CSC2429: Learning Gaussian Mixtures with the Sum-of-Squares Proof System

Alex Edmonds

March 7, 2018

These notes were prepared in conjunction with a presentation given for CSC2429 at the University of Toronto. They are based on a series of blog posts by Sam Hopkins [4].

1 Introduction

1.1 Mixtures of Gaussians

Let $\mathcal{D}_1, \ldots, \mathcal{D}_k$ be spherical Gaussians on $\mathbb{R}^d$ with means $\mu_1, \ldots, \mu_k$ and variance 1. We consider mixtures of Gaussians defined by the following sampling procedure:

1. Choose $i \in [k]$ uniformly at random;
2. Return a sample X drawn from $\mathcal{D}_i$

1.2 The Learning Problem

Given i.i.d. samples $X_1, \ldots, X_n$ drawn from an unknown mixture of Gaussians, our goal is to recover the means $\mu_1, \ldots, \mu_k$.

- Actually, we will recover the true cluster membership of most samples.

1.3 Parameter Regime

If the means $\mu_1, \ldots, \mu_k$ are allowed to be dense, then exponentially many samples may be required to learn the mixture.

- This is true even from the information-theoretic point of view.

For this reason, we consider a parameter Δ which denotes the separation between means, i.e. $||\mu_i - \mu_j|| \geq \Delta$ for $i \neq j$.

In the end, the parameter regime we would like is

- $n \sim \text{poly}(d)$
- runtime $\text{poly}(d)$
- Δ as small as possible

1.4 History

1. With radius of clusters $\sim \sqrt{d}$ and $\Delta > 4\sqrt{d}$, the problem is solved easily with a greedy clustering algorithm
2. [Dasgupta 1999]
   - polytime algorithm for $\Delta = \epsilon\sqrt{k}$.
   - Gaussians not necessarily spherical
   - ϵ is a parameter which describes the geometry of the variance
3. [Dasgupta-Schulman 2013]
   - expectation maximization (EM)
   - polytime algorithm for $\Delta \sim k^{1/4}$
4. [Regev-Vijayaraghavan 2017]
   - maximum likelihood estimation (MLE)
   - $\Delta = O(\sqrt{\log d})$
   - only poly(d) samples but the runtime is exp(d)

1.5 Main Theorem

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

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

for some universal constant C.

Special Case A. If $\Delta = k^\epsilon$ (where $\epsilon > 0$) and $t = 100/\epsilon$, then the algorithm:

- uses poly(k, d) samples;
- runs in poly(k, d) time;
- and recovers the correct clustering up to 1/poly(k) errors.

Special Case B. For some universal constant C', if $\Delta = C'\sqrt{\log k}$ and $t = O(\log k)$, then the algorithm:

- uses quasipoly(k, d) samples;
- runs in quasipoly(k, d) time;
- and recovers the correct clustering up to 1/poly(k) errors.

1.6 Related work

[Diakonikolas-Kamath-Kane-Li-Moitra-Stewart 2017]

- does not use SOS
- learning a single high-dimensional Gaussian rather than a mixture
- Gaussian is not necessarily spherical
  - estimate mean and variance matrix
- *outlier robust estimation:*
  - accurately estimate parameters even when an ϵ-fraction are corrupted by an adversary

[Kothari-Steurer 2017]

- does use SOS
- outlier robust estimation
- applies to distributions other than Gaussians; assumes moment bounds instead
- not just estimating means and variance, estimate other low-degree moments also

2 Setup ($d = 1$ dimension)

Throughout our discussion of $d = 1$, we make the following assumptions.

- We are given samples $X_1, \dots, X_n \in \mathbb{R}$.
- There is an unknown partition $\{S_1, \dots, S_k\}$ of $[n]$ into k parts of size $N = n/k$ such that each part $\{X_j\}_{j \in S_i}$ obeys the empirical moment bound

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

  where μ_i is the empirical mean $\mathbb{E}_{j \sim S_i} X_j$.
- $|\mu_i - \mu_j| \ge \Delta$ for $i \ne j$.

Since the latter two conditions hold with high probability, we only require that our algorithm succeeds when they are satisfied.

Goal. Up to permutation, obtain a partition $\{T_1, \dots, T_k\}$ such that $T_i \approx S_i$.

In particular, we want

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

Why we want to learn $\mathbb{E}_{a \sim \nu} aa^T$.

Let $a_1, \dots, a_k \in \{0,1\}^n$ be the 0/1 indicators for clusters $S_1, \dots, S_k$.

