Lecture 4: Clustering well-separated Gaussian mixtures

In this lecture, we will use SoS to develop efficient algorithms for clustering mixtures of Gaussain distributions, provided they satisfy certain separation assumptions. We'll also see our first example of an information-computation tradeoff within the sum-of-squares algorithm, wherein we can obtain improved clustering guarantees as we increase the degree of our sum-of-squares relaxation. Some bibliographic remarks will be deferred to the end.

These notes have not been reviewed with the same scrutiny applied to formal publications. There may be errors.

1 Mixtures of Gaussians with separated means

A distribution $\mathcal{D}$ is called a *mixture model* if it can be decomposed as a convex combination of k simpler distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$. That is, we can describe $X \sim \mathcal{D}$ as being sample by first choosing some $i \in [k]$ with probability λ_i, then sampling $X \sim \mathcal{D}_i$. In this lecture, we will be concerned with the problem of *clustering* samples from a mixture model. Formally,

Problem 1.1 (Clustering a mixture model)**.** Let $\mathcal{D}$ be a mixture of k probability distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ over $\mathbb{R}^d$, with mixing weights $\lambda_1, \dots, \lambda_k$. Given independent samples $X_1, \cdots, X_n \sim \mathcal{D}$, our goal is to *cluster* the samples, partitioning $[n]$ into sets $S_1, \dots, S_k$ such that $i \in S_c$ iff $X_i \sim \mathcal{D}_c$.

The requirement that $i \in S_c \iff X_i \sim \mathcal{D}_c$ will often be relaxed, and we'll settle for getting the partition right for most samples with decent probability.

Of course, in some cases the clusters may not be identifiable. We'll restrict our attention to the following special case, in the parameter regime where identifiability holds.

Problem 1.2 (Uniform mixture of Δ-separated isotropic Gaussians)**.** This is the special case of Problem 1.1 in which $\lambda_b = \frac{1}{k}$ and $\mathcal{D}_b = \mathcal{N}(\mu_b, \mathbb{1})$ for all $b \in [k]$, and furthermore $\|\mu_b - \mu_c\| \geqslant \Delta$ for all $i \neq c \in [k]$.

Problem 1.2 is known to be information-theoretically possible with $\mathrm{poly}(d, k)$ samples if $\Delta = \Omega(\sqrt{\log k})$ [RV17]. We would like to design an algorithm for this problem which uses only $n = \mathrm{poly}(k, d)$ samples and runs in time $\mathrm{poly}(k, d)$ as well.

Clustering vs. parameter estimation. Sometimes, and in particular in the isotropic Gaussian setting, the problem of *clustering* is algorithmically equivalent to the problem of *parameter estimation* for the mixture, where the goal is to *estimate μ_c and λ_c for each $c \in [k]$*, given access to the samples.

- If we can cluster the samples, then for each $c \in [k]$, the subset of samples $\{X_i\}_{i \in S_c}$ are independent samples from $\mathcal{D}_c$, and so long as $|S_c|$ is large enough, the empirical average $\overline{X}_c = \frac{1}{|S_c|} \sum_{i \in S_c} X_i$ is a good approximation to μ_c.
- If we can estimate the means, then we can try to assign each sample to the mean that was most likely to have generated it; this can be done on a sample-by-sample basis, but a more powerful (and time-intensive) approach is to create a global assignment of samples to means which takes into account whether the sample statistics are consistent with what one would expect.

The implementation of the second point is intentionally a bit vague, because there are a number of ways that one can do it. In this lecture we will see one way, which uses SoS.

Clustering with the SoS paradigm Perhaps you remember lecture 0, in which we showed how the SoS paradigm can be used to solve the *robust mean estimation* problem. Here, we'll solve the *clustering* problem in a similar way, exploiting this connection between mean estimation and clustering.

