Robust polynomial time tensor decomposition

Sam Hopkins

December 19, 2017

1 Introduction

Tensor decomposition has recently become an invaluable algorithmic primitive. It has seen much use in new algorithms with provable guarantees for fundamental statistics and machine learning problems. In these settings, some low-rank k-tensor $A = \sum_{i=1}^{r} a_i^{\otimes k}$ which we would like to decompose into components $a_1, \ldots, a_r \in \mathbb{R}^n$ is often not directly accessible. This could happen for many reasons; a common one is that $A = \mathbb{E}\, X^{\otimes k}$ for some random variable X, and estimating A to high precision may require too many independent samples from X.

In this lecture we will dig in to algorithms for robust tensor decomposition—that is, how to accomplish tensor decomposition efficiently in the presence of errors.

We will focus on *orthogonal tensor decomposition* where components $a_1, \ldots, a_r \in \mathbb{R}^n$ of the tensor $A = \sum_{i=1}^{r} a_i^{\otimes k}$ to be decomposed are orthogonal unit vectors. Tensor decomposition is already both algorithmically nontrivial and quite useful in this setting—the orthogonal setting is good enough to give the best known algorithms for Gaussian mixtures, some kinds of dictionary learning, and the stochastic blockmodel. As we saw before, via *whitening* if the covariance matrix $\sum_{i=1}^{n} a_i a_i^\top$ is known for non-orthogonl but linearly independent vectors $a_1, \ldots, a_n$ then decomposing the tensor $\sum_{i=1}^{n} a_i^{\otimes 3}$ reduces to orthogonal decomposition.

2 Jenrich's algorithm for orthogonal tensor decomposition

The algorithm.

> **Input**: $A = \sum_{i=1}^{r} a_i^{\otimes 3}$ for orthogonal unit vectors $a_1, \ldots, a_r \in \mathbb{R}^n$
>
> **Algorithm**: sample $g \sim \mathcal{N}(0, I)$ and compute the contraction $M = \sum_{i=1}^{r} \langle g, a_i \rangle a_i a_i^\top$. Output the top eigenvector of M.
>
> **Analysis**: clearly the top eigenvector is a_i for $i = \operatorname{argmax} \langle a_i, g \rangle$. By symmetry, each vector a_i is equally likely to be the output of the algorithm, so running the algorithm $n \log n$ times recovers all the vectors.

2.1 Robustness to $1/\operatorname{poly}(n)$ errors

Jenrich's algorithm is already robust to a small amount of error in the input.

Input: $B = A + C$, where A is as above and every entry of C has magnitude at most n^{-10}.

Algorithm: same as above.

Analysis: Now the matrix M takes the form

$$M = \sum_{i=1}^{r} \langle a_i, g\rangle a_i a_i^\top + C'$$

It is elementary to show that for every a_i,

$$\mathbb{P}\{\langle a_i, g\rangle \geqslant 200 \max_{j \neq i} |\langle a_i, g\rangle|, 200\|C'\|_{op}\} \geqslant n^{-O(1)} .$$

Suppose this occurs for a_1. Then there is a number c such that $M = c a_1 a_1^\top + M'$, where $\|M'\| \leqslant c/10$. Thus, the top eigenvalue of M is at least $99c/100$, and so $\langle a_i, v\rangle^2 \geqslant 0.9$ where v is the top eigenvector of M.

It follows that for every i, with probability $n^{-O(1)}$, the algorithm outputs b such that $\langle a_i, b\rangle^2 \geqslant 0.9$.

To turn this in to an algorithm to recover $a_1, \ldots, a_r$ to accuracy 0.9 we need a way to check that this $n^{-O(1)}$-probability event has occurred. This can be done by checking that the value $\langle B, v^{\otimes 3}\rangle \geqslant 0.7$; we leave the details to the reader.

Exercise: Describe and analyze an algorithm for orthogonal tensor decomposition in the above setting, which first uses Jenrich's algorithm to find candidate vectors $v_1, \ldots, v_m$ for some $m = n^{O(1)}$, then uses the values $\langle B, v_i^{\otimes 3}\rangle$ to determine which v_i's are close to some a_i.

