Clustering and Sum of Squares Proofs: Six Blog Posts on Unsupervised Learning

Sam Hopkins

January 5, 2018

This note contains six blog posts comprising a tutorial on the Sum of Squares method in unsupervised learning, using clustering in mixtures of Gaussians as an extended example. They originally appeared on the Windows on Theory blog in December, 2017. I would like to thank Boaz Barak, Daniel Freund, Gautam "G" Kamath, Ilya Mironov, and David Steurer for the many improvements they suggested. Remaining errors are, of course, my own fault.

Please do not hesitate to write me with questions or to report errors, at sam.bkh@gmail.com.

Sam Hopkins
San Diego, CA
January, 2018

Contents

1 Unsupervised Learning, Mixture Models, and Identifiability

I am excited to (temporarily) join the Windows on Theory family as a guest blogger!

This is the first post in a series which will appear on Windows on Theory in the coming weeks. The aim is to give a self-contained tutorial on using the Sum of Squares algorithm for *unsupervised learning problems*, and in particular in Gaussian mixture models. This will take several posts: let's get started.

In an unsupervised learning problem, the goal is generally to recover some parameters $\theta \in \mathbb{R}^d$ given some data $X_1, \ldots, X_n \sim P(X \mid \theta)$, where P is a probability distribution on, say, $\mathbb{R}^p$ which is parameterized by θ. The goal is to estimate θ by some estimator $\hat{\theta}(X_1, \ldots, X_n)$ which (a) does not require too many samples and (b) can be computed from those samples in polynomial time. This basic setup can be instantiated (sometimes with minor adaptations) to capture numerous important problems in statistics and machine learning: a few examples are

- component analysis and its many variants (PCA, ICA, Sparse PCA, etc.)
- Netflix problem / matrix completion / tensor completion
- dictionary learning / blind source separation
- community detection and recovery / stochastic block models
- many clustering problems

Excellent resources on any of these topics are just a Google search away, and our purpose here is not to survey the vast literature on unsupervised learning, or even on provable "TCS-style" algorithms for these problems. Instead, we will try to give a simple exposition of one technique which has now been applied successfully to many unsupervised learning problems: the Sum of Squares method for turning *identifiability proofs* into algorithms.

Identifiability is a concept from statistics. If one hopes for an algorithm which recovers parameters $\hat{\theta}(X_1, \ldots, X_n) \approx \theta$, it must at least be true that $X_1, \ldots, X_n$ uniquely (or almost uniquely) determine θ, with high probability: when this occurs we say θ is identifiable from the samples $X_1, \ldots, X_n$.

Classically, identifiability is often proved by analysis of a (typically) inefficient brute-force algorithm. First, one invents some property $Q(X_1, \ldots, X_n)$ of θ. For example, in a maximum-likelihood argument, one shows that $Pr(X_1, \ldots, X_n \mid \theta) > p$ for some p. Then, often via a concentration-of-measure argument, one shows that no θ' far from θ satisfies property Q. In the maximum-likelihood example, this would entail showing that if θ' is far from θ then $Pr(X_1, \ldots, X_n \mid \theta') < p$.

An identifiability argument like this typically has no implications for computationally-efficient algorithms, since finding θ which satisfies Q often appears to require brute-force search. Thus, often when the investigation turns to efficient algorithms, the identifiability argument is abandoned and more explicitly algorithmic techniques are brought to bear: convex relaxations, spectral methods, and even heuristic methods.

The method we present here for designing computationally-efficient algorithms begins with a return to identifiability proofs. The main insight is that if both

- the property Q and, more importantly,
- the proof that any θ' satisfying Q must be close to θ

are sufficiently simple, then identifiability of θ from $X_1, \ldots, X_n$ does imply the existence of an efficient algorithm to (approximately) recover θ from $X_1, \ldots, X_n$!

For us, simple has a formal meaning: the proof should be captured in the low-degree Sum of Squares proof system.

The algorithms which result in the end follow a familiar recipe: solve some convex relaxation whose constraints depend on $X_1, \ldots, X_n$ and round it to obtain $\hat{\theta}(X_1, \ldots, X_n)$. But the design and analysis of this relaxation are heavily informed by the simple identifiability proof described above. In particular, the convex programs which result will not be familiar relaxations of maximum likelihood problems!

In this series of blog posts, we are going to carry out this strategy from start to finish for a classic unsupervised learning problem: clustering mixtures of Gaussians. So that we can get down to business as quickly as possible, we defer a short survey of the literature on this "proofs-to-algorithms" method to a later post.

1.1 Mixtures of Gaussians

Gaussian mixtures are classic objects in statistics, dating at least to Pearson in 1894. The basic idea is: suppose you have a data set $X_1, \ldots, X_n \in \mathbb{R}^d$ which was generated in a heterogeneous fashion: some points were sampled from probability distribution $\mathcal{D}_1$, some from $\mathcal{D}_2$, and so on up to $\mathcal{D}_k$, but you do not know which points came from which distributions. Under what settings can you cluster the points by which distribution they came from, and perhaps also recover some properties of those distributions, such as their means $\mu_i = \mathbb{E}_{X \sim \mathcal{D}_i} X$?

Pearson, in 1894, was faced with a collection of body measurements of crabs. The distribution of one such attribute in the data – the ratio of forehead length to body width – curiously deviated from a Gaussian distribution. Pearson concluded that in fact two distinct species of crabs were present in his data set, and that the data should therefore be modeled as a mixture of two Gaussians. He *invented the method of moments* to discover the means of both the Gaussians in question.[1] In the years since, Gaussian mixtures have become a fundamental statistical modeling tool: algorithms to fit Gaussian mixtures to data sets are included in basic machine learning packages like sklearn.

Here is our mixture of Gaussians model, formally.

Mixtures of separated spherical Gaussians: Let

- $\Delta > 0$.
- $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ be such that $\|\mu_i - \mu_j\| \geq \Delta$ for every $i \neq j$.
- $\mathcal{N}_1(\mu_1, I), \ldots, \mathcal{N}_k(\mu_k, I)$ be d-dimensional spherical Gaussians, centered at the means μ_i.

[1] Morit Hardt on Gaussian mixtures and Pearson's approach: http://blog.mrtz.org/2014/04/22/pearsons-polynomial.html

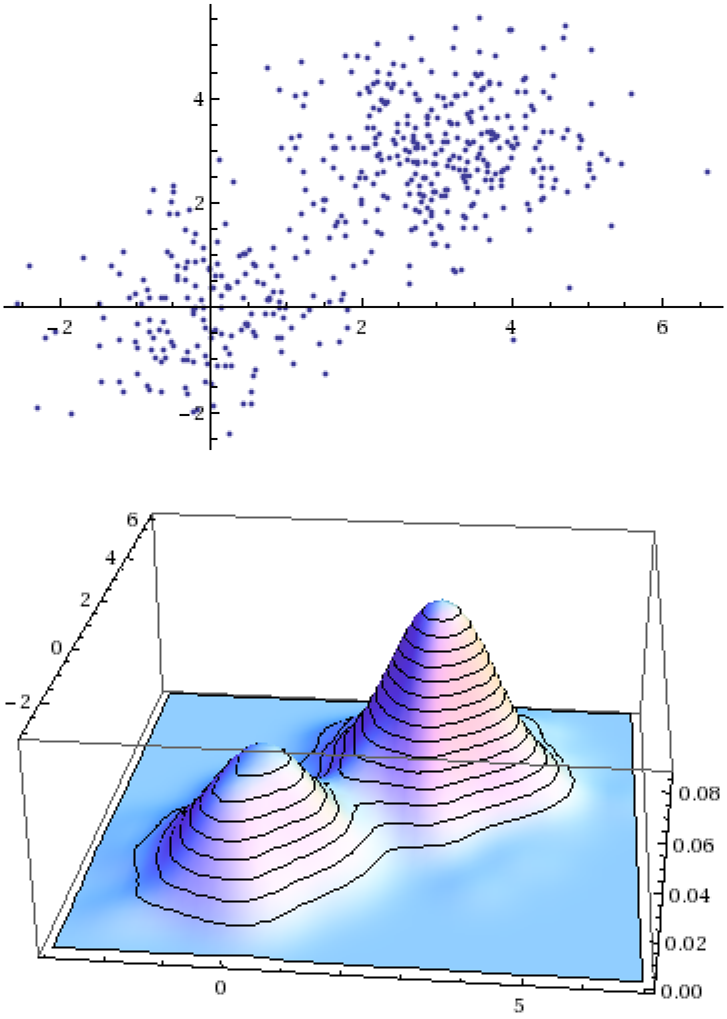

Figure 1: A mixture of 2 Gaussians in $\mathbb{R}^2$.

- $X_1, \ldots, X_n \in \mathbb{R}^d$ be iid samples, each drawn by selecting $j \sim [k]$ uniformly, then drawing $X \sim \mathcal{N}(\mu_j, I)$.
- $S_1, \ldots, S_k$ be the induced partition of $[n]$ into k parts, where $i \in S_j$ if samples X_i was drawn from $\mathcal{N}(\mu_j, I)$

The problem is: given $X_1, \ldots, X_n$, output a partition $T_1, \ldots, T_k$ of $[n]$ into k parts, each of size n/k, such that (up to renaming the clusters $1, \ldots, k$),

$$|S_i \cap T_i| \geq (1 - \delta) \cdot \frac{n}{k}$$

for each $i \in [k]$ and some small number $\delta > 0$.

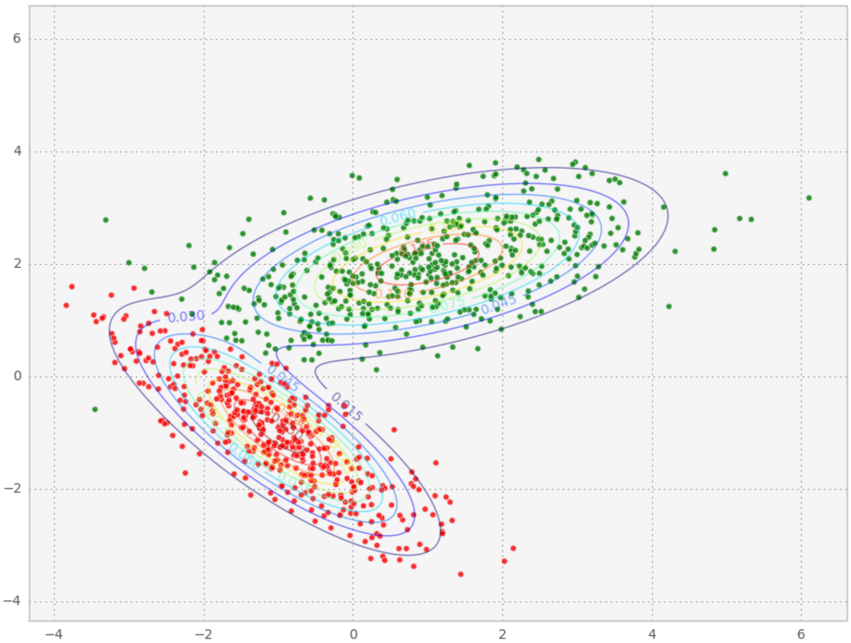

Figure 2: Another mixture of 2 Gaussians. The means have Euclidean distance about 3.5.

To avoid some minor but notationally annoying matters, we are going to work with a small variant of the model, where we assume that exactly n/k samples X_i came from each Gaussian $\mathcal{N}(\mu_j, I)$. We call a mixture like this "equidistributed".

We will work up to a proof of this theorem, which was proved recently (as of Fall 2017) in 3 independent works.

Theorem (Main Theorem (Hopkins-Li, Kothari-Steinhardt, Diakonikolas-Kane-Stewart)[2]). *For arbitrarily-large $t \in \mathbb{N}$ there is an algorithm requiring $n = d^{O(t)}k^{O(1)}$ samples from the equidistributed mixtures of Gaussians model and running in time $n^{O(t)}$ which outputs a partition $T_1, \ldots, T_k$ of $[n]$ into parts of size $N = n/k$ such that with high probability,*

$$\frac{|S_i \cap T_i|}{N} \geq 1 - k^{10} \cdot \left(\frac{C\sqrt{t}}{\Delta}\right)^t$$

for some universal constant C.[3]

In particular:

[2] See papers at: https://arxiv.org/abs/1711.07454 and https://arxiv.org/abs/1711.07465 and https://arxiv.org/abs/1711.07211

[3] To see how to apply the ideas in this tutorial to a much broader class of clustering problems, see my joint paper with Jerry Li and the recent paper of Pravesh Kothari and Jacob Steinhardt.

- If $\Delta = k^{\epsilon}$ for some $\epsilon > 0$, then by choosing $t = 100/\epsilon$ the algorithm recovers the correct clustering up to $1/\mathrm{poly}(k)$ errors in $\mathrm{poly}(k, d)$ time with $\mathrm{poly}(k, d)$-many samples.
- If $\Delta = C\sqrt{\log k}$ for a large-enough universal constant C, then choosing $t = O(\log k)$ gives a quasipolynomial-time algorithm (using quasipolynomially-many samples) to recover clusters up to $1/\mathrm{poly}(k)$ error.[4]

One (rather weak) consequence of the main theorem is that, for $n = d^{O(t)}k^{O(1)}$ samples, there is enough information in the samples to determine the underlying clustering, up to error $\delta(t, \Delta) = \frac{2^{O(t)} \cdot t^{t/2} \cdot k^{10}}{\Delta^t}$. Our strategy to prove the main theorem will start with proving the latter statement independently – as we have discussed, such an argument is often called a *proof of identifiability* because we say that the clusters are *identifiable* from the samples (up to $\delta(t, \Delta)$ errors).

