W(a, b; P_n^\omega) = n^{-1} \sum_{i=1}^{n} \min\bigl(|\xi_i - a|^2, |\xi_i - b|^2\bigr). \quad (3)$$

By the ergodic theorem,

$$W_n(1/4, 3/4) \to \mathbb{P} \min\bigl(|\xi_i - 1/4|^2, |\xi_i - 3/4|^2\bigr) \quad \text{almost surely}$$
$$= W(1/4, 3/4)$$
$$= 1/48.$$

Thus,

$$\limsup W_n(a_n^\omega, b_n^\omega) \le 1/48 \quad \text{almost surely}, \quad (4)$$

because $W_n(a_n^\omega, b_n^\omega) \le W_n(1/4, 3/4)$. I shall show that W_n can be replaced by W in (4) without increasing the left-hand side. From there it will follow that, with probability one, $|a_n^\omega - 1/4| \le \delta$ and $|b_n^\omega - 3/4| \le \delta$, eventually. You can apply the same argument to a sequence of δ values converging to zero, discarding a null set of ω at each δ, to establish the desired convergence, $a_n^\omega \to 1/4$ and $b_n^\omega \to 3/4$, for almost all ω.

I shall justify substituting W for W_n in (4) by proving that

$$\sup_{a, b} |W_n(a, b) - W(a, b)| \to 0 \quad \text{almost surely}. \quad (5)$$

This will follow from a uniform continuity property of the functions

$$m(a, b; x) = \min\bigl(|x-a|^2, |x-b|^2\bigr).$$

By compactness of the region defined by $0 \le a \le b \le 1$, there exists, for each positive ϵ, a finite collection $(a_1, b_1), \cdots, (a_s, b_s)$ such that

$$\min_i \sup_x |m(a, b; x) - m(a_i, b_i; x)| < \epsilon,$$

for each (a, b) in that region; each $m(a, b; \cdot)$ is uniformly close to one of the functions $m(a_i, b_i; \cdot)$. Integrate out with respect to the uniform distribution, and then with respect to the empirical measure to deduce that

$$|W(a, b) - W(a_i, b_i)| < \epsilon$$

and

$$|W_n(a, b) - W_n(a_i, b_i)| < \epsilon,$$

where i (the same i in both cases) depends on (a, b). Use this to bound the left-hand side of (5) by

$$2\epsilon + \max_i |W_n(a_i, b_i) - W(a_i, b_i)|.$$

Apply the ergodic theorem s times, once for each pair (a_i, b_i), to show that the maximum here converges almost surely to zero.

Convergence results like (5) are sometimes called uniform strong laws of large numbers, or generalized Glivenko–Cantelli theorems. For a survey of such results see [4] or [3]. A more powerful combinatorial method that applies only to independent sequences of observations was described in [11]. This method gives rates of convergence.

With some little ingenuity the proof of the special case just treated can be expanded to cover the general consistency theorem, as in [10]. First consider the effect of replacing the uniform distribution by any P with compact support. The function $W(a, b)$ need no longer take an explicit form, as in (1), but that affects the proof only superficially. As long as $W(a, b)$—or perhaps I should now write it as $W(a, b; P)$—has a unique minimum, the argument based on (2) still works. Of course you might have a little more trouble in finding δ explicitly.

When solving for k-means instead of 2-means, you will need to replace functions of two variables, such as $W(a, b; P)$, by functions of k variables. The compactness argument leading to (5) generalizes easily to higher dimensions.

For a general P on the real line, not necessarily with compact support but satisfying the conditions of the consistency theorem, a new complication enters. The uniform convergence (5) only holds when the cluster centers range over a bounded region. A preliminary argument is needed to show that the optimal empirical cluster centers must all

eventually lie in some fixed compact region. The argument works by showing that any cluster center that lies too far from the origin can be discarded without too great an increase in the within cluster sum of squares. This would result in a $(k-1)$-level quantizer with distortion almost equal to that of the best k-level quantizer, which would contradict the uniqueness of the optimal k-level quantizer for the population distribution. For details I refer you to [10, p. 138] or Section V of this paper.