Let ν be the uniform distribution on $\{a_1, \dots, a_k\}$.

Note that the matrix $\mathbb{E}_{a \sim \nu} aa^T$ reveals the cluster membership of samples since

$$\left[\mathop{\mathbb{E}}_{a \sim \nu} aa^T\right]_{s,t} := \frac{1}{k} \sum_{i \in [k]} a_{i,s} a_{i,t} = \begin{cases} \frac{1}{k}, & \text{if } X_s \text{ and } X_t \text{ in same cluster} \\ 0, & \text{otherwise} \end{cases}$$

Furthermore, the s^{th} and t^{th} rows of $\mathbb{E}_{a\sim\nu}\, aa^T$ agree iff X_s and X_t are in the same cluster.

What we learn instead:

$$\text{Learn a pseudo-expectation } \tilde{\mathbb{E}} ww^T.$$

3 As a semi-definite program

Think of $(w_1, \ldots, w_n) \in \mathbb{R}^n$ as the 0/1 indicator vector for a cluster T.

Let $\mathcal{A}$ be the set of equations and inequalities

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

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

3.1 Approximation Lemma. *Let $\tilde{\mathbb{E}}$ be a degree $O(t)$ pseudo-expectation solving*

$$\min ||\tilde{\mathbb{E}} ww^T|| \textit{ such that } \tilde{\mathbb{E}} \textit{ satisfies } \mathcal{A}$$

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

$$||\tilde{\mathbb{E}} ww^T - \mathop{\mathbb{E}}_{a\sim\nu} aa^T|| \leq ||\,\mathbb{E}\, aa^T|| \cdot \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t}\right)^{1/2}$$

Why min $||\tilde{\mathbb{E}}\mathbf{ww^T}||$?

For intuition, consider instead the case where ν is an actual distribution supported on vectors $a_1, \ldots, a_k \in \{0,1\}^n$ which are 0/1 indicators for clusters $S_1, \ldots, S_k$.
Using $\langle a_i, a_j\rangle = 0$ for $i \neq j$, obtain

$$\begin{aligned} ||\mathop{\mathbb{E}}_{a\sim\nu} aa^{\mathrm{T}}||^2 &= \left\langle \sum_{i\in[k]} \nu(a_i) a_i a_i^{\mathrm{T}}, \sum_{i\in[k]} \nu(a_i) a_i a_i^{\mathrm{T}} \right\rangle \\ &= \sum_{i\in[k]} \nu(a_i)^2 \cdot ||a_i||^4 = ||\nu|| \cdot (n/k)^2 \end{aligned}$$

which is minimized when ν is uniform.

Since the distribution we are trying to approximate has this structure, minimization of $||\tilde{\mathbb{E}} ww^T||$ provides a ‘soft’ way of enforcing this structure on $\tilde{\mathbb{E}}$.

To prove the Approximation Lemma, we need the following result:

3.2 Main Lemma. *Any degree $O(t)$ pseudo-expectation $\tilde{\mathbb{E}}$ which satisfies $\mathcal{A}$ must also satisfy*

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

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

Note, this result does not assume that $\tilde{\mathbb{E}}$ is necessarily the minimizer for $||\tilde{\mathbb{E}}ww^T||$. In fact, any $\tilde{\mathbb{E}}$ which satisfies $\mathcal{A}$ will give us one cluster since $\sum_{i\in[k]}\left(\frac{|T\cap S_i|}{N}\right)^2$ is a lower bound on $\max_{i\in[k]}\frac{|T\cap S_i|}{N}$. However, iterating this procedure to identify multiple clusters leads to difficult error analysis where analysis of later rounds must account for errors made in earlier rounds.

The Approximation Lemma is an immediate consequence of the Main Lemma:

Proof of Approximation Lemma. The uniform distribution ν over $a_1, a_2, \dots, a_k$ is a feasible solution and satisfies $||\,\mathbb{E}_{a\sim\nu}\, aa^T||^2 = n^2/k^3$.

Hence, the optimal pseudo-expectation $\tilde{\mathbb{E}}$ satisfies $||\tilde{\mathbb{E}}ww^T||^2 \leq n^2/k^3$.

