Information and Inference: A Journal of the IMA (2017) **6**, 389–415
doi:10.1093/imaiai/iax001
Advance Access Publication on 17 March 2017

Clustering subgaussian mixtures by semidefinite programming

DUSTIN G. MIXON
Department of Mathematics and Statistics, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA

AND

SOLEDAD VILLAR* AND RACHEL WARD
Department of Mathematics, University of Texas at Austin, Austin, Texas, USA
*Corresponding author: mvillar@math.utexas.edu

[Received on 10 May 2016; revised on 16 December 2016; accepted on 17 December 2016]

We introduce a model-free relax-and-round algorithm for k-means clustering based on a semidefinite relaxation due to Peng and Wei (2007, *SIAM J. Optim.*, **18**, 186–205). The algorithm interprets the output of the semidefinite program as a denoised version of the original data and then rounds this output to a hard clustering. We provide a generic method for proving performance guarantees for this algorithm, and we analyse the algorithm in the context of subgaussian mixture models. We also study the fundamental limits of estimating Gaussian centers by k-means clustering to compare our approximation guarantee to the theoretically optimal k-means clustering solution.

Keywords: clustering; machine learning; semidefinite programming; approximation algorithm.

1. Introduction

Consider the following mixture model: for each $t \in [k] := \{1, \ldots, k\}$, let $\mathcal{D}_t$ be an unknown subgaussian probability distribution over $\mathbb{R}^m$, with first moment $\gamma_t \in \mathbb{R}^m$ and second moment matrix with largest eigenvalue σ_t^2. For each t, an unknown number n_t of random points $\{x_{t,i}\}_{i\in[n_t]}$ is drawn independently from $\mathcal{D}_t$. Given the points $\{x_{t,i}\}_{i\in[n_t],t\in[k]}$ along with the model order k, the goal is to approximate the centers $\{\gamma_t\}_{t\in[k]}$. How large must $\Delta := \min_{a\neq b} \|\gamma_a - \gamma_b\|_2$ be relative to $\sigma_{\max} := \max_t \sigma_t$, and how large must $n_{\min} := \min_t n_t$ be relative to $n_{\max} := \max_t n_t$, to have sufficient signal for successful approximation?

For the most popular instance of this problem, where the subgaussian distributions are Gaussians, theoretical guarantees date back to the work of Dasgupta [13]. Dasgupta introduced an algorithm based on random projections, and showed that this algorithm well approximates centers of Gaussians in $\mathbb{R}^m$ that are separated by $\sigma_{\max}\sqrt{m}$. Since Dasgupta's seminal work, improved performance guarantees for several algorithmic alternatives have emerged, including expectation maximization [14], spectral methods [6,10–12,18,26], projections (random and deterministic) [4,20] and the method of moments [20]. Every existing performance guarantee has one of two forms:

(a) the algorithm correctly clusters all points according to Gaussian mixture component or

(b) the algorithm well approximates the center and covariance matrix of each Gaussian (a la Dasgupta [13]).

Results of type (a), which include [3,6,18,26], require the minimum separation between the Gaussians centers to have a multiplicative factor of $\operatorname{polylog} N$, where $N = \sum_{t=1}^k n_t$ is the total number of points.

This stems from a requirement that every point be closer to their Gaussian's center (in some sense) than the other centers, so that the problem of cluster recovery is well posed. We note that in the case of spherical Gaussians, such highly separated Gaussian components may be truncated so as to match a different data model known as the stochastic ball model, where the semidefinite program we use in this article is already known to be tight with high probability [5,16].

Most results of type (b) tend to be specifically tailored to exploit unique properties of the Gaussian distribution, and are thus not easily generalizable to other data models. For instance, if x has distribution $\mathcal{N}(\mu, \sigma^2 I_m)$ then $\mathbb{E}(\|x-\mu\|^2) = m\sigma^2$, and concentration of measure implies that in high dimensions, most of the points will reside in a thin shell with center μ and radius about $\sqrt{m}\sigma$. This sort of behavior can be exploited to cluster even concentric Gaussians as long as the covariances are sufficiently different [8,20]. However, algorithms that perform well even with no separation between centers require a sample complexity which is exponential in k [20]. Some results of type (b) like [2] and [17] take into consideration more general data models, in particular log-concave distributions and approximate the mixture centers for minimum separation of $k + \sqrt{k \log n}$ and $k^{3/2}$, respectively.

In this article, we provide a performance guarantee of type (b), but our approach is model free. In particular, we consider the k-means clustering objective:

$$\begin{aligned} &\text{minimize} \quad \sum_{t=1}^{k} \sum_{i \in A_t} \left\| x_i - \frac{1}{|A_t|} \sum_{j \in A_t} x_j \right\|_2^2 \\ &\text{subject to} \quad A_1 \cup \cdots \cup A_k = \{1, \ldots, N\}, \quad A_i \cap A_j = \emptyset \quad \forall i, j \in [k],\ i \neq j. \end{aligned} \tag{1}$$

Letting D denote the $N \times N$ matrix defined entrywise by $D_{ij} = \|x_i - x_j\|_2^2$, then a straightforward calculation gives the following 'lifted' expression for the k-means objective:

$$\sum_{t=1}^{k} \sum_{i \in A_t} \left\| x_i - \frac{1}{|A_t|} \sum_{j \in A_t} x_j \right\|_2^2 = \frac{1}{2} \operatorname{Tr}(DX), \qquad X_{ij} = \begin{cases} \frac{1}{|A_t|} & \text{if } i, j \in A_t \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The matrix X necessarily satisfies various convex constraints, and relaxing to such constraints leads to the following semidefinite relaxation of (1), first introduced by Peng and Wei [23]:

$$\begin{aligned} &\text{minimize} \quad \operatorname{Tr}(DX) \\ &\text{subject to} \quad \operatorname{Tr}(X) = k,\ X1 = 1,\ X \geq 0,\ X \succeq 0. \end{aligned} \tag{3}$$

Here, $X \geq 0$ means that X is entrywise non-negative, whereas $X \succeq 0$ means that X is symmetric and positive semidefinite.

As mentioned earlier, this semidefinite relaxation is known to be tight for a particular data model called the stochastic ball model [5,16,21]. In this article, we study its performance under subgaussian mixture models, which include the stochastic ball model and the Gaussian mixture model as special cases. The semidefinite relaxation (SDP) is not typically tight under this general model, but the optimizer can be interpreted as a denoised version of the data and can be rounded to produce a good estimate for the centers (and therefore produce a good clustering).

To see this, let P denote the $m \times N$ matrix whose columns are the points $\{x_{t,i}\}_{i \in [n_t], t \in [k]}$. Notice that whenever X has the form (2), then for each $t \in [k]$, PX has $|A_t|$ columns equal to the centroid of points assigned to A_t. In particular, if X is k-means-optimal then PX reports the k-means-optimal centroids (with

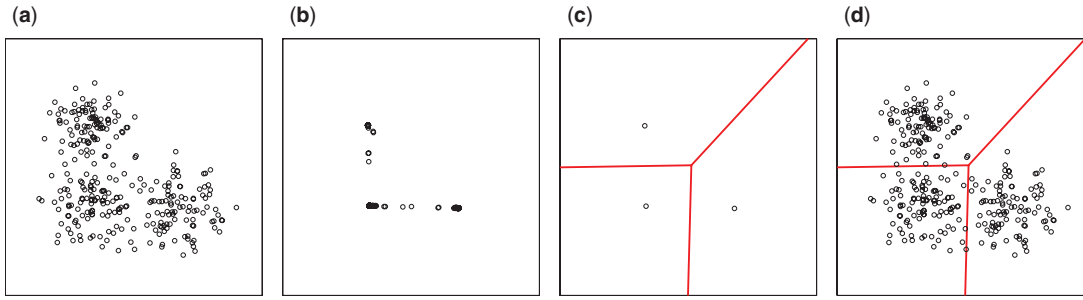

FIG. 1. **(a)** Draw 100 points at random from each of three spherical Gaussians over $\mathbb{R}^2$. These points form the columns of a 2×300 matrix P. **(b)** Compute the 300×300 distance-squared matrix D from the data in (a), and solve the k-means semidefinite relaxation (3) using SDPNAL+v0.3 [28]. (The computation takes about 16 s on a standard MacBook Air laptop.) Given the optimizer X, compute PX and plot the columns. We interpret this as a denoised version of the original data P. **(c)** The points in (b) land in three particular locations with particularly high frequency. Take these locations to be estimators of the centers of the original Gaussians. **(d)** Use the estimates for the centers in (c) to partition the original data into three subsets, thereby estimating the k-means-optimal partition.

appropriate multiplicities). Next, we note that every SDP-feasible matrix $X \geq 0$ satisfies $X^\top 1 = X1 = 1$ and so $X^\top$ is a stochastic matrix, meaning each column of PX is still a weighted average of columns from P. Intuitively, if the SDP relaxation (3) were close to being tight, then the SDP-optimal X would make the columns of PX close to the k-means-optimal centroids. Empirically, this appears to be the case (see Fig. 1 for an illustration). Overall, we may interpret PX as a denoised version of the original data P, and we leverage this strengthened signal to identify good estimates for the k-means-optimal centroids.

What follows is a summary of our relax-and-round procedure for (approximately) solving the k-means problem (1):

Relax-and-round k-means clustering procedure.
Given and $m \times N$ data matrix $P = [x_1 \cdots x_N]$, do:

(i) Compute distance-squared matrix D defined by $D_{ij} = \|x_i - x_j\|_2^2$.
(ii) Solve k-means semidefinite program (3), resulting in optimizer X.
(iii) Cluster the columns of the denoised data matrix PX.

For step (iii), we find there tends to be k vectors that appear as columns in PX with particularly high frequency, and so we are inclined to use these as estimators for the k-mean-optimal centroids (see Fig. 1, for example). Running Lloyd's algorithm for step (iii) also works well in practice. To obtain theoretical guarantees, we instead find the k columns of PX for which the unit balls of a certain radius centered at these points in $\mathbb{R}^m$ contain the most columns of PX (see Theorem 14 for more details). An implementation of our procedure is available on GitHub [25].

Our contribution. We study performance guarantees for the k-means semidefinite relaxation (3) when the point cloud is drawn from a subgaussian mixture model. Adapting ideas from Guédon and Vershynin [15], we obtain approximation guarantees comparable with the state of the art for learning mixtures of Gaussians, despite the fact that our algorithm is a generic k-means solver and uses no

model assumptions. To the best of our knowledge, no convex relaxation has been used before to provide theoretical guarantees for clustering mixtures of Gaussians. We also provide conditional lower bounds on how well a k-means solver can approximate the centers of Gaussians by (conditionally) quantifying how biased the k-means solution is as an estimator for Gaussian centers .

Organization of this article. In Section 2, we present a summary of our results and give a high-level explanation of our proof techniques. We also illustrate the performance of our relax-and-round algorithm on the MNIST dataset of handwritten digits. Our theoretical results consist of an approximation theorem for the SDP (proved in Section 3), a denoising consequence of the approximation (explained in Section 4) and a rounding step (presented in Section 5). We also study the fundamental limits for estimating Gaussian centers by k-means clustering (see Section 2.2).

2. Summary of results

This article has two main results. First, we present a relax-and-round algorithm for k-means clustering that well approximates the centers of sufficiently separated subgaussians. Second, we provide a conditional result on the minimum separation necessary for Gaussian center approximation by k-means clustering. The first result establishes that the k-means SDP (3) performs well with noisy data (despite not being tight), and the second result helps to illustrate how sharp our analysis is. This section discusses these results and then applies our algorithm to the MNIST dataset [**19**].

