CS 224 Fall 2023 Scribes: Sabarish Sainathan, Anna Trella
27 September 2023

Lecture 7: Sum-of-squares Application: Tensor Decomposition

1 Continuation of Lecture 6

Recall from last lecture we want to learn a mixture of Gaussians. [HL18, KS17]. We have d-dimensional samples

$$x_1, \dots, x_n \sim q, \quad \text{where} \quad q := \frac{1}{k} \sum_{j=1}^{k} \mathcal{N}(\mu_j, Id_d).$$

In order to learn the centers μ_j of these Gaussians, we designed a sum of square (SoS) program. See [Hop18] for a collection of blog posts on the SoS method. This program is designed to simulate the inefficient algorithm of brute-force constructing subsets S of appropriate size $N = n/k$ to find some subset of the data that "looks like" a Gaussian.

Set up our SoS program:

- **Variables:**

  $$a_1, \dots, a_n, \mu$$

  where $a_i = 1$ if we believe point i came from the component and 0 otherwise; μ is our final estimate for what the mean of that component is.

  We also define the following quantity:

  $$c_j := \frac{|S \cap S_j|}{N} = \frac{\sum_{i \in S_j} a_i}{N} \in [0, 1] \tag{1}$$

  this can be interpreted as the (normalized) *overlap* between S, the set of points we found, and S_j, the set of points in the jth component. Notice that $c_j = 1$ means we have learned the component perfectly.

- **Constraints:**
  - $a_i^2 = a_i$ (ensures a_is are just Boolean indicator variables)
  - $\sum_i a_i = N$ (ensures we select exactly $N = \frac{n}{k}$ points for the component estimate)

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

  Notice that the last constraint says, for the points I picked out, if I look at the projection in any direction u, the empirical tth moment of the data I picked out, should look like the empirical tth moment of a Gaussian (i.e., $\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,1)}[g^t] = (t-1)! \le t^{t/2}$), namely it should have the above bound)

- **Objective** (Max Entropy):

$$\min_{\tilde{\mathbb{E}}} \|\tilde{\mathbb{E}}[aa^T]\|_F$$

  we want to minimize over pseudo-distributions over solutions to this polynomial system where $a = (a_1, .., a_n)$.

Recall in the last lecture, under sum of squares, we showed that:

$$\sum_{j=1}^{k} c_j^2 \ge 1 - o(1)$$

In addition, we know that $\sum c_j = 1$. Using these the fact that $\sum c_j = 1$ and the sum of squares of c_js are close to 1, which implies one of the c_js is close to 1.

1.1 Motivation for Max Entropy

First Example: Let's pretend that $\widetilde{\mathbb{E}}$ is actually a real distribution — in fact, a deterministic distribution:

$$a_i = \mathbb{I}[x_i \text{ came from component } j] \tag{2}$$

Then

$$\widetilde{\mathbb{E}}[aa^\top] = aa^\top$$

For example, suppose $a = (1, 0, 1, 0, 0, 1)$, then:

$$aa^T = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

Since a can only pick out N points, this will be an $N \times N$ submatrix of 1's.

$$\implies \|\widetilde{\mathbb{E}}[aa^\top]\|_F^2 = N^2$$

Second Example: Pretend $\widetilde{\mathbb{E}}$ is actually a real distribution, but there is some randomness. Indeed, we'll sample $j \sim [k]$ and keep the definition of a_i above in Equation 2. Let $a_i^{(j)} = \mathbb{I}[x_i \text{ came from component } j]$.

$$\widetilde{\mathbb{E}}[aa^\top] = \mathbb{E}_j[a^{(j)}(a^{(j)})^\top] = \frac{1}{k}\Bigg[\underbrace{\begin{pmatrix} & \cdots & \end{pmatrix}}_{j=1} + \underbrace{\begin{pmatrix} & \cdots & \end{pmatrix}}_{j=2} + \cdots\Bigg]$$

This will result in a matrix with $N \times N$ blocks with only $1/k$ values, where this matrix tells us if any 2 points i, i' are in the same cluster.

