Proceedings of Machine Learning Research vol 65:1–34, 2017

Fast and robust tensor decomposition with applications to dictionary learning

Tselil Schramm TSCHRAMM@CS.BERKELEY.EDU
UC Berkeley

David Steurer DSTEURER@CS.CORNELL.EDU
Cornell and IAS

Abstract

We develop fast spectral algorithms for tensor decomposition that match the robustness guarantees of the best known polynomial-time algorithms for this problem based on the sum-of-squares (SOS) semidefinite programming hierarchy.

Our algorithms can decompose a 4-tensor with n-dimensional orthonormal components in the presence of error with constant spectral norm (when viewed as an n^2-by-n^2 matrix). The running time is n^5 which is close to linear in the input size n^4.

We also obtain algorithms with similar running time to learn sparsely-used orthogonal dictionaries even when feature representations have constant relative sparsity and non-independent coordinates.

The only previous polynomial-time algorithms to solve these problem are based on solving large semidefinite programs. In contrast, our algorithms are easy to implement directly and are based on spectral projections and tensor-mode rearrangements.

Or work is inspired by recent of Hopkins, Schramm, Shi, and Steurer (STOC'16) that shows how fast spectral algorithms can achieve the guarantees of SOS for average-case problems. In this work, we introduce general techniques to capture the guarantees of SOS for worst-case problems.

Keywords: tensor decomposition, dictionary learning, sum of squares, iterative projection method, orthogonal tensor, spectral algorithm.

1. Introduction

Tensor decomposition is the following basic inverse problem: Given a k-th order tensor $T \in (\mathbb{R}^d)^{\otimes k}$ of the form

$$T = \sum_{i=1}^{n} a_i^{\otimes k} + E, \tag{1.1}$$

we aim to approximately recover one or all of the unknown components $a_1, \ldots, a_n \in \mathbb{R}^d$. The goal is to develop algorithms that can solve this problem efficiently under the weakest possible assumptions on the order k, the components $a_1, \ldots, a_n$, and the error E.

Tensor decomposition is studied extensively across many disciplines including machine learning and signal processing. It is a powerful primitive for solving a wide range of other inverse / learning problems, for example: blind source separation / independent component analysis (Lathauwer et al. (2007)), learning phylogenetic trees and hidden Markov models (Mossel and Roch (2005)), mixtures of Gaussians (Hsu and Kakade (2013)), topic models (Anandkumar et al. (2012)), dictionary learning (Barak et al. (2015); Ma et al. (2016)), and noisy-or Bayes nets (Arora et al. (2016)).

A classical algorithm based on simultaneous diagonalization, Harshman (1970); Lathauwer et al. (1996) (often attributed to R. Jennrich) can decompose the input tensor (1.1) when the components are linearly independent, there is no error, and the order of the tensor is at least 3. Current research on algorithms for tensor decomposition aims to improve over the guarantees of this classical algorithm in two important ways:

Overcomplete tensors: What conditions allow us to decompose tensors when components are linearly dependent?

Robust decomposition: What kind of errors can efficient decomposition algorithms tolerate? Can we tolerate errors E with "magnitude" comparable to the low-rank part $\sum_{i=1}^{n} a_i^{\otimes k}$?

The focus of this work is on robustness. There are two ways in which errors arise in applications of tensor decomposition. The first is due to finite samples. For example, in some applications T is the empirical k-th moment of some distribution and the error E accounts for the difference between the empirical moment and actual moment ("population moment"). Errors of this kind can be made smaller at the expense of requiring a larger number of samples from the distribution. Therefore, robustness of decomposition algorithms helps with reducing sample complexity.

Another way in which errors arise is from modeling errors ("systematic errors"). These kinds of errors are more severe because they cannot be reduced by taking larger samples. Two important applications of tensor decomposition with such errors are learning Noisy-or Bayes networks (Arora et al. (2016)) and sparsely-used dictionaries (Barak et al. (2015)). For noisy-or networks, the errors arise due to non-linearities in the model. For sparsely-used dictionaries, the errors arise due to unknown correlations in the distribution of sparse feature representations. These examples show that robust tensor decomposition allows us to capture a wider range of models.

Robustness guarantees for tensor decomposition algorithms have been studied extensively (e.g., the work on tensor power iteration Anandkumar et al. (2014)). The polynomial-time algorithm with the best known robustness guarantees for tensor decomposition Barak et al. (2015); Ma et al. (2016) are based on the sum-of-squares (SoS) method, a powerful meta-algorithm for polynomial optimization problems based on semidefinite programming relaxations. Unfortunately, these algorithms are far from practical and have polynomial running times with large exponents. The goal of this work is to develop practical tensor decomposition algorithms with robustness guarantees close to those of SoS-based algorithms.

For the sake of exposition, we consider the case that the components $a_1, \ldots, a_n \in \mathbb{R}^d$ of the input tensor T are orthonormal. (Through standard reductions which we explain later, most of our results also apply to components that are spectrally close to orthonormal or at least linearly independent.) It turns out that the robustness guarantees that SoS achieves for the case that T is a 4-tensor are significantly stronger than its guarantees for 3-tensors. These stronger guarantees are crucial for applications like dictionary learning. (It also turns out that for 3-tensors, an analysis of Jennrich's aforementioned algorithm using matrix concentration inequalities gives robustness guarantees that are similar to those of SoS Ma et al. (2016); Arora et al. (2016).)

In this work, we develop an easy-to-implement, randomized spectral algorithm to decompose 4-tensors with orthonormal components even when the error tensor E has small but constant spectral norm as a d^2-by-d^2 matrix. This robustness guarantee is qualitatively optimal with respet to this norm in the sense that an error tensor E with constant spectral norm (as a d^2-by-d^2 matrix) could change each component by a constant proportion of its norm. To the best of our knowledge, the

only previous algorithms with this kind of robustness guarantee are based on SoS.[1] Our algorithm runs in time $d^{2+\omega} \leqslant d^{4.373}$ using fast matrix multiplication. Even without fast matrix multiplication, our running time of d^5 is close to linear in the size of the input d^4 and significantly faster than the running time of SoS. As we will discuss later, an extension of this algorithm allows us to solve instances of dictionary learning that previously could provably be solved only by SoS.

A related previous work Hopkins et al. (2016) also studied the question how to achieve similar guarantees for tensor decomposition as SoS using just spectral algorithms. Our algorithms follow the same general strategy as the algorithms in this prior work: In an algorithm based on the SoS semidefinite program, one solves a convex programming relaxation to obtain a "proxy" for a true, integral solution (such as a component of the tensor). Because the program is a relaxation, one must then process or "round" the relaxation or proxy into a true solution. In our algorithms, as in Hopkins et al. (2016), instead of finding solutions to SoS semidefinite programs, the algorithms find "proxy objects" that behave in similar ways with respect to the rounding procedures used by SoS-based algorithms. Since these rounding procedures tend to be quite simple, there is hope that generating proxy objects that "fool" these procedures is computationally more efficient than solving general semidefinite programs.

However, our algorithmic techniques for finding these "proxy objects" differ significantly from those in prior work. The reason is that many of the techniques in Hopkins et al. (2016), e.g., concentration inequalities for matrix-valued polynomials, are tailored to average-case problems and therefore do not apply in our setting because we do not make distributional assumption about the errors E.

The basic version of our algorithm is specified by a sequence of convex sets $\mathcal{X}_1, \ldots, \mathcal{X}_r \subseteq (\mathbb{R}^d)^{\otimes 4}$ of 4-tensors and proceeds as follows:

> Given a 4-tensor $T \in (\mathbb{R}^d)^{\otimes 4}$, compute iterative projections $T^{(1)}, \ldots, T^{(r)}$ to the convex sets $\mathcal{X}_1, \ldots, \mathcal{X}_r$ (with respect to euclidean norm) and apply Jennrich's algorithm on $T^{(r)}$.

It turns out that the SoS-based algorithm correspond to the case that $r = 1$ and $\mathcal{X}_1 = \mathcal{X}_{\mathrm{SOS}}$ is the feasible region of a large semidefinite program. For our fast algorithm, $\mathcal{X}_1, \ldots, \mathcal{X}_r$ are simpler sets defined in terms of singular values or eigenvalues of matrix reshapings of tensors. Therefore, projections boil down to fast eigenvector computations. We choose the sets $\mathcal{X}_1, \ldots, \mathcal{X}_r$ such that they contain $\mathcal{X}_{\mathrm{SOS}}$ and show that the iterative projection behaves in a similar way as the projection to $\mathcal{X}_{\mathrm{SOS}}$. In this sense our algorithm is similar in spirit to iterated projective methods like the Bregman method (e.g., Goldstein and Osher (2009)).

Dictionary learning. In this basic unsupervised learning problem, the goal is to learn an unknown matrix $A \in \mathbb{R}^{d \times n}$ from i.i.d. samples $y^{(1)} = Ax^{(1)}, \ldots, y^{(m)} = Ax^{(m)}$, where $x^{(1)}, \ldots, x^{(m)}$ are i.i.d. samples from a distribution $\{x\}$ over sparse vectors in $\mathbb{R}^n$. (Here, the algorithm has access only to the vectors $y^{(1)}, \ldots, y^{(m)}$ but not to $x^{(1)}, \ldots, x^{(m)}$.)

Dictionary learning, also known as sparse coding, is studied extensively in neuroscience (Olshausen and Field (1997)), machine learning (Evgeniou and Pontil (2007); Marc'Aurelio Ranzato et al. (2007)), and computer vision (Elad and Aharon (2006); Mairal et al. (2008); Yang et al. (2008)). Most algorithms for this problem used in practice do not come with strong provable guarantees. In recent years, several algorithms with provable guarantees have been developed for this

1. We remark that the aforementioned analysis Ma et al. (2016) of Jennrich's algorithm can tolerate errors E if its spectral norm as a non-square d^3-by-d matrix is constant. However, this norm of E can be larger by a $\sqrt{d}$ factor than its spectral norm as a square matrix.

problem (Agarwal et al. (2014); Arora et al. (2014, 2015); Barak et al. (2015); Ma et al. (2016); Hazan and Ma (2016)).

For the case that the coordinates of the distribution $\{x\}$ are independent (and non-Gaussian) there is a well-known reduction[2] of this problem to tensor decomposition where the components are the columns of A and the error E can be made inverse polynomially small by taking sufficiently many samples. (In the case of independent coordinates, dictionary learning becomes a special case of independent component analysis / blind source separation, where this reduction originated.) Variants of Jennrich's spectral tensor decomposition algorithm (e.g., Bhaskara et al. (2014); Ma et al. (2016)) imply strong provable guarantees for dictionary learning in the case that $\{x\}$ has independent coordinates even in the "overcomplete" regime when $n \gg d$ (using a polynomial number of samples).

More challenging is the case that $\{x\}$ has non-independent coordinates, especially if those correlations are unknown. We consider a model proposed in Barak et al. (2015) (similar to a model in Arora et al. (2014)): We say that the distribution $\{x\}$ is τ-*nice* if

1. $\mathbb{E}\, x_i^4 = 1$ for all $i \in [n]$,
2. $\mathbb{E}\, x_i^2 x_j^2 \leqslant \tau$ for all $i \neq j \in [n]$,
3. $\mathbb{E}\, x_i x_j x_k x_\ell = 0$ unless $x_i x_j x_k x_\ell$ is a square.

The conditions allow for significant correlations in the support set of the vector x. For example, we can obtain a τ-nice distribution $\{x\}$ by starting from any distribution over subsets $S \subseteq [n]$ such that $\mathbb{P}\{i \in S\} = p$ for all $i \in [n]$ and $\mathbb{P}\{j \in S \mid i \in S\} \leqslant \tau$ for all $i \neq j$ and choosing x of the form $x_i = p^{-1/4} \cdot \sigma_i$ if $i \in S$ and $x_i = 0$ if $i \notin S$, where $\sigma_1, \dots, \sigma_n$ are independent random signs.

An extension of our aforementioned algorithm for orthogonal tensor decomposition with spectral norm error allows us to learn orthonormal dictionaries from τ-nice distributions. To the best of our knowledge, the only previous algorithms to provably solve this problem use sum-of-squares (Barak et al. (2015); Ma et al. (2016)), which have large polynomial running time. Our algorithm recovers a 0.99 fraction of the columns of A up to error τ from $\widetilde{O}(n^3)$ samples, and runs in time $n^{3+O(\tau)}d^4$. By a standard reduction, our algorithm also works for non-singular dictionaries and the running time increases by a factor polynomial in the condition number of A.

1.1. Results

Tensor decomposition. For tensor decomposition, we give an algorithm with close to linear running time that recovers the rank-1 components of a tensor with orthonormal components, so long as the spectral norm of the square unfoldings of the error tensor is small.

Theorem 1 (Tensor decomposition with spectral norm error) *There exists a randomized spectral algorithm with the following guarantees: Given a tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$ *of the form* $\mathbf{T} = \sum_{i=1}^n a_i^{\otimes 4} + \mathbf{E}$ *such that* $a_1, \dots, a_n \in \mathbb{R}^d$ *are orthonormal and* $\mathbf{E}$ *has spectral norm at most* ε *as a* d^2*-by-*d^2 *matrix, the algorithm can recover one of the components with* ℓ_2*-error* $O(\varepsilon)$ *in time* $\widetilde{O}(d^{2+\omega+O(\varepsilon)})$ *with high probability, and recover* $0.99n$ *of the components with* ℓ_2*-error* $O(\varepsilon)$ *in time* $\widetilde{O}(d^{2+\omega+O(\varepsilon)})$ *with high probability.*

Furthermore, if $\varepsilon \leqslant O(\log\log n/\log^3 n)$*, the algorithm can recover all components up to error* $O(\varepsilon)$ *in time* $\widetilde{O}(d^{2+\omega} + nd^4)$ *with high probability. Here,* $\omega \leqslant 2.373$ *is the matrix multiplication exponent.*

2. The reduction only requires k-wise independence where k is the order to tensor the reduction produces.

The orthogonality condition may at first seem restrictive, but for most applications it is possible to take a tensor with linearly independent components and transform it to a tensor with orthogonal components, as we will do for our dictionary learning result below. Furthermore, our algorithm works as-is if the components are sufficiently close to orthonormal:

Corollary 2 *If we have that $a_1, \ldots, a_n$ are only approximately orthonormal in the sense that the a_i are independent and $\left\|\sum_i a_i a_i^\top - \mathrm{Id}_S\right\| \leqslant \eta$, where Id_S is the identity in the subspace spanned by the a_i, then we can recover b_i so that $\langle a_i, b_i\rangle^2 \geqslant 1 - O(\sqrt{\eta})$ with the same algorithm and runtime guarantees.*

These robustness guarantees are comparable to those of the sum-of-squares based algorithms in Barak et al. (2015); Ma et al. (2016) for the undercomplete case, which are the best known. Meanwhile, the sum-of-squares based algorithms require solving large semidefinite programs, while the running time of our algorithms is close to linear in the size of the input, and our algorithms are composed of simple matrix-vector multiplications.

On the other hand, our algorithms fail to work in the overcomplete case, when the rank grows above n, and the components are no longer linearly independent. One interesting open question is whether the techniques used in this paper can be extended to the overcomplete case.

Dictionary learning. Using our tensor decomposition algorithm as a primitive, we give an algorithm for dictionary learning when the sample distribution is τ-nice.

Theorem 3 (Dictionary learning) *Suppose that $A \in \mathbb{R}^{d\times n}$ is a dictionary with orthonormal columns, and that we are given random independent samples of the form $y = Ax$ for $x \sim \mathcal{D}$. Suppose furthermore that $\mathcal{D}$ is τ-nice, as defined above, for $\tau < c^*$ for some universal constant c^*.*

Then there is a randomized spectral algorithm that recovers orthonormal vectors $b_1, \ldots, b_k \in \mathbb{R}^d$ for $k \geqslant 0.99n$ with $\langle b_i, a_i\rangle^2 \geqslant (1 - O(\tau))$, and with high probability requires $m = \widetilde{O}(n^3)$ samples and time $\widetilde{O}(d^{2+\omega} + n^{1+O(\tau)}d^4 + md^4)$.

