

Asymptotic regularity of subdivisions of Euclidean domains by iterated PCA and iterated 2-means

Arthur Szlam

Keywords: spectral clustering, vector quantization, PCA, k -means, multiscale analysis.

1. Introduction and results

The clustering problem is to divide a data set A into “nice” pieces, where nice is usually defined by geometric conditions, such as the isoperimetry of the pieces. Solutions to the problem have widespread applications in machine learning and data compression, and many clustering algorithms have been proposed and studied [JMF99]. Perhaps the most standard method for data in $A \subset \mathbb{R}^d$ is k -means [Mac67, HW79]; this consists of choosing a number k , and minimizing the energy

$$\sum_{j=1}^k \int_{C_j} \|x - c_j\|^2 dx \quad (1)$$

over partitions $C_1 \cup \dots \cup C_k = A$, $C_i \cap C_j = \emptyset$ for $i \neq j$, and where c_j are the centroids $\frac{1}{|C_j|} \int_{C_j} x dx$.

It has been recently noted that the k -means algorithm is closely related to Principal Components Analysis, or PCA, see [DFK⁺, ZDG⁺01], where it is proved that PCA is the continuous relaxation of the k -means minimization. The principal vectors of $A \subset \mathbb{R}^d$ are the eigenvectors of the $d \times d$ centered covariance matrix $V(A)$ with i, j entry

$$V(A)_{ij} = \frac{1}{|A|} \int_A (x_i - c)(x_j - c),$$

where c is the centroid of A ; see [Jol86]. A simple division of A into two pieces is to take the eigenvector v of V with largest eigenvalue, and centering the coordinate system at c , divide the set into $\langle x, v \rangle \geq 0$ and $\langle x, v \rangle < 0$.¹ This is the “best” linear division of the ambient \mathbb{R}^d in the sense of maximizing the variance in the cut direction, i.e.

$$\operatorname{argmax}_{v \in \mathbb{R}^d} \frac{1}{|A|} \int_A |\langle x, v \rangle|^2.$$

¹If there are several principal components with the same variance (eigenvalue), we can pick one at random, or, if appropriate, pick the cut which minimizes some measure of boundary.

For complicated data and large k , finding the minimum of the energy (1) can be difficult; and the PCA method naturally divides the data only in two. A standard procedure for obtaining many clusters is to iteratively subdivide, obtaining 2^n partition elements at the n th subdivision. In this article we will analyze the asymptotic behavior of repeated subdivisions of a data set in \mathbb{R}^d by 2-means or by the sign of the inner product against the first principal vector; we will henceforth refer to these two methods as iterated 2-means and iterated PCA, respectively. Instead of assuming a generative model for the data, we will instead choose a geometric model.

Specifically, we will assume the data is a bounded open set in \mathbb{R}^d . This is not an especially realistic assumption, but on the other hand, it is realistic to expect data from applications to be locally parameterized by a number of real variables (perhaps with different numbers of parameters at different locations in the data). A mathematically simple example of being locally parameterized by \mathbb{R}^d is an open subset of \mathbb{R}^d , so working with this assumption can be thought of as a first step. The contribution of this paper is to prove that in the case of domains in \mathbb{R}^d , the asymptotic regularity of iterated PCA or iterated 2-means is provably good except at perhaps a very small set of points.

After proving the main results in Sections 2 and 3, we will briefly discuss the case of kernelized binary subdivision by PCA in Section 4, and in particular, iterated binary subdivision with the second eigenfunction of the Laplacian on the data set. Although it is not clear how to extend the main results to this interesting setting, computer experiments suggest that similar results hold. Finally, we will conclude with a number of open questions.

1.1. Results

Before stating the main results, we start with a bit of notation. We will be studying binary partitions of the bounded open set $A \subset \mathbb{R}^d$. For each point $x \in A$, let $U_n(x)$ be the element of the partition at level n containing the point x . Define $e(A)$, the eccentricity of A , to be $\max \lambda_i / \lambda_j$, where the λ are the eigenvalues of the covariance matrix of A .

