Learning to Cluster Using SOS

Simple (but still interesting) version of problem:

K = number of clusters

d = dimension

n = # of samples

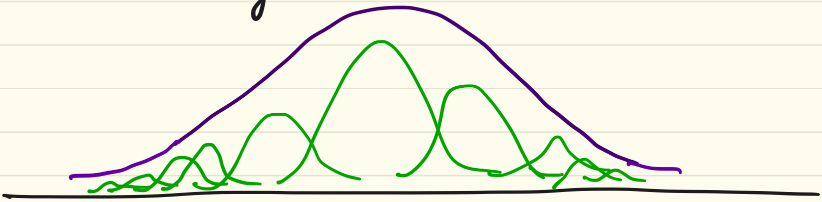
We have fixed spherical (but unknown) gaussians $\mathcal{D}_1, \ldots, \mathcal{D}_K$; with means $\mu_1, \ldots, \mu_K$, and variance 1.

Input is: n samples $X_1 \ldots X_n$ drawn from mixture

Output: means $\mu_1 \ldots \mu_K$

or clusters (say which distrib 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 (# samples) $\sim \text{poly}(d)$

runtime $\text{poly}(d)$

Δ as small as possible

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

$\boxed{\text{Info LB}}$: For $\Delta \leq \sqrt{\log d}$ clustering impossible using $\text{poly}(d)$ samples n samples

History

d dimensions, k gaussians

$n \sim poly(d)$
$k \sim d$

1. radius of clusters $\sim \sqrt{d}$

   $\Delta > 4\sqrt{d}$ greedy clustering
   easy

2. Dasgupta (spectral)

   poly time alg for $\Delta = \varepsilon\sqrt{k}$

3. Dasgupta-Schulman (EM)

   $\Delta \sim k^{1/4}$ poly time

4. Regev-Vijayaraghavan (MLE)

   $\Delta = O(\sqrt{\log d})$ poly(d) samples

   but runtime exp(d)

$\Delta = \| \mu_i - \mu_j \|$ (L_2 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 $\frac{1}{\mathrm{poly}(d)}$

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

* Also works for robust versions of the problem, and for more general distributions

Actual Theorem Statement

Fix $d \approx k$, $\exists t$ s.t. for $n = d^{O(t)} \cdot k^{O(1)}$, there is an algorithm running in time poly(n) that takes as input random samples $X_1, .., X_n \in \mathbb{R}^d$

$\Big[$ $S_1, .., S_k$ is the true partition of $[n]$ s.t. $\{X_i \mid i \in S_i\}$ is from $\mathcal{D}_i$, and $|S_i| = \frac{n}{k} = N$ $\Big]$

and outputs a partition $T_1, .., T_k$ 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 $t \sim O(\log k)$

$\Delta = k^{\varepsilon}$, 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 <u>necessary</u> to prove quasipoly sample complexity bounds

- We'll give low degree SOS sample complexity bounds, which will automatically [by SOS automatizability] 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

($d=1$, generalizes easily)

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

$$\mathop{\mathbb{E}}_{j \sim [N]} |Y_j - \bar{\mu}|^t \leq 1.1\, t^{t/2}$$

sample mean

[concentration of t^{th} empirical moments]

* Our algorithm will assume that the samples $X_1 \ldots X_n$ come from a true Partition $S_1 \ldots S_k$, $S_i \subseteq [n]$, $|S_i| = \frac{n}{k} = N$

s.t. $\forall i \in [k]$, $\mathop{\mathbb{E}}_{j \sim S_i} |X_j - \mu_i|^t \leq 2 \cdot t^{t/2}$

empirical avg of $\{X_j \mid j \in S_i\}$

An easy calculation shows this is true whp.

