ON THE MINIMUM OF THE MEAN-SQUARED ERROR IN 2-MEANS CLUSTERING

BERNHARD G. BODMANN AND CRAIG J. GEORGE

Abstract. We study the minimum mean-squared error for 2-means clustering when the outcomes of the vector-valued random variable to be clustered are on two spheres, the surface of two touching balls of unit radius in n-dimensional Euclidean space, and the underlying probability distribution is the normalized surface measure. For simplicity, we only consider the asymptotics of large sample sizes and replace empirical samples by the probability measure. The concrete question addressed here is whether a minimizer for the mean-squared error identifies the two individual spheres as clusters. Indeed, in dimensions $n \geq 3$, the minimum of the mean-squared error is achieved by a partition obtained from a separating hyperplane tangent to both spheres at the point where they touch. In dimension $n = 2$, however, the minimizer fails to identify the individual spheres; an optimal partition is associated with a hyperplane that does not contain the intersection of the two spheres.

1. Introduction

In many applications of data science, large sets of vectors need to be grouped into a small number of subsets whose elements are close to each other. This type of partitioning into subsets is also called *clustering* [13]. The subsets are often believed to be distinct constituents in a mixture of random vectors that are sampled from different distributions. In many cases, the distributions are from a known family that is parametrized by the expected value of the outcomes, and the outcomes concentrate near the expected value [17, 3]. Partitioning the observed set of vectors into subsets yields the empirical means, also called centroids, which provide an estimate for the expected values. On the other hand, once the expected values are accurately determined, one assumes that mapping each vector to the subset whose centroid is closest provides a good partition. This heuristic approach to the clustering problem is captured in an iterative algorithm by Lloyd [10], which aims to minimize an objective function that measures the Euclidean mean squared distance of the elements in each of the subsets from the respective centroid. Although the algorithm seems to work well in practice, known results lack general a priori performance guarantees [1, 9, 19, 4, 12] or show cases with slow convergence [21] even for two-dimensional clustering.

Another setting in which one tries to minimize the mean-squared distance is in vector quantization [2, 5], see also [20]. There, partitioning of the outcomes of a random vector

---

Key words and phrases. k-means clustering, performance guarantees; MSC (2010): 62H30.

This paper was supported in part by an REU portion of NSF grant DMS-1412524 and by NSF grant DMS-1715735.

is not explicitly motivated by an underlying assumption that it is a mixture. The main goal is to approximate the random vector by a quantized one, with a finite or discrete set of outcomes while minimizing the distortion, measured in the expected Euclidean squared norm of the quantization error or in terms of more general norms [6].

In this paper, we investigate the problem of minimizing the objective function appearing in Lloyd's algorithm for the special case of partitioning into two subsets. Optimality for the 2-means problem has already been considered in dimension $n = 2$ for the concrete examples of the uniform distribution on the disk and on the square [18]. We consider the example of random vectors governed by a probability measure ρ that is formed by taking the average of two probability measures that are uniform on two spheres, the surface of two balls of unit radius in n-dimensional Euclidean space. If the set S is the union of the two touching spheres and ρ the associated normalized surface measure, we wish to find the assignment $q : S \to \{c_1, c_2\}$ which maps S to $c_1, c_2 \in \mathbb{R}^n$ such that the mean-squared error $\int_S \|x - q(x)\|^2 d\rho(x)$ is minimized. The concrete question is then whether an optimizer to the mean-squared error assigns, up to sets of measure zero, a partition that singles out each individual sphere.

Earlier results prove that applying semidefinite programming to a convex relaxation of the objective function in Lloyd's clustering algorithm [16] is successful if the spheres are sufficiently separated [7, 8, 11], see also a separation requirement for more general, subgaussian clusters [14]. Indeed, in dimension $n = 1$, the desired result is achieved if and only if the spheres are separated by a sufficiently large distance. A unit sphere in dimension $n = 1$ is a set of two points at a distance of 2. The uniform probability measure on two symmetrically arranged spheres at a distance 2ϵ is $\rho = (1/4)\delta_{-2-\epsilon} + (1/4)\delta_{-\epsilon} + (1/4)\delta_{\epsilon} + (1/4)\delta_{2+\epsilon}$, where δ_a is for any $a \in \mathbb{R}$ a Dirac measure with support $\{a\}$. If we choose $0 < \epsilon < (\sqrt{3} - 1)/2$, then by exhausting all choices of partitions, it is seen that the set $S_1 = \{-\epsilon, \epsilon, 2 + \epsilon\}$ with mean $m_1 = (2 + \epsilon)/3$ and the set $S_2 = \{-2 - \epsilon\}$ with mean $m_2 = -2 - \epsilon$ provide an optimal partition of $\{-2 - \epsilon, -\epsilon, \epsilon, 2 + \epsilon\}$ for which the resulting mean-squared error is $2(1+\epsilon+\epsilon^2)/3 < 1$, whereas the symmetric choice $R_1 = \{\epsilon, 2+\epsilon\}$ and $R_2 = \{-\epsilon, -2 - \epsilon\}$ gives a mean-squared error of 1. On the other hand, if $\epsilon > (\sqrt{3} - 1)/2$, then the partitioning into R_1 and R_2 is indeed optimal for the mean-squared error.

It is tempting to attribute the failure to recover the individual spheres to the discrete nature of the "surface" measures in $\mathbb{R}$. A closer look shows that the concentration of the measure near the origin is the reason for the optimal partition formed by one sphere cannibalizing the other. As n grows, the measure ρ is less concentrated near the origin, and one expects this cannibalizing behavior to disappear. Here, we examine the question whether a successful partition can be obtained in dimensions $n \geq 2$ even if the spheres touch. This is the most challenging case in which separation can still be achieved theoretically. We consider the continuum limit, which means instead of sampling the distributions with finitely many outcomes, we assume data given in the form of uniform measures on the spheres.

Our results show that minimizing the mean-squared error in $\mathbb{R}^2$ leads to a non-symmetric partition, as in the case of dimension $n = 1$. Fortunately, in dimensions $n \geq 3$ the minimizer recovers the partition into individual spheres, as one hopes to achieve. In that case, the partition is symmetric (up to sets of measure zero); it is given by a separating hyperplane that is invariant under reflections mapping each sphere onto the other.

This paper is organized as follows: In Section 2, we present the main results. The proofs are either elementary and included there or relegated to the Section 3. A first part of the proofs establishes that optimal partitions for 2-means clustering are obtained from separating hyperplanes. The next part determines the location of the hyperplane.

Acknowledgment. Both authors would like to thank Dustin Mixon for suggesting the intriguing calculus exercise worked out in Section 3.2. Additional thanks go to the anonymous referee for comments that helped improve the presentation of this paper.

2. Optimal partitions for the mean-squared error

The problem we are concerned with is the minimization of the mean-squared error. Its value depends on the partition of the support of a probability measure ρ describing the outcomes of a mixture of random vectors.