Generalizing to multidimensional distributions—to vector quantizers, that is—presents only notational difficulties. No longer do the cluster centers have a natural ordering imposed upon them. Without some restriction on the domain of $W(\cdot; P)$, the minimum would not be achieved at a unique k-tuple of centers; relabeling of coordinates would give a new k-tuple with the same value for $W(\cdot; P)$. Working with the cluster centers as a set of points in $\mathbb{R}^d$ calls for some fancy footwork to convert the preceding arguments into forms applicable to convergence of sets. That was the approach I adopted in [10]; but now I find the solution offered later in this paper more natural.

Finally, what can be done to replace minimum mean-squared error as the criterion of optimality? Consider, for example, a criterion expressible in this form: choose the set A of optimal cluster centers to minimize an expectation $H(A) = Ph(A; x)$. The optimal empirical cluster centers minimize $H_n(A) = P_n^\omega h(A; x)$. As long as $H(\cdot)$ were well defined and finite, the ergodic theorem could be applied to deduce that $|H_n(A) - H(A)| \to 0$ almost surely, for each fixed A. With sufficient smoothness properties for the functions $h(\cdot; \cdot)$, this might be extended to a uniform convergence result

$$\sup_A |H_n(A) - H(A)| \to 0 \quad \text{almost surely,}$$

at least if A were to range over a compact region. Provided the optimal centers could be forced into this compact region, and provided $H(\cdot)$ had a unique minimum at a set A in this region, the argument at the beginning of this section could be carried over. I would expect the first of these requirements to present the greatest difficulty. The references already cited for uniform strong laws of large numbers offer criteria powerful enough to handle the uniform convergence result. The distortion measures discussed in [7] would be good cases to start with.

III. Representing Quantizers by Measures

In Section I an optimal k-level quantizer for P was defined as a map q that takes at most k values and minimizes the L^2 distance $\|X - q(X)\|$ for every X with distribution P. Section II treated quantizers as sets of k points in $\mathbb{R}^d$, representing q by its range set. Now I turn to a third method for representing a quantizer, by considering the image measure Pq^{-1}. Questions of convergence of quantizers become questions of convergence of measures; problems of existence of quantizers become problems of minimizing a distance between measures. I defer proofs of results stated in this section to Section V.

The method for defining an optimal k-level quantizer for P is equivalent to a slightly more general procedure. Find a pair of random vectors X, with distribution P, and Y, taking on at at most k-distinct values, to minimize the L^2 distance $\|X - Y\|$. Up to almost sure equivalence, Y must be a function of X; this function will define an optimal quantizer (3 Theorem). This construction suggests a method for defining a metric on the class $\mathcal{P}$ of all probability measures on $\mathbb{R}^d$ with finite second moment.

2 Definition: The distance $\Delta(P, Q)$ between two measures in $\mathcal{P}$ equals the infimum of the L^2 distance $\|X - Y\|$ over all pairs of random vectors X, with distribution P, and Y, with distribution Q.

With the L^2 norm $\|X - Y\|$ replaced by the L^1 norm $\|X - Y\|_1$, this distance would become a special case of the Vasershtein distance [1] (also known as the $\bar{\rho}$ metric [5]). Because the proof that Δ defines a metric on $\mathcal{P}$ follows Dobrushin [1, theorem 2] so closely, I omit the details.

Write $\mathcal{P}_k$ for the class of all probability measures on $\mathbb{R}^d$ supported by at most k points, then define

$$m_k(P) = \inf\{\Delta(P, Q): Q \in \mathcal{P}_k\}.$$

Call a measure Q in $\mathcal{P}_k$ optimal for P if $\Delta(P, Q) = m_k(P)$. Optimal quantizers can be identified with optimal measures in $\mathcal{P}_k$.

3 Theorem: Each optimal k-level quantizer q for a measure P in $\mathcal{P}$ defines an optimal measure Q in $\mathcal{P}_k$ by $Q = Pq^{-1}$. Every optimal measure for P can be constructed in this way.