The total runtime is thus $\widetilde{O}(n^3 d^4)$—in the theorem statement, we write it in terms of the number of samples m in order to separate the time spent processing the samples from the learning phase. We note that the sample complexity bound that we have, $m = \widetilde{O}(n^3)$, may very well be sub-optimal; we suspect that $m = \widetilde{O}(n^2)$ is closer to the truth, which would yield a better runtime.

We are also able to apply standard whitening operations (as in e.g. Anandkumar et al. (2013)) to extend our algorithm to dictionaries with linearly-independent, but non-orthonormal, columns, at the cost of polynomially many additional samples.

Corollary 4 *If $A \in \mathbb{R}^{d\times n}$ is a dictionary with linearly independent columns, then there is a randomized spectral algorithm that recovers the columns of A with guarantees similar to Theorem 3 given $\widetilde{O}(n^2 \cdot f(\mu))$ additional samples, where $\mu = \lambda_{\max}(AA^\top)/\lambda_{\min}(AA^\top)$ is the condition number of the covariance matrix, and f is a polynomial function.*

To our knowledge, our algorithms are the only remotely efficient dictionary learning algorithms with provable guarantees that permit τ-nice distributions in which the coordinates of x may be correlated by constant factors, the only other ones being the sum-of-squares semidefinite programming based algorithms of Barak et al. (2015); Ma et al. (2016).

2. Preliminaries

Throughout the rest of this paper, we will denote tensors by boldface letters such as $\mathbf{T}$, matrices by capital letters M, and vectors by lowercase letters v, when the distinction is helpful. We will use $A^{\otimes k}/u^{\otimes k}$ to denote the kth Kronecker power of a matrix/vector with itself. To enhance legibility, for $u \in \mathbb{R}^d$ we will at times abuse notation and use $u^{\otimes 4}$ to denote the order-4 tensor $u \otimes u \otimes u \otimes u$, the $d^2 \times d^2$ matrix $(u^{\otimes 2})(u^{\otimes 2})^\top$, and the dimension d^4 vector $u^{\otimes 4}$—we hope the meaning will be clear from context.

For a tensor $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$ and a partition of the modes $\{1,2,3,4\}$ into two ordered sets A and B, we let $T_{A,B}$ denote the reshaping of $\mathbf{T}$ as $d^{|A|}$-by-$d^{|B|}$ matrix, where the modes in A are used to index rows and the modes in B are used to index columns. For example, $T_{\{1,2\},\{3,4\}}$ is a d^2-by-d^2 matrix such that the entry at row (i,j) and column (k,ℓ) contains the entry $T_{i,j,k,\ell}$ of T. We remark that the order used to specify the modes matters—for example, for the rank-1 tensor $\mathbf{T} = a \otimes b \otimes a \otimes b$, we have that $T_{\{1,2\}\{3,4\}} = (a \otimes b)(a \otimes b)^\top$ is a symmetric matrix, while $T_{\{2,1\}\{3,4\}} = (b \otimes a)(a \otimes b)^\top$ is not. We use $\|T_{A,B}\|$ to denote the spectral norm (largest singular value) of the matrix $T_{A,B}$.

We will also make frequent use of the following lemma, which states that the distance between two points cannot increase when both are projected onto a closed, convex set.

Lemma *Let $C \subset \mathbb{R}^n$ be a closed convex set, and let $\Pi : \mathbb{R}^n \to C$ be the projection operator onto C in terms of norm $\|\cdot\|_2$, i.e. $\Pi(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{c \in C} \|x - c\|_2$. Then for any $x, y \in \mathbb{R}^n$,*

$$\|x - y\|_2 \geqslant \|\Pi(x) - \Pi(y)\|_2.$$

This lemma is well-known (see e.g. Rockafellar (1976)), but we will prove it for completeness in Appendix A.

3. Techniques

In this section we give a high-level overview of the algorithms in our paper, and of their analyses. We begin with the tensor decomposition algorithm, after which we'll explain the (non-trivial) extension to the dictionary learning application. At the very end, we will discuss the relationship between our algorithms and sum-of-squares relaxations.

Suppose have a tensor $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$, and that $\mathbf{T} = \mathbf{S} + \mathbf{E}$ where the *signal* $\mathbf{S}$ is a low-rank tensor with orthonormal components, $\mathbf{S} = \sum_{i \in [n]} a_i^{\otimes 4}$, and the *noise* $\mathbf{E}$ is an arbitrary tensor of noise with the restriction that for any reshaping of $\mathbf{E}$ into a square matrix E, $\|E\| \leqslant \varepsilon$. Our goal is to (approximately) recover the rank-1 components, $a_1, \ldots, a_n \in \mathbb{R}^d$, up to signs.

Failure of Jennrich's algorithm. To motivate the algorithm and analysis, it first makes sense to consider the case when the noise component $\mathbf{E} = 0$. In this case, we can run Jennrich's algorithm: if we choose a d^2-dimensional random vector $g \sim \mathcal{N}(0, \mathrm{Id})$, we can compute the contraction

$$M_g \stackrel{\text{def}}{=} \sum_{i,j=1}^{n} g_{ij} T_{ij} = \sum_{i=1}^{n} \langle g, a_i^{\otimes 2} \rangle \cdot a_i a_i^\top,$$

where T_{ij} is the i,jth $d \times d$ matrix slice of the tensor $\mathbf{T}$. Since the a_i are orthogonal, the coefficients $\langle g, a_i^{\otimes 2} \rangle$ are independent, and so we find ourselves in an ideal situation—M_g is a sum of

the orthogonal components we want to recover with independent Gaussian coefficients. A simple eigendecomposition will recover all of the a_i.

On the other hand, when we have a nonzero noise tensor $\mathbf{E}$, a random contraction along modes $\{1, 2\}$ results in the matrix

$$M_g = \sum_{i=1}^{n} \langle g, a_i^{\otimes 2} \rangle a_i a_i^\top + \sum_{i,j=1}^{n} g_{ij} \cdot E_{ij},$$

where the E_{ij} are $d \times d$ slices of the tensor $\mathbf{E}$. The last term, composed of the error, complicates things. Standard facts about Gaussian matrix series assert that the spectral norm of the error term behaves like $\|E_{\{1,2,3\}\{4\}}\|$, the spectral norm of a $d^3 \times d$ reshaping of E, whereas we only have control over square reshapings such as $\|E_{\{1,2\}\{3,4\}}\|$.[3] These can be off by polynomial factors. If the Frobenius norm of $\mathbf{E}$ is $\|\mathbf{E}\|_F^2 \approx \varepsilon^2 d^2$, which is the magnitude one would expect from a tensor whose square reshapings are full-rank matrices with spectral norm ε, then we have that necessarily

$$\|E_{\{1,2,3\}\{4\}}\|^2 \geqslant \frac{\|\mathbf{E}\|_F^2}{\operatorname{rank}\left(E_{\{1,2,3\}\{4\}}\right)} \geqslant \varepsilon^2 d,$$

since there are at most d nonzero singular values of rectangular reshapings of $\mathbf{E}$. In this case, unless $\varepsilon \ll 1/\sqrt{d}$, the components $a_i a_i^\top$ are completely drowned out by the contribution of the noise, and so the robustness guarantees leave something to be desired.

Basic idea. The above suggests that, as long as we allow the error $\mathbf{E}$ to have large Frobenius norm, an approach based on random contraction will not succeed. Our basic idea is to take $\mathbf{T}$, whose error has small spectral norm, and transform it into a tensor $\mathbf{T}'$ whose error has small Frobenius norm.

Because we do not know the decomposition of $\mathbf{T}$, we cannot access the error $\mathbf{E}$ directly. However, we do know that for any $d^2 \times d^2$ reshaping T of $\mathbf{T}$,

$$T = S + E,$$

where $S = \sum_{i=1}^{n} a_i^{\otimes 4}$, and $\|E\| \leqslant \varepsilon$. The rank of S is $n \leqslant d$, and all eigenvalues of S are 1. Thus, if we perform the operation

$$T^{>\varepsilon} = (T - \varepsilon \operatorname{Id})_+,$$

where $(\cdot)_+$ denotes projection to the cone of positive semidefinite matrices, we expect that the signal term S will survive, while the noise term E will be dampened. More formally, we know that T has n eigenvalues of magnitude $1 \pm \varepsilon$, and $d^2 - n$ eigenvalues of magnitude at most ε, and therefore $\operatorname{rank}(T^{>\varepsilon}) \leqslant n$. Also by definition, $\|T - T^{>\varepsilon}\| \leqslant \varepsilon$. Therefore, we have that

$$S + E = T = T^{>\varepsilon} + E'$$

with $\|E'\| \leqslant \varepsilon$, and thus

$$\|T^{>\varepsilon} - S\| = \|E - E'\|.$$

Since $S, T^{>\varepsilon}$ are both of rank at most n, and E', E have spectral norm bounded by ε, we have that

$$\|T^{>\varepsilon} - S\|_F^2 \leqslant \left(\operatorname{rank}(S) + \operatorname{rank}(T^{>\varepsilon})\right) \cdot \left(\|E\| + \|E'\|\right)^2 \leqslant 2n \cdot 4\varepsilon^2.$$

3. In fact, this observation was crucial in the analysis of Ma et al. (2016)—in that work, semidefinite programming constraints are used to control the spectral norm of the rectangular reshapings.

So, the Frobenius norm is no longer an impassable obstacle to the random contraction approach—using our upper bound on the Frobenius norm of our new error $\tilde{E} \stackrel{\text{def}}{=} T^{>\varepsilon} - S$, we have that the average squared singular value of $\tilde{E}_{\{1,2,3\}\{4\}}$ will be

$$\sigma^2_{avg}(\tilde{E}_{\{1,2,3\}\{4\}}) = \frac{\|\tilde{E}\|_F^2}{d} = O\left(\frac{n}{d}\varepsilon^2\right).$$

So while $\tilde{E}_{\{1,2,3\}\{4\}}$ may have large singular values, by Markov's inequality it cannot have too many singular values larger than $O(\varepsilon)$.

Finally, to eliminate these large singular values, we will project $\mathbf{T}^{>\varepsilon}$ into the set of matrices whose rectangular reshapings have singular values at most 1—because S is a member of this convex set, the projection can only decrease the Frobenius norm. After this, we will apply the random contraction algorithm, as originally suggested.

Variance of Gaussian matrix series. Recall that we wanted to sample a random d^2-dimensional Gaussian vector g, and the perform the contraction

$$M_g \stackrel{\text{def}}{=} \sum_{i,j=1}^{d} g_{ij} T^{>\varepsilon}_{ij} = \sum_{i=1}^{d} \langle g, a_i^{\otimes 2}\rangle \cdot a_i a_i^\top + \sum_{ij} g_{ij} \cdot \tilde{E}_{ij}.$$

The error term on the right is a matrix Gaussian series. The following lemma describes the behavior of the spectra of matrix Gaussian series:

Lemma *[See e.g. Tropp (2012)] Let $g \sim \mathcal{N}(0, \mathrm{Id})$, and let $A_1, \ldots, A_k$ be $n \times m$ real matrices. Define $\sigma^2 = \max\left\{\left\|\sum_i A_i A_i^\top\right\|, \left\|\sum_i A_i^\top A_i\right\|\right\}$. Then*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{k} g_i A_i\right\| \geqslant t\right) \leqslant (n+m) \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

For us, this means that we must have bounds on the spectral norm of both $\sum_{ij} \tilde{E}_{ij}\tilde{E}_{ij}^\top$ and $\sum_{ij} \tilde{E}_{ij}^\top \tilde{E}_{ij}$. This means that if we have performed the contraction along modes 1 and 2, so that the index i comes from mode 1 and the index j comes from mode 2, then it is not hard to verify that $\|\sum_{ij} \tilde{E}_{ij}\tilde{E}_{ij}^\top\| = \|\tilde{E}_{\{1,2,3\}\{4\}}\|^2$, and $\|\sum_{ij} \tilde{E}_{ij}^\top \tilde{E}_{ij}\| = \|\tilde{E}_{\{1,2,4\}\{3\}}\|^2$. So, we must control the maximum singular values of two different rectangular reshapings of $\tilde{E}$ simultaneously.

It turns out that for us it suffices to perform two projections in sequence—we first reshape $\mathbf{T}^{>\varepsilon}$ to the matrix $\mathbf{T}^{>\varepsilon}_{\{123\}\{4\}}$,project it to the set of matrices with singular values at most 1, and then reshape the result along modes $\{124\}\{3\}$, and project to the same set again. As mentioned before, because projection to a convex set containing S cannot increase the distance from S, the Frobenius norm of the new error can only decrease. What is less obvious is that performing the second projection will not destroy the property that the reshaping along modes $\{123\}\{4\}$ has spectral norm at most 1. By showing that each projection corresponds to either left- or right- multiplication of $\mathbf{T}^{>\varepsilon}_{\{123\}\{4\}}$ and $\mathbf{T}^{>\varepsilon}_{\{124\}\{3\}}$ by matrices of spectral norm at most 1, we are able to show that the second projection does not create large singular values for the first flattening, and so two projections are indeed enough. Call the resulting tensor $(\mathbf{T}^{>\varepsilon})^{\leqslant 1}$.

Now, if we perform a random contraction in the modes 1, 2, we will have

$$M_g = \sum_{ij} g_{ij} (\mathbf{T}^{>\varepsilon})^{\leqslant 1}_{ij} = \sum_i \langle a_i^{\otimes 2}, g\rangle a_i a_i^\top + \hat{E}_g,$$

where the spectral norm of $\|\hat{E}_g\| \leqslant \sqrt{\log n}$ with good probability. So, ignoring for the moment dependencies between $\langle a_i^{\otimes 2}, g\rangle$ and E_g, $\max_i |\langle g, a_i^{\otimes 2}\rangle| > 1.1 \cdot \|E_g\|$ with probability at least $n^{-.2}$, which will give a_i correlation 0.9 with the top eigenvector of M_g with good probability.

Improving accuracy of components. The algorithm described thus far will recover components b_i that are 0.9-correlated with the a_i, in the sense that $\langle b_i, a_i\rangle^2 \geqslant 0.9$. To boost the accuracy of the recovered components, we use a simple method which resembles a single step of tensor power iteration.

We'll use the closeness of our original tensor $\mathbf{T}$ to $\sum_i a_i^{\otimes 4}$ in spectral norm. We let T be a $d^2 \times d^2$ flattening of $\mathbf{T}$, and compute the vector

$$v = T(b_i \otimes b_i) = 0.9 \cdot a_i \otimes a_i + \sum_{j \neq i} \langle b_i, a_j\rangle^2 a_j \otimes a_j + E(b_i \otimes b_i).$$

Now, when the vector v is reshaped to a $d \times d$ matrix V, the term $E(b_i \otimes b_i)$ is a matrix of Frobenius norm (and thus spectral norm) at most ε. By the orthonormality of the a_j, the sum of the coefficients in the second term is at most 0.1, and so a_i is ε-close to the top eigenvector of V.

Recovering every component. Because the Frobenius norm of the error in $(\mathbf{T}^{>\varepsilon})^{\leqslant 1}$ is $O(\varepsilon\sqrt{n})$, there may be a small fraction of the components $a_i^{\otimes 4}$ that are "canceled out" by the error—for instance, we can imagine that the error term is $\hat{\mathbf{E}} = -\sum_{i=1}^{\varepsilon^2 n} a_i^{\otimes 4}$. So while only a constant fraction of the components $a_i^{\otimes 4}$ can be more than ε-correlated with the error, we may still be unable to recover some fixed ε^2-fraction of the a_i via random contractions.

