CS 224 Fall 2023 Scribes: Mary Letey & Alvaro Ribot
October 2nd 2023

Lecture 8: SoS for Tensor Decomposition (Conclusion)

Overview

- Recap SoS for noisy orthogonal tensor decomposition.
- Extension of noisy-orthogonal algorithm to over-complete setting for tensor decomposition.
- Overview of last few lectures on Sum-of-Squares methods.

Noisy Orthogonal Tensor Decomposition

We return to the problem of computing the orthogonal components of an order-3 tensor subject to some noise. As we shall see, for noise "small enough", this can be done using a SoS-based algorithm. Suppose we have a tensor of the form

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

for orthonormal $u_1, \ldots, u_k$ and error tensor E. Recall from Lecture 7 that for "small enough" error E, namely $\|E\|_{\text{SOS6}} = o(1)$, then we can recover the components u_i.

```
Algorithm 1: NOISY ORTHOGONAL TENSOR DECOMPOSITION
  Input: Noisy tensor $T$ as above with orthogonal components $u_i$.
  Output: Components $u_i$ of $T$.
  Maximise $\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$ over degree-6 pseudo expectations subject to the
    **high-entropy constraints** for $\|x\| = 1$.
  Let $\tilde{T} = \tilde{\mathbb{E}}[x^{\otimes 4}]$.
  Apply JENNRICH'S to $\tilde{T}$ repeatedly, i.e. computing the top eigenvector of
    $\tilde{T}(g, :, :) = \tilde{\mathbb{E}}[\langle g, x \otimes x\rangle x x^T]$ for poly($d$) many Gaussian samples
    $g \sim \mathcal{N}(0, \mathbb{I}_d)$.
```

The main result of today's lecture will be to prove the correctness of Algorithm 1 for this task. We will rely on the previously defined notion of **high-entropy** constraints of a pseudodistribution.

Definition 1. $\tilde{\mathbb{E}}$ *has* ***high entropy*** *if*

$$\left\|\tilde{\mathbb{E}}[xx^T]\right\|_{op} \leq 1/d$$
$$\left\|\tilde{\mathbb{E}}[(x \otimes x)(x \otimes x)^T]\right\|_{op} \leq 1/d$$
$$\left\|\tilde{\mathbb{E}}[(x \otimes x \otimes x)(x \otimes x \otimes x)^T]\right\|_{op} \leq 1/d$$

This makes sense since recall that the Frobenius norm of $\tilde{\mathbb{E}}$ is inversely related to entropy.

Why does STEP 3 in Algorithm 1 work? This will be proved by the following Theorem 1:

Theorem 1. *For any optimal pseudoexpectation $\tilde{\mathbb{E}}$ and some vector $a \in \mathbb{R}^d$ highly correlated with $\tilde{\mathbb{E}}$ (i.e. $\tilde{\mathbb{E}}[\langle a, x\rangle^4] \geq \frac{1}{d}(1 - o(1))$, then with probability at least 1/poly(d) the top eigenvector v of $\tilde{\mathbb{E}}[\langle g, x \otimes x\rangle xx^T]$ satisfies $\langle a, v\rangle^2 \geq 0.99$ for random $g \approx \mathcal{N}(0, \mathbb{I}_d)$.*

Remark Note that the optimality of $\tilde{\mathbb{E}}$ will be satisfied by STEP 1 of the algorithm. Further, the importance of this theorem is that we can then successively sample g and have a good probability of finding the components u_i of the tensor T regardless of error, so long as the optimised pseudoexpectation is highly correlated with the target components u_i of our tensor. Why should this be true?

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

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

We are now in a position to prove Theorem 1.