2.2 Larger errors and tensor norms

As before, our setting is orthonormal vectors $a_1, \ldots, a_n$ with a tensor $B = \sum_{i=1}^{n} a_i^{\otimes 3} + C$ for some error tensor C. Sample-efficient statistical estimation algorithms require tensor decomposition algorithms which tolerate nastier errors than C which has entries $\ll 1/\operatorname{poly}(n)$. But the question of how to measure such errors is subtle. We will need to make a detour to discuss a variety of tensor norms in which to measure errors.

ℓ_2-norm

The simplest norm we could consider treats a k-tensor as an n^k-length vector and measures its Euclidean norm. That is, if T is a 3-tensor,

$$\|T\|_2 = \left(\sum_{i,j,k \in [n]} T_{ijk}^2 \right)^{1/2} .$$

injective tensor norm

A rather different norm is the *injective tensor norm*, which is a natural tensor generalization of the spectral norm of a matrix. If T is a k-tensor,

$$\|T\|_{inj} = \max_{\|x\|_2=1} \left|\langle T, x^{\otimes k}\rangle\right| .$$

In words, the injective tensor norm treats T as the coefficients of a homogeneous degree-k polynomial and maximizes that polynomial over the unit sphere.

Exercise: Prove that for every tensor T,

$$\|T\|_2 \geqslant \|T\|_{inj} .$$

Unlike the ℓ_2-norm or the matrix spectral norm, the injective tensor norm is NP-hard to compute exactly, and there is evidence (in the form no-go results for SoS algorithms) suggesting NP-hard to approximate within $n^{o(1)}$ factors.

SoS norms

SoS norms are a family of efficiently computable relaxations of the tensor injective norm. The d-th SoS norm of a 3-tensor T is

$$\|T\|_{SoS_d} = \max_{\substack{\tilde{\mathbb{E}} \text{ is degree } d \\ \text{satisfies } \{\|x\|^2=1\}}} \tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle^2 .$$

(This is well defined so long as d is even and $d \geqslant 6$.) The d-th SoS norm can be computed in $n^{O(d)}$ time by solving the SoS semidefinite program. The definition generalizes naturally to k-tensors.

By the usual duality, $\|T\|_{SoS_d}$ is also the best upper bound certifiable on the maximum of the polynomial $\langle x^{\otimes k}, T\rangle$ over the unit sphere by degree-d SoS proofs. That is, if $c = \|T\|_{sos_d}$ is the least c such that there is an SoS proof

$$c - \langle T, x^{\otimes k}\rangle + q(x)(\|x\|^2 - 1) \geq 0$$

for some q of degree $\leqslant d$.

comparing norms

Exercise: Show that

$$\|T\|_2 \geqslant \|T\|_{SoS_6} \geqslant \|T\|_{SoS_8} \geqslant \ldots \geqslant \|T\|_{SoS_n} \geqslant \|T\|_{inj} .$$

Exercise (separation of norms): Suppose T is a 4-tensor whose entries are iid samples from $\mathcal{N}(0,1)$. Show that (with high probability) (a) $\|T\|_2 \approx n^2$ (easy), (b) $\|T\|_{inj} \leqslant \sqrt{n}\log(n)^{O(1)}$ (use a Chernoff/union bound argument), and (c)

$\|T\|_{SoS_8} \leqslant n \log(n)^{O(1)}$. As a bonus exercise, try to show (d) $\|T\|_{SoS_8} \geqslant n^{0.99}$ (this might be *very* difficult!). An easier (still nontrivial) version of this exercise is to show that with high probability

$$\max_{\tilde{\mathbb{E}} \text{ degree 4, satisfies } \{\|x\|^2=1\}} \tilde{\mathbb{E}}\langle T, x^{\otimes 4}\rangle \geqslant n^{0.99} .$$

(The only difference from the SoS_8 norm is the omission of the square and that the maximization is over $\tilde{\mathbb{E}}$ of degree 4.)

Exercise: Show that if k is even, the norm $\|T\|_{op}$ given by unfolding T to a $n^{k/2} \times n^{k/2}$ matrix and measuring its spectral norm satisfies $\|T\|_{op} \geqslant \|T\|_{sos_k}$.

non-algorithmic tensor decomposition under injective norm error

First, let us establish that if we are not concerned about efficient algorithms, tensor decomposition is possible when errors are bounded in injective norm.

