Rapid Clustering with a Deterministic Data Net *

Michelle Effros Leonard J. Schulman

Abstract

We consider the K-medians problem with the ℓ_2^2 distortion measure, also known as the problem of optimal fixed-rate vector quantizer design. We provide a deterministic approximation algorithm which works for all dimensions d and which, given a data set of size n, computes in time $f(K, \epsilon, d)n \log n$ (for some function f) a solution of cost at most $1 + \epsilon$ times optimal.

I Introduction

A striking lesson from the field of statistics is that important properties of a data set can be determined, with negligible error probability, by examining a very small random subset of the data. In the field of clustering algorithms, specifically, a key technique is to choose a random subset of the input points, cluster that set, and then extend the clustering to the entire original input. This approach has enabled algorithms whose complexity scales very slowly as a function of the size of the data set. (See discussion and references in [5].) Such scaling is important since very large data sets characterize many clustering applications. The sampling approach appears, however, to be critically dependent on a source of random bits.

In the present paper we show how to surmount this difficulty in the context of the K-medians clustering problem for ℓ_2^2 distortion, also known as the problem of optimal fixed-rate vector quantizer design. Our algorithm works in all dimensions d and finds a $(1 + \epsilon)$-optimal clustering in time quasilinear in n, the size of the data (or "training") set. Our central technique is the construction of a small, deterministic representation of the data set. The size of the representation is independent of the size of the input set (but not of d, ϵ, and K). Roughly speaking, this representation is analogous to the data sample constructed by a randomized algorithm. (The analogy is imperfect. On the one hand, our representation is even smaller, as a function of the number of input points, than the samples used in randomized methods. On the other hand, the representation is not simply handed off as a smaller data set to be clustered by a secondary algorithm; rather, it provides adequate representation of the geometric properties of the data set so that a very simple algorithm can then find a near-optimal clustering in quasilinear time.)

The ℓ_2^2 K-median problem is equivalently formulated as a problem in vector quantization. A quantizer of rate $\lg K$ is a data compression system representing each vector in $\mathbb{R}^d$ by one of K possible vectors, known as "codewords" or "reproduction values". In the rate- $\lg K$ quantization problem, we are given a pdf $p(\mathbf{x})$ on $\mathbb{R}^d$, an integer $K \geq 1$, and a distortion measure $\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$. The optimal vector quantizer is the set $\{\mu_1^\star, \ldots, \mu_K^\star\} \subset \mathbb{R}^d$ that achieves the minimal expected distortion $\int_{\mathbb{R}^d} p(\mathbf{x}) \min_{k \in \{1,\ldots,K\}} \rho(\mathbf{x}, \mu_k^\star) d\mathbf{x}$. While optimal scalar (one-dimensional) fixed- and variable-rate quantizer design for discrete distributions can be accomplished in polynomial time [4, 16, 17, 18, 13] optimal design even for fixed rate quantization in two dimensions is NP-hard [6]. We seek an ϵ-approximation algorithm for the fixed-rate case. Such an algorithm designs,

*California Institute of Technology, Pasadena, CA 91125. effros@caltech.edu, schulman@caltech.edu.

for any $\epsilon > 0$, a collection of K reproduction values yielding expected distortion within a factor of $(1+\epsilon)$ of the optimal expected distortion.

There has been a very large amount of work recently on various clustering tasks; this brief survey is restricted only to the most directly related work, pertaining to ℓ_2^2 K-median. Arora, Raghavan, and Rao [2] present a randomized algorithm giving a $(1+\epsilon)$-approximation to the planar ℓ_2 K-median problem in time $O(nKn^{O(1/\epsilon)}\log n)$. Kolliopoulos and Rao [12] extend that result, describing a randomized algorithm that gives a $(1+\epsilon)$-approximation to the d-dimensional ℓ_2 K-median problem in time $O(2^{O((1+\log(1/\epsilon)/\epsilon)^{d-1})}n\log n\log K)$. The problem considered in both of those papers differs from our problem both in their assumption that the distortion measure is a metric (we here focus primarily on the squared difference distortion measure) and in their requirement that the K medians (also known as reproduction values, codewords, centers, or centroids) sit at data points. Fernandez de la Vega, Karpinski, Kenyon, and Rabani [5] give a randomized approximation scheme for ℓ_2^2 K-median without restriction on codeword location (i.e., the same problem we are concerned with) running in time $O(g(K,\epsilon)n(\log n)^K)$ for some function g. (Similar results, though not for ℓ_2^2, appear in [3].)

Known deterministic algorithms for our problem are a factor of 2 approximation by Drineas, Frieze, Kannan, Vempala, and Vinay [6] running in time polynomial in n (and exponential in K and the dimension); a poly-time constant-factor approximation even for the case of arbitrary K by Jain and Vazirani [11]; and a poly-time approximation scheme by Ostrovsky and Rabani [14].

The contribution of this paper is a fundamentally new approach to clustering problems. The new approach is to to identify a *finite* set of points in space, which is an exhaustive list of candidates for the codewords. "Finite" means that the size of this set is a function of K, ϵ, and d but *not* of n, the size of the input set. "Exhaustive" means that only these points need to be considered as candidates for the locations of the codewords.

After such a set is identified, the subsequent computation is straightforward since all possible solutions can be tested and compared in time linear in n.

The principal technical work in the paper is devoted to the method for identifying the finite set of codeword-candidates (henceforth *net-points*), in time $f(K,\epsilon,d)n\log n$ for some function f.

We believe that the new "net-point" approach has the potential to yield deterministic solutions to a much wider variety of clustering criteria than the narrowly-defined case of ℓ_2^2 K-median, but we limit our claims in this manuscript to the case that has been proven. (The ℓ_2^2 version of K-median is widely motivated, in part because of applications involving mixtures of Gaussian sources, and in part due to metric embedding theorems [15].)

II Preliminaries

Given a distribution $p(\mathbf{x})$ on $\mathbb{R}^d$ and the squared error distortion measure $\rho(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||^2$, an optimal K-median clustering is K points $\mu_1^\star,\ldots,\mu_K^\star$ such that

$$\{\mu_1^\star,\ldots,\mu_K^\star\} = \arg\min_{\{\mu_1,\ldots,\mu_K\}} \int_{\mathbb{R}^d} p(\mathbf{x})\left[\min_{1\le k\le K}\rho(\mathbf{x},\mu_k)\right]d\mathbf{x}.$$

We call each cluster center μ_k a *codeword*, each collection of K codewords a *codebook*, and each solution $\{\mu_1^\star,\ldots,\mu_K^\star\}$ to the above minimization an optimal codebook. We use

$$\Delta = \int_{\mathbb{R}^d} p(\mathbf{x})\left[\min_{1\le k\le K}\rho(\mathbf{x},\mu_k^\star)\right]d\mathbf{x}$$

to denote the expected distortion of the optimal codebook. An optimal codebook for n points $\mathcal{T} = \{\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_n\}$ (known as a *data set* or *training set*) is simply an optimal codebook for the *empirical distribution* of $\mathcal{T}$ (which is the uniform distribution on the points of $\mathcal{T}$). The goal of an approximation algorithm is to find K points $\mu_1, ..., \mu_K$ such that

$$\int_{\mathbb{R}^d} p(\mathbf{x}) \left[\min_{1 \le k \le K} \rho(\mathbf{x}, \mu_k) \right] d\mathbf{x} \le (1+\epsilon)\Delta$$

The following notation is useful in what follows. For any $\mathcal{C} \subseteq \mathbb{R}^d$, define

$$\rho(\mathcal{C}, \mathbf{z}) = \int_{\mathcal{C}} p(\mathbf{x})\rho(\mathbf{x}, \mathbf{z}) d\mathbf{x} \qquad\qquad \mu(\mathcal{C}) = \arg\min_{\mathbf{z}} \rho(\mathcal{C}, \mathbf{z}).$$

III Deterministic Design of a Data Net

One ingredient of our algorithm is a routine performing the following task. Given a threshold θ, the routine partitions space into regions $\{A_m\}$ having two properties: the shape of each region is constrained to not be too "eccentric"; and the distortion of each region, $\rho(A_m, \mu(A_m))$, is no more than θ. We later bound the number of regions that are generated by our routine as a function of Δ and θ. With this routine in mind, here is the outline of our algorithm.

