Tensor Factorization via Matrix Factorization

Volodymyr Kuleshov∗ Arun Tejasvi Chaganty∗ Percy Liang
Department of Computer Science Stanford University Stanford, CA 94305

Abstract

Tensor factorization arises in many machine learning applications, such knowledge base modeling and parameter estimation in latent variable models. However, numerical methods for tensor factorization have not reached the level of maturity of matrix factorization methods. In this paper, we propose a new method for CP tensor factorization that uses random projections to reduce the problem to simultaneous matrix diagonalization. Our method is conceptually simple and also applies to non-orthogonal and asymmetric tensors of arbitrary order. We prove that a small number random projections essentially preserves the spectral information in the tensor, allowing us to remove the dependence on the eigengap that plagued earlier tensor-to-matrix reductions. Experimentally, our method outperforms existing tensor factorization methods on both simulated data and two real datasets.

1 Introduction

Given a tensor \( \hat{T} \in \mathbb{R}^{d \times d \times d} \) of the following form:

\[
\hat{T} = \sum_{i=1}^{k} \pi_i a_i \otimes b_i \otimes c_i + \text{noise},
\]

our goal is to estimate the factors \( a_i, b_i, c_i \in \mathbb{R}^d \) and factor weights \( \pi \in \mathbb{R}^k \). In machine learning and statistics, this tensor \( \hat{T} \) typically represents higher-order relationships among variables, and we would like to uncover the salient factors that explain these relationships. This problem of tensor factorization is an important problem rich with applications [1]: modeling knowledge bases [2], topic modeling [3], community detection [4], learning graphical models [5, 6]. The last three fall into a class of procedures based on the method of moments for latent-variable models, which are notable because they provide guarantees of consistent parameter estimation [7].

However, tensors, unlike matrices, are fraught with difficulties: identifiability is a delicate issue [8, 9, 10], and computing Equation 1 is in general NP-hard [11, 12]. In this work, we propose a simple procedure to reduce the problem of factorizing tensors to that of factorizing matrices. Specifically, we first project the tensor \( \hat{T} \) onto a set of random vectors, producing a set of matrices. Then we simultaneously diagonalize the matrices, producing an estimate of the factors of the original tensor. We can optionally refine our estimate by running the procedure using the estimated factors rather than random vectors. Our approach applies to orthogonal, non-orthogonal and asymmetric tensors of arbitrary order.

From a practical perspective, this approach enables us to immediately leverage mature algorithms for matrix factorization. Such algorithms often have readily available implementations that are numerically stable and highly optimized. In our experiments, we observed that they contribute to improvements in accuracy and speed over methods that deal directly with a tensor.

From a theoretical perspective, we consider both statistical and optimization aspects of our method. Most of our results pertain to the former: we provide guarantees on the accuracy of a solution as a function of the noise \( \epsilon \) (this noise typically comes from the statistical estimation of \( T \) from finite data) that are comparable to those of existing methods (Table 1). Algorithms based on matrix diagonalization have been previously criticized [7] to be extremely sensitive to noise due to a dependence on the smallest difference between eigenvalues (the eigengap). We show that this dependence can be entirely avoided using just \( O(\log k) \) tensor projections chosen uniformly at random. Furthermore,
our guarantees are independent of the algorithm used for diagonalizing the projection matrices.

The optimization aspects of our method, on the other hand, depend on the choice of joint diagonalization subroutine. Most subroutines enjoy local quadratic convergence rates [13, 14, 15] and so does our method. With sufficiently low noise, global convergence guarantees can be established for some joint diagonalization algorithms [16]. More importantly, local optima are not an issue for our method in practice, which is in sharp contrast to some other approaches, such as expectation maximization (EM).

Finally, we show that our method obtains accuracy improvements over alternating least squares and the tensor power method on several synthetic and real datasets. On a community detection task, we obtain up to a 15% reduction in error compared to a recently proposed approach [4], and up to an 8% reduction in error on a crowdsourcing task [17], matching or outperforming a state-of-the-art EM-based estimator on three of the four datasets.

Notation Let \([n] = \{1, \ldots, n\}\) denote the first \(n\) positive integers. Let \(e_i\) be the indicator vector which is 1 in component \(i\) and 0 in all other components. We use \(\otimes\) to denote the tensor product: if \(u, v, w \in \mathbb{R}^d\), then
\[
 u \otimes v \otimes w \in \mathbb{R}^{d \times d \times d}.
\]
For a third order tensor \(T \in \mathbb{R}^{d \times d \times d}\) we define vector and matrix application as,
\[
 T(x, y, z) = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} T_{ijk} x_i y_j z_k
\]
\[
 T(X, Y, Z)_{ijk} = \sum_{m=1}^{d} \sum_{n=1}^{d} T_{lmn} X_l Y_m Z_{nk},
\]
for vectors \(x, y, z \in \mathbb{R}^d\) and matrices \(X, Y, Z \in \mathbb{R}^{d \times k}\). The partial vector application (or projection) \(T(I, I, w)\) of a vector \(w \in \mathbb{R}^d\) returns a \(d \times d\) matrix:
\[
 T(I, I, w)_{ij} = \sum_{k=1}^{d} T_{ijk} w_k.
\]
We define the CP decomposition of a tensor \(T \in \mathbb{R}^{d \times d \times d}\) as
\[
 T = \sum_{i=1}^{k} \pi_i a_i \otimes b_i \otimes c_i,
\]
for \(\pi_i, a_i, b_i, c_i \in \mathbb{R}^d\). The rank of \(T\) is said to be \(k\). When \(a_i = b_i = c_i = u_i\) for all \(i\), and the \(u_i\)’s are orthogonal, we say \(T\) has a symmetric orthogonal factorization. 
\[
 T = \sum_{i=1}^{k} \pi_i u_i \otimes u_i \otimes u_i.
\]
Projecting a tensor \(T = \sum_{i=1}^{k} \pi_i a_i \otimes b_i \otimes c_i\) along \(w\) produces a matrix \(T(I, I, w) = \sum_{i=1}^{k} \pi_i (c_i^w) w a_i \otimes b_i\). We use \(\lambda_i = \pi_i(c_i^w) w\) to refer to the factor weights (or eigenvalues in the orthogonal setting) of the projected matrix.