To recover all components, we must subtract the components that we have found already and run the algorithm iteratively—if we have found $m = 0.99n$ components, then if we could perfectly subtract them from $\mathbf{T}$, we would end up with an even lower-rank signal tensor, and thus be able to make progress by truncating all but $0.1n$ eigenvalues in the first step.

The challenge is that we have recovered $b_1, \ldots, b_m$ that are only $(1-\varepsilon)$-correlated with the a_i, and so naively subtracting $\mathbf{T} - \sum_{i=1}^m b_i^{\otimes 4}$ can result in a Frobenius and spectral norm error of magnitude $\varepsilon\sqrt{m}$—thus the total error is still proportional to $\sqrt{n}$ rather than $\sqrt{0.1n}$.

In order to apply our algorithm recursively, we first orthogonalize the components we have found $b_1, \ldots, b_m$ to obtain new components $\tilde{b}_1, \ldots, \tilde{b}_m$. Because the b_i are close to the truly orthonormal a_i, the orthogonalization step cannot push too many of the $\tilde{b}_i$ more than $O(\varepsilon)$-far from the b_i—in fact, letting B be the matrix whose columns are the b_i, and letting A be the matrix whose columns are the corresponding a_i, we use that $\|A - B\|_F \leqslant O(\varepsilon)\sqrt{m}$, and that the matrix $\tilde{B}$ with columns $\tilde{b}_i$ is closer to B than A. We keep only the $\tilde{b}_i$ for which $\langle \tilde{b}_i^{\otimes 4}, \mathbf{T}\rangle \geqslant 1 - O(\varepsilon)$, and we argue that there must be at least $0.9m$ such $\tilde{b}_i$. Let $K \subset [m]$ be the set of indices for which this occurred.

Now, given that the two sets of orthogonal vectors $\{\tilde{b}_i\}_{i\in K}$ and $\{a_i\}_{i\in K}$ are all $O(\varepsilon)$ close, we are able to prove that

$$\left\| \sum_{i \in K} \tilde{b}_i^{\otimes 4} - a_i^{\otimes 4} \right\| \leqslant O(\sqrt{\varepsilon}).$$

So subtracting the $\tilde{b}_i^{\otimes 4}$ will not introduce a large spectral norm! Since we will only need to perform this recursion $O(\log n)$ times, allowing for some leeway in ε (by requiring $\sqrt{\varepsilon}\log n = o(1)$), we are able to recover all of the components.

Dictionary learning. In the dictionary learning problem, there is an unknown dictionary, $A \in \mathbb{R}^{d\times n}$, and we receive independent samples of the form $y = Ax$ for $x \sim \mathcal{D}$ for some distribution $\mathcal{D}$ over $\mathbb{R}^n$. The goal is, given access only to the samples y, recover A.

We can use our tensor decomposition algorithm to learn the dictionary A, as long as the columns of A are linearly independent. For the sake of this overview, assume instead that the columns of A are orthonormal. Then given samples $y^{(1)}, \dots, y^{(m)}$ for $m = \mathrm{poly}(n)$, we can compute the 4th moment tensor to accuracy ε in the spectral norm,

$$\frac{1}{m}\sum_{j=1}^{m}(y^{(j)})^{\otimes 4} \approx \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Ax)^{\otimes 4}\right].$$

If the right-hand side were close to $\sum_i a_i^{\otimes 4}$ in spectral norm, we would be done. However, for almost any distribution $\mathcal{D}$ which is supported on x with more than one nonzero coordinate, this is not the case. If we assume that $\mathbb{E}[x_i x_j x_k x_\ell] = 0$ unless $x_i x_j x_k x_\ell$ is a square, then we can calculate that any square reshaping of this tensor will have the form

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Ax)^{\otimes 4}\right] = \sum_i \mathbb{E}[x_i^4]\cdot a_i^{\otimes 4} + \sum_{i\neq j}\mathbb{E}[x_i^2 x_j^2]\cdot(a_i a_j^\top \otimes a_i a_j^\top + a_i a_i^\top \otimes a_j a_j^\top + a_i a_j^\top \otimes a_j a_i^\top).$$

If $\mathbb{E}[x_i^4] = \mathbb{E}[x_j^4]$ for all i, j, then the first term on the right is exactly the 4th order tensor that we want. The second term on the right can be further split into three distinct matrices, one for each configuration of the a_i, a_j. The $a_i a_i^\top \otimes a_j a_j^\top$ term and the $a_i a_j^\top \otimes a_j a_i^\top$ terms can be shown to have spectral norm at most $\max_{i\neq j}\mathbb{E}[x_i^2 x_j^2]$, and so as long as we require that $\max_{i\neq j}\mathbb{E}[x_i^2 x_j^2] \ll \varepsilon\,\mathbb{E}[x_k^4]$, these terms have spectral norm within the allowance of our tensor decomposition algorithm.

The issue is with the $a_i a_j^\top \otimes a_i a_j^\top$ term. This term factors into $(a_i \otimes a_i)(a_j \otimes a_j)^\top$, and because of this the entire sum $\sum_{i\neq j}\mathbb{E}[x_i^2 x_j^2]\cdot a_i a_j^\top \otimes a_i a_j^\top$ has rank at most $n \ll d^2$, but Frobenius norm as large as $\max_{i\neq j}\mathbb{E}[x_i^2 x_j^2]\cdot n$. If we require that the coordinates of $x \sim \mathcal{D}$ are independent, then we can see that this is actually close to a spurious rank-1 component, which can be easily removed without altering the signal term too much.[4] However, if we wish to let the coordinates of x exhibit correlations, we have very little information about the spectrum of this term.

In the sum-of-squares relaxation, this issue is overcome easily: because of the symmetries required of the SDP solution matrix X, $\langle X, a_i a_j^\top \otimes a_i a_j^\top\rangle = \langle X, a_i a_i^\top \otimes a_j a_j^\top\rangle$, so by linearity this low-rank error term cannot influence the objective function any more than the $a_i a_i^\top \otimes a_j a_j^\top$ term.

Inspired by this sum-of-squares analysis, we remove these unwanted directions as follows. Given the scaled moment matrix $M = \frac{1}{\mathbb{E}[x_1^4]}\mathbb{E}_{x\sim\mathcal{D}}\left[(Ax)^{\otimes 4}\right]$ (where the scaling serves to make the coefficients of the signal 1), we truncate the small eigenvalues:

$$M^{>\varepsilon} = (M_{\{1,2\}\{3,4\}} - \varepsilon\,\mathrm{Id})_+.$$

This removes the spectrum in the direction of the "nice" error terms, corresponding to $a_i a_i^\top \otimes a_j a_j^\top$ and $a_i a_j^\top \otimes a_j a_i^\top$. The fact that the rest of the matrix is low-rank means that we can apply an analysis similar to the analysis in the first step of our tensor decomposition to argue that $\|M^{>\varepsilon} - \sum_i a_i^{\otimes 4}\|_F \leqslant \|M - \sum_i a_i^{\otimes 4}\|_F$.

4. As was done in Hopkins et al. (2015), for example, albeit in a slightly different context.

Now, we re-shape $M^{>\varepsilon}$, so that if initially we had the flattening $\mathbf{M} \to M_{\{1,2\}\{3,4\}}$, we look at the flattening $M^{>\varepsilon}_{\{1,3\}\{2,4\}}$. In this flattening, the term $a_i a_j^\top \otimes a_i a_j^\top$ from $M^{>\varepsilon}$ is transformed to $a_i a_i^\top \otimes a_j a_j^\top$, and so the problematic error term from the original flattening has spectral norm ε in this flattening! Applying the projection

$$(M^{>\varepsilon}_{\{1,3\}\{2,4\}} - \varepsilon \operatorname{Id})_+$$

eliminates the problematic term, and brings us again closer to the target matrix $\sum_i a_i^{\otimes 4}$. Thus, we end up with a tensor that is close to $\sum_i a_i^{\otimes 4}$ in Frobenius norm, and we can apply our tensor decomposition algorithm.

Connection to sum-of-squares algorithms. We take a moment to draw parallels between our algorithm and tensor decomposition algorithms in the sum-of-squares hierarchy. In noisy orthogonal tensor decomposition, we want to solve the non-convex program

$$\operatorname{argmax} \langle X, \mathbf{T} \rangle \quad \text{s.t.} \quad \left\{ X \in \mathbb{R}^{d^2 \times d^2},\ X \succeq 0,\ \|X\| = 1,\ \operatorname{rank}(X) = n,\ X \in \operatorname{Span}\{u^{\otimes 4} : u \in \mathbb{R}^d\} \right\}.$$

The intended solution of this program is $X = \mathbf{S}$ (after which we can run Jennrich's algorithm to recover individual components). At first it may not be obvious that the maximizer of the above program is close to $\mathbf{S}$, but for any unit vector $x \in \mathbb{R}^d$,

$$\langle x^{\otimes 4}, \mathbf{T} \rangle = \sum_{i=1}^n \langle x, a_i \rangle^4 + (x \otimes x)^\top E (x \otimes x).$$

The error term is at most ε by our bound on $\|E\|$, and the first term is $\|x^\top A\|_4^4$, where A is the matrix whose columns are the a_i. Since by the orthonormality of the a_i, $\|x^\top A\|_2 \leqslant 1$, and the ℓ_4 norm is maximized relative to the ℓ_2 for vectors supported on a single coordinate, the x that maximize this must be ε-close to one of the a_i. In conjunction with the $\|X\| \leqslant 1$ constraint and the $\operatorname{rank}(X) = n$ constraint, we have that $X := \mathbf{S}$ is the maximizer.

The (somewhat simplified) corresponding sum-of-squares relaxation is the semidefinite program

$$\max \langle X, \mathbf{T} \rangle \quad \text{s.t.} \quad \left\{ X \in \mathbb{R}^{d^2 \times d^2},\ X \succeq 0,\ \|X\| \leqslant 1,\ \|X\|_F^2 = n,\ X_{ijk\ell} = X_{\pi(ijk\ell)}\ \forall \pi \in \mathcal{S}_4 \right\},$$ [5]

The constraints $\|X\|_F^2 = n$ and $X_{ijk\ell} = X_{\pi(ijk\ell)}$ together are a relaxation of the constraint that X be a rank-n matrix in the symmetric subspace $\operatorname{Span}\{u^{\otimes 4}\}$. Further, the constraint $\|X\| \leqslant 1$ is enforced in every rectangular $d \times d^3$ reshaping of X (this consequence of the SoS constraints is crucially used in Ma et al. (2016)).

To solve this semidefinite program, one should project $\mathbf{T}$ into the intersection of all of the convex feasible regions of the constraints. However, projecting to the intersection is an expensive operation in terms of runtime. Instead, we choose a subset of these constraints, and project $\mathbf{T}$ into the set of points satisfying each constraint sequentially, rather than simultaneously. These are not equivalent projection operations, but because we select our operations carefully, we are able to show that our

5. The program is actually over matrices indexed by all subsets of d of size at most 2, $\binom{d}{\leqslant 2}$, but for simplicity in this description we ignore this (and the interaction with the SoS variables corresponding to lower-degree monomials or lower-order moments).

$\mathbf{T}$ is close to the SDP optimum in a sense that is sufficient for successfully running Jennrich's algorithm.

In the first step of our algorithm, we change the objective function from $\langle X, \mathbf{T}\rangle$ to $\langle X, \mathbf{T} - \varepsilon \operatorname{Id}\rangle$. In the sum-of-squares SDP, this does not change the objective value dramatically—because the original objective value is at least n, and because the Frobenius norm constraint $\|X\|_F^2 = n$ constraint in conjunction with the sum-of-squares constraints implies that $\langle X, \operatorname{Id}\rangle = \varepsilon n$. Therefore this perturbation cannot decrease the objective by more than a multiplicative factor of ε. Then, we project a square reshaping of the objective to the PSD cone, $(T - \varepsilon \operatorname{Id})_+$—this corresponds to the constraint that $X \succeq 0$.[6] Finally, we project first to the set of matrices that have spectral norm at most 1 for one rectangular reshaping, then repeat for another rectangular reshaping. So after perturbing the objective very slightly, then choosing three of the convex constraints to project to in sequence, we end up with an object that approximates the maximizer of the SDP in a sense that is sufficient for our purposes.

Our dictionary learning pre-processing can be interpreted similarly. We first perturb the objective function by $\varepsilon \operatorname{Id}$, and project to the PSD cone. Then, in reshaping the tensor again, we choose another point that has the same projection onto any point in the feasible region (by moving along an equivalence class in the symmetry constraint). Finally, we perturb the objective by $\varepsilon \operatorname{Id}$ again, and again project to the PSD cone.

4. Decomposing orthogonal 4-tensors

Recall our setting: we are given a 4-tensor $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$ of the form $\mathbf{T} = \mathbf{S} + \mathbf{E}$ where $\mathbf{E}$ is a noise tensor and $\mathbf{S} = \sum_{i=1}^n a_i^{\otimes 4}$ for orthonormal vectors $a_1, \dots, a_n$. (We address the more general case of nearly orthogonormal vectos in Section 4.2.).

First, we have a pre-processing step, in which we go from a tensor with low spectral norm error to a tensor with low Frobenius norm error.[7]

Algorithm 1 (Preprocessing: spectral-to-Frobenius norm)
Input: *A tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$*, and an error parameter* ε*.*

1. *Reshape* $\mathbf{T}$ *to the* $d^2 \times d^2$ *matrix* $T \stackrel{\text{def}}{=} T_{\{1,2\}\{3,4\}}$.
2. *Truncate to 0 all eigenvalues of that have magnitude less than* ε*:*

$$T^{>\varepsilon} \stackrel{\text{def}}{=} (T - \varepsilon \operatorname{Id})_+,$$

where we have used $(M)_+$ *to denote projection to the PSD cone.*

Output: *The tensor* $T^{>\varepsilon}$.

Lemma 5 *Suppose that for some square reshaping* E *of* $\mathbf{E}$ *(without loss of generality along modes* $\{1,2\},\{3,4\}$*),* $\|E\| \leqslant \varepsilon$*. Say we are given access to* $\mathbf{T} = \mathbf{S} + \mathbf{E}$*, and we produce the matrix* $T^{>\varepsilon} = (T - \varepsilon \cdot \operatorname{Id})_+$ *as described in Algorithm 1. Then* $\|T' - S\|_F \leqslant 2\varepsilon\sqrt{2n}$*. This operation requires time* $\widetilde{O}(\min\{nd^4, d^{2+\omega}\})$.

6. For a proof that truncating the negative eigenvalues of a matrix is equivalent to projection to the PSD cone in Frobenius norm, see Fact 3.

7. We are approximating $T = S + E$ with $\operatorname{rank}(S) = n \ll d^2$, and $\|E\| \leqslant \varepsilon$, so for example if $E = \varepsilon \operatorname{Id}$ then we may have $\|T - S\|_F^2 = \varepsilon^2 d^2$, which is too large for us.

Proof Because $\|E\| \leqslant \varepsilon$, $T = S + E$ has only n eigenvalues of magnitude more than ε. So $\mathrm{rank}(T^{>\varepsilon}) \leqslant n$, and therefore $\mathrm{rank}(T^{>\varepsilon} - S) \leqslant 2n$. Furthermore, $S + E = T = T^{>\varepsilon} + \tilde{E}$ for a matrix $\tilde{E}$ of spectral norm at most ε. Therefore $\|T^{>\varepsilon} - S\| \leqslant 2\varepsilon$, and $\|T^{>\varepsilon} - S\|_F \leqslant 2\varepsilon\sqrt{2n}$.