1. Loosely bound Δ to obtain an estimate $\hat{\Delta}_0 \in [\Delta, 4n\Delta]$. Set $\hat{\Delta} := \hat{\Delta}_0$.
2. Repeatedly halve $\hat{\Delta}$ and run the space-partitioning routine with threshold $\epsilon\hat{\Delta}/K$, until one of two stopping conditions is triggered: either $\hat{\Delta} < \hat{\Delta}_0/4n$, or the number of regions, M, created by the space-partitioning routine, is greater than a certain function of ϵ, K, and d.
3. Use the regions $\{A_1, ..., A_M\}$ to define a set of net-points, as described at the end of this section. The number of net-points is bounded by a function of ϵ, K, and d.
4. Exhaustively consider each set of K net-points as a codebook, and output the best of these.

The Initial Estimate of Δ

Given an arbitrary training set $\mathcal{T}$, we wish to bound the distortion Δ associated with the optimal d-dimensional quantizer for the empirical distribution of $\mathcal{T}$. Any linear-complexity, deterministic algorithm for bounding Δ to within a factor that grows at most polynomially with the training set size n is good enough for our purposes. Here is a simple approach based on the literature.

We approximate Δ by the square of the solution to the Euclidean K-center problem. In that problem, we are given a set D of n "demand points" in R^d, and our goal is to find a set S of K "supply points" so that the maximum Euclidean distance between a demand point and its nearest neighbor in S is minimized. Equivalently, we wish to find a smallest r and set S of cardinality K, so that D is contained in the balls of radius r about members of S. We denote this smallest r by $r^\star$. We bound Δ as $\frac{1}{n}(r^\star)^2 \le \Delta \le (r^\star)^2$. The lower bound follows since even in the optimal solution, at least one training vector must observe distortion greater than or equal to $(r^\star)^2$ relative to its closest codeword. The upper bound follows since using the K-center problem's optimal supply points as a codebook in the K-median problem gives expected distortion bounded above by $(r^\star)^2$.

In [8] and [10, 9], Gonzalez and, independently, Hochbaum and Shmoys give an $O(nK)$ time algorithm that computes an S satisfying $r \le 2r^\star$. Feder and Greene improve the run time to $O(n \log K)$ in [7] and show that beating factor 1.822 is NP-hard. A survey of these results and related work appears in [1]. Applying a 2-approximation for the K-center problem to give $r^\star \le r \le 2r^\star$ and setting $\hat{\Delta} = r^2$ gives $\Delta \le \hat{\Delta} \le 4n\Delta$.

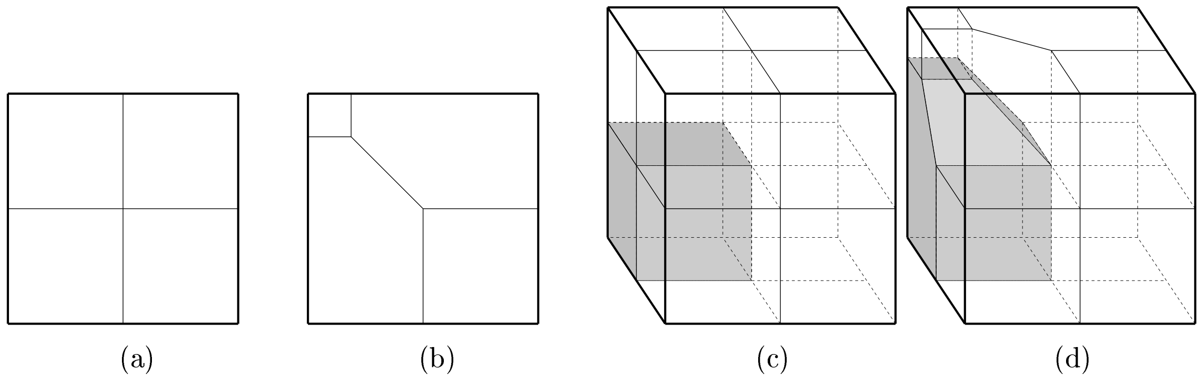

Figure 1: The region division and modification process in $\mathbb{R}^2$ and $\mathbb{R}^3$. Here (a) and (c) show an initial division, and (b) and (d) show the same division after modification. In (c) and (d), the bottom, rear subregion on the left is shaded to show its shape.

Designing Partition $\{A_1, \ldots, A_M\}$

We construct regions $A_1, \ldots, A_M$ such that $\{A_1, \ldots, A_M\}$ partitions $\mathbb{R}^d$ and $\rho(A_m, \mu(A_m)) \leq \epsilon\hat{\Delta}/K$ for all $m \in \{1, \ldots, M\}$. While our complexity results are expressed in terms of the cardinality n of a finite training set $\mathcal{T}$, we note that the basic strategy is applicable to continuous distributions. The procedure is iterative, constructing a series of increasingly fine partitions $\mathcal{A}_0, \mathcal{A}_1, \ldots$ culminating in the desired $\mathcal{A}_t = \{A_1, \ldots, A_M\}$. The initial conditions and procedure follow.

We begin with $\mathcal{A}_0 = \{a_1, a_2\}$ where a_2 is an axis-parallel d-dimensional hypercube centered at $\mu(\mathbb{R}^d)$, and $a_1 = a_2^c$ (the complement of a_2). We choose the bounding box separating a_1 from a_2 to be as small as possible subject to the constraint $\rho(a_1, \mu(\mathbb{R}^d)) \leq \epsilon\hat{\Delta}/K$. Given a distribution defined by a finite training set $\mathcal{T}$, the existence of a finite bounding box is immediate. When Δ is finite (the optimal clustering problem is trivial when Δ is infinite), the side-length for a_2 is bounded by the following argument. Since K is finite, $\rho(\mathbb{R}^d, \mu(\mathbb{R}^d)) < \infty$. If $p'(r)$ is the probability density function projected by p on the $(d-1)$-dimensional spheres of radius r, then from

$$\rho(\mathbb{R}^d, \mu(\mathbb{R}^d)) = \int_{\mathbb{R}^d} p(\mathbf{x})\rho(\mathbf{x}, \mu(\mathbb{R}^d))d\mathbf{x} = \int_0^\infty p'(r)||r - \mu(\mathbb{R}^d)||^2 dr$$

we see that $\int_{r_0}^\infty p'(r)||r - \mu(\mathbb{R}^d)||^2 dr$ tends to 0 as r_0 tends to ∞.