Proof. *Setup* Recall we set up $a \in \mathbb{S}^{d-1}$ with $\tilde{\mathbb{E}}[\langle a, x\rangle^4] \geq (1 - o(1))/d$ and $g \sim \mathcal{N}(0, \mathbb{I}_d)$, and we aim to better understand and bound the moment

$$M_g \equiv \tilde{\mathbb{E}}[\langle g, x \otimes x\rangle xx^T].$$

Decompose g as

$$g = \gamma a \otimes a + \gamma^{\perp}$$

for one-dimensional Gaussian $\gamma \sim \mathcal{N}(0, \mathbb{I})$. Note that this forces $\gamma^{\perp}$ to be distributed according to $\sim \mathcal{N}(0, \mathbb{I} - (a \otimes a)(a \otimes a)^T)$.

Big Picture We see now that

$$M_g = \gamma M_{a \otimes a} + M_{\gamma^{\perp}}.$$

We can think of γ as a gaussian in the direction spanned by $a \otimes a$ and so $M_{\gamma^\perp}$ should not contain too much information relating to a. Said another way, we can think of $\gamma M_{a \otimes a}$ as a **signal moment**, which we wish to bound by aa^T/d; and $M_{\gamma^\perp}$ as a **noise moment**, which we wish to bound by $\approx \sqrt{\log d}/d$. As long as γ is reasonably larger than $\sqrt{\log d}$, then the signal moment will dominate over the noise term. In this case, we would then have $M_g \approx aa^T/d$, giving precisely a as the top eigenvector, as desired. How often is γ is large enough, i.e. what is $\mathbb{P}(|\gamma| > \sqrt{\log d})$? This can be bounded approximately by some 1/poly(d) expression, giving us the guarantee of the theorem.

Signal Bound Recall our claim is that $M_{a \otimes a} = aa^T/d$ + some term with $o(1/d)$ operator norm. Thus analyse the components of $M_{a \otimes a}$, i.e. $v^T M_{a \otimes a} v$ for $\|v\| = 1$ gives projection of $M_{a \otimes a}$ along v. We have

$$\begin{aligned} a^T M_{a \otimes a} a &= a^T \tilde{\mathbb{E}}[\langle a \otimes a, x \otimes x \rangle x x^T] a \\ &= \tilde{\mathbb{E}}[\langle a, x \rangle^4] \\ &\geq \frac{1}{d}(1 - o(1)) \end{aligned}$$

by high correlation condition. To consider the non-a components of $M_{a \otimes a}$, take $b \in \mathbb{S}^{d-1}$ such that $b \perp a$ and consider

$$\begin{aligned} b^T M_{a \otimes a} b &= \tilde{\mathbb{E}}[\langle a \otimes a, x \otimes x \rangle b^T x x^T b] \\ &= \tilde{\mathbb{E}}[\langle a, x \rangle^2 \langle b, x \rangle^2] \\ &\leq \tilde{\mathbb{E}}[\langle a, x \rangle^2 (1 - \langle a, x \rangle)^2] \quad \text{since } a, b \text{ orthonormal and } \|x\| = 1 \text{ so } \langle a, x \rangle^2 + \langle b, x \rangle^2 \leq 1 \\ &= \tilde{\mathbb{E}}[\langle a, x \rangle^2] - \tilde{\mathbb{E}}[\langle a, x \rangle^4] \\ &\leq a^T \tilde{\mathbb{E}}[x x^T] a - \frac{1}{d}(1 - o(1)) \quad \text{by the high correlation condition} \\ &\leq \left\| \tilde{\mathbb{E}}[x x^T] \right\|_{\text{op}} - \frac{1}{d}(1 - o(1)) \\ &\leq o(1/d) \quad \text{by the high entropy condition on } \left\| \tilde{\mathbb{E}}[x x^T] \right\|_{\text{op}}. \end{aligned}$$

This analysis is not technically complete has we have not computed $a^T M_{a \otimes a} b$ or $b^T M_{a \otimes a} a$ terms, but we can be sure these are small since $M_{a \otimes a}$ is positive definite, so terms of the form $a^T M_{a \otimes a} b$ or $b^T M_{a \otimes a} a$ must be bounded by $\sqrt{(a^T M a)(b^T M b)} \leq \sqrt{(1/d) o(1/d)}$ since $a^T M a \leq 1/d$. $sqrt(1/d) o(1/d)$ is negligible compared to $1/d$ and so $a^T M a$ is still dominant.