Exercise: Suppose $B = \sum_{i=1}^n a_i^{\otimes 4} + C$ where $a_1, \ldots, a_n$ are orthonormal and $\|C\|_{inj} \leqslant o(1)$. Show that the maximizers $b_1, \ldots, b_n$ of the polynomial $f(x) = \langle B, x^{\otimes 4}\rangle$ over the unit sphere $\{x : \|x\| = 1\}$ satisfy $\langle b_i, a_i\rangle \geqslant 1 - o(1)$.

This shows that an exhaustive search algorithm (perhaps with some appropriate discretization of $\mathbb{R}^n$) finds a good decomposition of an orthogonal tensor with injective norm error.

2.3 Jenrich's and larger errors

How does Jenrich's algorithm perform with errors larger than $1/poly(n)$ in each coordinate? First, we give an example tensor decomposition problem which has errors bounded in injective norm but for which Jenrich's algorithm breaks down.

large random errors

As usual, let $B = \sum_{i=1}^n a_i^{\otimes 4} + C$, where $a_1, \ldots, a_n$ are orthonormal. Suppose C has iid entries from $\mathcal{N}(0, 1/n^{1.1})$. In a previous exercise, you showed that $\|C\|_{inj} \leqslant o(1)$ with high probability. How well does Jenrich's algorithm do to decompose B? The algorithm will sample $g \sim \mathcal{N}(0, I_{n^2})$ and compute

$$M_g = \sum_{i=1}^n \langle g, a_i \otimes a_i\rangle a_i a_i^\top + \sum_{i,j=1}^n g_{ij} C_{ij}$$

where $C_{ij} \in \mathbb{R}^{n\times n}$ is the ij-th matrix slice of the 4-tensor C. Without getting too rigorous, the distribution of each entry of the matrix $E = \sum_{i,j=1}^n g_{ij} C_{ij}$ is roughly Gaussian with variance $n^{0.9}$, and the entries are independent (conditioned on g). Thus the eigenvalues of E are roughly $n^{1.45}$ in magnitude, swamping the contribution of the matrix $\sum_{i=1}^n \langle g, a_i \otimes a_i\rangle a_i a_i^\top$. (Making this rigorous would take us on a tour of the beautiful theory of spiked Gaussian and Wigner matrices, but that would be too far afield for present.)

3 The high spectral entropy tensor decomposition algorithm

Ma, Shi, and Steurer introduced a method improve the performance of Jenrich's algorithm in the presence of larger errors. Their algorithm tolerates errors which are bounded in SoS norm.

3.1 Aside: decomposing tensors is the same thing as rounding moments

As usual, consider the goal of recovering $a_1, \ldots, a_r$ unit vectors in $\mathbb{R}^n$ from a tensor $T = \sum_{i=1}^r a_i^{\otimes 3} + C$. From now on, instead of applying Jenrich-like algorithms directly to input tensors, we will think of algorithms which work in two phases:

1. Solve a convex relaxation formed from the input tensor T to find moments of a (pseudo)distribution which is correlated with the vectors $a_1, \ldots, a_r$.

2. Round a moment tensor (usually the third or fourth moments) of the (pseudo)distribution to output vectors $b_1, \ldots, b_r$ correlated with $a_1, \ldots, a_r$.

Let us return to our first example of zero-error orthogonal tensor decomposition with input $A = \sum_{i=1}^r a_i^{\otimes 3}$. Rescaling, the tensor $\frac{1}{r}A$ is the third moment tensor of the finitely-supported distribution μ on the unit sphere which uniformly chooses one of the vectors $a_1, \ldots, a_r$. That is, $\mathbb{E}_{x \sim \mu} x^{\otimes 3} = \frac{1}{r}A$. Applying Jenrich's algorithm to this tensor (via the matrix $M = \mathbb{E}_{x \sim \mu} \langle x, g \rangle x x^\top$) was enough to recover the vectors $a_1, \ldots, a_r$. Here we did not even have to solve a convex relaxation to obtain a good moment tensor $\mathbb{E}\, x^{\otimes 3}$.