After permutation, we obtain a block-diagonal matrix that shows incidence of points in each component. We get:

$$\|\widetilde{\mathbb{E}}[aa^\top]\|_F^2 = k \cdot N^2 \cdot \frac{1}{k^2} = \frac{N^2}{k}$$

1.2 Lemma to Motivate Algorithm

Lemma 1. *For $\widetilde{\mathbb{E}}$ optimizing the SoS program:*

$$\|\widetilde{\mathbb{E}}[aa^\top] - \mathbb{E}_j[a^{(j)}(a^{(j)})^\top]\|_F^2 \quad \textit{is small.}$$

Proof. Let $\tilde{M} = \widetilde{\mathbb{E}}[aa^\top]$ and $M = \mathbb{E}_j[a^{(j)}(a^{(j)})^\top]$

$$\begin{aligned}\|\tilde{M} - M\|_F^2 &= \|\widetilde{M}\|_F^2 + \|M\|_F^2 - 2\langle\widetilde{M}, M\rangle \\ &\le \|M\|_F^2 + \frac{N^2}{k} - 2\langle\widetilde{M}, M\rangle \\ &= \frac{2N^2}{k} - 2\langle\widetilde{M}, M\rangle\end{aligned}$$

Now notice,

$$\begin{aligned}\langle\widetilde{M}, M\rangle &= \langle\widetilde{\mathbb{E}}[aa^\top], \mathbb{E}_j[a^{(j)}(a^{(j)})^\top]\rangle \\ &= \widetilde{\mathbb{E}}\mathbb{E}_j[\langle a, a^{(j)}\rangle^2]\end{aligned}$$

and recall from Equation 1,

$$\langle a, a_j\rangle = \sum_{i \in S_j} a_i = N \cdot c_j$$

Plugging in $\mathbb{E}_j[\langle a, a^{(j)}\rangle^2] = \frac{1}{k}\sum_{j=1}^{k}(Nc_j)^2$:

$$\langle \widetilde{M}, M\rangle = \widetilde{\mathbb{E}}\frac{1}{k}\sum_{j=1}^{k}(Nc_j)^2$$

$$= \frac{N^2}{k}\widetilde{\mathbb{E}}\sum_{j=1}^{k}c_j^2 = \frac{N^2}{k}(1-o(1))$$

$$\implies \|\tilde{M} - M\|_F^2 \leq \frac{2N^2}{k}o(1) \quad \text{which is sufficiently small.}$$

□

1.3 Algorithm

1. Solve for $\widetilde{\mathbb{E}}$ (run optimization problem).
2. Compute $\widetilde{\mathbb{E}}[aa^\top]$.
3. Read off the clustering structure from $\widetilde{\mathbb{E}}[aa^\top]$ (i.e., which points are in the same cluster j).
4. Compute empirical means of the clusters we have found.

2 SoS for Tensor Decomposition

In the first unit of this class, we tackled the problem of decomposing a tensor into the sum of rank-1 tensors. We'll consider here the same problem, albeit when the tensor we have access to is a highly noisy version of the true tensor.

Indeed, we'll suppose the following setting. Let $u_1, \ldots, u_k \in \mathbb{R}^d$ be orthonormal vectors. We have access to a tensor

$$T = \underbrace{\sum_{i=1}^{k} u_i^{\otimes 3}}_{\text{signal}} + \underbrace{E}_{\text{noise}},$$

where E is a noise tensor.

2.1 Aside: How do we quantify the "size" of a tensor?

Consider the following two norms, analogous to the Frobenius and operator norms over linear operators:

- **The Frobenius norm:**
$$\|T\|_F = \sqrt{\sum_{i,j,k} T_{ijk}^2}$$

- **The injective tensor norm:**
$$\|T\|_{\text{inj}} = \max_{\|x\|=1} \langle T, x^{\otimes 3} \rangle$$

  for symmetric tensor T. Note that there is no need for an absolute value within the maximum, since we may just as well take x to be $-x$. However, this norm is NP-hard to compute — even approximately to within a factor of $n^{o(1)}$.

