Tensor decomposition

In this lecture we will introduce the problem of tensor decomposition, discuss its usefulness as an algorithmic primitive in unsupervised learning, and give a general sum-of-squares framework for tensor decomposition. We will then show how to apply this framework to the problem of decomposing overcomplete tensors with random components.

1 Tensor decomposition

Tensor decomposition is a high-order generalization of matrix decomposition, where we decompose a tensor into rank-one factors. Formally, in the order-3 tensor decomposition problem, we are given the tensor $\mathbf{T} \in \mathbb{R}^{d_1 \otimes d_2 \otimes d_3}$, and we are promised that it has the form

$$\mathbf{T} = \sum_{i=1}^{n} a_i \otimes b_i \otimes c_i,$$

for vectors $a_i \in \mathbb{R}^{d_1}$, $b_i \in \mathbb{R}^{d_2}$, $c_i \in \mathbb{R}^{d_3}$, which are called the *components* of the tensor. The order-3 tensor decomposition problem is to recover the components $\{a_i, b_i, c_i\}_{i=1}^n$ given $\mathbf{T}$. The minimum n for which $\mathbf{T}$ can be decomposed in such a way is called the *rank* of $\mathbf{T}$. Throughout this lecture we will treat the *symmetric* case wherein $a_i = b_i = c_i$ for every $i \in [n]$, though much of what we discuss can be generalized to handle asymmetric (and higher-order) tensors as well.

We remark that many tensor-related problems such as computing rank or decomposition are NP-hard in the worst case, so typically we have to make some structural or distributional assumptions about $\mathbf{T}$ to have a hope of solving the problem in polynomial time.

Uniqueness of decompositions and applications to parameter estimation. Low-rank matrix decompositions are non-unique. Even a single rank-one factor $uv^\top$ can be rewritten as $(\mathbf{M}u)(\mathbf{M}v)^\top$ for any unitary matrix $\mathbf{M}$. We sometimes get around this symmetry by insisting on specific types of decompositions, like the eigendecomposition or singular value decomposition, which we know are unique (up to representation of rank > 1 subspaces) by the spectral theorem.

Tensors, on the other hand, often enjoy a uniqueness of the minimum rank decompoisition even without such orthogonality conditions. Further, tensors may have unique decompositions even when the rank n exceeds the dimension d! This is what we refer to as the *overcomplete* setting ($n > d$).

This makes tensor decomposition a useful primitive for high-dimensional parameter estimation tasks. As a concrete example, consider learning the parameters of a mixture of spherical Gaussians,

$$\mathcal{D} = \sum_{i \in [k]} w_i \mathcal{N}(\mu_i, \mathbb{1}),$$

where $\{\mu_i\}_{i \in [k]} \subset \mathbb{R}^d$, and $w \in \Delta^k$ is a probability distribution. A common paradigm for this problem is the *method of moments*, where one takes enough samples to accurately estimate the order-k moments of the distribution, and expresses them in terms of the parameters. For example, for $X \sim \mathcal{D}$,

$$\mathbf{E}[X] = \sum_{i \in [k]} w_i \mu_i, \ \mathbf{E}\left[XX^\top\right] = \sum_{i \in [k]} w_i \mu_i \mu_i^\top + \mathbb{1}.$$

However, even if we were given exact information about the second moment matrix $\mathbf{E}[XX^\top]$, this is not enough to learn the $\{w_i, \mu_i\}_{i\in[k]}$ because of the issue of rotation invariance of the matrix decomposition. Fortunately, the third-moment tensor can also be written in terms of the parameters of the distribution, and will let us extract information about $\sum_{i\in[k]} w_i\mu_i^{\otimes 3}$. Thus, we can use a tensor decomposition algorithm will tell us the parameters of the mixture model. Many fundamental high-dimensional algorithmic tasks use tensor decomposition as a basic stepping stone; we list a couple (without going into details):

1. Dictionary learning, where we have a dictionary $\mathbf{A} \in \mathbb{R}^{n\times d}$ and receive samples $\mathbf{A}v + \xi$, where v is sparse and ξ a noise vector. Here the goal is to learn the dictionary $\mathbf{A}$. This is achieved by approximately learning the polynomial $\sum_{i\in[n]} \langle a_i, u\rangle^k$ from samples and then running a tensor decomposition algorithm.

2. Topic modeling. This is similar to dictionary learning (the columns represent "topics"), but further assumptions are made on the model (e.g. the combination vector v is non-negative).

3. Learning phylogenetic trees. Here the idea is that we can use some types of property tests to learn the topology of the tree, at which point we can estimate moments of e.g. passing on genetic traits. This lets us learn the parameters of the tree from tensor decomposition, such as estimating transition matrices (from parent species to children species).

4. Independent component analysis (the "cocktail party problem") where we wish to denoise a mixture. The problem formulation is quite similar to dictionary learning.

Notice that for these learning applications, it is much harder to collect enough samples to approximate the true order-k tensor as k gets very large (since there are d^k entries), so it is ideal to learn the parameters efficiently from a low-order tensor, such as order-3. However, it gets easier to design algorithms for tensor decomposition the more moments we have access to (as the next section demonstrates), which makes it important to understand algorithms for low-order tensors.

2 Jennrich's algorithm

Now, we will discuss Jennrich's algorithm, a simple algorithm for decomposing undercomplete tensors. Later, we will see how Jennrich's algorithm is used, in combination with sum-of-squares relaxations, to decompose overcomplete tensors. For now, let $\mathbf{T} = \sum_{i\in[n]} a_i^{\otimes 3}$ be a symmetric third-order tensor for concreteness.

Orthonormal components. We first illustrate the main idea of Jennrich's algorithm in a very simple setting: suppose that the a_i are orthonormal. In this case, the algorithm is just to sample $g \sim \mathcal{N}(0, \mathbb{1}_d)$ and then write down the matrix

$$\mathbf{M}_g = \mathbf{T}[g, \cdot, \cdot] := \sum_{i=1}^{d} g_i \mathbf{T}_i$$

for $\mathbf{T}_i$ the ith $d \times d$ "slice" of $\mathbf{T}$. We call such a matrix a "random contraction" of $\mathbf{T}$. One can check that

$$\mathbf{M}_g = \sum_{i\in[n]} \langle a_i, g\rangle\, a_i a_i^\top,$$

and further when the $\mathbf{A}_{i:}$ are orthonormal, the $\{(\langle a_i, g\rangle, a_i)\}_{i=1}^n$ are also the unique eigendecomposition of $\mathbf{M}_g$ with probability 1. So, by computing the eigendecomposition of $\mathbf{M}_g$, we can recover the a_i.