At iteration t, the procedure divides a region $a \in \mathcal{A}_{t-1}$ for which $\rho(a, \mu(a)) > \epsilon\hat{\Delta}/K$ into 2^d sub-regions $\{b_1, \ldots, b_{2^d}\}$. (If there is no such region a then the procedure halts and outputs the current partition of space.) The regions $\{b_1, \ldots, b_{2^d}\}$ are non-overlapping hypercubes created by splitting a in half along each dimension. (See Figure 1(a).) (We show soon that our procedure guarantees that each region a for which $\rho(a, \mu(a)) > \epsilon\hat{\Delta}/K$ is a hypercube of dimension d.) If $\rho(b_j, \mu(b_j)) < \epsilon\hat{\Delta}/K$ for all $j \in \{1, \ldots, 2^d\}$ or $\rho(b_j, \mu(b_j)) \geq \epsilon\hat{\Delta}/K$ for at least two $j \in \{1, \ldots, 2^d\}$, then $\{b_1, \ldots, b_{2^d}\}$ is the final partition of a. In this case, we form $\mathcal{A}_t$ from $\mathcal{A}_{t-1}$ by replacing a by the list of regions $b_1, \ldots, b_{2^d}$. The procedure continues to iteration $t+1$. On the other hand if $\rho(b_j, \mu(b_j)) \geq \epsilon\hat{\Delta}/K$ for exactly one $j \in \{1, \ldots, 2^d\}$, then we modify $\{b_1, \ldots, b_{2^d}\}$ as described next, and form $\mathcal{A}_t$ from $\mathcal{A}_{t-1}$ by replacing a by the list of modified regions $b'_1, \ldots, b'_{2^d}$.

Modification of $\{b_1, \ldots, b_{2^d}\}$: without loss of generality, let b_1 be the single "heavy" sub-region. We modify $\{b_1, \ldots, b_{2^d}\}$ by shrinking b_1 and growing its neighbors. The procedure shrinks hypercube b_1 away from the center of a to form a new, smaller hypercube b'_1; we simultaneously grow the d neighbors of b_1 so that the resulting $\{b'_1, \ldots, b'_{2^d}\}$ gives a new partition of a as shown in Figure 1(b). (The region acquired by a neighbor b is the convex span of the boundary between b and b_1, and the new face of b'_1 parallel to that boundary.) The shrinking procedure on b_1 stops when either $\rho(b'_j, \mu(b'_j)) < \epsilon\hat{\Delta}/K$ for all $j \in \{1, \ldots, 2^d\}$ or $\rho(b'_1, \mu(b'_1)) \geq \epsilon\hat{\Delta}/K$ and $\max_{j\in\{2,\ldots,2^d\}} \rho(b'_j, \mu(b'_j)) = \epsilon\hat{\Delta}/K$. In order to make $\rho(b'_j, \mu(b'_j))$ a continuous function of the moving boundary for all j, we allow the probability of a training vector that lies on the boundary of two or more regions to be divided arbitrarily among those regions.

As indicated previously, the iterative formation of $\mathcal{A}_0, \mathcal{A}_1, \ldots$ stops at the t for which the distortions of all regions are less than or equal to $\epsilon\hat{\Delta}/K$. The regions of $\mathcal{A}_t$ are then labeled $A_1, \ldots, A_M$.

Bounding M

Theorem 1 *There exists a constant c such that the number of regions created through the above design procedure is bounded as $M \leq 2^{d+2}K/\epsilon + 2^dK^2dc^d$.*

The proof of Theorem 1 is the topic of the rest of this section. To begin, arrange a_2 and its sub-regions into a 2^d-ary tree $\mathcal{T}$ with a_2 at its root. For any $a \in \mathcal{A}_{t-1}$, any sub-region $b \subset a$ such that $b \in \mathcal{A}_t$ is a child of a. The tree leaves represent partition $A_2, \ldots, A_M$ of a_2. (We denote the outer region by $A_1 = a_1$; A_1 is not included in the tree.) Let $\mathcal{T}'$ be the subtree of $\mathcal{T}$ such that if a is the region represented by some node in $\mathcal{T}'$, then $\rho(a, \mu(a)) \geq \epsilon\hat{\Delta}/K$. We bound M by bounding the number $M' \leq (M-1)/2^d$ of leaves in $\mathcal{T}'$.

We use two types of arguments to bound the number of leaves in $\mathcal{T}'$. We call the first approach "charging by volume" and the second "charging by distortion." To describe these techniques, let $B(\mathbf{x}, r)$ denote the closed ball of radius r about $\mathbf{x}$, $\partial B(\mathbf{x}, r)$ denote the boundary of $B(\mathbf{x}, r)$, $B^o(\mathbf{x}, r) = B(\mathbf{x}, r) - \partial B(\mathbf{x}, r)$ denote the corresponding open ball, and

$$B'(\mathbf{x}_1, \mathbf{x}_2) = B(\mathbf{x}_1, 10||\mathbf{x}_1 - \mathbf{x}_2||) \cup B(\mathbf{x}_2, 10||\mathbf{x}_1 - \mathbf{x}_2||)$$

describe a union of radius-$(10||\mathbf{x}_1 - \mathbf{x}_2||)$ balls around $\mathbf{x}_1$ and $\mathbf{x}_2$.

Let A' be the region of some leaf of $\mathcal{T}'$. To charge A' by volume, we prove that there exist optimal codewords $\mu_i^\star$ and $\mu_j^\star$ such that $\mathrm{vol}(A' \cap B'(\mu_i^\star, \mu_j^\star))/\mathrm{vol}(B'(\mu_i^\star, \mu_j^\star)) > V_o$ for some fixed constant V_o; that is, A' occupies a constant fraction of $B'(\mu_i^\star, \mu_j^\star)$. (Here $\mathrm{vol}(A) = \int_A d\mathbf{x}$ for any $A \subseteq \mathbb{R}^d$.) At most K^2/V_o regions can have this property. To charge A' by distortion, we demonstrate that

$$\int_{A'} p(\mathbf{x}) \left[\min_{1\leq k\leq K} \rho(\mathbf{x}, \mu_k^\star)\right] d\mathbf{x} \geq D_o,$$

for some fixed constant D_o; that is, we show that region A' contributes at least distortion D_o to the optimal performance. Since the optimal code achieves distortion Δ, at most Δ/D_o regions can satisfy this property.

We bound M' by showing that every region A' at a leaf of $\mathcal{T}'$ can be charged either by volume or by distortion. In particular, the arguments that follow use $D_o = \epsilon\hat{\Delta}/(4K)$ and $V_o = c^{-d}/d$ for some constant c. We therefore find that $M \leq 2^dM' \leq 2^d(4K/\epsilon + K^2dc^d)$.

