Learning to Cluster Using SODS

Simple (but still interesting) version of problem:

\[ K = \text{number of clusters} \]
\[ d = \text{dimension} \]
\[ n = \text{\# of samples} \]

We have fixed spherical (but unknown) Gaussians \( D_1, \ldots, D_k \), with means \( \mu_1, \ldots, \mu_k \), and variance 1.

Input is: \( n \) samples \( X_1, \ldots, X_n \) drawn from mixture.

Output: means \( \mu_1, \ldots, \mu_k \) or clusters (say which distribution each sample is from).
Hard Info theoretically if means are too close together:

These 2 different mixtures look too similar.

$K$ gaussians have var distance $\sim 2^{-K}$ from single gaussian. Need $\exp(K)$ samples.

Regime of parameters we want: $K \sim d$

$n$ (the samples) $\sim \text{poly}(d)$

runtime $\text{poly}(d)$

$\Delta$ as small as possible

Separation Assumption: $\forall i \neq j \ |\mu_i - \mu_j| \geq \Delta$

Info LB: For $\Delta \leq \sqrt[100]{d}$ clustering impossible using $\text{poly}(d)$ samples.
n samples \( n = \text{poly}(d) \)
d dimensions, \( k \) gaussians \( k \sim d \)

1. radius \( q \) clusters \( \sim \sqrt{d} \)
   \[ \Delta > 4 \sqrt{d} \] greedy clustering easy

2. Dasgupta (spectral)
   polytime alg for \( \Delta = 3 \sqrt{k} \)

3. Dasgupta - Schulman (EM)
   \[ \Delta \sim K^{1/4} \] polytime

4. Regev - Vijayraghavan (MLE)
   \[ \Delta = O(\log d) \] poly \((d)\) samples
   but runtime \( \exp(d) \)

\[ \Delta = \| M_i - M_j \| \quad (L_2 \text{ distance in } \mathbb{R}^d) \]
New 3 papers STOC '18

[Hopkins-Li], [Kothari-Steinhardt],
[Diakonikolas-Kane-Stewart]

Theorem \exists constant C
For \( \Delta = c \sqrt{\log d} \), there is a
quasipoly(d)-time alg, error \( \frac{1}{\text{poly}(d)} \)

For \( \Delta = d^{\varepsilon} \), poly(d)-time alg, error \( \frac{1}{\text{poly}(d)} \)

* Also works for robust versions
of the problem, and for
more general distributions
Actual Theorem Statement: Fix $d = k$, \( t \in \mathbb{N} \) s.t. for \( n = d^{o(t)} \cdot k^{O(1)} \), there is an algorithm running in time \( \text{polynomial}(n) \) that takes as input random samples \( X_1, \ldots, X_n \in \mathbb{R}^d \) where \( S_1, \ldots, S_k \) is the true partition of \( [n] \) s.t. \( \{ X_i \mid i \in S_i \} \) is from \( \mathcal{D}_i \), and \( |S_i| = \frac{n}{k} = N \)

and outputs a partition \( T_1, \ldots, T_k \) of \( [n] \) s.t. \( |T_i| = N \), and whp \( \forall i \)

\[
\frac{|S_i \cap T_i|}{N} \geq 1 - k^{10} \cdot \left( \frac{2 \sqrt{t}}{\Delta} \right)^t
\]

For \( \Delta = O(\sqrt{\log d}) \), choose \( t \sim O(\log k) \)

\( \Delta = k^c \), choose \( t \sim 1000 \)
Overview \((d=1)\)

- In order to get an algorithm running in quasipoly time,
  need to show quasipoly many samples suffice.
  That is it is necessary to prove quasipoly sample complexity bounds.

- We'll give low degree \(\text{SOS}\) sample complexity bounds, which will automatically [by \(\text{SOS automata}\) realizability] give us an efficient learning algorithm.

⚠️ This isn't exactly how the algorithm goes...
but correct at a high level.
Important Property of the Samples \( \{d=1\}, \) (generally easy)

Let \( \mathcal{D} = N(\mu, 1) \) be gaussian

and \( Y_1, \ldots, Y_n \) be samples from \( \mathcal{D} \).