There exists a convenient relationship between these two norms:

Lemma 2. $\|T\|_F \geq \|T\|_{inj}$

Proof. We'll employ a degree-6 SoS proof (since the terms involve polynomials of degree at most 6). For any $x \in \mathbb{R}^d$,

$$\begin{aligned}
\langle T, x^{\otimes 3} \rangle^2 &= \left(\sum_{i,j,k} T_{ijk} x_i x_j x_k \right)^2 \\
&\leq \left(\sum_{i,j,k} T_{ijk}^2 \right) \cdot \left(\sum_{i,j,k} x_i^2 x_j^2 x_k^2 \right) \quad \text{(Cauchy-Schwarz)} \\
&= \left(\sum_{i,j,k} T_{ijk}^2 \right) \cdot 1 = \|T\|_F^2
\end{aligned}$$

Taking the supremum over the left-hand side shows that $\|T\|_{\text{inj}}^2 \leq \|T\|_F^2$. □

Consider a third, computationally tractable norm that interpolates between the Frobenius and injective tensor norms:

- **The SoS norm:**
$$\|T\|_{SoS_t} = \max_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}[\langle T, x^{\otimes 3} \rangle] \quad \text{for } t \geq 6 \text{ even}$$

where $\widetilde{\mathbb{E}}$ ranges over deg-t pseudo-expectations over the variable x satisfying $\|x\|^2 = 1$. Notice this norm is computationally tractable via a $d^{O(t)}$ algorithm — namely, by setting up a semi-definite program and running the ellipsoid method.

Claim 1. *As the degree t approaches infinity along the even numbers, the degree-t SoS norm is monotonically decreasing and approaches the injective tensor norm:*

$$\|T\|_{SoS_t} \searrow \|T\|_{inj}.$$

Additionally, the Frobenius norm is at least the degree-6 SoS norm:

$$\|T\|_F \geq \|T\|_{SoS_6}.$$

2.2 Q: How big does the noise E have to be before Jennrich's breaks?

When the scale of the noise is on the order of $1/d^c$ or smaller for sufficiently large c, Jennrich's algorithm succeeds. If c is too small, however, the scale of the noise dominates that of the signal, and Jennrich's cannot discern between the two.

Indeed, consider the setting in which

$$E_{ijk} \sim \mathcal{N}\left(0, d^{-2+\epsilon}\right),$$

for any $\epsilon > 0$. Here, $c = 1 - \epsilon/2 < 1$.

Recall that in Jennrich's algorithm, we consider a pair of contractions of the tensor T of the following sort:

- Sample $g \sim \mathcal{N}(0, I_d)$
- Contract to get matrix

$$M_g = T(g, :, :) = \sum_i \langle g, u_i \rangle u_i u_i^\top + E(g, :, :)$$

Every entry of $E(g, :, :) = \sum_{k=1}^d g_k E_{k::}$ is of the form

$$E(g, :, :)_{ij} = \sum_k g_k E_{kij}.$$

Conditioning on E, every entry has distribution $\mathcal{N}(0, \sum_k E_{kij}^2)$. The variance term concentrates to $d/d^{2-\epsilon} = d^{-1+\epsilon}$ with sufficiently large dimension d, so we make the approximation that $E(g,:,:)_{ij} \sim \mathcal{N}(0, d^{-1+\epsilon})$. Using the fact that with high probability

$$\|G\|_{\text{op}} \approx \sqrt{d}$$

where G is a random matrix with i.i.d. standard normal random variables, we obtain the high-probability lower bound

$$\|E(g,:,:)\|_{\text{op}} \approx \sqrt{d} \cdot \sqrt{\frac{1}{d^{1-\epsilon}}} = d^{\epsilon/2} >> 1.$$

In contrast, the scale of the signal is $O(1)$. Indeed, we have that $\langle g, u_k \rangle \sim \mathcal{N}(0,1)$ has scale 1 for each k. Since the matrices $u_k u_k^\top$ are mutually orthogonal,

$$\| \sum_k \langle g, u_k \rangle u_k u_k^\top \|_{\text{op}} = \max_{k \in [d]} |\langle g, u_k \rangle| \approx 2\sqrt{\log d}.$$

For sufficiently large d, it follows that the scale of the noise of M_g far exceeds the scale of the $O(\sqrt{\log d})$ signal.