Our main result in this lecture will be the following:

Theorem 1.3 ([HL18, KSS18]). *In the setting of Problem 1.2, there is a universal constant C such that for any even integer t, a degree-t SoS algorithm given $n = (d^t k)^{O(1)}$ samples runs in time $n^{O(1)}$ and with high probability returns a partition $W_1, \dots, W_k$ of $[n]$ such that for all $c \in [k]$, there exists a true cluster S_b for which*

$$\frac{|W_c \cap S_b|}{\frac{n}{k}} \geqslant 1 - \frac{k 2^{Ct} t^{t/2}}{\Delta^{t-1}}.$$

In particular, if $\Delta \geqslant k^\gamma$ for $\gamma > 0$ a fixed constant, then a degree-$O(1/\gamma)$ SoS algorithm can estimate the clusters up to error $1/\mathrm{poly}(k)$ given polynomially many samples ($d^{O(1/\gamma)} k^{O(1)}$ samples) and in polynomial time ($d^{O(1/\gamma^2)} k^{O(1/\gamma)}$ time). If $\Delta = \Omega(\sqrt{\log k})$, then a degree-$O(\log k)$ SoS algorithm can estimate the means up to error $1/\mathrm{poly}(k)$ given quasi-polynomially many samples in quasi-polynomial time. This is a *information-computation tradeoff* within SoS; the more computation time we are willing to use, the weaker our separation assumption (which is a kind of signal-to-noise ratio) becomes, and the fewer samples we require. One may ask if there is an *inherent* information-computation gap for this problem; as of very recently, it seems that the answer is no [LL21].

Prior to the algorithms I describe here, the best known polynomial time algorithm required $\Delta = \Omega(k^{1/4})$ [VW02].[1] We'll say more about the history in the bibliographic remarks below.

Remark 1.4. It is noteworthy that their algorithm makes use of higher order ($O(1/\gamma)$) moments of Gaussian (or sub-gaussian) distributions, whereas previous work only used second moments.

We'll once again apply the SoS algorithmic paradigm. We'll establish an SoS proof of identifiability for the clusters $S_1, \dots, S_k$. Then, we solve an SoS relaxation of the appropriate degree, after which we will apply a rounding algorithm to recover the true clusters.

2 Identifiability for the true clusters

Recall the setup. We have samples $X_1, \dots, X_n$ coming from the uniform mixture over $\mathcal{N}(\mu_1, \mathbb{1}), \dots, \mathcal{N}(\mu_k, \mathbb{1})$, with the clusters $S_1, \dots, S_k$ partitioning $[n]$ so that $S_c = \{i \in [n] \mid X_i \sim \mathcal{N}(\mu_c, \mathbb{1})\}$. In this section we will show that the clusters $S_1, \dots, S_k$ are identifiable from samples with high probability. To begin with, we describe some conditions that are satisfied by the true clusters with high probability.

Cluster conditions. Denote by $\bar{\mu}_c$ the empirical mean of samples in the cluster S_c. For some small $\tau > 0$,

(C1) The size of each cluster is close to its expectation:

$$(1-\tau)\frac{n}{k} \leqslant |S_c| \leqslant (1+\tau)\frac{n}{k}, \ \forall j \in [k]. \tag{1}$$

[1] Concurrent with these results is a comparable algorithmic result due to Diakonikolas et al. [DKS18], but it does not use SoS so the theorem statement differs.

(C2) The empirical means are close to population means: $\|\bar{\mu}_c - \mu_c\| \leqslant \tau$.

(C3) The empirical moments are subgaussian. To be specific, for large $t \in \mathbb{N}$, we require

$$\frac{1}{|S_c|}\sum_{i\in S_c}\langle X_i - \bar{\mu}_c, u\rangle^t \leqslant 2t^{t/2}\|u\|^t,\ \forall u \in \mathbb{R}^d,\ \ j \in [k]. \tag{2}$$

Conditions (C1) and (C2) make sense: we expect that these quantities will concentrate around their means. We comment on the moment condition (C3): note if $\mathcal{D}$ is a sub-gaussian distribution over $\mathbb{R}^d$ with mean vector μ and variance-proxy $\sigma = 1$, then by definition of subgaussianity

$$\mathbb{E}_{X\sim\mathcal{D}}\langle X - \mu, u\rangle^t \leqslant t^{t/2}\|u\|^t,\ \forall u \in \mathbb{R}^d.$$

Hence, this condition basically enforces that the uniform distribution over the samples in S_c is subgaussian, which is what we expect, since the samples in S_c are sampled from $\mathcal{N}(\mu_c, \mathbb{1})$.