Definition 2.1. Given a Borel probability measure ρ on $\mathbb{R}^n$ with support S and a Borel-measurable subset $S_1 \subset S$ with complement $S_2 = S \setminus S_1$, then the *mean squared error* associated with the partition $\{S_1, S_2\}$ of S is

$$\mathcal{E}(S_1) = \min_{c_1 \in \mathbb{R}^n} \int_{S_1} \|x - c_1\|^2 d\rho(x) + \min_{c_2 \in \mathbb{R}^n} \int_{S_2} \|x - c_2\|^2 d\rho(x).$$

Here, $\|x - c_i\|$ is the Euclidean distance between x and c_i in $\mathbb{R}^n$, $i \in \{1, 2\}$.

In this paper, we are concerned with a special case where ρ is the (normalized) surface measure for the union of two touching spheres,

$$\rho = \frac{1}{2}(\sigma_{-1} + \sigma_1)\,.$$

Here σ_a is the surface measure supported on $\mathbb{S}_a \equiv \{x \in \mathbb{R}^n : \|x - ae_1\| = 1\}$, where e_1 is the first canonical basis vector in $\mathbb{R}^n$. The measure σ_a is obtained from translating σ_0, so for any Borel measurable set A, $\sigma_a(A + ae_1) = \sigma_0(A)$, and for any orthogonal matrix O, $\sigma_0(A) = \sigma_0(O^{-1}(A))$.

The following are the main theorems in this paper:

Theorem 2.2. *Let the Borel measure be given by $\rho = \frac{1}{2}(\sigma_{-1} + \sigma_1)$ on $\mathbb{R}^n$ with support $S = \mathbb{S}_{-1} \cup \mathbb{S}_1$. Let S_1, S_2 form a partition of S into two Borel measurable subsets, then there exist $a \in \mathbb{R}$ and $T_1 = \{x \in \mathbb{R}^n : x_1 \leq a\}$ such that $\mathcal{E}(T_1) \leq \mathcal{E}(S_1)$. Moreover, if S_1 is minimal for the mean-squared error, then there is a choice of the cutoff a for which T_1 coincides with S_1 or S_2, up to a set of zero probability.*

In short, disregarding sets of zero probability, an optimal partition of S is given by two sets separated by a hyperplane orthogonal to e_1, at an offset a from the origin. The fact that an optimal partition comes from a separating hyperplane is well known [4], which we supplement with a symmetrization argument.

This result motivates abbreviating the mean-squared error for this special case, and studying its dependence on the cutoff,

$$E(n,a) = \mathcal{E}(\{x \in S : x_1 \le -a\})\,.$$

By the reflection symmetry of ρ with respect to the first coordinate, it is sufficient to consider $E(n,a)$ for $a \ge 0$. With this simplification, we can study the case of dimension $n = 2$ in elementary terms.

Theorem 2.3. *In dimension $n = 2$, the absolute minimum of $E(2,a)$ among $a \in [0,2)$ is attained at a non-zero cutoff a.*

Proof. Parametrizing the two circles by arc length gives by a direct computation for $a = 1 - \frac{\sqrt{3}}{2}$ the probabilities $\rho(\{x \in \mathbb{R}^2 : x_1 \le -1 + \sqrt{3}/2\}) = 5/12$ and $\rho(\{x \in \mathbb{R}^2 : x_1 > -1 + \sqrt{3}/2\}) = 7/12$. Choosing $c_1 = (\zeta_1, 0)$ and $c_2 = (\zeta_2, 0)$ with $\zeta_1 = -1 - 3/(5\pi)$ and $\zeta_2 = 5/7 + 3/(7\pi)$ gives for the mean-squared error

$$\begin{aligned} E\big(2, 1 - \tfrac{\sqrt{3}}{2}\big) \le \frac{1}{4\pi}\bigg(&\int_{\pi/6}^{11\pi/6} ((-1+\cos t - \zeta_1)^2 + \sin^2 t)dt \\ &+ \int_{-\pi/6}^{\pi/6} ((-1+\cos t - \zeta_2)^2 + \sin^2 t)dt + \int_0^{2\pi} ((\cos t + 1 - \zeta_2)^2 + \sin^2 t)dt\bigg) \\ = \frac{45\pi^2 - 30\pi - 9}{35\pi^2} < 0.987\,. \end{aligned}$$

This is less than $E(2,0) = 1$, so the absolute minimum is not attained at $a = 0$. □

To illustrate this result, we have computed the minimizing offset numerically and plotted the resulting partition of the two circles in Figure 1, together with the value of the mean-squared error associated with a given offset in Figure 2.

When the means of the two subsets $\{x \in \mathbb{R}^2 : x_1 \le -a\}$ and $\{x \in \mathbb{R}^2 : x_1 > -a\}$ then Theorem 2.2 reduces identifying the optimal mean-squared error to finding the minimum of a parameter integral.

In dimension $n = 3$, the mean-squared error can be computed explicitly.

Theorem 2.4. *In dimension $n = 3$, the absolute minimum of $E(3,a)$ among $a \in [0,2)$ occurs at $a = 0$.*

Proof. We parameterize the two spheres by spherical coordinates and normalize the measure by surface area. Based on Theorem 2.2, an optimal partition is obtained with a separating hyperplane orthogonal to the symmetry axis $\mathbb{R}e_1$. The associated probabilities are for $-2 \le a \le 2$: $\rho(\{x \in \mathbb{R}^2 : x_1 \le -a\}) = \frac{1}{2} - \frac{a}{4}$ and $\rho(\{x \in \mathbb{R}^2 : x_1 > -a\}) = \frac{1}{2} + \frac{a}{4}$. As shown in Theorem 3.4 below, the mean-squared error is obtained by choosing c_1 and c_2 to be the means of the two subsets, $c_1 = (\zeta_1, 0, 0)$, $c_2 = (\zeta_2, 0, 0)$ with $\zeta_1 = -1 - \frac{1}{2}a$, $\zeta_2 = 1 - \frac{1}{2}a$.

This choice results in

$$\begin{aligned} E(3,a) = \frac{1}{8\pi}\Bigg(&\int_0^{2\pi}\int_{\arccos(1-a)}^{\pi}((-1+\cos u-\zeta_1)^2+\sin^2 u)\sin u\,dudt \\ &+\int_0^{2\pi}\int_0^{\arccos(1-a)}((-1+\cos u-\zeta_2)^2+\sin^2 u)\sin u\,dudt \\ &+\int_0^{2\pi}\int_0^{\pi}((1+\cos u-\zeta_2)^2+\sin^2 u)\sin u\,dudt\Bigg) \\ =\ &\frac{1}{4}a^2+1\,. \end{aligned}$$

Thus $E(3,a)$ achieves its absolute minimum at $a=0$. □

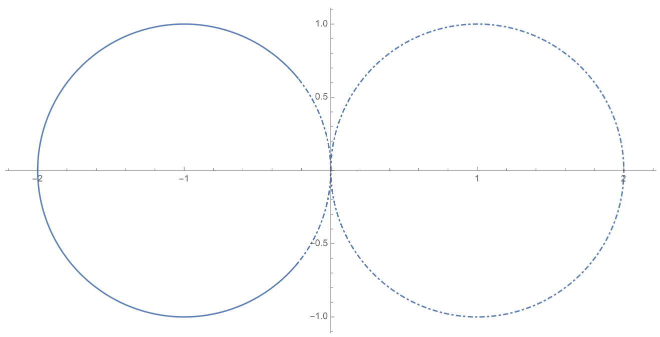

FIGURE 1. An optimal partition of the union of two circles. First set (solid) on left, second (dash-dotted) on right.