Suppose P has a unique optimal k-level quantizer q, corresponding to a measure Q in $\mathcal{P}_k$. Consider any other distribution P' that is close to P in the sense that $\Delta(P, P')$ is small. Then members of $\mathcal{P}_k$ that are optimal for P' must lie close to Q (9 Theorem). Put another way, if a sequence $\{P_n\}$ in $\mathcal{P}$ converges to P, that is $\Delta(P_n, P) \to 0$, then the measures corresponding to the optimal k-level quantizers also converge: $\Delta(Q_n, Q) \to 0$.

Think of this as a form of continuity for the map that sends distributions onto optimal quantizers. The convergence $\Delta(Q_n, Q) \to 0$ comes as close as possible to proving pointwise convergence of the corresponding quantizers $\{q_n\}$. After all, each q_n is defined only up to almost sure equivalence. It certainly does imply convergence of the quantization levels (6 Corollary) though.

Apply this argument with P_n equal to the empirical measure P_n^ω constructed from a stationary, ergodic sequence of observations on P. For almost every ω (7 Theorem), $\Delta(P_n^\omega, P) \to 0$, which implies that $\Delta(Q_n^\omega, Q) \to 0$. This provides another proof for the consistency theorem of Section I.

IV. Further Results on k-Means

The statement of the consistency theorem postulates the uniqueness of the optimal cluster centers for the underlying population distribution P. What happens when this condition is violated? One way to get nonuniqueness would be

for P to have fewer than k points in its support. In that case the whole theorem falls apart, because at least one cluster center is free to wander where it will. With the elimination of this possibility, however, something can be salvaged.

As long as the support of P contains at least k distinct points, Theorems 9 and 7 force the measure Q_n^ω, corresponding to an optimal k-level quantizer for the empirical measure P_n^ω, to converge almost surely towards the set $M(P)$ of optimal measures for P.

4 Example: Consider the optimal 2-level quantizers for a probability measure P that places mass $1/3$ at each of the vertices v_1, v_2, v_3 of an equilateral triangle. The optimal quantizer levels can be any of

a) $a_1 = (v_1 + v_2)/2, \quad a_2 = v_3$
b) $a_1 = (v_1 + v_3)/2, \quad a_2 = v_2$
c) $a_1 = (v_2 + v_3)/2, \quad a_2 = v_1.$

The empirical quantizer levels will approach this set of three possible population quantizer levels; the pair of empirical centers flips infinitely often between three possible configurations, moving closer to one of a), b), or c). The random measures $\{Q_n^\omega\}$ converge in distribution towards a uniform distribution over the three-point set $M(P)$. (To make sense of this statement, remember that Q_n^ω can be regarded as a random element of $\mathcal{P}_k$.) If you don't believe all this, you can prove it directly for yourself using elementary properties of the trinomial distribution. □

Any distribution P that is invariant under some group of symmetries on $\mathbb{R}^d$ will exhibit this sort of behavior. The compact set $M(P)$ of optimal measures inherits a group of symmetries; the random measures $\{Q_n^\omega\}$ converge in distribution to the invariant measure on $M(P)$. Consider the case of optimal 2-level quantizers for the spherical normal distribution if you want an example less trivial than the three-point distribution above.

What consequences do these results have for the construction of quantizers for distributions with symmetries? I would expect that the local optima found by the quantization algorithms would suffer from the same sort of rotational instability as the global minimum.

Whenever a consistency result has been proved, probabilists always start looking for a companion central limit theorem. For k-means, such a theorem holds if the population distribution has a smooth density, in addition to satisfying the conditions for the consistency theorem. For independent sampling, the one-dimensional case was solved in [6], the multidimensional case in [12].

Write $\boldsymbol{a}_n$ for the vector of optimal empirical cluster centers, and $\boldsymbol{a}$ for the optimal population centers, which are assumed to be uniquely determined. Then $n^{1/2}(\boldsymbol{a}_n - \boldsymbol{a})$ converges in distribution to a $N(0, V)$ distribution. The variance matrix V involves terms like the integrals of the population density over the faces of the optimal clusters. Let me sketch the method of proof for the 2-means problem considered at the beginning of Section II.