Proposition 1.1. *Let $U_n(x)$ be elements in the partition by iterated PCA of A . There exist constants $c_{PCA}(d) > 0$ depending only on the dimension so that there is an exceptional set $E(A)$ consisting of the countable union of $d - 2$ dimensional segments; and if $x \in A \setminus E$, then there exists $N = N(A, x)$ so that $e(U_n(x)) < c_{PCA}$ for all $n \geq N$.*

Proposition 1.2. *Let $U_n(x)$ be elements in the partition by iterated 2-means of A . There exist constants $c_{km}(d) > 0$ depending only on the dimension so that there is an exceptional set $E(A)$ consisting of the countable union of $d - 2$ dimensional segments; and if $x \in A \setminus E$, then there exists $N = N(A, x)$ so that $e(U_n(x)) < c_{km}$ for all $n \geq N$.*

Some remarks:

1. Note that the segments in the exceptional set E are isolated, in a sense which will be explained further at the end of Section 2.

2. There are indeed domains that have an exceptional set; see figure 1 for the schematics of an example.
3. The Propositions are asymptotic as opposed to quantitative, in the sense that they give no estimates on $N(A, x)$. On the other hand, no boundary regularity is used.
4. No attempt has been made to find the best constants, and the constants given by the proof below are super-exponential in d . It seems more likely that the correct constants are polynomial in d .

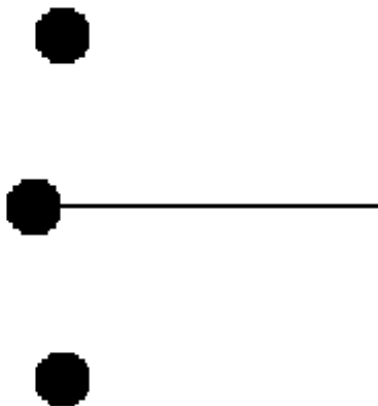


Figure 1: Schematic of a domain with an exceptional set. The thin line has small enough mass compared to the blobs to have negligible effect on variance computations. The first cut separates the top half from the bottom. The second cut passes between the two blobs, intersecting the first cut with a very small angle. The point on the corner with small angle will be an exceptional point.

2. Proof of Proposition 1.1

Throughout this section and the next, we will fix the domain being subdivided as $A \subset \mathbb{R}^d$. We will denote the metric on \mathbb{R}^d by ρ and the element of the partition containing x at level n by $U_n(x)$. The first part of the proof will be to reduce to the case of convex polygons by noting that each cut does not decrease the convexity, and that for large n , $U_n(x)$ is separated from ∂A .

Lemma 2.1. *If $S \subset L_{x,y}$ is a subset of the line segment between x and y , and if S is in U_n , and x and y are in U_{n+1} , $S \subset U_{n+1}$.*

Proof. This follows immediately from the fact that the boundary of the subdivision is a hyperplane. \square

Lemma 2.2. $\bigcap U_n(x) = \{x\}$

Proof. Let $U = \bigcap U_n(x)$. By Lemma 2.1, $U = A \cap \text{conv}(U)$ where $\text{conv}(U)$ is the set of convex combinations of points in U ; thus U has nonempty k dimensional interior for some $k \leq d$, and is contained in a k dimensional hyperplane K . Let $\epsilon > 0$ so that the d dimensional ball $B(x, \epsilon) \subset A$. Suppose there exists a point $y \neq x \in U$, and so $k > 0$; fix such a y in the k dimensional interior of U so that y is also in the interior of A (such a y exists in the k dimensional interior of $B(x, \epsilon) \cap U$). Rename a smaller number ϵ so that the d dimensional ball $B(y, \epsilon) \subset A$ and the k dimensional ball $B(y, \epsilon)$ is contained in the k dimensional interior of U , and $\|x - y\| > 2\epsilon$.

Now pick coordinates $a_1, \dots, a_k, b_1, \dots, b_{d-k}$ for \mathbb{R}^d so that a_i are parallel to K , and b_j are perpendicular to K , and so that the origin is at the center of mass of U_n . Furthermore, let a_1 be parallel to $L_{x,y}$. Let w be the diameter of U ; for large n , the diameter of U_n is less than $2w$. Suppose the maximum distance from K over points in U_n in the b_j direction is $h_j = h_j(n)$. We now bound the entries in the covariance matrix of $U_n(x)$:

$$\left| \int_{U_n} b_j^2 \right| \leq |U_n| h_j^2,$$

$$\left| \int_{U_n} a_i b_j \right| \leq \left| \int_{U_n} a_i^2 \right|^{\frac{1}{2}} \left| \int_{U_n} b_j^2 \right|^{\frac{1}{2}} \leq \sqrt{|U_n| 4w^2} \sqrt{|U_n| h_j^2} \leq 2|U_n| wh.$$