Let A' be the region of some leaf of $\mathcal{T}'$. Use s to denote the "linear dimension" of region A', where the linear dimension of $A' \subseteq \mathbb{R}^d$ is the side length for the largest axis-parallel, d-dimensional hypercube that lies within A'. We use s as a simple approximation for the the diameter of A'. Lemma 1 shows their relationship more precisely. The construction of A' further implies that A' cannot be long and narrow. This observation, made more precise in Lemma 2, is important in cases where we wish to charge by area. (The form of the lower bound addressed in Lemma 2 is not crucial as long as it is a positive function of the dimension d.) Proofs of these lemmas are omitted:

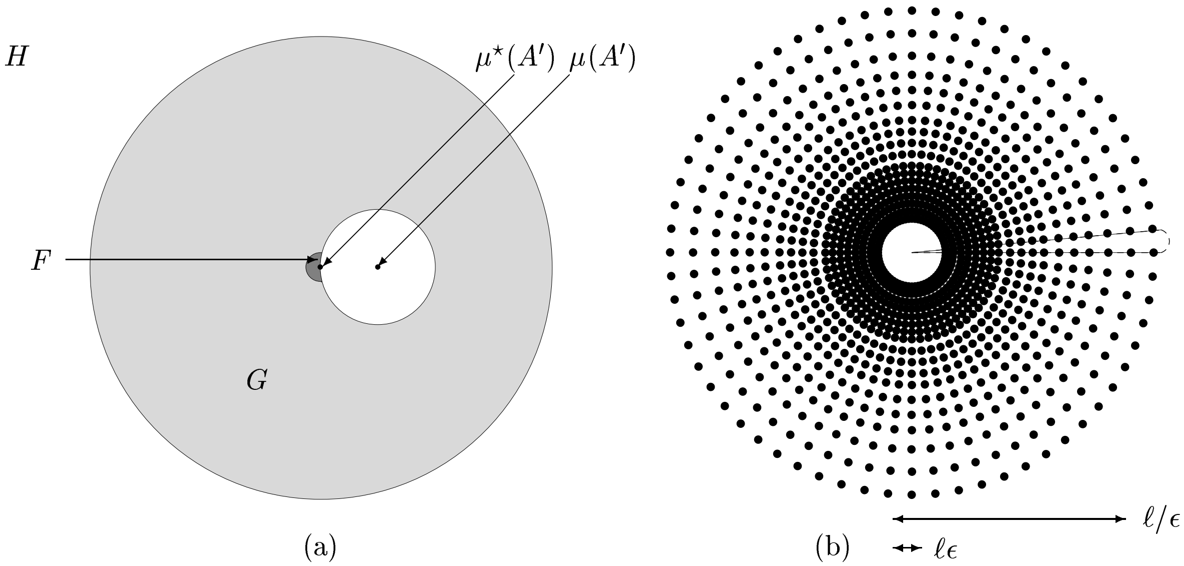

Figure 2: (a) The regions F, G, and H used in the counting argument to estimate M. (b) A sketch of the net-point construction for a single value of $(\mathbf{z}, \ell)$. In both (a) and (b), $d = 2$.

Lemma 1 $\mathrm{diam}(A') \le s\sqrt{d+3}$. □

Lemma 2 *There exists a constant c such that for any $\mathbf{x}_1, \mathbf{x}_2 \in A'$,* $\frac{\mathrm{vol}(A' \cap B(\mathbf{x}_1, ||\mathbf{x}_1 - \mathbf{x}_2||))}{\mathrm{vol}(B(\mathbf{x}_1, ||\mathbf{x}_1 - \mathbf{x}_2||))} \ge \frac{c^{-d}}{d}$. □

The argument that follows uses a careful case analysis to decide whether to charge a particular A' by volume or by distortion. That case analysis relies on the following definitions:

$$\begin{aligned}
\mu^\star(A') &= \arg\min_{1\le k\le K} ||\mu(A') - \mu_k^\star|| \\
\ell &= ||\mu(A') - \mu^\star(A')|| \\
F &= B\left(\mu^\star(A'), \frac{\ell}{4}\right) - B^o(\mu(A'), \ell) \\
G &= B(\mu^\star(A'), 4\ell) - \left(B\left(\mu^\star(A'), \frac{\ell}{4}\right) \cup B(\mu(A'), \ell)\right) \\
H &= \mathbb{R}^d - B(\mu^\star(A'), 4\ell).
\end{aligned}$$

Thus $\mu^\star(A')$ denotes the optimal codeword that is closest to centroid $\mu(A')$, ℓ denotes the distance of centroid $\mu(A')$ to its nearest optimal codeword, and F, G, and H are three non-overlapping regions in $\mathbb{R}^d$, as shown in Figure 2(a). Since F contains at least one optimal codeword, the following three values are well-defined.

$$\begin{aligned}
\mu_F^\star &= \arg\max_{\mu^\star \in \{\mu_1^\star, \ldots, \mu_K^\star\} \cap F} ||\mu^\star(A') - \mu^\star|| \\
\ell' &= ||\mu^\star(A') - \mu_F^\star|| \\
V &= \left\{\mathbf{x} \in \mathbb{R}^d : \min_{\mu^\star \in \{\mu_1^\star, \ldots, \mu_K^\star\} \cap F} \rho(\mathbf{x}, \mu^\star) = \min_{\mu^\star \in \{\mu_1^\star, \ldots, \mu_K^\star\}} \rho(\mathbf{x}, \mu^\star)\right\}.
\end{aligned}$$

Case 1. $s \le \ell/(2\sqrt{d+3})$**.** Charge A' by distortion. By Lemma 1 and the assumption, $\mathrm{diam}(A') \le s\sqrt{d+3} \le \ell/2$. So no $\mathbf{x} \in A'$ can be closer than $\ell - \ell/2$, to its closest optimal reproduction, giving

$$\int_{A'} p(\mathbf{x}) \left[\min_{1\le k\le K} \rho(\mathbf{x}, \mu_k^\star)\right] d\mathbf{x} \;\ge\; \int_{A'} p(\mathbf{x}) \left(\frac{\ell}{2}\right)^2 d\mathbf{x}$$

$$\geq \int_{A'} p(\mathbf{x}) \left(\mathrm{diam}(A')\right)^2 d\mathbf{x} \geq \rho(A', \mu(A')) \geq \epsilon\hat{\Delta}/K.$$

Case 2. $s > \ell/(2\sqrt{d+3})$ **and there is an optimal codeword** $\mu_j^\star$ **in** G**.** Charge A' by volume to the pair $\{\mu^\star(A'), \mu_j^\star\}$. We first show that $B(\mu(A'), \ell) \subseteq B'(\mu^\star(A'), \mu_j^\star)$ and that $B(\mu(A'), \ell)$ accounts for a constant fraction of the volume of $B'(\mu^\star(A'), \mu_j^\star)$. We then we show that A' accounts for a constant fraction of the volume of $B(\mu(A'), \ell)$.

By definition of G, $\ell/4 \leq ||\mu^\star(A') - \mu_j^\star|| \leq 4\ell$. Thus

$$\begin{aligned} B(\mu(A'), \ell) &\subset B(\mu^\star(A'), 10\ell/4) \subset B'(\mu^\star(A'), \mu_j^\star) \\ \frac{\mathrm{vol}(B(\mu(A'), \ell))}{\mathrm{vol}(B'(\mu^\star(A'), \mu_j^\star))} &\geq \frac{\ell^d}{2(10||\mu^\star(A') - \mu_j^\star||)^d} \geq \frac{\ell^d}{2(40\ell)^d} = \frac{1}{2(40^d)}. \end{aligned}$$

If A' lies entirely within $B(\mu(A'), \ell)$, then $s > \ell/(2\sqrt{d+3})$ and the shape of A' imply

$$\frac{\mathrm{vol}(A' \cap B(\mu(A'), \ell))}{\mathrm{vol}(B(\mu(A'), \ell))} \geq \frac{s^d/d}{\pi^{d/2}\ell^d/(d/2)!} > \frac{c_2^{-d}}{d}$$

for some constant c_2 independent of A'. (We here use the form of the volume equation that assumes d even. The equation is similar for d odd.) If A' extends outside of $B(\mu(A'), \ell)$, then there exists some point on $\mathbf{x} \in A'$ that lies on the outer boundary of $B(\mu(A'), \ell)$, and applying Lemma 2 with $\mathbf{x}_1 = \mu(A')$ and $||\mathbf{x}_1 - \mathbf{x}_2|| = \ell$ gives $\mathrm{vol}(A' \cap B(\mu(A'), \ell))/\mathrm{vol}(B(\mu(A'), \ell)) \geq c_1^{-d}/d$.