3.2 Orthogonal tensor decomposition with SoS-bounded errors

Theorem 3.1 (Ma-Shi-Steurer (weakened parameters for easier proof))**.** *There is $n^{O(d)}$-time time algorithm with the following guarantees. Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be orthonormal and let $A = \sum_{i=1}^r a_i^{\otimes 3}$. Let $T = A + C$ where $\|C\|_{sos_d} \leqslant o(1)$. The algorithm takes input T and outputs a (randomized) unit vector $b \in \mathbb{R}^n$ such that for every $i \leqslant r$,*

$$\mathbb{P}\{\langle a_i, b \rangle \geqslant 1 - o(1)\} \geqslant n^{-O(1)}$$

The first ingredient in the proof uses the SoS algorithm to find a pseudodistribution whose moments are correlated with those of the uniform distribution over $a_1, \ldots, a_r$.

Claim 3.2. In the setting of the above theorem, if $\tilde{\mathbb{E}}$ of degree d solves

$$\operatorname{argmax}_{\tilde{\mathbb{E}} \text{ satisfies } \|x\|^2 = 1} \tilde{\mathbb{E}} \langle T, x^{\otimes 3} \rangle$$

then $\tilde{\mathbb{E}} \sum_{i=1}^r \langle a_i, x \rangle^3 \geqslant 1 - o(1)$.

Proof. Let μ be the uniform distribution on $a_1, \ldots, a_r$. On the one hand, the maximum value of this optimization problem is at least

$$\mathop{\mathbb{E}}_{x \sim \mu} \sum_{i=1}^{r} \langle x, a_i \rangle^3 + \langle C, x^{\otimes 3} \rangle \geqslant 1 - o(1)\,.$$

where we have used the sos_d-bounedness of C.

On the other hand, any $\tilde{\mathbb{E}}$ which achieves objective value δ must satisfy

$$\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle x, a_i \rangle^3 \geqslant \delta - o(1)\,.$$

by similar reasoning. All together, the optimizer satisfies $\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle x, a_i \rangle^3 \geqslant 1 - o(1)$. □

It will be technically convenient also to assume that $\tilde{\mathbb{E}}$'s fourth moments are correlated with the fourth moments of the uniform distribution on $a_1, \ldots, a_r$. This is allowed, because if $\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle a_i, x \rangle^3 \geqslant 1 - o(1)$, then also

$$1 - o(1) \leqslant \tilde{\mathbb{E}} \sum_{i=1}^{r} \langle a_i, x \rangle^3 \leqslant \left(\sum_{i=1}^{r} \langle a_i, x \rangle^2 \right)^{1/2} \left(\sum_{i=1}^{r} \langle a_i, x \rangle^4 \right)^{1/2} \leqslant \left(\sum_{i=1}^{r} \langle a_i, x \rangle^4 \right)^{1/2}\,.$$

Thus we can assume access to a pseudodistribution with $\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle x, a_i \rangle^4 \geqslant 1 - o(1)$. We are hoping that $\tilde{\mathbb{E}}$'s moments look enough like those of μ that we can extract the a_i's from $\tilde{\mathbb{E}}$ using Jenrich's algorithm. Unfortunately, knowing only that $\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle x, a_i \rangle^4 \geqslant 1 - o(1)$ is not enough.

> **Exercise:** Construct a distribution ν on the unit sphere satisfying $\mathbb{E}_{x \sim nu} \sum_{i=1}^{r} \langle x, a_i \rangle^3 \geqslant 1 - o(1)$ but the top eigenvector v of the matrix M from Jenrich's algorithm applied to $\mathbb{E}_{x \sim \nu} x^{\otimes 3}$ satisfies $\langle v, a_i \rangle \leqslant o(1)$ with high probability for every a_i. *Hint: some spurious eigenvector in the matrix M should come from ν putting $o(1)$ probability on a vector having nothing to do with $a_1, \ldots, a_r$.*

3.3 High entropy saves the day

The key observation of Ma, Shi, and Steurer is that a distribution (or a pseudodistribution) on the unit sphere which is correlated with A and has high entropy (in a sense we will momentarily make precise) is enough like the uniform distribution on $a_1, \ldots, a_r$ that it can be rounded using Jenrich's algorithm. This should make sense in light of the preceding exercise. The counterexample ν (described in the hint) places probability $\gg 1/r$ on a single vector—a very low entropy thing to do! If we can force our pseudodistribution not to do something like this, we can remove spurious vectors appearing in the spectrum of the matrices in Jenrich's algorithm.