For each point z in $U_n \setminus B(y, \epsilon)$, by Lemma 2.1, the part of the line $L_{y,z}$ inside $B(y, \epsilon)$ is in U_n . Then we have the bounds

$$\frac{|B(y, \epsilon) \cap U_n|}{|U_n|} \geq \frac{|B(y, \epsilon) \cap (\bigcup_{z \in U_n} L_{y,z})|}{|\bigcup_{z \in U_n} L_{y,z}|} \geq \left(\frac{\epsilon}{w}\right)^d, \quad (2)$$

with equality when U_n is a spherical wedge with y at the center; we also have the same bound with $B(y, \epsilon)$ replaced by $B(x, \epsilon)$. Since either x or y is at a distance of greater than ϵ from the center of mass of U_n , we can then bound from below

$$\int_{U_n} a_1^2 \geq |B(x, \epsilon) \cap U_n| \epsilon^2 \gtrsim \left(\frac{\epsilon}{w}\right)^d |U_n| \epsilon^2.$$

Thus the principal vector of U_n asymptotically lies parallel to K . Since y is in the k dimensional interior of U , the bound on the ratio also shows that for large n , the projection of the center of mass of U_n onto K is in the interior of $\text{conv}(U)$, and stays uniformly bounded away from the boundary of $\text{conv}(U)$ in K as n increases. These two facts give a contradiction to the definition of U because the separating hyperplane would cut $\text{conv}(U)$, and the proof is complete. \square

Lemmas 2.1 and 2.2 together show that for each x , there is an $N(x)$ so that if $n > N(x)$, $U_n(x)$ is a convex polygon. $N(x)$ can be taken to be the first time $U_N(x)$ is separated from ∂A .

The next two lemmas rely heavily on convexity. They give an equivalence between L^∞ and L^2 measures of eccentricity. They are false in the non-convex case. If N is very large and ϵ is very small, $([-1, 1] \times [-2, 2]) \cup ([0, N] \times [-\epsilon, \epsilon])$ is a counterexample to both of the following lemmas.

Lemma 2.3. *There exists a constant $c_1 = c_1(d)$ so that if A is a convex polygon in \mathbb{R}^d with the origin at the center of mass of A , and x is a vector, then $\max_{y \in A} \langle x, y \rangle \geq c_1 \max_{y \in A} |\langle x, y \rangle|$*

Proof. Project A onto x ; but set 0 so that A lies entirely on the positive x axis. Let $p(x)$ be the density of the projection at x , and M be the maximum x value. Suppose that p takes its maximum value H at the point x_H . By convexity, the cone with peaks at a point projecting to the origin and at a point projecting to M and base at the cross section at x_H is contained in A . We wish to bound the integral $\int_0^M x p(x) dx$ from below; we do this by rearranging the mass of the cone as if x_H were at 0, and computing

$$\int_0^M x p(x) dx \geq \int_0^M x H \left(1 - \frac{x}{M}\right)^{d-1} dx = \frac{HM^2}{d(d+1)}.$$

On the other hand,

$$\int_0^M p(x) dx \leq HM,$$

so the moment of A has x coordinate greater than $M/d(d+1)$. We can bound the moment from above in the same manner. □

Lemma 2.4. *Suppose A is a convex polygon in \mathbb{R}^d . Let (x_1, \dots, x_d) be the coordinate system corresponding to the eigenvectors of the covariance matrix of A with 0 at the center of mass, and let λ_i be the corresponding eigenvalues. Then there exist constants c_2 and c_3 so that*

$$c_3 \sqrt{\lambda_i/\lambda_j} \max_{x \in A} |x_j| \geq \max_{x \in A} |x_i| \geq c_2 \sqrt{\lambda_i/\lambda_j} \max_{x \in A} |x_j|$$