Claim 2. *With high probability,*

$$\|\mathcal{N}(0, \sigma^2)^{d \times d \times d}\|_{SoS_6} \leq \sigma d^{3/4} \cdot \textit{polylog}(d)$$

In the noise setting defined above with $\epsilon = 0.1$, we use this claim to obtain $\|E\|_{SoS_6} \lesssim d^{-0.2}$ « 1. We will show that as long as $\|E\|_{SoS_6} << 1$, there is an algorithm to recover $u_1, \ldots, u_d$.

2.3 Problem Setup

Let's make this claim more concrete. Again, we have access to

$$T = \sum_{i=1}^{d} u_i^{\otimes 3} + E,$$

where the u_i's are orthonormal and the size of the norm is at most $\|E\|_{SoS_6} = o(1)$ (norm is vanishing).

Consider the polynomial

$$p_\ell(x) = \sum_{i=1}^{d} \langle x, u_i \rangle^\ell.$$

Goal: Our objective will be to maximize p_3.

2.4 Algorithm Attempt 1

This algorithm does not work, but introduces helpful ideas.

Define our SoS program to be:

- **Variables:**
$$x$$
- **Constraints:**
$$\|x\|^2 = 1$$
- **Objective:**
$$\max_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$$

Lemma 3.
$$\widetilde{\mathbb{E}}[p_3(x)] \geq 1 - o(1).$$

Proof. Consider the real distribution uniform over $\{u_1, \dots, u_d\}$, and call it $\mathbb{E}[\cdot]$. This is the ideal distribution because if we have acess to $\mathbb{E}[\cdot]$, we can just sample from this distribution to obtain the factors $\{u_1, \dots, u_d\}$.

$$\begin{aligned}
\mathbb{E}[p_3(x)] &= \mathbb{E}_{j\sim[d]}\left[\sum_i \langle u_i, x\rangle^3\right] \\
&= \frac{1}{d}\sum_{j=1}^{d}\left(\sum_i \langle u_i, u_j\rangle^3\right) \\
&= \frac{1}{d}\cdot d = 1.
\end{aligned}$$

This makes sense because the true components should be able to maximize p_3.

Now looking at our noisy objective:

$$\begin{aligned}
\widetilde{\mathbb{E}}[\langle T, x^{\otimes 3}\rangle] &\geq \mathbb{E}[\langle T, x^{\otimes 3}\rangle] \\
&= \mathbb{E}[p_3(x)] + \mathbb{E}[\langle E, x^{\otimes 3}\rangle] \geq 1 - o(1),
\end{aligned}$$

where we have used that
$$\|E\| = o(1).$$

Now,
$$\implies \tilde{\mathbb{E}}[p_3(x)] = \tilde{\mathbb{E}}[\langle T, x^{\otimes 3}\rangle] - \tilde{\mathbb{E}}[\langle E, x^{\otimes 3}\rangle] \leq 1 - o(1)$$

□

Lemma 4. *The optimal choice of* $\widetilde{\mathbb{E}}$ *also satisfies* $\widetilde{\mathbb{E}}[p_4(x)] \geq 1 - o(1)$.

Proof.

$$\begin{aligned}
1 - o(1) &\leq \widetilde{\mathbb{E}}[p_3(x)]^2 \\
&\leq \widetilde{\mathbb{E}}[p_3(x)^2] \quad \text{(using "pseudo-expectation Cauchy-Schwarz.")} \\
&= \widetilde{\mathbb{E}}\left[\left(\sum_i \langle u_i, x\rangle^3\right)^2\right]
\end{aligned}$$

In the second line, "pseudo-expectation Cauchy-Schwarz." is $\widetilde{\mathbb{E}}[p \cdot q] \leq \widetilde{\mathbb{E}}[p^2]^{1/2}\widetilde{\mathbb{E}}[q^2]^{1/2}$.

Now focusing on $\left(\sum_i \langle u_i, x\rangle^3\right)^2$, in degree-6 SoS:

$$\left(\sum_i \langle u_i, x\rangle^3\right)^2 = \left(\sum_i \langle u_i, x\rangle \cdot \langle u_i, x\rangle^2\right)^2$$

$$\leq \left(\sum_i \langle u_i, x\rangle^2\right) \cdot \left(\sum_i \langle u_i, x\rangle^4\right) \quad \text{(using Cauchy-Schwarz)}$$

$$= 1 \cdot \sum_i \langle u_i, x\rangle^4 = p_4(x)$$

□

2.5 How do we "round" the pseudo-distribution to a solution?

Idea 1: What if we used Jennrich's on the degree-3 object $\widetilde{\mathbb{E}}[x^{\otimes 3}]$?