We will require that our pseudodistribution's moment matrices do not have large eigenvalues. Notice that if μ is the uniform distribution over orthonormal vectors $a_1, \ldots, a_r$, then $\| \mathbb{E}_{x \sim \mu} x x^\top \| = 1/r$.

Claim 3.3. Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be orthonormal. If $\tilde{\mathbb{E}}$ is a degree-4 pseudodistribution satisfying $\{\|x\|^2 = 1\}$ and $\|\tilde{\mathbb{E}}\, xx^\top\|_{op}, \|\tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\|_{op} \leqslant 1/r$ with $\tilde{\mathbb{E}} \sum_{i=1}^r \langle a_i, x\rangle^4 \geqslant 1 - o(1)$, then for all but a $o(1)$-fraction of $a_1, \ldots, a_r$,

$$\tilde{\mathbb{E}}\langle a_i, x\rangle^4 \geqslant (1 - o(1))/r\,.$$

Proof. Suppose to the contrary that a $\delta = \Omega(1)$-fraction of $a_1, \ldots, a_r$ have $\tilde{\mathbb{E}}\langle a_i, x\rangle^4 \leqslant (1-\delta)/r$. Then there is some a_i with $\tilde{\mathbb{E}}\langle a_i, x\rangle^4 > 1/r$, by averaging. But for any unit vector a,

$$\tilde{\mathbb{E}}\langle a, x\rangle^4 \leqslant \left\|\tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\right\|_{op} \leqslant \frac{1}{r}\,.$$

□

Next we show how to exploit the constraints $\|\tilde{\mathbb{E}}\, xx^\top\|, \|\tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\| \leqslant 1/r$ to round a pseudodistribution $\tilde{\mathbb{E}}$ to produce estimates of the vectors $a_1, \ldots, a_r$.

Lemma 3.4. *Let $a \in \mathbb{R}^n$ be a unit vector and let $\tilde{\mathbb{E}}$ be a degree-6 pseudodistribution satisfying $\{\|x\|^2 = 1\}$ and $\|\tilde{\mathbb{E}}\, xx^\top\|_{op}, \|\tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\|_{op}, \|\tilde{\mathbb{E}}(x^{\otimes 3})(x^{\otimes 3})^\top\|_{op} \leqslant \frac{1}{r}$. Suppose $\tilde{\mathbb{E}}\langle x, a\rangle^4 \geqslant (1-o(1))/r$. Then with probability $n^{-O(1)}$, the top eigenvector v of the matrix $M_g \overset{\text{def}}{=} \tilde{\mathbb{E}}\langle x \otimes x, g\rangle xx^\top$ for $g \sim \mathcal{N}(0, \mathrm{Id})$ satisfies $\langle v, a\rangle^2 \geqslant 0.99$.*

Together with the preceding claim this is enough to prove a (slightly weakened) version of the theorem.[1]

> **Exercise.** Show that the distribution you constructed in the previous exercise violates assumptions on $\tilde{\mathbb{E}}$ above.

To prove Lemma 3.4 will require two claims.

Claim 3.5. Let $g \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \preceq \mathrm{Id}$. Then

$$\mathop{\mathbb{E}}_{g} \|\tilde{\mathbb{E}}\langle x \otimes x, g\rangle xx^\top\| \leqslant O(\log n)^{1/2}/r\,.$$

Proof sketch. We prove the case $\Sigma = \mathrm{Id}$ from which general $\Sigma \preceq \mathrm{Id}$ can be derived. In this case,

$$\tilde{\mathbb{E}}\langle x \otimes x, g\rangle xx^\top = \sum_{i \leqslant n} g_{ij}\, \tilde{\mathbb{E}}\, x_i x_j xx^\top$$

where $g_{ij} \sim \mathcal{N}(0, 1)$ are independent. Let $M_{ij} = \tilde{\mathbb{E}}\, x_i x_j xx^\top$. By standard matrix concentration bounds, the expected spectral norm of this matrix is at most

$$O(\log n)^{1/2} \cdot \left\|\sum_{i \leqslant n} M_{ij} M_{ij}^\top\right\|^{1/2}\,.$$

[1] It is possible to recover all of the vectors, rather than only a 0.99 fraction of them, by adding constraints on $\tilde{\mathbb{E}}$ and re-solving after each vector is found—see the paper of Ma, Shi, and Steurer for details.