Proof. Project A onto x_j , as before, but this time with 0 at the center of mass, and let p be the density. We have

$$\int_A x_i^2 = \frac{\lambda_i}{\lambda_j} \int_A x_j^2.$$

Let m_i and m_j be the maximal absolute values of x_i and x_j in A , and let $p(h) = H$ be the maximum value of p ; let \tilde{m}_j be the maximal absolute value of x_j in the same half line as h . Then we have

$$\int_A x_i^2 \leq H m_i^2 m_j.$$

On the other hand, we can estimate the variance in the j direction as above:

$$\int_A x_j^2 \geq \int_0^{\tilde{m}_j} x_j^2 H \left(1 - \frac{x_j}{\tilde{m}_j} \right)^{d-1} dx = \frac{2H\tilde{m}_j^3}{d(d+1)(d+2)}.$$

Thus

$$m_i^2 \geq \frac{\lambda_i \tilde{m}_j^3}{\lambda_j m_j^3} \frac{m_j^2}{d(d+1)(d+2)},$$

and

$$m_i \geq \sqrt{\frac{\lambda_i}{\lambda_j} c_1^3} \frac{m_j}{\sqrt{d(d+1)(d+2)}}.$$

The other inequality is proved similarly. \square

Proof of Proposition 1.1. Consider $U_n(x)$, where n is the smallest number large enough so that $\partial U_n(x) \cap \partial A = \emptyset$, and suppose $e(U_{n+k}(x)) > K^{2d}$ where K is a large number to be chosen later; and furthermore, temporarily suppose x is in the interior of U_n and pick k large enough so that $\partial U_{n+k} \cap \partial U_n = \emptyset$. The idea of the proof will be to show that none of the set of faces f_i of ∂U_{n+k} roughly parallel to the long directions of U_{n+k} could have been the PCA division of U_{n+k-1} . This will be done using lemmas 2.3 and 2.4; in short, the diameter of these faces are much larger than the possible diameter of U_{n+k-1} across the face. This forces U_{n+k-1} to lie in the simplicial cone generated by these faces. Any subset of the cone containing U_{n+k} will be shown to satisfy the same relative diameter estimates, and so all of the ancestors of U_{n+k} inside U_n thus lie in the cone. This would be a contradiction because $\partial U_{n+k} \cap \partial U_n = \emptyset$, and in particular, f_i are not in ∂U_n . To the details:

Suppose the principal values $e_d \leq \dots \leq e_1$ of U_{n+k} have been arranged in descending order. Because $e(U_{n+k}(x)) > K^{2d}$, there exists i so that $e_i/e_{i+1} > K^2$. Let V be the subspace of \mathbb{R}^d spanned by the $e_j, j \geq i$, and let $W = V^\perp$, and let $m = \max_{x \in A} |a|$, where $x = aw + bv$. V thus contains the ‘‘long’’ directions and W contains the ‘‘short’’ directions. By lemmas 2.3 and 2.4, and because the covariance of any two PCA coordinates is 0, for any $v \in V$,

$$\max_{s \in U_{n+k}} \left\langle \frac{v}{\|v\|}, s \right\rangle \geq c_1 c_2 K m.$$

Let f_i be the hyperplanes defined by the faces of U_{n+k} intersecting W , and let z_i be the normal vector to f_i such that $z_i \in f_i$; note that $\|z_i\| \leq m$ because there are points of W in f_i . Suppose $z_i = a_i w_i + b_i v_i$ where w_i and v_i are unit vectors in W and V , and $a_i, b_i \geq 0$. For each z_i there is a long vector $l_i \in U_{n+k}$, $l_i = p_i w_i + p'_i w'_i + q_i v_i + q'_i v'_i$, where $w'_i \in W$ is perpendicular to w_i , $v'_i \in W$ is perpendicular to v_i , and $q_i \geq c_1 c_2 K m$. By convexity, $\langle l_i, z_i \rangle \leq \|z_i\|^2$, and by Lemma 2.3 we were able to choose l_i so that $q_i \geq 0$. Then

$$a_i p_i + b_i q_i \leq a_i^2 + b_i^2,$$

and so

$$b_i q_i - b_i^2 \leq a_i^2 + a_i m,$$

and

$$b_i \leq \frac{a_i(a_i + m)}{q_i - b_i} \leq \frac{2a_i}{c_1 c_2 K - 1}. \quad (3)$$