For a vector of values \(\pi \in \mathbb{R}^k\), we use \(\pi_{\min}\) and \(\pi_{\max}\) to denote the minimum and maximum absolute values of the entries, respectively. Finally, we use \(\delta_{ij}\) to denote the indicator function, which equals 1 when \(i = j\) and 0 otherwise.

2 Background

In this section, we establish the context for tensor factorization, method of moments for estimating latent-variable models, and simultaneous matrix diagonalization.

2.1 Tensor factorization algorithms

Existing tensor factorization methods vary in their sensitivity to noise \(\epsilon\) in the tensor, their tolerance of non-orthogonality (as measured by the incoherence \(\mu\)) and in their convergence properties (Table 1). The robust tensor power method (TPM, [7]) is a popular algorithm with theoretical guarantees on global convergence. A recently-developed coordinate-descent method for orthogonal tensor factorization based on Givens rotations [18] is empirically more robust than the TPM; however it is limited to the full-rank setting and lacks a sensitivity analysis. A further limitation of both methods is that they only work for symmetric orthogonal tensors. Asymmetric non-orthogonal tensors could be handled by preprocessing and whitening, but this can be a major source of errors in itself [21]. Alternating least squares (ALS) and other gradient-
based methods [22] are simple, popular, and apply to
the non-orthogonal setting, but are known to easily
get stuck in local optima [23]. Anandkumar et al. [19]
explicitly show both local and global convergence
 guarantees for a slight modification of the ALS procedure
under certain assumptions on the tensor \( \mathbf{T} \).

Finally, some authors have also proposed using simul-
taneous diagonalization for tensor factorization: Lath-
dauwer [23] proposed a reduction, but it requires for-
ing a linear system of size \( O(d^4) \) and is quite com-
plex. Anandkumar et al. [20] performed multiple ran-
don projections, but only diagonalized two at a time
(SD2), leading to unstable results; the method also
only applies to orthogonal factors. Anandkumar et al.
[7] briefly remarked that using all the projections at
once was possible but did not pursue it. In contrast,
our method, has comparable bounds to the tensor
power method in the orthogonal setting (conven-
tionally \( \| \pi \|_1 = 1 \) is assumed), and the ALS method in
the non-orthogonal setting.

2.2 Parameter estimation in mixture models

Tensor factorization can be used for parameter esti-

mation for a wide range of latent-variable models such as
Gaussian mixture models, topic models, hidden
Markov models, etc. [7]. For illustrative purposes, we
focus on the single topic model [7], defined as follows:
For each of \( n \) documents, draw a latent “topic” \( h \in [k] \)
with probability \( \mathbb{P}[h = i] = \pi_i \) and three observed
words \( x_1, x_2, x_3 \in \{ e_1, \ldots, e_d \} \), which are condi-
tionally independent given \( h \) with \( \mathbb{P}[x_j = w \mid h = i] = u_{iw} \)
for each \( j \in \{ 1, 2, 3 \} \). The parameter estimation task is
to output an estimate of the parameters \((\pi, \{ u_i \}_{i=1}^k)\)
given \( n \) documents \( \{ (x_1^{(i)}, x_2^{(i)}, x_3^{(i)}) \}_{i=1}^n \)
(importantly, the topics are unobserved).

Traditional approaches typically use Expectation
Maximization (EM) to optimize the marginal log-
likelihood, but this algorithm often gets stuck in
local optima. The method of moments approach is
to cast estimation as tensor factorization: define the
empirical tensor \( \hat{T} = \frac{1}{n} \sum_{i=1}^n x_1^{(i)} \otimes x_2^{(i)} \otimes x_3^{(i)} \). It can be
shown that \( \hat{T} = \sum_{i=1}^k \pi_i u_i \otimes u_i \otimes u_i + \epsilon R \) (a refinement
of Equation 1), where \( \epsilon R \in \mathbb{R}^{d \times d \times d} \) is the statistical
noise which goes to zero as \( n \to \infty \). A tensor factoriza-
tion scheme that asymptotically recovers estimates of
\((\pi, \{ u_i \}_{i=1}^k)\) therefore provides a consistent estimator
of the parameters.

2.3 Simultaneous diagonalization

We now briefly review simultaneous matrix diagonal-
ization, the main technical driver in our approach. In
simultaneous diagonalization, we are given a set of
symmetric matrices \( M_1, \ldots, M_L \in \mathbb{R}^{d \times d} \) (see Section 6
for a reduction from the asymmetric case), where each
matrix can be expressed as
\[
M_i = U_i \Lambda_i U_i^\top + \epsilon_i R_i. \tag{2}
\]
The diagonal matrix \( \Lambda_i \in \mathbb{R}^{d \times d} \) and the noise \( \epsilon_i R_i \)
are individual to each matrix, but the non-singular
transform \( U \in \mathbb{R}^{d \times k} \) is common to all the matrices.
We also define the full-rank extensions,
\[
\bar{U} = [U \ U^\perp] \quad \bar{\Lambda}_i = \begin{bmatrix} \Lambda_i & 0 \\ 0 & 0 \end{bmatrix}, \tag{3}
\]
where the columns of \( U^\perp \in \mathbb{R}^{d-k \times d} \) span the orthog-
onal subspace of \( U \) and \( \bar{\Lambda}_i \in \mathbb{R}^{d \times d} \) has been appropri-
ately padded with zeros. Note that \( \bar{U} \bar{\Lambda}_i \bar{U}^\top = U_i \Lambda_i U_i^\top \).