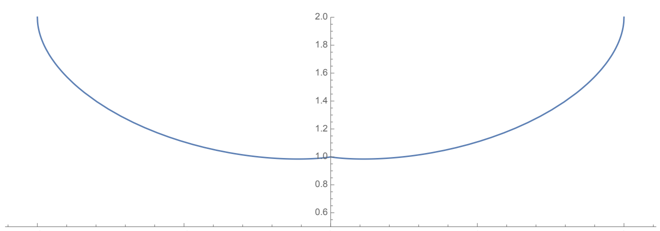

FIGURE 2. Value of $E(2,a)$ depending on cutoff $a\in[-2,2]$, with minimum achieved at two non-zero values of a.

Even in the absence of explicit computations for $E(n,a)$ in case $n>3$, we obtain the same monotonicity property as for $n=3$.

Theorem 2.5. *The inequality* $\frac{\partial}{\partial a}E(n,a)>0$ *holds for all* $a\in(0,2)$ *and* $n>3$. *Moreover,* $E(n,a)$ *attains a minimum at* $a=0$, *and this minimum is unique.*

Theorems 2.4 and 2.5 give us that the 2-means objective function E of two touching n-spheres is increasing in the variable a for the cutoff for $n \geq 3$ in the continuum limit. Thus, for dimensions $n \geq 3$, the optimal 2-means cutoff has a value of zero, so both n-spheres are recovered successfully.

The remainder of the paper is dedicated to the proofs of Theorems 2.2 and 2.5.

3. Proofs of main results on optimal partitions

The first part of this section establishes the proof that an optimal partition is given by a separating hyperplane that is orthogonal to the symmetry axis. The second part examines the offset of the optimal separating hyperplane.

3.1. Minimizing the mean-squared error by partitions with a separating hyperplane.

First, we consider a general Borel measure ρ with support S in $\mathbb{R}^n$. Given a partition $\{S_1, S_2\}$ of S, and $\rho(S_i) > 0$, then we call $m(S_i) = \int_{S_i} x d\rho(x)/\rho(S_i)$ the *mean* associated with the set S_i. If S_i is clear from the context, we also abbreviate $m_i = m(S_i)$.

By a direct computation, we have for any S_i with $\rho(S_i) > 0$ and $c_i \in \mathbb{R}^n$

$$\int_{S_i} \|x - c_i\|^2 d\rho(x) = \int_{S_i} \|x - m_i\|^2 d\rho(x) + \|c_i - m_i\|^2 \rho(S_i)\,,$$

so the minimum is achieved if and only if $c_i = m_i$.

Moreover, given $c_1, c_2 \in \mathbb{R}^n$, then among all the partitions, the partition into Voronoi regions is optimal, as shown in Lemma 3.2 below.

Definition 3.1. Given $c_1, c_2 \in \mathbb{R}^n$, we define the *Voronoi partition* $\{T_1, T_2\}$ of a Borel set S associated with the vectors c_1 and c_2 by the assignment

$$T_1 = \{x \in S : \|c_1 - x\| \leq \|c_2 - x\|\}\,, T_2 = S \setminus T_1\,.$$

From this definition, we see that this Voronoi partition consists of a closed half-space and its complement, with a separating hyperplane that is orthogonal to $c_1 - c_2$ and contains the midpoint $(c_1 + c_2)/2$.

Next, we note that given a partition into sets of non-zero probability, passing to the Voronoi partition associated with the means can only improve the mean-squared error. This fact is generally known, see for example [4, Proposition 3.1].

Lemma 3.2. *Let S_1, S_2 be a partition of S with $0 < \rho(S_1) < 1$ and associated means m_1 and m_2, then the Voronoi partition associated with m_1, m_2 satisfies*

$$\mathcal{E}(T_1) \leq \mathcal{E}(S_1)\,.$$

Proof. For any measurable partition S_1 and S_2 and $i \in \{1, 2\}$, choosing any $x \in T_i$ gives by the definition of the Voronoi partition $\|x - m_i\| \leq \min\{\|x - m_1\|, \|x - m_2\|\}$. Thus, the partition of S into T_1 and T_2 gives a mean-squared error that is bounded above by that associated with S_1 and S_2. □

In the following, we focus on properties of optimal partitions. These properties are also known, even in the more general context of k-means, see e.g. [4, Propositions 3.1 and 3.5] or [6, Section 4.1]. We have decided to include them here to keep the exposition self-contained.

Lemma 3.3. *If $\{S_1, S_2\}$ is a minimizing partition for the mean-squared error, then $0 < m(S_i) < 1$ for $i \in \{1, 2\}$ and $m(S_1) \neq m(S_2)$.*

Proof. Let $\{S_1, S_2\}$ be a minimizing partition. We know $0 < \rho(S_1) < 1$, otherwise S_1 or S_2 have unit measure and we can refine S_1 or S_2 and improve the mean-squared error.

Moreover, assuming an optimal partition into two sets S_1 and S_2 of non-zero probability and equal means $m_1 = m_2$, then any partition performs equally well, and we can choose a subset $R_1 \subset S_1$ with $0 < \rho(R_1) < 1$ such that the associated mean $r_1 \equiv m(R_1) \neq m_1$. By the characterization of the mean, then $\int_{R_1} \|x - r_1\|^2 d\rho(x) < \int_{R_1} \|x - m_1\|^2 d\rho(x)$. For the partition formed by R_1 and $R_2 = S \setminus R_1$, we then get that

$$\begin{aligned}\int_{R_1} \|x - r_1\|^2 d\rho(x) + \int_{R_2} \|x - m_1\|^2 d\rho(x) &< \int_{R_1} \|x - m_1\|^2 d\rho(x) + \int_{R_2} \|x - m_1\|^2 d\rho(x) \\ &= \mathcal{E}(S_1) \,.\end{aligned}$$

Now inserting the mean of R_2 instead of m_1 in the second term on the left shows that

$$\mathcal{E}(R_1) < \mathcal{E}(S_1) \,.$$

This contradicts optimality, so $m_1 = m_2$ cannot hold for a minimizing partition. □

Theorem 3.4. *Let ρ be a Borel measure on $\mathbb{R}^n$ with support S. If the partition $\{S_1, S_2\}$ is a minimizer for the mean-squared error, then the sets T_1 and T_2 in the Voronoi partition associated with the means $\{m(S_i)\}_{i=1}^2$ coincide with S_1 and S_2 up to changes involving subsets of the separating hyperplane or sets whose probability vanishes.*

Proof. We know $0 < \rho(S_1) < 1$, so both sets S_1 and S_2 have means under ρ.