Denote by Y the simplicial cone $\{y : \langle z_i, y \rangle \leq \|z_i\|^2\}$. Consider a point $y = pw + qv \in Y$ where we choose w so $p > 0$; assume that $p > 4m$. Let \bar{U}_{n+k} and \bar{Y} be the projection of U_{n+k} and Y onto the (w, v) plane. The maximal distance in the w direction over points in \bar{U}_{n+k} is still bounded by m , and there is a point with v coordinate greater than $c_1 c_2 K m$. Let $z = aw + bv$ be normal to and contained in the line in the boundary of \bar{U}_{n+k} intersecting the positive w axis. The inequality in equation (3) holds between a and b , and so,

$$p \leq \frac{|b||q|}{a} + 2a,$$

so

$$\frac{p}{2} \leq \frac{|b||q|}{a},$$

and

$$p \leq \frac{4q}{c_1 c_2 K - 1}. \quad (4)$$

Let ρ_{f_i} be the distance between the projection of points onto f_i . If

$$\max_{s \in U_{n+k}} \rho(s, f_i) < c_1 c_2 \max_{s \in U_{n+k}} \rho_{f_i}(s, z_i)/2, \quad (5)$$

then by Lemma 2.4 U_{n+k-1} was not split along f_i ; the same holds for U_{n+k-1} , etc. So suppose $y = pw + qv \in Y$, and assume first (4) holds. Then

$$\frac{|\langle y, z_i \rangle|}{\|z_i\|} \leq \frac{|ap| + \frac{2|aq|}{c_1 c_2 K - 1}}{|a|} \leq \frac{6|q|}{c_1 c_2 K - 1},$$

and the distance from y to f_i satisfies

$$\|z_i\| - \langle y, z_i \rangle / \|z_i\| \leq m + \frac{6|q|}{c_1 c_2 K - 1} \leq \frac{7|q|}{c_1 c_2 K - 1};$$

but

$$\rho_{f_i}(y, z_i)^2 = \|y\|^2 - \frac{|\langle y, z_i \rangle|^2}{\|z_i\|^2},$$

and as long as $\frac{7}{c_1 c_2 K - 1} < \frac{1}{4}$,

$$\rho_{f_i}(y, z_i) \geq |q|/2.$$

If equation (4) does not hold, $4m > p > \frac{4q}{c_1 c_2 K - 1}$,

$$\frac{|\langle y, z_i \rangle|}{\|z_i\|} \leq 6m,$$

and the distance from y to f_i is bounded by $7m$. On the other hand, there is a point $y' \in U_{n+k}$ with $\|y'\| \geq c_1 c_2 K m$, and as long as $\frac{c_1 c_2 K}{c_1 c_2 K - 1} < 2$, $\frac{|\langle y', z_i \rangle|}{\|z_i\|} \leq 5m$, and

$$\rho_{f_i}(y', z_i)^2 = \|y'\|^2 - \frac{|\langle y', z_i \rangle|^2}{\|z_i\|^2} \geq [(c_1 c_2 K)^2 - 25]m^2.$$

Thus if $c_1 c_2 (c_1 c_2 K - 1) > 28$ and $c_1 c_2 \sqrt{(c_1 c_2 K)^2 - 25} > 14$, equation (5) holds, and f_i could not have been the boundary of the PCA cut. Then $U_{n+k-1} \subset Y$; again equation (5) holds, except with the maximum on either side of the inequality taken over $s \in U_{n+k-1}$, and none of the f_i were the boundary of the PCA cut. We iterate, and because $\partial U_{n+k} \cap \partial U_n = \emptyset$ we arrive at a contradiction, because the f_i can never be PCA boundaries, but U_n is an ancestor of U_{n+k} .

Earlier we assumed x was in the interior of U_n . If x is in the interior of a face f of the boundary of U_n , the above argument goes through; we pick k so that $\partial U_{n+k} \cap \partial U_n = f$. Then there is at least one f_i as above. However, if x is on the boundary between two faces of ∂U_n , it is possible that these two faces are the f_i for all k . Here we have no control and give up, and assign such x to the exceptional set. □

To end this section we expand on the first remark after the statement of the Propositions. Call a partition element emancipated if its parent contains a boundary point, but it does not. The exceptional set was identified as a subset of the $d - 2$ dimensional faces of the eccentric emancipated partition elements.

