Lecture 5, Clustering Mixture Models

Alert: these notes are a work in progress, and have not been subjected to the usual scrutiny reserved for formal publications!

1 Introduction to Clustering

In a clustering problem, we are given n items and associated information. The goal is to find a good partition of the items into k parts, where the k parts correspond to some underlying similarity between some of the samples.

Example 1. Examples of clustering problems are:

- Given $G = (V, E)$ with $|V| = n$, we want to cluster the vertices into k disjoints parts $S_1, S_2, \dots, S_k$, so that we minimize edges between different clusters, i.e. $E(S_i, S_j)$ for $i \neq j$. This intuitively corresponds to identifying a structure of k communities in a graph.
- Given n vectors $x_1, \dots, x_n \in \mathbb{R}^d$, we want to cluster them respecting geometry. This includes clustering based on distance, but also more exotic cases like clustering for meshes (see [2], for example, for some clustering related to computer vision).
- Clustering in an n point metric space $\{d(x_i, x_j)\}_{i,j\in[n]}$, where clusters are formed by points that are close to each other.

There are many interesting clustering-related computational problems, both in theory and in practice. Some that we will *not* discuss include determining what the “right” number of clusters is. Today, we focus on the question of clustering when we already know the number of clusters k and that a good clustering of size k exists.

Specifically, we consider a clustering problem $\Theta = \{(S_1, \dots, S_k), X)\}$ where we receive X as an input, $S_1, \dots, S_k$ are a partition of $[n]$, and our goal is to find $S_1, \dots, S_k$. (X could be a graph, a set of vectors, etc.) We will assume that $|S_1| = \cdots = |S_k| = n/k$ and that the problem is identifiable. Formally, we require that there exists a map (maybe not efficiently computable) $\zeta : X \longrightarrow (T_1, \dots, T_k)$ such that, for all $(S_1, \dots, S_k, X)$, we must have $|T_i \cap S_i| \geq (1-\delta)n/k.$

Example 2. The main example for today clustering Gaussian mixture models. For k Gaussians distributions $D_1, D_2, \dots, D_k$ over $\mathbb{R}^d$, given n samples $X_1, X_2, \dots, X_n$, sampled independently from the mixture $\frac{1}{k}\sum_{i=1}^{k} D_i$, we want to (approximately) recover the sets $S_1, S_2, \dots, S_k$, where S_i corresponds to the samples drawn from D_i. (We describe this in more detail in Section 3).

So how do we cluster with SoS?

2 Clustering with SoS

Here, we describe a meta-algorithm for clustering with SoS.

Algorithm 3. *Suppose that for some* $((S_1, \ldots, S_k), X) \in \Theta$ *there exist an input-specific (that is, depending on* X*) axiom set* P_X *of degree at most* d *and size* $n^{O(d)}$ *on variables* $\omega_1, \omega_2, \ldots, \omega_n, z$ *such that:*

1. $\left\{ \omega_i^2 = \omega_i, \sum_i \omega_i = n/k \right\} \cup P_X \vdash \sum_{i \in S_a, j \in S_b} \omega_i \omega_j \leq \delta \left(\frac{n}{k} \right)^2$ *holds for any two different true clusters* S_a *and* S_b*.*
2. *Whenever* ω *is the indicator of some true cluster* S_a*, the axioms* P_X *are satisfied.*

Then, by finding a pseudoxpectation $\tilde{\mathbf{E}}$ *satisfying* $\left\{ w_i^2 = w_i, \sum_i w_i = n/k \right\} \cup P_X$ *for which* $||\tilde{\mathbf{E}}\omega||_2^2$ *is minimized, one can recover in time* $n^{O(d)}$ *sets* $T_1, T_2, \ldots, T_k$ *such that* $|S_i \cap T_i| \geq (1 - \delta k^{O(1)}) \frac{n}{k}$*.*

Before proving Algorithm 3, we give some intuition about it. Note that the condition $\sum_{i \in S_a, j \in S_b} \omega_i \omega_j$ implies that ω cannot simultaneously assign a lot of weight on two different clusters. Therefore, rounding with respect to ω or its higher moments (stay tuned!) will allow us to approximately recover the true clusters. As we will show the rounding is simple. Thus, all the "magic" in designing clustering SoS algorithms is in finding the right polynomials P_X. We give a very simple example of how to do this.

Example 4 (When we forget about the sweep line)**.** Suppose that our data $X_1, X_2, \ldots, X_n$ comes from the following model. There are k different unknown unit-length intervals $(a_i, a_i + 1)$ on a line, which are at least distance 10 apart (that is, $a_i + 11 < a_{i+1}$, for all i). Then, our n samples are formed by taking n/k points in each interval. Our goal is to recover the sets of points belonging to the same interval. Of course, this is a computationally trivial problem as we can just use a sweep line argument... Nevertheless, it serves as a good example of how to use SoS.