2.1 *Performance guarantee for the k-means SDP*

Our relax-and-round performance guarantee consists of three steps.

Step 1: Approximation. We adapt an approach used by Guédon and Vershynin to provide approximation guarantees for a certain semidefinite program under the stochastic block model for graph clustering [**15**].

Given the points $x_{t,1},\dots,x_{t,n_t}$ drawn independently from $\mathcal{D}_t$, consider the squared-distance matrix D and the corresponding minimizer X_D of the SDP (3). We first construct a 'reference' matrix R such that the SDP (3) is tight when $D = R$ with optimizer X_R. To this end, take $\Delta_{ab} := \|\gamma_a - \gamma_b\|_2$, let X_D denote the minimizer of (3) and let X_R denote the minimizer of (3) when D is replaced by the reference matrix R defined as

$$(R_{ab})_{ij} := \xi + \Delta_{ab}^2/2 + \max\left\{0, \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle\right\}, \tag{4}$$

where $r_{t,i} := x_{t,i} - \gamma_t$, and $\xi > 0$ is a parameter to be chosen later. Indeed, this choice of reference is quite technical, as an artifact of the entries in D being statistically dependent. Despite its lack of beauty, our choice of reference enjoys the following important property:

LEMMA 1 Let $1_a \in \mathbb{R}^N$ denote the indicator function for the indices i corresponding to points x_i drawn from the ath subgaussian. If $\gamma_a \neq \gamma_b$ whenever $a \neq b$, then $X_R = \sum_{t=1}^k (1/n_t) 1_t 1_t^\top$.

Proof. Let X be feasible for the SDP (3). Replacing D with R in (3), we may use the SDP constraints $X1 = 1$ and $X \geq 0$ to obtain the bound

$$\mathrm{Tr}(RX) = \sum_{i=1}^N \sum_{j=1}^N R_{ij} X_{ij} \geq \sum_{i=1}^N \sum_{j=1}^N \xi X_{ij} = \sum_{i=1}^N \xi \sum_{j=1}^N X_{ij} = N\xi = \mathrm{Tr}(RX_R).$$

Furthermore, since $\gamma_a \neq \gamma_b$ whenever $a \neq b$, and since $X \geq 0$, we have that equality occurs precisely for the X such that $(X_{ab})_{ij}$ equals zero whenever $a \neq b$. The other constraints on X then force X_R to have the claimed form (i.e., X_R is the unique minimizer). □

The previous lemma shows X_R is the planted solution of our clustering problem. It remains to demonstrate regularity in the sense that $X_D \approx X_R$ provided the subgaussian centers are sufficiently separated. For this, we use the following scheme:

- If $\langle R, X_D\rangle \approx \langle R, X_R\rangle$ then $\|X_D - X_R\|_F^2$ is small (Lemma 7).
- If $D \approx R$ (in some specific sense) then $\langle R, X_D\rangle \approx \langle R, X_R\rangle$ (Lemmas 8 and 9).
- If the centers are separated by $O(k\sigma_{\max})$, then $D \approx R$.

What follows is the result of this analysis:

THEOREM 2 Given $x_1, \ldots, x_N$ points drawn independently from a mixture of k subgaussian distributions in $\mathbb{R}^m$. Say that the subgaussian a, for $1 \leq a \leq k$ has center γ_a, maximum covariance σ_a^2 and n_a points have been drawn from it. Let $n_{\max} := \max_{1\leq a\leq k} n_a$ and, similarly, $n_{\min}$ and $\sigma_{\max}$. Let X_D the minimizer of (3) for D such that $D_{ij} = \|x_i - x_j\|^2$ and X_R the minimizer of (3) for $D = R$ the reference matrix defined in (4) (which coincides with the planted clusters as a consequence of Lemma 1). Fix $\epsilon, \eta > 0$. There exist universal constants C, c_1, c_2, c_3 such that if

$$\alpha = n_{\max}/n_{\min} \lesssim k \lesssim m \quad \text{and} \quad N > \max\{c_1 m, c_2 \log(2/\eta), \log(c_3/\eta)\}$$

then $\|X_D - X_R\|_{\mathrm{F}}^2 \leq \epsilon$ with probability $\geq 1 - 2\eta$ provided

$$\Delta_{\min}^2 \geq \frac{C}{\epsilon} k^2 \alpha \sigma_{\max}^2,$$

where $\Delta_{\min} = \min_{a\neq b} \|\gamma_a - \gamma_b\|_2$ is the minimal cluster center separation.

See Section 3 for the proof. Note that if we remove the assumption $\alpha \lesssim k \lesssim m$, we obtain the result $\Delta_{\min}^2 \geq \frac{C}{\epsilon}(\min\{k, m\} + \alpha) k\alpha\sigma_{\max}^2$.

Step 2: Denoising. Suppose we solve the SDP (3) for an instance of the subgaussian mixture model, where $\Delta_{\min}$ is sufficiently large. Then Theorem 2 ensures that the solution X_D is close to the ground truth. At this point, it remains to convert X_D into an estimate for the centers $\{\gamma_t\}_{t\in[k]}$. Let P denote the $m \times N$ matrix whose (a, i)th column is $x_{a,i}$. Then PX_R is an $m \times N$ matrix whose (a, i)th column is $\tilde{\gamma}_a$, the centroid of the ath cluster, which converges to γ_a as $N \to \infty$ (and does not change when i varies, for a fixed a), and so one might expect PX_D to have its columns be close to the γ_t's. In fact, we can interpret the columns of PX_D as a denoised version of the points (see Fig. 1).

To illustrate this idea, assume the points $\{x_{a,i}\}_{i\in[n]}$ come from $\mathcal{N}(\gamma_a, \sigma^2 I_m)$ in $\mathbb{R}^m$ for each $a \in [k]$. Then we have

$$\mathbb{E}\left[\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n} \|x_{a,i} - \gamma_a\|_2^2\right] = m\sigma^2. \tag{5}$$

Letting $c_{a,i}$ denote the (a,i)th column of PX_D (i.e., the ith estimate of γ_a), in Section 4 we obtain the following denoising result:

COROLLARY 3 If $k\sigma \lesssim \Delta_{\min} \leq \Delta_{\max} \lesssim K\sigma$, then

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 \lesssim K^2\sigma^2$$

with high probability as $n \to \infty$.

Note that Corollary 3 guarantees denoising in the regime $K \ll \sqrt{m}$. This is a corollary of a more technical result (Theorem 10), which guarantees denoising for certain configurations of subgaussians (e.g., when the γ_t's are vertices of a regular simplex) in the regime $k \ll m$.

At this point, we comment that one might expect this level of denoising from principal component analysis (PCA) when the mixture of subgaussians is sufficiently nice. To see this, suppose we have spherical Gaussians of equal entrywise variance σ^2 centered at vertices of a regular simplex. Then in the large-sample limit, we expect PCA to approach the $(k-1)$-dimensional affine subspace that contains the k centers. Projecting onto this affine subspace will not change the variance of any Gaussian in any of the principal components, and so one expects the mean squared deviation of the projected points from their respective Gaussian centers to be $(k-1)\sigma^2$.

In contrast, we find that in practice, the SDP denoises substantially more than PCA does. For example, Figs 1 and 2 illustrate cases in which PCA would not change the data, since the data already lies in $(k-1)$-dimensional space, and yet the SDP considerably enhances the signal. In fact, we observe empirically that the matrix X_D has low rank and that PX_D has repeated columns. This does not come as a complete surprise, considering SDP optimizers are known to exhibit low rank [7,22,24]. Still, we observe that the

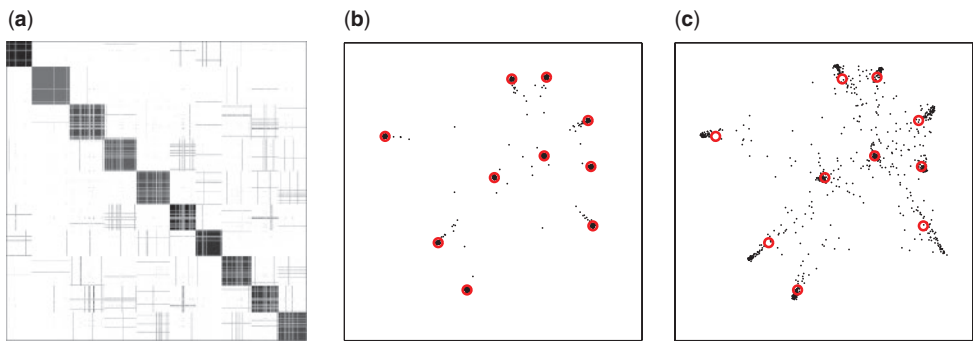

FIG. 2. **(a)** After applying TensorFlow [1] to learn a nine dimensional feature space of MNIST digits [19], determine the features of the first 1,000 images in the MNIST test set, compute the $1,000 \times 1,000$ matrix D of squared distances in feature space and then solve the k-means semidefinite relaxation (3) using SDPNAL+v0.3 [28]. (The computation takes about 6 min on a standard MacBook Air laptop.) Convert the SDP-optimizer X to a grayscale image such that white pixels denote zero entries. By inspection, this matrix is not exactly of the form (2), but it looks close, and it certainly appears to have low rank. **(b)** Letting P denote the $9 \times 1,000$ matrix whose columns are the feature vectors to cluster, compute the denoised data PX and identify the 10 most popular locations in $\mathbb{R}^9$ (denoted by larger circles) among the columns of PX (denoted by smaller dots). For the plot, we project the nine-dimensional data onto a random two-dimensional subspace. **(c)** The 10 most popular locations form our estimates for the centers of digits in feature space. We plot these locations relative to the original data, projected in the same 2-dimensional subspace as (b).

optimizer tends to have rank $O(k)$ when clustering points from the mixture model. This is not predicted by existing bounds, and we did not leverage this feature in our analysis, though it certainly warrants further investigation.

Step 3: Rounding. In Section 5, we present a rounding scheme that provides a clustering of the original data from the denoised results of the SDP (Theorem 14). In general, the cost of rounding is a factor of k in the average squared deviation of our estimates. Under the same hypothesis as Corollary 3, we have that there exists a permutation π on $\{1,\dots,k\}$ such that

$$\frac{1}{k}\sum_{i=1}^{k}\|v_i-\tilde{\gamma}_{\pi(i)}\|_2^2 \lesssim kK^2\sigma^2, \tag{6}$$

where v_i is what our algorithm chooses as the ith center estimate. Much like the denoising portion, we also have a more technical result that allows one to replace the right-hand side of (6) with $k^2\sigma^2$ for sufficiently nice configurations of subgaussians. As such, we can estimate Gaussian centers with mean squared error $O(k^2\sigma^2)$ provided the centers have pairwise distance $\Omega(k\sigma)$. In the next section, we indicate that model order dependence cannot be completely removed when using k-means to estimate the centers.

Before concluding this section, we want to clarify the nature of our approximation guarantee (6). Since centroids correspond to a partition of Euclidean space, our guarantee says something about how 'close' our k-means partition is to the 'true' partition. In contrast, the usual approximation guarantees for relax-and-round algorithms compare values of the objective function (e.g., the k-means value of the algorithm's output is within a factor of 2 of minimum). Also, the latter sort of optimal value-based approximation guarantee cannot be used to produce the sort of optimizer-based guarantee we want. To illustrate this, imagine a relax-and-round algorithm for k-means that produces a near-optimal partition with $k=2$ for data coming from a single spherical Gaussian. We expect every subspace of co-dimension 1 to separate the data into a near-optimal partition, but the partitions are very different from each other when the dimension $m\geq 2$, and so a guarantee of the form (6) will not hold.