Algorithm in the $d=1$ case [on input $X_1 \dots X_n$]

Let $\mathcal{A}$ be the following equations:

$$w_i^2 = w_i \qquad i \in [n]$$

$$\sum_{i \in [n]} w_i = N \qquad \left(N = \frac{n}{k}\right)$$

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

where $\mu = \frac{1}{N} \sum_{i \in [n]} w_i X_i$

[variables are vectors $w_1 \dots w_n$]

Solve SDP of degree $O(t)$ SOS lift of $\mathcal{A}$ to minimize $\| \underbrace{w w^T}_{n \times n \text{ matrix}} \|^2$

Frobenius norm

L_2^2 as a vector

Claim If vv^T is solution to degree $O(t)$ SOS lift, then $\|vv^T - aa^T\|^2 \le 2\left(\frac{n^2}{k^3} - \langle vv^T, aa^T\rangle\right)$

Proof Assume WLOG $\underbrace{X_1 \dots X_N}_{S_1}\ \underbrace{X_{N+1} \dots X_{2N}}_{S_2} \dots\dots \underbrace{X_{(k-1)N+1} \dots X_{kN}}_{S_k}$

(so first $N = \frac{n}{k}$ samples are drawn from $\mathcal{D}_1$,
next N samples " " " $\mathcal{D}_2$, and so on.)

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

$$a_1 = \begin{bmatrix} 1 \\ \\ \\ \\ \end{bmatrix} \quad a_2 = \begin{bmatrix} \\ 1 \\ \\ \\ \end{bmatrix} \quad \dots \quad a_k = \begin{bmatrix} \\ \\ \\ \left.1\right\}{\scriptstyle N} \end{bmatrix}$$

the Corresponding solutions to degree 2 variables ww^T $(w_i w_j)$ are:

$$a_1 a_1^T = \begin{bmatrix} \boxed{1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} \quad a_2 a_2^T = \begin{bmatrix} \ddots & & & \\ & \boxed{1} & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} \quad \dots \quad a_k a_k^T = \begin{bmatrix} \ddots & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \boxed{1} \end{bmatrix}$$

So the avg aa^T also satisfies A:

$$\begin{bmatrix} \boxed{1/k} & & & & \\ & \boxed{1/k} & & & \\ & & \boxed{1/k} & & \\ & & & \ddots & \\ & & & & \boxed{1/k} \end{bmatrix}$$

$\therefore \quad \|aa^T\|^2$ (Frobenius norm = L_2 norm as a vector)

$$= \frac{1}{K^2} \cdot K N^2 = \frac{1}{K} \frac{n^2}{K^2} = \frac{n^2}{K^3}$$

Say degree $O(t)$ SDP finds a solution vv^T satisfying degree $O(t)$ SOS of A, and minimizing $\|ww^T\|$. Then $\|vv^T\| \le \|aa^T\| \le \frac{n^2}{K^3}$

Then

$$\|vv^T - aa^T\|^2 = \|vv^T\|^2 + \|aa^T\|^2 - 2\langle vv^T, aa^T\rangle$$

$$\le 2\frac{n^2}{K^3} - 2\langle vv^T, aa^T\rangle$$

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

$\blacksquare$ end of claim

MAIN LEMMA Let $X_1 \dots X_n \in \mathbb{R}$, $S_1 \dots S_K$ a partition of $[n]$, $|S_i| = \frac{n}{K}$, s.t. $\forall i \in [K]$:

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

Let w be a solution to degree $O(t)$ SDP for A.
(so v is a degree $O(t)$ pseudodistrib)

Then w satisfies

$$\sum_{i \in [K]} \left(\frac{|T \cap S_i|}{N} \right)^2 \ge 1 - \frac{2^{O(t)}\, t^{t/2}\, K^2}{\Delta^t}$$

where $|T \cap S_i| = \sum_{j \in S_i} w_j$ [Proof after]

$\langle ww^T, aa^T \rangle =$

$$\begin{bmatrix} \quad & & \\ & & \\ & & \end{bmatrix} \begin{bmatrix} \boxed{1/K} & & & \\ & \boxed{1/K} & & \\ & & \ddots & \\ & & & \boxed{1/K} \end{bmatrix}$$

$$= \sum_{i \in [K]} \Big(\underbrace{\sum_{j \in S_i} w_j}_{(T \cap S_i)^2} \Big)^2 \ge K \cdot N^2 \left(1 - \frac{2^{O(t)}\, t^{t/2}\, K^2}{\Delta^t} \right)$$

