9/25/24

Lecture 7: SoS and Gaussian mixtures:

Let $x_1, \dots, x_n \in \mathbb{R}^d$ be samples from

$$q = \frac{1}{k} \sum_{j=1}^{k} N(\mu_j, I_d)$$

Define $\Delta \triangleq \min_{i \neq j} \|\mu_i - \mu_j\|_2$

$$N \triangleq \frac{n}{k} \quad (\approx \# \text{ pts in each component})$$

Let t (sos degree) be power of 2. Suppose

$$\Delta \gg \sqrt{t}\, k^{1/t}.$$

$\boxed{\text{SoS program}}$

Variables: $a_1, \dots, a_n$ (1-dimensional)

μ (d-dimensional)

Constraints:

1). $a_i^2 = a_i$ (Boolean indicators)

2). $\sum_{i=1}^{n} a_i = N$ (selects out enough points for one component)

3) $\frac{1}{N} \sum_i a_i x_i = \mu$ (μ is empirical mean of points selected)

4). $\frac{1}{N} \sum_i a_i \langle u, x_i - \mu \rangle^t \leq 2t^{t/2} \|u\|_2^t$ for all vectors u* (selected points have Gaussian moment bounds)

Let $S_j \subset [n]$ denote samples from $N(\mu_j, I_d)$

For convenience, we will pretend $|S_j| = N$ exactly $\forall j$.

---

* Seems like infinite constraints... see below for how to quantify over all $u \in S^{d-1}$

$\boxed{\text{For now, assume } d=1}$

So constraint 4) becomes $\frac{1}{N}\sum_i a_i (x_i-\mu)^t \le 2t^{t/2}$.

Warmup lemma: Let $S^\circ = S_j$, $\mu^\circ = \mu_j$ for any $j\in[k]$.

There is a deg-$O(t)$ proof that:

overlap b/t our points and S°

$$\left(\sum_{i\in S^\circ} a_i\right)^t (\mu-\mu^\circ)^t \le 2^{O(t)}\left(\sum_{i\in S^\circ} a_i\right)^{t-1}\cdot N\cdot t^{t/2}.$$

Pf: $\left(\sum_{i\in S^\circ} a_i\right)^t\cdot(\mu-\mu^\circ)^t$

$$= \left(\sum_{i\in S^\circ} a_i(\mu-\mu^\circ)\right)^t$$

$$= \left(\sum_{i\in S^\circ} a_i\left[(\mu-x_i)-(\mu^\circ-x_i)\right]\right)^t$$

---

by degree-t Hölder's inequality (SoS-able):

$$\left(\sum_i b_i c_i\right)^t = \Big(\sum_i \underbrace{b_i^{\frac{t-1}{t}}}_{L_p}\ \underbrace{b_i^{\frac{1}{t}}\ c_i}_{L_q}\Big)^t \le \left(\sum_i b_i\right)^{t-1}\left(\sum b_i\, c_i^t\right)$$

for $p=\frac{t}{t-1}$, $q=t$

so $\frac{1}{p}+\frac{1}{q}=1$

$$\le \left(\sum_{i\in S^\circ} a_i\right)^{t-1}\cdot\left(\sum_{i\in S^\circ} a_i\left[(\mu-x_i)-(\mu^\circ-x_i)\right]^t\right)$$

$(a-b)^t \le 2^t(a^t+b^t)$

$$(*)\qquad \le 2^t\left(\sum_{i\in S^\circ} a_i\right)^{t-1}\cdot\left(\sum_{i\in S^\circ} a_i\left[(\mu-x_i)^t+(\mu^\circ-x_i)^t\right]\right)$$

Note: $\sum_{i\in S^\circ} a_i(\mu-x_i)^t \le \sum_{\text{all } i} a_i(\mu_i-x_i)^t \le N\cdot 2t^{t/2}$

moment bound, i.e. constraint 4

$$\sum_{i \in S^\circ} a_i (\mu^\circ - x_i)^t \underset{\uparrow}{\le} \sum_{i \in S^\circ} (\mu^\circ - x_i)^t \underset{\uparrow}{\le} N \cdot 2 t^{t/2}$$

Booleanity, i.e. constraint 1

assuming tth empirical moment of actual samples from component concentrate

So

$$\left(\sum_{i \in S^\circ} a_i\right)^t (\mu - \mu^\circ)^t \le 2^{O(t)} \left(\sum_{i \in S^\circ} a_i\right)^{t-1} \cdot N \cdot t^{t/2} . \quad \square$$

Note, if we could "divide on both sides" and take t th roots, we would get

$$|\mu - \mu^\circ| \lesssim \left(\frac{1}{N} \sum_{i \in S^\circ} a_i\right)^{-1/t} \cdot \sqrt{t} \qquad (*)$$

i.e. if overlap between our points (chosen by a_i) and true points in component S° is large, then our μ is close to the mean of S°.