2.2 *Fundamental limits of k-means clustering*

In Section 5, we provide a rounding scheme that, when applied to the output of the k-means SDP, produces estimates of the subgaussian centers. But how good is our guarantee? Observe the following two issues: (i) the amount of tolerable noise σ and our bound on the error $\max_i\|v_i-\tilde{\gamma}_{\pi(i)}\|_2$ both depend on k. (ii) Our bound on the error does not vanish with N.

In this section, we give a conditional result that these issues are actually artifacts of k-means; that is, both of these would arise if one were to estimate the Gaussian centers with the k-means-optimal centroids (though these centroids might be computationally difficult to obtain). The main trick in our argument is that, in some cases, the so-called 'Voronoi means' to appear to serve as good a proxy for the k-means-optimal centroids. This trick is useful because the Voronoi means are much easier to analyse. We start by providing some motivation for the Voronoi means.

Given $\mathcal{X}=\{x_i\}_{i=1}^N\subseteq\mathbb{R}^m$, let $A_1^{(\mathcal{X})}\sqcup\cdots\sqcup A_k^{(\mathcal{X})}=\{1,\dots,N\}$ denote any minimizer of the k-means objective

$$\sum_{t=1}^{k}\sum_{i\in A_t}\left\|x_i-\frac{1}{|A_t|}\sum_{j\in A_t}x_j\right\|_2^2$$

and define the ***k*-means-optimal centroids** by

$$c_t^{(\mathcal{X})} := \frac{1}{|A_t^{(\mathcal{X})}|} \sum_{j \in A_t^{(\mathcal{X})}} x_j.$$

(Note that the k-means minimizer is unique for generic $\mathcal{X}$.) Given $\Gamma = \{\gamma_t\}_{t=1}^k$, then for each γ_a, consider the Voronoi cell

$$V_a^{(\Gamma)} := \left\{x \in \mathbb{R}^m : \|x - \gamma_a\|_2 < \|x - \gamma_b\|_2 \ \forall b \neq a\right\}.$$

Given a probability distribution $\mathcal{D}$ over $\mathbb{R}^m$, define the **Voronoi means** by

$$\mu_t^{(\Gamma,\mathcal{D})} := \mathop{\mathbb{E}}_{X \sim \mathcal{D}} \left[X \middle| X \in V_t^{(\Gamma)}\right].$$

With this background, we formulate the following conjecture:

CONJECTURE 4 (Voronoi means conjecture) Let $\Gamma = \{(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), (0, 0, \ldots, 1)\}$ be the standard $k-1$-simplex in $\mathbb{R}^k$. Draw N points $\mathcal{X}$ independently from a mixture $\mathcal{D}$ of equally weighted spherical Gaussians of equal variance centered at the points in Γ. Then

$$\min_{\substack{\pi:\,[k]\to[k] \\ \text{permutation}}} \max_{t \in \{1,\ldots,k\}} \left\| c_t^{(\mathcal{X})} - \mu_{\pi(t)}^{(\Gamma,\mathcal{D})} \right\|_2$$

converges to zero in probability as $N \to \infty$, i.e., the k-means-optimal centroids converge to the Voronoi means.

REMARK A previous version of this article stated a stronger version for the Voronoi Means Conjecture for stable isogons. That version of the conjecture was disproven by Boris Alexeev.

THEOREM 5 Let $\mathcal{D}$ be a mixture of equally weighted spherical Gaussians of equal variance centered at the points of the standard $k-1$-simplex $\Gamma = \{\gamma_t\}_{t=1}^k$. Then there exists $\alpha > 0$ such that $\mu_t^{(\Gamma,\mathcal{D})} = \alpha\gamma_t$ for each $t \in \{1, \ldots, k\}$.

See Section 6 for the proof. This theorem holds for a bigger class of center configurations Γ. In particular, it holds for stable isogons. In Section 6, we introduce the concept and prove the more general result. To interpret Theorem 5, consider k-means optimization over the distribution $\mathcal{D}$ instead of a large sample $\mathcal{X}$ drawn from $\mathcal{D}$. This optimization amounts to finding k points $C = \{c_t\}_{t=1}^k$ in $\mathbb{R}^m$ that minimize

$$\sum_{t=1}^k \mathop{\mathbb{E}}_{X \sim \mathcal{D}} \left[\|X - c_t\|_2^2 \middle| X \in V_t^{(C)}\right] \mathop{\mathbb{P}}_{X \sim \mathcal{D}} \left(X \in V_t^{(C)}\right). \tag{7}$$

Intuitively, the optimal C is a good proxy for the k-means-optimal centroids when N is large (and one might make this rigorous using the plug-in principle with the Glivenko–Cantelli Theorem). What

Theorem 5 provides is that, when Γ is the vertices of the $k-1$-simplex, the Voronoi means have the same Voronoi cells as do Γ. As such, if one were to initialize Lloyd's algorithm at the Gaussian centers to solve (7), the algorithm converges to the Voronoi means in one step. Overall, one should interpret Theorem 5 as a statement about how the Voronoi means locally minimize (7), whereas the Voronoi means conjecture is a statement about global minimization.

What follows is the main result of this subsection:

THEOREM 6 Let $\Gamma = \{\gamma_t\}_{t=1}^k \subseteq \mathbb{R}^k$ the $k-1$-simplex. Then for every $\sigma > 0$, either

$$\sigma \lesssim \Delta_{\min}/\sqrt{\log k} \qquad \text{or} \qquad \min_{t\in\{1,\dots,k\}} \|\mu_t^{(\Gamma,\mathcal{D})} - \gamma_t\|_2 \gtrsim \sigma\sqrt{\log k},$$

where $\mathcal{D}$ denotes the mixture of equally weighted spherical Gaussians of entrywise variance σ^2 centered at the members of Γ.

See Section 7 for the proof. In other words, Theorem 6 establishes that one must accept k-dependence in either the data's noise or the estimate's error.

2.3 *Numerical example: clustering the MNIST dataset*

In this section, we apply our clustering algorithm to the NMIST handwritten digits dataset [**19**]. This dataset consists of 70,000 different 28×28 grayscale images, reshaped as 784×1 vectors; 55,000 of them are considered training set, 10,000 are test set and the remaining 5,000 are validation set.

Clustering the raw data gives poor results (due to 4's and 9's being similar, for example), so we first learn meaningful features and then cluster the data in feature space. To simplify feature extraction, we used the first example from the TensorFlow tutorial [**1**]. This consists of a one-layer neural network $y(x) = \sigma(Wx + b)$, where σ is the softmax function, W is a 784×10 matrix to learn and b is a 10×1 vector to learn. As the tutorial shows, the neural network is trained for 1,000 iterations, each iteration using batches of 100 random points from the training set.

Training the neural network amounts to finding W and b that fit the training set well. After selecting these parameters, we run the trained neural network on the first 1,000 elements of the test set, obtaining $\{y(x_i)\}_{i=1}^{1,000}$, where each $y(x_i)$ is a 10×1 vector representing the probabilities of being each digit. Since $y(x_i)$ is a probability vector, its entries sum to 1, and so the feature space is actually nine dimensional.

For this experiment, we cluster $\{y(x_i)\}_{i=1}^{1,000}$ with two different algorithms: (i) MATLAB's built-in implementation of k-means++, and (ii) our relax-and-round algorithm based on the k-means semidefinite relaxation (3). (The results of the latter alternative are illustrated in Fig. 2.)

Since each run of k-means++ uses a random initialization that impacts the partition, we ran this algorithm 100 times. In fact, the k-means value of the output varied quite a bit: the all-time low was 39.1371, the all-time high was 280.4174 and the median was 108.2358; the all-time low was reached in 34 out of the 100 trials. Since our relax-and-round alternative has no randomness, the outcome is deterministic, and its k-means value was 39.1371, i.e., identical to the all-time low from k-means++. By comparison, the k-means value of the planted solution (i.e., clustering according to the hidden digit label) was 103.5430, and the value of the SDP (which serves as a lower bound on the optimal k-means value) was 38.5891. As such, not only did our relax-and-round alternative produce the best clustering that k-means++ could find, but it also provided a certificate that no clustering has a k-means value that is 1.5% better.

Recalling the nature of our approximation guarantees, we also want to know well the relax-and-round algorithm's clustering captures the ground truth. To evaluate this, we determined a labeling of the clusters for which the resulting classification exhibited a minimal misclassification rate. (This amounts to minimizing a linear objective over all permutation matrices, which can be relaxed to a generically tight linear program over doubly stochastic matrices.) For k-means++, the all-time low misclassification rate was 0.0971 (again, accomplished by 34 of the 100 trials), the all-time high was 0.4070, and the median was 0.2083. As one might expect, the relax-and-round output had a misclassification rate of 0.0971.

3. Proof of Theorem 2

By the following lemma, it suffices to bound $\mathrm{Tr}(R(X_D - X_R))$:

LEMMA 7 $\|X_D - X_R\|_{\mathrm{F}}^2 \leq \dfrac{5}{n_{\min}\Delta_{\min}^2}\,\mathrm{Tr}(R(X_D - X_R))$.

Proof. First, by Lemma 1, we have $\|X_R\|_{\mathrm{F}}^2 = k$. We also claim that $\|X_D\|_{\mathrm{F}}^2 \leq k$. To see this, first note that $X_D 1 = 1$ and $X_D \geq 0$, and so the ith entry of $X_D v$ can be interpreted as a convex combination of the entries of v. Let v be an eigenvector of X_D with eigenvalue μ, and let i index the largest entry of v (this entry is positive without loss of generality). Then $\mu v_i = (X_D v)_i \leq v_i$, implying that $\mu \leq 1$. Since the eigenvalues of X_D lie in $[0, 1]$, we may conclude that $\|X_D\|_{\mathrm{F}}^2 \leq \mathrm{Tr}(X_D) = k$. As such,

$$\begin{aligned}\|X_D - X_R\|_{\mathrm{F}}^2 &= \|X_D\|_{\mathrm{F}}^2 + \|X_R\|_{\mathrm{F}}^2 - 2\,\mathrm{Tr}(X_D X_R)\\ &\leq 2k - 2\,\mathrm{Tr}(X_D X_R)\\ &= 2k + 2\,\mathrm{Tr}((X_R - X_D)X_R) - 2\|X_R\|_{\mathrm{F}}^2\\ &= 2\,\mathrm{Tr}((X_R - X_D)X_R). \end{aligned} \tag{8}$$

We will bound (8) in two different ways, and a convex combination of these bounds will give the result. For both bounds, we let Ω denote the indices in the diagonal blocks, and Ω^c the indices in the off-diagonal blocks, and $\Omega_t \subset \Omega$ denote the indices in the diagonal block for the cluster t. In particular, A_Ω denotes the matrix that equals A on the diagonal blocks and is zero on the off-diagonal blocks. For the first bound, we use that $R_\Omega = \xi(11^\top)_\Omega$ and that $(X_R - X_D)_\Omega(11^T)_\Omega$ has non-negative entries (since both X_R and X_D have non-negative entries, $X_R 1 = X_D 1 = 1$ and $X_R = (X_R)_\Omega$). Recalling that $R_\Omega = \xi$, we have

$$\begin{aligned}2\,\mathrm{Tr}((X_R - X_D)X_R) &= \sum_{t=1}^{k} 2\,\mathrm{Tr}\left((X_R - X_D)(11^\top)_{\Omega_t}\frac{1}{n_t}\right)\\ &\geq \frac{2}{n_{\max}}\,\mathrm{Tr}\left((X_R - X_D)(11^\top)_\Omega\right)\\ &= -\frac{2}{\xi n_{\max}}\,\mathrm{Tr}\left((X_D - X_R)R_\Omega\right). \end{aligned} \tag{9}$$