Using standard concentration of measure arguments, one can actually show that the above (C1)-(C3) are satisfied with high probability, so long as n is large enough.[2] It turns out these conditions suffice for identifying the clusters, in the sense that, if some subset $W \subset [n]$ satisfies (C1) and (C3), then with high probability there exists a true cluster S_c such that $|W \cap S_c|/|S_c|$ is close to 1. This idea will be made rigorous in Lemma 2.2 below. Before that let us encode the above conditions into the polynomial system $\mathcal{A}$, which will be fully exploited in the SoS proof.

System 2.1 (Polynomial constraints on the indicator vector of a cluster)**.** Given samples $X_1, \dots, X_n \in \mathbb{R}^d$ and a small number $\tau > 0$, the following polynomial system $\mathcal{A}$ describes the indicator vector $w \in \{0,1\}^n$ of a cluster W, $w_i = \mathbf{1}_{i\in W}$, as well as the mean of the cluster $\mu \in \mathbb{R}^d$:

1. $w_i^2 = w_i$ for all $i \in [n]$, i.e., w is an indicator vector,
2. $(1-\tau)n/k \leqslant \sum_{i\in[n]} w_i \leqslant (1+\tau)n/k$, enforcing that $|W| \approx n/k$,
3. $\mu \sum_{i\in[n]} w_i = \sum_{i\in[n]} w_i X_i$, meaning that μ is the empirical mean of W,
4. $\sum_{i\in[n]} w_i \langle X_i - \mu, u\rangle^t \leqslant 2t^{t/2}\sum_{i\in[n]} w_i\|u\|^t$, $\forall u \in \mathbb{R}^d$, i.e., the empirical moments are subgaussian.

We now show that 2.1 in conjunction with the conditions (C1)-(C3) ensure that w is an indicator vector for some cluster S_c.

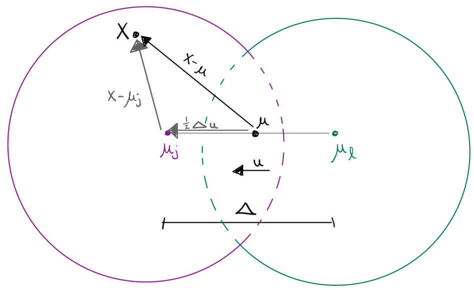
Lemma 2.2 (Lemma 4.20 from [FKP$^+$19])**.** *For $1 \leqslant j \leqslant k$, let a_c be the indicator vector of cluster S_c, and set $A = \sum_{c=1}^k a_c a_c^\top$. Suppose (C1)-(C3) are satisfied by the true clusters. Assume t is a power of 2, and w is a solution of $\mathcal{A}$ with $\tau \leqslant \Delta^{-t}$, then we have*

$$\max_{c\in[k]}\langle w, a_c\rangle \geqslant \frac{n}{k}\left(1 - \frac{2^{O(t)}t^{t/2}k}{\Delta^t}\right). \tag{3}$$

Notice that since w, a_j are a 0/1 vectors with $\frac{n}{k}(1 \pm \tau)$ nonzero entries, this implies that w and a_j agree on most of their entries when $k2^{O(t)} \ll (\Delta/\sqrt{t})^t$.

Proof. At a high level, we will use the fact that if w has significant mass on points in both S_j and S_ℓ for $j \neq \ell$, then the uniform distribution over points in the set W indicated by w cannot be subgaussian in the direction $u = \frac{\mu_j - \mu_\ell}{\|\mu_j - \mu_\ell\|}$. To see why, assume for the sake of illustration that W has half of its points in S_j, and half of its points in S_ℓ, in such a way that its mean is equidistant between μ_j and μ_ℓ: $\mu = \frac{1}{2}(\mu_j + \mu_\ell)$.