While identifiability itself does not carry immediate algorithmic consequences, our proof of identifiability will be somewhat special: it will be “simple” in a formal sense, namely, that it will be captured by a formal proof system of restricted power. This simplicity of the proof of identifiability will almost immediately imply the main theorem: the construction of a computationally-efficient algorithm from a simple proof of identifiability is the heart of the proofs-to-algorithms method.

1.2 Identifiability proof: 1 dimension

Our eventual goal is to work up to a proof in the low-degree Sum of Squares proof system that clusters $S_1, \ldots, S_k$ are identifiable from samples $X_1, \ldots, X_n$ from a mixture of Gaussians. Since we have not yet defined low-degree Sum of Squares proofs, for now we will focus on constructing an identifiability proof which avoids mathematical facts which we deem "too complicated". In particular, we will avoid any Chernoff/union bound style arguments.

To get to the heart of the matter it helps to simplify the setting. Our first simplification is to restrict attention to the $d = 1$ case, so that distributions $\mathcal{N}(\mu_i, 1)$ are univariate Gaussians with unit variance.

Before stating our first lemma, let's discuss the key property of a collection $Y_1, \ldots, Y_m$ of samples from a Gaussian $\mathcal{N}(0, 1)$ which we will use. Recall that if $Y \sim \mathcal{N}(0, 1)$ is a standard Gaussian, then for every $t \in \mathbb{N}$,

$$\mathbb{E}|Y|^t \leq t^{t/2}.$$

If $Y_1, \ldots, Y_m$ are samples from Y, then for $m = m(t)$ large enough, the empirical distribution of $Y_1, \ldots, Y_m$ inherits this property, up to some small fluctuations. Namely, with high probability we would have

$$\mathbb{E}_{i \sim [m]} |Y_i|^t \leq 1.1 \cdot t^{t/2}.$$

[4] Before these recent works, the best polynomial-time algorithms for the clustering mixtures of Gaussians could not tolerate any $\Delta < k^{1/4}$ (when $\Delta \geq k^{1/4}$ a simple greedy algorithm can be shown to solve the clustering problem to high accuracy). On the other hand, known lower bounds show that when $\Delta = C\sqrt{\log k}$, clustering is impossible (even using exponential time) with $k^{O(1)}$ samples, so one cannot hope to improve the guarantees of this theorem too much further [Regev-Vijayaraghavan `https://arxiv.org/abs/1710.11592`]. (That said, reducing the sample complexity and running time to $\mathrm{poly}(d, k)$ when $\Delta = C\sqrt{\log k}$ is a fascinating open problem.)

(We could have replaced 1.1 by any small constant greater than 1.) Here, the notation $i \sim [m]$ means that an index i is chosen uniformly among $\{1, \ldots, m\}$.

If $Y \sim \mathcal{N}(\mu, 1)$ for some $\mu \in \mathbb{R}$, then the same discussion applies immediately to $\mathbb{E}|Y - \mu|^t$ and $\mathbb{E}_{i\sim[m]}|Y_i - \mu|^t$. But even more is true: if $\overline{\mu}$ is the empirical mean of $Y_1, \ldots, Y_m$ (that is, $\overline{\mu} = \mathbb{E}_{i\sim[m]}Y_i$), then with high probability the t-th centered empirical moment also inherits the same bound:

$$\mathbb{E}_{i\sim[m]}|Y_i - \overline{\mu}|^t \leq 1.1 \cdot t^{t/2}.$$

In the Gaussian mixture setting, so long as enough samples are drawn from each Gaussian $\mathcal{N}(\mu_i, 1)$, each cluster will have t-th empirical moments satisfying the above bound (with high probability).

In our identifiability proof, we choose to forget that the samples $X_1, \ldots, X_n$ were drawn from Gaussians, and we remember only that they break up into underlying clusters, each of which satisfies that empirical moment bound. We do not even remember the "true" means of each Gaussian; instead we talk only about the empirical means. We will show that no clustering far from that underlying ground-truth clustering results in clusters which satisfy the empirical moment bound.

Lemma (1). *Let $X_1, \ldots, X_n \in \mathbb{R}$. Let $S_1, \ldots, S_k$ be a partition of $[n]$ into k pieces of size $N = n/k$ such that for each S_i, the collection of numbers $\{X_j\}_{j\in S_i}$ obeys the following moment bound:*

$$\mathbb{E}_{j\sim S_i}|X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$$

where μ_i is the average $\mathbb{E}_{j\sim S_i}X_j$ and t is some number in $\mathbb{N}$. Let $\Delta > 0$ be such that $|\mu_i - \mu_j| \geq \Delta$ for every $i \neq j$. Suppose t is large enough that $10\sqrt{t}k^{1/t} \leq \Delta$.

Let $S \subseteq [n]$ have size $|S| = N = n/k$ and be such that $\{X_i\}_{i\in S}$ obey the same moment-boundedness property:

$$\mathbb{E}_{j\sim S}|X_j - \mu_S|^t < 2 \cdot t^{t/2}$$

for the same $t \in \mathbb{N}$, where μ_S is the mean $\mu_S = \mathbb{E}_{j\sim S}X_j$. Then there exists an S_i such that

$$\frac{|S \cap S_i|}{N} \geq 1 - k \cdot \left(\frac{C\sqrt{t}}{\Delta}\right)^t.$$

for some universal constant C.

How do we interpret Lemma 1 as a statement of cluster identifiability? The lemma implies that the clusters are, up to $\delta(t, \Delta)$ errors, the only subsets of $[n]$ of size n/k which satisfy the t-th moment bound. This is our property Q, like we discussed earlier in this post. The true clustering $S_1, \ldots, S_k$ satisfies Q (i.e. if you group $X_1, \ldots, X_n$ by this ground-truth clustering, the resulting clusters will have bounded empirical t-th moments), and every clustering which satisfies this bounded t-th moment property must be close to the true clustering. Thus, the correct clusters could be identified by searching for subsets of $[n]$ which satisfy the t-th moment bound (never mind that such a search would naively require about 2^n time).

We said that to use the sum of squares method to turn our identifiability proof into an algorithm, both the property Q and the proof of identifiability need to be *simple*.

This t-th moment boundedness property will turn out to be simple enough. What about the proof of Lemma 1? By the end of this post we will prove Lemma 1 in a way which uses only Holder's inequality for the t vs $\frac{t}{t-1}$ norms and the triangle inequality for the t-norm. Later, we will show that these inequalities are simple in the correct formal sense: they are captured by a proof system of restricted power.

Our proof of Lemma 1 relies on the following key fact.

Fact (1)**.** *Let $S, S' \subseteq \mathbb{R}$ have $|S| = |S'| = N$. Let X denote a uniform sample from S and similarly for X'. Let $\mu = \mathbb{E}X$ and $\mu' = \mathbb{E}X'$. Suppose X, X' satisfy the t-th moment bound*

$$\mathbb{E}|X - \mu|^t \leq 2 \cdot t^{t/2} \textit{ and } \mathbb{E}|X' - \mu'|^t \leq 2 \cdot t^{t/2}.$$

Then

$$|\mu - \mu'| \leq 4\sqrt{t} \cdot \left(\frac{|S \cap S'|}{N}\right)^{-1/t}.$$

A slightly more general interpretation of this fact is that a pair of random variables X, X' on $\mathbb{R}$ which have bounded t-th moments and whose total variation distance $TV(X, X') \leq 1 - \alpha$ cannot have means which are too far apart: $|\mathbb{E}X - \mathbb{E}X'| \leq 4\sqrt{t}/\alpha^{1/t}$.

Proof of Fact 1. Let $\alpha = |S \cap S'|/N$. Observe that there is a coupling of the random variables X, X' so that $Pr(X = X') = \alpha$. The coupling chooses with probability α to select a uniform sample $Y \sim S \cap S'$, then lets $X = X' = Y$. With probability $1 - \alpha$, the random variable X is a uniform sample from $S \setminus S'$ and similarly for X'.

Let (X, X') be a coupled pair of samples. We expand a quantity related to the one we want to bound, and then apply Holder's inequality with the t and $\frac{t}{t-1}$ norms. (The underlying inner product space assigns functions $f, g : (X, X') \mapsto \mathbb{R}$ the inner product $\mathbb{E}_{(X,X')} f(X, X') \cdot g(X, X')$.)

$$\begin{aligned}
\alpha \cdot |\mu - \mu'| &= \mathbb{E}_{(X,X')}\left[\mathbf{1}_{X=X'} \cdot |(\mu - X) - (\mu' - X')|\right] \\
&\leq \left(\mathbb{E}(\mathbf{1}_{X=X'})^{t/(t-1)}\right)^{\frac{t-1}{t}} \cdot \left(\mathbb{E}|(\mu - X) - (\mu' - X')|^t\right)^{1/t} \\
&= \alpha^{1-1/t} \cdot \left(\mathbb{E}|(\mu - X) - (\mu' - X')|^t\right)^{1/t}.
\end{aligned}$$

Now we can apply the triangle inequality for the t-norm to the last term, followed by our t-th moment assumptions.

$$\left(\mathbb{E}|(\mu - X) - (\mu' - X')|^t\right)^{1/t} \leq \left(\mathbb{E}|\mu - X|^t\right)^{1/t} + \left(\mathbb{E}|\mu' - X'|^t\right)^{1/t} \leq 4\sqrt{t}.$$

Putting everything together, we get $|\mu - \mu'| \leq \frac{4\sqrt{t}}{\alpha^{1/t}}$. □

Keeping in mind our eventual goal of constructing a low-degree Sum of Squares proof, we record the observation that the only inequalities we required to prove Fact 1 were the t vs. $\frac{t}{t-1}$ Holder's inequality and the triangle inequality for the t-norm.

Armed with Fact 1, we can prove Lemma 1. The main idea in the proof is that if S_1 and S_2 are the two clusters with greatest intersection with S, then μ_S can only be close to one of μ_1, μ_2.

Proof of Lemma 1. Let $S \subseteq [n]$ have size $|S| = n/k = N$, with mean $\mu_S = \mathbb{E}_{i\sim S} X_i$ and t-th moment bound $\mathbb{E}_{i\sim S}|X_i - \mu_S|^t \leq 2t^{t/2}$. Without loss of generality, order the clusters so that S_1 has largest intersection with S, then S_2, and so on: that is $|S \cap S_1| \geq |S \cap S_2| \geq \ldots \geq |S \cap S_k|$. If $|S \cap S_1| = (1-\delta)N$, then $|S \cap S_2| \geq \frac{\delta}{k} N$, just by counting.

Since $|\mu_1 - \mu_2| \geq \Delta$, either $|\mu_1 - \mu_S| \geq \Delta/2$ or $|\mu_2 - \mu_S| \geq \Delta/2$. We claim it must be the second. Using Fact 1,

$$|\mu_1 - \mu_S| \leq \frac{4\sqrt{t}}{(1-\delta)^{1/t}} \leq 4\sqrt{t} \cdot k^{1/t} < \Delta/2,$$

since certainly $(1-\delta) \geq \frac{1}{k}$, and we assumed $10\sqrt{t}k^{1/t} \leq \Delta$. We conclude that $|\mu_2 - \mu_S| \geq \Delta/2$.

We have obtained $\frac{|S \cap S_2|}{N} \geq \frac{\delta}{k}$ and $|\mu_2 - \mu_S| \geq \Delta/2$. Putting these together with Fact 1, we find

$$\frac{\Delta}{2} \leq |\mu_2 - \mu_S| \leq 4\sqrt{t} \cdot \left(\frac{k}{\delta}\right)^{1/t} .$$

Rearranging, this reads $\delta \leq \frac{2^{O(t)} t^{t/2} k}{\Delta^t}$. □

1.3 Looking Ahead

This concludes our one-dimensional identifiability proof, which will form the core of our proof of the Main Theorem. In our next post we will lift this proof to a Sum of Squares proof (for which we will need to define Sum of Squares proofs). After that, with a Sum of Squares proof in hand, we will finish designing our mixture of Gaussians algorithm for the one-dimensional case. Then we will show that the same ideas, nearly unchanged, imply that the algorithm works in higher dimensions.

1.4 Image credits

- Figure 1: Mathematica Stack Exchange. https://mathematica.stackexchange.com/questions/15055/finding-distribution-parameters-of-a-gaussian-mixture-distribution

2 SoS Proofs in One Dimension

Welcome back.

In the last post, we introduced Gaussian mixture models and the clustering problem for Gaussian mixtures. We described *identifiability proofs* for unsupervised learning problems. Then we set ourselves some goals:

- Design a simple identifiability proof for clustering in Gaussian mixtures, saying that if $X_1, \ldots, X_n$ are (enough) samples from a d-dimensional mixture of k Gaussians, the ground-truth clustering of the X_i's by which Gaussian they were drawn from is identifiable from the samples.

- Formalize the simplicity of that identifiability proof by showing that it is captured by a formal proof system of restricted power: the Sum of Squares (SoS) proof system.

- Guided by our SoS identifiability proof, design an algorithm for clustering in Gaussian mixtures. (And hence prove Theorem 1 from last time.)

In the last post, we accomplished task (1) in 1-dimensional case. In this post we will get started on task (2), again in the 1-dimensional case.