For the second bound, we first write $n_{\min} X_R = 11^\top - (11^\top)_{\Omega^c} - \sum_{t=1}^k \left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t}$. Since $X_R 1 = 1 = X_D 1$, we then have

$$\begin{aligned}
2\,\mathrm{Tr}((X_R - X_D)X_R) &= \frac{2}{n_{\min}} \mathrm{Tr}\left((X_D - X_R)\left((11^\top)_{\Omega^c} + \sum_{t=1}^k \left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t} - 11^\top\right)\right) \\
&= \frac{2}{n_{\min}} \mathrm{Tr}\left((X_D - X_R)\left((11^\top)_{\Omega^c} + \sum_{t=1}^k \left(1 - \frac{n_{\min}}{n_t}\right)(11^\top)_{\Omega_t}\right)\right) \\
&\le \frac{2}{n_{\min}} \mathrm{Tr}((X_D - X_R)(11^\top)_{\Omega^c}) \\
&= \frac{2}{n_{\min}} \mathrm{Tr}(X_D(11^\top)_{\Omega^c}),
\end{aligned}$$

where the last and second-to-last steps use that $(X_R)_{\Omega^c} = 0$. Next, $X_D \ge 0$ and $R_{\Omega^c} \ge (\xi + \Delta_{\min}^2/2)(11^\top)_{\Omega^c}$, and so we may continue:

$$\begin{aligned}
2\,\mathrm{Tr}((X_R - X_D)X_R) &\le \frac{2}{n_{\min}(\xi + \Delta_{\min}^2/2)} \mathrm{Tr}(X_D R_{\Omega^c}) \\
&= \frac{2}{n_{\min}(\xi + \Delta_{\min}^2/2)} \mathrm{Tr}((X_D - X_R)R_{\Omega^c}),
\end{aligned} \tag{10}$$

where again, the last step uses the fact that $(X_R)_{\Omega^c} = 0$. At this point, we have bounds of the form $x \ge a y_1$ with $a < 0$ and $x \le b y_2$ with $b > 0$ (explicitly (9) and (10)), and we seek a bound of the form $x \le c(y_1 + y_2)$. As such, we take the convex combination for a, b such that $a^{-1}/(a^{-1} + b^{-1}) < 0$ and $b^{-1}/(a^{-1} + b^{-1}) > 0$

$$x \le \frac{a^{-1}}{a^{-1} + b^{-1}} a y_1 + \frac{b^{-1}}{a^{-1} + b^{-1}} b y_2 = \frac{1}{a^{-1} + b^{-1}}(y_1 + y_2).$$

Taking $a = -2/(\xi n_{\max})$ and $b = 2/(n_{\min}(\xi + \Delta_{\min}^2/2))$ and combining with (8) then gives

$$\|X_D - X_R\|_{\mathrm{F}}^2 \le 2\,\mathrm{Tr}((X_R - X_D)X_R) \le \left(\frac{\xi}{2}(n_{\min} - n_{\max}) + \frac{n_{\min}}{4}\Delta_{\min}^2\right)^{-1} \mathrm{Tr}((X_D - X_R)(R_\Omega + R_{\Omega^c}))$$

choosing $\xi > 0$ sufficiently small and simplifying yields the result. □

We will bound $\mathrm{Tr}(R(X_D - X_R))$ in terms of the following: for each $N \times N$ real symmetric matrix M, let $\mathcal{F}(M)$ denote the value of the following program:

$$\begin{aligned}
\mathcal{F}(M) = \ & \text{maximum} \quad |\mathrm{Tr}(MX)| \\
& \text{subject to} \quad \mathrm{Tr}(X) = k,\ X1 = 1,\ X \ge 0,\ X \succeq 0.
\end{aligned} \tag{11}$$

LEMMA 8 Put $\widetilde{D} := P_{1^\perp} D P_{1^\perp}$ and $\tilde{R} := P_{1^\perp} R P_{1^\perp}$. Then $\mathrm{Tr}(R(X_D - X_R)) \le 2\mathcal{F}(\tilde{D} - \tilde{R})$.

Proof. Since X_D and X_R are both feasible in (11), we have

$$-\operatorname{Tr}(\widetilde{D}X_D) + \operatorname{Tr}(\widetilde{R}X_D) \leq |\operatorname{Tr}((\widetilde{D}-\widetilde{R})X_D)| \leq \mathcal{F}(\widetilde{D}-\widetilde{R}),$$
$$\operatorname{Tr}(\widetilde{D}X_R) - \operatorname{Tr}(\widetilde{R}X_R) \leq |\operatorname{Tr}((\widetilde{D}-\widetilde{R})X_R)| \leq \mathcal{F}(\widetilde{D}-\widetilde{R})$$

and adding followed by reverse triangle inequality gives

$$2\mathcal{F}(\widetilde{D}-\widetilde{R}) \geq \Big(\operatorname{Tr}(\widetilde{D}X_R) - \operatorname{Tr}(\widetilde{D}X_D)\Big) + \Big(\operatorname{Tr}(\widetilde{R}X_D) - \operatorname{Tr}(\widetilde{R}X_R)\Big). \tag{12}$$

Write $X_{\widetilde{D}} := P_{1^\perp} X_D P_{1^\perp}$. Note that $X_D 1 = (X_D)^T 1$ implies $X_D = X_{\widetilde{D}} + (1/N)11^\top$, and so

$$\operatorname{Tr}(\widetilde{D}X_D) = \operatorname{Tr}(DX_{\widetilde{D}}) = \operatorname{Tr}\Big(D\big(X_D - (1/N)11^\top\big)\Big) = \operatorname{Tr}(DX_D) + \frac{1}{N}1^\top D1.$$

Similarly, $\operatorname{Tr}(\widetilde{D}X_R) = \operatorname{Tr}(DX_R) + \frac{1}{N}1^\top D1$, and so

$$\operatorname{Tr}(\widetilde{D}X_R) - \operatorname{Tr}(\widetilde{D}X_D) = \operatorname{Tr}(DX_R) - \operatorname{Tr}(DX_D) \geq 0,$$

where the last step follows from the optimality of X_D. Similarly, $\operatorname{Tr}(\widetilde{R}X_D) - \operatorname{Tr}(\widetilde{R}X_R) = \operatorname{Tr}(R(X_D - X_R))$, and so (12) implies the result. □

Now it suffices to bound $\mathcal{F}(\widetilde{D}-\widetilde{R})$. For an $n_1 \times n_2$ matrix X, consider the matrix norm

$$\|X\|_{1,\infty} := \sum_{i=1}^{n_1} \max_{1\leq j\leq n_2} |X_{i,j}| = \sum_{i=1}^{n_1} \|X_{i,\cdot}\|_\infty.$$

The following lemma will be useful:

LEMMA 9 $\mathcal{F}(M) \leq \min\big\{\|M\|_{1,\infty}, \min\{k,r\}\|M\|_{2\to 2}\big\}$, where $r = \operatorname{rank}(M)$.

Proof. The first bound follows from the classical version of Hölder's inequality (recalling that $X_{i,j} \geq 0$ and $X1 = 1$ by design):

$$|\operatorname{Tr}(MX)| \leq \sum_{i=1}^N \sum_{j=1}^N |M_{i,j}X_{i,j}| \leq \sum_{i=1}^N \|M_{i,\cdot}\|_\infty \left(\sum_{j=1}^N |X_{i,j}|\right) = \sum_{i=1}^N \|M_{i,\cdot}\|_\infty.$$

The second bound is a consequence of Von Neumann's trace inequality: if the singular values of X and M are, respectively, $\alpha_1 \geq \ldots \geq \alpha_N$ and $\beta_1 \geq \ldots \geq \beta_N$, then

$$|\operatorname{Tr}(MX)| \leq \sum_{i=1}^N \alpha_i\beta_i.$$

Since X is feasible in (11), we have $\alpha_1 \le 1$ and $\sum_{i=1}^{N} \alpha_i \le k$. Using that $\operatorname{rank}(M) = r$, we get

$$|\operatorname{Tr}(MX)| \le \sum_{i=1}^{k} \beta_i \le \min\{k, r\}\|M\|_{2\to 2}. \qquad \square$$

Proof of Theorem 2. Write $x_{t,i} = r_{t,i} + \gamma_t$. Then

$$\begin{aligned}(D_{ab})_{ij} &= \|x_{a,i} - x_{b,j}\|_2^2 \\ &= \|(r_{a,i} + \gamma_a) - (r_{b,j} + \gamma_b)\|_2^2 = \|r_{a,i} - r_{b,j}\|_2^2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle + \|\gamma_a - \gamma_b\|_2^2.\end{aligned}$$

Furthermore,

$$\|r_{a,i} - r_{b,j}\|_2^2 = \|r_{a,i}\|_2^2 - 2\langle r_{a,i}, r_{b,j}\rangle + \|r_{b,j}\|_2^2 = ((\mu 1^\top + G^\top G + 1\mu^\top)_{ab})_{ij},$$

where G is the matrix whose (a,i)th column is $r_{a,i}$, and μ is the column vector whose (a,i)th entry is $\|r_{a,i}\|_2^2$. Recall that

$$(R_{ab})_{ij} = \xi + \Delta_{ab}^2/2 + \max\left\{0, \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle\right\}.$$

Then $P_{1^\perp}(D - R)P_{1^\perp} = P_{1^\perp}G^\top G P_{1^\perp} + P_{1^\perp}FP_{1^\perp}$, where

$$(F_{ab})_{ij} = \begin{cases} \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle & \text{if} \quad 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle \le -\Delta_{ab}^2/2 \\ 0 & \text{otherwise.} \end{cases}$$

Considering Lemma 9 and that $\operatorname{rank}(G^\top G) \le m$, we will bound

$$\mathcal{F}(M) \le \min\{k, m\}\|P_{1^\perp}G^\top G P_{1^\perp}\|_{2\to 2} + \frac{1}{n_{\min}}\|P_{1^\perp}FP_{1^\perp}\|_{1,\infty}. \tag{13}$$

For the first term:

$$\|P_{1^\perp}G^\top G P_{1^\perp}\|_{2\to 2} \le \|G^\top G\|_{2\to 2} = \|G^\top\|_{2\to 2}^2.$$

Note that if the rows $X_i^{(t)}, i = 1, \ldots n_t$ of $G^\top$ come from a distribution with second moment matrix Σ_t then $X_i^{(t)}$ has the same distribution as $\Sigma_t^{1/2} g$, where g is an isotropic random vector. Then $\|G^\top\| \le \sigma_{\max}\|\tilde{G}^\top\|$, where the rows of $\tilde{G}^\top$ are isotropic random vectors.

By Theorem 5.39 in [27], we have that there exist c_1 and c_2 constants depending only on the subgaussian norm of the rows of G such that with probability $\ge 1 - \eta$:

$$\|G^\top\|_{2\to 2} \le \sigma_{\max}\left(\sqrt{N} + c_1\sqrt{m} + \sqrt{c_2 \log(2/\eta)}\right).$$

Note that by Corollary 3.35, when the rows of $G^\top$ are Gaussian random vectors, we have the result for $c_1 = 1$ and $c_2 = 2$.