Furthermore,

$$\begin{aligned} ||\tilde{\mathbb{E}}ww^T - \mathop{\mathbb{E}}_{a\sim\nu} aa^T||^2 &= ||\tilde{\mathbb{E}}ww^T||^2 + ||\mathop{\mathbb{E}}_{a\sim\nu} aa^T||^2 - 2\langle\tilde{\mathbb{E}}ww^T, \mathbb{E}\, aa^T\rangle \\ &\leq 2\left(\frac{n^2}{k^3} - \langle\tilde{\mathbb{E}}ww^T, \mathbb{E}\, aa^T\rangle\right) \end{aligned}$$

Using the Main Lemma, we obtain

$$\langle\tilde{\mathbb{E}}ww^T, \mathbb{E}\, aa^T\rangle = \frac{1}{k}\tilde{\mathbb{E}}\sum_{i\in[k]}|S_i\cap T|^2 \geq \frac{n^2}{k^3}\cdot\left(1 - \frac{2^{O(t)}t^{t/2}k^2}{\Delta^t}\right)$$

which, by substitution, implies

$$||\tilde{\mathbb{E}}ww^T - \mathop{\mathbb{E}}_{a\sim\nu} aa^T|| \leq \left(\frac{2n^2}{k^3}\right)^{1/2}\left(\frac{2^{O(t)}t^{t/2}k^2}{\Delta^t}\right)^{1/2} \leq ||\,\mathbb{E}\, aa^T||\cdot\left(\frac{2^{O(t)}t^{t/2}k^2}{\Delta^t}\right)^{1/2}$$

■

We now turn to proving the Main Lemma:

Proof of Main Lemma. Proof: Let $\tilde{\mathbb{E}}$ be a degree $O(t)$ pseudo-expectation which satisfies $\mathcal{A}$.

Assume $i \neq j$. Since $\mu_i = \mathbb{E}_{\ell\sim S_i} X_\ell$ and $\mu_j = \mathbb{E}_{\ell\sim S_j} X_\ell$ satisfy $|\mu_i - \mu_j|^2 \geq \Delta^2$, the SOS triangle inequality implies

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

Hence,

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

Recall that, by Lemma 1,

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

Therefore,

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

Since $\mathcal{A} \vdash_t w_i^2 \geq 1$ and thereby $\mathcal{A} \vdash_t |T \cap S_i| \leq N$, it follows that

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

Furthermore, by pseudo-expecation Cauchy-Schwartz,

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

Combined with the preceding, we get

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

Now by the pseudo-expectation Holder's inequality,

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

so

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

Cancelling out terms on both sides of

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

and then applying Cauchy-Schwartz results in

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

Finally, since

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

we may obtain

$$\begin{aligned}
\tilde{\mathbb{E}}\left[\sum_{i \in [k]} \left(\frac{|T \cap S_i|}{N}\right)^2\right] &= \frac{1}{N^2} \tilde{\mathbb{E}}\left[\sum_{i \in [k]} |T \cap S_i|^2\right] \\
&= \frac{1}{N^2}\left[\tilde{\mathbb{E}} \sum_{i,j \in [k]} |T \cap S_i||T \cap S_j| - \sum_{i \neq j} |T \cap S_i||T \cap S_j|\right] \\
&= \frac{1}{N^2}\left[N^2 - \frac{k^2 2^{O(t)} t^{t/2} N^2}{\Delta^t}\right] = 1 - \frac{k^2 2^{O(t)} t^{t/2}}{\Delta^t}
\end{aligned}$$

■

4 Rounding for the win

4.1 Rounding Lemma. *Let $A \in \{0,1\}^{n \times n}$ be the $0/1$ same-set indicator matrix for a partition $\{S_1, \ldots, S_k\}$ of $[n]$ into k parts of size $N = n/k$.*

- *i.e. $A_{ij} = 1$ iff $\exists p$ such that $i, j \in S_p$.*

Suppose also that $M \in \mathbb{R}^{n \times n}$ satisfies $||M - A|| \leq \epsilon ||A||$.

Then there is a polytime algorithm which, on input M, with probability $1 - O(k^2\epsilon)$, produces a partition $T_1, \ldots, T_k$ of $[n]$ into clusters of size N such that

$$\frac{S_i \cap T_i}{N} \geq 1 - O(\epsilon^2 k) \qquad \forall i$$

with appropriate permutation.

Once we have the Rounding Lemma, the Main Theorem will follow:

Proof of Main Theorem. Our algorithm is:

1. Given $X_1, \ldots, X_n$, solve

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

2. Then apply the rounding algorithm guaranteed by our lemma and output the resulting partition.

A concentration argument shows that the vectors $X_1, \ldots, X_n$ satisfy the empirical moment bound with probability 0.99 when $n \geq \text{poly}(k)$.