How do we encode the fact that the distances between points in the same cluster are small, but they are otherwise big? Imagining that ω was a true cluster indicator (or a distribution over such!), an intuitive set of axioms P_X is $\{\omega_i \omega_j (X_i - X_j)^2 \leq 1\}_{i \neq j}$. Indeed, this gives the desired inequality in Algorithm 3 as

$$\sum_{i \in S_a, j \in S_b} \omega_i \omega_j \leq \sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{(X_i - X_j)^2}{100} \leq \frac{1}{100} \sum_{i \in S_a, j \in S_b} \omega_i \omega_j (X_i - X_j)^2 \leq$$
$$\frac{1}{100} \sum_{i \in S_a, j \in S_b} 1 = \frac{1}{100} \left(\frac{n}{k} \right)^2.$$

All of the above inequalities follow in SoS. The first simply because $(X_i - X_j)^2 > 100$ and $\omega_i \omega_j = \omega_i^2 \omega_j^2$ (in SoS) and the second from the axioms P_X. So, using Algorithm 3, we manage to cluster to 99% accuracy... Of course, it might be disappointing that for this trivial problem, our SoS based approach doesn't give 100% accuracy. We can actually fix this by iterating

$$\frac{1}{100} \sum_{i \in S_a, j \in S_b} \omega_i \omega_j (X_i - X_j)^2 \leq \frac{1}{10^4} \sum_{i \in S_a, j \in S_b} \omega_i \omega_j (X_i - X_j)^4 \ldots$$

It is easy to show that all of these steps can be done in SoS, but now we move to the less trivial problem of rounding.

2.1 The Rounding Procedure

Now let's imagine we are given a pseudoexpectation $\tilde{\mathbf{E}}$ in variables $\omega_1, \ldots, \omega_n$ which satisfies P_X. One seemingly-ideal situation would be that it is actually the low-degree moments of a distribution on cluster indicators which is supported on a single cluster-indicator vector, say 1_{S_a}, the indicator vector for S_a. Then $\tilde{\mathbf{E}}\omega = 1_{S_a}$ and we could easily identify S_a.
It turns out that there are two main issues with this approach. First, it requires that $\tilde{\mathbf{E}}\omega$ is close to a cluster indicator. However, as we will see in a bit, $\tilde{\mathbf{E}}$ might actually correspond to a nontrivial distribution over cluster indicators. Second, even if we get lucky and $\tilde{\mathbf{E}}$ is concentrated on one cluster, what do we do after identifying that one cluster?

An alternative possibility is that $\tilde{\mathbf{E}}$ is the low-degree moments of the uniform distribution on cluster indicators, $1_{S_1}, \ldots, 1_{S_k}$. In this case we can hope that $\tilde{\mathbf{E}}$ contains information about all $S_1, \ldots, S_k$, but we can no longer read that information off of the first moments, as $\tilde{\mathbf{E}}\omega_i = 1/k$ for each i.

So, what could we do in this case? We can read the clusters off of the second moments $\tilde{\mathbf{E}}\omega\omega^\top$, since $\tilde{\mathbf{E}}\omega_i\omega_j$ is nonzero if and only if i, j belong to the same S_a.

We will make this general strategy work for (almost) any pseudoexpectation $\tilde{\mathbf{E}}$ which satisfies P_X, as we know that the very strong condition $\tilde{\mathbf{E}}\sum_{i\in S_a, j\in S_a}\omega_i\omega_j \leq \delta\left(\frac{n}{k}\right)^2$ holds, simply because $\tilde{\mathbf{E}}$ satisfies the axioms in Algorithm 3. As $\sum\omega_i = \frac{n}{k}$, it must also be the case that within clusters the moments are large. Indeed,

$$\sum_{a=1}^{k}\sum_{i,j\in S_a}\tilde{\mathbf{E}}\omega_i\omega_j = \left(\frac{n}{k}\right)^2 - \sum_{a\neq b}\sum_{i\in S_a, j\in S_b}\tilde{\mathbf{E}}\omega_i\omega_j \geq$$

$$\left(\frac{n}{k}\right)^2 - k^2\delta\left(\frac{n}{k}\right)^2 = n^2\left(\frac{1}{k^2} - \delta\right),$$

which is much larger than $\delta(n/k)^2$ when $\delta \ll \frac{1}{k^{O(1)}}$.

In fact, the only thing standing in our way is the first possibility we considered, that $\tilde{\mathbf{E}}$ could be supported on a single cluster indicator (or, say, all but one of them). We avoid this situation by requiring $||\tilde{\mathbf{E}}\omega||$ to be as small as possible – this forces $\tilde{\mathbf{E}}$ to spread out over all of the clusters.
So, how do we use these second moments? Intuitively, ω should be a "distribution over true cluster indicators". Thus, conditioning that we are in one of the clusters, we can recover other points in it as well. Formally, one rounding scheme that uses this idea is the following.

Algorithm 5. *Given* $\tilde{\mathbf{E}}$, *for* $a \in \{1, 2, \ldots k\}$:

1. *Pick* i_a *uniformly among the set of remaining items* $R = [n]\setminus\bigcup_{s<a}T_a$ *and compute the conditional pseudoexpectation* $\tilde{\mathbf{E}}[\cdot|\omega_{i_a} = 1]$.
2. *Include each element* j *of* R *independently in* T_a *with probability* $\tilde{\mathbf{E}}[\omega_j|\omega_{i_a} = 1]$.

Return the clusters $T_1, T_2, \ldots, T_k$.

We will prove that with high probability, there exists an indexing of $S_1, S_2, \ldots, S_k$ such that $|S_i \cap T_i| \geq \frac{n}{k}(1 - 16\delta k^8)$ holds with high probability. We prove this in several steps.