Noise Bound $\gamma^\perp$ has a weird covariance (as $\gamma^\perp \sim \mathcal{N}(0, \mathbb{I} - (a\otimes a)(a \otimes a)^T)$. We can make this nicer to work with if we take $g' \sim \mathcal{N}(0, (a\otimes a)(a\otimes a)^T)$ and

$$\begin{aligned} g_1 &= \gamma^\perp - g' \\ g_2 &= \gamma^\perp + g' \end{aligned}$$

so that $\gamma^\perp = (g_1 + g_2)/2$ with $g_1, g_2 \sim \mathcal{N}(0, \mathbb{I})$ (although g_1, g_2 are very much not independent). Then

$$M_{g^\perp} = \frac{1}{2} M_{g_1} + \frac{1}{2} M_{g_2}.$$

Recall we wish to show that $\left\|M_{\gamma^\perp}\right\|_{op} \leq \sqrt{\log d}/d$ so for $h \sim \mathcal{N}(0, \mathbb{I})$ it suffices to show that $\left\|\tilde{\mathbb{E}}[\langle h, x \otimes x\rangle x x^T]\right\|_{op} \leq \sqrt{\log d}/d$.

$$\begin{aligned} \|M_h\|_{\text{op}} &= \left\| \sum_{ij} h_{ij} A_{ij} \right\|_{\text{op}} \qquad \text{for } A_{ij} = \tilde{\mathbb{E}}[x_i x_j x x^T] \text{ as } \langle h, x \otimes x \rangle = \sum_{ij} h_{ij} x_i x_j \\ &\leq \left\| \sum_{ij} A_{ij}^2 \right\|_{\text{op}}^{1/2} \sqrt{\log d} \qquad \text{concentration inequality [Tro15, Thm. 4.1.1]} \\ &= \left\| B B^T \right\|_{\text{op}}^{1/2} \sqrt{\log d} \end{aligned}$$

where we define $B \in \mathbb{R}^{d \cdot d^3}$ by $B_{ijk\ell} = \tilde{\mathbb{E}}[x_i x_j x_k x_\ell]$, so that

$$\begin{aligned} \left(\sum_{ij} A_{ij}^2 \right)_{ab} &= \sum_c \sum_{ij} \tilde{\mathbb{E}}[x_i x_j x_a x_c] \tilde{\mathbb{E}}[x_i x_j x_c x_b] \\ &= \sum_c \sum_{ij} B_{aijc} B_{bijc} \\ &\equiv (BB^T)_{ab}. \end{aligned}$$

Now we can bound $\left\|BB^T\right\|_{\text{op}}^{1/2} = \|B\|_{\text{op}}$ as

$$\begin{aligned} \|B\|_{\text{op}} &= \sup_{z \in \mathbb{S}^d, z' \in \mathbb{S}^{d^3 - 1}} \left| z^T B z' \right| \\ &= \sup \tilde{\mathbb{E}}[\langle z, x \rangle \langle z', x^{\otimes 3} \rangle] \\ &\leq \tilde{\mathbb{E}}[\langle z, x \rangle^2]^{1/2} \tilde{\mathbb{E}}[\langle z, x \rangle^6]^{1/2} \qquad \text{by Cauchy-Schwartz} \\ &= \left| z^T \tilde{\mathbb{E}}[x x^T] z \right|^{1/2} \left| z' \tilde{\mathbb{E}}[x^{\otimes 3} (x^{\otimes 3})^T] z^T \right|^1 /2 \\ &\leq \frac{1}{\sqrt{d}} \frac{1}{\sqrt{d}} \qquad \text{by the high-entropy condition.} \end{aligned}$$

Finally we are done. □

Overcomplete Tensors

Now we see how this idea can also be used for the case where we wish to decompose a noiseless, overcomplete ($k >> d$) tensor

$$T = \sum_{i=1}^{k} u_i^{\otimes 3}$$

for $u_1, \cdots, u_k \in \mathbb{S}^{d-1}$ *so long as we have* $k << d^{3/2}$.