Recalling from last time, in our identifiability proof we remembered only two things things about our samples $X_1, \ldots, X_n$:

1. They break up into clusters $S_1, \ldots, S_k$ of equal size, $|S_i| = n/k = N$, such that for some $t \in \mathbb{N}$, each cluster obeys the empirical moment bound,

$$\mathbb{E}_{j \sim S_i} |X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$$

   where $\mu_i = \mathbb{E}_{j \sim S_i} X_j$ is the empirical mean of the cluster S_i, and

2. Those means are separated: $|\mu_i - \mu_j| \geq \Delta$.[5]

The key tool we used in our identifiability proof was Fact 1. We are going to give a Sum of Squares proof of that fact. Or rather: we are going to state a very similar fact, which concerns inequalities among low-degree polynomials, and give a Sum of Squares proof of that.

We are going to do things in a slightly unusual order, delaying definition of the SoS proof system till we have something concrete in mind to prove in it.

First, because SoS is a proof system to reason about inequalities among low-degree polynomials, we are going to formulate Fact 2, which is like Fact 1 except it will be explicitly about low-degree polynomials. Proving Fact 2 will be our goal.

Second, we will define the SoS proof system.

Finally, we will prove this Fact 2 in the low-degree SoS proof system.

2.1 Fact 2: an SoS version of Fact 1

Fact 1 concerns two subsets S, S' of $\mathbb{R}$. When we used Fact 1 to prove Lemma 1 in part 1, one of those sets S was one of the ground-truth clusters among $S_1, \ldots, S_k$, and one of them was a "candidate" cluster $S' \subseteq X_1, \ldots, X_n$ – Lemma 1 showed that the candidate cluster S' must in fact have been close to one of the true clusters $S_1, \ldots, S_k$.

We will design a system of polynomial equations whose solutions are in correspondence with candidate clusters S'. This is probably unavoidably foreign at first. We will offer some attempts at demystifying remarks later on, but for now we will forge ahead.

Let $X_1, \ldots, X_n \in \mathbb{R}$. Let $w_1, \ldots, w_n$ be some indeterminates; they will be the variables in our polynomials. We are going to think of them as 0/1 indicators of a subset $T \subseteq [n]$ which is a candidate cluster.

[5] Suspicious readers may note that our original Gaussian mixture model assumed that the *population* means are Δ-separated. Because we will draw enough samples that the empirical mean of each cluster is very close to the true (population) mean, this difference can be ignored for now and is easy to handle when we put everything together to analyze our final algorithm.

First, let's enforce that w is the indicator vector of a set of size $N = n/k$. Consider the equations

$$w_i^2 = w_i \text{ for all } i \in [n] \text{ and } \sum_{i \in [n]} w_i = N.$$

Any solution to these equations over $\mathbb{R}$ is a 0/1 indicator of a subset of $[n]$ of size N.

The second hypothesis Fact 1 places on a candidate cluster is the t-th moment bound. Let's enforce that with a polynomial inequality. First, we need some notation for the empirical mean μ of T. Denote by $\mu(w)$ the polynomial

$$\mu(w) = \frac{1}{N} \sum_{i \in [n]} w_i X_i.$$

Often we drop the w and just write μ. And now consider the inequality:

$$\frac{1}{N} \sum_{i \in [n]} w_i (X_i - \mu)^t \leq 2 \cdot t^{t/2}.$$

Belaboring the point somewhat: any solution over $\mathbb{R}$ to the equations and inequalities we have described would correspond to a subset of $[n]$ of which obeys the t-th empirical moment bound. (Here we assume that t is even, so that $|X_i - \mu|^t = (X_i - \mu)^t$.)

Although we don't have all the definitions we need in place yet, we will go ahead and state Fact 2. We introduce some suggestive notation. If $S \subseteq [n]$, we define the polynomial

$$|S \cap T|(w) = \sum_{i \in S} w_i.$$

Often we just write $|S \cap T|$.

Fact (2)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}$. Let $S \subseteq [n]$ have $|S| = N$; let $\mu_S = \mathbb{E}_{i \sim S} X_i$ be its mean. Let t be a power of 2. Suppose S satisfies*

$$\mathbb{E}_{i \sim S} |X_i - \mu_S|^t \leq 2 \cdot t^{t/2}.$$

Let $w_1, \ldots, w_n$ be indeterminates. Let $\mathcal{A}$ be the following set of equations and inequalities.

$$w_i^2 = w_i \text{ for } i \in [n]$$

$$\sum_{i \in [n]} w_i = N \qquad \frac{1}{N} \sum_{i \in [n]} w_i \cdot (X_i - \mu)^t \leq 2 \cdot t^{t/2}.$$

Then

$$\mathcal{A} \vdash_{O(t)} \left(\frac{|S \cap T|}{N} \right)^t \cdot (\mu - \mu_S)^t \leq 2^{O(t)} \cdot t^{t/2} \cdot \left(\frac{|S \cap T|}{N} \right)^{t-1}.$$

We have not yet defined the notation $\mathcal{A} \vdash_{O(t)} \ldots$. This notation means that that there is a degree $O(t)$ SoS proof of the inequality on the right-hand side using the axioms $\mathcal{A}$ – we will define this momentarily.

The purpose of stating Fact 2 now was just to convince the reader that there is a plausible version of Fact 1 which may be stated entirely as inequalities among polynomials. The reader is encouraged to compare the hypotheses and conclusions of Facts 1 and 2 (ignoring this $\mathcal{A} \vdash \ldots$ for now). The main differences are:[6]

a. The inequality in the conclusion of Fact 2 is raised to the t-th power, as compared to the conclusion of Fact 1.

b. The inequality in the conclusion of Fact 2 seems to have extra factors of $\frac{|S \cap T|}{N}$ on both sides.

Point (a) is needed just to make both sides of the inequality into polynomials in w; otherwise there would be fractional powers. The need for (b) is more subtle, and we will not be able to fully understand it for a while. For now, what we can say is: the inequality

$$\left(\frac{|S \cap T|}{N}\right) \cdot (\mu - \mu_S)^t \leq 2^{O(t)} \cdot t^{t/2}.$$

would be true for any w which solves $\mathcal{A}$, but that might not have an SoS proof.

One last difference is the factor $2^{O(t)}$ on the right-hand side of the conclusion. This is probably not inherent; it arises because we will use an easy-to-prove but lossy version of the triangle inequality in our SoS proof. In any case, it is dwarfed by the term $t^{t/2}$, so we are not too worried about it.

Enough dancing around these SoS proofs – in order to stop with the hand-waiving, we need to set up our proof system.

2.2 Sum of Squares Proofs

We cannot go any further without defining the formal proof system we will work in. The Sum of Squares proof system, henceforth "SoS", is a formal proof system for reasoning about systems of polynomial equations and inequalities over the real numbers. At its heart is the simple fact that if $p(x)$ is a polynomial in indeterminates x with real coefficients, then $p(x)^2 \geq 0$ for all $x \in \mathbb{R}$.

Plenty of elementary expositions on SoS proofs are available (see e.g. “SoS on the hypercube” (http://www.sumofsquares.org/public/lec01-2_definitions.html), “SoS on general domains” (http://www.sumofsquares.org/public/lec-definitions-general.html), and the first several sections of this paper by Ryan O’Donnell and Yuan Zhou (https://arxiv.org/pdf/1211.1958.pdf). We will define as much as we need to keep this tutorial self-contained but for expanded discussion we refer the reader to those resources and references therein.

Let $x_1, \ldots, x_n$ be some indeterminates. Let $\mathcal{P} = \{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$ be some polynomial inequalities in those indeterminates. If $r(x)$ is some other polynomial with real coefficients, it may be the case that for any x satisfying $\mathcal{P}$, also $r(x) \geq 0$; we would say that $\mathcal{P}$ implies $r(x) \geq 0$.

The key concept for us will be that $\mathcal{P}$ implies $r(x) \geq 0$ with a sum of squares proof. We say that $\mathcal{P}$ SoS-proves $r(x) \geq 0$ if there exist polynomials $b_S(x)$ for $S \subseteq [m]$ such that

$$r(x) = \sum_{S \subseteq [m]} b_S(x) \prod_{i \in S} p_i(x)$$

[6] To make Facts 1 and 2 look even more alike of course we could have introduced notations like $\mathbb{E}_{i \sim T}(X_i - \mu)^t$, but for concreteness we are keeping the w variables at least a little explicit.

and the polynomials $\{b_S\}$ are *sums of squares*—i.e. each of them has the form $\sum_j u_j(x)^2$ for some polynomials $u_j(x)$. Notice that if this equation holds, for any $x \in \mathbb{R}^n$ which satisfies the inequalities $\mathcal{P}$ the right-hand side must be nonnegative, so $r(x)$ is also nonnegative.

The polynomials b_S form an SoS proof that $\mathcal{P}$ implies $r(x) \geq 0$. If $\deg \left[b_S(x) \prod_{i \in S} p_i(x) \right]$ are all at most $d \in \mathbb{N}$, we say that the proof has degree d, and we write

$$\mathcal{P} \vdash_d r(x) \geq 0 .$$

When $\mathcal{P} = \emptyset$ we often just write $\vdash_d r(x) \geq 0$.

We commonly use a few shorthand notations.

- We write $\mathcal{P} \vdash r(x) \geq r'(x)$, by which we mean $\mathcal{P} \vdash r(x) - r'(x) \geq 0$.
- We include polynomial equations such as $x_i^2 = x_i$ in $\mathcal{P}$, by which we mean that $\mathcal{P}$ contains both $x_i^2 \leq x_i$ and $x_i^2 \geq x_i$.
- We write $\mathcal{P} \vdash r(x) = r'(x)$, by which we mean that both $\mathcal{P} \vdash r(x) \geq r'(x)$ and $\mathcal{P} \vdash r'(x) \geq r(x)$.

Although the definition suggests something static – there is a fixed collection of polynomials b_S forming an SoS proof – in practice we treat SoS proofs as dynamic objects, building them line by line much as we would any other proof. We are going to see an example of this very soon when we prove Fact 2.

It is time to begin to make good on the promise from the last section that we would get substantial mileage out of proving identifiability of mixtures of Gaussians using only simple inequalities. While it will take us several sections to completely make good on our promise, we can begin by giving SoS versions of the triangle and Holder's inequalities we used in our identifiability proof. We will prove one of these now to give the reader a sense of how such arguments go; since there is usually not much novelty in SoS proofs of such basic inequalities we will defer others till later.

We would like to emphasize that the following SoS proofs themselves have nothing to do with mixtures of Gaussians; instead they are part of a growing problem-independent toolkit of basic inequalities useful in designing SoS proofs of more interesting mathematical statements. The fact that one can have such a problem-independent toolkit is in part what makes the proofs-to-algorithms method so broadly useful.

SoS Triangle Inequality We will start with a fairly weak SoS triangle inequality, that will suffice for our needs. Much more sophisticated versions of this inequality are possible which allow various norms and do not lose the factor 2^t we do here.

Fact (SoS triangle inequality)**.** *Let x, y be indeterminates. Let t be a power of* 2*. Then*

$$\vdash_t (a+b)^t \leq 2^{t-1}(a^t + b^t) .$$

To prove the SoS triangle inequality we will want the following basic observation about composability of SoS proofs. (This is really a special case of a more general composibility result.)

Proposition (squaring SoS proofs). *Suppose* $\vdash_d p(x) \leq q(x)$ *and* p, q *are sums of squares. Then* $\vdash_{2d} p(x)^2 \leq q(x)^2$.

Proof of proposition. By hypothesis, $q(x) - p(x)$ is a sum of squares polynomial. Now,

$$q(x)^2 - p(x)^2 = (q(x) - p(x))(q(x) + p(x))$$

so it is a product of sum of squares polynomials, and hence itself a sum of squares. □

Proof of SoS triangle inequality. We start with the case $t = 2$. In this case, we have $(a+b)^2 = a^2+2ab+b^2$. We claim that $\vdash_2 2ab \leq a^2 + b^2$, since the polynomial $a^2 + b^2 - 2ab = (a-b)^2$ is a square. Hence, we find

$$\vdash_t (a+b)^2 = a^2 + 2ab + b^2 \leq 2(a^2 + b^2).$$

Now to prove the general case, we proceed by induction. We may suppose

$$\vdash_{t/2} (a+b)^{t/2-1} \leq 2^{t/2}(a^{t/2} + b^{t/2}).$$

By the proposition, this implies $\vdash_t (a+b)^t \leq 2^{t-2}(a^{t/2} + b^{t/2})^2$. Now we can apply the base case again to $(a' + b')^2$ for $a' = a^{t/2}$ and $b' = b^{t/2}$ to complete the argument. □

SoS Holder's inequality Holder's inequality poses a quandary for SoS proofs, because of the non-integral exponents in most p-norms (hence such norms do not naturally correspond to polynomials). Consequently, there are many SoS versions of Holder's inequality in the literature, choosing various ways to handle this non-integrality. The version we present here will be most useful for our mixtures of Gaussians proof. We will address the non-integral powers issue by imposing polynomial inequalities requiring that some of the underlying variables be Boolean.

Since the proof of this SoS Holder's inequality proceeds via a similar induction to the one we used for the SoS triangle inequality we just proved, we defer it to the end of this post.