First, minimality of $||\tilde{\mathbf{E}}\omega||_2^2$ implies that $\tilde{\mathbf{E}}\omega_i = \frac{1}{k}$ for each i. Indeed, this is true for the following reason. Since each cluster indicator 1_{S_a} satisfies all axioms, the (pseudo)-expectation $\tilde{\mathbf{E}}^a$ defined by $\tilde{\mathbf{E}}^a q(\omega) := q(1_{S_a})$ satisfies the axioms.[1] But then so does $\tilde{\mathbf{E}}' := \frac{1}{k}\sum_{a=1}^k \tilde{\mathbf{E}}^a$. Now, $\tilde{\mathbf{E}}'$ satisfies $\tilde{\mathbf{E}}\omega_i = \frac{1}{k}$ for all i. Since this clearly minimizes the two-norm $||\tilde{\mathbf{E}}\omega||_2^2$ among all pseudoexpectations for which $\sum_i \tilde{\mathbf{E}}\omega_i = \frac{n}{k}$, the choice of $\tilde{\mathbf{E}}$ implies the desired condition. We continue with two technical lemmas.

Proposition 6. *Suppose that $A \subseteq S_a$ is a set such that $|A| \geq \frac{n}{k}(1-\epsilon)$ for some $\epsilon < 1/2$ and let $i \in A$. Then,*

$$\mathbf{E}_{i\sim Unif(A)}\tilde{\mathbf{E}}[\sum_{j\in S_a}\omega_j|w_i = 1] \geq \frac{n}{k}(1-\epsilon-2k^2\delta),\ \textit{and}$$

$$\mathbf{E}_{i\sim Unif(A)}\tilde{\mathbf{E}}[\sum_{j\in[n]\setminus S_a}\omega_j|w_i = 1] \leq 2nk\delta$$

Proof. First, note that for any $i \in A$, we have

$$\tilde{\mathbf{E}}[\sum_{j\in[n]}\omega_j|\omega_i = 1] = \frac{n}{k},$$

as $\sum_{j\in[n]}\omega_j = \frac{n}{k}$ is one of the axioms. Thus, the two inequalities above are equivalent. On the other hand, note that

$$\mathbf{E}_{i\sim Unif(A)}\tilde{\mathbf{E}}[\sum_{j\in[n]\setminus S_a}\omega_j|w_i = 1] \leq \frac{|S_a|}{|A|}\mathbf{E}_{i\sim Unif(S_a)}\tilde{\mathbf{E}}[\sum_{j\in[n]\setminus S_a}\omega_j|w_i = 1] \leq$$

$$\frac{1}{1-\epsilon}\frac{k}{n}\sum_{i\in S_a}\tilde{\mathbf{E}}[\sum_{j\in[n]\setminus S_a}\omega_j|w_i = 1] \leq$$

$$(1+2\epsilon)\frac{k}{n}\sum_{b\neq a}\sum_{i\in S_a, j\in S_b}\tilde{\mathbf{E}}[\omega_i\omega_j]/\tilde{\mathbf{E}}[\omega_i] =$$

$$(1+2\epsilon)\frac{k^2}{n}\sum_{b\neq a}\sum_{i\in S_a, j\in S_b}\tilde{\mathbf{E}}[\omega_i\omega_j] \leq$$

$$(1+2\epsilon)\frac{k^2}{n}\sum_{b\neq a}\delta\left(\frac{n}{k}\right)^2 \leq (1+2\epsilon)\delta kn \leq 2\delta kn,$$

where in the last inequality we used the assumptions in Algorithm 3, before that the fact that $\tilde{\mathbf{E}}\omega_i = \frac{1}{k}$, and before that the definition of conditional pseudoexpectation. □

Proposition 7. *Suppose that we are at some step s of the rounding algorithm and there are at least $\frac{n}{k}(1-\epsilon)$ elements of some true cluster S_a in R. Conditioned on the fact that $i_s \in S_a$, with probability at least $\frac{1}{k^2}$, we will add to T_s at least $\frac{n}{k}(1-2k^2\epsilon-4k^4\delta)$ elements from S_a and at most $\frac{n}{k}4k^4\delta$ that are not in S_a.*

Proof. We only prove the second statement as the proofs are the same. Note that the previous proposition implies that in expectation, we add to T_a at most $\frac{n}{k}2k^2\delta$ elements. By Markov's inequality, we add more than $\frac{n}{k}4k^4\delta$ with probability at most $\frac{1}{2k^2}$. The other statement is equivalent, except that we consider the number of elements of S_a we don't add. Then, we do union bound over the two statements. □

[1] This is a real expectation, so it is also a pseudoexpectation.

Using the last proposition, we can easily prove by induction that the following things happen together with high probability as long as $\delta < \frac{1}{k^{10}}$. For all steps a, the following simultaneously hold:

1. Each i_a for $a \in [k]$ belongs to a different cluster $S_{\pi(a)}$.
2. For each $a \in [k]$, $\left|\left(S_{\pi(1)} \cup S_{\pi(2)} \cup \cdots S_{\pi(a)}\right) \setminus (T_1 \cup T_2 \cdots \cup T_a)\right| \leq 12\frac{n}{k}a\delta k^7$
3. For each $a \in [k]$, $|(T_1 \cup T_2 \cdots \cup T_a) \setminus \left(S_{\pi(1)} \cup S_{\pi(2)} \cup \cdots S_{\pi(a)}\right)| \leq 4\frac{n}{k}a\delta k^4$.