Remark *We can notice that a high-level takeaway from the above proof of Theorem 1 is that it relied on the compatibility of the high-entropy condition and the high-correlation condition, i.e. in the noisy orthogonal case, we have*

$$\left\| \tilde{\mathbb{E}}[xx^T] \right\|_{op} \le \frac{1}{d} \quad \textit{high entropy}$$

$$\tilde{\mathbb{E}}[\langle a, x\rangle^4] \ge (1 - o(1))\frac{1}{d} \quad \textit{high correlation}$$

In the previous section, it was important that both inequalities depend on $1/d$*. Recall that we needed to bound* $b^\top M_{a\otimes a} b$ *for* $b \perp a$ *and we got:*

$$b^\top M_{a\otimes a} b \le \left\| \tilde{\mathbb{E}}[xx^T] \right\|_{op} - \tilde{\mathbb{E}}[\langle a, x\rangle^4].$$

We shall see that this breaks for overcomplete tensors as follows.

Consider a pseudodistribution $\tilde{\mathbb{E}}$ uniform over $\{u_1, \cdots, u_k\}$ and take $a = u_i$. Then we have correlation

$$\tilde{\mathbb{E}}[\langle a, x\rangle^4] = \frac{1}{k}\sum_{j=1}^{k} \langle u_i, u_j\rangle^4 = \frac{1}{k}\left(1 + \sum_{j:j\neq i} \langle u_i, u_j\rangle^4\right) \approx \frac{1}{k}(1 + O(k/d^2)) = \frac{1}{k}(1 + o(1))$$

where we have used that, for $i \neq j$, $\langle u_i, u_j\rangle^4 \approx (1/\sqrt{d})^4 = 1/d^2$. On the other hand, we have entropy

$$\tilde{\mathbb{E}}[xx^T] = \frac{1}{k}\sum_{j=1}^{k} u_i u_i^\top \approx \frac{1}{d}\mathbb{I}_d$$

since, for $k >> d$, we are effectively computing a sample variance of a uniform distribution on $\mathbb{S}^{d-1}$. Therefore,

$$\left\| \tilde{\mathbb{E}}[xx^T] \right\|_{\text{op}} \approx \frac{1}{d} >> \frac{1}{k}$$

Intuitively, the issue here is that there are too few dimensions, so the u_i's have tiny but non-negligible correlations. This implies that $\left\|\tilde{\mathbb{E}}[xx^T]\right\|_{\text{op}}$ cannot be small enough for the previous argument to go through.

Big Picture Lift to higher dimensions. Suppose that instead of $T = \sum_{i=1}^{k} u_i^{\otimes 3}$ we had the tensor

$$T' = \sum_{i=1}^{k} u_i^{\otimes 6} = \sum_{i=1}^{k} \left(u_i^{\otimes 2}\right)^{\otimes 3}.$$

Now we have that $\{w_i \triangleq u_i^{\otimes 2}\}_{i=1,\ldots,k}$ are $k << d^{\frac{3}{2}}$ "random" vectors in d^2 dimensions. The idea is that we would like to use Algorithm 1 replacing x everywhere with $x \otimes x$. But there are some details that should be taken into consideration.

Entropy To be more precise, it turns out that $\sum_i w_i w_i^\top$ and $\mathbb{I}_{d^2}$ may differ significantly for some bad directions. However, we can choose an appropriate projection that deals with this problem. Let Π be the projection operator to the symmetric subspace spanned by the flattening of tensors of the form $u^{\otimes 2}$, see [HSSS16, Section C.0.4] for more details. If we use $w_i = \Pi(u_i^{\otimes 2})$ we obtain the following result.