Corollary (of claim + MAIN Lemma)

$$\| vv^T - aa^T \| \le \|aa^T\| \cdot \underbrace{\left(\frac{2^{O(t)} t^{t/2} K^2}{\Delta^t} \right)^{1/2}}_{\varepsilon}$$

Pf

By claim

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

By Main Lemma

$$\langle vv^T, aa^T \rangle \ge \frac{1}{K} N^2 \left(1 - \frac{2^{O(t)} t^{t/2} K^2}{\Delta^t} \right)$$

So

$$\|vv^T - aa^T\| \le \left[2\left(\underbrace{\frac{n^2}{K^3} - \frac{N^2}{K}}_{0} - \underbrace{\frac{N^2}{K}}_{= \frac{n^2}{K^3} = \|aa^T\|^2} \frac{2^{O(t)} t^{t/2} K^2}{\Delta^t} \right) \right]^{1/2}$$

$$= \|aa^T\| \cdot \left(\frac{2^{O(t)} t^{t/2} K^2}{\Delta^t} \right)^{1/2}$$

$\blacksquare$

<u>Recap</u> $X_1 \ldots X_n$, $S_1 \ldots S_k$

we solve degree $O(t)$ SOS SDP to get pseudodistribution 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

<u>Note</u> from aa^T we know $S_1 \ldots S_k$ and ww^T is very close entry-wise to aa^T

<u>Rounding Alg</u>

(1) Let $I = [n]$ be active indices

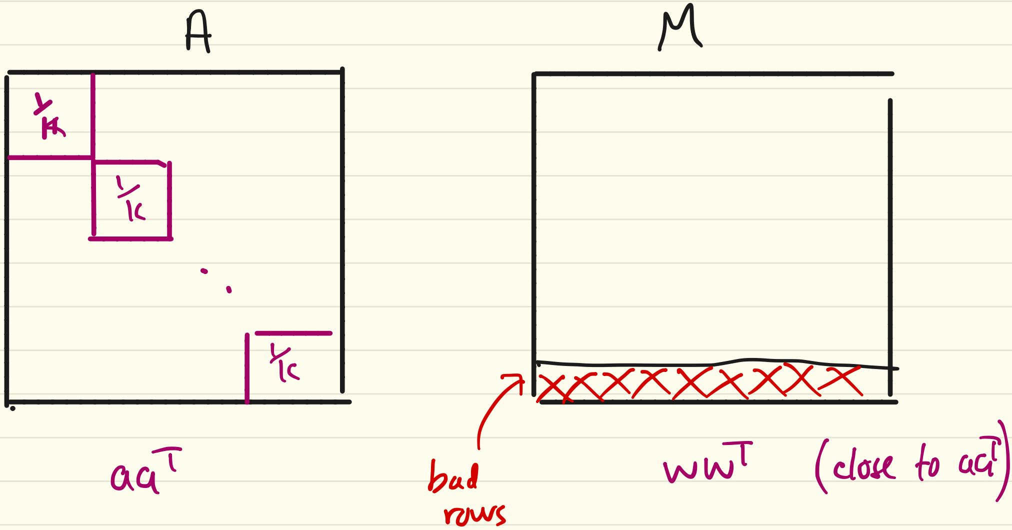
(2) Pick $i \sim I$ uniformly. Let $S \subseteq I$ be the indices j s.t. $\| M_i - M_j \| \leq \delta \sqrt{\frac{n}{k}}$
(so the rows that are almost the same as row i)

add S to list of clusters. Let $I = I \setminus S$

(3) If $|I| \geq n/2k$ go to (2)

(4) Assign remaining indices to clusters til all have size n/k

Sketch of Correctness Pf



A row i is <u>good</u> if $\|M_i - A_i\| < \frac{1}{100}\sqrt{\frac{n}{k}} = \frac{1}{100}\|A_i\|$

There are $\ll \varepsilon^2 n$ bad rows (by avg'ing)

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

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

since $\sum_{i\in n}\|M_i - A_i\|^2 = \|M-A\|^2 \leq \varepsilon^2\|A\|^2$