■

Proof of Main Theorem. In the case that the samples $X_1, \ldots, X_n$ do obey the moment bound, then the conclusion of the Approximation Lemma holds, namely

$$||\tilde{\mathbb{E}} ww^T - \mathop{\mathbb{E}}_{a \sim \nu} aa^T|| \leq ||\mathbb{E}\, aa^T|| \cdot \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t} \right)^{1/2}$$

where ν is the uniform distribution on the 0/1 indicator vectors for $S_1, \ldots, S_k$.

Now we may apply our rounding algorithm by taking

$$A = k\ \mathbb{E}_{a \sim \nu} aa^T \qquad\qquad M = k\ \tilde{\mathbb{E}} ww^T \qquad\qquad \epsilon = \left(\frac{2^{O(t)} t^{t/2} k^2}{\Delta^t} \right)^{1/2}$$

It follow that, up to permutation, the rounding algorithm produces a partition $T_1, \ldots, T_k$ of $[n]$ such that

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

■

It remains to prove the Rounding Lemma:

Proof of Rounding Lemma. We use consider following algorithm, with $\delta = 0.1$:

```
Rounding Algorithm
1. Let $\mathcal{I} = [n]$ be the set of active indices.
2. Pick $i \sim \mathcal{I}$ uniformly.
3. Let $T \subseteq \mathcal{I}$ be those indices $j$ for which $||M_j - M_i|| \leq \delta \cdot \sqrt{n/k}$.
4. Add $T$ to the list of clusters and let $\mathcal{I} := \mathcal{I} \setminus T$
5. If $|\mathcal{I}| > n/2k$, go to Step 2.
```

A row index i is considered *good* if

$$||M_i - A_i|| \leq \frac{\delta}{2} \cdot \sqrt{n/k} = \frac{\delta}{2}||A_i||$$

Otherwise, say that i is *bad*.

1. Good rows are not misclassified:

   Case A: If i, j are good indices in the same cluster S_p, then $A_i = A_j$ implies

   $$||M_i - M_j|| \leq ||M_i - A_i|| + ||M_j - A_j|| \leq \delta \cdot \sqrt{n/k}$$

   so, if the algorithm selects i, then j is placed in the same cluster.

   Case B: Let $\delta < 0.1$. Suppose i and j are good indices in distinct clusters S_p and S_q resp. Then $||A_i - A_j|| = 2\sqrt{n/k}$ implies

   $$\begin{aligned} ||M_i - M_j|| &\geq ||A_i - A_j|| - ||A_i - M_i|| - ||A_j - M_j|| \\ &\geq 2\sqrt{n/k} - \delta \cdot \sqrt{n/k} \geq \delta \cdot \sqrt{n/k} \end{aligned}$$

   so, if the algorithm selects i, then j is not placed in the same cluster.

2. Most rows are good:

   By hypothesis,

   $$\sum_{i \in [n]} ||M_i - A_i||^2 = ||M - A||^2 \leq \epsilon^2 ||A||^2$$

   Each bad index contributes $\frac{\delta^2}{4}||A_i||^2$ to the left side and $||A_i||^2 = ||A||^2/n$ so the number of bad indices is at most

   $$\left(\epsilon^2 ||A||^2\right) \Big/ \left(\frac{\delta^2 ||A||^2}{4n}\right) = 4\epsilon^2 n/\delta^2$$

3. Algorithm succeeds if bad indices are never chosen:

   Suppose the algorithm never chooses a bad index.

   Then, before post-processing, the clusters $T_1, \ldots, T_k$ satisfy that T_i contains all good indices in S_i.

   Hence, only bad indices are misclassified or not classified. This implies that at most $4\epsilon^2 n/\delta^2$ are moved in the post-processing step.

   In the end, the only misclassified are

   - those that were misclassified before post-processing;
   - those that were moved after post-processing because they were displaced by some misclassified point.

   It follows that at most $8\epsilon^2 n/\delta^2$ are misclassified.

   In particular, each S_i differs from each T_i on at most $8\epsilon^2 n/\delta^2$ points.

4. With high probability, bad indices are never chosen:

   Consider implementing the algorithm by drawing a list L of k^2 indices before seeing M.

   When the algorithm asks for a random index $i \in \mathcal{I}$, it is given the next element from L which is still in $\mathcal{I}$.