The goal is to find an invertible transform \( V^{-1} \in \mathbb{R}^{d \times d} \)
such that each \( V^{-1} M_i V^{-\top} \) is nearly diagonal. We
refer to the \( V^{-1} \) as inverse factors. When \( \epsilon = 0 \), this
problem admits a unique solution when there are at
least two matrices [24]. There are a number of objec-
tive functions for finding \( V \) [25, 13, 26], but in this
paper, we focus on a popular one that penalizes off-
diagonal terms:
\[
F(X) \triangleq \sum_{l=1}^L \text{off}(X^{-1} M_l X^{-\top}), \quad \text{off}(A) = \sum_{i \neq j} A_{ij}^2. \tag{4}
\]
An important setting of this problem, which we refer
to as the orthogonal case, is when we know the true
factors \( U \) to be orthogonal. In this case we constrain
our optimization variable \( X \) to be orthogonal as well,
i.e. \( X^{-1} = X^\top \).

In principle, we could just diagonalize one of the ma-
trices, say \( M_1 \) (assuming its eigenvalues are distinct)
to recover \( U \). However, when \( \epsilon > 0 \), this procedure
is unreliable and simultaneous diagonalization greatly
improves on robustness to noise, as we will witness in
Section 4.

There exist several algorithms for optimizing \( F(X) \).
In this paper, we will use the Jacobi method [27, 25] for
the orthogonal case and the QRJ1D algorithm [26] for
the non-orthogonal case. Both techniques are based on
same idea of iteratively constructing \( X^{-1} \) via a pro-
duct of simple matrices \( X^{-1} = B_T \cdots B_2 B_1 \), where at
each iteration \( t = 1, \ldots, T \), we choose \( B_t \) to minimize
\( F(X) \). Typically, this can be done in closed form.

The Jacobi algorithm for the orthogonal case is a sim-
ple adaptation of the Jacobi method for diagonaliz-
ing a single matrix. Each \( B_t \) is chosen to be a Givens ro-
tation [27] defined by two of the \( d \) axes \( i < j \in [d] \):
\[
B_t = (\cos \theta (\Delta_{ii} + \Delta_{jj}) + (\sin \theta)(\Delta_{ij} - \Delta_{ji}) \text{ for some}
\]
angle $\theta$, where $\Delta_{ij}$ is a matrix that is 1 in the $(i, j)$-th entry and 0 elsewhere. We sweep over all $i < j$, compute the best angle $\theta$ in closed form using the formula proposed by Cardoso and Souloumiac [25] to obtain $B_t$, and then update each $M_i$ by $B_t M_i B_t^\top$. The above can be done in $O(d^3L)$ time per sweep.

For the non-orthogonal case, the QR11D algorithm is similar, except that $B_t$ is chosen to be either a lower or upper unit triangular matrix ($B_t = I + a \Delta_{ij}$ for some $a$ and $i \neq j$). The optimal value of $a$ that minimizes $F(X)$ can also be computed in closed form (see [26] for details). The running time per iteration is the same as before.

3 Tensor factorization via simultaneous matrix diagonalization

We now outline our algorithm for symmetric third-order tensors. In Section 6, we describe how to generalize our method to arbitrary tensors. Observe that the projection of $T = \sum_i \pi_i u_i^{\otimes 3}$ along a vector $w$ is a matrix $T(I, I, w) = \sum_i \pi_i (w^\top u_i) u_i^{\otimes 2}$ that preserves all the information about the factors $u_i$ (assuming the $\pi_i (w^\top u_i)$’s are distinct). In principle one can recover the $u_i$ through an eigendecomposition of $T(I, I, w)$. However, this method is sensitive to noise: the error $\|u_i - \tilde{u}_i\|_2$ of an estimated eigenvector $\tilde{u}_i$ depends on the reciprocal of the smallest eigengap $\max_{j \neq i} 1/|\lambda_i - \lambda_j|$ of the projected matrix (recall that $\lambda_i = \pi_i (w^\top u_i)$), which can be large and lead to inaccurate estimates.

Instead, let us obtain the factorization of $T$ from projections along multiple vectors $w_1, w_2, \ldots, w_L$. The projections produce matrices of the form $M_i = \sum_{l=1}^L \lambda_{il} u_i^{\otimes 2}$, with $\lambda_{il} = \pi_i (w_i^\top u_i)$; they have common eigenvectors, and therefore can be simultaneously diagonalized. The advantage is, as we will show later, that simultaneous diagonalization is sensitive to the measure $\min_{i \neq j} \sum_{l=1}^L (\lambda_{il} - \lambda_{jl})^2 / \left(\sum_{l=1}^L (\lambda_{il} - \lambda_{jl})^2\right)$, which averages the minimum eigengap across the matrices $M_i$ (here, $\lambda_{il} = \pi_i (w_i^\top u_i)$).

A natural question to ask is along which vectors $(w_i)$ should we project? In Section 4 and Section 5 we show that (a) estimates of the inverse factors $(v_i)$ is a good choice (when the $(v_i)$ are approximately orthogonal, they are close to the factors $(u_i)$) and that (b) random vectors do almost as well. This suggests a simple two-step method: (i) first, we find approximations of the tensor factors by simultaneously diagonalizing a small number of random projections of the tensor; (ii) then we perform another round of simultaneous diagonalization on projections along inverse of these approximate factors. Algorithm 1 describes the approach. Its running time is $O(k^2 d^2 s)$, where $s$ is the number of sweeps for the simultaneous diagonalization algorithm.