[2] Also, notice that (C3), restricted to the case $t = 2$, is the same as the covariance condition we used in the identifiability proof for robust mean estimationin Lecture 0.



For points X in S_j,

$$\langle X-\mu, u\rangle = \langle X-\mu_j, \mu\rangle + \langle \tfrac{1}{2}(\mu_j-\mu_\ell), u\rangle \sim N(0,1) + \tfrac{1}{2}\|\mu_i-\mu_j\|,$$

where we've used that subtracting μ is equivalent to subtracting μ_j and adding $\frac{1}{2}(\mu_j-\mu_\ell)$, and that $X-\mu_j \sim \mathcal{N}(0,\mathbb{1})$ so $\langle X-\mu_j, u\rangle \sim \mathcal{N}(0,1)$ for any unit vector u. So the moments of $\langle X-\mu, u\rangle$ will grow at least like $(\frac{1}{2}\Delta)^t$, which will violate the subgaussianity constraint $\mathcal{A}(4)$.

Now, we will implement this intuition in our proof. We will show that

$$\sum_{j\in[k]} \left\langle w, a_j\right\rangle^2 \geqslant \frac{n^2}{k^2}(1-\varepsilon)\,, \tag{4}$$

For $\varepsilon = 2^{O(t)}t^{t/2}k/\Delta^t$. This is enough to imply our conclusion, because if (4) is true,

$$\max_{j\in[k]} \left\langle w, a_j\right\rangle \geqslant \frac{\sum_{j\in[k]}\left\langle w, a_j\right\rangle^2}{\sum_{j\in[k]}\left\langle w, a_j\right\rangle} \geqslant \frac{1}{\sum_{i=1}^n w_i}\frac{n^2}{k^2}(1-\varepsilon) \geqslant \frac{n}{k}(1-\varepsilon)(1-\tau) \geqslant \frac{n}{k}(1-\varepsilon-\tau), \tag{5}$$

where in the first inequality we use that $\max_{j\in[k]}\langle w, a_\ell\rangle \cdot \sum_{j\in[k]}\langle w, a_\ell\rangle \geqslant \sum_{j\in[k]}\langle w, a_j\rangle^2$, in the second inequality we applied (4), and finally we used the constraint that $\sum w_i = \frac{n}{k}(1\pm\tau)$.

Now, we'll prove (4). First, by applying constraint (2) in $\mathcal{A}$,

$$\left(\sum_{i\in[k]}\langle w, a_i\rangle\right)^2 = \left(\sum_{i=1}^n w_i\right)^2 \geqslant (1-\tau)^2\frac{n^2}{k^2} \geqslant (1-2\tau)\frac{n^2}{k^2}.$$

The left-hand side include the left-hand side of (4) as well as cross-terms $\langle w, a_j\rangle\langle w, a_i\rangle$, so it will suffice to show that these cross-terms do not contribute more than $O(\varepsilon)$ to the total,

$$\sum_{j\neq\ell}\langle w, a_j\rangle\langle w, a_\ell\rangle \leqslant \frac{n^2}{k^2}O(\varepsilon). \tag{6}$$

This is where our intuition about the subgaussianity in the direction $u = (\mu_j-\mu_\ell)/\|\mu_j-\mu_\ell\|$ comes in. Using that $\|\mu_j-\mu_\ell\|/\Delta \geqslant 1$,

$$\begin{aligned}\langle w, a_j\rangle\langle w, a_\ell\rangle &\leqslant \frac{\|\mu_j-\mu_\ell\|^t}{\Delta^t}\langle w, a_j\rangle\langle w, a_\ell\rangle \\ &= \frac{1}{\Delta^t}\cdot\langle w, a_j\rangle\langle w, a_\ell\rangle\langle \mu_j-\mu_\ell, u\rangle^t \\ &= \frac{1}{\Delta^t}\cdot\langle w, a_j\rangle\langle w, a_\ell\rangle\left(\left\langle \mu_j-\mu, u\right\rangle + \langle \mu-\mu_\ell, u\rangle\right)^t\end{aligned} \tag{7}$$