Take independent observations $\xi_1, \xi_2, \cdots$ on the uniform distribution on [0, 1]. Remember the notation

$$m(a, b; x) = \min(|x - a|^2, |x - b|^2).$$

From (3),

$$\begin{aligned} n^{1/2}(W_n(a, b) - W(a, b)) &= n^{-1/2} \sum_{i=1}^{n} m(a, b; \xi_i) - \mathbb{P}m(a, b; \xi_i) \\ &= X_n(a, b), \quad \text{say.} \end{aligned} \tag{6}$$

Concentrate on values of (a, b) close to the population optimal values $(1/4, 3/4)$, by writing

$$a = 1/4 + n^{-1/2}s, \qquad b = 3/4 + n^{-1/2}t.$$

A formal Taylor series expansion of m about $(1/4, 3/4)$ leads to

$$\begin{aligned} X_n(1/4 + n^{-1/2}s, 3/4 + n^{-1/2}t) &= X_n(1/4, 3/4) + 2n^{-1/2}sY_n + 2n^{-1/2}tZ_n \\ &\quad + \text{higher order terms,} \end{aligned} \tag{7}$$

where

$$Y_n = n^{-1/2} \sum_{i=1}^{n} (\xi_i - 1/4)\{\xi_i \le 1/2\} \tag{8}$$

and Z_n has a similar form. Read the factor $\{\xi_i \le 1/2\}$ in the summation here as an indicator function of the set where $\xi_i \le 1/2$. Combine (6) with (7) to get an approximation for W_n near the population optimum.

$$\begin{aligned} W_n(1/4 + n^{-1/2}s, 3/4 + n^{-1/2}t) &= W(1/4 + n^{-1/2}s, 3/4 + n^{-1/2}t) \\ &\quad + n^{-1/2}X_n(1/4, 3/4) \\ &\quad + 2n^{-1}(sY_n + tZ_n) + \text{higher order terms,} \\ &= 1/48 + n^{-1}(3s^2/8 - st/4 + 3t^2/8) \\ &\quad + n^{-1/2}X_n(1/4, 3/4) \\ &\quad + 2n^{-1}(sY_n + tZ_n) + \text{higher order terms,} \\ &= 1/48 + n^{-1/2}X_n(1/4, 3/4) \\ &\quad + n^{-1}(\text{quadratic in } s \text{ and } t) \\ &\quad + \text{higher order terms.} \end{aligned}$$

To accuracy of the order $n^{-1/2}$, the location of the minimum of W_n can be found by minimizing the quadratic term. This gives the optimal empirical centers

$$a_n^\omega = 1/4 + n^{-1/2}(\text{linear function of } Y_n \text{ and } Z_n) + \text{higher order terms,}$$

and

$$b_n^\omega = 3/4 + n^{-1/2}(\text{linear function of } Y_n \text{ and } Z_n) + \text{higher order terms.}$$

The linear functions of Y_n and Z_n here have an asymptotic joint normal distribution, because both Y_n and Z_n

have the form (8) of a normalized sum of independent random variables. This accounts for the asymptotic normality of the optimal empirical centers. See [12] for a more rigorous derivation of this result.

The argument becomes much more complicated when the distribution P does not have a unique optimal k-level quantizer. Hartigan [6] has conjectured the asymptotic distribution of the minimum value of W_n, the minimum distortion obtainable with a k-level quantizer, but has given no proof.

V. Proofs of the Results of Section III

The infimum in the definition of Δ is achieved. That is, for each P and Q in $\mathcal{P}$, there exist random vectors X with distribution P and Y with distribution Q such that $\Delta(P, Q) = \|X - Y\|$. Indeed, as shown by Shields in [9, appendix], the random vectors X and Y can be taken as the coordinate projections on $\mathbb{R}^d \times \mathbb{R}^d$: there exists a probability measure μ on $\mathbb{R}^{2d}$, with marginal measures $\mu X^{-1} = P$ and $\mu Y^{-1} = Q$, for which $\Delta(X, Y)^2 = \mu|X - Y|^2$. Put another way, there exists a family $\{Q(x, \cdot)\}$ of probability measures on $\mathbb{R}^d$ such that, for any X with distribution P, any random vector Y generated by using $Q(x, \cdot)$ for the conditional distribution of Y given $X = x$ will have distribution Q and satisfy $\|X - Y\| = \Delta(X, Y)$. This representation will be needed for the proof of 5 Theorem.