4 Perturbation analysis for orthogonal tensor factorization

In this section, we will focus on the orthogonal setting, returning to non-orthogonal factors in Section 5. For ease of exposition, we restrict ourselves to symmetric third-order orthogonal tensors: $T = \sum_{i=1}^k \pi_i u_i^{\otimes 3}$. Here the inverse factors $(v_i)$ are equivalent to the factors $(u_i)$, and we do not distinguish between the two. The proofs for this section can be found in Appendix B.

Our sensitivity analysis builds on the perturbation analysis result for the simultaneous diagonalization of matrices in Cardoso [28].

**Lemma 1** (Cardoso [28]). Let $M_i = U \Lambda_i U^\top + \epsilon R_i$, $i \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $\Lambda_i \in \mathbb{R}^{k \times k}$. Let $\hat{U} \in \mathbb{R}^{d \times d}$ be a full-rank extension of $U$ with columns $u_1, u_2, \ldots, u_d$ and let $\hat{U} \in \mathbb{R}^{d \times d}$ be the orthogonal minimizer of the joint diagonalization objective $F(\cdot)$. Then, for all $u_j, j \in [k]$, there exists a column $\hat{u}_j$ of $\hat{U}$ such that

$$\|\hat{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1}^d E_{ij}^2} + o(\epsilon),$$

(5)

where $E \in \mathbb{R}^{d \times k}$ is

$$E_{ij} \triangleq \frac{\sum_{l=1}^L (\lambda_{il} - \lambda_{jl}) u_j^\top R_i u_l}{\sum_{l=1}^L (\lambda_{il} - \lambda_{jl})^2}.$$  

(6)

when $i \neq j$ and $i \leq k$ or $j \leq k$. We define $E_{ij} = 0$ when $i = j$ and $\lambda_{ij} = 0$ when $i > k$.

In the tensor factorization setting, we jointly diagonalize projections $M_l$, $l = 1, 2, \ldots, L$ of the noisy
tensor $\hat{T}$ along vectors $w_i$: $\hat{M}_1 = \hat{T}(I,I,w_i) = \sum_{l=1}^L \pi_i(w_i^\top u_i)u_i^\delta + \epsilon R(I,I,w_i)$, where $R_l \triangleq R(I,I,w_i)$ has unit operator norm. Cardoso’s lemma provides bounds on the accuracy of recovering the $u_i$ via joint diagonalization; in particular, we can further rewrite Equation 6 in the tensor setting as:

$$E_{ij} = \frac{\sum_{l=1}^L w_i^\top p_{ij}^\top u_l}{\sum_{l=1}^L w_i^\top p_{ij}^\top w_i},$$

where $p_{ij} \triangleq (\pi_i u_i - \pi_j u_j)$ and $r_{ij} \triangleq R(u_i, u_j, I)$.

Equation 7 tells us that we can control the magnitude of the $E_{ij}$ (and hence the error on recovering the $u_i$) through appropriate choice of the projections $(w_i)$. Ideally, we would like to ensure that the projected eigenvalue, $\min_{i \neq j} w_i^\top p_{ij} = \min_{i \neq j} \pi_i (w_i^\top u_i - \pi_j (w_j^\top u_j))$, is bounded away from zero for at least one $M_i$ so that the denominator of Equation 7 does not blow up.

**Random projections** The first step of Algorithm 1 projects the tensor along random directions. The form of Equation 7 suggests that the error terms, $E_{ij}$ should concentrate over several projections and we will show that this is indeed the case. Consequently, the error terms will depend independently on the mean of $w_i^\top p_{ij}$, $\|p_{ij}\|^2 = \pi_i^2 + \pi_j^2 > \pi_{\min}^2$. Our final result is as follows:

**Theorem 1** (Tensor factorization with random projections). Let $w_1, \ldots, w_L$ be i.i.d. Gaussian vectors, $w_l \sim N(0, I)$, and let the matrices $\hat{M}_i \in \mathbb{R}^{d \times d}$ be constructed via projection of $\hat{T}$ along $w_1, \ldots, w_L$. Let $\hat{u}_i$ be estimates of the $u_i$ derived from the $\hat{M}_i$. Let $L \geq 16 \log(2d(k-1))/\delta^2$. Then, with probability at least $1 - \delta$, for every $u_i$ there exists a $\tilde{u}_i$ such that

$$\|\tilde{u}_i - u_i\|_2 \leq \left(\frac{2\sqrt{2} \|\pi\|_1 \pi_{\max}}{\pi_i} + C(\delta) \sqrt{\frac{\log(kd)}{\delta}}\right) \epsilon + o(\epsilon),$$

where $C(\delta) \triangleq O\left(\log(kd)/\delta \right)$. The first of the above two terms is the fundamental error in estimating a noisy tensor $\hat{T}$; the second term is due to the concentration of random projections and can be made arbitrarily small by increasing $L$.

**Plug-in projections** The next step of our algorithm projects the tensor along the approximate factors from step 2. Intuitively, if the $w_l$ are close to the eigenvectors $u_i$, then $w_i^\top p_{ij} = w_i^\top (\pi_i u_i - \pi_j u_j) \approx \pi_i \delta_i d_i$. Then for each $i \neq j$, there is some projection that ensures that $E_{ij}$ is bounded and does not depend on the projected eigengap $\min_{i \neq j} \pi_i (w_i^\top u_i - \pi_j (w_j^\top u_j))$.