Indeed, note that if the statements above are satisfied before some step a, then at step a we will choose an item i_a from a new set with probability at least

$$\frac{\frac{n}{k}(k-a) - 12\frac{n}{k}a\delta k^7}{\frac{n}{k}(k-a)} > 1 - 12\delta k^8,$$

so we proved the first bullet point by induction. Then, applying Proposition 7 with $\epsilon = 4(a-1)\delta k^4$ (since at most $\epsilon\frac{n}{k}$ elements of each true cluster that has not been selected have been added before step a), we know that with probability at least $\frac{1}{k^2}$, we will add to T_a at least $\frac{n}{k}(1 - 4k^4\delta - 8ak^6\delta) > \frac{n}{k}(1 - 12k^7\delta)$ elements of the true cluster of i_a and at most $4\frac{n}{k}\delta k^4$ elements not from the cluster (here we use $\delta < \frac{1}{k^{10}}$). Thus, the second and third bullets points continue to hold after step a.

All that is left to show to complete the proof is that everything happens with high probability simultaneously. Note, however, that we apply Proposition 7 exactly k times. By union bound, with probability at least $1 - k \times \frac{1}{k^2}$, every time the bounds on $|T_a \backslash S_{\pi(a)}|$ and $|S_{\pi(a)} \backslash T_a|$ hold. Similarly, at each step we select a new element with probability at least $1 - 12\delta k^8$, so this happens with probability at least $1 - 12\delta k^9$. Therefore, with probability at least $1 - 12\delta k^9 - k^{-1}$, the statement holds. As $\delta < \frac{1}{k^{10}}$ and $k = \omega(1)$ (as k and d are polynomially related), the induction is complete.

To finish, note that the last two bullet points clearly imply that

$$|S_{\pi(a)} \cap T_a| \geq \frac{n}{k} - |S_a \backslash T_a| \geq \frac{n}{k}(1 - 16\delta k^8),$$

as desired. Indeed, at most $12\delta k^7 \frac{n}{k}$ of the elements of S_a are in $T_1 \cup T_2 \cup \cdots \cup T_{a-1}$ by the third bullet point, and at most another $4\delta k^4 \frac{n}{k}$ of the elements of S_a are not in $T_1 \cup T_2 \cup \cdots \cup T_{a_1} \cup T_a$ by the second bullet point, so at worst T_a misses $16\delta k^8 \frac{n}{k}$ of the elements of $S_{\pi(a)}$.

The analysis of the rounding algorithm is finished.

3 Gaussian Mixture Models

So far, in Algorithm 3 we have seen that in order to design an SoS clustering algorithm, it is enough to find well-behaved input-specific polynomials P_X and construct the set of axioms

$$\{\omega_i^2 = \omega_i \; \forall i, \; \sum_i \omega_i = \frac{n}{k}\} \cup P_X.$$

This allows us to find an appropriate pseudoexpectation satisfying these axioms and perform our rounding scheme. Now, we turn to the main problem for today - clustering Gaussian mixtures.

3.1 Model and Preliminaries

Set-Up: In a mixture model, k unknown independent distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ over $\mathbb{R}^d$ are given. In our set-up, we focus on $\mathcal{D}_i = \mathcal{N}(\mu_i, \Sigma_i)$. The problem of interest is the following. The input $X_1, X_2, \ldots, X_n$, consists of n independent variables distributed according to $\frac{1}{k}\sum_i \mathcal{D}_i$. That is, for each $j \in [n]$ an independent cluster indicator $c_j \sim Unif(\{1, 2, \ldots, k\})$ is sampled and then X_j is sampled independently from $\mathcal{D}_{c_j}$. The task is the following. Let $S_c = \{j \in [n] \ : \ c_j = c\}$ for $c \in [k]$ be the true clusters. We need to recover (approximately) the sets $S_1, S_2, \ldots, S_k$. The input consists of the samples $X_1, X_2, \ldots, X_n$.

Different regularity assumptions imposed on this problem are studied. For this lecture, we assume a second-moment bound $\Sigma_i \preceq I$. We also make the simplifying assumption that every true cluster contains exactly $\frac{n}{k}$ samples. This condition can be easily removed using concentration inequalities for the random variables $|S_c|$. Implicit today will also be a polynomial relationship between the dimension and number of clusters, that is $k = d^{\Theta(1)}$.

Clearly, distances between means will play a crucial role. If two of the means, say μ_1 and μ_2 are nearly the same, it is information theoretically impossible to distinguish[2] $\mathcal{N}(\mu_a, I)$ and $\mathcal{N}(\mu_b, I)$. Thus, we introduce the parameters $\Delta_{a,b} = \mu_a - \mu_b$. Central to the analysis will be a lower bound Δ on $\min_{a \neq b} ||\Delta_{a,b}||_2$.

This immediately leads us to a first idea about an SoS clustering method.

