CS369H: Hierarchies of Integer Programming Relaxations Spring 2016-2017

Low correlation tensor decomposition via entropy maximization

Lecture and notes by Sam Hopkins *Scribes: James Hong*

Overview These notes are adapted from Sam Hopkin's notes for his talk on May 19, 2017 (Lecture 7 in CS369H).

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.

1.1 Contexts

Suppose $x \in R^n$ is a mixture of r Gaussians with centers $a_1, ..., a_r$ and we are allowed to see $x_1, ..., x_m$ IID samples of X. Can we recover the centers, given enough samples. Specifically, if the samples are drawn as follows: (1) sample $i \sim [r]$, (2) output $x \sim \mathcal{N}(a_i, I)$.

Some possible ways we could attempt this problem based on moments:

1. $\mathbb{E}x \approx \frac{1}{m}\sum_{i=1}^{m} x_i$

   This does not work and is probably 0 when $\sum_{i=1}^{r} a_i = 0$.

2. $\mathbb{E}xx^\top \approx \frac{1}{m}\sum_{i=1}^{m} x_i x_i^\top$

   This also does not work. Consider what happens if the centers are the coordinate vectors. Then $\mathbb{E}xx^\top = I$. Moreover, this occurs when $a_1, ..., a_r$ are a rotation of the coordinate basis. The result is not unique.

3. $\mathbb{E}x^{\otimes 3} \approx \frac{1}{m}\sum_{i=1}^{m} x_i^{\otimes 3} = \frac{1}{m}\sum_{i=1}^{m} x_i \otimes x_i \otimes x_i$

   Instead, we should estimate $\mathbb{E}x^{\otimes 3}$, which is unique if we know the exact centers. We must ensure that the tensor decomposition algorithm is robust to estimation errors. A more robust algorithm allows for lower sample complexity.

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.

2 Jenrich's algorithm for orthogonal tensor decomposition

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.

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 $i = \arg\max \langle a_i, g \rangle$. 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 $1/\mathbf{poly}(n)$ 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

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

The entries of C' are of the form:

$$C'_{ij} = \sum_{k=1}^{n} g_k C_{ijk}$$

It is elementary to show that for every a_i,

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

Suppose this occurs for a_1. Then there is a number c such that $M = ca_1a_1^\top + M'$, where $\|M'\| \leq c/10$. Thus, the top eigenvalue of M is at least $99c/100$, and so $\langle a_i, v\rangle^2 \geq 0.9$ 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 $\langle a_i, b\rangle^2 \geq 0.9$.

Intuitively, this means that we can still recover each a_i w.h.p and every a_i appears.

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, but we will ignore this issue for this lecture. (Same as in dictionary learning algorithm.)

Exercise 7.1. *Let v be the max eigenvector of M. Show that w.h.p, $\max_i \langle a_i, v\rangle^2 \geq 1 - o(1)$ and every a_i is still maximal w.p. $n^{-O(1)}$.*

2.2 Nonrobustness to larger errors

What about larger errors? If we think of $A = \sum_{i=1}^r a_i^{\otimes 3}$ as an n^3-dimensional vector, its 2-norm $\|A\|^2 = r$. Previously the error tensor C we considered had $\|C\|^2 \leq n^{-5}\|A\|^2$. Could we tolerate errors like $\|C\|^2 \leq 0.01\|A\|^2$?

Example: This example will show that Jenrich's algorithm will not work out of the box in the presence of such large errors. Suppose $r = n/2$ and $A = \sum_{i=1}^r a_i^{\otimes 3}$ for $a_1, \ldots, a_r$ orthogonal unit vectors in $\mathbb{R}^n$, as usual. Let $C = c^{\otimes 3}$, where c is a vector of norm $0.01n^{1/10}$. Then the matrix M computed by Jenrich's algorithm has the form

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

Because $\|c\| \gg \|a_i\|$, the probability that $\langle a_i, g\rangle > |\langle g, c\rangle|$ is exponentially small; the top eigenvector of M will instead by very close to c.