$$\overset{(i)}{\leqslant} \frac{2^{t-1}}{\Delta^t}\langle w, a_j\rangle \langle w, a_\ell\rangle \left(\left\langle \mu_j - \mu, u \right\rangle^t + \langle \mu - \mu_\ell, u\rangle^t \right)$$
$$= \frac{2^{t-1}}{\Delta^t}\langle w, a_\ell\rangle \sum_{i\in S_j} w_i \cdot \langle \mu_j - \mu, u\rangle^t + \frac{2^{t-1}}{\Delta^t}\langle w, a_j\rangle \sum_{i \in S_\ell} w_i \cdot \langle \mu_\ell - \mu, u\rangle^t, \tag{8}$$

where in (i) we used the triangle inequality: $(a+b)^t \leqslant 2^{t-1}(a^t + b^t)$. Now, we bound just one of the terms above (as they are symmetric). We will introduce the samples, twice use subgaussianity:

$$\sum_{i\in S_j} w_i \cdot \langle \mu_j - \mu, u\rangle^t = \sum_{i \in S_j} w_i \cdot \left(\langle \mu_j - X_i, u\rangle + \langle X_i - \mu, u\rangle \right)^t$$
$$\leqslant 2^{t-1} \sum_{i\in S_j} \langle \mu_j - X_i, u\rangle^t + 2^{t-1} \sum_{i \in S_j} w_i \cdot \langle X_i - \mu, u\rangle^t. \tag{9}$$

Since by (C2) we have $\|\bar{\mu}_j - \mu_j\| \leqslant \tau$ and by (C3) we have empirical subgaussianity within S_j, using the triangle inequality, the first sum we may bound by $|S_j| \cdot (2^t t^{t/2} + (2\tau)^t) \leqslant (1+\tau)\frac{n}{k}(2^t t^{t/2} + (2\tau)^t)$. For the second sum, we use the polynomial subgaussianity constraint $\mathcal{A}(4)$ to conclude that

$$\sum_{i\in S_j} w_i \cdot \langle X_i - \mu, u\rangle^t \leqslant \sum_{r\in[k]} \sum_{i \in S_r} w_i \cdot \langle X_i - \mu, u\rangle^t \leqslant 2t^{t/2} \sum_{i \in [n]} w_i \leqslant 2t^{t/2}\frac{n}{k}(1+\tau).$$

Together, this gives us that the right-hand side of (9) is at most $(1+\tau)(2^{t+1}t^{t/2} + (2\tau)^t)\frac{n}{k}$, and in turn the right-hand side of (8) is at most $2^{2t+2}\Delta^{-t}(\langle w, a_j\rangle + \langle w, a_\ell\rangle)\frac{n}{k}t^{t/2}$ (we have used that τ is small).

So putting it all together,

$$\sum_{j\neq \ell} \langle w, a_j\rangle\langle w, a_\ell\rangle \leqslant \frac{2^{2t+2}}{\Delta^t}\frac{n}{k}t^{t/2} \sum_{j \neq \ell} \langle w, a_\ell\rangle + \langle w, a_j\rangle \leqslant k \cdot \frac{2^{2t+2}}{\Delta^t}\frac{n}{k}t^{t/2} \cdot 2 \sum_{i\in[n]} w_i \leqslant \frac{2^{2t+4}}{\Delta^t}\frac{n^2}{k}t^{t/2} = O(\varepsilon)\frac{n^2}{k^2},$$

as desired. □

3 SoS-izing the proof of identifiability

Lemma 2.2 immediately suggests a procedure for recovering the ground-truth partition $S_1, \ldots, S_k$: First find a solution (w, S) to $\mathcal{A}$. According to Lemma 2.2, W should be very close to some S_j. Then we remove the points in W and repeat the above procedure for the remaining points, until we obtain k clusters. However, this does not result an efficient algorithm. Following the usual sum-of-squares paradigm, we will instead solve for a pseudoexpectation $\tilde{\mathbb{E}}$ of sufficiently high degree that satisfies $\mathcal{A}$, and round this pseudoexpectation to find a good clustering.