Passing to the Voronoi partition $\{T_1, T_2\}$ associated with these means $\{m(S_i)\}_{i=1}^2$ gives

$$\mathcal{E}(T_1) = \mathcal{E}(S_1) \,.$$

Using the inequality in the definition of the Voronoi partition, we see that if $R_1 = T_1 \cap S_2$ is non-empty, then so is $R_2 = T_2 \cap S_1$, and $\|x - m_i\| \leq \min\{\|x - m_1\|, \|x - m_2\|\}$ if $x \in R_i \subset T_i$, $i \in \{1, 2\}$. Hence, denoting the hyperplane $H = \{x \in \mathbb{R}^n : \|x - m_1\| = \|x - m_2\|\}$, on $R_1 \setminus H$ and $R_2 \setminus H$ strict inequality holds in the norm bounds, and we see that by the monotonicity of integrals, the equality $\mathcal{E}(T_1) = \mathcal{E}(S_1)$ forces both sets to have probability zero, $\rho(R_1 \setminus H) = \rho(R_2 \setminus H) = 0$. □

From now on, we specialize to $\rho = (\sigma_{-1} + \sigma_1)/2$. As a first result for this concrete choice of ρ, we show that the mean-squared error does not increase when passing to a suitable partition into half-spaces that are separated by a hyperplane orthogonal to e_1.

To obtain this, we note that choosing a partition that separates into half-spaces with a separating hyperplane that contains the symmetry axis $\mathbb{R}e_1$ is not optimal. Without loss of generality, we orient this hyperplane so that it is orthogonal to e_2.

Lemma 3.5. *Let $n \geq 2$, $\rho = \frac{1}{2}(\sigma_{-1} + \sigma_1)$ be the measure defined on $\mathbb{R}^n$ with support S, $S_1 = S \cap \{x \in \mathbb{R}^n : x_2 \geq 0\}$ and $T_1 = S \cap \{x \in \mathbb{R}^n : x_1 \geq 0\}$, then $\mathcal{E}(S_1) > \mathcal{E}(T_1)$.*

Proof. By symmetry, the mean of S_1 is $m(S_1) = \alpha e_2$. Also, we know that the mean is in the interior of the convex hull of S_1, so $0 < \alpha < 1$. Again using the symmetry between S_1 and S_2 as well as $\rho(S_1) = \rho(S_2) = 1/2$,

$$\mathcal{E}(S_1) = 2\int_{S_1} \|x - \alpha e_2\|^2 d\rho = 2\int_{S_1} \|x\|^2 d\rho - \alpha^2 = \int_S \|x\|^2 d\rho - \alpha^2 \,.$$

Next, comparing with the Voronoi partition corresponding to $\{\pm e_1\}$ and using symmetry properties, we have

$$\mathcal{E}(S_1) = 2(\int_{T_1} \|x - e_1\|^2 + 1/2) - \alpha^2 = \mathcal{E}(T_1) + 1 - \alpha^2 \,.$$

From $0 < \alpha < 1$, we then have $\mathcal{E}(S_1) > \mathcal{E}(T_1)$. □

We are now ready to prove Theorem 2.2, which states that an optimal partition coincides, up to sets of measure zero, with one obtained from a separating hyperplane that is orthogonal to $\mathbb{R}e_1$.

Proof of Theorem 2.2. Given a partition of S by S_1 and S_2 with means $m_i = m(S_i)$, $i \in \{1, 2\}$, we observe the following:

The algebra of Borel sets of the form $A_1 \times \mathbb{R}^{n-1}$ with $A_1 \subset \mathbb{R}$, is a sub-algebra of the Borel algebra of $\mathbb{R}^n$. The functions that are measurable with respect to this algebra depend only on the first coordinate. By the Radon-Nikodym theorem, there exist functions $d_i : \mathbb{R} \to \mathbb{R}$ such that for any $A = A_1 \times \mathbb{R}^{n-1}$,

$$\int_A d_i(x_1)\chi_{S_i}(x)d\rho(x) = \int_A \|x - m_i\|^2 \chi_{S_i}(x) d\rho(x) \,.$$

Next, using Fubini, if μ is the image measure of ρ under projection onto the first coordinate, $\mu(A_1) = \rho(A_1 \times \mathbb{R}^{n-1})$, then there is $f : \mathbb{R} \to [0, 1]$ such that

$$\int_{A_1} d_1 f d\mu = \int_{A_1 \times \mathbb{R}^{n-1}} \|x - m_1\|^2 \chi_{S_1}(x) d\rho(x)$$

and

$$\int_{A_1} d_2(1 - f) d\mu = \int_{A_1 \times \mathbb{R}^{n-1}} \|x - m_2\|^2 (1 - \chi_{S_1}(x)) d\rho(x) \,.$$

Next, we observe if f is the function associated with a partition S_1 and S_2 and $R_1 = \{x \in \mathbb{R} : d_1(x) \leq d_2(x)\}$, then letting $g = \chi_{R_1}$ gives that

$$\int_{\mathbb{R}} d_1 g d\mu + \int_{\mathbb{R}} d_2(1 - g) d\mu \leq \int_{\mathbb{R}} d_1 f d\mu + \int_{\mathbb{R}} d_2(1 - f) d\mu \,.$$

We conclude, setting $T'_1 = S \cap (R_1 \times \mathbb{R}^{n-1})$ and $T'_2 = S \setminus T'_1$ that

$$\int \|x - m_1\|^2 \chi_{T'_1} d\rho + \int \|x - m_2\|^2 \chi_{T'_2} d\rho \leq \mathcal{E}(S_1) \,.$$

Next, replacing m_1 and m_2 by the means $m'_i \equiv m(T'_i)$, $i \in \{1, 2\}$, does not increase the left-hand side, which shows that

$$\mathcal{E}(T'_1) \leq \mathcal{E}(S_1) \,.$$

Finally, setting $\{T_1, T_2\}$ to be the Voronoi partition associated with the means m'_1 and m'_2 implies

$$\mathcal{E}(T_1) \leq \mathcal{E}(S_1) \,.$$

Moreover, if S_1 is chosen as a minimizer for the mean-squared error, then necessarily $m_i = m'_i$, $i \in \{1, 2\}$, otherwise we would have strict inequality between $\mathcal{E}(T'_1)$ and $\mathcal{E}(S_1)$. This implies that the means m_i are on the symmetry axis $\mathbb{R}e_1$. Applying Theorem 3.4 now shows that, up to a set of probability zero, S_1 and S_2 are separated by a hyperplane. From the preceding lemma, optimality implies that the hyperplane does not contain the symmetry axis. If it is not orthogonal to e_1, then there is a set $A_1 \subset \mathbb{R}$ such that $0 < \rho(A_1 \times \mathbb{R}^{n-1} \cap S_1) < \rho(A_1 \times \mathbb{R}^{n-1} \cap S)/2$ and hence there is a subset $B_1 \subset A_1$ with $\mu(B_1) > 0$ for which $f(B_1) \subset (0, 1/2)$. This contradicts optimality, because changing from f to the characteristic function g would lower the mean-squared error. We conclude that the hyperplane is orthogonal to e_1. □

3.2. **The optimal offset of the separating hyperplane.** From here on, we consider the dependence of the mean-squared error on the offset of the separating hyperplane.