Independent components. Though we will not use it in what follows, we remark that a variation on this works more generally when the $\{a_i\}_{i\in[n]}$ are not necessarily orthonormal (or close to orthonormal), but merely linearly independent (so still $n \leqslant d$). The algorithm is as follows: we sample $g, g' \sim \mathcal{N}(0, \mathbb{1}_d)$ independently, and form $\mathbf{M}_g, \mathbf{M}_{g'}$. Denoting $\mathbf{D}_g := \mathrm{diag}\,\{\langle a_i, g\rangle\}_{i\in[n]}$, note that

$$\mathbf{M}_g = \mathbf{U}^\top \mathbf{D}_g \mathbf{U},\ \mathbf{M}_{g'} = \mathbf{U}^\top \mathbf{D}_{g'} \mathbf{U},$$

where the rows of $\mathbf{U}$ are given by the $\{a_i\}_{i\in[n]}$. Hence,

$$\mathbf{M}_g^{-1}\mathbf{M}_{g'} = \mathbf{U}^{-1}\left(\mathbf{D}_g^{-1}\mathbf{D}_{g'}\right)\mathbf{U}.$$

Now for any matrix $\mathbf{A} = \mathbf{U}^{-1}\mathbf{D}\mathbf{U}$ with $\mathbf{D}$ diagonal, the rows of $\mathbf{U}$ are the eigenvectors of $\mathbf{A}$ (up to scaling), so we can use the eigendecomposition of $\mathbf{M}_g^{-1}\mathbf{M}_{g'}$ to recover the a_i up to scaling, then solve a linear system to recover the scales (so long as the a_i are linearly independent the solution will be unique).

3 Running Jennrich on "lifted" tensors.

Recall that we are interested in overcomplete third-order tensor decomposition (say $n \gg d$), so Jennrich's algorithm does not work as-is. Suppose however that $n \leqslant d^2$, Then though the a_i are not linearly independent, the vectors $a_i^{\otimes 2} \in \mathbb{R}^{d^2}$ very well may be (for example, when the a_i are chosen independently from $\mathcal{N}(0, \frac{1}{d}\mathbb{1})$ or uniformly from $\mathbb{S}^{d-1}$, the $a_i^{\otimes 2}$ are linearly independent with probability 1).

Now, if we also had access to the sixth-order tensor in the components $\sum_{i\in[n]} a_i^{\otimes 6}$, then thinking of this as a 3-tensor in the components $a_i^{\otimes 2} \in \mathbb{R}^{d^2}$, we could form $\sum_{i\in[n]} \left\langle a_i^{\otimes 2}, g\right\rangle \left(a_i^{\otimes 2}\right)\left(a_i^{\otimes 2}\right)^\top$ via a random contraction and use it to recover the $\{a_i^{\otimes 2}\}_{i\in[n]}$ as before.

In the learning applications we mentioned above, often times it is possible to obtain access to the order-6 tensor by estimating the order-6 moments; however as mentioned above, estimating higher-moment tensors is sample-intensive. So we would like a way to *lift* the third-order tensor to a sixth-order tensor, without needing any additional samples.

The following sections we will use sum-of-squares to "lift" a third-order tensor $\mathbf{T}$ to a surrogate for the order-6 tensor. We will quantify when we can treat a "lifted" higher-order tensor as the true higher-order tensor we want (the "signal") plus additional information which will hopefully cancel for random instances (the "noise"). For this, it will be helpful to see a variation on Jennrich's algorithm which succeeds when the tensor has the structure of being a rank-1 component in the presence of noise:

Noisy rank-1 tensor. We can generalize the above to the setting where we observe a rank-1 tensor with the addition of some "well-behaved" noise. That is, we have $\mathbf{T} = a^{\otimes 3} + \mathbf{E}$, for $\mathbf{E} \in (\mathbb{R}^d)^{\otimes 3}$ and $a \in \mathbb{R}^d$ (say that we have scaled $\mathbf{T}$ so that a is a unit vector). We will show that in this case, the top eigenvector of a random contraction of $\mathbf{T}$ is correlated with a with not-too-small a probability.

Lemma 3.1. *If we are given $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 3}$ such that $\mathbf{T} = a^{\otimes 3} + \mathbf{E}$ for $a \in \mathbb{R}^d$ a unit vector and $\mathbf{E}$ a symmetric 3-tensor whose $d^2 \times d$ reshaping E has $\|E\|_{op} \leqslant \lambda$ and $Ea = 0$,[1] then for any $\varepsilon > 0$, with probability at least $\Omega(d^{-\alpha^2})$ over $g \sim \mathcal{N}(0, \mathbb{1}_d)$, Furthermore,*

$$\mathbf{M}_g = \rho \cdot aa^\top + N,$$

for $N \in \mathbb{R}^{d\times d}$ with $\rho \geqslant \frac{\alpha}{\lambda}(1 - o(1)) \cdot \|N\|_{op}$, so that if $\alpha \geqslant (1 + o(1))\lambda$, a is the top eigenvector of $\mathbf{M}_g$.[2]

[1] This condition can in fact be relaxed, but it will significantly simplify the proof.

[2] If we loosen the requirement that $Ea = 0$, then a will instead be closely correlated to the top eigenvector.

Before proving the lemma, let us make a couple of observations. Firstly, with this lemma in hand, we have an algorithm for producing a list of vectors which, with high probability, includes a: we simply try $O(d^{(1+\varepsilon)\lambda})$ independent random contractions of $\mathbf{T}$, and find the top eigenvector of each. If we had a procedure for testing whether a vector v is indeed close to a, this would amount to a decomposition algorithm.