Case 3. $s > \ell/(2\sqrt{d+3})$**,** $\{\mu_1^\star, \ldots, \mu_K^\star\} \cap G = \phi$**, and** $A' \cap B(\mu^\star(A'), 3\ell') \neq \phi$**.** Charge A' by volume to $\{\mu^\star(A'), \mu_F^\star\}$. If $\ell' = 0$, then A' occupies the full volume of $B'(\mu^\star(A'), \mu_j^\star)$. Otherwise, apply Lemma 2 with $\mathbf{x}_2 = \mu(A')$ and $\mathbf{x}_1 \in B'(\mu^\star(A'), 3\ell') \cap A'$. Then $||\mathbf{x}_1 - \mathbf{x}_2|| \geq \ell - 3\ell' \geq \ell'$, and $\mathrm{vol}(A' \cap B(\mathbf{x}_2, \ell'))/\mathrm{vol}(B(\mathbf{x}_2, \ell')) \geq c_1^{-d}/d$. Since $B(\mathbf{x}_2, \ell')$ has radius 1/10 that of each of the balls in $B'(\mu^\star(A'), \mu_j^\star)$ and $B(\mathbf{x}_2, \ell') \subset B'(\mu^\star(A'), \mu_j^\star)$, we have the desired result.

Case 4. $s > \ell/(2\sqrt{d+3})$**,** $\{\mu_1^\star, \ldots, \mu_K^\star\} \cap G = \phi$**,** $A' \cap B(\mu^\star(A'), 3\ell') = \phi$**, and** $A' \subseteq V$**.** Charge A' by distortion. We achieve a bound as

$$\begin{aligned} \int_{A'} \left[\min_{1\leq k\leq K} \rho(\mathbf{x}, \mu_k^\star)\right] d\mathbf{x} &= \int_{A'} p(\mathbf{x}) \left[\min_{\mu^\star\in\{\mu_1^\star,\ldots,\mu_K^\star\}\cap F} \rho(\mathbf{x}, \mu^\star)\right] d\mathbf{x} \qquad (1) \\ &\geq \int_{A'} p(\mathbf{x}) \left[\frac{1}{4}\rho(\mathbf{x}, \mu^\star(A'))\right] d\mathbf{x} \geq \frac{1}{4}\rho(A', \mu(A')) \geq \frac{1}{4}\frac{\epsilon\hat{\Delta}}{K}. \qquad (2) \end{aligned}$$

Here (1) follows from $A' \subseteq V$. The definition of ℓ' implies that all optimal codewords in F lie in a ball of radius ℓ' around $\mu^\star(A')$; since A' does not intersect the ball of radius $3\ell'$ around $\mu^\star(A')$, $(\min_{\mu^\star\in\{\mu_1^\star,\ldots,\mu_K^\star\}\cap F} ||\mathbf{x} - \mu^\star||)/||\mathbf{x} - \mu^\star(A')|| \geq 1/2$, giving the first inequality. The last two inequalities follow by definition of $\mu(A')$ and design of A', respectively.

Case 5. $s > \ell/(2\sqrt{d+3})$**,** $\{\mu_1^\star, \ldots, \mu_K^\star\} \cap G = \phi$**,** $A' \cap B(\mu^\star(A'), 3\ell') = \phi$**, and** $A' \not\subseteq V$**.** Here $A' \cap V \neq \phi$ by definition of $\mu^\star(A')$, and $A' \cap V^c \neq \phi$ by assumption. Thus there exist codewords $\mu_i^\star \in \{\mu_1^\star, \ldots, \mu_K^\star\} \cap F$ and $\mu_j^\star \in \{\mu_1^\star, \ldots, \mu_K^\star\} \cap H$ such that the intersection between the Voronoi regions for $\mu_i^\star$ and $\mu_j^\star$ runs through A'. Charge A' by volume to $\{\mu_i^\star, \mu_j^\star\}$. Let $\ell'' = ||\mu_i^\star - \mu_j^\star||$. By definition of F and H, $\ell'' \geq 15\ell/4$. Since the boundary between the Voronoi cells for $\mu_i^\star$ and $\mu_j^\star$ runs through A', $\max_{\mathbf{x}\in A'} ||\mu_i^\star - \mathbf{x}|| \geq \ell''/2$, giving $\max_{\mathbf{x}\in A'} ||\mu(A') - \mathbf{x}|| \geq \ell''/2 - 5\ell/4 \geq \ell''/6$. Applying Lemma 2 with $\mathbf{x}_1 = \mu(A')$ and $||\mathbf{x}_1 - \mathbf{x}_2|| \geq \ell''/6$ tells us that A' occupies some fraction c^{-d}/d of the ball of radius $\ell'/6$ around $\mu(A')$. Since that ball falls entirely within the ball of radius $10\ell'$ around $\mu^\star(A')$ and occupies a constant fraction of its volume, we have the desired result.

Building the Data Net $P \cup Z$

We use the above region construction to lay down our data net. Let $Z = \{\mu(A_m)\}_{m=1}^{M}$ and $L = \{||\mathbf{z} - \mathbf{z}'|| : \mathbf{z}, \mathbf{z}' \in Z, \mathbf{z} \neq \mathbf{z}'\}$; we assume that $\mu(A_i) \neq \mu(A_j)$ for all $i \neq j$. For each $(z, \ell) \in Z \times L$, we create a "cloud" of net-points around z; the inner radius of the cloud is $\ell\epsilon$, the outer radius of the cloud is ℓ/ϵ, all net-points lie along lines of polar spacing ϵ, and the radial spacing is adjusted to match the polar spacing locally. See Figure 2(b). We use P to denote the complete set of net-points and use $P \cup Z$ as our deterministic data net. We bound $|Z \cup P|$ as

$$N = |Z| + |P| \leq M + \frac{c^d M^2}{\epsilon^{d+1}} \leq \left(\frac{2^{d+2}K}{\epsilon} + 2^d K^2 dc^d \right) + \frac{c^d}{\epsilon^{d+1}} \left(\frac{2^{d+2}K}{\epsilon} + 2^d K^2 dc^d \right)^2 .$$

IV The Deterministic Data Net is Good

To show that the data net $P \cup Z$ is good, we bound the performance of a quantized version $\{Q(\mu_1^\star), \ldots, Q(\mu_K^\star)\}$ of optimal codebook $\{\mu_1^\star, \ldots, \mu_K^\star\}$. Here $Q : \mathbb{R}^d \to P \cup Z$ is defined as

$$Q(\mathbf{x}) = \arg \min_{\mathbf{q} \in P \cup Z} \left[||\mathbf{x} - \mathbf{q}|| + \pi(\mathbf{q}) \right],$$

where penalty function $\pi(\mathbf{q}) = 0$ when $\mathbf{q} \in Z$ and $\pi(\mathbf{q}) = \epsilon\ell$ for each $\mathbf{q} \in P$ created in a cloud with parameter ℓ. While we don't know $\{\mu_1^\star, \ldots, \mu_K^\star\}$ and can't find $\{Q(\mu_1^\star), \ldots, Q(\mu_K^\star)\}$, we can find the optimal codebook with codewords in $P \cup Z$, which is at least as good.