A First Naive Approach: We can repeat the same idea as in the SoS proof of identifying interval clustering on a line by encoding the fact that $||X_i - X_j||$ is large if and only if X_i and X_j are in different clusters. Explicitly, we form the axioms $\{\omega_i \omega_j ||X_i - X_j||^2 \leq c\}$ for an appropriately chosen c (so that with high probability, distances between pairs of samples from different clusters are greater than distances between pairs of samples from the same cluster). It turns out that this is indeed a viable approach, but it only works whenever $\Delta_{a,b} >> k^{1/4}$ holds for any $a \neq b$. Can we do better?

It turns out that we can. A purely information-theoretic result (leading to an exponential time algorithm) is the following.

Theorem 8 ([6]). *There exists an (exponential time) algorithm with the following guarantee. If $\Delta \gg \sqrt{\log k}$, using $(dk)^{O(1)}$ samples, one can cluster to 99% accuracy.*

What about polynomial-time algorithms? A very recent result establishes that the above guarantee is actually efficiently achievable for spherical Gaussians.

Theorem 9 ([5]). *For any positive constant $c > 0$, one can cluster with 99% accuracy a mixture of k spherical Gaussians (that is, Gaussians with covariance I) for which $\Delta > (\log k)^{\frac{1}{2}+c}$ in time $poly(d, k)$ when $n = poly(d, k)$.*

When we allow for a quasipolynomial time, we can again achieve $\sqrt{\log k}$, even assuming only $\Sigma_a \preceq I$.

Theorem 10 ([4, 3, 1]). *There exists an algorithm using $d^{O(\log k)}$ samples and time which clusters with 99% accuracy when $\Delta \gg \sqrt{\log k}$.*

[2] See, for example p.13 here for the KL divergence between two Gaussians.

The theorem that we will prove is the following. Theorem 10 also follows from our proof, if we are a little more careful with parameters.

Theorem 11 ([4, 3, 1])**.** *For any $\epsilon > 0$, if $\Delta > k^{\epsilon}$, there exists a $poly(n, d)$ clustering algorithm that achieves 99% accuracy.*

4 Gaussian Clustering with SoS

The idea behind improving the naive approach is to use more information about Gaussian random variables. Since Gaussian random variables have very thin tails, the higher moments also behave well. Namely, we have the following fact about single dimensional Gaussians. If $N \sim \mathcal{N}(0, 1)$, then it is well known that for any t,

$$\mathbf{E}|N|^t = \frac{2^{t/2}\Gamma(\frac{t+1}{2})}{\sqrt{\pi}} = O(t)^{t/2}.$$

It turns out that a much stronger high-probability version of this equality holds, as we will see in a bit. This motivates the construction of axioms which use the relatively small average t-th moment within clusters and the much larger t-th moment between different clusters. This leads us to the following axioms.

4.1 First Set of Axioms and SoS Proof

Axioms $P_{X_1,X_2,\dots,X_n}$: For every unit vector v, we have the inequality $A^v_{X_1,X_2,\dots,X_n}$ given by

$$\frac{k}{n}\sum_{i=1}^{n} \omega_i \langle X_i - \frac{k}{n}\sum_{j=1}^{n}\omega_j X_j, v\rangle^t \le O(t)^{t/2},$$

where t is some parameter which we will choose later. At a first glance, we haven't made much progress. We have an uncountable collection of axioms - one inequality for each unit vector v in $\mathbb{R}^d$. We will deal with this problem later, however, and we will move on to an SoS proof. That is, we need to show two things.

1. True Clusters Satisfy Axioms: That is, the true clusters satisfy each inequality

$$\frac{k}{n}\sum_{i} \omega_i \langle X_i - \frac{k}{n}\sum_{j}\omega_j X_j, v\rangle^t \le O(t)^{t/2}.$$

We will not prove the statement, but will at least give some intuition what this expression looks like for true clusters. Let ω be the indicator of some cluster S_a, corresponding to $\mathcal{N}(\mu_a, \Sigma_a)$. Let $X_i = \mu_a + Y_i$ for $i \in S_a$ where $Y_i \sim \mathcal{N}(0, \Sigma_a)$. Then,

$$\frac{k}{n}\sum_{i} \omega_i \langle X_i - \frac{k}{n}\sum_{j}\omega_j X_j, v\rangle^t = \frac{k}{n}\sum_{i\in S_a} \langle Y_i + \mu_a - \frac{k}{n}\sum_{j\in S_a}(Y_j + \mu_a), v\rangle^t =$$

$$\frac{k}{n}\sum_{i\in S_a} \langle Y_i - \frac{k}{n}\sum_{j\in S_a} Y_j, v\rangle^t.$$

This is just the empirical t-th moment of a d-dimensional Gaussian; standard concentration tools imply that it concentrates to its expectation uniformly for all unit v with $poly(d)$ samples (for constant t).

2. The Axioms Imply a Small Overlap between Different Clusters: That is, we need to show that for any two different clusters S_a and S_b, the axioms P_X imply that

$$\sum_{i \in S_a, j \in S_b} \omega_i \omega_j \leq \delta \left(\frac{n}{k}\right)^2$$

for $\delta = \frac{1}{poly(k)}$. This can be proved in SoS as follows.