However, $\|A\|^2 \approx n$, while $\|C\| \leq n^{0.9}$. So measured in 2 norm, the error C is small compared to A.

We might even consider error tensors C which are large in metrics other than the 2-norm.

For instance, we will define the injective tensor norm below.

$$||T||_{inj} = \max_{||x||=1} \langle T, x^{\otimes 3}\rangle$$

Note that if $T = \sum_{i=1}^r a_i^{\otimes 3}$, then $\langle T, x^{\otimes 3}\rangle = \sum_{i=1}^r \langle a_i, x\rangle^3$.

Exercise 7.2. *Suppose we instead defined $T = \sum_{i=1}^r a_i^{\otimes 3} + R$ where entries of R are IID Gaussian. How large can R be for the problem to still be well defined? Explain why Jenrich's algorithm will fail.*

3 The high spectral entropy tensor decomposition algorithm

Ma, Shi, and Steurer introduced a method to get around the problem that Jenrich's algorithm has in the presence of larger errors.

3.1 Aside: SoS norms

For a k-tensor T and $d \in \mathbb{N}$, then sos_d norm of T is

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

As $d \to \infty$, the sos_d norm becomes the *injective tensor norm*. But we will be interested in the constant-d setting, which as usual will correspond to polynomial time algorithms. By the usual duality, $\|T\|_{sos_d}$ is also the best bound certifiable on the polynomial $\langle x^{\otimes k}, T\rangle$ by degree-d SoS proofs. That is, if $c = \|T\|_{sos_d}$, there is an SoS proof

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

for some q of degree $\leq d$. (And this is not true for any $c' < c$.)

Exercise 7.3. *Show that $||T||_{sos_k}$ is a norm.*

Exercise 7.4. *Show that the 2-norm of a k-tensor (that is, its 2-norm as a large vector in $\mathbb{R}^{n^k}$) is an upper bound on its sos_d norm, for any $d \geq k$.*

Exercise 7.5. *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 $\|T\|_{op} \geq \|T\|_{sos_k}$.*

Exercise 7.6. *What is the analogue of the sos norm for matrices? Prove that they collapse to $||M||_{inj}$.*

One note here is that computing $||T||_{inj}$ is intractable. For matrices, we can use the power method. For the problem,

$$\max_{||x||=1} \langle M, x^{\otimes 2}\rangle = \max_{||x||=1} x^\top M x$$

which can be computed by $x \sim \mathcal{N}(0, I)$, $M^l x \approx$ max eigenvector. We can apply a similar method for orthogonal tensors.

3.2 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 $M = \mathbb{E}_{x\sim\mu}\langle x, g\rangle xx^\top$) 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 $\mathbb{E}x^{\otimes 3}$.

3.3 Orthogonal tensor decomposition with SoS-bounded errors

Theorem 7.1 (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 $\|C\|_{sos_d} \leq o(1)$. The algorithm takes input T and outputs a (randomized) unit vector $b \in \mathbb{R}^n$ such that for every $i \leq r$,*

$$\Pr\{\langle a_i, b\rangle \geq 1 - o(1)\} \geq n^{-O(1)}$$

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

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

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

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

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

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

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

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

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

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

Exercise 7.7. *Show that* $\left(\tilde{\mathbb{E}}\sum_{i=1}^{r}\langle a_i, x\rangle^2\right)^{1/2} \leq 1.$

Thus we can assume access to a pseudodistribution with $\tilde{\mathbb{E}}\sum_{i=1}^{r}\langle x, a_i\rangle^4 \geq 1 - o(1)$. 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 $\tilde{\mathbb{E}}\sum_{i=1}^{r}\langle x, a_i\rangle^4 \geq 1 - o(1)$ is not enough.

3.4 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.