To compute $T^{>\varepsilon}$, we can compute the top n eigenvectors of T. Since $O(\varepsilon) \cdot \lambda_n \geqslant \lambda_{n+1}$, where λ_n, λ_{n+1} are the nth and $(n+1)$st eigenvalues, we can compute this in time $\widetilde{O}(\min\{nd^4, d^{2+\omega}\})$ via subspace power iteration (see for example Hardt and Price (2014)). ■

Now, we can run our main algorithm:

Algorithm 2
Input: *A tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$.

1. *Project* $\mathbf{T}$ *to the set of tensors whose rectangular reshapings along modes* $\{1,2,3\},\{4\}$ *have spectral norm at most* 1 *(obtaining a new tensor* $\hat{\mathbf{T}}$ *with* $\|\hat{\mathbf{T}}_{\{1,2,3\}\{4\}}\| \leqslant 1$*).*
2. *Project* $\hat{\mathbf{T}}$ *to the set of tensors whose rectangular reshapings along modes* $\{1,2,4\},\{3\}$ *have spectral norm at most* 1*, obtaining a new tensor* $\mathbf{T}^{\leqslant 1} = \mathbf{S} + \mathbf{E}^{\leqslant 1}$.
3. *Sample* $g \sim \mathcal{N}(0, \mathrm{Id}_{d^2})$*, and compute the random flattening* $M_g \stackrel{\text{def}}{=} \sum_{j=1}^{d^2} g_j \mathbf{T}_j^{\leqslant 1}$.

Output: u_L *and* u_R*, the top left- and right- unit singular vectors of* M_g.

Under the appropriate conditions on $\mathbf{T}$, with probability $\widetilde{O}(n^{-\varepsilon})$, Algorithm 2 will output a vector that is 0.9-correlated with $a_j^{\otimes 2}$ for some $j \in [n]$.

Theorem 6 *Suppose we are given a* 4*-tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$*, and* $\mathbf{T} = \sum_i a_i^{\otimes 4} + E$*, where* $a_1, \ldots, a_n \in \mathbb{R}^d$ *are orthonormal vectors and* $\|E\|_F \leqslant \eta\sqrt{n}$.

Then running Algorithm 2 $\widetilde{O}(n^{1+O(\eta)})$ *times allows us to recover* $m \geqslant 0.99n$ *unit vectors* $u_1, \ldots, u_m \in \mathbb{R}^d$ *such that for each* $i \in [m]$*, there exists* $j \in [n]$ *such that*

$$\langle u_i, a_j \rangle^2 \geqslant 0.99.$$

Recovering one component requires time $\widetilde{O}(d^{2+\omega} n^{O(\eta)})$*, and recovering* m *components requires time* $\widetilde{O}(\max\{mn^{O(\eta)}d^4, d^{2+\omega}\})$.

We can then post-process the vectors to obtain a vector that has correlation $1 - \varepsilon$ with a_j—the details are given in Section 4.3 below.

We will prove Theorem 6 momentarily, but first, we bring the reader's attention to a nontrivial technical issue left unanswered by Theorem 6. The issue is that we can only guarantee that Algorithm 2 recovers $0.99n$ of the vectors, and the set of recoverable vectors is invariant under the randomness of the algorithm. That is, as a side effect of the error-reducing step 2, the $\mathbf{S}$ part of $\mathbf{T}$ may also be adversely affected. For that reason, in $\tilde{O}(n)$ runs of Algorithm 2, we can only guarantee that recover a constant fraction of the components, and we must iteratively remove the components we recover in order to continue to make progress. This removal must be handled delicately to ensure that the Frobenius norm of the error shrinks at each step, so the conditions of Theorem 6 continue to be met (for shrinking values of n). The overall algorithm, which uses Algorithm 2 as a subroutine, will be given in Section 4.3.

We will now prove the correctness of Algorithm 2 step-by-step, tying details together at the end of this subsection. First, we argue that in step 1, the truncation of the large eigenvalues cannot increase the Frobenius norm of the error.

Lemma 7 *Suppose that we define* $\mathbf{T}^{\leqslant 1}$ *to be the result of projecting* $\mathbf{T} = \mathbf{S} + \mathbf{E}$ *to the set of tensors whose rectangular reshapings along modes* $\{1,2,3\},\{4\}$ *have spectral norm at most* 1*, then projecting the result to the set of tensors whose rectangular reshapings along modes* $\{1,2,3\},\{4\}$ *have spectral norm at most* 1*. Then* $\|\mathbf{T}^{\leqslant 1} - \mathbf{S}\|_F \leqslant \|E\|_F$*, and*

$$\|\mathbf{T}^{\leqslant 1}_{\{1,2,3\}\{4\}}\| \leqslant 1, \quad \textit{and} \quad \|\mathbf{T}^{\leqslant 1}_{\{1,2,4\}\{3\}}\| \leqslant 1\,.$$

This operation requires time $\widetilde{O}(d^{2+\omega})$*.*

Proof To establish the first claim, we note that the tensor $\mathbf{T}^{\leqslant 1}$ was obtained by two projections of different rectangular reshapings of the matrix $S + E$ to the set of rectangular matrices with singular value at most 1. This set is closed, convex, and contains S, and so the error can only decrease in Frobenius norm (see Lemma 19 in Appendix A for a proof),

$$\|\mathbf{T}^{\leqslant 1} - \mathbf{S}\|_F \leqslant \|\mathbf{T} - \mathbf{S}\|_F = \|\mathbf{E}\|_F.$$

It is not hard to see that each projection step can be accomplished by reshaping the tensor to the appropriate rectangular matrix, then truncating all singular values larger than 1 to 1. Now, we establish the remaining claims. For convenience, define $\hat{T}^{\leqslant 1}$ to be the matrix $(S+E)_{\{123\}\{4\}}$ after restricting singular values of magnitude > 1 to 1. Now, we reshape $\hat{T}^{\leqslant 1}$ to a new matrix $B \stackrel{\text{def}}{=} \hat{T}^{\leqslant 1}_{\{1,2,4\}\{3\}}$, which has the d^2 blocks $B_1 = (\hat{T}^{\leqslant 1})^\top_1, \ldots, B_{d^2} = (\hat{T}^{\leqslant 1})^\top_{d^2}$. Say that the singular value decomposition of B is $B = U\Sigma V^\top$. Define $\widehat{\Sigma^{-1}}$ to be the diagonal matrix with entries equal to those of Σ^{-1} when the value is < 1 and with ones elsewhere. When we truncate the large singular values of B this is equivalent to multiplying by $P = U\widehat{\Sigma^{-1}}U^\top$. The result is the matrix PB, with blocks $PB_1 = P(\hat{T}^{\leqslant 1})^\top_1, \ldots, PB_{d^2} = P(\hat{T}^{\leqslant 1})^\top_{d^2}$. By definition, the singular values of $PB = T^{\leqslant 1}_{\{3\}\{1,2,4\}}$ are at most 1. Also, $T^{\leqslant 1}_{\{4\}\{123\}} = \hat{T}^{\leqslant 1}(P \otimes \mathrm{Id}_{d^2})$, and by the submultiplicativity of the norm, $\|\hat{T}^{\leqslant 1}(P \otimes \mathrm{Id}_{d^2})\| \leqslant \|\hat{T}^{\leqslant 1}\| \cdot \|P \otimes \mathrm{Id}_{d^2}\| \leqslant 1$, and the first reshaping still has spectral norm at most 1.

Finally, each reshaping step takes $O(d^4)$ time. Since we are only interested in truncating large singular values, it suffices for us to compute the SVD corresponding to singular values between $\sqrt{n}$ and 1. This can be done via subspace power iteration, which here involves the multiplication of a $d^3 \times d$ matrix and a $d \times d$ matrix (with intermediate orthogonalization steps for the $d \times d$ matrix, see Hardt and Price (2014)), which requires time $\widetilde{O}(d^{2+\omega})$, where ω is the matrix multiplication constant. Going forward the representation of the matrix will be as the original matrix, with the subtracted SVD corresponding to large singular values. ∎

We will need to argue that if the Frobenius norm of the error matrix is small, this is a sufficient condition under which we succeed. For this, we will use the following two lemmas. The first tells us that with probability $\widetilde{\Omega}(n^{-O(\varepsilon)})$ over the choice of g, we will have for some $i \in [n]$ that

$$M_g = c \cdot a_i a_i^\top + N,$$

where $|c| \geqslant \|N\|$ and furthermore $\|Na_i\|$ and $\|N^\top a_i\|$ are small:

Lemma 8 *Let* $g \sim \mathcal{N}(0, \mathrm{Id}_{d^2})$*. Suppose that* $\|T_{\{1,2,3\}\{4\}}\|, \|T_{\{1,2,3\}\{4\}}\| \leqslant 1$*, and that* $\|T - \sum_i a_i^{\otimes 4}\|_F \leqslant \varepsilon\sqrt{n}$*. Define the matrix* M_g *to be the flattening of* $\mathbf{T} = \sum_i a_i^{\otimes 4} + \mathbf{E}$ *along* g *in the*

modes $\{1\}$ *and* $\{2\}$*, and let* $|c|$ *be the magnitude of* a_j*'s projection onto* M_g*, i.e.*

$$M_g \stackrel{\text{def}}{=} \sum_{j=1}^{d^2} \langle e_j, g\rangle \cdot T_j \;=\; c \cdot a_j a_j^\top + N\,.$$

Then for a $1 - 3\delta$ *fraction of* $j \in [n]$*,*

$$\mathop{\mathbb{P}}_{g}\left[\; |c| \geqslant (1+\beta)\|N\|,\; \|N^\top a_j\|, \|Na_j\| \leqslant (\varepsilon/\delta)(c+\sqrt{2}+o(1))\right] = \widetilde{\Omega}\left(n^{-\left(\frac{1+\beta}{1-(1+\beta)\varepsilon/\delta}\right)^2}\right).$$

In particular, if $\delta = \Omega(1)$*,* $\beta = O(\varepsilon)$*,* $\beta < 1$*, then this probability is* $\widetilde{\Omega}(n^{-(1+O(\varepsilon))})$*.*

The proof consists primarily of the application of concentration inequalities, and we provide it below in Section 4.4.

The second lemma states that if M_g indeed has the form above, the top singular vectors of M_g must be close to the component a_i.

Lemma 9 *Let* M_g *be an* $n \times n$ *matrix, and* $a_1 \in \mathbb{R}^n$*, and suppose that*

$$M_g = c \cdot a_1 a_1^\top + N$$

with $|c| \geqslant (1+\beta)\|N\|$ *for* $\beta > 0$*, and* $\|Na_1\|, \|N^\top a_1\| \leqslant \varepsilon|c|$ *so that the relationship* $\frac{2\varepsilon(1+\beta)}{\beta} < 0.01$ *holds. Then letting* u *be a top singular vector of* M_g*, it follows that*

$$\langle u, a_1\rangle^2 \geqslant 0.99.$$

The proof requires some careful calculations, but is not complicated, and we will prove it below in Section 4.4.

Finally, we are ready to stitch these arguments together and prove that Algorithm 2 works.

Proof [Proof of Theorem 6] After reshaping and truncating $\mathbf{T}$ in step 1 of the algorithm, by Lemma 7 the matrix $T^{\leqslant 1} = \sum_i a_i^{\otimes 4} + E$ has the properties that

$$\|T^{\leqslant 1}_{\{1,2,3\}\{4\}}\|, \|T^{\leqslant 1}_{\{1,2,4\}\{3\}}\| \leqslant 1,$$

and also that still $\|E\|_F \leqslant \eta\sqrt{n}$.

We can now apply Lemma 8 with $\delta = \frac{1}{300}$ and $\beta = 400\eta/\delta = O(\eta)$ to conclude that for at least a 0.99-fraction of the $i \in [n]$, with probability at least $\widetilde{O}(n^{-1-O(\eta)})$, we will have

$$M_g = c \cdot a_i a_i^\top + N,$$

where $\|N\| \leqslant (1+\beta)c$ and $\|Na_i\|, \|N^\top a_i\| \leqslant 48\eta \cdot |c|$. Applying Lemma 9, we have that either the left- or right- top unit singular vector u of M_g has correlation at least

$$\langle u, a_i\rangle^2 \geqslant 0.99,$$

as desired.

For runtime, by our arguments in Lemma 7 step 1 takes time $\widetilde{O}(d^{2+\omega})$. After this, with either representation of our matrix $T^{\leqslant 1}$ (whether we compute the full truncated SVD or have the original

matrix minus the subtracted SVD), performing power iteration to find the top eigenvector with the flattening M_g takes time $\widetilde{O}(d^4)$, and finding a single component takes $\widetilde{O}(n^{-O(\eta)})$ samples of random contractions. Since we can reuse $T^{\leqslant 1}$ with new random contractions, the total runtime for recovering one component is $\widetilde{O}(d^{2+\omega}n^{-O(\eta)})$, and by the independence of the runs recovering m components requires $\widetilde{O}(d^{2+\omega}) + mn^{-O(\eta)} \cdot \widetilde{O}(d^4)$ time. ∎

With the core of our algorithm in place, we now take care of the remaining technical issues: recovery precision, working with near-orthonormal vectors, and recovering the full set of component vectors.

4.1. Postprocessing for closer vectors

Because the precision of recovery will be important in not amplifying the error, we begin with our precision-amplifying postprocessing algorithm.

Algorithm 3 (Postprocessing for error reduction)
Input: *A tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$*, a vector* $u \in \mathbb{R}^{d^2}$*, and an error parameter* $\varepsilon \geqslant \|E_{\{12\}\{34\}}\|$.

1. *Compute the matrix-vector product* $a \stackrel{\text{def}}{=} T_{\{1,2\}\{3,4\}}(u \otimes u)$.
2. *Reshape* $a \in \mathbb{R}^{d^2}$ *to a* $d \times d$ *matrix* A*, and compute the top left- and right- singular vectors* v_L *and* v_R *of* A.

Output: *If for one of* $v \in \{v_L, v_R\}$, $(v^{\otimes 2})^\top T v^{\otimes 2} \geqslant (1-3\varepsilon)^2 - \varepsilon$*, output* v.

Lemma 10 *Suppose that* v *is a unit vector with* $\langle v, a_i\rangle^2 \geqslant 0.99$*, and* $T = \sum_i a_i^{\otimes 4} + E$ *for* $\|E\| \leqslant \varepsilon$ *and* $a_1, \ldots, a_n$ *orthonormal. Then if we let* A *be the reshaping of* $T(v \otimes v)$ *to a* $d \times d$ *matrix, and if we let* u_L, u_R *be the top left- and right- unit singular vectors of* M*, then*

$$\langle u_L, a_i\rangle^2 \geqslant 1 - 3\varepsilon \quad \text{or} \quad \langle u_R, a_i\rangle^2 \geqslant 1 - 3\varepsilon.$$

In other words, Algorithm 3 succeeds. Further, the time required is $\widetilde{O}(d^4)$.

Proof [Proof of Lemma 10] For convenience and without loss of generality, let $i := 1$, and let $\alpha \stackrel{\text{def}}{=} 1 - \langle a_1, v\rangle^2 \leqslant 0.01$. Because $a_1, \ldots, a_n$ are orthonormal, we can write $v = \sum_j \langle a_j, v\rangle \cdot a_j + w$, where $w \perp a_j$ for all $j \in [n]$. By assumption, $\langle a_1, v\rangle^2 \geqslant 1 - \alpha$, and therefore $\sum_{j>1}\langle v, a_j\rangle^2 + \|w\|^2 \leqslant \alpha$. Now,

$$\begin{aligned}(T-E)(v\otimes v) &= \sum_j a_j^{\otimes 2}(a_j^{\otimes 2})^\top \left(\sum_j \langle a_j, v\rangle a_j \otimes \sum_j \langle a_j, v\rangle a_j\right)\\ &= \sum_{j,k,\ell} a_j^{\otimes 2}\langle a_j, a_k\rangle\langle a_j, a_\ell\rangle\langle a_k, v\rangle\langle a_\ell, v\rangle\end{aligned}$$

by the orthonormality of the a_i,

$$= \sum_j a_j^{\otimes 2}\langle a_j, v\rangle^2 .$$

Therefore, defining M to be the $n \times n$ reshaping of $T(v \otimes v)$ and defining N to be the $n \times n$ reshaping of $E(v \otimes v)$,

$$M = (1-\alpha)a_1 a_1^\top + \sum_{j>1} \langle a_j, v\rangle^2 a_j a_j^\top + N,$$

where $\|N\|_F \leqslant \varepsilon$, since $\|E\| \leqslant \varepsilon$.