**Theorem 2** (Tensor factorization with plug-in projections). Let $w_1, \ldots, w_k$ be approximations of $u_1, \ldots, u_k$: $\|w_i - u_i\|_2 = O(\epsilon)$, and let $\hat{M} \in \mathbb{R}^{d \times d}$ be constructed via projection of $\hat{T}$ along $w_1, \ldots, w_k$. Let $\hat{u}_i$ be estimates of the $u_i$ derived from the $\hat{M}_i$. Then, for every $u_i$, there exists a $\tilde{u}_i$ such that

$$\|\tilde{u}_i - u_i\|_2 \leq \frac{2\sqrt{2} \|\pi\|_1 \pi_{\max}}{\pi_i} \epsilon + o(\epsilon).$$

Note that Theorem 1 says that with $O(d)$ random projections, we can recover the eigenvectors $u_i$ with almost the same precision as if we used approximate eigenvectors, with high probability. Moreover, as $L \to \infty$, there is no gap between the precision of the two methods. Theorem 2 on the other hand suggests that we can tolerate errors on the order of $O(\epsilon)$ without significantly affecting the error in recovering $\hat{u}_i$. In practice, we find that using the plug-in estimates allows us to improve accuracy with fewer random projections.

## 5 Perturbation analysis for non-orthogonal tensor factorization

We now extend our results to the case when the tensor $T$ has a non-orthogonal symmetric CP decomposition: $T = \sum_{i=1}^k \pi_i u_i^\delta \otimes \pi_j v_j^\delta \otimes \pi_k w_k^\delta$, where the $u_i$ are not orthogonal and $k \leq d$. We parameterize the non-orthogonality using incoherence: $\mu \triangleq \max_{i \neq j} \|u_i - u_j\|_2$ and the norm of the inverse factor $\|V^{\top}\|_2$ where $V \triangleq U^{-1}$. Compared to the orthogonal setting, our bounds reveal an $O\left(\frac{\|V^{\top}\|_2}{\mu}\right)$ dependence on incoherence when $\mu \leq \frac{1}{2d}$. Proofs for this section are found in Appendix C.

We base our analysis on the perturbation result by Afsari [24].

**Lemma 2** (Afsari [24]). Let $M_i = U A_i U^{\top} + \epsilon R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $A_l \in \mathbb{R}^{k \times k}$. Let $\hat{U} \in \mathbb{R}^{d \times d}$ be a full-rank extension of $U$ with columns $u_1, u_2, \ldots, u_d$ and let $V = \hat{U}^{-1}$, with rows $v_1, v_2, \ldots, v_d$. Let $\hat{U} \in \mathbb{R}^{d \times d}$ be the minimizer of the joint diagonalization objective $F(\cdot)$ and let $V = \hat{U}^{-1}$.

Then, for all $u_j$, $j \in [k]$, there exists a column $\hat{u}_j$ of $\hat{U}$ such that

$$\|\hat{u}_j - u_j\|_2 \leq \epsilon \left(\max_{l=1}^d E_{ij} + o(\epsilon)\right),$$

where the entries of $E \in \mathbb{R}^{d \times k}$ are bounded by

$$|E_{ij}| \leq \frac{1}{1 - \rho_{ij}^2} \left(\frac{1}{\|\lambda_i^2\|_2^2} + \frac{1}{\|\lambda_j^2\|_2^2}\right) \left(\sum_{l=1}^d v_l^\top R_l v_j \lambda_{jl} + \sum_{l=1}^d v_l^\top R_l v_j \lambda_{il}\right),$$

where $\rho_{ij} \triangleq \frac{\lambda_{ij}^2}{\lambda_i^2 + \lambda_j^2}$. The entries of $E$ are bounded by $|E_{ij}| \leq \frac{1}{1 - \rho_{ij}^2} \left(\frac{1}{\|\lambda_i^2\|_2^2} + \frac{1}{\|\lambda_j^2\|_2^2}\right) \left(\sum_{l=1}^d v_l^\top R_l v_j \lambda_{jl} + \sum_{l=1}^d v_l^\top R_l v_j \lambda_{il}\right)$. 


when \( i \neq j \) and \( E_{ij} = 0 \) when \( i = j \) and \( \lambda_{ii} = 0 \) when \( i > k \). Here \( \lambda_i = (\lambda_{1i}, \lambda_{2i}, \ldots, \lambda_{L_i}) \in \mathbb{R}^L \) and \( \rho_{ij} = \frac{\lambda_i^\top \lambda_j}{||\lambda_i||_2 ||\lambda_j||_2} \) is the modulus of uniqueness, a measure of how ill-conditioned the problem is.

In the orthogonal case, we had a dependence on the eigengap \( \lambda_i - \lambda_j \). Now the error crucially depends on the modulus of uniqueness, \( \rho_{ij} \). The non-orthogonal simultaneous diagonalization problem has a unique solution iff \( |\rho_{ij}| < 1 \) for all \( i \neq j \) [24]. In the orthogonal case, \( \rho_{ij} = 0 \). It can be shown that \( \rho_{ij} \) can once again be controlled by appropriately choosing the projections \((w_i)\).

To get a handle on the difficulty of the problem, let us assume that the vectors \( u_i \) are incoherent: \( u_i^\top u_j \leq \mu \) for all \( i \neq j \). Intuitively, the problem is easy when \( \mu \approx 0 \) and hard when \( \mu \approx 1 \). In the results that follow, we require \( \mu \leq \frac{1}{2d} \).