Secondly, let us foreshadow how this lemma will be used in the context of sum-of-squares. Given an overcomplete 3-tensor $\mathbf{T}$, we will write a properly constrained sum-of-squares program over indeterminates $u \in \mathbb{R}^d$ and solve for a pseudoexpectation operator $\tilde{\mathbf{E}} : \mathbb{R}[u]^{\leqslant 10} \to \mathbb{R}$ which will satisfy that the order-6 pseudomoment tensor $\tilde{\mathbf{T}}_6 = \tilde{\mathbf{E}} u^{\otimes 6}$ will in fact be of the form $\tilde{\mathbf{T}}_6 = \sum_{i=1}^m a_i^{\otimes 6} + \mathbf{E}$, for $\mathbf{E}$ a 6-tensor whose $d^4 \times d^2$ reshapings have operator norm at most 1. Then, we will be able to run Jennrich's algorithm poly(d) times as a *rounding algorithm* on $\tilde{\mathbf{T}}_6$ to produce a list of vectors which contains a vector close to $a_i^{\otimes 2}$ for each $i \in [n]$, after which we will be able to design a testing procedure to find the true components a_i.

Proof of Lemma 3.1. We notice that

$$\mathbf{M}_g = \sum_{i=1}^d g_i \cdot \mathbf{T}_i = \sum_{i=1}^d g_i \cdot (a(i) \cdot aa^\top + \mathbf{E}_i) = \langle g, a \rangle \cdot aa^\top + \sum_{i=1}^d g_i \mathbf{E}_i.$$

Now, we can split $g = \langle a, g \rangle a + h$ for $\langle h, a \rangle = 0$. Notice that in fact $\sum_{i=1}^d g_i \mathbf{E}_i = \sum_{i=1}^d h_i \mathbf{E}_i$, by the symmetry of $\mathbf{E}$ (as $\mathbf{E}[g, \cdot, \cdot]$ is simply a reshaping of $\mathbf{E}[\cdot, \cdot, g] = Eg$, and then applying linearity). So by the independence of h and $\langle g, a \rangle$, the quantities $\rho = \langle g, a \rangle$ and $\| \sum_{i=1}^d g_i \mathbf{E}_i \|_{op}$ are independent as well.[3]

Using a concentration inequality for Gaussian matrix series (see e.g. Theorem 1.2 in [Tro12]), we have the following:

Claim 3.2.

$$\mathbf{Pr}_{g \sim \mathcal{N}(0, \mathbb{1})} \left[\left\| \sum_{i=1}^d g_i \mathbf{E}_i \right\|_{op} \geqslant \lambda \sqrt{2(1+\delta) \log d} \right] \leqslant d^{-\delta}.$$

Furthermore, one can show that $\langle g, a \rangle \sim \mathcal{N}(0, 1)$. So using a Gaussian anticoncentration statement, we have that $\mathbf{Pr}_{g \sim \mathcal{N}(0, \mathbb{1})}[\langle g, a \rangle \geqslant \sqrt{2(1+\varepsilon) \log d}] \geqslant \Omega(d^{-(1+\varepsilon)})$.

Hence, with probability at least $\Omega(d^{-\alpha^2})$, we have that $N = \sum_{i=1}^d g_i \mathbf{E}_i$ has $\|N\|_{op} \leqslant \lambda \sqrt{2(1 + \log d)}$ and $\rho \geqslant \alpha \sqrt{2 \log d}$ hold simultaneously. This gives our first conclusion. For our second conclusion, we simply notice that a is in the null space of N by our assumption that $Ea = 0$. □

4 Rounding with Jennrich and pseudodistributions

The goal of this section is to answer the following question:

> *When is it possible to design a Jennrich-based rounding algorithm, when we only have access to a pseudoexpectation operator in lieu of "true" moments?*

In the previous section we saw that it would be helpful if we had access to the true sixth-moment tensor of the $\{a_i\}_{i \in [n]}$. In its place, we will only be able to obtain a pseudoexpectation operator subject to helpful polynomial constraints which will allow us to apply Jennrich's algorithm in the style of Lemma 3.1.

[3] In the case when we do not have independence but instead have a bound on, say, the norm $\|Ea\|$, one can make a more careful argument accounting for this, at the cost of some error terms.

Throughout this section we will be concerned with general conditions under which SoS can be used to decompose an undercomplete tensor with components $\{b_i\}_{i\in[n]}$.[4] Ultimately, we will use our SOS framework to recover the $\{b_i\}_{i\in[n]}$ via simulating higher moments.

We will prove Theorem 4.1 (based on Theorem 4.1 of [MSS16]), which is an SoS version of a variant of Lemma 3.1 (in the setting where $\lambda < 1$ is small and without the stipulation that $Ea = 0$). In particular, Theorem 4.1 shows that as long as $\tilde{\mathbf{E}}$ is a degree-$O(k)$ pseudoexpectation operator over a vector variable $u \in \mathbb{R}^d$, satisfying for some unit vector b,

$$\|u\|_2^2 \leqslant 1 \text{ and } \tilde{\mathbf{E}}\left[\langle b, u\rangle^{k+2}\right] = \Omega\left(\frac{1}{\varepsilon\sqrt{k}}\right)\left\|\tilde{\mathbf{E}}\left[uu^\top\right]\right\|_{\text{op}}, \tag{1}$$

then with reasonable probability the top eigenvector of a contracted matrix (a random Gaussian contraction of the order-$k+2$ pseudoexpectation) is $1 - O(\varepsilon)$ correlated with b.

Notice that if $b_1, \dots, b_n \in \mathbb{R}^d$ are unit vectors which are isotropic ($\mathbf{E}_{i\sim[n]} b_i b_i^\top \preceq \mathbb{1}$) and near-orthogonal so that $|\langle b_i, b_j\rangle| \leqslant \delta$ if $i \neq j$,[5] then taking $\tilde{\mathbf{E}}u^{\otimes k+2}$ to be the uniform mixture tensor $\mathbf{E}_{i\sim[n]} b_i^{\otimes k+2}$, we have that $\tilde{\mathbf{E}}uu^\top \preceq \frac{1}{n}\mathbb{1}$, and

$$\mathbf{E}_{i\in[n]}\tilde{\mathbf{E}}[\langle u, b_i\rangle^{k+2}] = \mathbf{E}_{i,j\sim[n]}\langle b_i, b_j\rangle^{k+2} = \frac{1}{n} \pm O(n\delta^{k+2}).$$

In particular, if $\delta^{k+2}n \ll \frac{1}{n}$, then taking $\tilde{\mathbf{E}}$ to be consistent with the moments of the uniform mixture over the b_i, by an averaging argument there must exist some $\ell \in [n]$ such that $b = b_\ell$ satisfies (1) so long as $k = \Omega(\frac{1}{\varepsilon^2})$. So we will later use SoS to find a proxy for this moment tensor.