It is an exercise to show that our assumptions on spectral norms of moments of $\tilde{\mathbb{E}}$ imply

$$\left\| \sum_{i \leqslant n} M_{ij} M_{ij}^\top \right\|^{1/2} \leqslant 1/r\,.$$

The claim follows. □

Claim 3.6. The matrix $\tilde{\mathbb{E}}\langle x, a\rangle^2 xx^\top$ can be expressed as

$$\tilde{\mathbb{E}}\langle x, a\rangle^2 xx^\top = \tfrac{1}{r} aa^\top + E$$

where $\|E\|_{op} \leqslant o(1/r)$.

Proof sketch. To save on notation, without loss of generality suppose that $a = e_1$. Consider the submatrix M of $\tilde{\mathbb{E}}\, x_1 xx^\top$ given by rows and columns $2, \ldots, n$. This matrix has spectral norm

We assumed that

$$\tilde{\mathbb{E}}\, x_1^4 \geqslant (1 - o(1))/r$$

but at the same time

$$\tilde{\mathbb{E}}\, x_1^2 \sum_{i=1}^{r} x_i^2 \leqslant \tilde{\mathbb{E}}\, x_1^2 \leqslant 1/r$$

by our eigenvalue bounds on $\|\tilde{\mathbb{E}}\, xx^\top\|$. So,

$$\tilde{\mathbb{E}}\, x_1^2 \sum_{i=2}^{r} x_i^2 \leqslant o(1/r)\,.$$

Let $v \in \mathbb{R}^n$ be a unit vector orthogonal to a. Then

$$\tilde{\mathbb{E}}\, x_1^2 \langle x, v\rangle^2 \leqslant \tilde{\mathbb{E}}\, x_1^2 \|\Pi^\perp x\|^2 \leqslant o(1/r)$$

where $\Pi^\perp$ is the projector to last $n-1$ coordinates. Since $\tilde{\mathbb{E}}\, x_1^2 xx^\top \succeq 0$, this implies that $\|e_1 e_1^\top / r - \tilde{\mathbb{E}}\, x_1^2 xx^\top\| \leqslant o(1/r)$. □

Proof sketch of Lemma 3.4. We sample the vector g as $g = \xi \cdot a + g'$, where g' is a unit-variance multivariate Gaussian in the subspace orthogonal to $a \otimes a$, and ξ is a unit-variance univariate Gaussian. Furthermore, ξ and g' are independent. So we can write M_g as

$$M_g = \xi\, \tilde{\mathbb{E}}\langle x, a\rangle^2 xx^\top + \tilde{\mathbb{E}}\langle x \otimes x, g'\rangle xx^\top\,.$$

By Markov's inequality, our claims above, and Gaussian anti-concentration, with probability $n^{-O(1)}$ we can write

$$M_g = \xi aa^\top / r + E$$

where $\|E\| \leqslant 0.001\xi/r$ and $\xi > 0$. The lemma follows. □

4 What if the errors are not bounded in SoS norm?

Many tensors do not have errors bounded in SoS norm but should nonetheless be easy to decompose. For example, consider the tensor $T = \sum_{i=1}^r a_i^{\otimes 3} + c^{\otimes 3}$, where as usual the $a_1, \ldots, a_r$ are orthonormal but c has norm 100. The tensor $c^{\otimes 3}$ does not have SoS norm $\ll 1$, but at least intuitively this should not present a real difficulty in decomposing this tensor. However, the solution to $\operatorname{argmax}_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$ will put all its weight on c, so the resulting $\tilde{\mathbb{E}}$ will (probably) not contain any information about $a_1, \ldots, a_r$.

There are likely many kinds of errors not $\ll 1$ in SoS norm but which do not present a problem for tensor decomposition. Hopkins and Steurer study the setting that the input tensor is correlated—in the Euclidean sense—with the target orthogonal tensor. Tensor decomposition in this setting can be used to obtain algorithms for statistical inference problems with very tight sample complexity guarantees (see the paper for more).