For bounding the second term in (13), the triangle inequality gives $\|P_{1^\perp} F P_{1^\perp}\|_{1,\infty} \leq 4\|F\|_{1,\infty}$. To get a handle on $\|F\|_{1,\infty}$, we first compute the expected value of its entries using that $|2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle|$ obeys a folded subgaussian distribution, coming from a subgaussian with variance at most $8\sigma_{\max}^2 \Delta_{ab}^2$:

$$\begin{aligned}
\mathbb{E}|(F_{ab})_{ij}| &\leq \left(\Delta_{ab}^2/2 + \mathbb{E}|2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle|\right) \mathbb{P}\left(2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle < -\Delta_{ab}^2/2\right) \\
&\leq \left(\frac{\Delta_{ab}^2}{2} + \frac{4\sigma_{\max}\Delta_{ab}}{\sqrt{\pi}}\right) \exp\left(-\frac{\Delta_{ab}^2}{64\sigma_{\max}^2}\right) \\
&\leq \Delta_{ab}^2 \exp\left(-\frac{\Delta_{ab}^2}{64\sigma_{\max}^2}\right) \text{ assuming } \Delta_{\min}^2 > 16k\sigma_{\max}^2 \quad k \geq 2 \\
&\leq \Delta_{ab}^2 \frac{64^2\sigma_{\max}^4}{\Delta_{ab}^4} \text{ using } \mathrm{e}^{-x} \leq \frac{1}{x^2} \text{ for } x > 0. \\
&\leq -\frac{256\sigma_{\max}^2}{k} \text{ using again } \Delta_{\min}^2 > 16k\sigma_{\max}^2 \quad k \geq 2 \\
&= O(\sigma_{\max}^2/k).
\end{aligned}$$

Now we can write $F = 2(L - L^\top)$, where $L_{a,i} := (L_{ab})_{ij} \in \{\langle r_{a,i}, \gamma_a - \gamma_b\rangle, 0\}$ has independent rows, and $\mathbb{E}|(L_{ab})_{ij}| \leq \mathbb{E}|(F_{ab})_{ij}| = O(\sigma_{\max}^2/k)$. We can then bound

$$\|F\|_{1,\infty} \leq 4\|L\|_{1,\infty} \leq \|L^{small}\|_{1,1},$$

where $L^{small} \in \mathbb{R}^{N\times k}$ is a submatrix of distinct columns.

Then we have a high-probability estimate:

$$\mathbb{P}(\|L^{small}\|_{1,1} > t) \leq \mathbb{P}\left(2k\sum_{a=1}^{k}\sum_{i=1}^{n_a}|L_{a,i}| > t\right) \leq \mathbb{P}\left(\sum_{a=1}^{k}\sum_{i=1}^{n_a}\left(|L_{a,i}| - \mathbb{E}|L_{a,i}|\right) > \frac{t}{2k} - c_3\sigma_{\max}^2 n_{\max}\right).$$

Using that $L_{a,i}$ are independent subgaussian random variables, we know there exist constants $c_4, c_5 \geq 0$ such that

$$\mathbb{P}\left(\sum_{a=1}^{k}\sum_{i=1}^{n_a}\left(|L_{a,i}| - \mathbb{E}|L_{a,i}|\right) > u\right) \leq c_4 \exp\left(-c_5\frac{u^2}{N}\right).$$

So, choosing $t = 2c_3 k n_{\max}\sigma_{\max}^2 + \sqrt{\frac{N}{c_5}\log\frac{c_4}{\eta}}$, we get that with probability at least $1 - \eta$

$$\|P_{1^\perp} F P_{1^\perp}\|_{1,\infty} \leq 8c_3 k n_{\max}\sigma_{\max}^2 + 4\sqrt{\frac{N}{c_5}\log\frac{c_4}{\eta}}.$$

Putting everything together, we get that there exist constants C_1, C_2, C_3 such that with probability at least $1 - 2\eta$

$$\begin{aligned}\|X_D - X_R\|_{\mathrm{F}}^2 &\le \frac{5}{n_{\min}\Delta_{\min}^2}\operatorname{Tr}(R(X_D - X_R)) \\ &\le \frac{10}{n_{\min}\Delta_{\min}^2}\mathcal{F}(\widetilde{D} - \widetilde{R}) \\ &\le C_1\frac{\min\{k,m\}\left(\sqrt{N} + c_1\sqrt{m} + \sqrt{c_2\log(2/\eta)}\right)^2\sigma_{\max}^2}{n_{\min}\Delta_{\min}^2} + C_2\frac{kn_{\max}\sigma_{\max}^2}{n_{\min}\Delta_{\min}^2} + C_3\frac{\sqrt{N\log c_4/\eta}}{n_{\min}\Delta_{\min}^2}.\end{aligned}$$

If, additionally, we require $N > \max\{c_1 m, c_2\log(2/\eta), \log(c_4/\eta)\}$, we get

$$\|X_D - X_R\|_{\mathrm{F}}^2 \le C\frac{k\alpha\sigma_{\max}^2(\alpha + \min\{k,m\})}{\Delta_{\min}^2}.$$

Rearranging gives the result. □

4. Denoising

In the special case where each Gaussian is spherical with the same entrywise variance σ^2 and the same number n of samples, the main result says:

$$\|X_D - X_R\|_{\mathrm{F}}^2 \lesssim \frac{k^2\sigma^2}{\Delta_{\min}^2}$$

with high probability as $n \to \infty$.

Let P denote the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$. Then PX_R is an $m \times N$ matrix whose (a,i)th column is $\tilde{\gamma}_a$, a good estimate of γ_a, and so one might expect PX_D to have its columns be close to the $\tilde{\gamma}_a$'s. This is precisely what the following theorem gives:

THEOREM 10 Suppose $\sigma \lesssim \Delta_{\min}/\sqrt{k}$. Let P denote the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$, and let $c_{a,i}$ denote the (a,i)th column of PX_D. Then

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim \frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2}\cdot k\sigma^2$$

with high probability as $n \to \infty$. Here, the ath column of Γ is $\tilde{\gamma}_a - \frac{1}{k}\sum_{b=1}^{k}\tilde{\gamma}_b$.

The proof can be found at the end of this section. For comparison,

$$\mathbb{E}\left[\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|x_{a,i} - \gamma_a\|_2^2\right] = m\sigma^2 \tag{14}$$

meaning the $c_{a,i}$'s serve as 'denoised' versions of the $x_{a,i}$'s provided $\|\Gamma\|_{2\to 2}$ is not too large compared with $\Delta_{\min}$. The following lemma investigates this provision:

LEMMA 11 For every choice of $\{\tilde{\gamma}_a\}_{a=1}^k$, we have

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \geq \frac{1}{2}$$

with equality if $\{\tilde{\gamma}_a\}_{a=1}^k$ is a simplex. More generally, if the following are satisfied simultaneously:

(i) $\sum_{a=1}^k \tilde{\gamma}_a = 0$,
(ii) $\|\tilde{\gamma}_a\|_2 \asymp 1$ for every $a \in \{1,\dots,k\}$ and
(iii) $|\langle \tilde{\gamma}_a, \tilde{\gamma}_b\rangle| \lesssim 1/k$ for every $a, b \in \{1,\dots,k\}$ with $a \neq b$,

then

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \lesssim 1.$$

See the end of the section for the proof. Plugging these estimates for $\|\Gamma\|_{2\to 2}^2/\Delta_{\min}^2$ into Theorem 10 shows that the $c_{a,i}$'s in this case exhibit denoising to an extent that the m in (14) can be replaced with k:

$$\frac{1}{N}\sum_{a=1}^k \sum_{i=1}^n \|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim k\sigma^2.$$

For more general choices of $\{\tilde{\gamma}_a\}_{a=1}^k$, one may attempt to estimate $\|\Gamma\|_{2\to 2}$ in terms of $\Delta_{\max}$, but this comes with a bit of loss in the denoising estimate:

COROLLARY 12 If $k\sigma \lesssim \Delta_{\min} \leq \Delta_{\max} \lesssim K\sigma$, then

$$\frac{1}{N}\sum_{a=1}^k \sum_{i=1}^n \|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim K^2\sigma^2$$

with high probability as $n \to \infty$.

Indeed, this does not guarantee denoising unless $k \lesssim K \leq \sqrt{m}$. To prove this corollary, apply the following string of inequalities to Theorem 10:

$$\|\Gamma\|_{2\to 2}^2 \leq \|\Gamma\|_{\mathrm{F}}^2 \leq k\Delta_{\max}^2 \lesssim kK^2\sigma^2,$$

where the second inequality uses the following lemma:

LEMMA 13 If $\sum_{a=1}^k \tilde{\gamma}_a = 0$, then $\|\tilde{\gamma}_a\|_2 \leq \Delta_{\max}$ for every a.

Proof. Fix a. Then

$$\min_{b\neq a}\left\langle \tilde{\gamma}_b, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2}\right\rangle \leq \frac{1}{k-1}\sum_{\substack{b=1\\ b\neq a}}^{k}\left\langle \tilde{\gamma}_b, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2}\right\rangle = \frac{1}{k-1}\left\langle \sum_{b=1}^{k}\tilde{\gamma}_b - \tilde{\gamma}_a, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2}\right\rangle = -\frac{1}{k-1}\|\tilde{\gamma}_a\|_2.$$

Let $b(a)$ denote the minimizer. Then Cauchy–Schwarz gives

$$\Delta_{\max} \geq \|\tilde{\gamma}_a - \tilde{\gamma}_{b(a)}\|_2 \geq \left\langle \tilde{\gamma}_a - \tilde{\gamma}_{b(a)}, \frac{\tilde{\gamma}_a}{\|\tilde{\gamma}_a\|_2}\right\rangle \geq \|\tilde{\gamma}_a\|_2 + \tfrac{1}{k-1}\|\tilde{\gamma}_a\|_2 \geq \|\tilde{\gamma}_a\|_2. \qquad \square$$

Proof of Theorem 10. Without loss of generality, we have $\sum_{a=1}^{k}\tilde{\gamma}_a = 0$. Write

$$\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i} - \tilde{\gamma}_a\|_2^2 = \|P(X_D - X_R)\|_{\mathrm{F}}^2 \leq \|P\|_{2\to 2}^2\|X_D - X_R\|_{\mathrm{F}}^2. \tag{15}$$

Decompose $P = \Gamma \otimes 1^\top + G$, where 1 is n dimensional and G has i.i.d. entries from $\mathcal{N}(0,\sigma^2)$. Observe that

$$\|\Gamma \otimes 1^\top\|_{2\to 2}^2 = \|(\Gamma \otimes 1^\top)(\Gamma \otimes 1^\top)^\top\|_{2\to 2} = \|n\Gamma\Gamma^\top\|_{2\to 2} = n\|\Gamma\|_{2\to 2}^2. \tag{16}$$

Also, Corollary 5.35 in [27] gives that

$$\|G\|_{2\to 2} \lesssim (\sqrt{N} + \sqrt{m})\sigma \lesssim \sqrt{N}\sigma \tag{17}$$

with probability $\geq 1 - \mathrm{e}^{-\Omega_m(N)}$. The result then follows from estimating $\|P\|_{2\to 2}$ with (16) and (17) by triangle inequality, plugging into (15) and then applying Theorem 2. $\square$

Proof of Lemma 11. Since $\|\Gamma x\|_2 \leq \|\Gamma\|_{2\to 2}\|x\|_2$ for every x, we have that

$$\|\Gamma\|_{2\to 2}^2 \geq \frac{\|\tilde{\gamma}_a - \tilde{\gamma}_b\|_2^2}{2}$$

for every a and b, and so

$$\frac{\|\Gamma\|_{2\to 2}^2}{\Delta_{\min}^2} \geq \frac{1}{2}\cdot\frac{\Delta_{\max}^2}{\Delta_{\min}^2} \geq \frac{1}{2}.$$