Claim 1: If a_i's were real indicators of a set S satisfying $(*)$ for every center $\mu^\circ = \mu_j$, then

Component S_{j^*} with largest overlap with S satisfies $|S \cap S_{j^*}| = (1-\delta)N$ for $\delta \le k t^{t/2} \cdot O(1/\Delta)^t$.
$(\ll 1)$

<u>Pf</u>: Note $1-\delta \geq \frac{1}{k}$ by averaging,

and $|S \cap S_j| \geq \frac{\delta}{k} \cdot N$ for some $j \neq j^*$. So:

$$|\mu_{j^*} - \mu| \underset{\text{Eq. }(*)\text{ applied to comp. }j^*}{\lesssim} (1-\delta)^{-1/t} \sqrt{t} \lesssim k^{1/t} \sqrt{t} \ll \Delta/2,$$

so $|\mu_j - \mu| > \frac{\Delta}{2}$. Thus, by $(*)$ applied to comp. j,

$$\frac{\Delta}{2} < |\mu_j - \mu| \lesssim \left(\frac{\delta}{k}\right)^{-1/t} \sqrt{t},$$

so $$\delta^{1/t} \lesssim \frac{k^{1/t}\sqrt{t}}{\Delta} \quad (\ll 1),$$

and thus $\delta \le k t^{t/2} \cdot O(1/\Delta)^t$ as claimed. $\square$

i.e. a_i's must have $1-o(1)$ overlap with some component!

IOU's:

- SoS version of claim 1?
- rounding SoS solution?

closely related

- $d > 1$?

Issue with Claim 1 is it breaks symmetry across clusters. Makes it unclear how to round.

Claim 2 (symmetric version of Claim 1 — still not SoS):

If a_i's are indicator of S, then

$$\sum_{j=1}^{k} \left(\frac{|S_j \cap S|}{N} \right)^2 \geq 1 - k^2 t^{t/2} \cdot O(1/\Delta)^t$$

Note: This implies Claim 1. Define $c_j \triangleq \frac{|S_j \cap S|}{N}$ so $\sum_j c_j = 1$. Thus

$$1 - k^2 t^{t/2} \cdot O(1/\Delta)^t \leq \sum_j c_j^2 \leq (\max_j c_j) \cdot \overbrace{\sum_j c_j}^{\substack{1 \\ \shortparallel}} = \max_j c_j,$$

i.e. exists j s.t. $\frac{1}{N}|S_j \cap S| \geq 1 - k^2 t^{t/2} \cdot O(1/\Delta)^t$, which recovers Claim 1 w/ extra (but unimportant) k factor.

Pf: Define $c_j = \frac{|S_j \cap S|}{N}$. Then

$$1 = \left(\sum_j c_j \right)^2 = \sum_j c_j^2 + 2 \sum_{i<j} \underbrace{c_i c_j}_{\text{we'll show these are small}}$$

$$c_i^{1/t} c_j^{1/t} \le c_i^{1/t} c_j^{1/t} \underbrace{\frac{|\mu_i - \mu| + |\mu_j - \mu|}{\Delta}}_{\ge 1 \text{ by triangle ineq.}}$$

$$\le c_i^{1/t} \frac{|\mu_i - \mu|}{\Delta} + c_j^{1/t} \frac{|\mu_j - \mu|}{\Delta}$$

Recall by warmup lemma, in particular $(\circledast)$,

$$|\mu_i - \mu| \lesssim c_i^{-1/t} \cdot \sqrt{t}, \quad \text{so}$$

$$\lesssim \frac{\sqrt{t}}{\Delta},$$

thus $\sum_{i \ne j} c_i \cdot c_j \le k^2 \cdot O\left(\frac{\sqrt{t}}{\Delta}\right)^t$ as desired. $\square$

Next "SoS-ize" Claim 2 (the following is pedantic and can be ignored upon first reading)

<u>Claim 3</u> (SoS version of Claim 2):

For any deg-t pseudodistribution over $\{a_i\}, \mu$,

$$\tilde{\mathbb{E}}\left[\sum_{j=1}^{k}\left(\frac{1}{N}\sum_{i \in S_j} a_i\right)^2\right] \ge 1 - k^2 t^{t/2} \cdot O(1/\Delta)^t$$

Pf: Define $c_j \triangleq \frac{1}{N} \sum_{i \in S_j} a_i$ (now a deg-1 polynomial).