Since not knowing $\{\mu_1^\star, \ldots, \mu_K^\star\}$ implies that we cannot directly address the question of how well $Q(\mu_i^\star)$ approximates $\mu_i^\star$ for each i, we instead visit each $\mathbf{q} \in P \cup Z$ and show how well $\mathbf{q}$ would serve as an approximation for any $\mu^\star$ that might map to it. For any $\mathbf{q} \in P \cup Z$, let $V(\mathbf{q}) = \{\mathbf{x} \in \mathbb{R}^d : Q(\mathbf{x}) = \mathbf{q}\}$. Break $\mathbb{R}^d$ into regions near to and far from $\mathbf{q}$, here given by $\mathrm{Near}(\mathbf{q})$ and $\mathrm{Near}(\mathbf{q})^c$. For $\mathbf{x} \in \mathrm{Near}(\mathbf{q})$, the cost of reproducing $\mathbf{x}$ by $\mathbf{q}$ rather than optimal codeword $\mu^\star \in V(\mathbf{q})$ increases the expected distortion by at most a small additive constant. For $\mathbf{x} \in \mathrm{Near}(\mathbf{q})^c$, the cost of mapping $\mathbf{x}$ to $\mathbf{q}$ rather than its optimal codeword $\mu^\star \in V(\mathbf{q})$ modifies the expected distortion by at most a small multiplicative constant. If $\{\mathcal{C}_1^\star, \ldots, \mathcal{C}_K^\star\}$ denotes the Voronoi cells for optimal codewords $\{\mu_1^\star, \ldots, \mu_K^\star\}$, then these results together imply $\sum_{i=1}^{K} \rho(\mathcal{C}_i^\star, Q(\mu_i^\star)) \leq c^d \epsilon \hat{\Delta} + \Delta/(1-\epsilon)$.

For any $\mathbf{q} \in Z \cup P$, let $\mathbf{z}_1$ be a closest element of $Z - \{\mathbf{q}\}$ and let $\ell_1 = ||\mathbf{q} - \mathbf{z}_1||$. Define $\mathbf{z}_2$ as a closest element of $Z \cap B(\mathbf{z}_1, 2\varepsilon\ell_1)^c$ if that set is not empty and write $\mathbf{z}_2 = \phi$ otherwise; let $\ell_2 = ||\mathbf{z}_1 - \mathbf{z}_2||$ if $\mathbf{z}_2 \neq \phi$. Let H be the half-space of points closer to $\mathbf{q}$ than to $\mathbf{z}_1$. Define

$$\mathrm{Near}(\mathbf{q}) = \begin{cases} B(\mathbf{q}, \ell_1/4) & \text{if } \mathbf{q} \in Z \\ H & \text{if } \mathbf{q} \in P \text{ and } \mathbf{z}_2 = \phi \\ H \cap B(\mathbf{q}, \max\{\ell_1, \ell_2\}/4) & \text{if } \mathbf{q} \in P \text{ and } \mathbf{z}_2 \neq \phi. \end{cases}$$

Theorem 1 shows that for the part of $\mathcal{C}_i^\star$ that is close to $\mathbf{q}$, the cost of representing $\mu_i^\star \in V(\mathbf{q})$ by $\mathbf{q}$ is bounded by a small additive constant.

Theorem 1 (Additive analysis) *There exists a constant c such that if $\mu_i^\star \in V(\mathbf{q})$, then $\rho(\mathcal{C}_i^\star \cap \mathrm{Near}(\mathbf{q}), \mathbf{q}) \leq c^d \epsilon \hat{\Delta}$.*

Proof: We show that for every $\mathbf{q} \in Z \cup P$, $\mathrm{Near}(\mathbf{q})$ intersects $M' \leq c^d$ of the regions in $\{A_1, \ldots, A_M\}$. Let $\{A_1, \ldots, A_{M'}\}$ be those intersecting regions. The definition of $\mathrm{Near}(\mathbf{q})$ gives $||\mathbf{x} - \mathbf{q}|| \leq (1 + \epsilon)||\mathbf{x} - \mu(A_m)||$ for all $m \in \{1, \ldots, M'\}$ and $\mathbf{x} \in \mathrm{Near}(\mathbf{q}) \cap A_m$, and thus

$$\rho(\mathrm{Near}(\mathbf{q}), \mathbf{q}) \leq \sum_{m=1}^{M'} (1+\epsilon)^2 \rho(\mathrm{Near}(\mathbf{q}) \cap A_m, \mu(A_m)) \leq c^d \epsilon \hat{\Delta}/K.$$

When $\mathbf{q} \in Z$, each A_m that intersects $\mathrm{Near}(\mathbf{q})$ must cross the annulus between $B(\mathbf{q}, \ell_1/4)$ and $B(\mathbf{q}, \ell_1)$. Thus by the shape of A_m, A_m has linear dimension at least $3\ell_1/(4\sqrt{d+3})$ and occupies an angular section of fraction at least c^{-d} of $\partial B(\mathbf{z}, 5\ell_1/8)$.

When $\mathbf{q} \in P$ and $\mathbf{z}_2 = \phi$, $Z \subseteq B(\mathbf{z}_1, 2\varepsilon\ell_1)$. Any A_m intersecting $\mathrm{Near}(\mathbf{q})$ has linear dimension $\geq (1/2 - 2\varepsilon)\ell_1/\sqrt{d+3}$ and occupies an angular section of fraction at least c^{-d} of $\partial B(\mathbf{z}_1, \ell_1/4)$.

When $\mathbf{q} \in P$ and $\mathbf{z}_2 \neq \phi$, let $Z_1 = Z \cap B(\mathbf{z}_1, 2\varepsilon\ell_1)$ and let $Z_2 = Z - Z_1$. For any A_m such that $\mu(A_m) \in Z_1$, the previous argument follows. For any A_m such that $\mu(A_m) \in Z_2$ and $A_m \cap \mathrm{Near}(p) \neq \phi$, A_m intersects $B(\mathbf{q}, \max\{\ell_1, \ell_2\}/4)$ and $||\mu(A_m) - \mathbf{q}|| \geq \max\{\ell_1, \ell_2 - \ell_1\} \geq \max\{\ell_1, \ell_2/2\} \geq \max\{\ell_1, \ell_2\}/2$. Consequently, the linear dimension of A_m is at least $\max\{\ell_1, \ell_2\}/(2\sqrt{d+3})$, and A_m covers a constant fraction of $\partial B(\mathbf{q}, 3\max\{\ell_1, \ell_2\}/8)$. □

Theorem 4 proves that if $\mu_i^\star \in V(\mathbf{q})$, then $||\mathbf{x} - \mathbf{q}|| \leq ||\mathbf{x} - \mu^\star||/(1 - 4\epsilon)$ for all $x \notin \mathrm{Near}(\mathbf{q})$, giving $\rho(\mathcal{C}_i^\star \cap \mathrm{Near}(\mathbf{q})^c, \mathbf{q}) \leq \rho(\mathcal{C}_i^\star \cap \mathrm{Near}(\mathbf{q})^c, \mu^\star)/(1 - \epsilon)$. First we show Lemmas 2 and 3.