Note that for large finite n , all the partition elements in the n th generation whose closures contain a given point x are separated from the boundary of the domain. If not, there would be a sequence of partition elements whose closures contain x and some point in the boundary; we can take these partition elements to be nested descendants (at each subdivision, ask the question: are there infinitely many descendants of this set containing x and a boundary point? if yes, take this set, else take its sibling). In this case we can follow the proof of Lemma 2.2 and via a compactness argument achieve a contradiction. Thus given any exceptional point x , there is a ball centered at x so that any other exceptional point in the ball is in a $d - 2$ face of an emancipated partition element containing x ; and by the above argument there are only finitely many such partition elements. In particular, it is not possible to have sequences of exceptional faces approaching any other exceptional face.

3. Proof of Proposition 1.2

The analysis in the previous section can be used to prove a similar result for iterated 2-means.

Since the division by 2-means cuts by hyperplanes, we again have the use of Lemma 2.1. Our proof will follow the same outline as the proof of Proposition 1.1; except we will need two more lemmas.

Lemma 3.1. *In subdivision by iterated 2-means, $\bigcap U_n(x) = \{x\}$*

Proof. As before, let $U = \bigcap U_n(x)$, and suppose there is some point $y \neq x \in U$. Since the subdivision procedure divides by hyperplanes, $U = A \cap \text{conv}(U)$ where $\text{conv}(U)$ is the set of convex combinations of points in U ; thus U has nonempty k dimensional interior for some $k \leq d$, and is contained in a k dimensional hyperplane K . The argument in Lemma 2.2 leading to equation (2) goes through unchanged, and as before, this means that the projection of the center of mass of U_n onto K for large n is contained in the k dimensional interior of $\text{conv}(U)$ in K , and that the distance of the center of mass to K tends to 0. U_n has two children, U_{n+1} and W_{n+1} . We claim that the projection of the center of mass of W_{n+1} is also in the k dimensional interior of $\text{conv}(U)$ for some large n (and so small h). This leads to a contradiction; because the separating hyperplane between the two centers would cut $\text{conv}(U)$. To check the claim, let u_n be the center of mass of U_n and w_n be the center of mass of W_n . Because A is bounded, there are simultaneously convergent subsequences $u_{n_j} \rightarrow u$ and $w_{n_j} \rightarrow w$. In fact, $u = w$, else for large enough j , and u_{n_j} and w_{n_j} close enough to u and w , the division of U_{n_j} into U_{n_j+1} and W_{n_j+1} separates u and w . This cannot happen because $W_{n_j} \subset U_{n_j-1}$ for all n_j . Thus there are w_{n_j} arbitrarily close to u in the interior of U , and the Lemma is proved. \square

We also need:

Lemma 3.2. *If A is a convex polygon in \mathbb{R}^d with diameter D , and C_1 and C_2 are the centroids of the 2-means partition of A , then there is a constant c_4 so that $\rho(C_1, C_2) \geq c_4 D$*

Proof. Let A_1 and A_2 be a partition of A with centroids C_1 and C_2 , and let $s = \rho(C_1, C_2)$. By assumption there are points y_1 and y_2 in A distance D apart; and so wlog

$$\|y_1 - C_1\| > D/2 - s,$$

and $y_1 \in A_1$. Some manipulation shows that the energy of the partition is

$$\int_{A_1} \|x - C_1\|^2 dx + \int_{A_2} \|x - C_2\|^2 dx = \int_A \|x\|^2 - m_1 \|C_1\|^2 - m_2 \|C_2\|^2.$$

Center the coordinate system at C_1 ; then $m_1 \|C_1\|^2 + m_2 \|C_2\|^2 < ms^2$. Since C_1 is the centroid of the convex polygon A_1 , by Lemma 2.3 there is a point y'_1 in A_1 in the $-y_1$ direction with $\|y'_1\| > c_1(D/2 - s)$. Consider the hyperplane H passing through C_1 perpendicular to y_1 . Now we compute the energy of the subdivision of A by H into A'_1 and A'_2 . By an argument as in 2.3, we find that the centroids C'_1 and C'_2 satisfy

$$\|C'_i\| > \frac{c_1(D - 2s)}{2(d+1)d}.$$

Using the A'_i with larger mass we find that for the subdivision by H to have smaller energy, it suffices that

$$2ms^2 < m \left[\frac{c_1(D - 2s)}{2(d+1)d} \right]^2,$$

and so it suffices that

$$D > \frac{(2\sqrt{2} + 1)(d + 1)d}{c_1} s.$$

□

The proof of Proposition 1.2 is obtained by following the proof of Proposition 1.1, but using Lemma 3.1 in place of Lemma 2.2 and adjusting (5) to make use of Lemma 3.2.