Fact (SoS Holder's inequality). *Let* $w_1, \ldots, w_n$ *be indeterminates and let* $\mathcal{W}$ *be the collection of equations*

$$\mathcal{W} = \{w_i^2 - w_i = 0 \, : \, i \in [n]\}.$$

Note that the only solutions to $\mathcal{W}$ *are* $\{0,1\}^n$. *Let* $p_1(w), \ldots, p_n(w)$ *be polynomials of degree at most* ℓ. *Let* t *be a power of* 2. *Then*

$$\mathcal{W} \vdash_{O(t\ell)} \left(\sum_{i \in [n]} w_i \cdot p_i(w) \right)^t \leq \left(\sum_{i \in [n]} w_i \right)^{t-1} \cdot \sum_{i \in [n]} p_i(w)^t.$$

and

$$\mathcal{W} \vdash_{O(t\ell)} \left(\sum_{i \in [n]} w_i \cdot p_i(w) \right)^t \leq \left(\sum_{i \in [n]} w_i \right)^{t-1} \cdot \sum_{i \in [n]} w_i \cdot p_i(w)^t.$$

SoS Boolean Inequalities We will also want one more SoS inequality for our proof of Fact 1. In linear programming relaxations of Boolean problems, it is common to replace an integrality constraint $x \in \{0,1\}$ with the linear inequalities $0 \le x \le 1$. SoS can derive the latter inequalities.

Fact (SoS Boolean Inequalities)**.** $x^2 = x \vdash_2 0 \le x \le 1$.

Proof of Fact. For the inequality $x \ge 0$, just use the axiom $x \ge x^2$. For the inequality $x \le 1$, write

$$1 - x = x - x^2 + (1-x)^2. \quad \square$$

2.3 Proof of Fact 2

Without further ado, we will prove Fact 2 by lifting the proof of Fact 1 into the SoS proof system. Though slightly notationally cumbersome, the proof follows that of Fact 1 nearly line by line—the reader is encouraged to compare the two proofs.

Proof of Fact 2. We write out what we want to bound in terms of $X_1, \ldots, X_n$, then apply Holder's inequality and the triangle inequality.

$$\left(\sum_{i \in S} w_i\right)^t \cdot (\mu - \mu_S)^t = \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]\right)^t .$$

We deploy our SoS Holder's inequality to obtain

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]\right)^t \le \left(\sum_{i \in S} w_i\right)^{t-1} \cdot \sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]^t .$$

Next we can use our equations $w_i^2 - w_i$ to conclude that in fact

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]\right)^t \le \left(\sum_{i \in S} w_i^2\right)^{t-1} \cdot \sum_{i \in S} w_i^2 \left[(\mu - X_i) - (\mu_S - X_i)\right]^t .$$

The polynomial $\left(\sum_{i \in S} w_i^2\right)^{t-1}$ is a sum of squares, as is $2^t(a^t + b^t) - (a+b)^t$ via our SoS triangle inequality; applying this with $a = (\mu - X_i)$ and $b = -(\mu_S - X_i)$ we obtain

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]\right)^t \le 2^t \left(\sum_{i \in S} w_i^2\right)^{t-1} \cdot \sum_{i \in S} w_i^2 (\mu - X_i)^t + w_i^2 (\mu_S - X_i)^t .$$

We can add the sum of squares $2^t \left(\sum_{i \in S} w_i^2\right)^{t-1} \cdot \sum_{i \notin S} w_i^2 (\mu - X_i)^t + w_i^2 (\mu_S - X_i)^t$ to obtain

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i)\right]\right)^t \le 2^t \left(\sum_{i \in S} w_i^2\right)^{t-1} \cdot \sum_{i \in [n]} w_i^2 (\mu - X_i)^t + w_i^2 (\mu_S - X_i)^t .$$

Using the equations $w_i^2 - w_i$, which, as in the SoS Boolean inequalities fact can be used to prove $w_i^2 \leq 1$, we obtain

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i) \right] \right)^t \leq 2^t \left(\sum_{i \in S} w_i^2 \right)^{t-1} \cdot \sum_{i \in [n]} w_i (\mu - X_i)^t + (\mu_S - X_i)^t .$$

Finally using $\mathbb{E}_{i \in S} (\mu_S - X_i)^t \leq 2 \cdot t^{t/2}$ and $\mathcal{A} \vdash_{O(t)} \sum_{i \in [n]} w_i (X_i - \mu)^t \leq 2 \cdot t^{t/2} \cdot N$, we get

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i) \right] \right)^t \leq 2^{O(t)} \cdot \left(\sum_{i \in S} w_i^2 \right)^{t-1} \cdot t^{t/2} \cdot N .$$

Last of all, using $w_i^2 - w_i = 0$ to simplify the term $\sum_{i \in S} w_i^2$,

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S} w_i \left[(\mu - X_i) - (\mu_S - X_i) \right] \right)^t \leq 2^{O(t)} \cdot \left(\sum_{i \in S} w_i \right)^{t-1} \cdot N \cdot t^{t/2} .$$

The fact follows by rearranging. □

2.4 SoS proof of Holder's inequality

The last thing we haven't proved is our SoS Holder's inequality. We will need an SoS Cauchy-Schwarz inequality to prove it.

Fact (SoS Cauchy-Schwarz). *Let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be indeterminates. Then*

$$\vdash_2 \left(\sum_{i \in [n]} x_i y_i \right)^2 \leq \left(\sum_{i \in [n]} x_i^2 \right) \left(\sum_{i \in [n]} y_i^2 \right) .$$

Proof of SoS Cauchy-Schwarz. It is not hard to check that

$$\left(\sum_{i \in [n]} x_i^2 \right) \left(\sum_{i \in [n]} y_i^2 \right) - \left(\sum_{i \in [n]} x_i y_i \right)^2 = \sum_{i,j \in [n]} (x_i y_j - x_j y_i)^2$$

which is a sum of squares. □

Proof of SoS Holder's inequality. We start with the case $t = 2$. Using SoS Cauchy-Schwarz, we obtain

$$\vdash_{2\ell+2} \left(\sum_{i \in [n]} w_i \cdot p_i(w) \right)^2 \leq \left(\sum_{i \in [n]} w_i^2 \right) \cdot \left(\sum_{i \in [n]} p_i(w)^2 \right) .$$

This follows from our SoS Cauchy-Schwarz inequality by substituting w_i for x_i and $p_i(w)$ for y_i; we proved a fact about a sum of squares polynomial in x, y which implies a corresponding fact about a sum of squares in variables w_i and $p_i(w)$. The latter in turn is a sum of squares in w.

To finish the first half of the $t = 2$ case, we just need to replace w_i^2 with w_i on the right-hand side. By adding the polynomials $(w_i^2 - w_i) \cdot \left(-\sum_{i\in[n]} p_i(w)^2\right)$ via the equations $\mathcal{W}$, we obtain

$$\mathcal{W} \vdash_{2\ell+2} \left(\sum_{i\in[n]} w_i \cdot p_i(w)\right)^2 \leq \left(\sum_{i\in[n]} w_i\right) \cdot \left(\sum_{i\in[n]} p_i(w)^2\right).$$

To establish the second inequality for the $t = 2$ base case, we start by again adding multiples of $(w_i^2 - w_i)$ to get

$$\mathcal{W} \vdash_{2\ell+2} \left(\sum_{i\in[n]} w_i \cdot p_i(w)\right)^2 = \left(\sum_{i\in[n]} w_i^2 \cdot p_i(w)\right)^2.$$

Then the inequality follows from Cauchy-Schwarz and again adding some multiples of $(w_i^2 - w_i)$.

Now it's time for the induction step. We can assume

$$\mathcal{W} \vdash_{O(t/2\cdot\ell)} \left(\sum_{i\in[n]} w_i \cdot p_i(w)\right)^{t/2} \leq \left(\sum_{i\in[n]} w_i\right)^{t/2-1} \cdot \left(\sum_{i\in[n]} w_i p_i(w)^{t/2}\right).$$

By again adding multiples of $w_i^2 - w_i$, we obtain

$$\mathcal{W} \vdash_{O(t/2\cdot\ell)} \left(\sum_{i\in[n]} w_i \cdot p_i(w)\right)^{t/2} \leq \left(\sum_{i\in[n]} w_i^2\right)^{t/2-1} \cdot \left(\sum_{i\in[n]} w_i^2 p_i(w)^{t/2}\right)$$

Now both sides are sums of squares. So, by squaring, we find

$$\mathcal{W} \vdash_{O(t\cdot\ell)} \left(\sum_{i\in[n]} w_i \cdot p_i(w)\right)^{t} \leq \left(\sum_{i\in[n]} w_i^2\right)^{t-2} \cdot \left(\sum_{i\in[n]} w_i^2 \cdot p_i(w)^{t/2}\right)^2.$$

The proof is finished by applying Cauchy-Schwarz to the last term and cleaning up by adding multiples of $w_i^2 - w_i$ as necessary. □

2.5 Looking Ahead

In the next post, we will use Fact 2 to deduce an SoS version of Lemma 1 (from part 1). Subsequently, we will finish designing an algorithm for one-dimensional clustering, proving the Main Theorem (from part 1) in the one-dimensional case. Then we will get high dimensional.

3 SoS Proofs in One Dimension, continued

3.1 The Story So Far

Let's have a brief recap. We are designing an algorithm to cluster samples from Gaussian mixture models on $\mathbb{R}^d$. Our plan is to do this by turning a simple identifiability proof into an algorithm. For us, "simple" means that the proof is captured by the low degree Sum of Squares (SoS) proof system.

We have so far addressed only the case $d = 1$ (which will remain true in this post). In part 1 we designed our identifiability proof, not yet trying to formally capture it in SoS. The proof was simple in the sense that it used only the triangle inequality and Holder's inequality. In part 2 we defined SoS proofs formally, and stated and proved an SoS version of one of the key facts in the identifiability proof (Fact 2).

In this post we are going to finish up our SoS identifiability proof. In the next post, we will see how to transform the identifiability proof into an algorithm. Setting

We recall our setting formally. Although our eventual goal is to cluster samples $X_1, \ldots, X_n$ sampled from a mixture of k Gaussians, we decided to remember only a few properties of such a collection of samples, which will hold with high probability.

The properties are:

1. They break up into clusters $S_1, \ldots, S_k$ of equal size, $|S_i| = n/k = N$, such that for some $t \in \mathbb{N}$, each cluster obeys the empirical moment bound,

   $$\mathbb{E}_{j \sim S_i} |X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$$

   where $\mu_i = \mathbb{E}_{j \sim S_i} X_j$ is the empirical mean of the cluster S_i, and

2. Those means are separated: $|\mu_i - \mu_j| \geq \Delta$.

The main statement of cluster identifiability was Lemma 1. Our main goal in this post is to state and prove an SoS version of Lemma 1. We have already proved the following Fact 2, an SoS analogue of Fact 1 which we used to prove Lemma 1.

3.2 Remaining obstacles to an SoS version of Lemma 1

We are going to face a couple of problems.

1. The statement and proof of Lemma 1 are not sufficiently symmetric for our purposes – it is hard to phrase things like "there exists a cluster S_i such that..." as statements directly about polynomials. We will handle this by giving more symmetric version of Lemma 1, with a more symmetric proof.

2. Our proof of Lemma 1 uses the conclusion of Fact 1 in the form

   $$|\mu - \mu'| \leq 4\sqrt{t} \cdot \left(\frac{|S \cap S'|}{N} \right)^{-1/t}$$

   whereas Fact 2 concludes something slightly different:

   $$\mathcal{A} \vdash_{O(t)} \left(\frac{|S \cap T|}{N} \right)^t \cdot (\mu - \mu_S)^t \leq 2^{O(t)} \cdot t^{t/2} \cdot \left(\frac{|S \cap T|}{N} \right)^{t-1} .$$

   The difference in question is that the polynomials in Fact 2 are degree t, and $\frac{\sum_{i \in S} w_i}{N}$ appears on both sides of the inequality. If we were not worried about SoS proofs, we could just cancel terms in the second inequality and take t-th roots to obtain the first, but these operations are not necessarily allowed by the SoS proof system.

One route to handling this would be to state and prove a version of Lemma 1 which concerns only degree t. This is probably possible but definitely inconvenient. Instead we will exhibit a common approach to situations where it would be useful to cancel terms and take roots but the SoS proof system doesn't quite allow it: we will work simultaneously with SoS proofs and with their dual objects, pseudodistributions.

We will tackle issues (1) and (2) in turn, starting with the (a)symmetry issue.

3.3 Lemma 1 reformulated: maintaining symmetry

We pause here to record an alternative version of Lemma 1, with an alternative proof. This second version is conceptually the same as the one we gave in part 1, but it avoids breaking the symmetry among the clusters $S_1, \ldots, S_k$, whereas this was done at the very beginning of the first proof, by choosing the ordering of the clusters by $|S \cap S_i|$. Maintaining this symmetry requires a slight reformulation of the proof, but will eventually make it easier to phrase the proof in the Sum of Squares proof system. In this proof we will also avoid the assumption $10\sqrt{t}k^{1/t} \leq \Delta$, however, we will pay a factor of k^2 rather than k in the final bound.

Lemma (Alternative version of Lemma 1)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}$. Let $S_1, \ldots, S_k$ be a partition of $[n]$ into k pieces of size $N = n/k$ such that for each S_i, the collection of numbers $\{X_j\}_{j \in S_i}$ obeys the following moment bound:*

$$\mathbb{E}_{j \sim S_i} |X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$$

where μ_i is the average $\mathbb{E}_{j \sim S_i} X_j$ and t is some number in $\mathbb{N}$. Let $\Delta > 0$ be such that $|\mu_i - \mu_j| \geq \Delta$ for every $i \neq j$.