For the second part, let $\{\tilde{\gamma}_a\}_{a=1}^k$ be a simplex. Without loss of generality, $\{\tilde{\gamma}_a\}_{a=1}^k$ is centered at the origin, each point having unit 2-norm. Then $\langle\tilde{\gamma}_1,\tilde{\gamma}_2\rangle = -1/(k-1)$, and so

$$\Delta_{\min}^2 = \|\tilde{\gamma}_1 - \tilde{\gamma}_2\|_2^2 = \|\tilde{\gamma}_1\|_2^2 + \|\tilde{\gamma}_2\|_2^2 - 2\langle\tilde{\gamma}_1,\tilde{\gamma}_2\rangle = \frac{2k}{k-1}.$$

Next, we write

$$\Gamma^\top\Gamma = \frac{k}{k-1}I - \frac{1}{k-1}11^\top$$

and conclude that $\|\Gamma\|_{2\to 2}^2 = \|\Gamma^\top \Gamma\|_{2\to 2} = k/(k-1)$. Combining with our expression for $\Delta_{\min}^2$ then gives the result. For the last part, pick a and b such that $\Delta_{\min} = \|\tilde{\gamma}_a - \tilde{\gamma}_b\|_2$. Then

$$\Delta_{\min}^2 = \|\tilde{\gamma}_a\|_2^2 + \|\tilde{\gamma}_b\|_2^2 - 2\langle \tilde{\gamma}_a, \tilde{\gamma}_b\rangle \gtrsim 2 - 2/k.$$

Also, Gershgorin implies

$$\|\Gamma\|_{2\to 2}^2 = \|\Gamma^\top \Gamma\|_{2\to 2} \lesssim 1 + (k-1)/k$$

and so combining these estimates gives the result. □

5. Rounding

THEOREM 14 Take $\epsilon < \Delta_{\min}/8$, suppose

$$\#\Big\{(a,i) : \|c_{a,i} - \tilde{\gamma}_a\|_2 > \epsilon\Big\} < \frac{n}{2}$$

and consider the graph G of vertices $\{c_{a,i}\}_{i=1,\,a=1}^{n,\,k}$ such that $c_{a,i} \leftrightarrow c_{b,j}$ if $\|c_{a,i} - c_{b,j}\|_2 \le 2\epsilon$. For each $i = 1, \ldots, k$, select the vertex v_i of maximum degree (breaking ties arbitrarily) and update G by removing every vertex w such that $\|w - v_i\|_2 \le 4\epsilon$. Then there exists a permutation π on $\{1, \ldots, k\}$ such that

$$\|v_i - \tilde{\gamma}_{\pi(i)}\|_2 \le 3\epsilon$$

for every $i \in \{1, \ldots, k\}$.

Proof. Let $B(x, r)$ denote the closed 2-ball of radius r centered at x. For each i, we will determine $\pi(i)$ at the conclusion of iteration i. Denote $R_1 := \{1, \ldots, k\}$ and $R_{i+1} := R_i \setminus \{\pi(i)\}$ for each $i = 2, \ldots, k-1$. We claim that the following hold at the beginning of each iteration i:

(i) $< n/2$ vertices lie outside $\bigcup_{a\in R_i} B(\tilde{\gamma}_a, \epsilon)$,

(ii) $\ge n/2$ vertices lie inside $B(v_i, 2\epsilon)$ and

(iii) there exists a unique $a \in R_i$ such that $\|v_i - \tilde{\gamma}_a\|_2 \le 3\epsilon$.

First, we show that for each i, (i) and (ii) together imply (iii). Indeed, there are enough vertices in $B(v_i, 2\epsilon)$ that one of them must reside in $B(\tilde{\gamma}_a, \epsilon)$ for some $a \in R_i$. Furthermore, this a is unique since $\epsilon < \Delta_{\min}/6$. By triangle inequality, we have $\|v_i - \tilde{\gamma}_a\|_2 \le 3\epsilon$, and so we put $\pi(i) := a$.

We now prove (i) and (ii) by induction. When $i = 1$, we have (i) by assumption. For (ii), note that each $B(\tilde{\gamma}_a, \epsilon)$ contains $\ge n/2$ of the vertices, and by triangle inequality, each has degree $\ge n/2 - 1$ in G. As such, the vertex v_1 of maximum degree will have degree $\ge n/2 - 1$, thereby implying (ii).

Now suppose (i), (ii) and (iii) all hold for iteration $i < k$. By triangle inequality, (iii) implies $B(\tilde{\gamma}_{\pi(i)}, \epsilon) \subseteq B(v_i, 4\epsilon)$. As such, the ith iteration removes all vertices in $B(\tilde{\gamma}_{\pi(i)}, \epsilon)$ so that (i) continues to hold for iteration $i + 1$. Next, $\epsilon < \Delta_{\min}/8$ and (iii) together imply that the removal of vertices in $B(v_i, 4\epsilon)$ preserves the vertices in $B(\tilde{\gamma}_a, \epsilon)$ for every $a \in R_{i+1}$, and their degrees are still $\ge n/2 - 1$ by the same triangle argument as before. Thus, (ii) holds for iteration $i + 1$. □

COROLLARY 15 Suppose $k \lesssim m$ and denote $S := \|\Gamma\|_{2\to 2}/\Delta_{\min}$. Pick $\epsilon \asymp Sk\sigma$. Perform the rounding scheme of Theorem 14 to columns of PX_D. Then with high probability, $\{v_i\}_{i=1}^k$ satisfies

$$\|v_i - \tilde{\gamma}_{\pi(i)}\|_2 \lesssim Sk\sigma$$

for some permutation π, provided $\sigma \lesssim \Delta_{\min}/(Sk)$.

By Lemma 11, we have $S \lesssim 1$ in the best-case scenario. In this case, our rounding scheme works in the regime $\sigma \lesssim \Delta_{\min}/k$. (Note that denoising is guaranteed in the regime $\sigma \lesssim \Delta_{\min}/\sqrt{k}$.) In general, the cost of rounding is a factor of k in the average squared deviation of our estimates:

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim S^2k\sigma^2, \qquad \text{whereas} \qquad \frac{1}{k}\sum_{i=1}^{k}\|v_i - \tilde{\gamma}_{\pi(i)}\|_2^2 \lesssim S^2k^2\sigma^2.$$

On the other hand, we are not told which of the points in $\{c_{a,i}\}_{i=1,\,a=1}^{n\quad k}$ correspond to any given $\tilde{\gamma}_a$, whereas in rounding, we know that each v_i corresponds to a distinct $\tilde{\gamma}_a$.

Proof of Corollary 15. Draw (a, i) uniformly from $\{1, \dots, k\} \times \{1, \dots, n\}$ and take X to be the random variable $\|c_{a,i} - \tilde{\gamma}_a\|_2^2$. Then Markov's inequality and Theorem 10 together give

$$\begin{aligned}\#\Big\{(a,i) : \|c_{a,i} - \tilde{\gamma}_a\|_2 > \epsilon\Big\} &= N \cdot \mathbb{P}(X > \epsilon^2)\\ &\le \frac{N}{\epsilon^2}\cdot\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim \frac{N}{\epsilon^2}\cdot S^2k\sigma^2 \lesssim \frac{n}{2}.\end{aligned}$$

For Theorem 14 to apply, it suffices to ensure $\epsilon < \Delta_{\min}/8$, which follows from $\sigma \lesssim \Delta_{\min}/(Sk)$. □

6. Proof of Theorem 5

In this section, we prove a more general version of Theorem 5. In particular, we prove it for Γ any stable isogon. Examples of stable isogons include regular and quasi-regular polyhedra, as well as highly symmetric frames [9].

DEFINITION 16 We say $\Gamma \subseteq \mathbb{R}^m$ is a **stable isogon** if

(si1) $|\Gamma| > 1$,

(si2) the symmetry group $G \le O(m)$ of Γ acts transitively on Γ and

(si3) for each $\gamma \in \Gamma$, the stabilizer G_γ has the property that

$$\{x \in \mathbb{R}^m : Qx = x \ \ \forall Q \in G_\gamma\} = \text{span}\{\gamma\}.$$

LEMMA 17 Let G be the symmetry group of a stable isogon $\Gamma \subseteq \mathbb{R}^m$, and let K and H denote the subgroups of G that fix $W = \text{span}(\Gamma)$ and its orthogonal complement $W^\perp$, respectively. Then

(i) G is the direct sum of H and K,

(ii) H is finite,

(iii) H acts transitively on Γ and

(iv) K is isomorphic to the orthogonal group $O(m-r)$, where r is the dimension of W.

Proof. Pick $Q \in G$. Then Q permutes the points in Γ, and the permutation completely determines how Q acts on W by linearity. In particular, W is invariant under the action of G, which in turn implies the same for $W^\perp$. Let V denote an $m \times r$ matrix whose columns form an orthonormal basis for W, and let $V_\perp$ denote an $m \times (m-r)$ matrix whose columns form an orthonormal basis for $W^\perp$. Then Q can be expressed as

$$Q = \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix}.$$

We see that $Q \in H$ when $B = I$ and $Q \in K$ when $A = I$. Let Q_H denote the 'projection' of Q onto H, obtained by replacing B with I, and similarly for Q_K.

For (i), it suffices to show that H and K are normal subgroups of G, that $H \cap K = \{I\}$, and that G is generated by H and K. The first is obtained by observing that K is the kernel of the homomorphism $Q \mapsto Q_H$, and similarly, H is the kernel of $Q \mapsto Q_K$. The second follows from the observation that $Q \in H \cap K$ implies $A = I$ and $B = I$. The last follows from the observation that every $Q \in G$ can be factored as $Q_H Q_K$.

For (ii), we note that A is completely determined by how Q permutes the points in Γ, of which $\leq k!$ possibilities are available.

For (iii), we know that by (si2), for every pair $\gamma, \gamma' \in \Gamma$, there exists $Q \in G$ such that $G\gamma = \gamma'$. Consider the factorization $Q = Q_H Q_K$. Since $Q_K \gamma = \gamma$, we therefore have $Q_H \gamma = \gamma'$, meaning H also acts transitively on Γ.

For (iv), we first note that $B \in O(m-r)$ is necessary to have $Q \in O(m)$. Now pick any $B \in O(m-r)$. Then

$$Q = \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix} \tag{18}$$

has the effect of fixing each point in Γ, meaning $Q \in G$ (and therefore $Q \in K$). As such, K is the set of all Q of the form (18). □

LEMMA 18 For any stable isogon Γ, we have $\sum_{\gamma \in \Gamma} \gamma = 0$.

Proof. By (si1), we have $|\Gamma| > 1$. In the special case where $|\Gamma| = 2$ and the points in Γ are linearly dependent, write $\Gamma = \{\gamma_1, \gamma_2\}$ with $\gamma_1 \neq \gamma_2$. By (si2), we know there exists $Q \in G$ such that $Q\gamma_1 = \gamma_2$, and so $\|\gamma_1\|_2 = \|\gamma_2\|_2$. This combined with the assumed linear dependence gives $\gamma_1 = \pm\gamma_2$, and since $\gamma_1 \neq \gamma_2$, we conclude $\gamma_1 + \gamma_2 = 0$, as desired.