Encoding the subgaussian condition. Note that we cannot directly make use of $\mathcal{A}$ to find a degree-$O(t)$ pseudoexpectation operator in polynomial time, since there are infinitely many inequality constraints in $\mathcal{A}(4)$ (one for each $u \in \mathbb{R}^d$). Although in the proof above we only used the t-th empirical moments are bounded in the $\binom{k}{2}$ directions of $\mu_j - \mu_\ell$, $j \neq \ell \in [k]$, this is still problematic because the μ_j's are unknown parameters that we are trying to estimate, so we don't have access to them when we are trying to encode a polynomial system as part of our algorithm. To deal with this issue, we introduce the notion of a "t-explicitly bounded distribution" (also known as *t-certifiably subgaussian* in the literature) below:

Definition 3.1 (t-explicitly bounded)**.** Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ. For $\sigma > 0$ and $t \in \mathbb{N}$, we say that $\mathcal{D}$ is *t-explicitly bounded with variance proxy σ* if for every even number $s \leqslant t$, there is a degree-s SoS proof of the inequality:

$$\vdash_s \mathbb{E}_{X\sim\mathcal{D}} \langle X-\mu, u\rangle^s \leqslant (\sigma s)^{s/2} \|u\|^s. \tag{10}$$

Equivalently, the polynomial $(\sigma s)^{s/2}\|u\|^s - \mathbb{E}_{X\sim\mathcal{D}}\langle X-\mu,u\rangle^s$ can be written as a sum of squares. In this lecture we will assume $\sigma = 1$ and just call the distribution t-explicitly bounded, we also assume that t is even to avoid some technical difficulties.

Remark 3.2 (Examples of t-explicitly bounded distributions)**.** Any normal distribution with identity covariance matrix is t-explicitly bounded for any $t \in \mathbb{N}$. The rotation of product distributions with bounded t-th moments are also t-explicitly bounded.

Moreover, [KSS18] proved that σ-Poincaré distributions are t-explicitly bounded. We say a distribution $\mathcal{D}$ is σ-Poincaré if it satisfies the following Poincaré inequality: For all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbf{Var}_{X\sim\mathcal{D}}[f(X)] \leqslant \sigma^2 \mathbb{E}_{X\sim\mathcal{D}}\left[\|\nabla f(X)\|^2\right]. \tag{11}$$

Together, these examples comprise many commonly considered distributions.

We'll briefly describe the proof that for any $\zeta \in \mathbb{R}^d$, the distribution $\mathcal{N}(\zeta, \mathbb{1})$ is t-explicitly bounded. This boils down to the fact that the following matrix inequality holds for any $s \leqslant t/2$:

$$\mathbf{E}_{X\sim\mathcal{N}(\zeta,\mathbb{1})}(X-\zeta)^{\otimes s}((X-\zeta)^{\otimes s})^\top \preceq \mathbf{E}_{X\sim\mathcal{N}(0,\mathbb{1})} X^{\otimes s}(X^{\otimes s})^\top. \tag{12}$$

The fact that this implies t-subgaussian behavior is given by taking the quadratic form of the left- and right-hand side with $u^{\otimes s}$ for any unit vector $u \in \mathbb{R}^d$.