Theorem 4.1 (SoS version of Lemma 3.1)**.** *Let $\tilde{\mathbf{E}}$ be a degree-$O(k)$ pseudoexpectation satisfying the conditions (1). Then with probability $d^{-O(k)}$ over the choice of the Gaussian vector g, the top eigenvector v of $\mathbf{M}_g$ satisfies $\langle b, v\rangle^2 = 1 - O(\varepsilon)$.*

Proof. For any vector $v \in (\mathbb{R}^d)^{\otimes k}$, we will use the notation $\mathbf{M}_v := \tilde{\mathbf{E}}[\langle v, u^{\otimes k}\rangle \cdot uu^\top]$ throughout.

The Davis-Kahan theorem says that if $\mathbf{M}$ and unit vector b satisfy $\left\|\mathbf{M} - bb^\top\right\|_{\text{op}} \leqslant \varepsilon$, then their top eigenvectors have correlation $1 - O(\varepsilon)$. Applying this theorem, we wish to show with probability $d^{-O(k)}$ over $g \sim \mathcal{N}(0, \mathbb{1}_{d^k})$, there is a multiple of $\mathbf{M}_g$ at most ε away in operator norm from $bb^\top$. It will be convenient for us to write an orthogonal decomposition of $\mathbf{M}_g$ by projecting g against $b^{\otimes k}$, via

$$\mathbf{M}_g = \rho\mathbf{M}_{b^{\otimes k}} + \mathbf{M}_h, \text{ for } h := g - \left\langle g, b^{\otimes k}\right\rangle b^{\otimes k} \text{ and } \rho := \left\langle g, b^{\otimes k}\right\rangle. \tag{2}$$

Using Gaussian anticoncentration, with probability at least $d^{-O(k)}$, we have $\rho = \left\langle g, b^{\otimes k}\right\rangle \geqslant \sqrt{k\log d}$. Conditioning on this event, and letting $t = \tilde{\mathbf{E}}\left[\langle b, u\rangle^{k+2}\right]$,

$$\begin{aligned}\mathbf{E}_g\left[\left\|\frac{1}{\rho t}\mathbf{M}_g - bb^\top\right\|_{\text{op}}\middle|\rho \geqslant \sqrt{k\log d}\right] &\leqslant \left\|\frac{1}{t}\mathbf{M}_{b^{\otimes k}} - bb^\top\right\|_{\text{op}} + \mathbf{E}_g\left[\frac{1}{\rho t}\|\mathbf{M}_h\|_{\text{op}}\middle|\left\langle g, b^{\otimes k}\right\rangle \geqslant \sqrt{k\log d}\right], \\ &\leqslant \left\|\frac{1}{t}\mathbf{M}_{b^{\otimes k}} - bb^\top\right\|_{\text{op}} + \frac{1}{t\sqrt{k\log d}}\mathbf{E}_g\left[\|\mathbf{M}_h\|_{\text{op}}\right], \end{aligned} \tag{3}$$

where we have used that h is independent of ρ.

Now, we will bound each term with a separate claim. We begin with the first term:

[4] It may be useful to keep in mind that these b_i vectors which we wish to recover will be a polynomial transformation of the original component vectors $\{a_i\}_{i\in[n]}$ in the tensor decomposition, namely $b_i = a_i^{\otimes 2}$ for all $i \in [n]$.

[5] This is satisfied by random unit vectors with $\delta = O(\frac{\log n}{\sqrt{d}})$ with high probability.

Claim 4.2. For $t = \tilde{\mathbf{E}}\left[\langle b, u\rangle^{k+2}\right]$,

$$\left\|\mathbf{M}_{b^{\otimes k}} - tbb^\top\right\|_{\mathrm{op}} = O(\varepsilon t).$$

Proof. Let $\mathbf{\Pi} = bb^\top$ be the orthogonal projection onto the span of b; note $\mathbf{\Pi}\mathbf{M}_{b^{\otimes k}}\mathbf{\Pi} = tbb^\top$. Since $t\Pi = \Pi\mathbf{M}_{b^{\otimes k}}\Pi$, it suffices for us to show that the operator norm of the left- and right-projections to the subspace $(\mathbb{1} - \Pi)$ are small. That is,

$$\begin{aligned}\left\|\mathbf{M}_{b^{\otimes k}} - t\Pi\right\|_{\mathrm{op}} &\leqslant 2\left\|(\mathbb{1}-\Pi)\mathbf{M}_{b^{\otimes k}}\Pi\right\|_{\mathrm{op}} + \left\|(\mathbb{1}-\Pi)\mathbf{M}_{b^{\otimes k}}\Pi(\mathbb{1}-\Pi)\right\|_{\mathrm{op}} \\ &\leqslant 2\left\|\mathbf{\Pi}\mathbf{M}_{b^{\otimes k}}\mathbf{\Pi}\right\|_{\mathrm{op}}^{\frac{1}{2}}\left\|(\mathbb{1}-\mathbf{\Pi})\mathbf{M}_{b^{\otimes k}}(\mathbb{1}-\mathbf{\Pi})\right\|_{\mathrm{op}}^{\frac{1}{2}} + \left\|(\mathbb{1}-\mathbf{\Pi})\mathbf{M}_{b^{\otimes k}}(\mathbb{1}-\mathbf{\Pi})\right\|_{\mathrm{op}},\end{aligned} \tag{4}$$

Where we have used the triangle inequality in the first line and in the second line the Cauchy-Schwarz inequality in $\mathbf{M}_{b^{\otimes k}}$ which is positive semidefinite for even k. Next, we bound the operator norm of $\mathbf{M}_{b^{\otimes k}}$ outside of $\mathbf{\Pi}$:

$$\begin{aligned}\left\|(\mathbb{1}-\mathbf{\Pi})\mathbf{M}_{b^{\otimes k}}(\mathbb{1}-\mathbf{\Pi})\right\|_{\mathrm{op}} &= \left\|\tilde{\mathbf{E}}\left[\langle b, u\rangle^k(\mathbb{1}-\mathbf{\Pi})uu^\top(\mathbb{1}-\mathbf{\Pi})\right]\right\|_{\mathrm{op}} \\ &\leqslant \left\|\tilde{\mathbf{E}}\left[\langle b, u\rangle^k\left(1-\langle b, u\rangle^2\right)\right]\right\|_{\mathrm{op}} \\ &\leqslant \frac{2}{k-2}\tilde{\mathbf{E}}\left[\langle b, u\rangle^2\right] \leqslant \frac{2}{k-2}\left\|\tilde{\mathbf{E}}\left[uu^\top\right]\right\|_{\mathrm{op}}.\end{aligned}$$