Now, we have that

$$\|M\| \geqslant 1 - \alpha - \varepsilon,$$

and that if we choose η so that $1 - \eta \geqslant 2\alpha \geqslant 1/50$,

$$\|M - \eta \cdot a_1 a_1^\top\| \leqslant 1 - \alpha - \eta + \varepsilon.$$

Thus, $\|M - \eta a_1 a_1^\top\| \leqslant \|M\| - \eta + 2\varepsilon$, and we have by Fact 2 (see Appendix A for a proof) that the top unit eigenvector u of M is such that

$$\langle u, a_1\rangle^2 \geqslant \frac{\eta - 2\varepsilon}{\eta} \geqslant 1 - \frac{2\varepsilon}{\eta}.$$

Choosing $\eta = 49/50$, the result follows. ■

4.2. Near-orthonormal components

We'll now dispense with the discrepancy between the orthonormal and near-orthonormal cases.

Fact 1 *If $S = \sum_i a_i a_i^\top = \mathrm{Id} + E$ for $\|E\| \leqslant \varepsilon$, then $\tilde{a}_1 = S^{-1/2}a_1, \ldots, \tilde{a}_n = S^{-1/2}a_n$ are orthonormal, $\langle \tilde{a}_i, a_i\rangle^2 \geqslant (1-\varepsilon)\|a_i\|^2$, and*

$$\left\| \sum_i \tilde{a}_i^{\otimes 4} - a_i^{\otimes 4} \right\| \leqslant 4\sqrt{\varepsilon}.$$

Proof The fact that the $\tilde{a}_i$ are orthonormal follows because they are independent and have Gram matrix Id. Using the fact that the eigenvalues of S are between $(1-\varepsilon)^{-1/2}$ and $(1+\varepsilon)^{-1/2}$, quantity $\langle a_i, \tilde{a}_i\rangle^2 = \left(a_i^\top S^{-1/2} a_i\right)^2 \geqslant \frac{\|a_i\|^2}{1+\varepsilon} \geqslant \|a_i\|^2(1-\varepsilon)$. Finally,

$$\sum_i \tilde{a}_i^{\otimes 4} - a_i^{\otimes 4} = (S^{-1/2})^{\otimes 2} \left(\sum_i a_i^{\otimes 4} \right) (S^{-1/2})^{\otimes 2} - \sum_i a_i^{\otimes 4}$$

and because $\|(S^{-1/2})^{\otimes 2} - \mathrm{Id}\| \leqslant \sqrt{\varepsilon}$, and $\sum_i a_i^{\otimes 4} \preceq \sum_{ij} a_i a_i^\top \otimes a_j a_j^\top$, this difference has spectral norm at most $3\varepsilon \|\sum_i a_i^{\otimes 4}\| \leqslant 3\sqrt{\varepsilon}(1+\varepsilon)^2 \leqslant 4\sqrt{\varepsilon}$. ■

4.3. Full Recovery

Now we give the full algorithm, which will remove the components we find in each step from the tensor without amplifying the spectral norm of the error too much.

Algorithm 4 (Full tensor decomposition)
Input: *A tensor* $\mathbf{T} \in (\mathbb{R}^d)^{\otimes 4}$*, and the error parameter* $\varepsilon \geqslant \|E_{\{1,2\}\{3,4\}}\|$.

1. *Initialize the set of known components* $K = \emptyset$*, and the set of components under inspection* $B = \emptyset$.
2. *Initialize a working copy of* $\mathbf{T}$*,* $\mathbf{T}^{(0)}_{work}$*, and keep a clean copy of* $\mathbf{T}$ *called* $\mathbf{T}_{clean}$.
3. *For* $t = 0, \dots, 100 \log n$*,*
   (a) *Preprocess* $\mathbf{T}^{(t)}_{work}$ *with Algorithm 1, then run Algorithm 2 with* $\mathbf{T}^{(t)}_{work}$ $\widetilde{O}(n)$ *times, then postprocess with Algorithm 3 using* $\mathbf{T}_{clean}$ *and error parameter* ε*. If this produces an output vector* v*, add* v *to* B *(unless* K *already contains a vector that is* $1 - \varepsilon$ *correlated with* v*).*
   (b) *Let* $B = b_1, \dots, b_m$*, and abuse notation by letting* B *also be the matrix whose* i*th row is* b_i*. Compute the singular value decomposition* $B = U\Sigma V^\top$*, and compute the orthonormalized set* $\tilde{B} = \{\tilde{b}_i\} = \{U\Sigma^{-1}U^\top b_i\}_{i=1}^m$.
   (c) *Remove from* $\tilde{B}$ *any* $\tilde{b}_i$ *for which* $\langle \mathbf{T}_{clean}, \tilde{b}_i^{\otimes 4} \rangle < (1 - 6\varepsilon)^2 - \varepsilon$.
   (d) *Update the known components: set* $K := K \cup \tilde{B}$*, and set* $B, \tilde{B} := \emptyset$.
   (e) *Update the working tensor by removing known components: set* $\mathbf{T}^{(t+1)}_{work} := \mathbf{T}^{(t)}_{work} - \sum_{\tilde{b} \in \tilde{B}} \tilde{b}^{\otimes 4}$.

Output: *The set of known components* K.

Theorem 11 *Given* $\mathbf{T} = \sum_{i=1}^n a_i^{\otimes 4} + E$ *where the* a_i *are orthonormal and* $\|E_{\{1,2\}\{3,4\}}\| \leqslant \varepsilon$*, then if* $\varepsilon < O(\eta^2/\log^2 n)$*, with probability* $1 - o(1)$*, Algorithm 4 recovers orthonormal vectors* $b_1, \dots, b_n$ *so that there exists a permutation* $\pi : [n] \to [n]$ *such that for each* $i \in [n]$*,*

$$\langle a_i, b_{\pi(i)} \rangle^2 \geqslant 1 - 3\varepsilon.$$

Furthermore, this requires runtime $\widetilde{O}(n^{1+O(\eta)} d^{2+\omega})$.

First, we prove that if we have an *orthonormal* basis that approximates $a_1, \dots, a_k$, we can subtract it without introducing a large spectral norm error—this motivates and justifies steps 3(b)–3(e).

Lemma 12 *Let* $a_1, \dots, a_k \in \mathbb{R}^d$ *and* $b_1, \dots, b_k \in \mathbb{R}^d$ *be two sets of orthonormal vectors, such that* $\langle a_i, b_i \rangle^2 \geqslant 1 - \varepsilon$*. Then*

$$\left\| \sum_i a_i^{\otimes 4} - b_i^{\otimes 4} \right\|_2 \leqslant 4\sqrt{\varepsilon}$$

Proof Define the matrices $U, V \in \mathbb{R}^{d^2 \times k}$ so that the ith column of U (or V) is equal to $a_i^{\otimes 2}$ ($b_i^{\otimes 2}$ respectively). We have that

$$\sum_i a_i^{\otimes 4} - b_i^{\otimes 4} = UU^\top - VV^\top = (U - V)(U + V)^\top.$$

So it suffices for us to bound $\|U - V\| \cdot \|U + V\|$.

By the subadditivity of the norm, $\|U+V\| \leqslant \|U\| + \|V\| = 2$. Meanwhile, the singular values of $U - V$ are the square roots of the eigenvalues of $\|(U-V)^\top(U-V)\|$, and so we bound

$$\begin{aligned}(U-V)^\top(U-V) &= UU^\top + VV^\top - U^\top V - V^\top U \\ &= 2\,\mathrm{Id}_k - U^\top V - V^\top U, \end{aligned} \tag{4.1}$$

where the second line follows because U and V have orthonormal columns. Now, by assumption we know that

$$U^\top V = (1-\varepsilon)\cdot \mathrm{Id}_k + E,$$

where for $i \neq j$, $E_{ij} = \langle b_i, a_j\rangle^2$ and $E_{ii} = \langle b_i, a_i\rangle^2 - (1-\varepsilon)$, and by the orthonormality of the a_j,

$$\sum_j |E_{ij}| = \sum_j \langle b_i, a_j\rangle^2 = \varepsilon.$$

So the 1-norm of the rows of E is at most ε. By the orthonormality of the b_j, the same holds for the 1-norm of the columns, $\sum_i |E_{ij}|$. It follows that $\|E\| \leqslant \varepsilon$. Therefore,

$$U^\top V + V^\top U = 2(1-\varepsilon)\,\mathrm{Id}_k + \hat{E},$$

where $\|\hat{E}\| \leqslant 2\varepsilon$. Returning to (4.1), we can conclude that $\|(U-V)\| \leqslant 2\sqrt{\varepsilon}$, and we have our result. ∎

Now, we will prove that by orthogonalizing, we do not harm too many components $\tilde{b}_i$.

Lemma 13 *Suppose $a_1, \ldots, a_k \in \mathbb{R}^d$ are orthonormal vectors, and $u_1, \ldots, u_k \in \mathbb{R}^d$ are unit vectors such that $\|u_i - a_i\|_2^2 \leqslant \varepsilon$. Let U be the $d \times k$ matrix whose ith column is u_i, $U = X\Sigma Y$ be the singular value decomposition of U, and let $\tilde{u}_i = X\Sigma^{-1}X^\top u_i$. Then for a $1-\delta$ fraction of $i \in [k]$,*

$$\langle u_i, \tilde{u}_i\rangle \geqslant 1 - \varepsilon/2\delta.$$

Proof For convenience, let A be the $d \times k$ matrix whose ith column is a_i, and let $\tilde{U} = X\Sigma^{-1}X^\top U$, and $\tilde{u}_i = X\Sigma^{-1}X^\top u_i$. Let $\mathbb{X}$ be the space of all real $d \times k$ real matrices with orthonormal columns, and notice that $\tilde{U}, A \in \mathbb{X}$ and that $\tilde{U}$ is closer to U than A. Indeed, for any matrix X with orthonormal columns,

$$\|X-U\|_F^2 \;=\; k + \|U\|_F^2 - 2\langle U, X\rangle \;\geqslant\; k + \|U\|_F^2 - 2\|X\|\|U\|_* \;=\; k + \|U\|_F^2 - 2\|U\|_* \;=\; \|\tilde{U}-U\|_F^2,$$

Where we have used that the spectral norm and nuclear norm are dual. Therefore,

$$\|\tilde{U}-U\|_F^2 \;\leqslant\; \|A-U\|_F^2 \;=\; \sum_i \|a_i - u_i\|_2^2 \;=\; \varepsilon\cdot k$$

And on average, $\varepsilon \geqslant \|\tilde{u}_i - u_i\|_2^2$, so by Markov's inequality, for at least $(1-\delta)k$ of the u_i, $\langle u_i, \tilde{u}_i\rangle \geqslant 1 - \varepsilon/2\delta$. ∎

Finally, we are ready to prove that Algorithm 4 works.

Proof [Proof of Theorem 11] We claim that in the tth iteration of step 3, with high probability we have at most $0.45^t n$ components remaining to be found, and that $\mathbf{T}^{(t)}_{work} = \mathbf{T} + F^{(t)}$ where $\|F^{(t)}\| \leqslant 8t\sqrt{\varepsilon}$. For $t = 0$, this is easily true.

Now assume this holds for t, and we will prove it for $t+1$. Since by assumption $t\sqrt{\varepsilon}\log n \leqslant 100\sqrt{\varepsilon}\log n \leqslant O(\eta)$, applying Lemma 5 and Theorem 6 to the running of preprocessing Algorithm 1 and the main step Algorithm 2 with $\mathbf{T}^{(t)}_{work}$ and Lemma 10 to the running of the postprocessing Algorithm 3 with $\mathbf{T}_{clean}$, in step 3(a) with high probability we will find $m \geqslant 0.9n_t$ vectors $b_1, \dots, b_m$ so that $\langle b_i, a_i\rangle^2 \geqslant 1 - 3\varepsilon$. Furthermore, this takes a total of $\widetilde{O}(mn^{O(\eta)}d^{2+\omega})$ time.

By Lemma 13, in step 3(c) we will remove no more than a half of the $\tilde{b}_i$, while maintaining $\langle \tilde{b}_i, a_i\rangle^2 \geqslant 1 - 3\varepsilon$ (where we are abusing notation by re-indexing conveniently), so that $n_{t+1} \geqslant 0.45n_t$. Finally, by Lemma 12, we have that

$$\left\| \sum_{i=1}^{|\tilde{B}|} a_i^{\otimes 4} - \tilde{b}_i^{\otimes 4} \right\|_2 \leqslant 4\sqrt{3\varepsilon},$$

So that in step 3(e),

$$\begin{aligned}\mathbf{T}^{(t+1)}_{work} &= \left(\mathbf{T}^{(t)}_{work} - \sum_{i=1}^{|\tilde{B}|} a_i^{\otimes 4}\right) + \left(\sum_{i=1}^{|\tilde{B}|} a_i^{\otimes 4} - \sum_{\tilde{b}_i \in \tilde{B}} \tilde{b}_i^{\otimes 4}\right) \\ &= \mathbf{T}^{(t-1)}_{work} + F,\end{aligned}$$

for a matrix F with $\|F\| \leqslant 8\sqrt{\varepsilon}$. By induction, this implies that $\mathbf{T}^{(t+1)}_{work} = \mathbf{T} + F^{(t+1)}$ where $\|F^{(t+1)}\| \leqslant \|F\| + \|F^{(t)}\| \leqslant (t+1)8\sqrt{\varepsilon}$.

Taking a union bound over the high-probability success of Algorithm 2, we have that after $t = O(\log n)$ steps we have found all of the components. We have spent a total of $\widetilde{O}(n^{1+O(\eta)}d^{2+\omega})$ time in step 3(a). Finally, steps 3(b)-3(e) of Algorithm 4 require no more than $\widetilde{O}(d^3)$ time, and since the entire loop runs $\widetilde{O}(1)$ times, we have our result. ■

4.4. Supporting Lemmas

Now we circle back and prove the omitted supporting lemmas.

Proof [Proof of Lemma 8] For convenience, fix $j := 1$. Let $g^{(1)}$ be the component of g in the direction $a_1^{\otimes 2}$, and let $g^{(>1)}$ be the component of g orthogonal to $a_1^{\otimes 2}$. Notice that $g^{(1)}, g^{(>1)}$ are independent.