The first step is to include the geometry of the clustering problem via the t-th moments, which we know should be small. Namely, we have

$$\sum_{i \in S_a, j \in S_b} \omega_i \omega_j = \sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{\langle \mu_a - \mu_b, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}.$$

Now, we try to incorporate the fact that all points in a cluster are close to the mean by approximating μ_ℓ by the weighted average defined by ω. That is, for $\mu(\omega) := \frac{k}{n} \sum_j \omega_i X_i$, we have

$$\sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{\langle \mu_a - \mu_b, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} = \sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{\langle \mu_a - \mu(\omega) + \mu(\omega) - \mu_b, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}.$$

Now, from the homework, we know that there exists an SoS proof of the triangle inequality $||p + q||_s^s \leq 2^{O(s)}(||p||_s^s + ||q||_s^s)$. Using it, we obtain

$$\sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{\langle \mu_a - \mu(\omega) + \mu(\omega) - \mu_b, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} \leq$$

$$2^{O(t)} \sum_{i \in S_a, j \in S_b} \omega_i \omega_j \frac{\langle \mu_a - \mu(\omega), \Delta_{a,b} \rangle^t + \langle \mu(\omega) - \mu_b, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} =$$

$$2^{O(t)} \sum_{j \in S_b} \sum_{i \in S_a} \omega_i \omega_j \frac{\langle \mu_a - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} + 2^{O(t)} \sum_{j \in S_a} \sum_{i \in S_b} \omega_i \omega_j \frac{\langle \mu_b - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}.$$

Now, since $\sum_i \omega_i = \frac{n}{k}$, for any polynomial $p(\omega)$, there exists an SoS proof that $\sum_i \omega_i p(\omega) = \frac{n}{k} p(\omega)$, again as in the homework. Therefore, the above expression becomes

$$2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle \mu_a - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} + 2^{O(t)} \frac{n}{k} \sum_{j \in S_b} \omega_j \frac{\langle \mu_b - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}.$$

We only bound the first term due to symmetry. We now want to incorporate our axioms. Again using the triangle inequality,

$$2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle \mu_a - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} = 2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle \mu_a - X_i + X_i - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} \leq$$

$$2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle \mu_a - X_i, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} + 2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle X_i - \mu(\omega), \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}$$

Since the entire expression $\frac{\langle \mu_a - X_i, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}}$ is just a real number, $\omega_i^2 = \omega_i$ clearly proves that

$$2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \omega_i \frac{\langle \mu_a - X_i, \Delta_{a,b} \rangle^t}{||\Delta_{a,b}||_2^{2t}} \leq 2^{O(t)} \frac{n}{k} \sum_{i \in S_a} \frac{\langle \mu_a - X_i, \frac{\Delta_{a,b}}{||\Delta_{a,b}||_2} \rangle^t}{||\Delta_{a,b}||_2^t}.$$

Using that $\frac{\Delta_{a,b}}{||\Delta_{a,b}||_2} = 1$, we know that with high probability, each term $\langle \mu_a - X_i, \frac{\Delta_{a,b}}{||\Delta_{a,b}||_2}\rangle^t$ is of order $O(t)^{t/2}$, so the last expression is bounded with high probability by

$$\frac{n}{k}\sum_{i\in S_a}\frac{O(t)^{t/2}}{||\Delta_{a,b}||_2^t} = \left(\frac{n}{k}\right)^2\frac{O(t)^{t/2}}{||\Delta_{a,b}||_2^t}.$$

Similarly, for the other term, the axioms $P_{X_1,X_2,\ldots,X_n}$ directly imply that

$$2^{O(t)}\frac{n}{k}\sum_{i\in S_a}\omega_i\frac{\langle X_i-\mu(\omega),\Delta_{a,b}\rangle^t}{||\Delta_{a,b}||^{2t}} = 2^{O(t)}\frac{n}{k}||\Delta_{a,b}||_2^{-t}\sum_{i\in S_a}\omega_i\langle X_i-\frac{k}{n}\sum_{i=1}^n\omega_iX_i,\frac{\Delta_{a,b}}{||\Delta_{a,b}||_2}\rangle^t \le$$

$$\left(\frac{n}{k}\right)^2\frac{O(t)^{t/2}}{||\Delta_{a,b}||_2^t}.$$

Therefore, the entire expression is bounded in SoS by

$$\left(\frac{n}{k}\right)^2\frac{O(t)^{t/2}}{||\Delta_{a,b}||_2^t}.$$

Now, we see that if $||\Delta_{a,b}|| > k^\epsilon$, it is enough to choose $t >> \frac{1}{\epsilon}$ such that

$$\frac{O(t)^{t/2}}{||\Delta_{a,b}||_2^t} = \frac{1}{poly(k)}.$$

In other words, we have managed to design a $poly(n,d)$ algorithm for separation $\Delta = k^\epsilon$ for any constant ϵ, improving the $k^{1/4}$ barrier of the naive approach.

4.2 Making the Axiom Space Finite

Recall that we constructed an uncountable axiom space to prove that $\sum_{i\in S_a,j\in S_b}\omega_i\omega_j \le \delta\left(\frac{n}{k}\right)^2$. However, for an efficient algorithm we want the axiom space to be of size $(nd)^{O(1)}$. How can we do this?

A natural idea is to discretize the unit sphere by, say, an ϵ-net and approximate each unit vector v by a (linear) combination of the vectors in the respective net. This, however, will still result in an axiom family of exponential size as ϵ-nets on the unit sphere generally have exponential size.