Let $S \subseteq [n]$ have size $|S| = N = n/k$ and be such that $\{X_i\}_{i \in S}$ obey the same moment-boundedness property:

$$\mathbb{E}_{j \sim S} |X_j - \mu_S|^t < 2 \cdot t^{t/2}.$$

for the same $t \in \mathbb{N}$, where $\mu_S = \mathbb{E}_{j \sim S} X_j$. Then

$$\sum_{i \in [k]} \left(\frac{|S_i \cap S|}{N} \right)^2 \geq 1 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}.$$

We remark on the conclusion of this alternative version of Lemma 1. Notice that $c_i = |S_i \cap S|/N$ are nonnegative numbers which sum to 1. The conclusion of the lemma is that $\sum_{i \in [n]} c_i^2 \geq 1 - \delta$ for $\delta = \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}$. Since the sum of their squares is at least $1 - \delta$, one obtains

$$1 - \delta \leq \sum_{i \in [k]} c_i^2 \leq \max_{i \in [k]} c_i \cdot \sum_{i \in [k]} c_i = \max_{i \in [k]} c_i,$$

matching the conclusion of our first version of Lemma 1 up to an extra factor of k.

Proof of alternative version of Lemma 1. Let $S \subseteq [n]$ again have size $|S| = n/k = N$ with mean $\mu_S = \mathbb{E}_{i \sim S} X_i$ and t-th moment bound $\mathbb{E}_{i \sim S} |X_i - \mu_S|^t \leq 2t^{t/2}$. Since $S_1, \ldots, S_k$ partition $[n]$,

$$\sum_{i,j \in [k]} |S_i \cap S| \cdot |S_j \cap S| = \left(\sum_{i \in [k]} |S_i \cap S| \right)^2 = N^2 .$$

We will endeavor to bound $|S_i \cap S| \cdot |S_j \cap S|$ for every pair $i \neq j$. Since $|\mu_i - \mu_S| + |\mu_j - \mu_S| \geq \Delta$,

$$\left(\frac{|S_i \cap S|}{N} \right)^{1/t} \cdot \left(\frac{|S_j \cap S|}{N} \right)^{1/t} \leq \frac{|\mu_i - \mu_S| + |\mu_j - \mu_S|}{\Delta} \cdot \left(\frac{|S_i \cap S|}{N} \right)^{1/t} \cdot \left(\frac{|S_j \cap S|}{N} \right)^{1/t} .$$

Certainly $|S_i \cap S|/N \leq 1$ and similarly for S_j, so this is at most

$$\frac{1}{\Delta} \left[|\mu_i - \mu_S| \left(\frac{|S_i \cap S|}{N} \right)^{1/t} + |\mu_j - \mu_S| \left(\frac{|S_j \cap S|}{N} \right)^{1/t} \right] .$$

Using Fact 1, this in turn is at most $\frac{2}{\Delta} \cdot 4\sqrt{t}$. So, we obtained

$$|S_i \cap S| \cdot |S_j \cap S| \leq \frac{2^{O(t)} t^{t/2}}{\Delta^t} \cdot N^2$$

for every $i \neq j$.

Putting this together with our first bound on $\sum_{i,j \in [k]} |S_i \cap S| \cdot |S_j \cap S|$, we get

$$\sum_{i \in [k]} |S_i \cap S|^2 \geq N^2 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t} \cdot N^2 . \quad \square$$

Now that we have resolved the asymmetry issue in our earlier version of Lemma 1, it is time to move on to pseudodistributions, the dual objects of SoS proofs, so that we can tackle the last remaining hurdles to proving an SoS version of Lemma 1.

3.4 Pseudodistributions and duality

Pseudodistributions are the convex duals of SoS proofs. As with SoS proofs, there are several expositions covering elementary definitions and results in detail (e.g. the lecture notes of Barak and Steurer, here and here). We will define what we need to keep the tutorial self-contained but refer the reader elsewhere for further discussion. Here we follow the exposition in those lecture notes.

As usual, let $x_1, \ldots, x_n$ be some indeterminates. For a finitely-supported function $\mu : \mathbb{R}^n \mapsto \mathbb{R}$ and a function $f : \mathbb{R}^n \to \mathbb{R}$, define

$$\tilde{\mathbb{E}}_\mu f = \sum_{x \in \text{supp}(\mu)} \mu(x) \cdot f(x) .$$

If μ defines a probability distribution, then $\tilde{\mathbb{E}}_\mu$ is the operator sending a function to its expectation under μ.

A finitely-supported μ is a degree d pseudodistribution if

1. $\tilde{\mathbb{E}}_\mu 1 = \sum_{x \in \mathrm{supp}(\mu)} \mu(x) = 1$
2. $\tilde{\mathbb{E}}_\mu p(x)^2 \geq 0$ for every polynomial $p(x)$ of degree at most $d/2$.

When μ is clear from context, we usually suppress it and write $\tilde{\mathbb{E}} f$. Furthermore, if $\tilde{\mathbb{E}} : \{\text{polynomials}\} \to \mathbb{R}$ is an operator and $\tilde{\mathbb{E}} = \tilde{\mathbb{E}}_\mu$ for some pseudodistribution μ, we often abuse terminology and call $\tilde{\mathbb{E}}$ a pseudoexpectation.

If $\mathcal{P} = \{p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$ is a family of polynomial inequalities and μ is a degree d pseudodistribution, we say μ satisfies $\mathcal{P}$ if for every $S \subseteq [m]$ and $q(x)$ such that $\deg \left[q(x)^2 \cdot \prod_{i \in S} p_i(x)\right] \leq d$ one has

$$\tilde{\mathbb{E}}_\mu \left[q(x)^2 \cdot \prod_{i \in S} p_i(x) \right] \geq 0.$$

We are not going to rehash the basic duality theory of SoS proofs and pseudodistributions here, but we will need the following basic fact, which is easy to prove from the definitions.

Fact (weak soundness of SoS proofs)**.** *Suppose $\mathcal{P} \vdash_\ell r(x) \geq 0$ and that μ is a degree d pseudodistribution which satisfies $\mathcal{P}$. Then for every SoS polynomial h, if $\deg h + \ell \leq d$ then $\tilde{\mathbb{E}} h(x) r(x) \geq 0$.*

We call this "weak soundness" because somewhat stronger statements are available, which more readily allow several SoS proofs to be composed. See Barak and Steurer's notes[7] for more.

The following fact exemplifies what we mean in the claim that pseudodistributions help make up for the inflexibility of SoS proofs to cancel terms in inequalities.

Fact (pseudoexpectation Cauchy-Schwarz)**.** *Let $\tilde{\mathbb{E}}$ be a degree d pseudoexpectation on indeterminates $x_1, \ldots, x_n$. Let $p(x)$ and $q(x)$ be polynomials of degree at most $d/2$. Then*

$$\tilde{\mathbb{E}} p(x) q(x) \leq \left(\tilde{\mathbb{E}} p(x)^2\right)^{1/2} \left(\tilde{\mathbb{E}} q(x)^2\right)^{1/2}.$$

As a consequence, if $\tilde{\mathbb{E}}$ has degree dt and t is a power of 2, by induction

$$\tilde{\mathbb{E}} p(x) \leq \left(\tilde{\mathbb{E}} p(x)^t\right)^{1/t}.$$

Proof of pseudoexpectation Cauchy-Schwarz. For variety, we will do this proof in the language of matrices rather than linear operators. Let M be the matrix indexed by monomials among $x_1, \ldots, x_n$ of degree at most d, with entries $M(x^\alpha, x^\beta) = \tilde{\mathbb{E}} x^\alpha x^\beta$. If p is a polynomial of degree at most $d/2$, we can think of p as a vector indexed by monomials (whose entries are the coefficients of p) such that $p^\top M q = \tilde{\mathbb{E}} p(x) q(x)$. Hence,

$$\tilde{\mathbb{E}} p(x) q(x) = \langle M^{1/2} p, M^{1/2} q \rangle \leq \|M^{1/2} p\| \|M^{1/2} q\| = \left(\tilde{\mathbb{E}} p(x)^2\right)^{1/2} \left(\tilde{\mathbb{E}} q(x)^2\right)^{1/2}. \quad \square$$

We will want a second, similar fact.

Fact (pseudoexpectation Holder's)**.** *Let p be a degree ℓ sum of squares polynomial, $t \in \mathbb{N}$, and $\tilde{\mathbb{E}}$ a degree $O(t\ell)$ pseudoexpectation. Then*

$$\tilde{\mathbb{E}} p(x)^{t-2} \leq \left(\tilde{\mathbb{E}} p(x)^t\right)^{\frac{t-2}{t}}.$$

The proof of pseudoexpectation Holder's is similar to several we have already seen; it can be found as Lemma A.4 in this paper by Barak, Kelner, and Steurer (https://arxiv.org/abs/1312.6652).

[7] http://www.sumofsquares.org/public/lec-definitions-general.html

3.5 Lemma 2: an SoS version of Lemma 1

We are ready to state and prove our SoS version of Lemma 1. The reader is encouraged to compare the statement of Lemma 2 to the alternative version of Lemma 1. The proof will be almost identical to the proof of the alternative version of Lemma 1.

Lemma (2). *Let $X_1, \ldots, X_n \in \mathbb{R}$. Let $S_1, \ldots, S_k$ be a partition of $[n]$ into k pieces of size $N = n/k$ such that for each S_i, the collection of numbers $\{X_j\}_{j \in S_i}$ obeys the following moment bound:*

$$\mathbb{E}_{j \sim S_i} |X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$$

where μ_i is the average $\mathbb{E}_{j \sim S_i} X_j$ and t is a power of 2 in $\mathbb{N}$. Let $\Delta > 0$ be such that $|\mu_i - \mu_j| \geq \Delta$ for every $i \neq j$.

Let $w_1, \ldots, w_n$ be indeterminates. Let $\mathcal{A}$ be the following set of equations and inequalities.

$$\begin{aligned} & w_i^2 = w_i \text{ for } i \in [n] \\ & \sum_{i \in [n]} w_i = N \\ & \frac{1}{N} \sum_{i \in [n]} w_i \cdot (X_i - \mu)^t \leq 2 \cdot t^{t/2}. \end{aligned}$$

As before $\mu = \mu(w)$ is the polynomial $\frac{1}{N} \sum_{i \in [n]} w_i X_i$. Thinking of the variables w_i as defining a set $T \subseteq [n]$ via its 0/1 indicator, let $|T \cap S_j|$ be the formal expression

$$|T \cap S_j| = \sum_{i \in S_j} w_i.$$

Let $\tilde{\mathbb{E}}$ be a degree $O(t)$ pseudoexpectation which satisfies $\mathcal{A}$. Then

$$\tilde{\mathbb{E}} \left[\sum_{i \in [k]} \left(\frac{|T \cap S_i|}{N} \right)^2 \right] \geq 1 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}.$$

Proof of Lemma 2. We will endeavor to bound $\tilde{\mathbb{E}} |T \cap S_i| \cdot |T \cap S_j|$ from above for every $i \neq j$. Since we want to use the degree t polynomials in Fact 2, we get started with

$$\tilde{\mathbb{E}} |T \cap S_i| |T \cap S_j| \leq \left(\tilde{\mathbb{E}} |T \cap S_i|^t |T \cap S_j|^t \right)^{1/t}$$

by (repeated) pseudoexpectation Cauchy-Schwarz.

Since $\mu_i = \mathbb{E}_{\ell \sim S_i} X_\ell$ and $\mu_j = \mathbb{E}_{\ell \sim S_j} X_\ell$ are Δ-separated, i.e. $|\mu_i - \mu_j|^t \geq \Delta^t$, we also have

$$\vdash_t (\mu_i - \mu)^t + (\mu_j - \mu)^t \geq 2^{-t} \left[(\mu_i - \mu) - (\mu_j - \mu) \right]^t \geq 2^{-t} \Delta^t$$

where the indeterminate is μ and we have only used the SoS triangle inequality. Hence,

$$\tilde{\mathbb{E}} |T \cap S_i|^t |T \cap S_j|^t \leq \tilde{\mathbb{E}} \left[\frac{(\mu_i - \mu)^t + (\mu_j - \mu)^t}{2^{-t} \Delta^t} |T \cap S_i|^t |T \cap S_j|^t \right].$$

Applying Fact 2 and soundness to the right-hand side, we get

$$\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t \leq 2^{O(t)}t^{t/2}\Delta^{-t} \cdot N \cdot \left(\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^{t-1} + \tilde{\mathbb{E}}|T \cap S_i|^{t-1}|T \cap S_j|^t\right).$$

Now using that $\mathcal{A} \vdash_t w_i^2 \leq 1$ and hence $\mathcal{A} \vdash_t |T \cap S_i| \leq N$ and similarly for $|T \cap S_j|$, we get

$$\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t \leq 2^{O(t)}t^{t/2}\Delta^{-t} \cdot N^2 \cdot \tilde{\mathbb{E}}|T \cap S_i|^{t-1}|T \cap S_j|^{t-1}.$$

By pseudoexpectation Cauchy-Schwarz

$$\tilde{\mathbb{E}}|T \cap S_i|^{t-1}|T \cap S_j|^{t-1} \leq (\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t)^{1/2}(\tilde{\mathbb{E}}|T \cap S_i|^{t-2}|T \cap S_j|^{t-2})^{1/2}$$

which, combined with the preceding, rearranges to

$$\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t \leq \frac{2^{O(t)}t^tN^4}{\Delta^{2t}}\tilde{\mathbb{E}}|T \cap S_i|^{t-2}|T \cap S_j|^{t-2}.$$

By pseudoexpectation Holder's,

$$\tilde{\mathbb{E}}|T \cap S_i|^{t-2}|T \cap S_j|^{t-2} \leq \left(\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t\right)^{(t-2)/t}.$$

All together, we got

$$\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t \leq \frac{2^{O(t)}t^tN^4}{\Delta^{2t}}\left(\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t\right)^{(t-2)/t}.$$

Now we no longer have to worry about SoS proofs; we can just cancel the terms on either side of the inequality to get

$$\tilde{\mathbb{E}}|T \cap S_i||T \cap S_j| \leq \left(\tilde{\mathbb{E}}|T \cap S_i|^t|T \cap S_j|^t\right)^{1/t} \leq \frac{2^{O(t)}t^{t/2}N^2}{\Delta^t}.$$

Putting this together with

$$\tilde{\mathbb{E}}\sum_{ij \in [k]}|T \cap S_i||T \cap S_j| = \tilde{\mathbb{E}}\left(\sum_{i \in [n]} w_i\right)^2 = N^2$$

finishes the proof. □

4 Clustering One-Dimensional Gaussian Mixtures

Last time we finished our SoS identifiability proof for one-dimensional Gaussian mixtures. In this post, we are going to turn it into an algorithm.