Claim 1. *If $k << d^2$, then*

$$\left\|\frac{1}{k}\sum_i w_i w_i^\top\right\|_{op} = \frac{1}{k}(1 + o(1))$$

The proof is again based on a concentration inequality, that needs some notion of incoherence. The idea is that, as long as d is sufficiently large, this incoherence parameter is bounded and we can apply the concentration inequality. Intuitively this relies on the fact that for a "small" number of vectors sampled from a high-dimensional space, they are effectively orthogonal, i.e. their inner products can be bounded. For more details, see [Ver18, Theorem 5.62].

Correlation The only thing that remains to be proved is the correlation lower bound

Claim 2. *For optimal $\tilde{\mathbb{E}}$, with high probability over the u_i's,*

$$\tilde{\mathbb{E}}\left[\langle w_i, x^{\otimes 2}\rangle^4\right] \geq 1 - o(1)$$

so, by the averaging argument from last time, for $1 - o(1)$ fraction of the i's we have

$$\tilde{\mathbb{E}}\langle w_i, x^{\otimes 2}\rangle \geq \frac{1}{k}(1 - o(1)).$$

Note Have

$$\tilde{\mathbb{E}}\left[\langle w_i, x^{\otimes 2}\rangle^4\right] = \tilde{\mathbb{E}}\left[\langle u_i, x\rangle^8\right].$$

We will prove a "baby version" of Claim 2 for intuition, as follows.

Claim 3.

$$\tilde{\mathbb{E}}\left[\langle u_i, x\rangle^4\right] = 1 - o(1).$$

Proof. *Setup* We will give a low-degree SoS proof starting from $\sum_i \langle u_i, x\rangle^3 \geq 1 - o(1)$. Using Cauchy-Schwarz we have

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

Note that $\|x\| = 1$ and, using that $\|u_i\| = 1$, we obtain

$$\begin{aligned}\left\| \sum_i \langle u_i, x\rangle^2 u_i \right\|^2 &= \sum_i \langle u_i, x\rangle^4 \|u_i\|^2 + \sum_{i \neq j} \langle u_i, x\rangle^2 \langle u_j, x\rangle^2 \langle u_i, u_j\rangle \\ &= \sum_i \langle u_i, x\rangle^4 + \sum_{i \neq j} \langle u_i, x\rangle^2 \langle u_j, x\rangle^2 \langle u_i, u_j\rangle\end{aligned}$$

Therefore,

$$\sum_i \langle u_i, x\rangle^4 \geq \sum_i \langle u_i, x\rangle^3 - \sum_{i \neq j} \langle u_i, x\rangle^2 \langle u_j, x\rangle^2 \langle u_i, u_j\rangle.$$

So it suffices to bound

$$\left\| \sum_{i \neq j} \langle u_i, u_j\rangle (u_i \otimes u_j)(u_i \otimes u_j)^\top \right\|_{\text{op}} = o(1).$$

A Lazy Bound Let $M \in \mathbb{R}^{d^2 \times \binom{k}{2}}$ have columns consisting of $u_i \otimes u_j$'s. We can bound its operator norm and obtain that

$$\left\| \sum_{i \neq j} \langle u_i, u_j\rangle (u_i \otimes u_j)(u_i \otimes u_j)^\top \right\|_{\text{op}} = \left\| M \text{diag}(\{\langle u_i, u_j\rangle\}_{i \neq j}) M^\top \right\|_{\text{op}} \leq \frac{k^2}{d^{5/2}}$$

which is $o(1)$ if $k << d^{5/4}$. □

Remark *Nevertheless, PSET2 Question #3, gives a proof of the "right" bound, i.e. getting* $k << d^{3/2}$.