Lemma 2 *If $V(\mathbf{q}) \neq \emptyset$, then $\mathbf{q} \in V(\mathbf{q})$.*

Proof: Suppose $\mathbf{x} \in V(\mathbf{q})$. Then $||\mathbf{x} - \mathbf{q}|| + \pi(\mathbf{q}) \leq ||\mathbf{x} - \mathbf{z}||$ and $||\mathbf{x} - \mathbf{q}|| + \pi(\mathbf{q}) \leq ||\mathbf{x} - \mathbf{p}|| + \pi(\mathbf{p})$ for all $\mathbf{z} \in Z$ and $\mathbf{p} \in P$. By the triangle inequality $||\mathbf{x} - \mathbf{q}'|| - ||\mathbf{x} - \mathbf{q}|| \leq ||\mathbf{q} - \mathbf{q}'||$ for all $\mathbf{q}'$, so $\pi(\mathbf{q}) \leq ||\mathbf{q} - \mathbf{z}||$ and $\pi(\mathbf{q}) \leq ||\mathbf{q} - \mathbf{p}|| + \pi(\mathbf{p})$ for all $\mathbf{z} \in Z$ and $\mathbf{p} \in P$. □

Lemma 3 *If $\mathbf{z} \in Z$ generates a cloud of parameter ℓ, then for any $\mathbf{q} \in B(\mathbf{z}, \ell(1-\varepsilon)/\varepsilon) \cap P$, $\mathrm{diam}(V(\mathbf{q})) = O(\varepsilon \max\{\ell, ||\mathbf{q} - \mathbf{z}||\})$.*

Proof: Any $\mathbf{q} \in B(\mathbf{z}, (1+\epsilon)\epsilon\ell)$ lies between $\mathbf{x}$ and a radius-$\epsilon\ell$ sphere of net-points. Any other $\mathbf{q} \in B(\mathbf{z}, (1-\epsilon)(\ell/\epsilon))$ lies among net-points spaced at order $\epsilon||\mathbf{q} - \mathbf{z}||$ distances. □

Theorem 4 (Multiplicative analysis) *There exists a constant c such that for each $\mathbf{q} \in P \cup Z$ and $\mathbf{x} \notin \mathrm{Near}(\mathbf{q})$, $B(\mathbf{x}, (1 - c\varepsilon)||\mathbf{x} - \mathbf{q}||) \cap V(\mathbf{q}) = \emptyset$.*

Proof: If $\mathbf{q} \in Z$, then $V(\mathbf{q}) \subseteq B(\mathbf{q}, \varepsilon\ell_1)$. So if $\mathbf{x} \notin \mathrm{Near}(\mathbf{q}) = B(\mathbf{q}, \ell_1/4)$, then $||\mathbf{x} - \mathbf{q}|| - \epsilon\ell_1 \geq ||\mathbf{x} - \mathbf{q}|| - \epsilon(4||\mathbf{x} - \mathbf{q}||) = (1 - 4\epsilon)||\mathbf{x} - \mathbf{q}||$, and $V(\mathbf{q})$ is disjoint from $B(\mathbf{x}, (1 - 4\varepsilon)||\mathbf{x} - \mathbf{q}||)$.

If $\mathbf{q} \in P$ and $\mathbf{z}_2 = \phi$, then $\mathbf{q}$ is created in a cloud about some $\mathbf{z} \in B(\mathbf{z}_1, 2\varepsilon\ell_1)$. Suppose that $V(\mathbf{q}) \subseteq L = \{\mathbf{y} \in \mathbb{R}^d : ||\mathbf{y} - \mathbf{z}|| \geq ||\mathbf{y} - \mathbf{q}|| + \ell_1(1 - \varepsilon)\}$ (a set with a hyperbolic boundary). We get the desired result by a geometric argument that shows that for $\mathbf{x} \notin H$, $B(\mathbf{x}, (1 - 4\varepsilon)||\mathbf{x} - \mathbf{q}||)$ does not intersect L. Suppose instead that $V(\mathbf{q})$ extends outside L; then the following argument gives a contradiction. If $V(\mathbf{q}) \cap L^c \neq \phi$, then $\pi(\mathbf{q}) \leq \epsilon\ell_1$, point $\mathbf{q}$ is generated in a cloud of parameter $\ell \leq \ell_1$ around $\mathbf{z}$, and $\mathbf{q}$ is not on the inner sphere of the cloud in which it is generated. In this case, the cloud generating $\mathbf{q}$ has points on the sphere of radius $||\mathbf{z} - \mathbf{q}||$ around $\mathbf{z}$ and also on the sphere of radius $(1 - \epsilon)||\mathbf{z} - \mathbf{q}||$ around $\mathbf{z}$. These points have the same penalty as $\mathbf{q}$ and are spaced regularly enough so $V(\mathbf{q})$ is contained in a wedge-shaped region entirely contained within L.

If $\mathbf{q} \in P$, $\mathbf{z}_2 \neq \phi$, and $\mathbf{q}$ is created by some $\mathbf{z} \in B(\mathbf{z}_1, 2\varepsilon\ell_1)$, then the reasoning from the previous case implies $V(\mathbf{q}) \subseteq L$. That handles the multiplicative distortion for $\mathbf{x} \notin H$; it remains to handle the multiplicative distortion for $\mathbf{x} \notin B(\mathbf{q}, \max\{\ell_1, \ell_2\}/4)$. Observe that $\mathbf{q} \in B(\mathbf{z}_1, \ell_2/(2\varepsilon))$; therefore, $\mathbf{q}$ is internal to the cloud of parameter ℓ_2 about $\mathbf{z}_1$. So by Lemma 3, $V(\mathbf{q})$ has diameter $O(\varepsilon \max\{\ell_1, \ell_2\})$ and so $B(\mathbf{x}, (1 - 4\varepsilon)||\mathbf{x} - \mathbf{q}||) \cap B(\mathbf{q}, \max\{\ell_1, \ell_2\}/4)$.

If $\mathbf{q} \in P$, $\mathbf{z}_2 \neq \phi$, $\mathbf{q}$ is created by some $\mathbf{z} \notin B(\mathbf{z}_1, 2\varepsilon\ell_1)$, and $\ell_2 \leq 4\ell_1$, then the existence of $\mathbf{z}_2$ again implies that $\mathbf{q} \in B(\mathbf{z}_1, \ell_2/(2\varepsilon))$ and is internal to the cloud of parameter ℓ_2 about $\mathbf{z}_1$; thus by the previous argument, $V(\mathbf{q})$ has diameter $O(\varepsilon \max\{\ell_1, \ell_2\}) = O(\varepsilon\ell_1)$. If $\mathbf{x} \notin H \cap B(\mathbf{q}, \max\{\ell_1, \ell_2\}/4)$, then in particular $||x - q|| \geq \ell_1/4$. So for a suitable c, dictated by the bound on $\mathrm{diam}(V(q))$, $B(\mathbf{x}, (1 - c\varepsilon)||\mathbf{x} - \mathbf{q}||) \cap V(\mathbf{q}) = \emptyset$.