Here, we used that $\|u\|^2 \leqslant 1$ is a polynomial constraint, and the inequality $x^{k-2}(1-x^2) \leqslant \frac{2}{k-2}$, which has a degree-$O(k)$ SOS proof. Furthermore, observe that $\mathbf{\Pi}\mathbf{M}_{b^{\otimes k}}\mathbf{\Pi} \preceq \mathbf{M}_{b^{\otimes k}} \preceq \tilde{\mathbf{E}}\left[uu^\top\right]$, and all steps are certifiable by degree-$O(k)$ SOS. Going back to (4),

$$\left\|\mathbf{M}_{b^{\otimes k}} - tbb^\top\right\|_{\mathrm{op}} \leqslant \left(\frac{2}{k-2} + 2\sqrt{\frac{2}{k-2}}\right)\left\|\tilde{\mathbf{E}}\left[uu^\top\right]\right\|_{\mathrm{op}} = O(\varepsilon t).$$

In the last step we used the assumption (1). □

Now, to bound the second term, we show that $\|\mathbf{M}_h\|_{op}$ is not too large in expectation.

Claim 4.3. For $t = \tilde{\mathbf{E}}\left[\langle b, u\rangle^{k+2}\right]$,

$$\mathbf{E}_g\left[\|\mathbf{M}_h\|_{\mathrm{op}}\right] = O\left(\varepsilon t\sqrt{k\log d}\right).$$

Proof. For any pseudoexpectation respecting $\|u\|_2^2 \leqslant 1$,

$$\mathbf{E}_g\left[\|\mathbf{M}_h\|_{\mathrm{op}}\right] = \mathbf{E}_g\left[\sum_{j=1}^{d^k} h_j \cdot (u^{\otimes k})_j \cdot uu^\top\right] = O\left(\sqrt{k\log d}\right)\left\|\tilde{\mathbf{E}}\left[u^{\otimes k+1}u^\top\right]\right\|_{\mathrm{op}} = O\left(\sqrt{k\log d}\right)\left\|\tilde{\mathbf{E}}\left[uu^\top\right]\right\|_{\mathrm{op}},$$

Where the first equality again uses concentration of Gaussian matrix series ([Tro12], Theorem 1.2),[6] and the second equality uses that $\|\tilde{\mathbf{E}}u^{\otimes k+1}u^\top\|_{op} \leqslant \|\tilde{\mathbf{E}}uu^\top\|_{op}$ for any degree-$4(k+2)$ pseudoexpectation operator (proof can be found in [MSS16], Theorem 6.1). The conclusion is immediate upon applying the condition (1). □

Plugging Claims 4.2 and 4.3 into (3) and applying Markov's inequality concludes the proof. □

[6] Because h is not a standard normal vector one has to be careful here; however, one can show that using h in place of g actually improves the bound.

5 SOS framework for tensor decomposition

We are now ready to give an algorithm which uses Theorem 4.1 in an SOS framework for tensor decomposition. Here we will show that as long as our SoS relaxation satisfies the constraint[7]

$$\sum_{i=1}^{n} \langle b_i, u \rangle^4 \geqslant 1 - \varepsilon$$

(as well as $\|u\|^2 \leqslant 1$), Jennrich's algorithm will recover a list of vectors, one of which is close to some component b_i. In the tensor decomposition algorithm of Ma Shi and Steurer [MSS16], one can recover approximations to all of the components by additionally using an SoS relaxation to (1) test if each recovered vector is indeed close to some b_i, and (2) after recovering some b_i, one re-solve the SoS SDP with the added condition $\langle b_i, u \rangle^2 < 0.01$ to make sure that the next component which is found is far from b_i.

Now, we can state the algorithm and give its analysis in the following theorem. At a high level, the SoS relaxation will look for an order-k tensor that is close to the moments of the uniform mixture over $b_1 = P(a_1)^{\otimes k}, b_2 = P(a_2)^{\otimes k}, \dots, b_n = P(a_n)^{\otimes k}$, for k a sufficiently large integer. The map P is intended to ensure that $P(a_i)$ are linearly independent and close to orthonormal. It might be helpful to keep in mind the example $P(u) = u^{\otimes 2}$, which as discussed above lifts $n \leqslant d^2$ vectors in d dimensions into d^2 dimensions where they may be linearly independent.[8] Then, Jennrich's algorithm is run on the pseudomoment tensor $\tilde{\mathbf{E}} P(u)^{\otimes k+2}$ to find a vector v which may be correlated with $b_i = P(a_i)$ for one of the a_i.

Theorem 5.1 (SoS Algorithm for recovering one lifted tensor component). *Let $P : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a degree-$O(k)$ lifting polynomial, for $k \ll \frac{1}{\varepsilon}$. For a set of vectors $\{a_i\}_{i \in [n]}$, define for all $i \in [n]$, $b_i = P(a_i)$, and suppose all $\|b_i\|_2 \geqslant 1 - \varepsilon$ and $\left\| \sum_{i \in [n]} b_i b_i^\top \right\|_{\mathrm{op}} \leqslant 1 + \varepsilon$. Finally, let $\mathcal{A}$ be any polynomial system of inequalities which implies*

$$\mathcal{A} \vdash_{O(k)} \left\{ \sum_{i \in [n]} \langle b_i, P(u) \rangle^4 \geqslant (1 - \varepsilon) \|P(u)\|^4 \right\}. \tag{5}$$

Then, consider the following algorithm, taking as input $\mathcal{A}$ and P.