Random projections Intuitively, random projections are isotropic and hence we expect the projections \( \lambda_i \) and \( \lambda_j \) to be nearly orthogonal to each other. This allows us to show that \( \rho_{ij} \leq O(\mu) \), which matches our intuitions on the difficulty of the problem. Our final result is the following:

**Theorem 3** (Non-orthogonal tensor factorization with random projections). Let \( w_1, \ldots, w_L \) be i.i.d. random Gaussian vectors, \( w_i \sim N(0, I) \), and let the matrices \( \tilde{M}_l \in \mathbb{R}^{d \times d} \) be constructed via projection of \( \tilde{T} \) along \( w_1, \ldots, w_L \). Assume incoherence \( \mu \leq \frac{1}{2d} \) on \((u_i): u_i^\top u_j \leq \mu \). Let \( L_0 \leq \left( \frac{50}{1 - \mu^2} \right)^2 \) and let \( L \geq L_0 \log(15d(k - 1)/\delta)^2 \). Then, with probability at least \( 1 - \delta \), for every \( u_i \), there exists a \( \tilde{u}_i \) such that

\[
\| \tilde{u}_j - u_j \|_2 \leq O \left( \frac{\sqrt{\|\pi\|_1 l \mu_{\text{max}}}}{\mu_{\text{min}}} \frac{||V||_2^2}{1 - \mu^2 (1 + C(\delta))} \epsilon + o(\epsilon) \right),
\]

where \( C(\delta) \triangleq \left( \log(kd/\delta) \sqrt{\frac{d}{L}} \right) \).

Once again, the error decomposes into a fundamental recovery error and a concentration term. Note that the error is sensitive to the smallest factor weight, \( \mu_{\text{min}} \). This dependence arises from the sensitivity of the non-orthogonal factorization method to the \( \lambda_i \) with the smallest norm and is unavoidable.

**Theorem 4** (Non-orthogonal tensor factorization with plug-in projections). Let \( w_1, \ldots, w_k \) be approximations of \( v_1, \ldots, v_k \): \( \|w_l - v_l\|_2 \leq O(\epsilon) \), and let the matrices \( \tilde{M}_l \in \mathbb{R}^{d \times d} \) be constructed via projection of \( \tilde{T} \) along \( w_1, \ldots, w_k \). Also assume that the \( u_i \) are incoherent: \( u_i^\top u_j \leq \mu \leq \frac{1}{27} \) when \( j \neq i \). Then, for every \( u_j \), there exists a \( \tilde{u}_j \) such that

\[
\| \tilde{u}_j - u_j \|_2 \leq O \left( \frac{\sqrt{\|\pi\|_1 l \mu_{\text{max}}}}{\mu_{\text{min}}} \frac{||V||_2^2}{1 - \mu^2 (1 + C(\delta))} \epsilon + o(\epsilon) \right).
\]

6 Asymmetric and higher-order tensors

In this section, we present simple extensions to the algorithm to asymmetric and higher order tensors.

**Asymmetric tensors** We use a reduction to handle asymmetric tensors. Observe that the \( l \)-th projection \( M_l \) of an asymmetric tensor has the form \( M_l = \sum_i \lambda_i u_i v_i^\top = U \Lambda_l V^\top \), for some diagonal (not necessarily positive) matrix \( \Lambda_l \) and common \( U, V \), not necessarily orthogonal. For each \( M_l \), define another matrix \( N_l = \begin{pmatrix} 0 & M_l^\top \\ M_l & 0 \end{pmatrix} \) and observe that

\[
\begin{pmatrix} 0 & M_l^\top \\ M_l & 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} V^\top & V \\ U & -U \end{pmatrix} \begin{pmatrix} \Lambda_l & 0 \\ 0 & -\Lambda_l \end{pmatrix} \begin{pmatrix} V^\top & V \\ U & -U \end{pmatrix}^\top.
\]

(The \( N_l \) are symmetric matrices with common (in general, non-orthogonal) factors. Therefore, they can be jointly diagonalized and from their components, we can recover the components of the \( (M_l) \). This reduction does not change the modulus of uniqueness of the problem: the factor weights remain unchanged.

**Higher order tensors** Finally, if we have a higher order (say fourth order) tensor \( T = \sum_i \pi_i a_i \otimes b_i \otimes c_i \otimes d_i \) then we can first determine the \( a_i, b_i \) by projecting into matrices \( T(I, I, w, u) = \sum_i \pi_i (w^\top c_i)(u^\top d_i) a_i \otimes b_i \), and then determine the \( c_i, d_i \) by projecting along the first two components. Our bounds only depend on the dimension of the matrices being simultaneously diagonalized, and thus this reduction does not introduce additional error. Intuitively, we should expect that additional modes of a tensor should provide more information and thus help estimation, not hurt it. However, note that as the tensor order increases, the noise in the tensor will presumably increase as well.

7 Convergence properties.

The convergence of our algorithm depends on the choice of joint diagonalization subroutine. Theoretically, the Jacobi method, the QRJ1D algorithm, and
other algorithms are guaranteed to converge to a local minimum at a quadratic rate \cite{27,14,29}. The question of global convergence is currently open \cite{30,25}. Empirically though, these algorithms have been found in the literature to converge reliably to global minima \cite{27,25,30} and to corroborate this claim, we conducted a series of experiments \cite{16}.

We first examined convergence to global minima in the orthogonal setting. In 1000 trials of the Jacobi algorithm on random sets of matrices for various $\epsilon$ and $d = L = 15$, we found that the objective values formed a Gaussian distribution around $\epsilon$ (the best accuracy that can be achieved). Then, on each of our real crowdsourcing datasets, we ran our algorithm from 1000 random starting points; in every case, the algorithm converged to the same solution (unlike EM). This suggests that our diagonalization algorithm is not sensitive to local optima. To complement this empirical evidence, we also established that the Jacobi algorithm will converge to the global minimum when $\epsilon$ is sufficiently small and when the algorithm is initialized with the eigendecomposition of a single projection matrix \cite{16}.

We also performed similar experiments in the non-orthogonal setting using the QRJ1D algorithm. Unlike Jacobi, QRJ1D suffers from local optima, which is expected since the general CP decomposition problem is NP-hard. However, local optima appear to only affect matrices with bad incoherence values, and in several real world experiments (see below), non-orthogonal methods fared better their orthogonal counterparts.