Another idea is to find some polysized family of axioms Q_X such that $Q_X \vdash P_X$ (where P_X is the uncountable family constructed in the previous section). How should we do this? We start by exploring the case for small t.

The case $t = 2$: We want axioms which imply that

$$\frac{k}{n}\sum_{i=1}^n\omega_i\langle X_i-\mu(\omega),v\rangle^2 \le O(1)$$

holds for any unit v. Note, however, that the above expression is simply

$$v^T\left[\frac{k}{n}\sum_{i=1}^n\omega_i(X_i-\mu(\omega))(X_i-\mu(\omega))^T\right]v.$$

Thus, it is enough that the spectral norm of $M = \left[\frac{k}{n}\sum_{i=1}^{n}\omega_i(X_i-\mu(\omega))(X_i-\mu(\omega))^T\right]$ is at most $O(1)$. But we already know how to encode this condition from the lecture on Robust Mean Estimation Claim 2.7 here. We simply need to add the matrix of slack variables B and the axiom

$$\frac{k}{n}\sum_{i=1}^{n}\omega_i(X_i-\mu(\omega))(X_i-\mu(\omega))^T = O(1)I - BB^T.$$

What about the next case, $t=4$?
The case $t=4$: Going through the same logic, we compute

$$\frac{k}{n}\sum_{i=1}^{n}\omega_i\langle X_i-\mu(\omega),v\rangle^4 = (v^{\otimes 2})^T\left[\frac{k}{n}\sum\omega_i\left[(X_i-\mu(\omega))^{\otimes 2}\right]\left[(X_i-\mu(\omega))^{\otimes 2}\right]^T\right](v^{\otimes 2}).$$

Seems like we just need an SoS axiom implying that

$$\frac{k}{n}\sum\omega_i\left[(X_i-\mu(\omega))^{\otimes 2}\right]\left[(X_i-\mu(\omega))^{\otimes 2}\right]^T \preceq O(1)I.$$

Except that there is a caveat. This fact does not hold for true clusters. Namely, for an indicator ω of cluster a, if we set $Y_i = X_i - \frac{k}{n}\sum_{j\in S_a}X_j$ for $i\in S_a$, the above expression becomes

$$\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes 2}\right]\left[Y_i^{\otimes 2}\right]^T.$$

Now, consider how this two-form acts on the unit vector $u\in\mathbb{R}^{d\otimes d}$ given by $u_{i,j} = \frac{1[i=j]}{\sqrt{d}}$. The respective quantity is

$$u^t\left[\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes 2}\right]\left[Y_i^{\otimes 2}\right]^T\right]u =$$

$$\frac{k}{n}\sum_{i\in S_a,j_1,j_2\in[d]}Y_{i,j_1}^2Y_{i,j_2}^2\frac{1}{d} = \frac{k}{nd}\sum_{i\in S_a,j_1,j_2\in[d]}(X_{i,j_1}-\frac{k}{n}\sum_{t\in S_a}X_{t,j_1})^2(X_{i,j_2}-\frac{k}{n}\sum_{t\in S_a}X_{t,j_2})^2.$$

Note that in expectation, (when the covariance matrix is I)

$$\mathbf{E}((X_{i,j_1}-\frac{k}{n}\sum_{t\in S_a}X_{t,j_1})^2) = (1-\frac{k}{n})^2\mathbf{E}(X_{i,j_1}) + \sum_{t\neq i}\frac{k^2}{n^2}\mathbf{E}(X_{t,j_1}) = \Theta(1).$$

As we are taking the sum of $\frac{n}{k}d^2$ summands of order 1, in expectation the above expression evaluates to $\Theta(d)=\omega(1)$ rather than a constant. So, is approach doomed?

It turns out that it is not. Even if the matrix

$$\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes 2}\right]\left[Y_i^{\otimes 2}\right]^T$$

has singular values of non-constant order, it is still the case that when it acts on unit vectors of the form $v^{\otimes 2}$, its spectral norm is constant with high probability. That is,

$$(v^{\otimes 2})^t\left[\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes 2}\right]\left[Y_i^{\otimes 2}\right]^T\right](v^{\otimes 2}) \leq 1 \qquad (1)$$

holds with high probability (we defer this proof to Lemma 12).

Something more can be show to hold true: in fact, as long as $n \geq (kd)^{O(1)}$, standard concentration results imply that every entry of the matrix $\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes t/2}\right]\left[Y_i^{\otimes t/2}\right]^\top$ concentrates, and consequently that, with high probability,

$$\frac{k}{n}\sum_{i\in S_a}\left[Y_i^{\otimes t/2}\right]\left[Y_i^{\otimes t/2}\right]^T \preceq \mathbf{E}_{Z\sim\mathcal{N}(0,I_d)}\left[\left[Z^{\otimes t/2}\right]\left[Z^{\otimes t/2}\right]^T\right] + 0.01I$$