By the orthogonality of the a_i, our matrix M_g can be written as

$$\begin{aligned}M_g &= \langle g^{(1)}, a_1^{\otimes 2}\rangle \cdot a_1 a_1^\top + \sum_{j=1}^{d^2} (g_j^{(1)} + g_j^{(>1)}) \cdot (S - a_i^{\otimes 4} + E)_j \\ &= \langle g^{(1)}, a_1^{\otimes 2}\rangle \cdot a_1 a_1^\top + \left(\sum_{j=1}^{d^2} g_j^{(1)} \cdot E_j\right) + \left(\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j\right),\end{aligned}$$

where T_j is the jth matrix slice of T. For convenience, we can refer to the two sums on the right as

$$N = \left(\sum_{j=1}^{d^2} g_j^{(1)} \cdot E_j\right) + \left(\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j\right).$$

First, we get a lower bound on the probability that the coefficient of $a_1 a_1^\top$ is large. Let $\mathcal{G}_1(\alpha)$ be the event that $|\langle g^{(1)}, a_1^{\otimes 2}\rangle| = \|g^{(1)}\| \geqslant \sqrt{2\alpha \log n}$. By standard tail estimates on univariate Gaussians, we have that

$$\mathbb{P}[\mathcal{G}_1(\alpha)] \geqslant \tilde{O}(n^{-\alpha}).$$

Now, we bound $\|N\|$. Define the event $\mathcal{E}_{>1}(\rho)$ to be the even that

$$\mathcal{E}_{>1}(\rho) \stackrel{\text{def}}{=} \left\{\left\|\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j\right\| \leqslant \sqrt{2(1+\rho)\log d}\right\}$$

By Lemma 18, we can conclude that

$$\mathbb{P}\left[\mathcal{E}_{>1}(\rho)\right] \geqslant 1 - d^{-\rho}\,.$$

To bound $\|N\|$, it thus remains to understand the term

$$\sum_j g_j^{(1)} E_j = \langle g, a_1^{\otimes 2}\rangle \cdot \sum_j a_i^{\otimes 2}(j) \cdot E_j = \langle g, a_1^{\otimes 2}\rangle \cdot (a_i a_i^\top \otimes \mathrm{Id}_{d^2})E, \tag{4.2}$$

where the quantity $(a_i a_i^\top \otimes \mathrm{Id}_{d^2})E$ corresponds to the contraction of E along two modes by the vector $a_i^{\otimes 2}$. We make the following observation:

Observation 1 *If $P_1, \ldots, P_n$ are orthogonal projections from $\mathbb{R}^{n^4} \to K$ for some convex set K, then for a $1-\delta$ fraction of $i \in [n]$,*

$$\|P_i E\|_F \leqslant \varepsilon/\delta.$$

Proof This follows from the fact that

$$\varepsilon^2 n \geqslant \|E\|_F^2 \;\geqslant \sum_i \|P_i E\|_F^2,$$

and then by an application of Markov's inequality. ■

Now, note that $\sum_j a_i a_i^\top \otimes \mathrm{Id}_{d^2}$ for $i \in [n]$ are orthogonal projectors from $\mathbb{R}^{n^4}$ to $\mathbb{R}^{n^2}$. Thus it follows that for a $1-\delta$ fraction of $i \in [n]$, and without loss of generality assuming that $i = 1$ is among them, $\|(a_1 a_1^\top \otimes \mathrm{Id}_{d^2})E\|_F \leqslant \varepsilon/\delta$. Therefore for any unit vectors $u, v \in \mathbb{R}^d$, returning to (4.2),

$$\begin{aligned}\left|u^\top\left(\sum_{j=1}^{d^2} g_j^{(1)} \cdot E_j\right)v\right| &= \left|\langle g, a_1^{\otimes 2}\rangle \cdot \langle uv^\top, (a_1 a_1^\top \otimes \mathrm{Id}_{d^2})E\rangle\right| \\ &\leqslant \|g^{(1)}\| \cdot \left\|(a_1 a_1^\top \otimes \mathrm{Id}_{d^2})E\right\|_F \cdot \|uv^\top\|_F \;\leqslant\; \frac{\varepsilon}{\delta}\|g^{(1)}\|\,.\end{aligned}$$

Thus, combining the above we have a two-part upper bound on $\|N\|$.

Finally, define the event $\mathcal{E}_{a_1,E}(\theta)$ to be the event that

$$\mathcal{E}_{a_1,E}(\theta) \stackrel{\text{def}}{=} \left\{ \left\| \left(\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j \right) a_1 \right\|_2, \left\| \left(\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j \right)^\top a_1 \right\|_2 \leqslant \frac{\varepsilon}{\delta} \cdot \sqrt{2(1+\theta)} \right\}$$

Examining this form, we can split

$$\begin{aligned} \left(\sum_{j=1}^{d^2} g_j^{(>1)} \cdot T_j \right) a_1 &= \sum_{j=1}^{d^2} g_j^{(>1)} \cdot (S - a_1^{\otimes 4})_j a_1 + \sum_{j=1}^{d^2} g_j^{(>1)} \cdot E_j a_1 \\ &= \sum_{j=1}^{d^2} g_j^{(>1)} \cdot E_j a_1 \end{aligned}$$

where the last line follows because the a_i are orthogonal. We note that $\sum_j g_j^{(>1)} E_j a_1$ is a Gaussian contraction of the form $(a_1 \otimes \mathrm{Id}_{d^3})E$. Again appealing to Observation 1 and to the fact that the $a_i \otimes \mathrm{Id}_{d^3}$ are orthogonal projections, we conclude that for a $1-\delta$ fraction of $i \in [n]$, $\|(a_1 \otimes \mathrm{Id}_{d^3})E\|_F \leqslant \varepsilon/\delta$. Without loss of generality we assume that this is true for $i = 1$, from which it follows by Lemma 18 that

$$\mathbb{P}\left[\left\| \sum_{j=1}^{d^2} g_j^{(>1)} \cdot E_j a_1 \right\|_2 \leqslant \frac{\varepsilon}{\delta} \sqrt{2(1+\theta)} \right] \geqslant 1 - d^{-\theta}.$$

We can apply the same arguments to $\left(\sum_j g_j^{(>1)} E_j \right)^\top a_1$, and we conclude that

$$\mathbb{P}\left[\mathcal{E}_{a_1,E}(\theta) \right] \geqslant 1 - 2d^{-\theta}.$$

Now, by the union bound $\mathcal{E}_{>1}(\rho)$ and $\mathcal{E}_{a_1,E}$ both occur with probability at least $1 - d^{-\rho} - 2d^{-\theta}$. Also, we notice that $\mathcal{E}_{>1} \cup \mathcal{E}_{a_1,E}$ and $\mathcal{G}_1$ are independent. Therefore, for $\rho, \theta \geqslant \log\log n / \log n$,

$$\mathbb{P}[\mathcal{G}_1(\alpha), \mathcal{E}_{>1}(\rho), \mathcal{E}_{a_1,E}(\theta)] \geqslant \tilde{O}(n^{-\alpha}).$$

Conditioning on $\mathcal{E}_{>1}$ and $\mathcal{G}_1$,

$$M_g = c \cdot a_1 a_1^\top + N,$$

where $|c| \geqslant \sqrt{2\alpha \log n}$, and N is a matrix of norm at most $(\varepsilon/\delta)c + \sqrt{2(1+\rho)\log d}$, such that $\|Na_1\|, \|N^\top a_1\| \leqslant (\varepsilon/\delta)(c + \sqrt{2(1+\theta)})$.

We now set α so that $|c| \geqslant \beta\|N\|$. This occurs when

$$\alpha \geqslant \left(\beta \frac{1}{1 - \beta(\varepsilon/\delta)} \right)^2 (1+\rho).$$

Choosing $\beta = 1 + \beta'$, $\rho, \theta = \log\log n / \log n$, we have our conclusion for a_1, and by symmetry for all other a_i in the $1 - 3\delta$ fraction of $i \in [n]$ for which the Frobenius norms of the contractions are small. ∎

Proof [Proof of Lemma 9] Assume without loss of generality that $c \geqslant 0$. Choose $\kappa = \frac{2\varepsilon|c|}{\delta} \leqslant |c|\left(1 - \frac{1}{1+\beta}\right)$. When $\kappa \cdot a_1 a_1^\top$ is subtracted from M_g, then given any unit vector $v \in \mathbb{R}^d$ with $|\langle v, a_1\rangle| = \alpha$, we can write $v = \alpha a_1 + w$ where $\langle w, a_1\rangle = 0$ and $\|w\| = \sqrt{1-\alpha^2}$. Examining the action of $M_g - \kappa a_1 a_1^\top$ on v,

$$\begin{aligned}\|(M_g - \kappa \cdot a_1 a_1^\top)v\|_2^2 &= \|(c-\kappa)\alpha a_1 + Nv\|_2^2 \\ &\leqslant (c-\kappa)^2\alpha^2 + (c-\kappa)\alpha a_1^\top Nv + (c-\kappa)\alpha v^\top N a_1 + v^\top N^\top N v\end{aligned}$$

applying the Cauchy-Schwarz inequality and our bounds on $\|N^\top a_1\|, \|Na_1\|$,

$$\leqslant (c-\kappa)^2\alpha^2 + 2(c-\kappa)\alpha\varepsilon c + v^\top N^\top N v\,.$$

Now expanding the $v^\top N^\top N v$ term along the components of v,

$$\begin{aligned}v^\top N^\top N v &= (\alpha a_1 + w)^\top N^\top N(\alpha a_1 + w) \\ &\leqslant (\alpha\|Na_1\| + \|w\|\|N\|)^2\end{aligned}$$

and since $\|w\| = \sqrt{1-\alpha^2}$, $\|Na_1\| \leqslant \varepsilon c$, and $\|N\| \leqslant c/(1+\beta) \leqslant c(1-\beta+2\beta^2)$,

$$\leqslant \left(\alpha\varepsilon c + \sqrt{1-\alpha^2}(1-\beta+2\beta^2)c\right)^2,$$

and putting these together,

$$\|(M_g - \kappa \cdot a_1 a_1^\top)v\|_2^2 \leqslant (c-\kappa)^2\alpha^2 + 2(c-\kappa)\alpha\varepsilon c + \left(\alpha\varepsilon c + \frac{\sqrt{1-\alpha^2}}{1+\beta}c\right)^2$$

It is easy to see that when $c/(1+\beta) < c - \kappa$, this quantity is maximized at $\alpha = 1$, and so by our choice of κ we have that

$$\|(M_g - \kappa \cdot a_1 a_1^\top)v\|_2^2 \leqslant (c-\kappa)^2 + 2\varepsilon c(c-\kappa) + \varepsilon^2 c^2 = (c(1+\varepsilon) - \kappa)^2$$

and thus $\|M_g - \kappa a_1 a_1^\top\| \leqslant (1+\varepsilon)c - \kappa$.

Now we will lower bound $\|M_g\|$.

$$\begin{aligned}\|M_g\| \;\geqslant\; a_1^\top M_g a_1 &= c + a_1^\top N a_1 \\ &\geqslant c - \|a_1\|\|Na_1\| \\ &\geqslant c(1-\varepsilon)\,.\end{aligned}$$

Where we have applied the Cauchy-Schwarz inequality, and the assumption that $\|Na_1\| \leqslant \varepsilon c$. It follows that

$$\|M_g - \kappa a_1 a_1^\top\| \leqslant \|M_g\| + 2\varepsilon c - \kappa.$$

Finally applying Fact 2, we can conclude that for either the left- or right-singular unit vector u of M_g,

$$\langle a_i, u\rangle^2 \geqslant \frac{\kappa - 2\varepsilon c}{\kappa} \;\geqslant 1 - \delta\,.$$

Choosing $\delta = \frac{2\varepsilon(1+\beta)}{\beta}$ as small as possible, we have our result. ■

5. Learning Orthonormal Dictionaries

Here, we show how to use our tensor decomposition algorithm to learn dictionaries with orthonormal basis vectors.

Problem 1 *Given access to a dictionary $A \in \mathbb{R}^{d\times d}$ with independent columns $a_1, \dots, a_d$, in the form of samples $y^{(1)} = Ax^{(1)}, \dots, y^{(m)} = Ax^{(m)}$ for independent $x^{(i)}$, recover A.*

Below, in Section 5.3, we will prove that $\widetilde{O}(n^3)$ samples suffice to estimate the 4th moment tensor within $o(1)$ spectral norm error. Computing this matrix from $\widetilde{O}(n^3)$ samples takes $\widetilde{O}(n^3 d^4)$ time. Thus we can equivalently formulate the problem as follows:

Problem 2 *Given access to $A \in \mathbb{R}^{d\times n}$ with independent columns $a_1, \dots, a_n$, via a noisy copy of the 4th moment tensor $\mathbf{T} = \mathbb{E}[(Ax)^{\otimes 4}] + E$, recover the columns $a_1, \dots, a_n$.*

Finally, given access to a sufficiently large number of samples, we can reduce to the case where the columns of A are orthogonal:

Lemma 14 *Suppose that the samples $y = Ax$ are generated from a distribution over x for which $\mathbb{E}[x_i^2] = \mathbb{E}[x_j^2]$ for all $i, j \in [n]$, and that the columns of A are independent. Then there exists an efficient reduction from the case when A has independent columns to the case when A has orthogonal columns, with sample complexity growing polynomially with the condition number of $\Sigma = AA^\top$.*

The proof is straightforward, involving a transformation by the empirical covariance matrix, and we give it below in Section 5.2.

Now, we reduce the dictionary learning problem to tensor decomposition. In Section 3, we explained that the 4th moment tensor itself may be far from our target tensor $\sum_i a_i^{\otimes 4}$, due to the presence of a low-rank, high-Frobenius norm component. The sum-of-squares algorithm for this problem can overcome this difficulty by exploiting the SDP's symmetry constraints. In our algorithm, we will exploit this symmetry manually to go from $\mathbb{E}[(Ax)^{\otimes 4}]$ to a tensor that approximates $\sum_i a_i^{\otimes 4}$ well in Frobenius norm.

Algorithm 5 (Preprocessing for dictionary learning)
Input: *A noisy copy of the fourth moment tensor $\mathbf{T} = \mathbb{E}[(Ax)^{\otimes 4}] + E$, truncation parameter ε.*

1. *Reshape $\mathbf{T}$ to $T_{\{1,2\}\{3,4\}}$, and perform the eigenvalue truncation*

$$T^{>\varepsilon} \stackrel{\text{def}}{=} (T - \varepsilon \operatorname{Id})_+,$$

   where $+$ denotes projection to the PSD cone.
2. *Compute the truncation of a different reshaping of $T^{>\varepsilon}$,*

$$\tilde{T} = (\sigma(T^{>\varepsilon}_{\{1,3\}\{2,4\}}) - \varepsilon \operatorname{Id})_+$$

Output: *The tensor $\tilde{T}$ as an approximation of $\sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4}$.*

Our claim is that this produces a tensor that is close to $\sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4}$ in Frobenius norm.

Theorem 15 *If the x are independent and distributed so that $\mathbb{E}[x_i x_j x_k x_\ell] = 0$ unless $x_i x_j x_k x_\ell$ is a square, and so that for all $i, j \in [n]$, $\mathbb{E}[x_i^2 x_j^2] \leqslant \alpha\, \mathbb{E}[x_1^4]$ for $\alpha < 1$, then given access to $\mathbf{T} = \mathbb{E}[(Ax)^{\otimes 4}] + E$ where $\|E\| \leqslant \alpha$, Algorithm 5 with $\varepsilon = 3\alpha$ returns a tensor $\tilde{T}$ such that*

$$\left\| \tilde{T} - \sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4} \right\|_F \leqslant 9\alpha\sqrt{n}.$$

Proof For convenience denote by $T \stackrel{\text{def}}{=} (\mathbb{E}[(Ax)^{\otimes} 4] + E)_{\{1,2\}\{3,4\}}$, and define $S = \sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4}$. We have that

$$\begin{aligned} T - E \;=\; \mathbb{E}[(Ax)^{\otimes 4}] &= \sum_{i,j,k,\ell=1}^{d} \mathbb{E}[x_i x_j x_k x_\ell] \cdot (a_i \otimes a_j)(a_k \otimes a_\ell)^\top \\ &= \sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4} + \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot \left((a_i^{\otimes 2})(a_j^{\otimes 2})^\top + a_i a_i^\top \otimes a_j a_j^\top + a_j a_i^\top \otimes a_i a_j^\top \right) \end{aligned}$$

The latter term can be split into three distinct matrices—the first is a potentially low-rank matrix, and may have large eigenvectors.[8] The latter two terms have small spectral norm.