Proof of 3 Theorem: To each Q in $\mathcal{P}_k$ there is a k-level quantizer s such that

$$\Delta(P, Q) \geq \|X - s(X)\|,$$

for any X with distribution P. The construction is easy. Find random vectors such that $\Delta(P, Q) = \|X - Y\|$. Suppose Y takes the values $a_1, a_2, \cdots, a_k$. Define the clusters

$$C_i = \left\{x \in \mathbb{R}^d : |x - a_i| \leq |x - a_j| \text{ , for all } j\right\}.$$

It doesn't matter that these clusters overlap, although it does necessitate some nit picking in the definition of s. If x belongs to C_i and does not belong to C_j, for any j less than i, define $s(x)$ to equal a_i.

Suppose q is an optimal k-level quantizer for P. Then

$$\begin{aligned}\Delta(P, Q) = \|X - Y\| &\geq \|X - s(X)\| \\ &\geq \|X - q(X)\| \geq \Delta\left(P, Pq^{-1}\right) \geq m_k(P).\end{aligned}$$

Since these inequalities hold for every Q in $\mathcal{P}_k$, the measure Pq^{-1} must achieve the lower bound $m_k(P)$.

Conversely, suppose Q is optimal for P (that is, $\|X - Y\| = m_k(P)$). Then $\|X - Y\| = \|X - s(X)\|$. This implies that $Y = s(X)$ almost surely, except possibly for those sample points where $Y = a_i$ and X lands on the boundary between C_i and some C_j. By showing that this can occur only with probability zero, I shall prove that $s(X)$ has distribution Q, the desired result. (Compare the method of proof with Lloyd's [8] discussion of his Method I.)

First notice that a_i must equal the conditional mean $\mu_i = \mathbb{P}(X \mid Y = a_i)$, for otherwise $\|X - Y\|$ could be decreased by the amount $|a_i - \mu_i|^2 \mathbb{P}\{Y = a_i\}$ by shifting the center a_i to μ_i. (I assume here that $\mathbb{P}\{Y = a_i\} \neq 0$. The same argument would work if Y took on fewer than k-distinct values with positive probability.) This constraint forces X to place no probability on the cluster boundaries. For suppose there were a set B of positive probability for which $Y = a_i$ and X fell on the boundary common to C_i and C_j. Define Y^* to equal Y at sample points not in B, and to equal a_j on B. Because $\|X - Y^*\| = \|X - Y\|$, this Y^* would correspond to a Q^* optimal for P; but a_j would not equal the conditional expectation $\mathbb{P}(X \mid Y^* = a_j)$. □

Since the definition of the Δ metric involves an infimum over a large class of pairs of random vectors, it might seem that checking convergence in that metric would require messing about with many L^2 convergences. The next theorem gives a much cleaner characterization, relating Δ convergence to the well studied concept of weak convergence. This characterization simplifies the task of proving convergence of empirical measures.

5 Theorem: $\Delta(P_n, P) \to 0$, if and only if $\{P_n\}$ converges weakly to P and $P_n|x|^2 \to P|x|^2$.

Proof: Suppose $\Delta(P_n, P) \to 0$. Choose any X with distribution P then, using the conditional distributions described at the start of this section, construct random vectors $\{X_n\}$ with distributions $\{P_n\}$ such that $\|X_n - X\| = \Delta(X_n, X)$. Convergence in L^2 norm of $\{X_n\}$ to X implies weak convergence of $\{P_n\}$ to P and convergence of second moments.