4. Kernel PCA and eigenfunctions of the Laplacian

Subdivision by iterated PCA is a subset of spectral partitioning, in which some kernel K is associated to the data, and the eigenfunctions of the kernel determine the subdivision. In the case of subdivision by PCA, the kernel was just the Gram matrix of the centered data. With a more general kernel, one can think of this as mapping the data to a different space where the inner product of two data points x and y is defined by $\langle x, y \rangle = x^t K y$. A popular choice of kernel is some sort of “Laplacian” on the data, as in [SM00, KVV04, Chu97]. Here we briefly discuss subdivisions of a bounded domain in \mathbb{R}^d into the sets where the second eigenfunction of the Neumann Laplacian or Neumann heat kernel is positive and where it is negative. In this case, the first eigenfunction is always the constant function, and all other eigenfunctions integrate to 0. Because of this, if K_t is the heat kernel on the domain at time t , dividing the domain A according to the second eigenfunction of K_t is exactly embedding A into L^2 by $x \mapsto K_t(x, \cdot)$, centering, and performing the standard PCA division in the embedding space.

The eigenfunctions of the Laplacian describe the resonant modes of random walks on the data in question, and so it is reasonable that there is a characterization of the partitions in terms of random walk. Here we mention such a characterization for bounded domains; compare with the ideas in [MS01] for the discrete case. Let A be a bounded domain in \mathbb{R}^d with enough boundary regularity so that Green’s identities hold. Let \mathcal{A} be the collection of subdomains of A with the property that for each $A \in \mathcal{A}$ Green’s identities hold on A . Let \mathcal{P} be pairs of subdomains $\{A_1, A_2\}$ of A such that $A_i \in \mathcal{A}$, $\bar{A}_1 \cup \bar{A}_2 = \bar{A}$, $A_1 \cap A_2 = \emptyset$. Let $P_{\mathcal{N}}$ be the pair of nodal domains for the first non-constant Neumann eigenfunction ϕ of A , i.e. $P_{\mathcal{N}}$ is the pair of domains $\{\{x \in A | \phi(x) > 0\}, \{x \in A | \phi(x) < 0\}\}^2$.

Given a pair $\{A_1, A_2\} \in \mathcal{P}$, and a point $x \in A$, let A_x be the subdomain containing x . Let $B_x(t)$ be Brownian motion started from x , killed on $\bar{D}_1 \cap \bar{D}_2$, and reflected on ∂A . Then

$$P_{\mathcal{N}} = \underset{\mathcal{P}}{\operatorname{argmin}} \lambda,$$

²If the first eigenspace is multidimensional, just pick a vector in the eigenspace, or choose the vector that minimizes the boundary area.

where for each x in A_1 and A_2 there exists C_x such that

$$\lim_{t \rightarrow \infty} \Pr(B_x(t) \in A_x) e^{\lambda t} = C_x$$

Although not stated as such, this theorem is proved for \mathbb{R}^2 in [BB99], and Theorem 1.7 of [HS89] and Theorem 14a of [Whi57] complete the proof in \mathbb{R}^d .

As before, we are interested in the shape of the partition elements after many iterations. Very little seems to be known about the asymptotics, even for domains in \mathbb{R}^2 . In the graph setting, there is [KVV04], which compares the quality of the clustering obtained by the partition to the “best” clustering of the graph in the sense of conductance, using Cheeger’s inequality as the major tool. It seems that with the extra structure afforded by a Euclidean domain, it should be possible to prove much stronger theorems than in [KVV04]. Indeed, computer experiments suggest that in \mathbb{R}^2 , the elements in the subdivision seem to limit to polar or cartesian rectangles.³

In figure 2 we show several levels in the decomposition of a domain in \mathbb{R}^2 using the second eigenvector method sampled at gridpoints with approximately 100,000 samples. Note that the intersections between boundaries is always at right angles (this is due to the Neumann condition), and how in the acute corners, the decomposition becomes a polar grid.