The probability that L intersects with $\mathcal{I}$ as long as $|\mathcal{I}| > n/2k$ is at least $1 - O(1/k)$.

Furthermore, since there are only $4\epsilon^2 n/\delta^2$ bad indices,

$$\Pr(L \text{ does not contain a bad index}) \leq \sum_{i=1}^{k^2} \Pr(L_i \text{ is a bad index}) = \frac{4\epsilon^2 k^2}{\delta^2}$$

With $\delta = 0.01$, we may concude that the probability of our algorithm encountering a bad index is at most $O(\epsilon^2 k^2) + O(1/k) = O(\epsilon^2 k^2)$ [since $k = \Omega(1/\epsilon)$].

■

5 Moving to higher dimensions...

5.1 Identifiability

Recall that previously we used the t^{th} moment bound which was defined for an arbitary set S as

$$\mathop{\mathbb{E}}_{j \in S} |X_j - \mu|^t \leq 2 \cdot t^{t/2} \qquad (*)$$

where $\mu = \mathbb{E}_{j \in S} X_j$ is the empirical mean of the cluster S_i.

Three important properties were used:

1. With high probability, $(*)$ holds for all true clusters $S_1, \ldots, S_k$;
2. Supposing $(*)$ holds for all true clusters $S_1, \ldots, S_k$ as well as for some arbitary set T, then $T \approx S_i$ for some i;
3. $(*)$ may be expressed in the language of SOS.

In higher dimensions, it is tempting to generalize the moment bound as follows.

$$\forall u \in \mathbb{R}^d, \qquad \mathop{\mathbb{E}}_{h \sim S} \left\langle X_j - \mu, \frac{u}{||u||} \right\rangle^t \leq 2 \cdot t^{t/2} \qquad (\diamond)$$

where $\mu = \mathbb{E}_{j \in S} X_j$ is the empirical mean of the cluster S_i. Indeed, this is enough for identifiability (Properties 1 and 2):

- If S and S' are clusters which both satisfy $(\diamond)$, then large $|S \cap S'|$ implies small $|\mu - \mu'|$ [where μ and μ' are the respective empirical means].
- To show this, reduce to the 1-dimensional case by projecting onto the line (μ, μ'). In particular, $(\diamond)$ implies that

$$\left\{ \left\langle X_j - \mu, \frac{\mu - \mu'}{||\mu - \mu'||} \right\rangle^t \right\}_{j \in S} \qquad \text{and} \qquad \left\{ \left\langle X_j - \mu, \frac{\mu - \mu'}{||\mu - \mu'||} \right\rangle^t \right\}_{j \in S'}$$

  are sets of points which each satisfy the 1-dimensional t^{th} moment bound

Unfortunately, $(\diamond)$ cannot be expressed in the language of SOS:

- Having $||u||$ in the denominator means that these are not low degree polynomials.
  - Just multiply through by $||u||$ and raise to a power to get a polynomial inequality.

- The moment bound has the quantifier "$\forall u \in \mathbb{R}^d$"
    - This is a more serious issue.

We want to enforce

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

by way of a low degree polynomial inequality.

- Does there exist a low degree polynomial inequality which certifies $(\star)$?
    - Probably not. Use a stronger certificate instead.

When $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ are sufficiently-many samples from a d-dimensional Gaussian, the following inequality holds w.h.p.:

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

Moreover, it implies $(\star)$.

5.2 The multi-dimensional setup

We are given samples $X_1, \ldots, X_n \in \mathbb{R}^d$. Assume

1. There is an unknown partition $\{S_1, \ldots, S_k\}$ of $[n]$ into k pieces of size $N = n/k$ such that each collection $\{X_j\}_{j \in S_i}$ has bounded moments by way of

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

   where μ_i is the empirical mean $\mathbb{E}_{j \sim S_i} X_j$.

2. $||\mu_i - \mu_j|| \geq \Delta$ for $i \neq j$.

Goal. Up to permutation, obtain a partition $\{T_1, \ldots, T_k\}$ such that $T_i \approx S_i$. In particular, we want

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

5.3 Multi-dimensional setup – SDP

Again, think of $(w_1, \ldots, w_n)$ as the indicator vector for a cluster T.

Let $\mathcal{B}$ be the set of equations and inequalities

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

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

Objective: $\min ||\tilde{\mathbb{E}} w w^T||$

As before, we use semi-definite programming to obtain a degree $O(t)$ pseudo-expectation $\tilde{\mathbb{E}}$ which minimizes $||\tilde{\mathbb{E}}ww^T||$ subject to $\mathcal{B}$.