1. *Compute a degree-$O(k)$ pseudoexpectation $\tilde{\mathbf{E}}$ satisfying the constraints*

$$\mathcal{A} \cup \left\{ \|P(u)\|_2^2 \in [1 - \varepsilon, 1] \right\}, \left\| \tilde{\mathbf{E}} \left[P(u) P(u)^\top \right] \right\|_{\mathrm{op}} \leqslant \frac{1 + \varepsilon}{n}. \tag{6}$$

2. *For $t \in [T]$ with $t = d^{O(k)}$:*

   (a) *Let $g^{(t)} \sim \mathcal{N}(0, \mathbb{1}_d^{\otimes k})$, and compute the top eigenvector v_t of*

$$\tilde{\mathbf{E}} \left[\left\langle g^{(t)}, P(u)^{\otimes k} \right\rangle P(u) P(u)^\top \right].$$

   (b) *Add v^t to the list of potential vectors which correlate with some b_i.*

3. *Output the list $\{v^t\}_{t \in [T]}$.*

[7] Notice that we may not know the entries of $\sum_{i=1}^n b_i^{\otimes 4}$, so we will later have to make sure that this constraint is implied by constraints that we can add.

[8] However, sometimes $\sum (a_i^{\otimes 2})(a_i^{\otimes 2})$ may have a large operator norm, in which case it is helpful to make P more complicated (e.g. $P(u) = \Pi u^{\otimes 2}$ for a matrix Π that projects away from the span of large eigenvectors).

This algorithm succeeds with high probability, and runs in time $d^{O(k)}$. It outputs a set of vectors $\{v^t\}_{t\in[T]} \subset \mathbb{R}^{d'}$, such there exists some $i \in [n]$ and $t \in [T]$ such that

$$\left\|b_i^{\otimes 2} - v^t\right\|_2^2 = O\left(\frac{1}{\sqrt{k}}\right).$$

Remark 5.2. The constraint $\|\tilde{\mathbf{E}}P(u)P(u)^\top\|_{op} \leqslant \frac{1+\varepsilon}{n}$ is an *entropy constraint* which encourages $\tilde{\mathbf{E}}$ to behave like the uniform distribution over the b_i, and enforces the properties that we will need for Jennrich rounding to succeed. Notice that this is not a polynomial constraint in u; however it is convex so we can still add it to our convex relaxation and solve in polynomial time.

Remark 5.3. The error guarantee stated here improves as k is taken larger. However, one can often get away with choosing k small, as in many cases one can use vectors which are only $\Omega(1)$-correlated with a component as a warm start for gradient descent or a similar algorithm, and obtain improved guarantees (see e.g. [AGJ14]).

Proof. First, notice that the uniform distribution on $\{b_j\}_{j\leqslant n}$ satisfies all constraints given by step 1.

We will require the following technical lemma from prior work [BKS15a], which we state here without proof. However, the proof follows from repeating the arguments we will make in Section 6.1 to higher powers, so we omit the proof of the more general claim.

Lemma 5.4. *Let $b_1, \dots, b_n$ be unit vectors with $\left\|\sum_{i\in[n]} b_i b_i^\top\right\|_{\text{op}} \leqslant 1 + \varepsilon$. Then for all even k, there is a degree-$O(k)$ SOS proof in the variable x that*

$$\left\{\|x\|_2^2 \leqslant 1, \sum_{i\in[n]} \langle b_i, x\rangle^4 \geqslant 1 - \varepsilon\right\} \vdash \left\{\sum_{i\in[n]} \langle b_i, x\rangle^k \geqslant 1 - O(k\varepsilon)\right\}.$$

Now, by averaging there exists some remaining index i satisfying

$$\tilde{\mathbf{E}}\left[\langle b_i, P(u)\rangle^k\right] \geqslant \frac{1-\varepsilon}{n} \geqslant \frac{1 - O(k\varepsilon)}{(1+\varepsilon)} \left\|\tilde{\mathbf{E}}\left[P(u)P(u)^\top\right]\right\|_{\text{op}}.$$

Applying Theorem 4.1 now implies the conclusion with the desired probability, since the conditions (1) are satisfied (where we have used that $O(k\varepsilon) \ll 1$ and have taken the parameter ε from Theorem 4.1 as $\Theta(\frac{1}{\sqrt{k}})$). □

6 Overcomplete 3rd-order tensor decomposition

Here, we specialize to the case when $\mathbf{T} = \sum_{i=1}^n a_i^{\otimes 3}$ for $a_i \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1})$, and we show that the algorithm outlined above succeeds when $n \ll d^{1.5}$.

The overcomplete 3rd-order tensor decomposition algorithm of [MSS16] prescribes that we apply Theorem 5.1 with an appropriate choice of polynomial constraints $\mathcal{A}$ and an appropriate polynomial P. However, note that we cannot directly encode a polynomial constraint of the form

$$\sum_{i\in[n]} \langle b_i, P(u)\rangle^4 \geqslant 1 - \varepsilon$$

into our SOS algorithm, because we do not have explicit access to the $\{b_i = P(a_i)\}_{i\in[n]}$. So, we need to find a polynomial in the explicit tensor $\mathbf{T}$ that we have access to which implies this constraint. Fortunately, the natural objective of maximizing the *injective norm* $\mathbf{T}[u,u,u]$ for $\|u\|_2^2 \leqslant 1$ turns out to be such an explicit polynomial constraint.

6.1 Maximizing the injective norm

This section is based on [GM15]. We present an observation that $\|\mathbf{T}\|_{\text{inj}} := \sup_{\|v\|_2=1} \mathbf{T}[v,v,v]$ is maximized when v has large correlation with some a_i as long as $n \ll d^{1.5}$. As we will see, this will imply that if we add the constraint $\mathbf{T}[u,u,u] \geqslant 1-\varepsilon$ to our SoS relaxation, this will imply that $\sum_{i=1}^n \langle b_i, u^{\otimes 2}\rangle \geqslant 1-\varepsilon'$. Observe

$$\begin{aligned}(\mathbf{T}[u,u,u])^2 &= \left(\sum_{i\in[n]} \langle a_i, u\rangle^3\right)^2 = \left(\left\langle \sum_{i\in[n]} \langle a_i, u\rangle^2 a_i, u \right\rangle\right)^2 \\ &\leqslant \left\|\sum_{i\in[n]} \langle a_i, u\rangle^2 a_i\right\|_2^2 \|u\|_2^2 \\ &= \sum_{i\in[n]} \langle a_i, u\rangle^4 \|a_i\|_2^2 + \sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle \langle a_i, u\rangle^2 \left\langle a_j, u\right\rangle^2 \text{ for all } \|u\|_2^2 = 1.\end{aligned} \tag{7}$$

We bound the two terms in (7) in two different ways. Roughly speaking, the first term is reduced by similar Cauchy-Schwarz arguments as have already been used to understanding the operator norm of a degree-6 matrix in the components, and the operator norm of $\sum_{i\in[n]} a_i a_i^\top$, both of which follow straightforwardly from concentration. The second term is captured by nested applications of matrix Bernstein after symmetrization.