8 Experiments

In the orthogonal setting, we compare our algorithms (OJD0, which uses random projections, and OJD1 which uses with plug-in) with the tensor power method (TPM), alternating least squares (ALS), and with the method of de Lathauwer \cite{23}. In the non-orthogonal setting, we compare de Lathauwer, alternating least squares (ALS), non-linear least squares (NLS), and our non-orthogonal methods (NOJD0 and NOJD1).

Random versus plug-in projections We generated random tensors $T = \sum_{i=1}^{k} \pi u_i^{\otimes 3} + \epsilon R$ with Gaussian entries in $\pi, R$ and $u_i$ distributed uniformly in the sphere $S^{d-1}$. In Figure 1, we plot the error $\sum_{i=1}^{k} \frac{1}{\pi} \|u_i - \tilde{u}_i\|_2$ (averaged over 1000 trials) of using $L$ random projections (blue line), versus using $L$ random projections followed by plug-in (green line). The accuracy of random projections tends to a limit that is immediately achieved by the plug-in projections, as predicted by our theory. In the orthogonal setting, plug-in reduces the total number of projected matrices

$L$ required to achieve the limiting error by three-fold (20 vs. 60 when $d = 10$). In the non-orthogonal setting, the difference between the two regimes is much smaller.

Synthetic accuracy experiments We generated random tensors for various $d, k, \epsilon$ using the same procedure as above. We vary $\epsilon$ and report the average error $\sum_{i=1}^{k} \frac{1}{\pi} \|u_i - \tilde{u}_i\|_2$ across 50 trials.

Our method realizes its full potential in the full-rank non-orthogonal setting, where OJD0 and OJD1 are up to three times more accurate than alternative methods (Figure 2, top). In the (arguably easier) under-complete case, our methods do not achieve more than a 10% improvement, and overall, all algorithms fare similarly (Figure 4 in the supplementary material). Alternating least squares displayed very poor performance, and we omit it from our graphs.

In the full rank setting, there is little difference in performance between our method and Lathauwer (Figure 2, bottom). In both the full and low-rank cases (Figure 2, bottom and Figure 5 in the supplementary material), we consistently outperform the standard approaches, ALS and NLS, by 20–50%. Although we do not always outperform Lathauwer (a state-of-the-art method), NOJD0 and NOJD1 are faster and much simpler to implement.

We also tested our method on estimating the single topic model from Section 2.2. For $d = 50$ and $k = 10$, over 50 trials in which model parameters were generated uniformly at random in $S^{d-1}$, OJD0 and OJD1 obtained error rates of 0.05 and 0.055 respectively, followed by TPM (0.62 error), and Lathauwer (0.65 error).

We refer the readers to the supplementary material for additional experiments on asymmetric tensors and on
algorithm running time.

Community detection in a social network
Next, we use our method to detect communities in a real Facebook friend network at an American university [31] using a recently developed estimator based on the method of moments [4]. We reproduce a previously proposed methodology for assessing the performance of this estimator on our Facebook dataset [31]: ground truth communities are defined by the known dorm, major, and high school of each student; empirical and true community membership vectors $\hat{c}_i, c_i$ are matched using a similarity threshold $t > 0$; for a given threshold, we define the recovery ratio as the number of true $c_i$ to which an empirical $\hat{c}_i$ is matched and we define the accuracy to be the average $\ell_1$ norm distance between $c_i$ and all the $\hat{c}_i$ that match to it. See [31] for more details. By varying $t > 0$, we obtain a tradeoff curve between the recovery ratio and accuracy (Figure 3). Our OJD1 method determines the top 10 communities more accurately than TPM; finding smaller communities was equally challenging for both methods.

Table 2: Crowdsourcing experiment results

| Dataset | Web  | RTE | Birds | Dogs |
|---------|------|-----|-------|------|
| TPM     | 82.25| 88.75| 87.96 | 84.01|
| OJD     | 82.33| 90.00| 89.81 | 84.01|
| NOJD    | 83.49| 90.50| 89.81 | 84.26|
| ALS     | 83.15| 88.75| 88.89 | 84.26|
| LATH    | 83.00| 88.75| 88.89 | 84.26|
| MV+EM   | 83.68| 92.75| 88.89 | 83.89|
| Size    | 2665 | 800 | 106   | 807  |

Label prediction from crowdsourcing data
Lastly, we use our algorithm to infer the true labels of data points within several datasets based on crowdsourcing annotations by real workers. We incorporate our algorithm into a recently proposed estimator based on the method of moments [17] and evaluate the resulting approach on the same datasets that were used to validate this estimator (except one, which we could not obtain). In addition to previously defined methods, we also compare to the expectation maximization algorithm initialized with majority voting by the workers (MV+EM). We measure the label prediction accuracy. Overall, NOJD1 outperforms all other tensor-based methods on three out of four datasets and results in accuracy gains of up to 1.75% (Table 2). Our orthogonal method outperforms the TPM on every dataset but one, and in two cases even outperforms ALS and Lathauwer, even though they are not affected by whitening. Most interestingly, on two datasets, at least one of our methods matches or outperforms the EM estimator, unlike any of the other tensor methods.

9 Discussion
We have presented a simple and efficient method for tensor factorization based on three ideas: simultaneous matrix diagonalization, random projections, and plugin estimates. While simultaneous diagonalization algorithms for tensor factorization have been proposed in the past, they have either been computationally too expensive [23] or numerically unstable [20]. We overcome both these limitations using $O(\log(k))$ random projections of the tensor. Note that our use of random projections is atypical: instead of using projections for dimensionality reduction (e.g. [32]), we use it to reduce the order of the tensor. Finally, we improve estimates of the factors retrieved with random projections by using them as plugin estimates, a common technique in statistics to improve statistical efficiency [33]. Extensive empirical experiments show that our approach results in a factorization algorithm that is both more efficient and more accurate than the state-of-the-art.
References

[1] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.