Then $\tilde{\mathbb{E}}ww^T$ provides an approximation of the same-cluster matrix, which allows us to determine cluster membership using the same algorithm as before.

5.4 Main Lemma

This time, we rely on the following multi-dimensional version of our earlier Main Lemma:

5.1 Main Lemma (multi-dimensional). *Any degree $O(t)$ pseudo-expectation $\tilde{\mathbb{E}}$ which satisfies $\mathcal{B}$ must also satisfy*

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

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

Largely, the proof of the Main Lemma will follow the proof of the 1-d case.

The main difference will be in leveraging the new moment bound. Doing so will rely on the following proposition:

5.2 Moment Inequality. *Let $u \in \mathbb{R}^d$ be an indeterminate. Then,*

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

Proof of Moment Inequality. We may expand $\mathbb{E}\langle\mathcal{Y},u\rangle^t$ as

$$\mathbb{E}\langle\mathcal{Y},u\rangle^t = \mathbb{E}\sum_{|\alpha|=t} u^\alpha \mathcal{Y}^\alpha = \sum_{|\alpha|=t} u^\alpha\,\mathbb{E}\,\mathcal{Y}^\alpha = \sum_{\substack{|\alpha|=t\\ \alpha\text{ even}}} u^\alpha\,\mathbb{E}\,\mathcal{Y}^\alpha$$

where α is a multi-index over $[n]$, $u^\alpha = \prod_i u_i^{\alpha_i}$, and "even" means that every element of α is even.

Each monomial u^α on the RHS is a square.

$\mathbb{E}\,\mathcal{Y}^\alpha \leq t^{t/2}$ holds by a standard moment bound but this only involves constants. Hence,

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

■

Proof of the Main Lemma relies on the following intermediate fact.

5.3 Mean-bound Lemma (analogue of Lemma 1). *Suppose $S \subseteq [n]$, $|S| = N$, has empirical mean $\mu_S = \mathbb{E}_{i\sim S} X_i$.*

Then, under our previous assumptions,

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

Note that this is the polynomial version of the inequality

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

Proof of the Mean-bound Lemma. First, expand in terms of $X_1, \dots, X_n$.

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

Apply SOS Holder's inequality to the RHS to obtain

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

Using $\vdash_t (a-b)^t \leq 2^t(a^t + b^t)$ and $w_i = w_i^2 \leq 1$, we get

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

So that we can eventually apply Cauchy-Schwarz, square both sides and apply $\vdash_2 (a+b)^2 \leq 2(a^2+b^2)$ to the RHS.

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

It is now enough to show

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

and

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

First we show

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

Working from the LHS, we have

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

Once again, square both sides and apply $\vdash_2 (a+b)^2 \leq 2(a^2+b^2)$ to get

$$\begin{aligned}\frac{1}{2N^2}\left(\sum_{i\in S} \langle X_i - \mu_S, \mu - \mu_S\rangle^t\right)^2 &= \mathop{\mathbb{E}}_{\mathcal{Y}\sim\mathcal{N}(0,I)} \langle \mathcal{Y}, \mu - \mu_S\rangle^{2t} \\ &\quad + \left(\frac{1}{N}\sum_{i\in S} \langle X_i - \mu_S, \mu - \mu_S\rangle - \mathop{\mathbb{E}}_{\mathcal{Y}\sim\mathcal{N}(0,I)} \langle \mathcal{Y}, \mu - \mu_S\rangle^t\right)^2\end{aligned}$$

We bound each term on the RHS. By the Proposition,

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

For the second term, we have

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

where M is the matrix

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

Apply Cauchy-Schwarz to get

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

This allow us to conclude

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

It remains to show

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

but this follows a similar argument with

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

in place of $||M||^2 \leq 1$.

■

References

[1] S. Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (New York, 1999)*, pages 634–644. IEEE Computer Soc., Los Alamitos, CA, 1999.

[2] S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *J. Mach. Learn. Res.*, 8:203–226, 2007.

[3] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 655–664. IEEE Computer Soc., Los Alamitos, CA, 2016.

[4] S. Hopkins. Clustering and sum of squares proofs. https://windowsontheory.org/2017/12/11/clustering-and-sum-of-squares-proofs-part-1, 2017.

[5] P. K. Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.

[6] O. Regev and A. Vijayaraghavan. On learning mixtures of well-separated gaussians. *CoRR*, abs/1710.11592, 2017.