We first introduce some additional notation. When the mean-square error is computed, the measure ρ can be replaced by an effective measure on $\mathbb{R}$ obtained from projecting onto the first coordinate.

We first consider the projection of σ_0. With the normalization constant

$$A_n := \left(\int_{-1}^{1} (1 - x^2)^{\frac{n-3}{2}} dx \right)^{-1} = \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{n-1}{2})},$$

the resulting measure μ_n on Borel sets in $[-1, 1]$ is given by [15]

$$d\mu_n(x) := A_n (1 - x^2)^{\frac{n-3}{2}} dx.$$

The probability that σ_0 assigns to $\{x \in \mathbb{S}_0 : x_1 \leq 1 - a\}$, $a \in [-1, 1]$, is equal to the probability of $\{x \in \mathbb{R} : x \leq 1 - a\}$ under μ_n,

$$M_n^-(a) := \int_{-1}^{1-a} d\mu_n(x).$$

This is the mass of part of the first sphere, obtained by a separating hyperplane between the two centers of the touching spheres, at a distance of $1 - a$ from the center of the first sphere. From the normalization convention, the total mass of the measure obtained from two spheres is two, so the complementary mass remaining is

$$M_n^+(a) := 2 - M_n^-(a).$$

The mean of the first piece is

$$C_n^-(a) := \frac{\int_{-1}^{1-a} x d\mu_n(x)}{M_n^-(a)},$$

and that of the second piece, relative to $C_n^-(0) = 0$, accordingly

$$C_n^+(a) := \frac{2 - \int_{-1}^{1-a} x d\mu_n(x)}{M_n^+(a)}.$$

With the help of Fubini-Tonelli, the integration over $\mathbb{R}^n$ giving the mean-squared error can be reduced to an integral with respect to μ_n. The contributions to the mean-squared error are split into 3 terms,

$$E_-(n,a) := \int_{-1}^{1-a} (1 - x^2 + (x - C_n^-(a))^2) d\mu_n(x),$$

$$E_\pm(n,a) := \int_{1-a}^{1} (1 - x^2 + (x - C_n^+(a))^2) d\mu_n(x),$$

and

$$E_+(n,a) := \int_{-1}^{1} (1 - x^2 + (2 + x - C_n^+(a))^2) d\mu_n(x).$$

In each of these cases the integrand is the squared distance of a point on either of the two spheres from the respective mean of the partition. The resulting mean-squared error is obtained by summing the three contributions and dividing by the total mass,

$$E(n,a) = \frac{1}{2}\big[E_-(n,a) + E_\pm(n,a) + E_+(n,a)\big].$$

Lemma 3.6. *Let $n \geq 2$ and $a \in [0,2]$, then $E(n,a)$ is expressed in terms of C_n^-, M_n^-, C_n^+, and M_n^+ according to*

$$E(n,a) = 3 - \frac{1}{2}\left((C_n^-(a))^2 M_n^-(a) + (C_n^+(a))^2 M_n^+(a)\right).$$

Proof. From normalization, we have the identities $\int_{-1}^{1} d\mu_n(x) = 1$ and $\int_{-1}^{1-a} d\mu_n(x) = 1 - \int_{1-a}^{1} d\mu_n(x)$; from symmetry, $\int_{-1}^{1} x d\mu_n(x) = 0$ and $\int_{-1}^{1-a} x d\mu_n(x) = -\int_{1-a}^{1} x d\mu_n(x)$. With the expression for $C_n^-(a)$ and $M_n^-(a)$,

$$\begin{aligned} E_-(n,a) &= M_n^-(a) - 2C_n^-(a) \int_{-1}^{1-a} x d\mu_n(x) + (C_n^-(a))^2 M_n^-(a) \\ &= M_n^-(a) - \left(C_n^-(a)\right)^2 M_n^-(a) \end{aligned}$$

The integrals in the other terms are converted similarly, including $C_n^+(a)$ and $M_n^+(a)$,

$$\begin{aligned} E_\pm(n,a) &= \int_{1-a}^{1} d\mu_n(x) - 2C_n^+(a) \int_{1-a}^{1} x d\mu_n(x) + (C_n^+(a))^2 \int_{1-a}^{1} d\mu_n(x) \\ &= 1 - M_n^-(a) + 2C_n^+(a) C_n^-(a) M_n^-(a) + (C_n^+(a))^2 (1 - M_n^-(a)) \\ &= 1 - M_n^-(a) + 2C_n^+(a) C_n^-(a) M_n^-(a) + (C_n^+(a))^2 (M_n^+(a) - 1). \end{aligned}$$

Because the last term is integrated over the entire sphere, the normalization and symmetry yield

$$\begin{aligned} E_+(n,a) &= \int_{-1}^{1} d\mu_n(x) - 2(2 - C_n^+(a)) \int_{-1}^{1} x d\mu_n(x) + (2 - C_n^+(a))^2 \int_{-1}^{1} d\mu_n(x) \\ &= 1 + (2 - C_n^+(a))^2 \\ &= 5 - 4C_n^+(a) + (C_n^+(a))^2. \end{aligned}$$

Adding together $E_-(n,a), E_\pm(n,a)$, and $E_+(n,a)$ and dividing by 2 gives, after collecting terms,

$$\begin{aligned} E(n,a) =& \frac{1}{2}\big[M_n^-(a) - \left(C_n^-(a)\right)^2 M_n^-(a) \\ & + 1 - M_n^-(a) + 2C_n^+(a)C_n^-(a)M_n^-(a) + (C_n^+(a))^2(M_n^+(a) - 1) \\ & + 5 - 4C_n^+(a) + (C_n^+(a))^2\big] \\ =& \frac{1}{2}\big[6 - \left(C_n^-(a)\right)^2 M_n^-(a) \\ & + 2C_n^+(a)C_n^-(a)M_n^-(a) + (C_n^+(a))^2 M_n^+(a) \\ & - 4C_n^+(a)\big]. \end{aligned}$$

We simplify further by converting between M_n^- and M_n^+,

$$\begin{aligned} E(n,a) =& \frac{1}{2}\big[6 - \left(C_n^-(a)\right)^2 M_n^-(a) \\ & + 2C_n^+(a)\left(2 - C_n^+(a)M_n^+(a)\right) + (C_n^+(a))^2 M_n^+(a) \\ & - 4C_n^+(a)\big] \\ =& \frac{1}{2}\big[6 - \left(C_n^-(a)\right)^2 M_n^-(a) \\ & - 2(C_n^+(a))^2 M_n^+(a) + (C_n^+(a))^2 M_n^+(a)\big]. \end{aligned}$$

Thus,

$$E(n,a) = 3 - \frac{1}{2}\left((C_n^-(a))^2 M_n^-(a) + (C_n^+(a))^2 M_n^+(a)\right).$$

□

Lemma 3.7. *The derivative* $\frac{\partial}{\partial a}E(n,a)$ *is expressed in terms of* M_n^-, M_n^+ *and* a *as*

$$\begin{aligned}\frac{\partial}{\partial a}E(n,a) =& \frac{2A_n(2a-a^2)^{\frac{n-3}{2}}}{(M_n^-(a)M_n^+(a))^2}\Big[(1-a)(M_n^-(a))^3 \\ &+ (2a-1)(M_n^-(a))^2 \\ &+ \frac{A_n}{n-1}(2-a)(2a-a^2)^{\frac{n-1}{2}}(M_n^-(a))^2 \\ &+ \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a) \\ &+ 2\frac{A_n}{n-1}(a-1)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) \\ &- \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\Big].\end{aligned}$$

Proof. Note that $\int_{-1}^{1-a} x d\mu_n(x) = -\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}$ by direct integration.