Recall the only thing we have proved about $\hat{\mathbb{E}}$ is that for all $j \in [k]$,

$$c_j^t \cdot (\mu - \mu_j)^t \leq O(t)^{t/2} \cdot c_j^{t-1}. \qquad (1)$$

Next, note that

$$\Delta^t \leq (\mu_i - \mu_j)^t$$
$$= [(\mu_i - \mu) - (\mu_j - \mu)]^t$$
$$\leq 2^t [(\mu_i - \mu)^t + (\mu_j - \mu)^t],$$

So $$\frac{(\mu_i - \mu)^t + (\mu_j - \mu)^t}{(\Delta/2)^t} \leq 1 \qquad (2)$$

Combining (1) and (2) yields

$$c_i^t c_j^t \overset{(2)}{\leq} c_i^t c_j^t \frac{(\mu_i - \mu)^t + (\mu_j - \mu)^t}{(\Delta/2)^t}$$

$$\leq (2/\Delta)^t \cdot \left[c_j^t \underline{c_i^t (\mu_i - \mu)^t} + c_i^t \underline{c_j^t (\mu_j - \mu)^t} \right]$$

$$\overset{(1)}{\leq} \left(2\sqrt{t}/\Delta\right)^t \left(c_j^t c_i^{t-1} + c_i^t c_j^{t-1}\right)$$

$$\overset{c_i, c_j \leq 1}{\leq} 2 \cdot \left(2\sqrt{t}/\Delta\right)^t c_i^{t-1} c_j^{t-1}$$

So we have proved in deg-$O(t)$ SoS that

$$c_i^t c_j^t \leq O\left(\sqrt{t}/\Delta\right)^t c_i^{t-1} c_j^{t-1}.$$

To avoid working with odd powers, square both sides to get

$$c_i^{2t} c_j^{2t} \leq O(t/\Delta^2)^t c_i^{2t-2} c_j^{2t-2}$$

Thus,

$$\hat{\mathbb{E}}\left[c_i^{2t} c_j^{2t}\right] \leq O(t/\Delta^2)^t \hat{\mathbb{E}}\left[c_i^{2t-2} c_j^{2t-2}\right] \quad (\star\star)$$

Q: How do we simulate taking t-th roots in SoS?

A: "pseudo-expectation Cauchy-Schwarz / Hölder's inequalities"

Fact ("Cauchy-Schwarz"):

$$\hat{\mathbb{E}}\left[p(x)\, q(x)\right] \leq \hat{\mathbb{E}}\left[p(x)^2\right]^{1/2} \cdot \hat{\mathbb{E}}\left[q(x)^2\right]^{1/2}$$

for any p, q of degree $\leq t/2$ and $\hat{\mathbb{E}}$ any deg-t pseudo-expectation.

Fact ("Hölder's"):

$$\tilde{\mathbb{E}}\left[p(x)^{t-2}\right] \le \tilde{\mathbb{E}}\left[p(x)^t\right]^{\frac{t-2}{t}}$$

for any deg-l sum of squares polynomial p and $\tilde{\mathbb{E}}$ any deg-tl pseudo-expectation.

Pf: Pset 2. $\square$

Applying pseudo-exp. Hölder's to (**), we get

$$\tilde{\mathbb{E}}\left[c_i^{2t} c_j^{2t}\right] \le O(t/\Delta^2)^t\, \tilde{\mathbb{E}}\left[c_i^{2t} c_j^{2t}\right]^{\frac{t-1}{t}}$$

now we can divide freely

$$\Rightarrow \tilde{\mathbb{E}}\left[c_i^{2t} c_j^{2t}\right] \le O(t/\Delta^2)^{t^2}$$