We are now finally in a position to construct an algorithm for overcomplete tensor decomposition given the results above. As mentioned before, the idea is to use Algorithm 1 replacing x by $\Pi(x \otimes x)$ in the high entropy constraints and run Jenrich's algorithm many times on

$$\tilde{T} \triangleq \tilde{\mathbb{E}}\left[\Pi(x \otimes x)^{\otimes 4}\right].$$

See Algorithm 2 for the details.

```
Algorithm 2: SOS FOR TENSOR LIFTING
  Input: Overcomplete tensor $T$ as above with $k$ components $u_i \in \mathbb{S}^{d-1}$ for
         $d << k << d^{3/2}$.
  Output: Components $u_i$ of $T$.
  Let $L(u) = \Pi(u \otimes u)$ be the lifting map. Connect this with previous notation
    $w_i = L(u_i)$ used above.
  Solve $\max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}[\langle T, x^{\otimes 3} \rangle]$ over degree-12 pseudoexpectations $\tilde{\mathbb{E}}$ such that
    • $\tilde{\mathbb{E}}$ satisfies $\|x\|^2 = 1$.
    • $\left\| \tilde{\mathbb{E}}[L(x)L(x)^T] \right\|_{\text{op}} \le \frac{1}{k}(1 + o(1))$.
    • $\left\| \tilde{\mathbb{E}}[(L(x) \otimes L(x))(L(x) \otimes L(x))^T] \right\|_{\text{op}} \le \frac{1}{k}(1 + o(1))$.
    • $\left\| \tilde{\mathbb{E}}[(L(x)^{\otimes 3})(L(x)^{\otimes 3})^T] \right\|_{\text{op}} \le \frac{1}{k}(1 + o(1))$.
  Define lifted tensor $\tilde{T} = \tilde{\mathbb{E}}[L(x)^{\otimes 4}]$
  Run JENNRICH'S to $\tilde{T}$ repeatedly, i.e. computing the top eigenvector of
    $\tilde{T}(g, :, :) = \tilde{\mathbb{E}}[\langle g, L(x) \otimes L(x) \rangle L(x) L(x)^T]$ for poly($d^2$) many Gaussian
    samples $g \in \mathbb{R}^{d^4}$. This step gives $\approx k$ vectors $L(x) \in \mathbb{R}^{d^2}$.
  Reshape $L(x)$ solutions above from 1x$d^2$ vector to $d$x$d$ matrix and take top
    eigenvector to give $u_i$.
```

SoS Retrospective

We have seen applications of SoS to

- **Robust mean estimation**
- **Robust regression** [KKM20]
- Learning **mixtures of Gaussians** down to the optimal separation threshold [KS17, HL17]
  - More efficient algorithm that only implicitly works with moment tensor [LL21]
- **Tensor decomposition** with maximal noise [GM17, MSS16, SS17]
- **Overcomplete** tensor decomposition with (conjectured) optimal number of components [GM17, MSS16, HSSS16, DdL$^+$22]
  - Fast spectral algorithms "inspired" by SoS [HSSS16, DdL$^+$22]

Takeaway - signatures of tractibility: For robustness or clusters, things are uniquely identified in moment bounds through sum-of-squares. For tensor algorithms, with non-convex optimisation landscape benign without noise, then controlled noise can't mess things up too much according to SoS.

Next Steps SoS algorithms are provably correct but are "first-pass" alghorithms, i.e. are often not practical or computationally tractable. We will see further examples of more efficient algorithms for the above tasks in later lectures. However, it is good to keep in mind the theoretical takeaways and signatures of tractibility as intuition that we gained from SoS techniques.

References

[DdL$^+$22] Jingqiu Ding, Tommaso d'Orsi, Chih-Hung Liu, Stefan Tiegel, and David Steurer. Fast algorithm for overcomplete order-3 tensor decomposition, 2022.

[GM17] Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions, 2017.

[HL17] Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs, 2017.

[HSSS16] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors, 2016.

[KKM20] Adam Klivans, Pravesh K. Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression, 2020.

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

[LL21] Jerry Li and Allen Liu. Clustering mixtures with almost optimal separation in polynomial time, 2021.

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares, 2016.

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning, 2017.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities, 2015.

[Ver18] Roman Vershynin. *High-Dimensional Probability*. Cambrdige University Press, 2018.