Differentiating term by term yields

$$\begin{aligned}\frac{\partial}{\partial a}E(n,a) =& -\frac{1}{2(M_n^-(a))^2}\Big[2\frac{A_n^2}{n-1}(1-a)(2a-a^2)^{n-2}M_n^-(a) \\ &+ \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}A_n(2a-a^2)^{\frac{n-3}{2}}\Big] \\ &- \frac{1}{2(M_n^+(a))^2}\Big[2\left(2+\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\right)A_n(1-a)(2a-a^2)^{\frac{n-3}{2}}M_n^+(a) \\ &- \left(2+\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\right)^2 A_n(2a-a^2)^{\frac{n-3}{2}}\Big] \\ =& -\frac{A_n(2a-a^2)^{\frac{n-3}{2}}}{2(M_n^-(a))^2}\Big[2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) \\ &+ \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\Big] \\ &- \frac{A_n(2a-a^2)^{\frac{n-3}{2}}}{2(M_n^+(a))^2}\Big[2\left(2+\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\right)(1-a)M_n^+(a) \\ &- \left(2+\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\right)^2\Big].\end{aligned}$$

Combining terms and simplifying gives

$$\begin{aligned}\frac{\partial}{\partial a}E(n,a) =& \frac{2A_n(2a-a^2)^{\frac{n-3}{2}}}{(M_n^-(a)M_n^+(a))^2}\Big[-2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a)\\ &-\left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\\ &+\left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a)\\ &+(2a-1)(M_n^-(a))^2\\ &+(1-a)(M_n^-(a))^3\\ &+\frac{A_n}{n-1}(2-a)(2a-a^2)^{\frac{n-1}{2}}(M_n^-(a))^2\Big].\end{aligned}$$

Finally, rearranging terms gives the claimed expression for $\frac{\partial}{\partial a}E(n,a)$.

□

To prove that for any fixed $n>3$, the function $a\mapsto E(n,a)$ is increasing for $a\in(0,2)$, it suffices to show that $\partial E(n,a)/\partial a$ is positive for all $a\in(0,2)$ and $n>3$. This will be the centerpiece of the proof of Theorem 2.5. To prepare this, we use the simplified expression for $\partial E(n,a)/\partial a$ given in the preceding lemma and find an estimate for M_n^- that is obtained by studying the monotonicity properties of the function $n\mapsto M_n^-(a)$ for a fixed.

Lemma 3.8. *The expression $M_n^-(a)$ is continuous in both $n\in[3,\infty)$ and $a\in[0,2]$, and $\frac{\partial}{\partial n}M_n^-(a)>0$ for $n>3$ and $a\in(0,1)$ (and negative for $n>3$ and $a\in(1,2)$).*

Proof. First, note that by Leibniz integral rule and integrability of $x^\alpha\ln x$, $\alpha>1$, at 0,

$$\frac{\partial}{\partial n}\int_{-1}^{1-a}(1-x^2)^{\frac{n-3}{2}}dx=\int_{-1}^{1-a}\frac{\partial}{\partial n}(1-x^2)^{\frac{n-3}{2}}dx=\int_{-1}^{1-a}\ln(1-x^2)(1-x^2)^{\frac{n-3}{2}}dx\,.$$

Thus, taking the partial derivative with respect to n, we obtain

$$\frac{\partial}{\partial n}M_n^-(a)=\int_{-1}^{1-a}\ln(1-x^2)d\mu_n(x)-\int_{-1}^{1-a}d\mu_n(x)\int_{-1}^{1}\ln(1-x^2)d\mu_n(x).$$

Consequently, we have $\frac{\partial}{\partial n}M_n^-(0)=\frac{\partial}{\partial n}M_n^-(1)=\frac{\partial}{\partial n}M_n^-(2)=0$. Next, we show that $\frac{\partial}{\partial n}M_n^-(a)>0$ for $a\in(0,1)$. To this end, we find critical points of $a\mapsto\frac{\partial}{\partial n}M_n^-(a)$.

By

$$\frac{\partial}{\partial a}\frac{\partial}{\partial n}M_n^-(a)=\frac{(2a-a^2)^{\frac{n-3}{2}}}{\int_{-1}^{1}(1-x^2)^{\frac{n-3}{2}}dx}\int_{-1}^{1}\left(\ln(1-x^2)-\ln(2a-a^2)\right)d\mu_n(x),$$

we have that $\frac{\partial}{\partial a}\frac{\partial}{\partial n}M_n^-(a) = 0$ if and only if

$$a \in \left\{0, 1 \pm \sqrt{1 - \exp\left(\int_{-1}^{1} \ln(1-x^2) d\mu_n(x)\right)}, 2\right\}.$$

Hence, for $a \in \left(0, 1 - \sqrt{1 - \exp\left(\int_{-1}^{1} \ln(1-x^2) d\mu_n(x)\right)}\right)$, we have $\frac{\partial}{\partial n}M_n^-(a)$ is increasing in a. To see this, take $0 < \omega \leq \int_{-1}^{1} \ln(1-x^2) d\mu_n(x)$, set $\omega^* = 1 - \sqrt{1-\exp(\omega)}$ and verify $\frac{\partial}{\partial a}\frac{\partial}{\partial n}M_n^-(\omega^*) > 0$. Similarly, for $a \in \left(1 - \sqrt{1 - \exp\left(\int_{-1}^{1} \ln(1-x^2) d\mu_n(x)\right)}, 1\right)$, we have $\frac{\partial}{\partial n}M_n^-(a)$ is decreasing in a. Therefore, by $\frac{\partial}{\partial n}M_n^-(0) = \frac{\partial}{\partial n}M_n^-(1) = 0$, we see $\frac{\partial}{\partial n}M_n^-(a) > 0$ for all $a \in (0,1)$. Repeating this for $a \in [1,2]$ gives $\frac{\partial}{\partial n}M_n^-(a) < 0$ for all $a \in (1,2)$. Thus we have shown $M_n^-(a)$ is increasing in $n > 3$ for $a \in (0,1)$ (and decreasing in $n > 3$ for $a \in (1,2)$). □

Corollary 3.1. For $a \in [0,1]$, we then have the inequalities $M_3^-(a) \leq M_n^-(a) \leq 1$ for all $n > 3$.