6.1.1 2-to-4 norm

We consider the first term in (7), namely $\sum_{i\in[n]} \langle a_i, u\rangle^4 \|a_i\|_2^2$.

Lemma 6.1. *There is a degree-$O(1)$ SoS proof in the variables $u \in \mathbb{R}^d$ that*

$$\sum_{i\in[n]} \langle a_i, u\rangle^4 \|a_i\|_2^2 \leqslant 1 + \widetilde{O}\left(\frac{1}{\sqrt{d}} + \frac{n^2}{d^3}\right).$$

Proof. We use the following assumption, which can be shown via standard concentration arguments to hold with high probability for $a_i \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1})$.

Assumption 6.2. The vectors $\{a_i\}_{i\in[n]}$ satisfy the following bounds.

1. $\left\langle a_i, a_j\right\rangle^2 \leqslant \widetilde{O}\left(\frac{1}{d}\right)$ for all $i \neq j \in [n]$.
2. $\|a_i\|_2^2 \in 1 \pm \widetilde{O}\left(\frac{1}{\sqrt{d}}\right)$ for all $i \in [n]$.

To bound the first term of (7), we use the same sequence of Cauchy-Schwarz inequalities:

$$\begin{aligned}\left(\sum_{i\in[n]} \langle a_i, u\rangle^4\right)^2 &= \left(\left\langle \sum_{i\in[n]} \langle a_i, u\rangle^3 a_i, u\right\rangle\right)^2 \leqslant \left\|\sum_{i\in[n]} \langle a_i, u\rangle^3 a_i\right\|_2^2 \\ &= \sum_{i\in[n]} \langle a_i, u\rangle^6 \|a_i\|_2^2 + \sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle \langle a_i, u\rangle^3 \left\langle a_j, u\right\rangle^3.\end{aligned} \tag{8}$$

Let $\mathbf{A}_1$ be the matrix whose rows are the $\{a_i^{\otimes 3}\}_{i\in[n]}$. Then letting $y = u^{\otimes 3}$ we have $\|\mathbf{A}_1\|_2^2 = \sum_{i\in[n]} \langle a_i, u\rangle^6$. It suffices to bound the operator norm of $\mathbf{A}_1^\top \mathbf{A}_1$; note that it has entries

$$[\mathbf{A}_1]_{ij} = \begin{cases} \|a_i\|_2^6 & i = j \\ \left\langle a_i, a_j\right\rangle^3 & i \neq j \end{cases}.$$

By Assumption 6.2 and the Gershgorin disk theorem, the operator norm is bounded by $1 + \widetilde{O}\left(\frac{n}{d^{1.5}}\right)$. Note that the Gershgorin strategy is only able to capture this fact with degree 6 polynomials. Next, we bound the second term of (8):

$$\left(\sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle \langle a_i, u\rangle^3 \left\langle a_j, u\right\rangle^3\right)^2 \leqslant \left(\sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle^2 \langle a_i, u\rangle^2 \left\langle a_j, u\right\rangle^2\right)\left(\sum_{i\neq j\in[n]} \langle a_i, u\rangle^4 \left\langle a_j, u\right\rangle^4\right)$$
$$\leqslant \widetilde{O}\left(\frac{1}{d}\right)\left(\sum_{i\in[n]} \langle a_i, u\rangle^2\right)^2\left(\sum_{i\in[n]} \langle a_i, u\rangle^4\right)^2 \leqslant \widetilde{O}\left(\frac{n^2}{d^3}\right).$$

Here, we used that the random matrix $\sum_{i\in[n]} a_i a_i^\top$ has operator norm $\widetilde{O}\left(\frac{n}{d}\right)$ with high probability. This concludes the proof. □

6.1.2 Cross terms

Lemma 6.3. *There is a degree-$O(1)$ SoS proof in the variabes $u \in \mathbb{R}^d$ that*

$$\sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle \langle a_i, u\rangle^2 \left\langle a_j, u\right\rangle^2 \leqslant \widetilde{O}\left(\frac{n}{d^{1.5}}\right).$$

Proof. To handle the cross terms of (7), namely $\sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle \langle a_i, u\rangle^2 \left\langle a_j, u\right\rangle^2$, we observe that this is the quadratic form of $u^{\otimes 2}$ in the matrix $\mathbf{A}_2 := \sum_{i\neq j\in[n]} \left\langle a_i, a_j\right\rangle (a_i \otimes a_j)(a_i \otimes a_j)^\top$, so it suffices to bound the operator norm of $\mathbf{A}_2$. Letting $\{\sigma_i\}_{i\in[n]}$ be random signs, we have (in distribution) that

$$\mathbf{A}_2 = \sum_{i\neq j\in[n]} \sigma_i\sigma_j \left\langle a_i, a_j\right\rangle (a_i \otimes a_j)(a_i \otimes a_j)^\top.$$

Letting $\{\sigma_i'\}_{i\in[n]}$ be independent random signs, Theorem 1 of [PMS95] says the concentration of $\mathbf{A}_2$ in any norm is up to constants the same as concentration of

$$\mathbf{A}_2' = \sum_{i\neq j\in[n]} \sigma_i\sigma_j' \left\langle a_i, a_j\right\rangle (a_i \otimes a_j)(a_i \otimes a_j)^\top = \sum_{i\in[n]} \sigma_i \left(\sum_{j\in[n],j\neq i} \sigma_j' \left\langle a_i, a_j\right\rangle (a_i \otimes a_j)(a_i \otimes a_j)^\top\right).$$