Claim 1

$$\left\| \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot a_i a_i^\top \otimes a_j a_j^\top \right\| \leqslant \alpha\, \mathbb{E}[x_1^4], \quad \text{and} \quad \left\| \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot a_j a_i^\top \otimes a_i a_j^\top \right\| \leqslant \alpha\, \mathbb{E}[x_1^4]\,.$$

We'll prove this claim below. Now, again for convenience define the matrix N to be the remaining term, $N = \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2](a_i^{\otimes 2})(a_j^{\otimes 2})^\top$. From Claim 1 and by our assumption on $\|E\|$,

$$T = S + N + \hat{E},$$

for $\|\hat{E}\| \leqslant 3\alpha$. On the other hand, if we let B be the $d \times n$ matrix whose ith column is $a_i^{\otimes 2}$, and we let X be the $n \times n$ matrix whose i, jth entry is $\mathbb{E}[x_i^2 x_j^2]$, then $S + N = BXB^\top$, and so $\operatorname{rank}(S + N) \leqslant n$.

It follows that when we perform the eigenvalue truncation in step 1 of Algorithm 5,

$$T^{<\varepsilon} = (T - 3\alpha \cdot \mathrm{Id})_+,$$

then we have that $\operatorname{rank}(T^{<\varepsilon}) \leqslant n$ as well. Also by definition of truncation, $T = T^{<\varepsilon} + \tilde{E}$, and because to begin with we had $T \succeq 0$, $\|\tilde{E}\| \leqslant 3\alpha$. Putting the above together, it follows that

$$\|T^{<\varepsilon} - (S + N)\|_F = \|\tilde{E} - \hat{E}\|_F \leqslant 6\alpha\sqrt{2n},$$

where we have used that $\operatorname{rank}(T^{<\varepsilon} - (S + N)) \leqslant 2n$ and $\|\tilde{E} - E\| \leqslant 6\alpha$. Now, we recall the reshaping operation on tensors from step 2 of Algorithm 5—in going from the reshaping $\{1, 2\}\{3, 4\}$ to $\{1, 3\}\{2, 4\}$, the rank-1 tensor $(a \otimes b)(c \otimes d)^\top)$ is reshaped to $(a \otimes c)(b \otimes d)^\top$. Let $\sigma(\cdot)$ denote this reshaping operation. Reshaping does not change the Frobenius norm. So by linearity, and since σ fixes S,

$$\|\sigma(T^{<\varepsilon}) - S - \sigma(N)\|_F = \|T^{<\varepsilon} - (S + N)\|_F \leqslant 6\alpha\sqrt{2n}.$$

8. For instance, in the case when $\mathbb{E}[x_i^2 x_j^2] = p^2$, this term is rank-1 and has spectral norm pn.

We now remark that σ maps N to one of the bounded-norm matrices from Claim 1,

$$\sigma(N) = \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot a_i a_i^\top \otimes a_j a_j^\top \preceq \alpha \cdot \mathrm{Id}\,.$$

Furthermore, because the positive semidefinite cone is a closed convex set, and because projection to closed convex sets can only decrease distances (see Lemma 19),

$$\begin{aligned}
\left\|\sigma(T^{<\varepsilon}) - S - \sigma(N)\right\|_F &= \left\|\sigma(T^{<\varepsilon}) - S - \alpha \cdot \mathrm{Id} - (\sigma(N) - \alpha \cdot \mathrm{Id})\right\|_F \\
&\geqslant \left\|\left(\sigma(T^{<\varepsilon}) - S - \alpha \cdot \mathrm{Id}\right)_+ - (\sigma(N) - \alpha \cdot \mathrm{Id})_+\right\|_F \\
&\geqslant \left\|\left(\sigma(T^{<\varepsilon}) - S - \alpha \cdot \mathrm{Id}\right)_+\right\|_F \\
&\geqslant \left\|\left(\sigma(T^{<\varepsilon}) - \alpha \cdot \mathrm{Id}\right)_+ - S\right\|_F\,,
\end{aligned}$$

where to obtain the last inequality we used that S is positive semidefinite. Therefore, step 3 of the algorithm ensures that $\tilde{T}$ is close to S in Frobenius norm, as desired. ■

Now we prove that the spectral norms of the symmetrizations of the tensor have small spectral norm.

Proof [Proof of Claim 1] The first matrix that we are interested in is PSD, and can dominated by a tensor power of the identity:

$$0 \preceq \sum_{i \neq j} \frac{\mathbb{E}[x_i^2 x_j^2]}{\mathbb{E}[x_1^4]} \cdot a_i a_i^\top \otimes a_j a_j^\top \preceq \alpha \sum_{i,j} a_i a_i^\top \otimes a_j a_j^\top \preceq \alpha \cdot \left(\sum_i a_i a_i^\top\right) \otimes \left(\sum_j a_j a_j^\top\right) \preceq \alpha \cdot \mathrm{Id}\,.$$

For the second matrix, if we let A be the $d^2 \times n^2$ matrix whose i,jth column is $a_i \otimes a_j$, and let M be the $n^2 \times n^2$ matrix whose i,jth diagonal entry is $\mathbb{E}[x_i^2 x_j^2]$, then

$$\sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot a_j a_i^\top \otimes a_i a_j^\top = AM\Pi A^\top,$$

where Π is the permutation matrix that takes the i,jth row to the j,ith row. By assumption, $\|M\| \leqslant \max_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \leqslant \alpha\, \mathbb{E}[x_1^2]$, and the columns of A are orthonormal, so $\|A\| = 1$. It follows by the submultiplicativity of the spectral norm that

$$\left\|\sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot a_j a_i^\top \otimes a_i a_j^\top\right\| \leqslant \alpha\, \mathbb{E}[x_1^4]\,.$$

This gives us the claim. ■

When $\mathbb{E}[x_i^4] = p$ for all $i \in [n]$, applying Algorithm 2 with $\tilde{T}$ a total of $\widetilde{O}(n)$ times will allow us to recover $m \geqslant n/2$ vectors $b_1, \ldots, b_m$ with $\langle a_i, b_i \rangle^2 \geqslant 0.99$. The following subsections contain the details regarding the refinement of the approximation, and the sample complexity bounds for estimating the 4th moment tensor.

5.1. Postprocessing to refine approximation

We now analyze the postprocessing algorithm Algorithm 3 for the context of dictionary learning, in which our tensor has the form $\mathbf{T} = \mathbb{E}[(Ax)^{\otimes 4}]$. We claim that, despite not having bounded spectral norm error away from $\sum_i \mathbb{E}[x_i^4] \cdot a_i^{\otimes 4}$, the postprocessing algorithm still succeeds.

Lemma 16 *Suppose that we are given* $\mathbf{T} = \mathbb{E}[(Ax)^{\otimes 4}]$*, where* x *is distributed so that* $\mathbb{E}[x_i x_j x_k x_\ell] = 0$ *unless* $x_i x_j x_k x_\ell$ *is a square,* $\mathbb{E}[x_i^4] = p$ *for all* $i \in [n]$*, and* $\mathbb{E}[x_i^2 x_j^2] \leqslant \alpha p$*. Suppose furthermore that we have a unit vector* u *such that* $\langle b, a_i\rangle^2 \geqslant 0.99$ *for some* $i \in [n]$*. Then applying Algorithm 3 to* u *and* $\mathbf{T}$ *with error parameter* $1/2$ *returns a vector* v *such that*

$$\langle v, a_i\rangle^2 \geqslant 1 - 16\alpha.$$

Proof Without loss of generality, let $i := 1$ so that $\langle b, a_i\rangle^2 = 1 - \eta \geqslant 0.99$ (henceforth, we use i as an ordinary index). We have that

$$\mathbb{E}[(Ax)^{\otimes 4}](u \otimes u) = p \sum_i \langle u, a_i\rangle^2 a_i^{\otimes 2} + \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \left(\langle u, a_j\rangle^2 \cdot a_i^{\otimes 2} + \langle u, a_i\rangle\langle u, a_j\rangle (a_i \otimes a_j + a_j \otimes a_i) \right)$$

Define M_u to be the reshaping of $\mathbb{E}[(Ax)^{\otimes 4}](u \otimes u)$ to a $d^2 \times d^2$ matrix. We must understand the spectrum of M_u, and for now we turn our attention to the second sum. Splitting the second sum into distinct parts, we have by the orthonormality of the a_i that

$$\sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot \langle u, a_j\rangle^2 \cdot a_i a_i^\top \preceq \max_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot \sum_{i \neq j} a_i a_i^\top \preceq \alpha p \cdot \mathrm{Id},$$

where the last line is by our assumption on $\mathbb{E}[x_i^2 x_j^2]$. Finally, for any $w \in \mathbb{R}^d$,

$$w^\top \left(\sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot \langle u, a_i\rangle\langle u, a_j\rangle (a_i a_j^\top + a_j a_i^\top) \right) w = \sum_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \cdot 2\langle u, a_i\rangle\langle u, a_j\rangle\langle w, a_i\rangle\langle w, a_j\rangle$$

Applying Cauchy-Schwarz and pulling out the maximum multiplier,

$$\leqslant 2 \max_{i \neq j}(E[x_i^2 x_j^2]) \cdot \left(\sum_{i \neq j} \langle u, a_i\rangle^2 \langle w, a_j\rangle^2 \right)^{1/2} \left(\sum_{i \neq j} \langle u, a_j\rangle^2 \langle w, a_i\rangle^2 \right)^{1/2}$$

Now noticing that the two parenthesized terms are actually identical, then adding a positive quantity and factoring,

$$\leqslant 2 \max_{i \neq j}(E[x_i^2 x_j^2]) \cdot \left(\left(\sum_i \langle u, a_i\rangle^2 \right) \left(\sum_j \langle w, a_j\rangle^2 \right) \right) = 2 \max_{i \neq j} \mathbb{E}[x_i^2 x_j^2] \leqslant 2p\alpha.$$

An identical proof, up to signs, gives us a lower bound of $2p\alpha$.

Therefore,

$$\frac{1}{p}M_u = \sum_i \langle u, a_i\rangle^2 a_i a_i^\top + E,$$

For a matrix E with $\|E\| \leqslant 4\alpha$.

It remains to argue that the top eigenvector of M_u is a_1. We have that

$$\begin{aligned}\|p^{-1}M_u\| &\geqslant a_1^\top M_u a_1 \\ &\quad a_1^\top \sum_i \langle u, a_i\rangle^2 a_i a_i^\top a_1 + a_1^\top E a_1 \\ &\geqslant 1 - \eta - 4\alpha,\end{aligned}$$

where the last line follows from the orthonormality of the a_i and our bound on $\|E\|$. Meanwhile, for any unit vector $w \in \mathbb{R}^d$ and any $\varepsilon < 1 - 2\eta$,

$$w^\top \left(p^{-1}M - \varepsilon \cdot a_1 a_1^\top\right) w = (1 - \eta - \varepsilon)\langle a_1, w\rangle^2 + \sum_{i>1}\langle u, a_i\rangle^2\langle a_i, w\rangle^2 + w^\top E w$$

and since $\max_{i>1}\langle a_i, u\rangle^2 \leqslant \eta$,

$$\begin{aligned}&\leqslant (1 - \eta - \varepsilon)\langle a_1, w\rangle^2 + \eta \cdot (1 - \langle a_1, w\rangle^2) + 4\alpha \\ &\leqslant 1 - \eta - \varepsilon + 4\alpha,\end{aligned}$$

where the last line follows because w is a unit vector, and we chose ε so that $1-\eta-\varepsilon > \eta$. Therefore

$$\|p^{-1}M_u - \varepsilon a_1 a_1^\top\| \leqslant \|p^{-1}M_u\| - \varepsilon + 8\alpha.$$

Applying Fact 2, we conclude that if v is the top eigenvector of M_u, then $\langle v, a_1\rangle^2 \geqslant \frac{\varepsilon - 8\alpha}{\varepsilon} \geqslant 1 - \frac{8\alpha}{\varepsilon}$. Since we can choose $\varepsilon = \frac{1}{2}$ and still have that $\varepsilon < 0.98 < 1 - 2\eta$, the conclusion follows. ■

5.2. From independent columns to orthonormal columns

We now use standard techniques to prove that one can reduce from a dictionary with independent columns to a dictionary with orthogonal columns, given sufficiently many samples.

Proof [Proof of Lemma 14] Note that the expected covariance matrix of the samples is equal to a scaled version of the covariance matrix,

$$\mathbb{E}[(Ax)(Ax)^\top] = \mathbb{E}[x_1]^2 \cdot \Sigma.$$

Since we have assumed that $\|x\|_2^2 \leqslant n$, the Frobenius norm of $(Ax)(Ax)^\top$ is bounded by n for every sample, and $\mathbb{E}[((Ax)(Ax)^\top)^2] \leqslant n^2$. By applying a matrix Bernstein inequality (see e.g. Tropp (2012)), we have that so long as we have $m \geqslant \widetilde{O}((n/\beta)^2)$ samples, the empirical covariance matrix $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m y^{(i)}(y^{(i)})^\top$ will approximate Σ within β in spectral norm.

So given sufficiently many samples, we can compute a good spectral approximation of $\Sigma^{-1/2}$, $\hat{\Sigma}^{-1/2}$ with

$$\left\|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\right\| \leqslant \varepsilon.$$

Assuming access to such a $\hat{\Sigma}^{-1/2}$, we can transform A to $\tilde{A} = \hat{\Sigma}^{-1/2}A$. Now, for matrices X, Y, Z of suitable dimensions,

$$YXY - ZXZ = \frac{1}{2}((Y-Z)X(Y+Z) + (Y+Z)X(Y-Z))\,.$$

Applying this to $\tilde{A}\tilde{A}^\top - \mathrm{Id} = \hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\Sigma\Sigma^{-1/2}$,

$$\begin{aligned}\left\|\tilde{A}\tilde{A}^\top - \mathrm{Id}\right\| &\leqslant \|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\| \cdot \|\Sigma\| \cdot (\|\hat{\Sigma}^{-1/2}\| + \|\Sigma^{-1/2}\|) \\ &\leqslant \varepsilon \cdot \|\Sigma\| \cdot (2+\varepsilon)\|\Sigma^{-1/2}\| \\ &\leqslant O\left(\varepsilon \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)^{1/2}}\right)\,.\end{aligned}$$

Defining $\eta \stackrel{\text{def}}{=} \|\tilde{A}\tilde{A}^\top - \mathrm{Id}\,\|$, we have that $\tilde{A}\tilde{A}^\top = (1 \pm \eta)\,\mathrm{Id}$, and the columns of $\tilde{A}$ are near-orthonormal. Similarly, we can transform our samples

$$y^{(i)} = Ax^{(i)} \to \tilde{y}^{(i)} = \hat{\Sigma}^{-1/2}Ax^{(i)}.$$

Now, define $S_{diff} = \left((\hat{\Sigma}^{-1/2})^{\otimes 2} - (\Sigma^{-1/2})^{\otimes 2}\right)$ and $S_{sum} = \left((\hat{\Sigma}^{-1/2})^{\otimes 2} + (\Sigma^{-1/2})^{\otimes 2}\right)$. We can factor the difference

$$\begin{aligned}\left\|\mathbb{E}[(\tilde{A}x)^{\otimes 4} - (\Sigma^{-1/2}Ax)^{\otimes 4}]\right\| &= \left\|\frac{1}{2}S_{diff}\,\mathbb{E}[(Ax)^{\otimes 4}]S_{sum} + \frac{1}{2}S_{sum}\,\mathbb{E}[(Ax)^{\otimes 4}]S_{diff}\right\| \\ &\leqslant \|S_{sum}\| \cdot \|S_{diff}\| \cdot \left\|\mathbb{E}[(Ax)^{\otimes 4}]\right\| \\ &\leqslant \left(\|\hat{\Sigma}^{-1/2}\|^2 + \|\Sigma^{-1/2}\|^2\right) \cdot \|S_{diff}\| \cdot \left\|\mathbb{E}[(Ax)^{\otimes 4}]\right\| \\ &\leqslant (2+\varepsilon)\|\Sigma^{-1}\| \cdot \|S_{diff}\| \cdot \left\|\mathbb{E}[(Ax)^{\otimes 4}]\right\|\end{aligned}$$

Applying the identity

$$A^{\otimes 2} - B^{\otimes 2} = \frac{1}{2}(A-B)\otimes(A+B) + \frac{1}{2}(A-B)\otimes(A+B),$$

to S_{diff}, we can get that

$$\begin{aligned}\left\|\mathbb{E}[(\tilde{A}x)^{\otimes 4} - (\Sigma^{-1/2}Ax)^{\otimes 4}]\right\| &\leqslant (2+\varepsilon)\|\Sigma^{-1}\| \cdot \|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\| \left(\|\hat{\Sigma}^{-1/2}\| + \|\Sigma^{-1/2}\|\right) \left\|\mathbb{E}[(Ax)^{\otimes 4}]\right\|\,, \\ &\leqslant 9 \cdot \|\Sigma^{-3/2}\| \cdot \varepsilon \left\|\mathbb{E}[(Ax)^{\otimes 4}]\right\|\,.\end{aligned}$$

Since we can choose the number of samples so as to make this last quantity as small as we would like, as a function of the condition number, and then appealing to Fact 1, the reduction is complete. ■

5.3. Sample complexity bounds

Below is our bound on the sample complexity of approximating the 4th moment tensor, which we believe may be loose.