Lemma 3.9. *For all $n \geq 3$ and for all $a \in [1,2)$, $M_n^-(a) \geq \frac{A_n}{n-1}\left(2a-a^2\right)^{\frac{n-1}{2}}$.*

Proof. We make the change of variables $y = 1 + x$ with $dy = dx$, in $M_n^-(a) = \int_{-1}^{1-a} A_n(1-x^2)^{\frac{n-3}{2}}dx$ to obtain $M_n^-(a) = \int_0^{2-a} A_n(2y-y^2)^{\frac{n-3}{2}}dy$. Repeating integration by parts on parts $(2-y)^{\frac{n-3}{2}}$ and $y^{\frac{n-3}{2}}dy$ yields the formula

$$\int_0^{2-a} A_n\,(2-y)^{\frac{n-3}{2}}\,y^{\frac{n-3}{2}}dy = 2\frac{A_n}{n-1}\sum_{k=0}^{\infty}\left(\prod_{j=0}^{k}\frac{n-2j-1}{n+2j-1}\right)(2a-a^2)^{\frac{n-(2k+3)}{2}}(2-a)^{2k+1}.$$

By

$$\begin{aligned}
\left|\left(\prod_{j=0}^{K}\frac{n-2j-1}{n+2j-1}\right)\right| &= \left|(-1)^K\left(\prod_{j=0}^{K}\left(1-\frac{n-1}{j+\frac{1}{2}(n-1)}\right)\right)\right| \\
&\leq \left|\left(\prod_{j=0}^{K}\exp\left(-\frac{n-1}{j+\frac{1}{2}(n-1)}\right)\right)\right| \\
&= \left|\exp\left(-\sum_{j=0}^{K}\frac{n-1}{j+\frac{1}{2}(n-1)}\right)\right| \\
&\to 0
\end{aligned}$$

as $K \to \infty$, we see the alternating series converges. Moreover, since the first term is always positive, the sum converges to a function always greater than zero for $a \in (1,2)$ (by property of alternating series). Lastly, we see that for each odd $n \geq 3$, there are exactly

$\frac{n-1}{2}$ positive terms and for even $n \geq 4$, there are $\frac{n-2}{2}$ positive terms prior to a convergent alternating series (which starts at a positive term).
Consequently,

$$\begin{aligned} M_n^-(a) &\geq 2\frac{A_n}{n-1}(2a-a^2)^{\frac{n-3}{2}}(2-a) \\ &= \frac{2}{a}\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}, \end{aligned}$$

which is greater than or equal to $\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}$ (by maximizing the denominator for $a \in [1,2)$). ☐

Lemma 3.9 gives estimates on $M_n^-(a)$ that we combined with the expression for $E(n,a)$ and $\partial E(n,a)/\partial a$ from Lemmas 3.6 and 3.7 to show the main inequality $\frac{\partial}{\partial a}E(n,a) > 0$ for $a \in [1,2)$ in the proof of Theorem 2.5.

Proof of Theorem 2.5. We recall the simplified expressions

$$E(n,a) = 3 - \frac{1}{2}\left((C_n^-(a))^2 M_n^-(a) + (C_n^+(a))^2 M_n^+(a)\right)$$

and

$$\begin{aligned} \frac{\partial}{\partial a}E(n,a) =& \frac{2A_n(2a-a^2)^{\frac{n-3}{2}}}{(M_n^-(a)M_n^+(a))^2}\Big[(1-a)(M_n^-(a))^3 \\ &+ (2a-1)(M_n^-(a))^2 \\ &+ \frac{A_n}{n-1}(2-a)(2a-a^2)^{\frac{n-1}{2}}(M_n^-(a))^2 \\ &+ \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a) \\ &+ 2\frac{A_n}{n-1}(a-1)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) \\ &- \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\Big]. \end{aligned}$$

To show the desired inequality, we need only show that the factor in brackets is positive:

$$\begin{aligned} L(n,a) =& \left((1-a)M_n^-(a) + 2a - 1\right)\left(M_n^-(a)\right)^2 \\ &+ \left(\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) - 2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}\right)M_n^-(a) \\ &+ \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a) - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1} > 0 \end{aligned}$$

for $a \in (0,2)$ and $n > 3$.

We distinguish two cases, depending on the value of a.

Case I: If $a \in (0,1)$, by Corollary 3.1, we replace $M_n^-(a)$ with $M_3^-(a) = \frac{1}{2}(2-a)$ for all positive terms. That is,

$$
\begin{aligned}
L(n,a) \geq & \big((1-a)M_n^-(a) + 2a - 1\big)\left(M_n^-(a)\right)^2 \\
& + \frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\left(\frac{1}{2}(2-a)\right)^2 \\
& - 2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) \\
& + \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\left(\frac{1}{2}(2-a)\right) - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}.
\end{aligned}
$$

Moreover, we see $(1-a)M_n^-(a) + 2a - 1 \geq (1-a)M_3^-(a) + 2a - 1 = a/2 + a^2/2 \geq 0$.
Hence, the first term can be estimated as well by eliminating $M_n^-(a)$, resulting in the lower bound

$$
\begin{aligned}
L(n,a) \geq & \left(\frac{a}{2} + \frac{a^2}{2}\right)\left(\frac{1}{2}(2-a)\right)^2 \\
& + \frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\left(\frac{1}{2}(2-a)\right)^2 \\
& - 2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) \\
& + \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\left(\frac{1}{2}(2-a)\right) \\
& - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}.
\end{aligned}
$$

By Lemma 3.8, we also have that $M_n^-(a) \leq M_n^-(0) = 1$.
Using this estimate for the remaining negative factor multiplying M_n^- gives a further lower

bound from which all quantities other than a have been eliminated,

$$\begin{aligned}
L(n,a) \geq & \frac{1}{8}(1+a)a(2-a)^2 \\
& + \frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}\left(\frac{1}{2}(2-a)\right)^2 \\
& - 2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}} \\
& + \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}\left(\frac{1}{2}(2-a)\right) \\
& - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1} \\
= & \frac{1}{8}(1+a)a(2-a)^2 \\
& + \frac{1}{8}\frac{A_n}{n-1}(2a^4-8a^3+24a^2-16a)(2a-a^2)^{\frac{n-1}{2}} \\
& - \frac{1}{2}\left(\frac{A_n}{n-1}\right)^2 a(2a-a^2)^{n-1}.
\end{aligned}$$

Finally, by the second and third term decreasing in $a \in (0,1)$, we have

$$\begin{aligned}
L(n,a) &\geq \frac{1}{8}(1+a)a(2-a)^2 + \frac{1}{4}\frac{A_n}{n-1} - \frac{1}{2}\left(\frac{A_n}{n-1}\right)^2 \\
&= \frac{1}{8}\left((1+a)a(2-a)^2 + \frac{2A_n}{n-1} - \left(\frac{2A_n}{n-1}\right)^2\right) \\
&\geq \frac{1}{8}(1+a)a(2-a)^2 > 0.
\end{aligned}$$

Consequently for $a \in (0,1]$, $\frac{\partial}{\partial a}E(n,a) > 0$.