So in order to encode the constraint that $\sum_i w_i(X_i - \mu)$ is t-subgaussian, we will replace $\mathcal{A}(4)$ with the polynomial constraint

$$\sum_{i\in[n]} w_i((X_i-\mu)^{\otimes t/2})((X_i-\mu)^{\otimes t/2}) = 2\cdot\left(\sum_{i\in[n]} w_i\right)\cdot \mathbf{E}_{X\sim\mathcal{N}(0,\mathbb{1})} X^{\otimes t/2}(X^{\otimes t/2})^\top - BB^\top,$$

for B a matrix of indeterminate variables of dimension $d^{t/2}\times d^{t/2}$. This constraint encodes the t-subgaussianity of the w-cluster W as a set of $d^{O(t)}$ polynomial equalities, and the fact that it is feasible follows from the fact that each $\mathcal{D}_j$ is Gaussian, plus an argument that (12) is satisfied (up to a factor of 2) by the emprical samples X_i for $i \in S_j$ with high probability so long as $n = d^{\Omega(t)}$; this is ensured by the condition $n = d^{\Omega(t)}$ in Theorem 1.3.[3] Call this new system of equations $\widehat{\mathcal{A}}$. Since $\widehat{\mathcal{A}}$ has only $d^{O(t)} + \mathrm{poly}(n)$ constraints, finding a pseudoexpectation $\tilde{\mathbb{E}}$ which satisfies $\widehat{\mathcal{A}}$ takes $d^{O(t)} + \mathrm{poly}(n)$ time.

After introducing the new polynomial system $\widehat{\mathcal{A}}$, one can prove a SoS version of Lemma 2.2.

Lemma 3.3 (Lemma 5.3 from [HL18])**.** *Under the same assumptions as Lemma 2.2, let $\tilde{\mathbb{E}}$ be a degree-$O(t)$ pseudoexpectation that satisfies $\widehat{\mathcal{A}}$, then*

$$\tilde{\mathbb{E}} \sum_{j\in[k]} \left\langle w, a_j\right\rangle^2 \geqslant \frac{n^2}{k^2}\left(1 - \frac{2^{O(t)} t^{t/2} k}{\Delta^t}\right). \tag{13}$$

Sketch of proof. Notice that each inequality that appeared in the proof of (4) in Lemma 2.2 can be SoS-ized by applying the usual SoS tools, including SoS versions of Cauchy-Schwarz, Hölder's inequality, and the triangle inequality. □

[3]See Lemma 4.1 in [HL18].

4 Rounding the pseudomoments

Finally, we will show that we can use our pseudoexpectation satisfying $\widehat{\mathcal{A}}$ to recover the cluster centers. Here, we will make one final modification to our algorithm: we will search for the pseudoexpectation satisfying $\widehat{\mathcal{A}}$ which minimizes the Frobenius norm $\|\widetilde{\mathbb{E}} ww^\top\|_F$. This is a convex objective, so we can solve for $\widetilde{\mathbb{E}}$ in polynomial time.

Lemma 4.1. *For the degree-$O(t)$ pseudoexpectation operator $\widetilde{\mathbb{E}}$ satisfying $\widehat{\mathcal{A}}$ which minimizes $\|\widetilde{\mathbb{E}} ww^\top\|_F$, the matrix $M = \widetilde{\mathbb{E}} ww^\top$ is close to the block matrix $A = \frac{1}{k}\sum_{j\in[k]} a_i a_i^\top$, in the sense that $\|A - M\|_F^2 \leqslant \varepsilon \|A\|_F^2$ for $\varepsilon = \frac{2^{O(t)} k t^{t/2}}{\Delta^t}$.*

Once we prove this claim, the algorithm is easy: A is a block matrix whose k blocks correspond exactly to the k clusters, and A and M agree on all but an ε-fraction of entries. So $M = \widetilde{\mathbb{E}} ww^\top$ is essentially a block matrix whose blocks correspond to the clusters, and we can effectively read these off. Finally, to estimate the mean μ_j, one can take the empirical mean of the samples in S_j.

Proof of Lemma 4.1. Note that $M \succeq 0$, and $\mathrm{Tr} M = \widetilde{\mathbb{E}} \sum_{i\in[n]} w_i^2 = \frac{n}{k}(1 \pm \tau)$ by $\mathcal{A}(1)$ and (2).