The MSS algorithm is as follows:

1. Solve $\arg\max \tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$ s.t.

$$\deg\tilde{\mathbb{E}} = 6 \text{ satisfies } \{\|x\|^2 = 1\} \tag{1}$$
$$\|\tilde{\mathbb{E}}xx^\top\|_{op} \leq \tfrac{1}{r} \tag{2}$$
$$\|\tilde{\mathbb{E}}(x\otimes x)(x\otimes x)^\top\|_{op} \leq \tfrac{1}{r} \tag{3}$$

2. Apply Jenrich's algorithm to $\tilde{\mathbb{E}}x^{\otimes 4}$.

Exercise 7.8. *Show that the above is a convex program.*

We 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 $\|\mathbb{E}_{x\sim\mu}xx^\top\| = 1/r$.

Claim 7.2. *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 $\|\tilde{\mathbb{E}}xx^\top\|_{op}, \|\tilde{\mathbb{E}}(x\otimes x)(x\otimes x)^\top\|_{op} \leq 1/r$ with $\tilde{\mathbb{E}}\sum_{i=1}^{r}\langle a_i, x\rangle^4 \geq 1 - o(1)$, then for all but a $o(1)$-fraction of $a_1, \ldots, a_r$,*

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

Proof. Suppose to the contrary that a $\delta = \Omega(1)$-fraction of $a_1, \ldots, a_r$ have $\tilde{\mathbb{E}}\langle a_i, x\rangle^4 \leq (1-\delta)/r$. 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,

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

□

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

Lemma 7.1. *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 $\|\tilde{\mathbb{E}}xx^\top\|_{op}, \|\tilde{\mathbb{E}}(x\otimes x)(x\otimes x)^\top\|_{op}, \|\tilde{\mathbb{E}}(x^{\otimes 3})(x^{\otimes 3})^\top\|_{op} \leq \frac{1}{r}$. Suppose $\tilde{\mathbb{E}}\langle x, a\rangle^4 \geq (1 - o(1))/r$. Then with probability $n^{-O(1)}$, the top eigenvector v of the matrix $M_g \stackrel{def}{=} \tilde{\mathbb{E}}\langle x\otimes x, g\rangle xx^\top$ for $g \sim \mathcal{N}(0, Id)$ satisfies $\langle v, a\rangle^2 \geq 0.99$.*

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

To prove Lemma 7.1 will require two claims.

Claim 7.3. *Let $g \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \preceq Id$. Then*

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

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

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

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

$$O(\log n)^{1/2} \cdot \|\sum_{i\leq n} M_{ij}M_{ij}^\top\|^{1/2}$$

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

$$\|\sum_{i\leq n} M_{ij}M_{ij}^\top\|^{1/2} \leq 1/r$$

The claim follows. □

Exercise 7.9. *Show that* $\|\sum_{i\leq n} M_{ij}M_{ij}^\top\|^{1/2} \leq \dfrac{1}{r}$.

Claim 7.4. *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 $\|E\|_{op} \leq o(1/r)$.

Proof sketch. To save on notation, without loss of generality suppose that $a = e_1$. Consider the submatrix M of $\tilde{\mathbb{E}}x_1xx^\top$ given by rows and columns $2, \ldots, n$. This matrix has spectral norm

We assumed that

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

but at the same time

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

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

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

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

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

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 $\|e_1e_1^\top/r - \tilde{\mathbb{E}}x_1^2xx^\top\| \le o(1/r)$. □

Proof sketch of Lemma 7.1. 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 $\|E\| \le 0.001\xi/r$ 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 $a_1, \dots, a_r$ 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 $\max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$ will put all its weight on c, so the resulting $\tilde{\mathbb{E}}$ will (probably) not contain any information about $a_1, \dots, a_r$.

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. 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

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

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 $a_1, \dots, a_r$ 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 $\tilde{\mathbb{E}}\sum_{i=1}^{r}\langle x, a_i\rangle^3 \geq \delta$, it would be enough to show how to take input T and produce such a pseudodistribution. For this we have the following lemma.