We will we design a convex program to exploit the SoS identifiability proof, and in particular Lemma 2. Then we describe a (very simple) rounding procedure and analyze it, which will complete our description and analysis of the one-dimensional algorithm.

Let's look at the hypothesis of Lemma 2. It asks for a pseudoexpectation of degree $O(t)$ which satisfies the inequalities $\mathcal{A}$. First of all, note that the inequalities $\mathcal{A}$ depend only on the vectors $X_1, \ldots, X_n$ to be clustered, and in particular not on the hidden partition $S_1, \ldots, S_k$, so they are fair game to use in our algorithm. Second, it is not too hard to check that the set of pseudoexpectations satisfying $\mathcal{A}$ is convex, and in fact the feasible region of a semidefinite program with $n^{O(t)}$ variables!

It is actually possible to design a rounding algorithm which takes any pseudoexpectation satisfying $\mathcal{A}$ and produces a cluster, up to about a $2^{O(t)} t^{t/2} \text{poly}(k)/\Delta^t$-fraction of misclassified points. Then the natural approach to design an algorithm to find all the clusters is to iterate:

(1) find such a pseudoexpectation $\tilde{\mathbb{E}}$ via semidefinite programming

(2) round to find a cluster T

(3) remove all the points in T from $X_1, \ldots, X_n$, go to (1).

This is a viable algorithm, but analyzing it is a little painful because misclassifications from early rounds of the rounding algorithm must be taken into account when analyzing later rounds, and in particular a slightly stronger version of Lemma 2 is needed, to allow some error from early misclassifications.

We are going to avoid this pain by imposing some more structure on the pseudoexpectation our algorithm eventually rounds, to enable our rounding scheme to recover all the clusters without re-solving a convex program. This is not possible if one is only promised a pseudoexpectation $\tilde{\mathbb{E}}$ which satisfies $\mathcal{A}$: observe, for example, that one can choose the pseudodistribution μ to be a probability distribution supported on one point $w \in \{0,1\}^n$, the indicator of cluster S_1. This particular $\tilde{\mathbb{E}}$ is easy to round to extract S_1, but contains no information about the remaining clusters $S_2, \ldots, S_k$.

We are going to use a trick reminiscent of entropy maximization to ensure that the pseudoexpectation we eventually round contains information about all the clusters $S_1, \ldots, S_k$. Our convex program will be:

$$\min \|\tilde{\mathbb{E}} w w^\top\| \text{ such that } \tilde{\mathbb{E}} \text{ satisfies } \mathcal{A},$$

where $\|\tilde{\mathbb{E}} w w^\top\|$ is the Frobenious norm of the matrix $\tilde{\mathbb{E}} w w^\top$.

It may not be so obvious why $\|\tilde{\mathbb{E}} w w^\top\|$ is a good thing to minimize, or what it has to do with entropy maximization. We offer the following interpretation. Suppose that instead of pseudodistributions, we were able to minimize $\|\mathbb{E}_{a \sim \mu} a a^\top\|$ over all μ which are supported on vectors $a_1, \ldots, a_k \in \{0,1\}^n$ which are the 0/1 indicators of the clusters $S_1, \ldots, S_k$. Such a distribution μ is specified by nonnegative $\mu(a_i) \in \mathbb{R}$ for $i \in [k]$ which sum to 1, and the Frobenious norm is given by

$$\|\mathbb{E}_{a \sim \mu} a a^\top\|^2 = \left\langle \sum_{i \in [k]} \mu(a_i) a_i a_i^\top, \sum_{i \in [k]} \mu(a_i) a_i a_i^\top \right\rangle = \sum_{i \in [k]} \mu(a_i)^2 \cdot \|a_i\|^4,$$

where we have used orthogonality $\langle a_i, a_j \rangle = 0$ if $i \neq j$. Since all the clusters have size n/k, we have $\|a_i\|^4 = (n/k)^2$, and we have $\|\mathbb{E}_{a \sim \mu} a a^\top\| = \|\mu\| \cdot (n/k)^2$, where $\|\mu\|$ is the 2-norm, or collision probability, of μ. This collision probability is minimized when μ is uniform.

We can analyze our convex program via the following corollary of Lemma 2.

Corollary (1)**.** *Let $X_1, \ldots, X_n$ and $S_1, \ldots, S_k$ be as in Lemma 2. Let $\tilde{\mathbb{E}}$ be the degree $O(t)$ pseudoexpectation solving*

$$\min \|\tilde{\mathbb{E}} ww^\top\| \text{ such that } \tilde{\mathbb{E}} \text{ satisfies } \mathcal{A}.$$

Let μ be the uniform distribution over vectors $a_1, \ldots, a_k \in \{0,1\}^n$ where a_i is the $0/1$ indicator of cluster S_i. Then

$$\|\tilde{\mathbb{E}} ww^\top - \mathbb{E}_\mu aa^\top\| \leq \|\mathbb{E} aa^\top\| \cdot \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}\right)^{1/2}.$$

where $\| \cdot \|$ is the Frobenious norm.

Proof of Corollary 1. The uniform distribution μ over $a_1, \ldots, a_k$ is a feasible solution to the convex program with $\|\mathbb{E}_{a \sim \mu} aa^\top\|^2 = (n/k)^2 \cdot 1/k$, by the calculation preceding the corollary. So if $\tilde{\mathbb{E}}$ is the minimizer, we know $\|\tilde{\mathbb{E}} ww^\top\|^2 \leq n^2/k^3$.

We expand the norm:

$$\|\tilde{\mathbb{E}} ww^\top - \mathbb{E}_\mu aa^\top\|^2 = \|\tilde{\mathbb{E}} ww^\top\|^2 + \|\mathbb{E}_\mu aa^\top\|^2 - 2\langle \tilde{\mathbb{E}} ww^\top, \mathbb{E} aa^\top \rangle \leq 2\left(\frac{n^2}{k^3} - \langle \tilde{\mathbb{E}} ww^\top, \mathbb{E} aa^\top \rangle\right).$$

To bound the last term we use Lemma 2. In the notation of that Lemma,

$$\langle \tilde{\mathbb{E}} ww^\top, \mathbb{E} aa^\top \rangle = \frac{1}{k} \tilde{\mathbb{E}} \sum_{i \in [k]} |S_i \cap T|^2 \geq \frac{1}{k} \cdot N^2 \cdot \left(1 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}\right).$$

Remember that $N = n/k$. Putting these together, we get

$$\|\tilde{\mathbb{E}} ww^\top - \mathbb{E}_\mu aa^\top\| \leq \|\mathbb{E} aa^\top\| \cdot \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}\right)^{1/2}. \quad \square$$

We are basically done with the algorithm now. Observe that the matrix $\mathbb{E}_\mu aa^\top$ contains all the information about the clusters $S_1, \ldots, S_k$. In fact the clusters can just be read off of the rows of $\mathbb{E} aa^\top$.

Once one has in hand a matrix $M = \tilde{\mathbb{E}} ww^\top$ which is close in Frobenious norm to $\mathbb{E} aa^\top$, extracting the clusters is still a matter of reading them off of the rows of the matrix (choosing rows at random to avoid hitting one of the small number of rows which could be wildly far from their sisters in $\mathbb{E} aa^\top$).

We will prove the following fact at the end of this post.

Fact (rounding)**.** *Let $S_1, \ldots, S_k$ be a partition of $[n]$ into k parts of size $N = n/k$. Let $A \in \{0,1\}^{n \times n}$ be the $0/1$ indicator matrix for same-cluster membership. That is, $A_{ij} = 1$ if i and j are in the same cluster S_ℓ. Suppose $M \in \mathbb{R}^{n \times n}$ satisfies $\|M - A\| \leq \epsilon \|A\|$.*

There is a polynomial-time algorithm which takes M and with probability at least $1 - O(k^2 \epsilon^2)$ produces a partition $T_1, \ldots, T_k$ of $[n]$ into clusters of size N such that, up to a permutation of $[k]$,

$$\frac{|S_i \cap T_i|}{N} \geq 1 - O(\epsilon^2 k).$$

4.1 Putting things together for one-dimensional Gaussian mixtures

Now we sketch a proof of the Main Theorem in the case $d = 1$. Our algorithm is: given $X_1, \ldots, X_n$, solve

$$\arg\min \|\tilde{\mathbb{E}} w w^\top\| \text{ such that } \tilde{\mathbb{E}} \text{ is degree } O(t) \text{ and satisfies } \mathcal{A}$$

then apply the rounding algorithm from the rounding fact to $\tilde{\mathbb{E}} w w^\top$ and output the resulting partition.

If the vectors $X_1, \ldots, X_n$ satisfy the hypothesis of Lemma 2, then by Corollary 1, we know

$$\|\tilde{\mathbb{E}} w w^\top - \mathbb{E}_\mu a a^\top\| \leq \|\mathbb{E}_\mu a a^\top\| \cdot \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t} \right)^{1/2}$$

where μ is the uniform distribution over 0/1 indicators for clusters $S_1, \ldots, S_k$. Hence the rounding algorithm produces a partition $T_1, \ldots, T_k$ of $[n]$ such that

$$\frac{|S_i \cap T_i|}{N} \geq 1 - O\left(\frac{2^{O(t)} t^{t/2} k^3}{\Delta^t} \right).$$

Since a standard Gaussian has $\mathbb{E}|X|^t \leq t^{t/2}$, elementary concentration shows that the vectors $X_1, \ldots, X_n$ satisfy the hypotheses of Lemma 1 with probability 0.99 so long as $n \geq \text{poly}(k)$.

4.2 Rounding algorithm

The last thing to do in this post is prove the rounding algorithm Fact. This has little to do with SoS; the algorithm is elementary and combinatorial. We provide it for completeness.

The setting is: there is a partition $S_1, \ldots, S_k$ of $[n]$ into k parts of size n/k. Let $A \in \{0,1\}^{n \times n}$ be the 0/1 indicator matrix for cluster membership; i.e. $A_{ij} = 1$ if and only if i and j are in the same cluster S_ϵ. Given a matrix M such that $\|M - A\| \leq \epsilon \|A\|$, the goal is to recover a partition $T_1, \ldots, T_k$ of $[n]$ which is close to $S_1, \ldots, S_k$ up to a permutation of $[k]$.

Let M_i be the i-th row of M and similarly for A_i. Let $\delta > 0$ be a parameter we will set later.

The algorithm is:

(1) Let $\mathcal{I} = [n]$ be the set of active indices.

(2) Pick $i \sim \mathcal{I}$ uniformly.

(3) Let $S \subseteq \mathcal{I}$ be those indices j for which $\|M_j - M_i\| \leq \delta \cdot \sqrt{n/k}$.

(4) Add S to the list of clusters and let $\mathcal{I} := \mathcal{I} \setminus S$.

(5) If $|\mathcal{I}| > n/2k$, go to (2).

(6) (postprocess) Assign remaining indices to clusters arbitrarily, then move indices arbitrarily from larger clusters to smaller ones until all clusters have size n/k.

Fact (rounding). *If $\delta = 0.01$ then with probability at least $1 - O(k^2\epsilon^2)$ the rounding algorithm outputs disjoint clusters $T_1, \ldots, T_k$, each of size n/k, such that up to a permutation of $[k]$, $|T_i \cap S_i| \geq (1 - O(\epsilon^2 k)) \cdot \frac{n}{k}$.*

Proof. Call an index $i \in [n]$ *good* if $\|M_i - A_i\| \leq \frac{\delta}{2} \cdot \sqrt{n/k} = \frac{\delta}{2}\|A_i\|$. An index is bad if it is not good. By hypothesis $\sum_{i \in [n]} \|M_i - A_i\|^2 = \|M - A\|^2 \leq \epsilon^2 \|A\|^2$. Each bad index contributes at least $\frac{\delta^2}{4}\|A_i\|^2$ to the left side and $\|A_i\|^2 = \|A\|^2/n$, so there are at most $4\epsilon^2 n/\delta^2$ bad indices.

If i, j are good indices and both are in the same cluster S_ϵ, then if the algorithm chooses i, the resulting cluster will contain j. If $\delta < 0.1$, then also if j is good but is in some other cluster $S_{\epsilon'}$, the cluster formed upon choosing i will not contain j. Thus if the algorithm never chooses a bad index, before postprocessing the clusters $T_1, \ldots, T_k$ it outputs will (up to a global permutation of $[k]$) satisfy that T_i contains all the good indices in S_i. Hence in this case only bad indices can be misclassified, so the postprocessing step moves at most $4\epsilon^2 n/\delta^2$ indices, and in the end the cluster T_i again errs from S_i on at most $8\epsilon^2 n/\delta^2$ indices.