More formally, the goal is to decompose an orthogonal tensor $A = \sum_{i=1}^r a_i^{\otimes 3}$, and the input is a tensor T such that

$$\frac{\langle T, A\rangle}{\|T\|\|A\|} \geqslant \delta = \Omega(1)\,.$$

By standard linear algebra, up to scaling we can think of $T = A + B$ where $\langle A, B\rangle = 0$ and $\|B\| = O(\|A\|)$. Note that the condition $\langle A, B\rangle = 0$ cannot be dropped: if $T = A + B$ and we do not require $\langle A, B\rangle = 0$, then setting $B = -A$ would destroy all the information about A in the input T.

Even assuming $\langle A, B\rangle = 0$, it is possible in this setting that not all the vectors $a_1, \ldots, a_r$ can be recovered. For example, if $B = a_1^{\otimes 3} - a_2^{\otimes 3}$, then $\langle A, B\rangle = 0$ but $A + B$ contains no information about a_2. We will have to set our sights on recovering just some of the vectors.

In light of the lemma on rounding pseudodistributions $\tilde{\mathbb{E}}$ having $\tilde{\mathbb{E}} \sum_{i=1}^r \langle x, a_i\rangle^3 \geqslant \delta$, it would be enough to show how to take input T and produce such a pseudodistribution. For this we have the following lemma.

Lemma 4.1 (Hopkins-Steurer)**.** *Let T satisfy*

$$\frac{\langle T, A\rangle}{\|T\|\|A\|} \geqslant \delta = \Omega(1)\,.$$

The solution to the following convex program

$$\min_{\tilde{\mathbb{E}}\ \textit{degree 4}} \|\tilde{\mathbb{E}}\, x^{\otimes 3}\| \ \textit{such that} \tag{4.1}$$

$$\tilde{\mathbb{E}}\ \textit{satisfies}\ \{\|x\|^2 = 1\} \tag{4.2}$$

$$\tilde{\mathbb{E}}\langle x^{\otimes 3}, T\rangle \geqslant \frac{\delta \cdot \|T\|}{\sqrt{r}} \tag{4.3}$$

$$\|\tilde{\mathbb{E}}\, xx^\top\|_{op} \leqslant \tfrac{1}{r} \tag{4.4}$$

$$\|\tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\|_{op} \leqslant \tfrac{1}{r}\,. \tag{4.5}$$

satisfies $\tilde{\mathbb{E}} \sum_{i=1}^r \langle a_i, x\rangle^3 \geqslant \operatorname{poly}(\delta)$.

Before we prove the lemma—how should we interpret this convex program? The objective function may be unfamiliar, but we can obtain some good intuition if we think about what $\| \mathbb{E}_{x\sim\mu} x^{\otimes 3}\|$ means for μ a distribution supported on orthonormal vectors $a_1, \ldots, a_r$ (but not necessarily the uniform distribution on those vectors). In this case,

$$\| \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\|^2 = \langle \sum_{i=1}^r \mu(i) a_i^{\otimes 3}, \sum_{i=1}^r \mu(i) a_i^{\otimes 3}\rangle = \sum_{i=1}^r \mu(i)^2 = \textsc{collision-probability}(\mu)\,.$$

The collision probability is an ℓ_2 version of entropy—as μ becomes closer to uniform, the collision probability decreases. It is also closely related to the Rényi entropy.

It is a good exercise to convince yourself that in the motivating example from before—$T = \sum_{i=1}^r a_i^{\otimes 3} + c^{\otimes 3}$ where $\|c\| = 100$—the distribution μ of minimal collision probability which obtains $\langle \mathbb{E}_{x\sim\mu} x^{\otimes 3}, T\rangle \geqslant \delta$ also has $\mathbb{E}_{x\sim\mu} \sum_{i=1}^r \langle x, a_i\rangle^3 \geqslant \mathrm{poly}(\delta)$, for small enough constants $\delta > 0$.

The lemma follows from the following general fact

Theorem 4.2 (Appears in this form in Hopkins-Steurer)**.** *Let C be a convex set and $Y \in C$. Let P be a vector with $\langle P, Y\rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$. Then, if we let Q be the vector that minimizes $\|Q\|$ subject to $Q \in C$ and $\langle P, Q\rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$, we have*

$$\langle Q, Y\rangle \geqslant \delta/2 \cdot \|Q\| \cdot \|Y\|\,. \tag{4.6}$$

Furthermore, Q satisfies $\|Q\| \geqslant \delta\|Y\|$.