Then for \( N = N(t) \) (large enough, whp

\[
\mathbb{E} \left| Y_j - \mu \right|^t \leq 1.1 \, t^{\frac{t}{2}}
\]

\( j \sim [N] \uparrow \)

\text{sample mean} \quad \left[ \text{concentration of} \quad \begin{aligned} \text{empirical moments} \end{aligned} \right]

\* Our algorithm will assume

that the samples \( X_1, \ldots, X_n \) come from

a true partition \( S_1, \ldots, S_k \), \( S_i \subset [n] \), \( |S_i| = \frac{n}{k} = N \)

s.t. \( \forall i \in [k], \quad \mathbb{E} \left| X_j - \mu_i \right|^t \leq 2 \, t^{\frac{t}{2}} \)

\( j \sim S_i \quad \text{empirical avg of} \quad \{X_j \mid j \in S_i\} \)

An easy calculation shows this is true whp.
Let $A$ be the following equations:

\[ w_i^2 = w_i \quad i \in [n] \]

\[ \sum_{i \in [n]} w_i = N \quad (N = \frac{A}{\kappa}) \]

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

where \( \mu = \frac{1}{N} \sum_{i \in [n]} w_i x_i \)

[Variables are vectors $w_1 \ldots w_n$]

Solve SDP of degree $O(t)$ SOS lift of $A$ to minimize $||WW^t||^2$
Claim: If $\mathbf{w} \mathbf{w}^T$ is a solution to degree 0(s) SOS lift, then $\|\mathbf{v} \mathbf{v}^T - a a^T\|^2 \leq 2 \left( \frac{n^2}{K^2} - \langle \mathbf{v} \mathbf{v}^T, a a^T \rangle \right)$

Proof: Assume w.l.o.g. $X_1 \ldots X_N X_{N+1} \ldots X_{2N} \ldots X_{(K-1)N+1} \ldots X_{KN}$

(so first $N = \frac{n}{K}$ samples are drawn from $\mathcal{D}$
next $N$ samples " " " $\mathcal{D}_z$ and so on.)

Then these values for $\mathbf{w}$ satisfy $\mathcal{A}$:

$$
\begin{bmatrix}
1 \\
1 \\
\vdots \\
1
\end{bmatrix}
$$

The corresponding solutions to degree $z$ variables $\mathbf{w} \mathbf{w}^T (\mathbf{w}, \mathbf{w})$ are:

$$
\begin{bmatrix}
1 \\
1 \\
\vdots \\
1
\end{bmatrix}
$$

So the avg $a a^T$ also satisfies $\mathcal{A}$:

$$
\begin{bmatrix}
\frac{1}{K} \\
\frac{1}{K} \\
\vdots \\
\frac{1}{K}
\end{bmatrix}
$$
\[ \| a a^T \|^2 (\text{Frobenius norm} = \| \cdot \|_F) \]

\[ = \frac{1}{K^2} \cdot K N^2 = \frac{1}{K} \cdot \frac{N^2}{K^2} = \frac{N^2}{K^3} \]

Say degree \( O(t) \) SDP finds a solution \( vv^T \) satisfying degree \( O(t) \) sos of \( A \), and minimizing \( \| vv^T \| \). Then \( \| vv^T \| \leq \| a a^T \| \leq \frac{N^2}{K^3} \)

Then

\[ \| vv^T - a a^T \|^2 = \| vv^T \|^2 + \| a a^T \|^2 - 2 \langle vv^T, a a^T \rangle \]

\[ \leq 2 \frac{N^2}{K^3} - 2 \langle vv^T, a a^T \rangle \]

\[ = 2 \left( \frac{N^2}{K^3} - \langle vv^T, a a^T \rangle \right) \]

end of claim
**MAIN LEMMA** Let \( X_1, \ldots, X_n \in \mathbb{R}, S_1, \ldots, S_k \) a partition of \([n]\), \( |S_i| = \frac{n}{k} \), s.t. \( \forall i \in [k] \):

\[
\mathbb{E} \left( |X_j - \mu_i|^t \right) = 2 \cdot t^{\frac{k}{2}}
\]

\( j \sim S_i \)

Let \( W \) be a solution to degree \( o(t) \) SDP for \( A \).