In the remaining case, Γ contains two points (say, γ_1 and γ_2) that are linearly independent. Fix $\gamma_0 \in \Gamma$. By Lemma 17(iii), the orbit $H\gamma_0$ is all of Γ. Consider the map from H onto Γ given by $f\colon U \mapsto U\gamma_0$ for all $U \in H$. The preimage of any member of Γ is a left coset of the stabilizer H_{γ_0} (which is finite by Lemma 17(ii)), and so

$$\sum_{U\in H} U\gamma_0 = \sum_{\gamma\in\Gamma} \sum_{U\in f^{-1}(\gamma)} U\gamma_0 = |H_{\gamma_0}| \sum_{\gamma\in\Gamma} \gamma. \tag{19}$$

Now pick any $Q \in G$, and consider the factorization $Q = Q_H Q_K$ with $Q_H \in H$ and $Q_K \in K$ (this exists uniquely by Lemma 17(i)). Since $U\gamma_0 \in \operatorname{span}(\Gamma)$ for every $U \in H$, we have that $Q_K U\gamma_0 = U\gamma_0$, and so

$$Q \sum_{U\in H} U\gamma_0 = \sum_{U\in H} Q_H U\gamma_0 = \sum_{U\in H} U\gamma_0, \tag{20}$$

where the last step follows from the fact that multiplication by $Q_H \in H$ permutes the members of H. Put $x = \sum_{U\in H} U\gamma_0$. Then (20) gives that $Qx = x$ for every $Q \in G$, and therefore $Qx = x$ for every $Q \in G_{\gamma_1} \cup G_{\gamma_2}$. By (si3), this then implies that $x \in \operatorname{span}\{\gamma_1\} \cap \operatorname{span}\{\gamma_2\}$, i.e., $x = 0$. Combining with (19) then gives the result. □

LEMMA 19 Let $\mathcal{D}$ be a mixture of equally weighted spherical Gaussians of equal variance centered at the points of a stable isogon $\Gamma = \{\gamma_t\}_{t=1}^k$. If $X \sim \mathcal{D}$ and Q is any member of the symmetry group G of Γ, then $QX \sim \mathcal{D}$.

Proof. The probability density function of $\mathcal{D}$ is given by

$$f(x) = \frac{1}{k} \sum_{t=1}^{k} \frac{1}{(\sqrt{2\pi}\sigma)^m} \mathrm{e}^{-\|x-\gamma_t\|_2^2/2\sigma^2}.$$

As such, $f(Qx) = f(x)$ since $\|Qx - \gamma_t\|_2 = \|x - Q^{-1}\gamma_t\|_2$ and $Q^{-1} \in G$ permutes the γ_t's. □

THEOREM 20 Let $\mathcal{D}$ be a mixture of equally weighted spherical Gaussians of equal variance centered at the points of a stable isogon $\Gamma = \{\gamma_t\}_{t=1}^k$. Then there exists $\alpha > 0$ such that $\mu_t^{(\Gamma,\mathcal{D})} = \alpha\gamma_t$ for each $t \in \{1, \ldots, k\}$.

Proof. Pick $Q \in G_{\gamma_t}$. Then $x \in V_t^{(\Gamma)}$ precisely when $x \in QV_t^{(\Gamma)}$. As such, Lemma 19 gives

$$\begin{aligned}
\mu_t^{(\Gamma,\mathcal{D})} = \mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[X \middle| X \in V_t^{(\Gamma)}\right] &= \mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[X \middle| X \in QV_t^{(\Gamma)}\right] \\
&= \mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[X \middle| Q^{-1}X \in V_t^{(\Gamma)}\right] \\
&= \mathop{\mathbb{E}}_{Y\sim\mathcal{D}}\left[QY \middle| Y \in V_t^{(\Gamma)}\right] = Q \mathop{\mathbb{E}}_{Y\sim\mathcal{D}}\left[Y \middle| Y \in V_t^{(\Gamma)}\right] = Q\mu_t^{(\Gamma,\mathcal{D})}.
\end{aligned}$$

By (si3), this then implies that $\mu_t^{(\Gamma,\mathcal{D})} \in \operatorname{span}\{\gamma_t\}$.

At this point, we have $\mu_t^{(\Gamma,\mathcal{D})} = \alpha_t \gamma_t$, where $\alpha_t = \langle \mu_t^{(\Gamma,\mathcal{D})}, \gamma_t \rangle / \|\gamma_t\|_2^2$. For any given $s \in \{1, \ldots, k\}$, pick Q such that $Q\gamma_t = \gamma_s$ (which exists by (si2)). Then Lemma 19 again gives

$$\begin{aligned}
\langle \mu_s^{(\Gamma,\mathcal{D})}, \gamma_s \rangle &= \left\langle \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[X \middle| X \in V_s^{(\Gamma)}\right], \gamma_s \right\rangle \\
&= \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[\langle X, \gamma_s \rangle \middle| X \in V_s^{(\Gamma)}\right] \\
&= \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[\langle X, \gamma_s \rangle \middle| X \in QV_t^{(\Gamma)}\right] \\
&= \mathop{\mathbb{E}}_{Y\sim\mathcal{D}} \left[\langle QY, \gamma_s \rangle \middle| Y \in V_t^{(\Gamma)}\right] = \mathop{\mathbb{E}}_{Y\sim\mathcal{D}} \left[\langle Y, \gamma_t \rangle \middle| Y \in V_t^{(\Gamma)}\right] = \langle \mu_t^{(\Gamma,\mathcal{D})}, \gamma_t \rangle,
\end{aligned}$$

meaning $\alpha_t = \alpha$ for all t. It remains to show that $\langle \mu_t^{(\Gamma,\mathcal{D})}, \gamma_t \rangle > 0$. To this end, note that $\|x - \gamma_t\|_2^2 < \|x - \gamma_s\|_2^2$ precisely when $\langle x, \gamma_t \rangle > \langle x, \gamma_s \rangle$, and so $x \in V_t^{(\Gamma)}$ if and only if $t = \arg\max_a \langle x, \gamma_a \rangle$. By Lemma 18,

$$\max_{a\in\{1,\ldots,k\}} \langle x, \gamma_a \rangle \geq \frac{1}{k} \sum_{a=1}^{k} \langle x, \gamma_a \rangle = \frac{1}{k} \left\langle x, \sum_{a=1}^{k} \gamma_a \right\rangle = 0,$$

with equality only if the maximizer is not unique, and so we conclude that $x \in V_t^{(\Gamma)}$ only if $\langle x, \gamma_t \rangle > 0$. As such,

$$\langle \mu_t^{(\Gamma,\mathcal{D})}, \gamma_t \rangle = \left\langle \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[X \middle| X \in V_t^{(\Gamma)}\right], \gamma_t \right\rangle = \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[\langle X, \gamma_t \rangle \middle| X \in V_t^{(\Gamma)}\right] > 0,$$

as desired. □

Combining Theorem 20 with the following lemma proves Theorem 5.

LEMMA 21 The standard $k-1$-simplex in $\mathbb{R}^k$ is a stable isogon for $k > 2$.

Proof. It suffices to observe that the symmetry group of Γ is the symmetric group S_k which acts transitively, implying (si1) and (si2). For (si3), pick $\gamma \in \Gamma$ and let i denote its non-zero entry. For any $\gamma_1 \neq \gamma_2 \in \Gamma$ with respective non-zero entries j, l different from i consider $Q \in G_\gamma$ that permutes γ_1 with γ_2 and leaves everything else fixed. Then $Q \in G_\gamma$, proving the statement. □

7. Proof of Theorem 6

We start with the following:

LEMMA 22 Let $\Gamma \subseteq \mathbb{R}^d$ denote the vertices of the $d-1$ simplex. For any given $c \geq 0$, consider the mixture $\mathcal{D}_c$ of equally weighted spherical Gaussians of unit entrywise variance centered at the members of $c\Gamma$. Let $V_1^{(\Gamma)}$ denote the Voronoi region corresponding to the first identity basis element, and define $\alpha_d \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ by

$$\alpha_d(c) := \mathop{\mathbb{E}}_{X\sim\mathcal{D}_c} \left[X_1 \middle| X \in V_1^{(\Gamma)}\right] = d \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_d=-\infty}^{x_1} x_1 f(x; c) \, \mathrm{d}x_d \cdots \mathrm{d}x_1, \tag{21}$$

where $f(\cdot; c)$ denotes the probability density function of $\mathcal{D}_c$:

$$f(x; c) := \frac{1}{d} \sum_{t=1}^{d} \frac{1}{(2\pi)^{d/2}} \mathrm{e}^{-\|x - c\gamma_t\|_2^2/2}.$$

Then $\alpha_d(c) \geq \alpha_d(0)$ for all $c \geq 0$.

See the end of this section for the proof. For context, our proof of Theorem 6 requires a bound on $\alpha_d(c)$ for general $c \geq 0$, and so Lemma 22 allows us to pass to the easier-to-estimate quantity $\alpha_d(0)$. The following lemmas estimate $\alpha_d(0)$:

LEMMA 23 If $g \sim \mathcal{N}(0, I)$ in $\mathbb{R}^d$ then $\mathbb{E}\|g\|_\infty \gtrsim \sqrt{\log d}$.

Proof. When $d = 1$, $\|g\|_\infty$ has half-normal distribution, and so its expected value is $\sqrt{2/\pi}$. Otherwise, $d \geq 2$. Since $\|g\|_\infty \geq \max_i g(i)$, we will estimate $\mathbb{E} \max_i g(i)$. To this end, take z such that $\mathbb{P}(g(1) \geq z) = 1/d$, denote $j := \arg\max_i g(i)$, and condition on the event that $g(j) < z$, which occurs with probability $(1 - 1/d)^d$:

$$\begin{aligned} \mathbb{E}g(j) &= \mathbb{E}\Big[g(j)\Big|g(j) < z\Big] \cdot (1 - 1/d)^d + \mathbb{E}\Big[g(j)\Big|g(j) \geq z\Big] \cdot \Big(1 - (1 - 1/d)^d\Big) \\ &\geq \frac{1}{2}\mathbb{E}\Big[g(j)\Big|g(j) < 0\Big] \cdot (1/4) + z \cdot \Big(1 - (1 - 1/d)^d\Big). \end{aligned} \tag{22}$$

Since $g(j) \geq \frac{1}{d}\sum_{i=1}^d g(i)$, we have

$$\mathbb{E}\Big[g(j)\Big|g(j) < 0\Big] \geq \mathbb{E}\Big[\frac{1}{d}\sum_{i=1}^{d} g(i)\Big|g(i) < 0 \ \forall i\Big] = -\sqrt{\frac{2}{\pi}}.$$

With this, we may continue (22) to get

$$\mathbb{E}g(j) \geq -\frac{1}{8}\sqrt{\frac{2}{\pi}} + z \cdot (1 - \mathrm{e}^{-1}) \gtrsim z \gtrsim \sqrt{\log d},$$

where the last step follows from rearranging $1/d = \mathbb{P}(g(1) \geq z) \geq \mathrm{e}^{-O(z^2)}$. □

LEMMA 24 The function α_d defined by (23) satisfies $\alpha_d(0) \gtrsim \sqrt{\log d}$.

Proof. Note that

$$\begin{aligned} \alpha_d(0) = d &\int_{x_1=0}^{\infty} \int_{x_2=-x_1}^{x_1} \cdots \int_{x_d=-x_1}^{x_1} x_1 f(x; 0) \ \mathrm{d}x_d \cdots \mathrm{d}x_1 \\ + d &\int_{x_1=-\infty}^{0} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_d=-\infty}^{x_1} x_1 f(x; 0) \ \mathrm{d}x_d \cdots \mathrm{d}x_1 \\ + d &\int_{x_1=0}^{\infty} \int_{x_2=-\infty}^{-x_1} \cdots \int_{x_d=-\infty}^{-x_1} x_1 f(x; 0) \ \mathrm{d}x_d \cdots \mathrm{d}x_1, \end{aligned} \tag{23}$$

where the last two terms cancel out. Let $A_1^{(\Gamma)}$ the domain of the first integral (idem $A_t^{(\Gamma)}$).