Therefore, it is enough to add the slack variables $B \in \mathbb{R}^{d^{\frac{t}{2}}\times d^{\frac{t}{2}}}$ and our axiom $Q_X(\omega, B)$ becomes

$$\frac{k}{n}\sum \omega_i\left[(X_i-\mu(\omega))^{\otimes t/2}\right]\left[(X_i-\mu(\omega))^{\otimes t/2}\right]^T = \mathbf{E}_{Z\sim\mathcal{N}(0,I_d)}\left[\left[Z^{\otimes t/2}\right]\left[Z^{\otimes t/2}\right]^T\right]+0.01I-BB^\top$$

(Note that $\mathbf{E}_{Z\sim\mathcal{N}(0,I_d)}\left[\left[Z^{\otimes t/2}\right]\left[Z^{\otimes t/2}\right]^T\right]$ is just a real matrix and can be computed in time $d^{O(t)}$). This axiom implies the SoS inequality in additional indeterminates $v = v_1, \ldots, v_d$

$$(v^{\otimes t/2})^T\left[\frac{k}{n}\sum \omega_i\left[(X_i-\mu(\omega))^{\otimes t/2}\right]\left[(X_i-\mu(\omega))^{\otimes t/2}\right]^T\right](v^{\otimes t/2}) \leq$$

$$(v^{\otimes t/2})^T\left[\mathbf{E}_{Z\sim\mathcal{N}(0,I)}\left[\left[Z^{\otimes t/2}\right]\left[Z^{\otimes t/2}\right]^T\right]+0.01I\right]v^{\otimes t/2} \leq O(t)^{t/2}\|v\|_2^t,$$

as desired. The respective SoS proofs are now in indeterminates (ω, B). We know that with high probability there exists a some B_a for each true cluster S_a such that $(1_{S_a}, B_a)$ satisfies the constructed axioms.

References

[1] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. "List-decodable robust mean estimation and learning mixtures of spherical gaussians". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 1047–1060.

[2] Fengtao Fan, Fuhua Cheng, Conglin Huang, Yong Li, Jianzhong Wang, and Shuhua Lai. "Mesh clustering by approximating centroidal Voronoi tessellation". In: *Symposium on Solid and Physical Modeling*. 2009.

[3] Samuel B. Hopkins and Jerry Li. "Mixture Models, Robustness, and Sum of Squares Proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2018. Los Angeles, CA, USA: Association for Computing Machinery, 2018, pp. 1021–1034. ISBN: 9781450355599. DOI: 10.1145/3188745.3188748. URL: https://doi.org/10.1145/3188745.3188748.

[4] Pravesh Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (2018).

[5] Allen Liu and Jerry Li. "Clustering Mixtures with Almost Optimal Separation in Polynomial Time". In: STOC 2022. 2022. ISBN: 9781450392648.

[6] Oded Regev and Aravindan Vijayaraghavan. "On Learning Mixtures of Well-Separated Gaussians". In: (Oct. 2017), pp. 85–96. DOI: 10.1109/FOCS.2017.17.

A Omitted Details

Lemma 12. *Let* $(Y_1, Y_2, \dots, Y_s) \sim \mathcal{N}0, \Sigma)^{\otimes s}$, *where* $s = \frac{n}{k}$ *and* $\Sigma \preceq I$. *Then, for any unit vector* v,

$$\frac{1}{s}\mathbf{E}\sum_{i=1}^{s}\langle Y_i - \frac{1}{s}\sum_{j=1}^{s} Y_j, v\rangle^4 \leq O(1).$$

Proof. First, note that it is enough to consider $\Sigma = I$ as otherwise for standard Gaussian vectors Z_i, we have

$$\langle Y_i - \frac{1}{s}\sum_{j=1}^{s} Y_j, v\rangle = \langle \Sigma^{1/2} Z_i - \frac{1}{s}\sum_{j=1}^{s}\Sigma^{1/2} Z_j, v\rangle = \langle Z_i - \frac{1}{s}\sum_{j=1}^{s} Z_j, \Sigma^{1/2} v\rangle,$$

but $||\Sigma^{1/2}v|| \leq ||v|| \leq 1$. Now, for standard Gaussians, denoting $T_i = Y_i - \frac{1}{s}\sum_{j=1}^{s} Y_j \sim N(0, (1-o(1))I)$, we have the expression $\frac{1}{s}\mathbf{E}\sum_{i=1}^{s}\langle T_i, v\rangle^4$. It is enough to show that $\mathbf{E}\langle T_i, v\rangle^4 = O(1)$. Note, however, that

$$\mathbf{E}\langle T_i, v\rangle^4 = \sum_{p,q,r,\ell}(T_i)_p(T_i)_q(T_i)_r(T_i)_\ell v_p v_q v_r v_\ell.$$

Since trivially $(T_i)_p$, $(T_i)_q$ are independent mean 0 Gaussians whenever $p \neq q$, the above expression is equivalent to

$$\sum_{p,q}\mathbf{E}(T_i)_p^2(T_i)_q^2 v_p^2 v_q^2 \leq \sum_{p,q} v_p^2 v_q^2 = ||v||_2^4 \leq 1,$$

as desired. □

Using standard Gaussian concentration results, we also obtain the respective high-probability bounds. Intuitively, this works as

$$\frac{1}{s}\mathbf{E}\sum_{i=1}^{s}\langle Y_i - \frac{1}{s}\sum_{j=1}^{s} Y_j, v\rangle^4$$

is a symmetric function of ds iid Gaussians, so its gradient is relatively small. See more in [4].