Case II: If $a \in [1,2)$, we re-examine $L(n,a)$ and apply Lemma 3.9.

By the inequality

$$\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}} \geq \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1},$$

we have

$$
\begin{aligned}
L(n,a) &= \big((1-a)M_n^-(a)+2a-1\big)\left(M_n^-(a)\right)^2 \\
&\quad + \left(\frac{A_n}{n-1}(2a-a^2)^{\frac{n-1}{2}}M_n^-(a) - 2\frac{A_n}{n-1}(1-a)(2a-a^2)^{\frac{n-1}{2}}\right)M_n^-(a) \\
&\quad + \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a) - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1} \\
&\geq \big((1-a)M_n^-(a)+2a-1\big)\left(M_n^-(a)\right)^2 \\
&\quad + \left(\frac{A_n}{n-1}\right)^2(2-a)(2a-a^2)^{n-1}\left(M_n^-(a)\right)^2 + \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1}M_n^-(a) \\
&\quad + 2\left(\frac{A_n}{n-1}\right)^2(a-1)(2a-a^2)^{n-1}M_n^-(a) - \left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1} \\
&= \big((1-a)M_n^-(a)+2a-1\big)\left(M_n^-(a)\right)^2 \\
&\quad + \left(\frac{A_n}{n-1}\right)^2(2a)\left[M_n^-(a) - \frac{1}{2}\left(M_n^-(a)\right)^2\right](2a-a^2)^{n-1} \\
&\quad + \left(\frac{A_n}{n-1}\right)^2\left[2\left(M_n^-(a)\right)^2 - M_n^-(a) - 1\right](2a-a^2)^{n-1}.
\end{aligned}
$$

Using Lemma 3.9 in the last inequality and recalling that if $a \in (1,2)$, then $M_n^-(a) < M_n^-(1) = 1/2$, we further estimate

$$
(1-a)M_n^-(a)+2a-1 > \frac{3}{2}a - \frac{1}{2} > 0
$$

which gives

$$
\begin{aligned}
L(n,a) &\geq \big((1-a)M_n^-(a)+2a-1\big)\left(\frac{A_n}{n-1}\right)^2(2a-a^2)^{n-1} \\
&\quad + \left(\frac{A_n}{n-1}\right)^2(2a)\left[M_n^-(a) - \frac{1}{2}\left(M_n^-(a)\right)^2\right](2a-a^2)^{n-1} \\
&\quad + \left(\frac{A_n}{n-1}\right)^2\left[2\left(M_n^-(a)\right)^2 - M_n^-(a) - 1\right](2a-a^2)^{n-1}.
\end{aligned}
$$

Thus, combining terms, we obtain a lower bound

$$\begin{aligned} L(n,a) \geq \left(\frac{A_n}{n-1}\right)^2 \Big[&(1-a)M_n^-(a) + 2a - 1 \\ &+ 2aM_n^-(a) - a\left(M_n^-(a)\right)^2 \\ &+ 2\left(M_n^-(a)\right)^2 - M_n^-(a) - 1\Big](2a-a^2)^{n-1} \\ = \left(\frac{A_n}{n-1}\right)^2 &\left[aM_n^-(a) + 2(a-1) + (2-a)\left(M_n^-(a)\right)^2\right](2a-a^2)^{n-1}, \end{aligned}$$

consisting of strictly positive terms if $1 < a < 2$.

Consequently, we see for $n > 3$ and $a \in (1,2)$,

$$\frac{\partial}{\partial a} E(n,a) > 0$$

We conclude that for $a \in (0,2)$ and $n > 3$, $E(n,a)$ is strictly increasing, thus attaining its unique minimum at $a = 0$. □

References

[1] J. A. Bucklew and G. L. Wise, Multidimensional asymptotic quantization theory with r^{th} power distortion measures, *IEEE Trans. Inform. Theory* **28**(2):239–247, 1982.

[2] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[3] S. Dasgupta, Learning mixtures of Gaussians, in: *40th Annual Symposium on Foundations of Computer Science 1999*, pages 634–644. IEEE, 1999.

[4] Q. Du, V. Faber, and M. Gunzburger, Centroidal Voronoi tessellations: Applications and algorithms, *SIAM Review*, 41(4):637–676, 1999.

[5] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Springer International Series in Engineering and Computer Science, Springer, **159**, Berlin, pp. 732, 1991.

[6] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*. Lecture Notes in Math. 1730. Springer, Berlin, pp. 203, 2000.

[7] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. On the tightness of an SDP relaxation of k-means, *arXiv preprint arXiv:1505.04778*, 2015.

[8] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. Probably certifiably correct k-means clustering, *Mathematical Programming*, pages 1–38, 2015.

[9] J. C. Kieffer, Exponential rate of convergence for Lloyd's method I, *IEEE Trans. on Inform. Theory*, Special issue on quantization, **28**(2):205–210, 1982.

[10] S. P. Lloyd, Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28**(2):129–137, 1982.

[11] X. Li, Y. Li, S. Ling, T. Strohmer, and K. Wei, When Do Birds of a Feather Flock Together? K-Means, Proximity, and Conic Programming, *arXiv preprint arXiv:1710.06008*, 2017.

[12] Y. Lu and H. H. Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants, *arXiv preprint arXiv:1612.02099*, 2016.

[13] D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, 2003.

[14] D. G. Mixon, S. Villar and R. Ward, Clustering subgaussian mixtures with k-means, in: 2016 IEEE Information Theory Workshop (ITW), Cambridge, 2016, pp. 211-215.

[15] C. E. Mueller and F. B. Weissler, Hypercontractivity for the heat semigroup for ultraspherical polynomials and on the n-sphere, *Journal of Functional Analysis* **48** (2): 252–283, 1982.

[16] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming, *SIAM Journal on Optimization*, 18(1):186–205, 2007.
[17] D. Pollard, A central limit theorem for k-means clustering, *Ann. Probab.*, **10**, 919–926, 1982.
[18] M. K. Roychowdhury, Optimal quantizers for some absolutely continuous probability measures, *arXiv preprint arXiv:1608.03815*, 2016.
[19] S. Z. Selim and M. A. Ismail. k-means-type algorithms: A generalized convergence theorem and characterization of local optimality, *IEEE Transactions on pattern analysis and machine intelligence*, 6(1):81–87, 1984.
[20] H. Steinhaus, Sur la division des corps matériels en parties, *Bull. Acad. Polon. Sci.* **4** (12): 801–804, 1957.
[21] A. Vattani. k-means requires exponentially many iterations even in the plane, *Discrete and Computational Geometry*, 45(4):596–616, 2011.

Mathematics Department, University of Houston, Houston, TX 77204-3008.
E-mail address: bgb@math.uh.edu

Mathematics Department, University of Houston, Houston, TX 77204-3008.
E-mail address: cm4il243@gmail.com