It is straightforward to verify that $x \in A_t^{(\Gamma)}$ precisely when

$$j = \arg\max_{i\in\{1,\dots,d\}} |x(i)| \qquad \text{and} \qquad \gamma_t = \operatorname{sign}(x(j)) \cdot e_j,$$

and so $x \in A_t^{(\Gamma)}$ implies $\langle x, \gamma_t\rangle = \|x\|_\infty$. Letting $g \sim \mathcal{N}(0, I)$ in $\mathbb{R}^d$, then

$$\alpha_d(0) = d\mathbb{E}\Big[\langle g, \gamma_1\rangle \text{ s.t. } g \in A_1^{(\Gamma)}\Big] = d\mathbb{E}\Big[\|g\|_\infty \text{ s.t. } g \in A_1^{(\Gamma)}\Big] \le \mathbb{E}\Big[\|g\|_\infty \Big| A_1^{(\Gamma)}\Big] = \mathbb{E}\|g\|_\infty,$$

where the third step occurs because $\mathbb{P}\left(A_1^{(\Gamma)}\right) \le 1/d$. The last step follows from the fact that $\|\Pi g\|_\infty$ has the same distribution for every signed permutation Π since, and this implies that the random variable is independent of the event $g \in A_1^{(\Gamma)}$. The result then follows from Lemma 23. □

We are now ready to prove the theorem of interest:

Proof of Theorem 6. Take Γ to be the standard orthoplex of dimension $d = k$. Observe that Theorem 5 along with a change of variables in (23) gives

$$\mu_t^{(\Gamma,\mathcal{D})} = \alpha\gamma_t = \sigma\alpha_d(1/\sigma)\gamma_t.$$

This then implies

$$\begin{aligned}\|\mu_t^{(\Gamma,\mathcal{D})} - \gamma_t\|_2 = |\langle \mu_t^{(\Gamma,\mathcal{D})} - \gamma_t, \gamma_t\rangle| &\ge |\langle \mu_t^{(\Gamma,\mathcal{D})}, \gamma_t\rangle| - |\langle \gamma_t, \gamma_t\rangle| \\ &= \sigma\alpha_d(1/\sigma) - \Delta_{\min}/\sqrt{2} \ge \sigma\alpha_d(0) - \Delta_{\min}/\sqrt{2}, \qquad (24)\end{aligned}$$

where the last step follows from Lemma 22. At this point, we consider two cases. In the first case, (24) $\ge \sigma\alpha_d(0)/2$, which implies

$$\|\mu_t^{(\Gamma,\mathcal{D})} - \gamma_t\|_2 \ge \sigma\alpha_d(0) - \Delta_{\min}/\sqrt{2} \ge \sigma\alpha_d(0)/2 \gtrsim \sigma\sqrt{\log k},$$

where the last step follows from Lemma 24. Since this bound is independent of t, we then get

$$\min_{t\in\{1,\dots,k\}} \|\mu_t^{(\Gamma,\mathcal{D})} - \gamma_t\|_2 \gtrsim \sigma\sqrt{\log k}.$$

In the remaining case, we have (24) $< \sigma\alpha_d(0)/2$ which one may rearrange to get

$$\sigma < \frac{2}{\alpha_d(0)} \cdot \frac{\Delta_{\min}}{\sqrt{2}} \lesssim \Delta_{\min}/\sqrt{\log k}.$$

□

Proof of Lemma 22. We will show that $\alpha_d(c)$ has a non-negative derivative, and the result will follow from the mean value theorem. First, we write $\alpha_d(c) = \sum_{t=1}^d I_t(c)$, where

$$I_t(c) := \frac{1}{(2\pi)^{d/2}} \int_{x\in V_1^{(\Gamma)}} x_1\, \mathrm{e}^{-\|x - c\gamma_t\|_2^2/2}\, \mathrm{d}x. \qquad (25)$$

We claim that $I_s(\cdot) = I_t(\cdot)$ whenever $s, t \in \{2, \dots, d\}$. This can be seen by changing variables in (25). As such, we have $\alpha_d(c) = I_1(c) + (d-1)I_2(c)$, where we take $\gamma_2 = e_2$ without loss of generality. At this point, we factor out the x_1 dependence in the integrands of $I_1(c)$ and observe that

$$\int_{x_2=-\infty}^{x_1} \cdots \int_{x_d=-\infty}^{x_1} \mathrm{e}^{-(x_2^2+\cdots+x_d^2)/2}\,\mathrm{d}x_d \cdots \mathrm{d}x_2 = \left(\int_{-\infty}^{x_1} \mathrm{e}^{-z^2/2}\,\mathrm{d}z\right)^{d-1} = (\pi/2)^{(d-1)/2}(\operatorname{erf}(x_1/\sqrt{2})+1)^{d-1}$$

to get

$$I_1(c) = \frac{1}{2^{d-1/2}\sqrt{\pi}} \int_{-\infty}^{\infty} x\,\mathrm{e}^{-(x-c)^2/2}(\operatorname{erf}(x/\sqrt{2})+1)^{d-1}\,\mathrm{d}x.$$

Similarly,

$$I_2(c) = \frac{1}{2^{d-1}\pi} \int_{x_1=-\infty}^{\infty} x_1\,\mathrm{e}^{-x_1^2/2}\left(\int_{x_2=-\infty}^{x_1} \mathrm{e}^{-(x_2-c)^2/2}\,\mathrm{d}x_2\right)(\operatorname{erf}(x_1/\sqrt{2})+1)^{d-2}\,\mathrm{d}x_1.$$

At this point, we apply differentiation under the integral sign to get

$$I_1'(c) = \frac{1}{2^{d-1/2}\sqrt{\pi}} \int_{-\infty}^{\infty} x(x-c)\,\mathrm{e}^{-(x-c)^2/2}(\operatorname{erf}(x/\sqrt{2})+1)^{d-1}\,\mathrm{d}x,$$
$$I_2'(c) = \frac{1}{2^{d-1}\pi} \int_{-\infty}^{\infty} x\,\mathrm{e}^{-x^2/2}\left(-\,\mathrm{e}^{-(x-c)^2/2}\right)(\operatorname{erf}(x/\sqrt{2})+1)^{d-2}\,\mathrm{d}x.$$

To continue, note that

$$\frac{\mathrm{d}}{\mathrm{d}x}(\operatorname{erf}(x/\sqrt{2})+1)^{d-1} = \frac{2(d-1)}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}(\operatorname{erf}(x/\sqrt{2})+1)^{d-2}.$$

With this, we integrate by parts to change the expression for $I_2'(c)$:

$$I_2'(c) = \frac{1}{d-1}\left(-I_1'(c) + \frac{1}{2^{d-1/2}\sqrt{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-(x-c)^2/2}(\operatorname{erf}(x/\sqrt{2})+1)^{d-1}\,\mathrm{d}x\right).$$

Overall, we have

$$\alpha_d'(c) = I_1'(c) + (d-1)I_2'(c) = \frac{1}{2^{d-1/2}\sqrt{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-(x-c)^2/2}(\operatorname{erf}(x/\sqrt{2})+1)^{d-1}\,\mathrm{d}x,$$

which is non-negative since the integrand is everywhere non-negative. □

Acknowledgments

The authors thank Pablo Parrilo for suggesting SDPNAL+ as a solver for the k-means semidefinite program, and David Bowie for the music that sustained our investigation of stable isogons and the Voronoi Means Conjecture.

Funding

D.G.M. was supported by an AFOSR Young Investigator Research Program award, NSF Grant No. DMS-1321779, and AFOSR Grant No. F4FGA05076J002. R.W. was supported in part by an NSF CAREER grant and ASOFR Young Investigator Award 9550-13-1-0125. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K.,Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
2. Achlioptas, D. & McSherry, F. (2005) On Spectral Learning of Mixtures of Distributions. *International Conference on Computational Learning Theory* (P. Auer and R. Meir eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 458–469.
3. Achlioptas, D. & McSherry, F. (2007) Fast computation of low-rank approximations. *J. ACM*, **54**, 9.
4. Sanjeev, A. & Kannan, R. (2001) Learning mixtures of arbitrary gaussians. *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, pp. 247–257.
5. Awasthi, P., Bandeira, A., Charikar, M., Krishnaswamy, R., Villar, S. & Ward, R. (2015) Relax, no need to round: Integrality of clustering formulations. *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. pp. 191–200.
6. Awasthi, P. & Sheffet, O. (2012) Improved spectral-norm bounds for clustering. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (A. Gupta, K. Jansen, J. D. P. Rolim & R. Servedio eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–49.
7. Barvinok, A. I. (1995) Problems of distance geometry and convex properties of quadratic maps. *Discrete Comput. Geom.*, **13**, 189–202.
8. Belkin, M. & Sinha, K. (2015) Polynomial learning of distribution families. *SIAM J. Comput.*, **44**, 889–911.
9. Broome, H. & Waldron, S. (2013) On the construction of highly symmetric tight frames and complex polytopes. *Linear Algebra Appl.*, **439**, 4135–4151.
10. Brubaker, S. C. & Vempala, S. (2008) Isotropic pca and affine-invariant clustering. *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08. Washington, DC, USA: IEEE Computer Society, pp. 551–560.
11. Chaudhuri, K. & Rao, S. (2008) Beyond Gaussians: Spectral Methods for Learning Mixtures of Heavy-Tailed Distributions. *COLT* (R. Servedio and T. Zhang eds) Vol. 4. New York, NY, USA: ACM, p. 1.
12. Chaudhuri, K. & Rao, S. (2008) Learning Mixtures of Product Distributions Using Correlations and Independence. *COLT*. Vol. 4, No. 1, pp. 9–20.
13. Dasgupta, S. (1999) Learning mixtures of gaussians. *Foundations of Computer Science, 1999. 40th Annual Symposium on*. Washington, DC, USA: IEEE Computer Society, pp. 634–644.
14. Dasgupta, S. & Schulman, L. (2007) A probabilistic analysis of em for mixtures of separated, spherical gaussians. *J. Mach. Learn. Res.*, **8**, 203–226.
15. Guédon, O. & Vershynin, R. (2016) Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, **165**, 1025–1049.
16. Iguchi, T., Mixon, D. G., Peterson, J. & Villar, S. (2015) Probably certifiably correct k-means clustering. *Mathematical Programming*, 1–38.
17. Kannan, R., Salmasian, H. and Vempala, S. (2008). The spectral method for general mixture models. *SIAM Journal on Computing,* **38**, 1141–1156.

18. KUMAR, A. & KANNAN, R. (2010) Clustering with spectral norm and the k-means algorithm. *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. Washington, DC, USA: IEEE Computer Society, pp. 299–308.

19. LECUN, Y., CORTES, C. & BURGES, C. J. (1998) The MNIST database of handwritten digits. Available at http://yann.lecun.com/exdb/mnist/ (last accessed 18 February 2017).

20. MOITRA, A. & VALIANT, G. (2010) Settling the polynomial learnability of mixtures of gaussians. *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. Washington, DC, USA: IEEE Computer Society, pp. 93–102.

21. NELLORE, A. & WARD, R. (2015) Recovery guarantees for exemplar-based clustering. *Inf. Comput.*, **245**, 165–180.

22. PATAKI, G. (1998) On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.*, **23**, 339–358.

23. PENG, J. & WEI, Y. (2007) Approximating k-means-type clustering via semidefinite programming. *SIAM J. Optim.*, **18**, 186–205.

24. SHAPIRO, A. (1982) Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, **47**, 187–199.

25. VILLAR, S. Implementation of the *k*-means semidefinite program, 2016. Software available from github.com/solevillar, last accessed 18 February 2017.

26. VEMPALA, S. & WANG, G. (2004) A spectral algorithm for learning mixture models. *J. Comput. Sys. Sci.*, **68**, 841–860.

27. VERSHYNIN, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications* (Y. Eldar & G. Kutyniok eds). Cambridge University Press.

28. YANG, L., SUN, D. & TOH, K.-C. (2015) Sdpnal+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Math. Program. Comput.*, **7**, 331–366.