Proposition 17 *Given samples of the form $y^{(i)} = Ax^{(i)}$ for $x^{(i)} \sim \mathcal{D}$, as long as $\beta \geqslant \mathbb{E}[x_i^8]$ dominates the expectation of any other order-8 monomial in x, and any monomial with odd multiplicity has expectation 0, and the entries of x are bounded by κ, then with high probability given $m \geqslant \widetilde{O}(\max\{\beta n^3, (\kappa n)^2\})$ samples,*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (y^{(i)})^{\otimes 4} - \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [(Ax)^{\otimes 4}] \right\| \leqslant o(1).$$

Proof Our matrix has the form

$$M = \sum_{ijk\ell} x_i x_j x_k x_\ell \cdot (a_i \otimes a_j)(a_k \otimes a_\ell)^\top,$$

And the a_i are orthonormal, so

$$\mathbb{E}[MM^\top] = \sum_{\substack{i,j,i',j' \\ k,\ell}} \mathbb{E}[x_i x_j x_{i'} x_{j'} x_k^2 x_\ell^2] \cdot (a_i \otimes a_j)(a_{i'} \otimes a_{j'})^\top.$$

This is because of the orthonormality of the a_i, which guarantees that terms in the product $MM^\top$ in which we have an inner product between two non-identical vectors drop out to 0.

If we define A to be the $d^2 \times n^2$ matrix whose columns are the Kronecker products $a_i \otimes a_j$ for all $i, j \in [n]$, and if for each pair $k, \ell \in [n]$ we define $X^{(k,\ell)}$ be the $n^2 \times n^2$ matrix whose $(i,j),(i',j')$th entry is $\mathbb{E}[x_i x_j x_i x_{j'} x_k^2 x_\ell^2]$, we can realize $E[MM^\top]$ as

$$\mathbb{E}[MM^\top] = A \left(\sum_{k,\ell} X^{(k,\ell)} \right) A^\top.$$

Because we assumed that $E[x_i x_j x_{i'} x_{j'} x_k^2 x_\ell^2] = 0$ unless every index appears with even multiplicity, the entry of $X^{(k,\ell)}$ will be 0. This happens only on the diagonal, unless $i' = j'$ and $i = j$, or for $X^{(k,\ell)}$ in the intersection of the (k,ℓ)th row and the (ℓ,k)th column and the (ℓ,k)th row and the (k,ℓ)th column. So we split each $X^{(k,\ell)}$ into a diagonal part $D^{(k,\ell)}$, an intersection part corresponding to the (k,ℓ) and (ℓ,k) intersections $C^{(k,\ell)}$, and the rest of the off-diagonal part $R^{(k,\ell)}$. and it follows that

$$\| \mathbb{E}[MM^\top] \| \leqslant n^2 \|A\|^2 \max_{k,\ell} \left(\left\| D^{(k,\ell)} \right\| + \left\| C^{k,\ell} \right\| + \left\| R^{(k,\ell)} \right\| \right).$$

Because every entry is bounded by $\mathbb{E}[x_i^8] \leqslant \beta$, and the $D^{(k,\ell)}$ are diagonal, the $\|D\|$ term contributes β. The C matrices have Frobenius norm 2β, and the R matrices have only n^2 nonzero entries, so $\|R\|_F \leqslant \beta n$. Therefore,

$$\| \mathbb{E}[MM^\top] \| \leqslant 3\beta n^3.$$

For each sample $x^{(i)}$, we have that $\| \frac{1}{m} (Ax^{(i)})^{\otimes 4} \|_F \leqslant \frac{\kappa^2 n^2}{m}$, and we have by the above reasoning that $\| \mathbb{E}[\frac{1}{m^2} (Ax^{(i)})^{\otimes 4} (Ax^{(i)})^{\otimes 4}] \| \leqslant 3\, \mathbb{E}[x_i^8] \frac{n^3}{m^2}$. So the variance of the empirical 4-tensor is $\sqrt{n^3/m}$, and the absolute bound on the norm of any summand is n^2/m. Applying a matrix Bernstein inequality (see e.g. Tropp (2012)), we have that as long as we have $m \gg \max\{\kappa^2 n^2 \log n, \beta n^3 \log n\}$ samples, with high probability we approximate $\mathbb{E}[(Ax)^{\otimes 4}]$ within spectral norm $o(1)$. ■

Acknowledgments

T.S. is supported by an NSF Graduate Research Fellowship (NSF award no 1106400). D.S. is supported by a Microsoft Research Fellowship, a Alfred P. Sloan Fellowship, an NSF award, and the Simons Collaboration for Algorithms and Geometry.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 123–137. JMLR.org, 2014.

Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *NIPS*, pages 926–934, 2012.

Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham Kakade. A tensor spectral approach to learning mixed membership community models. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 867–881. JMLR.org, 2013.

Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15 (1):2773–2832, 2014.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *CoRR*, abs/1401.0579, 2014.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 113–149. JMLR.org, 2015.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. Provable learning of noisy-or networks. *CoRR*, abs/1612.08795, 2016.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *STOC*, pages 143–151. ACM, 2015.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *STOC*, pages 594–603. ACM, 2014.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

Andreas Argyriou Theodoros Evgeniou and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, pages 41–48. MIT Press, 2007.

Tom Goldstein and Stanley Osher. The split bregman method for l1-regularized problems. *SIAM journal on imaging sciences*, 2(2):323–343, 2009.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *NIPS*, pages 2861–2869, 2014.

Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis. 1970.

Elad Hazan and Tengyu Ma. A non-generative framework and convex relaxations for unsupervised learning. In *NIPS*, pages 3306–3314, 2016.

Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Speeding up sum-of-squares for tensor decomposition and planted sparse vectors. *CoRR*, abs/1512.02337, 2015.

Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *STOC*, pages 178–191. ACM, 2016.

Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*, pages 11–19. ACM, New York, 2013.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. Blind source separation by simultaneous third-order tensor diagonalization. In *EUSIPCO*, pages 1–4. IEEE, 1996.

Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Trans. Signal Processing*, 55(6-2): 2965–2973, 2007.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *FOCS*, pages 438–446. IEEE Computer Society, 2016.

Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert, and Jean Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *Computer Vision–ECCV 2008*, pages 43–56. Springer, 2008.

Y Marc'Aurelio Ranzato, Lan Boureau, and Yann LeCun. Sparse feature learning for deep belief networks. *Advances in neural information processing systems*, 20:1185–1192, 2007.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *STOC*, pages 366–375. ACM, 2005.

Roberto I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010. ISSN 1083-589X/e.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. doi: 10.1137/0314056. URL http://dx.doi.org/10.1137/0314056.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

Appendix A. Useful Tools

Lemma 18 (Concentration of random tensor contractions Ma et al. (2016)) *Let g be a standard Gaussian vector in $\mathbb{R}^k$, $g \sim \mathcal{N}(0, \mathrm{Id}_k)$. Let A be a tensor in $(\mathbb{R}^k) \otimes (\mathbb{R}^\ell) \otimes (\mathbb{R}^m)$, and call the three modes of A α, β, γ respectively. Let A_i be a $\ell \times m$ slice of A along mode α. Then,*

$$\mathbb{P}\left[\left\|\sum_{i=1}^k g_i A_i\right\| \geqslant t \cdot \max\left\{\left\|A_{\{\alpha\beta\}\{\gamma\}}\right\|, \left\|A_{\{\alpha\gamma\}\{\beta\}}\right\|\right\}\right] \leqslant (m+\ell)\exp\left(-\frac{t^2}{2}\right).$$

Proof We compute the expectation and variance of our matrix,

$$\mathop{\mathbb{E}}_g\left[\sum_{i=1}^k g_i A_i\right] = 0, \qquad \text{and} \qquad \left\|\mathop{\mathbb{V}}_g\left[\sum_{i=1}^k g_i A_i\right]\right\| = \max\left\{\left\|\sum_{i=1}^k A_i A_i^\top\right\|, \left\|\sum_{i=1}^k A_i^\top A_i\right\|\right\},$$

The two variance terms correspond to $\|A_{\{\alpha\beta\}\{\gamma\}}\|^2$ and $\|A_{\{\alpha\gamma\}\{\beta\}}\|^2$ respectively. We can now apply concentration results for matrix Gaussian series to conclude the proof Oliveira (2010). ■

The following lemma states that distances can only decrease under projections to a convex set, and is well-known (see e.g. Rockafellar (1976)).

Lemma 19 *Let $C \subset \mathbb{R}^n$ be a closed convex set, and let $\Pi : \mathbb{R}^n \to C$ be the projection operator onto C in terms of norm $\|\cdot\|_2$, i.e. $\Pi(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{c \in C} \|x - c\|_2$. Then for any $x, y \in \mathbb{R}^n$,*

$$\|x - y\|_2 \geqslant \|\Pi(x) - \Pi(y)\|_2.$$

Proof If we let $D_x = x - \Pi(x)$, $D_y = y - \Pi(y)$,

$$\begin{aligned}\|x - y\|^2 &= \|D_x - D_y + \Pi(x) - \Pi(y)\|^2 \\ &= \|D_x - D_y\|^2 + \|\Pi(x) - \Pi(y)\|^2 + 2\langle D_x - D_y, \Pi(x) - \Pi(y)\rangle\end{aligned}$$

Now the conclusion will follow from the fact that

$$\langle D_x - D_y, \Pi(x) - \Pi(y)\rangle \geqslant 0.$$

This is because, by definition of Π,

$$\Pi(x) = \operatorname{argmin}_{c \in C} \|x - c\|_2^2 = \operatorname{argmin}_{p \in \mathbb{R}^n} \frac{1}{2}\|x - p\|_2^2 + \mathbb{I}_C(p),$$

where $\mathbb{I}_C(\cdot)$ is the convex function defined to be ∞ on elements not in C and 0 otherwise. From the strong convexity of the last expression the projection is unique. From the optimality conditions, it follows that $\Pi(x)$ is the unique point $p \in \mathbb{R}^n$ such that $x - p \in \partial\, \mathbb{I}_C(p)$, where $\partial\, \mathbb{I}_C(p)$ is the set of subgradients of $\mathbb{I}_C$ at p.

By definition of the subgradient and by the convexity of $\mathbb{I}_C$, for any $p, q \in \mathbb{R}^n$ and for $g_p \in \partial\, \mathbb{I}_C(p)$,$g_q \in \partial\, \mathbb{I}_C(q)$,

$$\begin{aligned} \mathbb{I}_C(p) + \langle g_p, q - p\rangle &\leqslant \mathbb{I}_C(x) \\ \langle g_p, q - p\rangle &\leqslant \mathbb{I}_C(q) - \mathbb{I}_C(p) \\ -\langle g_q, q - p\rangle &\leqslant -\mathbb{I}_C(q) + \mathbb{I}_C(p) \\ \langle g_p - g_q, p - q\rangle &\geqslant 0\,. \end{aligned}$$

Now, taking $p = \Pi(x)$ and $q = \Pi(y)$, we have that $D_x = x - \Pi(x) \in \partial\, \mathbb{I}_C(\Pi(x))$, and $D_y = y - \Pi(y) \in \partial\, \mathbb{I}_C(\Pi(y))$, so from the above,

$$\langle D_x - D_y, \Pi(x) - \Pi(y)\rangle \geqslant 0,$$

as desired. ■

Fact 2 *Let $v \in \mathbb{R}^n$, and suppose that $\|M - vv^\top\| \leqslant \|M\| - \varepsilon\|v\|^2$. Then if u, w are the top unit left- and right-singular vectors of M,*

$$\langle u, v\rangle^2 \geqslant \varepsilon \cdot \|v\|^2 \quad \textit{or} \quad \langle w, v\rangle^2 \geqslant \varepsilon \cdot \|v\|^2$$

Proof We have that

$$\|M\| - \varepsilon\|v\|^2 \geqslant |u^\top(M - vv^\top)w| \geqslant |u^\top M w| - |u^\top vv^\top w| = \|M\| - |\langle u, v\rangle\langle w, v\rangle|,$$

where the second inequality is the triangle inequality. Rearranging, the conclusion follows. ■

Fact 3 *If A is an $n \times n$ symmetric matrix with eigendecomposition $\sum_{i\in[n]} \lambda_i u_i u_i^\top$ for orthonormal $u_1, \ldots, u_n \in \mathbb{R}^n$ and eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_n$, then the projection of A to the PSD cone is equal to $\sum_{i\in[n]} \mathbb{I}[\lambda_i \geqslant 0] \cdot \lambda_i u_i u_i^\top$.*

Proof Let $\hat{A} = A + B$ be the projection (in Frobenius norm) of A to the PSD cone. Because $u_1, \ldots, u_n$ are orthonormal, we may choose an orthonormal basis $V = \{v_i\}_{i=1}^{n^2}$ for $\mathbb{R}^{n^2}$ that includes $v_1 = u_1 \otimes u_1, v_2 = u_2 \otimes u_2, \ldots, v_n = u_n \otimes u_n$, as the first n basis vectors. Now, viewing A, B as vectors in $\mathbb{R}^{n^2}$, we can write $A = \sum_{i=1}^n \lambda_i v_i$ for λ_i the eigenvalues of A, and write $B = \sum_{i=1}^{n^2} \beta_i v_i$ for some scalars $\beta_1, \ldots, \beta_{n^2}$.

For any eigenvector u_i of A, we have that $u_i^\top \hat{A} u_i = \langle v_i, A + B\rangle = \lambda_i + \beta_i$ by the orthonormality of the v_i. Therefore, if $\lambda_i < 0$ we must have $\beta_i \geqslant |\lambda_i|$, since $\hat{A}$ is PSD. We also have that $\|A - \hat{A}\|_F^2 = \|B\|_F^2 = \sum_{i=1}^{n^2} \beta_i^2$, so $B = \sum_{i=1}^n \mathbb{I}[\lambda_i < 0] \cdot |\lambda_i| \cdot u_i u_i^\top$ minimizes the Frobenius norm of the difference, which (after checking to see that $A + B$ has all non-negative eigenvalues) concludes the proof.

■