We use this symmetrization trick so we can apply matrix Bernstein twice to $\mathbf{A}_2'$ to bound its operator norm. Consider one term of $\mathbf{A}_2'$, which can be written as $(\sigma_i a_i a_i^\top) \otimes \left(\sum_{j\in[n],j\neq i} \sigma_j' \left\langle a_i, a_j\right\rangle a_j a_j^\top\right)$. We apply matrix Bernstein to $\sum_{j\in[n],j\neq i} \sigma_j' \left\langle a_i, a_j\right\rangle a_j a_j^\top$, which requires an operator norm bound on individual terms and a second moment bound. Under Assumption 6.2, each $\sigma_j' \left\langle a_i, a_j\right\rangle a_j a_j^\top$ is operator norm bounded by $\widetilde{O}(\frac{1}{\sqrt{d}})$, and

$$\sum_{j\in[n],j\neq i} \left\langle a_i, a_j\right\rangle^2 \left\|a_j\right\|_2^2 a_j a_j^\top \preceq \widetilde{O}\left(\frac{n}{d^2}\right)\mathbb{1}.$$

Thus, we conclude with high probability that each

$$(a_i a_i^\top) \otimes \left(\sum_{j\in[n],j\neq i} \sigma_j' \left\langle a_i, a_j\right\rangle a_j a_j^\top\right) \preceq \widetilde{O}\left(\frac{\sqrt{n}}{d}\right)(a_i a_i^\top) \otimes \mathbb{1}.$$

Next, we use matrix Bernstein once more on the random sum $\sum_{i\in[n]} \sigma_i a_i a_i^\top$. Note that the termwise operator norm bound is tightly concentrated near 1 and the second moment bound is the operator norm bound of $\sum_{i\in[n]} a_i a_i^\top$ which is $\widetilde{O}(\frac{n}{d})$. Hence, the random sum $\sum_{i\in[n]} \sigma_i a_i a_i^\top$ is bounded by $\widetilde{O}\left(\sqrt{\frac{n}{d}}\right)$. Combining these, we have the following conclusion. □

Putting together Lemma 6.1 and Lemma 6.3 implies that the injective norm is maximized near the $\{a_i\}_{i\in[n]}$. As an aside, carefully going through the proofs and using that the left hand side of (7) is bounded from below by $1-\varepsilon$, as well as the expanding of terms done in (8), implies a lower bound on $\sum_{i\in[n]}\langle a_i, x\rangle^6 \|a_i\|_2^2$ as well. We state without proof that applying the same strategy of this section "once more" to lift to a higher power (e.g. 8 rather than 6) yields the following result.

Proposition 6.4. *Let* $\mathbf{T} = \sum_{i\in[n]} a_i^{\otimes 3}$ *for* $a_i \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1})$ *independently. Let* $\mathcal{A} = \{\mathbf{T}(u,u,u) \geqslant 1-\varepsilon, \|u\|_2^2 = 1\}$. *Then, we have for* $\delta = \widetilde{O}(\frac{n}{d^{1.5}})$,

$$\mathcal{A} \vdash \left\{ \sum_{i\in[n]} \left\langle a_i^{\otimes 2}, u^{\otimes 2} \right\rangle^4 \geqslant 1 - O(\varepsilon) - \delta. \right\}$$

6.2 Rounding from maximizing the injective norm

Throughout this section, let $\delta = \widetilde{O}(\frac{n}{d^{1.5}})$; we should think δ is smaller than some constant. We also define the following projection operator $\mathbf{\Pi} = \Phi\Phi^\top$, where $\Phi = \frac{1}{\sqrt{d}} \sum_{j\in[d]} e_i^{\otimes 2} \in (\mathbb{R}^d)^{\otimes 2}$. We will apply Theorem 5.1 as our tensor decomposition algorithm, with the constraint set $\mathcal{A}$ as in Proposition 6.4 and the polynomial transform P defined as

$$P(a) = (\mathbb{1}_{d^2} - \mathbf{\Pi})\, a^{\otimes 2}.$$

For all intents and purposes the reader should just think $P(a) = a^{\otimes 2}$, but for technical reasons relating to Gaussian concentration we need to project out a particular subspace via the operator $\mathbf{\Pi}$. In order to apply Theorem 5.1, it suffices to show that the conditions of the algorithm hold, namely that (5) holds, all $P(a_i)$ have sufficiently large norm, and the operator norm of $\sum_{i\in[n]} P(a_i)P(a_i)^\top$ is bounded. The latter two claims follow straightforwardly by Gaussian concentration, and we omit them for simplicity. To see the first claim, we have by Proposition 6.4

$$\begin{aligned}
\mathcal{A} \vdash \sum_{i\in[n]} \langle P(u), P(a_i)\rangle^4 &\geqslant \sum_{i\in[n]} \left(\left\langle u^{\otimes 2}, a_i^{\otimes 2} \right\rangle - \frac{1}{d} \right)^4 \\
&\geqslant \left(1 - \frac{1}{d}\right) \sum_{i\in[n]} \langle u, a_i\rangle^8 - O\left(\frac{1}{d}\right) \geqslant 1 - O(\varepsilon) - \delta.
\end{aligned}$$

The first inequality used Gaussian concentration, the second used that $(x-y^2)^4 \geqslant (1-y^2)x^4 - O(y^2)$ has a low-degree SOS proof, and the third used Proposition 6.4. Finally, since all the conditions of Theorem 5.1 are met using the discussion of this section, we can recover the components to sufficiently small constant accuracy as desired.

7 Conclusion

We conclude with some bibliographic remarks. These notes primarily focus on the algorithm of Ma, Shi, and Steurer [MSS16]. The perspective in section 3 is also articulated and used in [SS17]. Jennrich's algorithm has been discovered and re-discovered many times; often people cite Harshman [H+70] as the first occurrence in the literature. For more details on applications of tensor decomposition to parameter estimation problems, we suggest for example [MR05, BKS15b, AGH+14].

Contact. Comments are welcome at tselil@stanford.edu.

References

[AGH+14] A Anandkumar, R Ge, D Hsu, SM Kakade, and M Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014. 11

[AGJ14] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *CoRR*, abs/1402.5180, 2014. 8

[BKS15a] Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 143–151, 2015. 8

[BKS15b] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015. 11

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*, pages 829–849, 2015. 9

[H+70] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970. 11

[MR05] Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375, 2005. 11

[MSS16] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 438–446, 2016. 5, 6, 7, 8, 11

[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *Conference on Learning Theory*, pages 1760–1793. PMLR, 2017. 11

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012. 4, 6