There's an issue here. Suppose $\widetilde{\mathbb{E}}$ is an actual distribution that places $\frac{1}{\sqrt{d}}$ mass on an asbitrary unit vector $w \perp u_1, \ldots, u_d$ and $\frac{1-1/\sqrt{d}}{d} \sim 1/d$ mass on each of $u_1, \ldots, u_d$.

$$\widetilde{T} = \widetilde{\mathbb{E}}[x^{\otimes 3}] = \frac{1 - 1/\sqrt{d}}{d} \sum_i u_i^{\otimes 3} + \frac{1}{\sqrt{d}} w^{\otimes 3}$$

has contraction

$$\widetilde{T}(g, :, :) = \frac{1 - 1/\sqrt{d}}{d} \sum_i \langle g, u_i\rangle u_i^{\otimes 2} + \frac{1}{\sqrt{d}} \langle g, w\rangle w^{\otimes 2}$$

The second term outweighs the first (important) one by a ratio of $\sqrt{d}$, which means the eigenvectors of $\widetilde{T} = \widetilde{\mathbb{E}}[x^{\otimes 3}]$ are useless!

In a deeper sense, this distribution is problematic because it's a very "low-entropy" distribution — lots of weight placed on one vector w. If we had something closer to the uniform distribution, this might work.

2.6 Algorithm Attempt 2

This time, the algorithm will work.

Consider the same SoS program as before in Section 2.4, but

$$\max_{\widetilde{\mathbb{E}}} \widetilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$$

over degree-6 pseudo-expectations $\widetilde{\mathbb{E}}$ which:

1. Satisfy program constraints — that is, $\|x\|^2 = 1$.
2. Have the max-entropy properties

$$\|\widetilde{\mathbb{E}}[xx^\top]\|_{\text{op}} \leq \frac{1}{d}$$
$$\|\widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top]\|_{\text{op}} \leq \frac{1}{d}$$

Indeed, note that if $\widetilde{\mathbb{E}}$ is uniform over the directions $u_1, \ldots, u_d$, then

$$\widetilde{\mathbb{E}}[xx^\top] = \frac{1}{d}\sum_i u_i u_i^\top = \frac{1}{d} \cdot I_d.$$

Claim: If we run Jennrich's on $\widetilde{\mathbb{E}}[x^{\otimes}3]$ for pseudo-distributions optimizing this new program, we will recover almost all the components.

Lemma 5. *For this optimal $\widetilde{\mathbb{E}}$, for $1 - o(1)$ fraction of $i \in [d]$,*

$$\widetilde{\mathbb{E}}[\langle u_i, x\rangle^4] \geq \frac{1}{d} \cdot (1 - o(1)).$$

Proof. Suppose for sake of contradiction that for $\delta = \Omega(1)$ fraction of i's we have

$$\widetilde{\mathbb{E}}\langle u_i, x\rangle^4 \leq \frac{1-\delta}{d}.$$

By averaging, for some other i, we have

$$\widetilde{\mathbb{E}}\langle u_i, x\rangle^4 > \frac{1}{d}$$

However, notice that:

$$\widetilde{\mathbb{E}}\langle u_i, x\rangle^4 = (u_i \otimes u_i)^\top \widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top](u_i \otimes u_i)$$
$$\implies \|(u_i \otimes u_i)^\top \widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top](u_i \otimes u_i)\|_{op}$$
$$\leq \|(u_i \otimes u_i)^\top\|_{op} \|\widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top]\|_{op} \|(u_i \otimes u_i)\|_{op}$$
$$= \|\widetilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^\top]\|_{op} \leq \frac{1}{d} \quad \text{(by the high entropy constraint)}$$

And we have a contradiction on $\widetilde{\mathbb{E}}\langle u_i, x\rangle^4$. □

References

[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.

[Hop18] Samuel B Hopkins. Clustering and sum of squares proofs: Six blog posts on unsupervised learning, 2018. https://www.samuelbhopkins.com/clustering.pdf.

[KS17] Pravesh K. Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.