Now to show that weak convergence plus convergence of second moments implies convergence in the Δ metric. By the Skorohod representation [2], there exist random vectors X_n (with distribution P_n) and X (with distribution P) such that X_n converges almost surely to X. From Fatou's lemma,

$$\liminf \mathbb{P}\left(2|X_n|^2 + 2|X|^2 - |X_n - X|^2\right) \geq \mathbb{P}4|X|^2.$$

From this and the convergence $\mathbb{P}|X_n|^2 \to \mathbb{P}|X|^2$ deduce that X_n converges to X in L^2 norm. □

6 Corollary: If a sequence $\{Q_n\}$ in $\mathcal{P}_k$ converges in the Δ metric to some Q in $\mathcal{P}_k$ then with a suitable labeling $\{a_{n1}, \cdots, a_{nk}\}$ for the support points of Q_n and $\{a_1, \cdots, a_k\}$ for the support points of Q,

$$a_{ni} \to a_i, \qquad \text{for each } i.$$

Proof: For any tiny δ set $f_i(x) = [1 - \delta^{-1}|x - a_i|]^+$. This function is bounded and continuous. In order that $Q_n(f_i) \to Q(f_i)$, the measure Q_n must eventually land one of its support points within a distance δ of a_i, for each i. □

5 Theorem corresponds to Dobrushin [1, theorem 2]. The construction he used in the second half of his proof parallels Dudley's [2] method for obtaining the Skorohod representation.

7 Theorem: Let P_n^ω be the empirical measure constructed from a stationary, ergodic sequence of observations on P.

For almost every ω,

$$\Delta(P_n^\omega, P) \to 0.$$

Proof: The idea behind this result goes back at least as far as [13]. Write C for the class of all bounded, continuous, real functions on $\mathbb{R}^d$. Let C_0 be a countable subclass such that every f in C can be obtained as a pointwise limit, $f_i \uparrow f$, of functions in C_0. For each fixed f_i, the ergodic theorem ensures

$$P_n^\omega(f_i) \to P(f_i) \quad \text{almost surely.} \tag{9}$$

Discard at most countably, many null sets of ω to be assured that the convergence in (9) holds with probability one for every f_i in C_0 simultaneously. For an ω at which this holds, $\liminf P_n^\omega(f) \geq \lim P_n^\omega(f_i) = P(f_i)$, for every i. Let i tend to infinity. Then $\liminf P_n^\omega(f) \geq P(f)$. Apply the same argument with $-f$ in place of f to get weak convergence of $\{P_n^\omega\}$ to P, for almost all ω. Cast out another null set on which $\{P_n^\omega |x|^2\}$ might not converge to $P|x|^2$ to be left with a set of ω on which $\Delta(P_n^\omega, P) \to 0$, by virtue of 5 Theorem. □

In [10], the hard step in proving the consistency theorem for k-means came with showing that all the optimal cluster centers eventually lie in some compact region of $\mathbb{R}^d$. The argument needed there appears in a disguised form in the next lemma. Before proving that lemma, I need to dispose of one trivial case that would otherwise continue to demand attention. Clearly it makes no sense to look for k-level quantizers for a distribution concentrating on fewer than k points. So from now on I shall insist that P does not degenerate in this way. Existence of a unique optimal k-level quantizer demands at least this much. Equivalently, I can require that $m_k(P) < m_{k-1}(P)$: if P does not concentrate on $k-1$ or fewer points then an optimal $(k-1)$-level quantizer can always be improved by adding in one more quantization level.

8 Lemma: Suppose the support of P contains at least k distinct points. Then if $m_k \leq \alpha < m_{k-1}(P)$, the set $K(\alpha) = \{Q \in \mathcal{P}_k: \Delta(P, Q) \leq \alpha\}$ is compact.

Proof: I shall show that, for a suitably large value of R, all the measures in $K(\alpha)$ have support in the closed ball $S(5R)$ of radius $5R$ centered at the origin of $\mathbb{R}^d$. Since a measure in $\mathcal{P}_k$ is specified by a set of k points in $\mathbb{R}^d$ and a set of k nonnegative weights summing to one, $K(\alpha)$ can then be expressed as a continuous image (by virtue of 6 Corollary) of a product of $S(5R)^k$ and a compact simplex in $\mathbb{R}^k$, which makes $K(\alpha)$ itself compact.

The R needs to satisfy two conditions. First choose r so that $PS(r) > 0$, then choose R to achieve

a) $(R-r)^2 PS(r) > \alpha$

b) $2(P|x|^2\{|x| > 2R\})^{1/2} < m_{k-1}(P) - \alpha$.