Consider implementing the algorithm by drawing a list of k^2 indices before seeing M (i.e. obliviously), then when the algorithm requires random index $i \in \mathcal{I}$ we give it the next index in our list which is in $\mathcal{I}$ (and halt with no output if no such index exists). It's not hard to see that this implementation fails only with probability at most $O(1/k)$. Furthermore, by a union bound the list contains a bad index only with probability $O(k^2\epsilon^2/\delta^2)$. Choosing $\delta = 0.01$ thus completes the proof. □

5 Warming up for high dimensions

Last time we finished our algorithm design and analysis for clustering one-dimensional Gaussian mixtures. Clustering points on $\mathbb{R}$ isn't much of a challenge. In this post we will finally move to the high-dimensional setting. We will see that most of the ideas and arguments so far carry over nearly unchanged.

In keeping with the method we are advocating throughout the posts, the first thing to do is return to the non-SoS cluster identifiability proof from Part 1 and see how to generalize it to collections of points in dimension $d > 1$. We encourage the reader to review that proof.

5.1 Generalizing the non-SoS Identifiability Proof

Our first step in designing that proof was to correctly choose a property of a collection of samples $X_1, \ldots, X_n$ from a Gaussian mixture which we would rely on for identifiability. The property we chose was that the points $X_1, \ldots, X_n$ break into clusters $S_1, \ldots, S_k$ of equal size so that each cluster has bounded empirical t-th moments and the means of the clusters are separated.

Here is our first attempt at a high-dimensional generalization: $X_1, \ldots, X_n$ break into k clusters $S_1, \ldots, S_k$ of equal size $N = n/k$ such that

(1) for each cluster S_i and $u \in \mathbb{R}^d$,

$$\mathbb{E}_{j \sim S_i} \langle X_j - \mu_i, u \rangle^t \leq 2 \cdot t^{t/2}$$

where $\mu_i = \mathbb{E}_{j \sim S_i} X_j$ is the empirical mean of cluster S_i, and

(2) those means are separated: $\|\mu_i - \mu_j\| \geq \Delta$ for $i \neq j$.

The first property says that every one-dimensional projection of every cluster has Gaussian t-th moments. The second should be familiar: we just replaced distance on the line with ℓ_2 distance in $\mathbb{R}^d$.

The main steps in our one-dimensional non-SoS identifiability proofs were Fact 1 and Lemma 1. We will give an informal discussion on their high-dimensional generalizations; for the sake of brevity we will skip a formal non-SoS identifiability proof this time and go right to the SoS proof.

The key idea is: for any pair of sets $S, S' \subseteq [n]$ such that $\{X_j\}_{i \in S}$ and $\{X_j\}_{i \in S'}$ satisfy the empirical t-th moment bound (2) with respect to empirical means μ and μ' respectively, if $|\mu - \mu'| \geq \Delta$, then by the one-dimensional projections

$$\left\{ \left\langle X_j - \mu, \frac{\mu - \mu'}{\|\mu - \mu'\|} \right\rangle \right\}_{j \in S} \text{ and } \left\{ \left\langle X_j - \mu, \frac{\mu - \mu'}{\|\mu - \mu'\|} \right\rangle \right\}_{j \in S'}$$

are collections of numbers in $\mathbb{R}$ which satisfy the hypotheses of our one-dimensional identifiability arguments. All we did was choose the right one-dimensional projection of the high-dimensional points $\{X_j\}$ to capture the separation between μ and μ'. (The reader is encouraged to work this out for themselves; it is easiest shift all the points so that without loss of generality $\mu = 0$.)

5.2 Obstacles to High-dimensional SoS Identifiability

We are going to face two main difficulties in turning the high-dimensional non-SoS identifiability proofs into SoS proofs.

(1) The one-dimensional projections above have $\|\mu - \mu'\|$ in the denominator, which is not a low-degree polynomial. This is easy to handle, and we have seen similar things before: we will just clear denominators of all inequalities in the proofs, and raise both sides to a high-enough power that we get polynomials.

(2) The high-dimensional t-th moment bound has a "for all u" quantification. That is, if $w = w_1, \ldots, w_n$ are indeterminates as in our one-dimensional proof, to be interpreted as the $0/1$ indicators for membership in a candidate cluster $T \subseteq [n]$, we would like to enforce

$$\max_{u \in \mathbb{R}^d} \frac{1}{N} \sum_{i \in [n]} w_i \langle X_i - \mu, u \rangle^t \leq 2 \cdot t^{t/2} \cdot \|u\|^t .$$

Because of the max, this is not a polynomial inequality in w. This turns out to be a serious problem, and it will require us to strengthen our assumptions about the points $X_1, \ldots, X_n$.

In order for the SoS algorithm to successfully cluster $X_1, \ldots, X_n$, it needs to *certify* that each of the clusters it produces satisfies the t-th empirical moment property. Exactly why this is so, and whether it would also be true for non-SoS algorithms, is an interesting topic for discussion. But, for the algorithm to succeed, in particular a short certificate of the above inequality must exist! It is probably not true that such a certificate exists for an arbitrary collection of points in $\mathbb{R}^d$ satisfying the t-th empirical moment bound. Thus, we will add the existence of such a certificate as an assumption on our clusters.

When $Y_1, \ldots, Y_m$ are sufficiently-many samples from a d-dimensional Gaussian, the following matrix inequality is a short certificate of the t-th empirical moment property:

$$\left\| \frac{1}{N} \sum_{i \in [m]} \left(Y_i^{\otimes t/2}\right) \left(Y_i^{\otimes t/2}\right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2}\right) \left(Y^{\otimes t/2}\right)^\top \right\|_F \leq 1$$

where the norm $\| \cdot \|_F$ is Frobenious norm (spectral norm would have been sufficient but the inequality is easier to verify with Frobenious norm instead, and this just requires taking a few more samples). This inequality says that the empirical t-th moment matrix of $Y_1, \ldots, Y_m$ is close to its expectation in Frobenious norm. It certifies the t-th moment bound, because for any $u \in \mathbb{R}^d$, we would have

$$\mathbb{E}_{i \sim [m]} \langle Y_i, u \rangle^t \leq \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \langle Y, u \rangle^t + \|u\|^t \leq 2 \cdot t^{t/2} \cdot \|u\|^t$$

by analyzing the quadratic forms of the empirical and true t-th moment matrices at the vector $u^{\otimes t/2}$.

In our high-dimensional SoS identifiability proof, we will remember the following things about the samples $X_1, \ldots, X_n$ from the underlying Gaussian mixture.

1. $X_1, \ldots, X_n$ break into clusters $S_1, \ldots, S_k$, each of size $N = n/k$, so that if $\mu_i = \mathbb{E}_{j \sim S_i} X_j$ is the empirical mean of the i-th cluster, $\|\mu_i - \mu_j\| \geq \Delta$ if $i \neq j$, and

2. For each cluster S_i:

$$\left\| \mathbb{E}_{j \sim S_i} \left([X_j - \mu_i]^{\otimes t/2}\right) \left([X_j - \mu_i]^{\otimes t/2}\right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2}\right) \left(Y^{\otimes t/2}\right)^\top \right\|_F^2 \leq 1$$

.

5.3 Algorithm for High-Dimensional Clustering

Now we are prepared to describe our high-dimensional algorithm for clustering Gaussian mixtures. For variety's sake, this time we are going to describe the algorithm before the identifiability proof. We will finish up the high-dimensional identifiability proof, and hence the analysis of the following algorithm, in the next post, which will be the last in this series.

Given a collection of points $X_1, \ldots, X_n \in \mathbb{R}^d$, let $\mathcal{B}$ be the following set of polynomial inequalities in indeterminates $w_1, \ldots, w_n$:

- $w_i^2 = w_i$ for all $i \in [n]$
- $\sum_{i \in [n]} w_i = N$
- $$\left\| \frac{1}{N} \sum_{i \in [n]} w_i \left([X_i - \mu]^{\otimes t/2}\right) \left([X_i - \mu]^{\otimes t/2}\right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2}\right) \left(Y^{\otimes t/2}\right)^\top \right\|_F^2 \leq 1$$

  where as usual $\mu(w) = \frac{1}{N} \sum_{i \in [n]} w_i X_i$.

The algorithm is: given $X_1, \ldots, X_n, k, t$, find a degree-$O(t)$ pseudoexpectation $\tilde{\mathbb{E}}$ of minimal $\|\tilde{\mathbb{E}} w w^\top\|_F$ satisfying $\mathcal{B}$. Run the rounding procedure from the one-dimensional algorithm on $\tilde{\mathbb{E}} w w^\top$.

6 High dimensional clustering: analysis

Last time we developed our high-dimensional clustering algorithm for Gaussian mixtures. In this post we will make our SoS identifiability proofs high-dimensional. In what is hopefully a familiar pattern by now, these identifiability proofs will also amount to an analysis of the clustering algorithm from part 5. At the end of this post is a modest discussion of some of the literature on the SoS proofs-to-algorithms method we have developed in this series.

6.1 Setting

We decided to remember the following properties of a collection $X_1, \ldots, X_n$ of samples from a d-dimensional Gaussian mixture model.

(1) $X_1, \ldots, X_n$ break up into clusters $S_1, \ldots, S_k \subseteq [n]$ which partition $[n]$, and each S_i has size exactly $|S_i| = N = n/k$.

(2) Each cluster has bounded moments, and this has a small certificate: for some $t \in \mathbb{N}$, if $\mu_i = \mathbb{E}_{j \sim S_i} X_j$ is the mean of the i-th cluster,

$$\left\| \mathbb{E}_{j \sim S_i} \left([X_j - \mu_i]^{\otimes t/2} \right) \left([X_j - \mu_i]^{\otimes t/2} \right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^\top \right\|_F^2 \leq 1.$$

(3) The means are separated: if $i \neq j$ then $\|\mu_i - \mu_j\| \geq \Delta$.

Our identifiability proof follows the template we have laid out in the non-SoS proof from part 1 and the one-dimensional SoS proof from later parts. The first thing to do is prove a key fact about a pair of overlapping clusters with bounded moments.

6.2 Fact 3

The next fact is the high-dimensional analogue of Fact 2. (We are not going to prove a high-dimensional analogue of Fact 1; the reader should at this point have all the tools to work it out for themselves.) We remind the reader of the key family polynomial inequalities.

Given a collection of points $X_1, \ldots, X_n \in \mathbb{R}^d$, let $\mathcal{B}$ be the following set of polynomial inequalities in indeterminates $w_1, \ldots, w_n$:

- $w_i^2 = w_i$ for all $i \in [n]$
- $\sum_{i \in [n]} w_i = N$
- $$\left\| \frac{1}{N} \sum_{i \in [n]} w_i \left([X_i - \mu]^{\otimes t/2} \right) \left([X_i - \mu]^{\otimes t/2} \right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^\top \right\|_F^2 \leq 1$$

where as usual $\mu(w) = \frac{1}{N} \sum_{i \in [n]} w_i X_i$. As before, for a subset $S \subseteq [n]$, we use the notation $|T \cap S| = |T \cap S|(w) = \sum_{i \in S} w_i$.

Fact (3)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$. Let $S \subseteq [n]$ have $|S| = N$; let $\mu_S = \mathbb{E}_{i \sim S} X_i$ be its mean. Let t be a power of 2. Suppose S satisfies*

$$\left\| \mathbb{E}_{j \sim S} \left([X_j - \mu_S]^{\otimes t/2} \right) \left([X_j - \mu_S]^{\otimes t/2} \right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^\top \right\|_F^2 \leq 1.$$

Then

$$\mathcal{B} \vdash_{O(t)} \left(\frac{|T \cap S|}{N} \right)^{2t} \|\mu - \mu_S\|^{4t} \leq 2^{O(t)} \cdot t^t \cdot \left(\frac{|T \cap S|}{N} \right)^{2(t-1)} \cdot \|\mu - \mu_S\|^{2t}.$$

The main difference between the conclusions of Fact 3 and Fact 2 is that both sides of the inequality are multiplied by $\|\mu - \mu_S\|^t$, as compared to Fact 2. As we will see, this is because an additional dependence on the vector-valued polynomial $(\mu - \mu_S)(w)$ is introduced by the need to project the high-dimensional vectors X_i onto the line $\mu - \mu_S$. We have already tackled a situation where an SoS-provable inequality seemed to require cancelling terms of left and right in order to be useful (i.e. when we used Fact 2 to prove Lemma 2), and similar ideas will work here.

The main innovation in proving Fact 3 is the use of the t-th moment inequality in $\mathcal{B}$. Other than that, we follow the proof of Fact 2 almost line by line. The main proposition needed to use the t-th moment inequality is:

Proposition. $\vdash_t \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \langle Y, u \rangle^t \leq t^{t/2} \cdot \|u\|^t$.