Applying pseudo-exp Cauchy-Schwarz, we have

$$\tilde{\mathbb{E}}\left[c_i c_j\right] = \tilde{\mathbb{E}}\left[c_i c_j \cdot 1\right]$$

$$\le \tilde{\mathbb{E}}\left[(c_i c_j)^2\right]^{1/2} \cdot \cancel{\tilde{\mathbb{E}}\left[1^2\right]^{1/2}}$$

and repeating this $\log_2 2t$ times, get

$$\tilde{\mathbb{E}}(c_i c_j) \le \tilde{\mathbb{E}}(\underbrace{c_i^{2t}\, c_j^{2t}})^{1/2t}.$$

Thus, $\tilde{\mathbb{E}}[c_i c_j] \le O(t/\Delta^2)^{t/2} \quad \forall\, i \ne j,$

and thus

$$\tilde{\mathbb{E}}\left[\sum_j c_j^2\right] = \tilde{\mathbb{E}}\left[\underbrace{\left(\sum_j c_j\right)^2}_{=1} - \sum_{i \ne j} c_i c_j\right]$$

$$\ge 1 - k^2 t^{t/2} O(1/\Delta)^t$$

as desired. $\square$

Rounding:

Can't just output $\tilde{\mathbb{E}}[\mu]$ b/c $\{a_i\}$'s don't preference any particular component...

How do we know $\{a_i\}$'s are indicating a fixed component, or a dist over components?

Trick: entropy maximization

Want pseudo-dist over $\{a_i\}$'s to resemble <u>uniform distribution</u> over <u>true indicators</u> $\{a_i^{(j)}\}$'s, where

$$a_i^{(j)} \triangleq \mathbb{1}[x_i \text{ from } N(\mu_j, Id)]$$

This distribution has high "entropy" as quantified by

$$\left\| \mathbb{E}_j \left[a^{(j)} (a^{(j)})^T \right] \right\|_F^2 \qquad \text{(ENT)}$$

Note:

$$\text{(ENT)} = \sum_{j,j'=1}^{k} \frac{1}{k^2} \langle a^{(j)} (a^{(j)})^T, a^{(j')} (a^{(j')})^T \rangle$$

$$= \sum_{j,j'=1}^{k} \frac{1}{k^2} \underbrace{\langle a^{(j)}, a^{(j')} \rangle^2}_{= 0 \text{ if } j \neq j'}$$

b/c components disjoint,

$$= \frac{1}{k^2} \sum_{j=1}^{k} \| a^{(j)} \|_2^4 = \frac{N^2}{k}$$

We pick the pseudo-distribution solving

$$\min_{\tilde{\mathbb{E}}} \left\| \tilde{\mathbb{E}} \left[\underbrace{(a_1, \ldots, a_n)}_{\triangleq a} (a_1, \ldots, a_n)^T \right] \right\|_F^2$$

subject to $\tilde{\mathbb{E}}$ satisfying constraints of the program.

Lemma: This $\hat{\mathbb{E}}$ satisfies

$$\left\| \hat{\mathbb{E}}[aa^T] - \mathbb{E}_j\left[a^{(j)}(a^{(j)})^T\right]\right\|_F^2 \qquad (\dagger)$$

$$\leq \left\|\mathbb{E}_j\left[a^{(j)}(a^{(j)})^T\right]\right\|_F^2 \Big(\underbrace{k^2 t^{t/2}\cdot O(1/\delta)^t}_{<<1}\Big)$$

Pf: Because unif distribution over $\{\{a^j_i\}_i, \mu_j\}_j$ is a feasible solution, $\|\hat{\mathbb{E}}[aa^T]\|_F^2 \leq \frac{N}{k}$, so

$$(\dagger) = \frac{2N^2}{k} - \frac{2}{k}\sum_{j=1}^{k} \hat{\mathbb{E}}\left[\langle a, a^{(j)}\rangle^2\right]$$

$$= \frac{2N^2}{k} - \frac{2}{k}\sum_{j=1}^{k}\hat{\mathbb{E}}\Big[\big(\underbrace{\textstyle\sum_{i\in S_j} a_i}_{= Nc_j}\big)^2\Big]$$

$$= \frac{2N^2}{k}\Big(1 - \sum_j c_j^2\Big)$$

$$\leq \frac{N^2}{k}\left(k^2 \cdot t^{t/2} \cdot O(1/\Delta)^t\right).$$

$\square$

Note,

$$\hat{\mathbb{E}}_j\left[a^{(j)} (a^{(j)})^T \right] = \begin{array}{c c} & n \\ n & \begin{pmatrix} \boxed{\frac{1}{k}} & & & & \\ & \boxed{\frac{1}{k}} & & & \\ & & \boxed{\frac{1}{k}} & & \\ & & & \boxed{\frac{1}{k}} & \\ & & & & \ddots \end{pmatrix} \end{array}$$

(after row/col permutation),

so Lemma implies that we can read off clustering from $\hat{\mathbb{E}}[aa^T]$!