Given Q in $K(\alpha)$, choose Y with distribution Q and X with distribution P so that $\alpha \geq \Delta(P, Q) = \|X - Y\|$. The random vector Y must take at least one value within the closed ball $S(R)$, for otherwise

$$\|X - Y\|^2 \geq \mathbb{P}(R-r)^2\{|X| \leq r\},$$

contradicting a).

Suppose that Y takes values $y_1, y_2, \cdots, y_k$, where at least y_1 lies in $S(R)$. I shall show that $S(5R)$ contains every y_i. Suppose, to the contrary, that $|y_k| > 5R$, for example. Define a random vector Y^*, which takes on at most $k-1$ values, by setting it equal to Y, if $Y \neq y_k$, and equal to y_1, if $Y = y_k$. From the inequality

$$|X - y_1| \leq |X - y_k| + 2|X|\{|X| > 2R\}$$

deduce that

$$\begin{aligned}\|X - Y^*\| &\leq \|X - Y\| + 2\|X\{|X| > 2R\}\| \\ &< \alpha + (m_{k-1}(P) - \alpha).\end{aligned}$$

This contradicts the defining property of $m_{k-1}(P)$. (Compare this argument with Lloyd's [8] discussion of his Method I.) □

Incidentally, 8 Lemma proves that optimal quantizers do exist. The set

$$M(P) = \{Q \in \mathcal{P}_k: \Delta(P, Q) = m_k(P)\}$$

equals the intersection of the nonempty, compact sets $K(\alpha)$, for $m_k(P) < \alpha < m_{k-1}(P)$. It therefore cannot be empty.

For the P appearing in the statement of the continuity theorem, $M(P)$ contains only one element, by assumption. A form of the result can be proved without this assumption however. Recall that the distance between a measure Q^* and the set $M(P)$ is defined as $\Delta(Q^*, M(P)) = \inf\{\Delta(Q^*, Q): Q \in M(P)\}$.

9 Theorem: For $n = 1, 2, \cdots$, let Q_n in $\mathcal{P}_k$ be optimal for a probability measure P_n. Suppose $\Delta(P_n, P) \to 0$, where P has support containing at least k distinct points. Then $\Delta(Q_n, M(P)) \to 0$.

Proof: Define $N(\delta) = \{Q \in \mathcal{P}_k: \Delta(Q, M(P)) \geq \delta\}$. I shall prove that, for each $\delta > 0$,

$$\inf\{\Delta(P, Q): Q \in N(\delta)\} > m_k(P) \tag{10}$$

and

$$\Delta(P, Q_n) \to m_k(P). \tag{11}$$

Together these imply the stated result.

By definition, the continuous function $\Delta(P, \cdot)$ is strictly greater than $m_k(P)$ at all points of $\mathcal{P}_k$ not in $M(P)$. For any fixed α between $m_k(P)$ and $m_{k-1}(P)$, this continuous function must achieve its lower bound on the compact set $N(\delta)K(\alpha)$; everywhere outside $K(\alpha)$ it exceeds α. Combine these two bounds to get (10).

The triangle inequality ensures that $m_k(P)$ varies continuously with P. Thus,

$$\begin{aligned}m_k(P) &\leq \Delta(P, Q_n) \leq \Delta(P, P_n) + \Delta(P_n, Q_n) \\ &= \Delta(P, P_n) + m_k(P_n) \to m_k(P).\end{aligned}$$

□

REFERENCES

[1] R. L. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theory of Probability and its Applications*, vol. 15, 458–486, 1970.

[2] R. M. Dudley, "Distances of probability measures and random variables," *Annals of Mathematical Statistics*, vol. 39, 1563–1572, 1968.

[3] J. Elker, D. Pollard, and W. Stute, "Glivenko–Cantelli theorems for classes of convex sets," *Advances in Applied Probability*, vol. 11, 820–833, 1979.

[4] P. Gaenssler and W. Stute, "Empirical processes: A survey of results for independent and identically distributed random variables," *Annals of Probability*, vol. 7, 193–243, 1979.