Proof of Proposition. Expanding the polynomial $\mathbb{E}_{Y \sim \mathcal{N}(0,I)} \langle Y, u \rangle^t$ we get

$$\mathbb{E}_{Y \sim \mathcal{N}(0,I)} \langle Y, u \rangle^t = \sum_{|\alpha| = t, \alpha \text{ even}} u^\alpha \cdot \mathbb{E} Y^\alpha$$

where α is a multi-index over $[n]$ and "even" means that every index in α occurs with even multiplicity. (Other terms vanish by symmetry of Y.) Since α is always even in the sum, the monomial u^α is a square, and $\mathbb{E} Y^\alpha \leq t^{t/2}$ by standard properties of Gaussian moments. Hence,

$$\vdash_t \sum_{|\alpha| = t, \alpha \text{ even}} u^\alpha \cdot \mathbb{E} Y^\alpha \leq t^{t/2} \cdot \sum_{|\alpha| = t, \alpha \text{even}} u^\alpha = t^{t/2} \cdot \|u\|^t$$

□

Proof of Fact 3. As usual, we write things out in terms of $X_1, \ldots, X_n$, then apply Holder's inequality and the triangle inequality. First of all,

$$\left(\sum_{i \in S} w_i \right)^t \|\mu - \mu_S\|^{2t} = \left(\sum_{i \in S} w_i \langle \mu - \mu_S, \mu - \mu_S \rangle \right)^t = \left(\sum_{i \in S} w_i \langle (\mu - X_i) - (\mu_S - X_i), \mu - \mu_S \rangle \right)^t .$$

By SoS Holder's inequality, we get

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i \in S} w_i \right)^t \|\mu - \mu_S\|^{2t} \leq \left(\sum_{i \in S} w_i \right)^{t-1} \cdot \sum_{i \in S} w_i \langle (\mu - X_i) - (\mu_S - X_i), \mu - \mu_S \rangle^t .$$

By the same reasoning as in Fact 2, using $\vdash_t (a-b)^t \leq 2^t(a^t+b^t)$ and $w_i = w_i^2 \leq 1$ we get

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i\in S} w_i\right)^t \|\mu - \mu_S\|^{2t} \leq 2^t \cdot \left(\sum_{i\in S} w_i\right)^{t-1} \cdot \left[\sum_{i\in[n]} w_i \langle X_i - \mu, \mu - \mu_S\rangle^t + \sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t\right].$$

By our usual squaring and use of $(a+b)^2 \leq 2(a^2+b^2)$, we also get

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i\in S} w_i\right)^{2t} \|\mu - \mu_S\|^{4t} \leq 2^{O(t)} \cdot \left(\sum_{i\in S} w_i\right)^{2(t-1)} \cdot \left[\left(\sum_{i\in[n]} w_i \langle X_i - \mu, \mu - \mu_S\rangle^t\right)^2 + \left(\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t\right)\right].$$

(We want both sides to be squared so that we are set up to eventually use SoS Cauchy-Schwarz.) We are left with the last two terms, which are t-th moments in the direction $\mu - \mu_S$. If we knew

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i\in[n]} w_i \langle X_i - \mu, \mu - \mu_S\rangle^t\right)^2 \leq 2^{O(t)} t^t \cdot \|\mu - \mu_S\|^{2t} \cdot N^2$$

and similarly

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t\right)^2 \leq 2^{O(t)} t^t \cdot \|\mu - \mu_S\|^{2t} \cdot N^2$$

then we would be done.

We start with the second inequality. We write the polynomial on the LHS as

$$\frac{1}{N}\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t = \mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t + \left(\frac{1}{N}\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t - \mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t\right)$$

Squaring again as usual, it would be enough to bound both $\left(\mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t\right)^2$ and $\left(\frac{1}{N}\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t - \mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t\right)^2$. For the former, using the Proposition above we get

$$\vdash_{2t} \left(\mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t\right)^2 \leq 2^{O(t)} t^t \|\mu - \mu_S\|^{2t}.$$

For the latter, notice that

$$\left(\frac{1}{N}\sum_{i\in S}\langle X_i - \mu_S, \mu - \mu_S\rangle^t - \mathbb{E}_{Y\sim\mathcal{N}(0,I)}\langle Y, \mu - \mu_S\rangle^t\right)^2 = \left\langle M, \left[(\mu - \mu_S)^{\otimes t/2}\right]\left[(\mu - \mu_S)^{\otimes t/2}\right]^\top\right\rangle$$

where M is the matrix

$$M = \mathbb{E}_{j\sim S}\left([X_j - \mu_S]^{\otimes t/2}\right)\left([X_j - \mu_S]^{\otimes t/2}\right)^\top - \mathbb{E}_{Y\sim\mathcal{N}(0,I)}\left(Y^{\otimes t/2}\right)\left(Y^{\otimes t/2}\right)^\top.$$

Hence by SoS Cauchy-Schwarz, we get

$$\vdash_{O(t)} \left\langle M, \left[(\mu - \mu_S)^{\otimes t/2}\right]\left[(\mu - \mu_S)^{\otimes t/2}\right]^\top\right\rangle^2 \leq \|M\|_F^2 \cdot \|\mu - \mu_S\|^{2t} \leq \|\mu - \mu_S\|^{2t}.$$

Putting these together, we get

$$\mathcal{B} \vdash_{O(t)} \left(\sum_{i \in S} \langle X_i - \mu_S, \mu - \mu_S \rangle^t \right)^2 \leq 2^{O(t)} t^t \cdot \|\mu - \mu_S\|^{2t} \cdot N^2,$$

the second of the inequalities we wanted. Proving the first one is similar, using the hypothesis

$$\left\| \frac{1}{N} \sum_{i \in [n]} w_i \left([X_i - \mu]^{\otimes t/2} \right) \left([X_i - \mu]^{\otimes t/2} \right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^\top \right\|_F^2 \leq 1$$

in place of $\|M\|_F^2 \leq 1$. □

6.3 Lemma 3

The last thing is to use Fact 3 to prove Lemma 3, our high-dimensional SoS identifiability lemma. Predictably, it is almost identical to our previous proof of Lemma 2 using Fact 2.

Lemma (3)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$. Let $S_1, \ldots, S_k$ be a partition of $[n]$ into k pieces of size $N = n/k$ such that for each S_i, the collection of vectors $\{X_j\}_{j \in S_i}$ obeys the following moment bound:*

$$\left\| \mathbb{E}_{j \sim S_i} \left([X_j - \mu_i]^{\otimes t/2} \right) \left([X_j - \mu_i]^{\otimes t/2} \right)^\top - \mathbb{E}_{Y \sim \mathcal{N}(0,I)} \left(Y^{\otimes t/2} \right) \left(Y^{\otimes t/2} \right)^\top \right\|_F^2 \leq 1.$$

where μ_i is the average $\mathbb{E}_{j \sim S_i} X_j$ and t is some number in $\mathbb{N}$ which is a power of 2. Let $\Delta > 0$ be such that $\|\mu_i - \mu_j\| \geq \Delta$ for every $i \neq j$.

Let $w_1, \ldots, w_n, \mu$ be indeterminates. Let $\mathcal{B}$ be the set of equations and inequalities defined above. Thinking of the variables w_i as defining a set $T \subseteq [n]$ via its 0/1 indicator, let $|T \cap S_j|$ be the formal expression $|T \cap S_j| = \sum_{i \in S_j} w_i$. Let $\tilde{\mathbb{E}}$ be a degree $O(t)$ pseudoexpectation which satisfies $\mathcal{B}$. Then

$$\tilde{\mathbb{E}} \left[\sum_{i \in [k]} \left(\frac{|T \cap S_i|}{N} \right)^2 \right] \geq 1 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}.$$

Proof of Lemma 3. Let $\tilde{\mathbb{E}}$ satisfy $\mathcal{B}$. As in Lemmas 1 and 2, we will endeavor to bound $\tilde{\mathbb{E}} |T \cap S_i|^t |T \cap S_j|^t$ for each $i \neq j$.

By Δ-separation,

$$\vdash_{2t} \Delta^{2t} \leq \|\mu_i - \mu_j\|^{2t} = \|(\mu_i - \mu) - (\mu_j - \mu)\|^{2t} \leq 2^{2t} \left(\|\mu_i - \mu\|^{2t} + \|\mu_j - \mu\|^{2t} \right),$$

where we implicitly used the SoS triangle inequality $\vdash_t (a - b)^t \leq 2^t (a^t + b^t)$ on each coordinate of the vectors $\mu_i - \mu$ and $\mu_j - \mu$.

So,

$$\tilde{\mathbb{E}} |T \cap S_i|^t |T \cap S_j|^t \leq \tilde{\mathbb{E}} \left[\frac{\|\mu_i - \mu\|^{2t} + \|\mu_j - \mu\|^{2t}}{2^{2t} \Delta^{2t}} |T \cap S_i|^t |T \cap S_j|^t \right].$$

Now we are going to bound $\tilde{\mathbb{E}} \|\mu_i - \mu\|^{2t} |T \cap S_i|^t |T \cap S_j|^t$. By symmetry the same argument will apply to the same expression with i and j exchanged.

We apply Fact 3:

$$\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \|\mu_i - \mu\|^{2t} \leq 2^{O(t)} t^t \cdot N^2 \cdot \tilde{\mathbb{E}}|T \cap S_i|^{t-2} |T \cap S_j|^t \|\mu_i - \mu\|^t .$$

Then we use pseudoexpectation Cauchy-Schwarz to get

$$\tilde{\mathbb{E}}|T \cap S_i|^{t-2} |T \cap S_j|^t \|\mu_i - \mu\|^t \leq \left(\tilde{\mathbb{E}}|T \cap S_i|^{t-4} |T \cap S_j|^t\right)^{1/2} \left(\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \|\mu_i - \mu\|^{2t}\right)^{1/2} .$$

Putting these two together and canceling $\left(\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \|\mu_i - \mu\|^{2t}\right)^{1/2}$ we get

$$\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \|\mu_i - \mu\|^{2t} \leq 2^{O(t)} t^{2t} \cdot N^4 \tilde{\mathbb{E}}|T \cap S_i|^{t-4} |T \cap S_j|^t .$$

Also, clearly $\mathcal{B} \vdash_{O(1)} |T \cap S_j| \leq N$. So we get

$$\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \|\mu_i - \mu\|^{2t} \leq 2^{O(t)} t^{2t} \cdot N^8 \cdot \tilde{\mathbb{E}}|T \cap S_i|^{t-4} |T \cap S_j|^{t-4} \|\mu_i - \mu\|^t .$$

We started out with the goal of bounding $\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t$. We have found that

$$\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t \leq \frac{2^{O(t)} t^{2t} N^8}{\Delta^{4t}} \tilde{\mathbb{E}}|T \cap S_i|^{t-4} |T \cap S_j|^{t-4} .$$

Applying pseudoexpectation Holder's inequality, we find that

$$\tilde{\mathbb{E}}|T \cap S_i|^{t-4} |T \cap S_j|^{t-4} \leq \left(\tilde{\mathbb{E}}|T \cap S_i|^t \tilde{\mathbb{E}}|T \cap S_j|^t\right)^{(t-4)/t} .$$

Rearranging things, we get

$$\left(\tilde{\mathbb{E}}|T \cap S_i|^t |T \cap S_j|^t\right)^{1/t} \leq \frac{2^{O(t)} t^{t/2} N^2}{\Delta^t} .$$

Now proceeding as in the proof of Lemma 2, we know

$$\tilde{\mathbb{E}} \left((\sum_{i \in [k]} \frac{|T \cap S_i|}{N} \right)^2 = 1$$

so

$$\tilde{\mathbb{E}} \sum_{i \in [k]} \left(\frac{|T \cap S_i|}{N} \right)^2 \geq 1 - \frac{2^{O(t)} t^{t/2} k^2}{\Delta^t} .$$

$\square$

Remark 6.1. Remark: We cheated ever so slightly in this proof. First of all, we did not state a version of pseudoexpectation Holder's which allows the exponent $(t-4)/t$, just one which allows $(t-2)/t$. The correct version can be found in Lemma A.4 of this paper. That inequality will work only when t is large enough; I think $t \geq 4$ suffices. To handle smaller t probably one must remove the square from both sides of Fact 3, which will require a hypothesis which does not use the squared Frobenius norm. This is possible; see e.g. my paper with Jerry Li (https://arxiv.org/abs/1711.07454).

6.4 Putting Things Together

The conclusion of Lemma 3 is almost identical to the conclusion of Lemma 2, and so the rest of the analysis of the high-dimensional clustering algorithm proceeds exactly as in the one-dimensional case. At the end, to show that the clustering algorithm works with high probability to cluster samples from a separated Gaussian mixture model, one uses straightforward concentration of measure to show that if $X_1, \ldots, X_n$ are enough samples from a Δ-separated mixture model, then the satisfy the hypotheses of Lemma 3 with high probability. This concludes our "proof" of the main theorem from way back in part 1.

6.5 Related literature

The reader interested in further applications of the Sum of Squares method to unsupervised learning problems may consult some of the following works.

- [Barak, Kelner, Steurer] on dictionary learning
- [Potechin, Steurer] on tensor completion
- [Ge, Ma] on random overcomplete tensors

Though we have not seen it in these posts, often the SoS method overlaps with questions about tensor decomposition. For some examples in this direction, see the [Barak, Kelner, Steurer] dictionary learning paper above, as well as

- [Ma, Shi, Steurer] on tensor decomposition

The SoS method can often be used to design algorithms which have more practical running times than the large SDPs we have discussed here. (This often requires further ideas, to avoid solving large semidefinite programs.) See e.g.:

- [Hopkins, Schramm, Shi, Steurer] on extracting spectral algorithms (rather than SDP-based algorithms) from SoS proofs
- [Schramm, Steurer] developing a sophisticated spectral method for dictionary learning via SoS proofs
- [Hopkins, Kothari, Potechin, Raghavendra, Schramm, Steurer] with a meta-theorem on when it is possible to extract spectral algorithms from SoS proofs

Another common tool in constructing SoS proofs for unsupervised learning problems which we did not see here are concentration bounds for random matrices whose entries are low-degree polynomials in independent random variables. For some examples along these lines, see

- [Hopkins, Shi, Steurer] on tensor principal component analysis
- [Rao, Raghavendra, Schramm] on random constraint satisfaction problems

as well as several of the previous papers.