Lemma 7.2 (Hopkins-Steurer)**.** *Let T satisfy*

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

The solution to the following convex program

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

$$\tilde{\mathbb{E}} \text{ satisfies } \{\|x\|^2 = 1\} \tag{5}$$

$$\tilde{\mathbb{E}}\langle x^{\otimes 3}, T\rangle \geq \frac{\delta \cdot \|T\|}{\sqrt{r}} \tag{6}$$

$$\|\tilde{\mathbb{E}}xx^\top\|_{op} \leq \tfrac{1}{r} \tag{7}$$

$$\|\tilde{\mathbb{E}}(x\otimes x)(x\otimes x)^\top\|_{op} \leq \tfrac{1}{r} \tag{8}$$

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

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,

$$\|\mathbb{E}_{x\sim\mu}x^{\otimes 3}\|^2 = \langle\sum_{i=1}^{r}\mu(i)a_i^{\otimes 3}, \sum_{i=1}^{r}\mu(i)a_i^{\otimes 3}\rangle = \sum_{i=1}^{r}\mu(i)^2 = \text{COLLISION-PROBABILITY}(\mu)$$

The collision probability is an ℓ_2 version of entropy—as μ becomes closer to uniform, the collision probability decreases. It is 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 $\langle\mathbb{E}_{x\sim\mu}x^{\otimes 3}, T\rangle \geq \delta$ also has $\mathbb{E}_{x\sim\mu}\sum_{i=1}^{r}\langle x, a_i\rangle^3 \geq \text{poly}(\delta)$, for small enough constants $\delta > 0$.

The lemma follows from the following general fact

Theorem 7.2 (Appears in this form in Hopkins-Steurer)**.** *Let $\mathcal{C}$ be a convex set and $Y \in \mathcal{C}$. Let P be a vector with $\langle P, Y\rangle \geq \delta \cdot ||P|| \cdot ||Y||$. Then, if we let Q be the vector that minimizes $||Q||$ subject to $Q \in \mathcal{C}$ and $\langle P, Q\rangle \geq \delta \cdot ||P|| \cdot ||Y||$, we have*

$$\langle Q, Y\rangle \geq \delta/2 \cdot ||Q|| \cdot ||Y||\,. \tag{9}$$

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

Proof. By construction, Q is the Euclidean projection of 0 into the set $\mathcal{C}' := \{Q \in \mathcal{C} \mid \langle P, Q\rangle \geq \delta||P|| \cdot ||Y||\}$. 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, $||Y-Q||^2 \leq ||Y-0||^2$ (using that $Y \in \mathcal{C}'$). Thus, $\langle Y, Q\rangle \geq ||Q||^2/2$. On the other hand, $\langle P, Q\rangle \geq \delta||P||\cdot||Y||$ means that $||Q|| \geq \delta||Y||$ by Cauchy–Schwarz. We conclude $\langle Y, Q\rangle \geq \delta/2 \cdot ||Y|| \cdot ||Q||$. □

Now we can prove the lemma.

Proof of lemma. Consider the convex set

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

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

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

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

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

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

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

which proves the lemma. □

To turn the above into an algorithm requires a version of Lemma 7.1 suitable for this low-correlation regime, stated below. The proof uses mostly the same ideas as that of Lemma 7.1.

Lemma 7.3 (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. $\tilde{\mathbb{E}} \sum_{i=1}^{r} \langle x, a_i\rangle^3 \geq \delta$.
2. $\|\tilde{\mathbb{E}}xx^\top\|_{op} \leq \frac{1}{r}$ *(this is a convex constraint!).*
3. $\|\tilde{\mathbb{E}}(x\otimes x)(x\otimes x)^\top\| \leq \frac{1}{r}$ *(this is also a convex constraint!).*

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

$$\Pr\{\langle a_i, b\rangle \geq poly(\delta)\} \geq n^{-poly(1/\delta)}$$

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.*