Final IOU: ... this was all for $d=1$!

Warmup lemma and main Claim 3 easy to generalize, e.g.

<u>Before:</u>

$$\left(\sum_{i \in S_j} a_i\right)^t (\mu - \mu_j)^t \leq 2^{O(t)} \left(\sum_{i \in S_j} a_i\right)^{t-1} N \cdot t^{t/2}.$$

<u>After</u>

$$\left(\sum_{i \in S_j} a_i\right)^t \|\mu - \mu_j\|_2^t \leq 2^{O(t)} \left(\sum_{i \in S_j} a_i\right)^{t-1} N \cdot t^{t/2}.$$

But $\|\mu-\mu_j\|_2^2 = \langle \mu-\mu_j, \mu-\mu_j \rangle$, so can just "project" data along $\mu-\mu_j$ direction and reduce to 1D proof.

(need to be careful b/c $\mu-\mu_j$ is not a real vector because μ is an SoS variable)

trickier: how to impose constraint

$$\frac{1}{N}\sum_{i=1}^{n} a_i \langle u, x_i-\mu\rangle^t \le 2t^{t/2}\|u\|_2^t$$

for all $u \in \mathbb{R}^d$?

Because we will apply this to $u = \mu-\mu_j$, need this to make sense even when u is not a real vector...

Idea: Constrain via

$$(\star\star\star)\quad \underbrace{\left\| \frac{1}{N}\sum_{i=1}^{n} a_i (x_i-\mu)^{\otimes t/2}\left[(x_i-\mu)^{\otimes t/2}\right]^T - \mathbb{E}_{g\sim N(0,Id)}\left(g^{\otimes t/2}\left(g^{\otimes t/2}\right)^T\right)\right\|_F^2 \le 1}_{\approx d^t \text{ polynomial constraints}}$$

(satisfied by $a_i = a_i^{(j)}$ and $\mu = \mu_j$, if n large enough) i.e. pick out subset s.t. empirical order-t moments are close to those of $N(0, I_d)$.

<u>Fact</u>: For an SoS variable u,

$$\mathop{\mathbb{E}}_{g \sim N(0, I_d)} \langle g, u \rangle^t \leq t^{t/2} \cdot \|u\|_2^t$$

has a deg-t SoS proof in u.

<u>Pf</u>:

$$\mathop{\mathbb{E}}_{g} \langle g, u \rangle^t = \sum_{\substack{\text{deg-}t \text{ monomials } \alpha \\ \text{s.t. every} \\ \text{variable appears} \\ \underline{\text{even}} \text{ \# times}}} u_\alpha \, \mathbb{E}[g_\alpha]$$

$$\leq t^{t/2} \sum_{\alpha} u_\alpha$$

$$= t^{t/2} \|u\|_2^t. \qquad \square$$

i.e. $N(0, I_d)$ is $\boxed{\text{"certifiably } t\text{-hypercontractive"}}$

Note, if we take constraint (★★★) and hit it on both sides with $\left[(\mu - \mu_j)^{\otimes t/2}\right]^T \cdots (\mu - \mu_j)^{\otimes t/2}$, we get:

$$\frac{1}{N}\sum_{i=1}^{n} a_i \langle \mu - \mu_j , x_i - \mu \rangle^t - \mathbb{E}_g \langle \mu - \mu_j , g \rangle^t \le \|\mu - \mu_j\|_2^t$$

$$\Downarrow \text{ (using Fact above)}$$

$$\frac{1}{N}\sum_{i=1}^{n} a_i \langle \mu - \mu_j , x_i - \mu \rangle^t \le \left(1 + t^{t/2}\right) \|\mu - \mu_j\|_2^t$$

$$\le O(t)^{t/2} \|\mu - \mu_j\|_2^t ,$$

which is sufficient to prove high-dim generalization of warmup lemma and its consequences.