If $\mathbf{q} \in P$, $\mathbf{z}_2 \neq \phi$, and $\mathbf{q}$ is created by some $\mathbf{z} \notin B(\mathbf{z}_1, 2\varepsilon\ell_1)$, and $\ell_2 > 4\ell_1$ (so $\mathbf{z} \notin B(\mathbf{z}_1, 4\ell_1)$), then we want to show that $\mathrm{diam}(V(\mathbf{q})) = O(\varepsilon\ell_1)$. This suffices since (always) $B(\mathbf{q}, \ell_1/4) \subseteq \mathrm{Near}(\mathbf{q})$. Let ℓ be the parameter of the cloud that generated $\mathbf{q}$. Note $\ell/\varepsilon \geq \ell_2 - \ell_1 \geq 3\ell_2/4$, so $\ell \geq 3\varepsilon\ell_2/4$. If $\ell > \ell_1/\varepsilon$ then $\pi(\mathbf{q}) > \ell_1$ so by lemma 2, $V(\mathbf{q})$ is empty and we're done. It remains to consider the case $3\varepsilon\ell_2/4 \leq \ell \leq \ell_1/\varepsilon$. There is a cloud about $\mathbf{z}_1$ of parameter ℓ, hence of inradius at most ℓ_1 and outradius at least $3\ell_2/4$ which in turn is at least $3\ell_1$. This cloud's penalty is $\varepsilon\ell$, equal to the penalty of $\mathbf{q}$. If $\ell \leq \ell_1(1-\varepsilon)/\varepsilon$ then the inradius is $\leq \ell_1(1-\varepsilon)$, so $\mathbf{q}$ is surrounded in all directions, at a spacing of $\varepsilon\ell_1$, by points of penalty equal to its own; hence $\mathrm{diam}(V(\mathbf{q})) = O(\varepsilon\ell_1)$. On the other hand if $\ell_1(1-\varepsilon)/\varepsilon < \ell \leq \ell_1/\varepsilon$ then $\pi(\mathbf{q}) > \ell_1(1-\varepsilon)$ so $V(\mathbf{q})$ cannot extend past the hyperplane perpendicular to $\overline{\mathbf{z}_1\mathbf{q}}$, at distance $\varepsilon\ell_1$ from $\mathbf{q}$ and distance $(1-\varepsilon)\ell_1$ from $\mathbf{z}_1$; while outside of $B(z_1, \ell_1)$, $\mathbf{q}$ is surrounded in all directions, at a spacing of $\varepsilon\ell_1$, by points of penalty equal to its own. So in this case too, $\mathrm{diam}(V(\mathbf{q})) = O(\varepsilon\ell_1)$. □

V Complexity

In this analysis K, d, and ϵ are treated as constants. (The runtime is exponential in each; details deferred.) We follow the outline given in section III. However, we start with the following preprocessing step. For each of the 2^d ways of orienting each of the coordinate axes, take all the points of the data set, label them with the largest (judged by this orientation) of their d coordinates, and then sort the entire data set by these labels. Retain these 2^d sorted lists for use below. The preprocessing requires time $O(2^d n \log n) = O(n \log n)$.

Step 1 takes time $O(n)$. In step 2, $\hat{\Delta}$ is halved at most $O(\log n)$ times. For each value of $\hat{\Delta}$ the space partitioning routine is run until it forms at most $2^{d+2}K/\epsilon + 2^d K^2 d c^d$ (i.e., a constant number of) regions. So the time spent per value of $\hat{\Delta}$ is proportional to the time required to split one region a and replace it with the regions $\{b_1, \ldots, b_{2^d}\}$ or the modified regions $\{b'_1, \ldots, b'_{2^d}\}$. Just one linear-time sweep over the data suffices to compute the distortions of each of $b_1, \ldots, b_{2^d}$. If exactly one of these (say b_1) exceeds distortion $\epsilon\hat{\Delta}/K$, then form the list of the points in b_1, sorted by their *maximum* coordinate as measured from an origin at the common vertex of a and b_1. (This sorted list can be obtained in linear time, by filtering out the points of b_1 from the appropriate sorted list created earlier.) Now shrink the boundaries of b_1, transferring points of b_1, in the stated order, to appropriate neighboring regions, until the halting condition is reached. Each transfer of a point (and subsequent update of the distortions of b_1 and the neighboring regions) can be performed in constant time. Thus the runtime per value of $\hat{\Delta}$ is $O(n)$, so the entire runtime of step 2 is $O(n \log n)$. Step 3 takes constant time, and step 4 takes time $O(n)$. Our total runtime is $O(n \log n)$.

Comment: This analysis relies only in a fairly casual way on the fact that the data set is finite. In fact, one of the design principles of our algorithm, which leads to its quasilinear runtime, is that the algorithm should access the data only through some elementary types of queries, and that the runtime should be expressible in terms of the numbers of these queries. Those queries are: given a polyhedral region A and a point x, compute $\rho(A, \mu(A))$ or $\rho(A, x)$; and given a polyhedral region A, sweep some of the walls of A until $\rho(A, \mu(A))$ reaches some prescribed value.

References

[1] P. K. Agarwal and M. Sharir. Efficient algorithms for geometric optimization. ACM Computing Surveys, 30:412–458, 1998.

[2] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for Euclidean k-medians and related problems. In *Annual ACM Symposium on Theory of Computing*, pages 106–113, Dallas, Texas, 1998.

[3] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proc. 34th ACM STOC*, pages 250–257, 2002.

[4] J. D. Bruce. *Optimum Quantization.* PhD thesis, M.I.T., Cambridge, MA, May 1964.

[5] W. Fernandez de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. In *STOC*, 2003.

[6] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1999.

[7] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proc. 20th Ann. ACM Symp. Theory Comput.*, pages 434–444. ACM, 1988.

[8] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:293–306, 1985.

[9] D. S. Hochbaum and D. B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. J. Assoc. Comput. Mach., 33(3):533–550, 1986.

[10] D. S. Hochbaum and D. B. Smoys. A best possible heuristic for the k-center problem. Math. Oper. Res., 10:180–184, 1985.

[11] K. Jain and V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *J. ACM*, 48:274–296, 2001.

[12] S. G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the Euclidean k-median problem. In *Proc. European Symposium on Algorithms*, pages 378–389. Springer-Verlag LNCS, 1999.

[13] D. Muresan and M. Effros. Quantization as histogram segmentation: globally optimal scalar quantizer design in network systems. In *Proceedings of the Data Compression Conference*, pages 302–311, Snowbird, Utah, March 2002.

[14] R. Ostrovsky and Y. Rabani. Polynomial time approximation schemes for geometric clustering problems. *J. ACM*, 49(2):139–156, 2002.

[15] L. J. Schulman. Clustering for edge-cost minimization. In *Proc. 32'nd Ann. Symp. on Theory of Computing (STOC)*, pages 547–555, 2000.

[16] D. K. Sharma. Design of absolutely optimal quantizers for a wide class of distortion measures. *IEEE Transactions on Information Theory*, IT-24(6):693–702, November 1978.

[17] X. Wu. *Algorithmic approach to mean-square quantization.* PhD thesis, University of Calgary, 1988.

[18] X. Wu and K. Zhang. Quantizer monotonicities and globally optimal scalar quantizer design. *IEEE Transactions on Information Theory*, IT-39(3):1049–1053, May 1993.