(\text{so } v \text{ is a degree } o(t) \text{ pseudodistribution})

Then \( W \) satisfies

\[
\sum_{i \in [k]} \left( \frac{|T \cap S_i|}{N} \right)^2 \geq 1 - 2 \frac{t}{\Delta^t} \frac{K^2}{N^2}
\]

where \( |T \cap S_i| = \sum_{j \in S_i} w_j \)

\( \left[ \text{Proof after} \right] \)

\[
\langle WW^T, aa^T \rangle = \sum_{i \in [k]} \left( \sum_{j \in S_i} w_j \right)^2 \frac{1}{(T \cap S_i)^2}
\]

\[
= \sum_{i \in [k]} \left( \sum_{j \in S_i} w_j \right)^2 \frac{1}{(T \cap S_i)^2} \geq k \cdot N^2 \left( 1 - \frac{2^{o(t)}}{\Delta^t} \frac{t^{\frac{k}{2}}}{K^2} \right)
\]
Corollary (of claim + MAIN Lemma)

$$\|vv^T - aa^T\| \leq \|aa^T\| \cdot \left(\frac{2^{o(t)}}{\Delta^t} \frac{t^2K^2}{\varepsilon}\right)^{\frac{1}{2}}$$

By claim

$$\|vv^T - aa^T\|^2 \leq 2\left(\frac{n^2}{K^3} - \langle vv^T, aa^T \rangle\right)$$

By MAIN Lemma

$$\langle vv^T, aa^T \rangle \geq \frac{1}{K} \cdot \frac{n^2}{K} \cdot \left(1 - \frac{2^{o(t)} t^2 K^2}{\Delta^t}\right)$$

So

$$\|vv^T - aa^T\| = \left[2\left(\frac{n^2}{K^3} - \frac{n^2}{K} - \frac{n^2}{K} \cdot \frac{2^{o(t)} t^2 K^2}{\Delta^t}\right)\right]^{\frac{1}{2}}$$

$$= \frac{n^2}{K^3} = \|aa^T\|^2$$

$$= \|aa^T\| \cdot \left(\frac{2^{o(t)} t^2 K^2}{\Delta^t}\right)^{\frac{1}{2}}$$
Recall $X_1, \ldots, X_n$, $S_1, \ldots, S_k$

we solve degree $O(t)$ SOS SDP to get pseudo-distribution $ww^T$ s.t.

$$\|ww^T - aa^T\| \leq \|aa^T\| \cdot \text{(small fraction)}$$

so $ww^T$ is very close to the "good" solution $aa^T$

Note from $aa^T$ we know $S_1, \ldots, S_k$

and $ww^T$ is very close entry-wise to $aa^T$

Rounding Alg

1. Let $I = [n]$ be active indices
2. Pick $i \sim I$ uniformly $\&$ let $S \subseteq I$ be the indices $j$ s.t. $\|M_i - M_j\| < \frac{\varepsilon}{\sqrt{k}}$
   (so the rows that are almost the same as row $i$

   add $S$ to list of clusters $\&$ let $J = I \setminus S$

3. If $|J| \geq \frac{n}{2k}$ go to (2)
4. Assign remaining indices to clusters $\tilde{J}$ till all have size $\frac{n}{k}$
A row $i$ is good if $\| M_i - A_i \| < \frac{\varepsilon}{100} \sqrt\frac{n}{k} = \frac{1}{100} \| A_i \|$

There are $\ll \varepsilon^2 n$ bad rows (by averaging).

If rounding alg never picks a bad row, then alg will cluster all good rows correctly.

Prob. round alg never picks a bad row $\leq k^2 \varepsilon^2$

since $\sum_{i \in n} \| M_i - A_i \|^2$

$= \| M - \bar{A} \|^2 \leq \varepsilon^2 \| A \|^2$