Proof. By construction, Q is the Euclidean projection of 0 into the set $C' := \{Q \in C \mid \langle P, Q\rangle \geqslant \delta\|P\| \cdot \|Y\|\}$. It's a basic geometric fact (sometimes called Pythagorean inequality) that a Euclidean projection into a set decreases distances to points into the set. Therefore, $\|Y - Q\|^2 \leqslant \|Y - 0\|^2$ (using that $Y \in C'$). Thus, $\langle Y, Q\rangle \geqslant \|Q\|^2/2$. On the other hand, $\langle P, Q\rangle \geqslant \delta\|P\| \cdot \|Y\|$ means that $\|Q\| \geqslant \delta\|Y\|$ by Cauchy–Schwarz. We conclude $\langle Y, Q\rangle \geqslant \delta/2 \cdot \|Y\| \cdot \|Q\|$. □

Now we can prove the lemma.

Proof of lemma. Consider the convex set

$$C = \{\tilde{\mathbb{E}} \text{ degree-4 satisfying } \|x\|^2 = 1, \| \tilde{\mathbb{E}}\, xx^\top\|_{op} \leqslant \tfrac{1}{r}, \| \tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top\|_{op} \leqslant \tfrac{1}{r}\}\,.$$

The uniform distribution μ over $a_1, \ldots, a_r$ is in C, and T satisfies

$$\langle T, \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\rangle \geqslant \delta \cdot \|T\| \cdot \| \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\|\,.$$

Let $\tilde{\mathbb{E}}$ be the solution to the convex program in the lemma. According to the the theorem on correlation-preserving projections,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\rangle \geqslant \delta/2 \cdot \| \tilde{\mathbb{E}}\, x^{\otimes 3}\| \cdot \| \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\| \geqslant \delta^2/2 \cdot \| \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\|^2 = \delta^2/(2r)\,.$$

where in the last step we have used that the collision probability of μ is $1/r$. Rearranging,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \mathop{\mathbb{E}}_{x\sim\mu} x^{\otimes 3}\rangle = \frac{1}{r} \cdot \tilde{\mathbb{E}} \sum_{i=1}^{r} \langle a_i, x\rangle^3$$

which proves the lemma. □

To turn the above into an algorithm requires a version of Lemma 3.4 suitable for this low-correlation regime, stated below. The proof uses mostly the same ideas as that of Lemma 3.4.

Lemma 4.3 (Hopkins-Steurer)**.** *For every $0 < \delta < 1$ there is a polynomial time algorithm with the following guarantees. Suppose $\tilde{\mathbb{E}}$ is a degree-4 pseudoexpectation in the variables $x_1, \ldots, x_n$ satisfying $\{\|x\|^2 = 1\}$. Furthermore, suppose that*

1. *$\tilde{\mathbb{E}} \sum_{i=1}^r \langle x, a_i\rangle^3 \geqslant \delta$.*
2. *$\| \tilde{\mathbb{E}}\, x x^\top \|_{op} \leqslant \frac{1}{r}$ (this is a convex constraint!).*
3. *$\| \tilde{\mathbb{E}}(x \otimes x)(x \otimes x)^\top \| \leqslant \frac{1}{r}$ (this is also a convex constraint!).*

Then for at least $r' = \mathrm{poly}(\delta) r$ vectors $a_1, \ldots, a_{r'}$, the algorithm takes input $\tilde{\mathbb{E}}$ and produces a unit vector b such that

$$\mathbb{P}\{\langle a_i, b\rangle \geqslant \mathrm{poly}(\delta)\} \geqslant n^{-\mathrm{poly}(1/\delta)}\,.$$

5 Bibliography

[Hopkins-Steurer]: Efficient Bayesian estimation from few samples: community detection and related problems. Samuel B. Hopkins, David Steurer. *In submission.*
[Ma-Shi-Steurer]: Polynomial-time Tensor Decompositions with Sum-of-Squares. Tengyu Ma, Jonathan Shi, David Steurer. *FOCS 2016.*