5. Conclusions and further work

We have shown that the asymptotic regularity of the clusters (or Voronoi cells) found by iterated 2-means or iterated PCA subdivision is good in the sense of eccentricity if the initial data is a bounded open subset of \mathbb{R}^d . On the other hand, there are many questions which remain unaddressed:

1. What are the best constants in Propositions 1.1 and 1.2?
2. Is there a quantitative statement of the Propositions, relating N such that $U_n(x)$ is regular for $n > N$ to the boundary regularity of A and the distance of x from the boundary?
3. Is there a similar theorem for k -means, where $k > 2$?
4. Is there a kernelized version of the Propositions? Of particular interest is the case of the heat kernel.
5. Or, closely related, is there a statement of the Propositions for the image of a bounded open set under a smooth map, or for manifolds or varifolds?

6. Acknowledgements

Thanks to the NSF for generous support by DMS-0811203, and to the reviewers for their helpful suggestions.

³If the subdivision is carried out on a right 45 triangle, all the partition elements will be right 45 triangles; however, this situation is unstable.

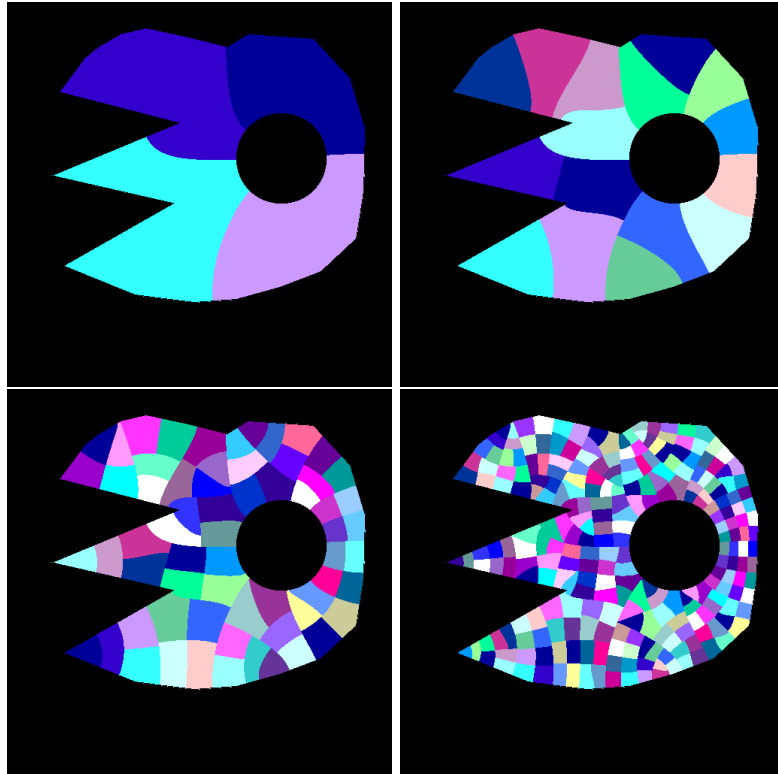


Figure 2: Levels 2, 4, 6, and 8 in the iterated subdivision of a domain in \mathbb{R}^2 sampled at approximately 100,000 grid points. Each subdomain was partitioned into the positive and negative set of first nontrivial eigenfunction of the Neumann Laplacian on that subdomain. The colors represent different regions.

References

- [BB99] Rodrigo Bañuelos and Krzysztof Burdzy. On the “hot spots” conjecture of J. Rauch. *J. Funct. Anal.*, 164(1):1–33, 1999.
- [Chu97] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.
- [DFK⁺] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33.
- [HS89] Robert Hardt and Leon Simon. Nodal sets for solutions of elliptic equations. *J. Differential Geom.*, 30(2):505–522, 1989.
- [HW79] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

- [JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.
- [Jol86] T. Jolliffe. *Principal Components Analysis*. Springer-Verlag, 1986.
- [KVV04] R. Kannan, S. Vempala, and A. Vetta. On clusterings: good, bad and spectral. *J. ACM*, 51(3):497–515 (electronic), 2004.
- [Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, 1:281–296, 1967.
- [MS01] M. Meila and J. Shi. Learning segmentation by random walks. In *NIPS*, 2001.
- [SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [Whi57] Hassler Whitney. *Geometric integration theory*. Princeton University Press, Princeton, N. J., 1957.
- [ZDG⁺01] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering, 2001.