[5] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\bar{d}$ distance with applications to information theory," *Annals of Probability*, vol. 3, 315–328, 1975.

[6] J. A. Hartigan, "Asymptotic distributions for clustering criteria," *Annals of Statistics*, vol. 6, 117–131, 1978.

[7] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 18, 84–95, 1980.

[8] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129–137, 1982, this issue, (Originally a 1957 Bell Labs memorandum).

[9] P. Papantoni-Kazakos and R. M. Gray, "Robustness of estimators on stationary observations," *Annals of Probability*, vol. 7, 989–1002, 1979.

[10] D. Pollard, "Strong consistency of k-means clustering," *Annals of Statistics*, vol. 9, 135–140, 1981.

[11] ——, "Limit theorems for empirical processes," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 57, 181–195, 1981.

[12] ——, "A central limit theorem for k-means clustering," *Annals of Probability*, to be published.

[13] V. S. Varadarajan, "On the convergence of probability distributions," *Sankhya*, vol. 19, 23–26, 1958.

Exponential Rate of Convergence for Lloyd's Method I

JOHN C. KIEFFER

***Abstract*—For a random variable with finite second moment and log-concave density, a unique quantizer exists which produces the minimum expected encoding error, using squared-error distortion. An algorithm given by Lloyd (Lloyd's Method I) yields a sequence of quantizers which converges to the optimum quantizer. Using results of Fleischer, it is shown that the convergence takes place exponentially fast if the logarithm of the density is not piecewise affine. As a consequence the number of iterations of Lloyd's algorithm needed to obtain the optimum distortion correct to n decimal places approaches infinity no faster than linearly in n. Another consequence is that if the output of a stationary information source at each time is distributed according to the given density, the source output can be encoded at each time i using the quantizer obtained on the ith iteration of Lloyd's method obtaining the same asymptotic behavior one would have obtained if the optimum quantizer had been used to encode at each time.**

I. INTRODUCTION

LET σ, τ be extended real numbers with $\sigma < \tau$. Let p: $(\sigma, \tau) \to (0, \infty)$ be a log-concave probability density with $\int_\sigma^\tau x^2 p(x)\,dx < \infty$. Let N be a fixed positive integer, greater than one. We call a map Q: $(\sigma, \tau) \to (-\infty, \infty)$ an N-level quantizer, if and only if there are real numbers $y_1, \cdots, y_N$ and real numbers $x_1 < \cdots < x_{N-1}$ in (σ, τ) such that

$$Q(y) = y_i, \qquad x_{i-1} \le y < x_i, \quad i = 1, \cdots, N,$$

where we take $x_0 = \sigma$, $x_N = \tau$. (We adopt this convention from now on; that is, if $u = (u_1, \cdots, u_{N-1})$ is a sequence of $(N-1)$ real numbers, it will be understood that in addition there is a u_0 defined to be σ and a u_N defined to be τ.)

It is well-known [2] [9] that if X is a (σ, τ)-valued random variable with density p, there is a unique N-level quantizer Q^* for which

$$E\{(X - Q^*(X))^2\} \le E\{(X - Q(X))^2\},$$

for every N-level quantizer Q. Thus, if one wishes to encode X with an N-level quantizer, the best encoder, in the sense of squared-error distortion, is Q^*.

Lloyd's Method I [6] is a way of finding Q^*. It involves applying a certain iterative procedure to an initial quantizer Q_0, obtaining a sequence of quantizers $Q_0, Q_1, Q_2, \cdots$, which are successive approximations to Q^*. In the limit, $Q_n \to Q^*$. We now describe this method.

Let R denote the real line and let O_N be the set of all $(x_1, \cdots, x_{N-1}) \in R^{N-1}$ such that $\sigma < x_1 < \cdots < x_{N-1} < \tau$.

Manuscript received March 10, 1981; revised October 6, 1981. This work was supported by the National Science Foundation under Grant ECS-7821335.

The author is with the Department of Mathematics and Statistics, University of Missouri—Rolla, MO 65401.

0018-9448/82/0300-0205$00.75 ©1982 IEEE