Now, we have that

$$\begin{aligned} \|M - A\|_F^2 &= \|M\|_F^2 + \|A\|_F^2 - 2\langle M, A\rangle \\ &\leqslant 2(\|A\|_F^2 - \langle M, A\rangle), \end{aligned}$$

where the inequality follows because $\widetilde{\mathbb{E}}$ was chosen to minimize the Frobenius norm of M, and A corresponds to the (with high probability) feasible choice of $\widetilde{\mathbb{E}}$ as the actual expectation of the distribution where w is chosen from the uniform mixture over $\{a_j\}_{j\in[k]}$. But now, notice that

$$\langle M, A\rangle = \frac{1}{k}\widetilde{\mathbb{E}}\left[\sum_{j\in[k]} \langle w, a_j\rangle^2\right] \geqslant \frac{n^2}{k^3}(1-\varepsilon),$$

for $\varepsilon = 2^{O(t)} t^{t/2} k/\Delta^t$, where to obtain the inequality we have applied Lemma 3.3, the SoS-version of Lemma 2.2. The lemma now follows because $\|A\|_F^2 = (1 \pm \tau)\frac{n^2}{k^3}$. □

5 Conclusion

Putting it all together, we have now seen the proof of Theorem 1.3. We note that the algorithms presented here can be generalized to the case of non-uniform mixing weights, and to the case when the mixture is over any $\mathcal{D}_1, \ldots, \mathcal{D}_k$ which are t-explicitly bounded. By results of [KSS18], the property of being t-explicitly bounded holds for the large family of distributions which satisfy a Poincaré inequality.

Bibliographic remarks. The algorithm given here is based on the concurrent works of Hopkins-Li and Kothari-Steinhardt [HL18, KSS18]; here we are borrowing from the presentations of [HL18, FKP⁺19].

The study of Problem 1.1 can be traced back to Pearson [Pea94]. Prior to these works, the best algorithm for learning Gaussian mixture model with isotropic components required $\Delta \geqslant k^{1/4}$, via *single-linkage clustering*, which is a simple greedy algorithm. In this parameter regime, every pair of samples from the same cluster are closer to each other in Euclidean distance than are every pair of samples from distinct clusters (with high probability), so the clusters can be identified using this information about sample second moments. The single-linkage clustering algorithm of Vempala and Wang [VW02] was built upon several pioneering works [Das99, DS07, AK+05].

Standard information-theoretic arguments [RV17] show that it's possible to identify the cluster means from $n = \text{poly}(k, d)$ samples when Δ is $\Omega(\sqrt{\log k})$, but prior to 2018 only exponential-time algorithms were known. As is evident from this timeline, this computational-to-statistical gap stood open for a long time until the breakthrough works of Hopkins and Li [HL18], Kothari, Steinhardt and Steurer [KSS18], and Diakonikolas, Kane, and Stewart [DKS18]. The two former results use SoS algorithms, while the latter uses a spectral filtering approach and is significantly different from the approach here.

The SoS and pseudoexpectation inequalities needed in the proof of Lemma 3.3 may be found in Section 7 of [HL18] and Appendix A of [BKS14].

These lecture notes are partially based on scribe notes written by Kangjie Zhou.

Contact. Comments are welcome at tselil@stanford.edu.

References

[AK+05] Sanjeev Arora, Ravi Kannan, et al. Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):69–92, 2005. 8

[BKS14] Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40, 2014. 8

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999. 8

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018. 2, 8

[DS07] Sanjoy Dasgupta and Leonard J Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007. 8

[FKP+19] Noah Fleming, Pravesh Kothari, Toniann Pitassi, et al. *Semialgebraic proofs and efficient algorithm design*. now the essence of knowledge, 2019. 3, 7

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018. 2, 6, 7, 8

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018. 2, 6, 7, 8

[LL21] Jerry Li and Allen Liu. Clustering mixtures with almost optimal separation in polynomial time. *arXiv preprint arXiv:2112.00706*, 2021. 2

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. 7

[RV17] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017. 1, 8

[VW02] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 113–122. IEEE, 2002. 2, 8