[2] M. Nickel, V. Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning (ICML)*, pages 809–816, 2011.

[3] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. Liu. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory (COLT)*, pages 867–881, 2013.

[5] Y. Halpern and D. Sontag. Unsupervised learning of noisy-or Bayesian networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

[6] A. Chaganty and P. Liang. Estimating latent-variable graphical models using moments and likelihoods. In *International Conference on Machine Learning (ICML)*, 2014.

[7] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. Technical report, arXiv, 2013.

[8] J. B. Kruskal. Three-ways arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Applications*, 18:95–138, 1977.

[9] d. S. V and L. L. Tensor rank and the Ill-Posedness of the best Low-Rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30:1084–1127, 2008.

[10] J. Brachat, P. Comon, B. Mourrain, and E. Tsigaridas. Symmetric tensor decomposition. *Linear Algebra and its Applications*, 433(11):1851–1872, 2010.

[11] J. Hoastad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4), 1990.

[12] C. J. Hillar and L. Lim. Most tensor problems are NP-Hard. *Journal of the ACM (JACM)*, 60, 2013.

[13] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002.

[14] A. Ziehe, P. Laskov, G. Nolte, and K. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research (JMLR)*, 5:777–800, 2004.

[15] R. Vollgraf and K. Obermayer. Quadratic optimization for simultaneous matrix diagonalization. *IEEE Transactions on Signal Processing*, 54(9): 3270–3278, 2006.

[16] V. Kuleshov, A. Chaganty, and P. Liang. Simultaneous diagonalization: the asymmetric, low-rank, and noisy settings. Technical report, arXiv, 2015.

[17] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. Technical report, arXiv, 2014.

[18] U. Shalit and G. Chechik. Coordinate-descent for learning orthogonal matrices through given rotations. In *International Conference on Machine Learning (ICML)*, 2014.

[19] A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. Technical report, arXiv, 2014.

[20] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*, 2012.

[21] S. A. Joint diagonalization: Is non-orthogonal always preferable to orthogonal? In *Computational Advances in Multi-Sensor Adaptive Processing*, pages 305–308, 2009.

[22] P. Comon, X. Luciani, and A. L. D. Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7):393–405, 2009.

[23] L. D. Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal of Matrix Analysis and Applications*, 28(3):642–666, 2006.

[24] B. Afsari. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1148–1171, 2008.

[25] J. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

[26] B. Afsari. Simple LU and QR based non-orthogonal matrix joint diagonalization. In *In-
dependent Component Analysis and Blind Signal Separation, pages 1–7, 2006.

[27] A. Bunse-Gerstner, R. Byers, and V. Mehrmann. Numerical methods for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 14(4):927–949, 1993.

[28] J. Cardoso. Perturbation of joint diagonalizers. Technical report, T’el’ecom Paris, 1994.

[29] A. Yeredor, A. Ziehe, and K. Müller. Approximate joint diagonalization using a natural gradient approach. *Independent Component Analysis and Blind Signal Separation*, 1:86–96, 2004.

[30] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. Independent component analysis and (simultaneous) third-order tensor diagonalization. *Signal Processing, IEEE Transactions on*, 49(10):2262–2271, 2001.

[31] F. Huang, U. N. Niranjan, M. U. Hakeem, and A. Anandkumar. Fast detection of overlapping communities via online tensor methods. Technical report, arXiv, 2013.

[32] H. N, M. P, and T. J. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:217–288, 2011.

[33] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

[34] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.
A Experiments

A.1 Synthetic experiments

Orthogonal tensors We start by generating random tensors $T = \sum_i \pi u_i^{\otimes 3} + \epsilon R$ with Gaussian entries in $\pi, R$ and $u_i$ distributed uniformly in the unit sphere $S^{d-1}$. We let $d = 25, 50, 100$ and in each case consider two regimes: undercomplete tensors with $k = 0.2d$ and full rank tensors, $k = d$. We vary $\epsilon$ and report the average error $\|\hat{u}_i - u_i\|_2$ across all eigenvectors $u_i$ and across 50 trials. In the orthogonal setting, we compare our algorithms (OJD0 uses random projections, OJD1 is with plugin) with the tensor power method (TPM), alternating least squares (ALS), and with the method of de Lathauwer [23]. Alternating least squares displayed very poor performance, and we omit it from our graphs. In the undercomplete case (Figure 4, right), all algorithms fare similarly and errors are within 10% of each other. Our method realizes its full potential in the full-rank setting, where OJD0 and OJD1 are up to three times more accurate than alternative methods (Figure 4, left).

Non-orthogonal tensors In the non-orthogonal setting, we compare de Lathauwer, alternating least squares (ALS), non-linear least squares (NLS), and our non-orthogonal methods (NOJD0 and NOJD1). We follow the same experimental setup as above and summarize our experiments in Figure 5. In the undercomplete setting, Lathauwer’s algorithm has the highest accuracy, about a 10% more than our approach (Figure 5, right). In the full rank setting, there is little difference in performance between our method and Lathauwer’s. In both settings, we consistently outperform the standard approaches, ALS and NLS, by 20-50% (Figure 5, left). Although we do
not always outperform Lathauwer’s state-of-the-art method, NOJD0 and NOJD1 are faster and much simpler to implement.

Asymmetric tensors  Lastly, we evaluate the extension of our algorithm to tensors of size $50 \times 50 \times 50$ having three distinct sets of asymmetric components (one in each mode). We find that performance is consistent with the symmetric setting, in both orthogonal and non-orthogonal regimes; our method outperforms its competitors by at least 25%, and in the non-orthogonal setting, it achieves an error reduction of up to 70% over Lathauer (Figure 6).

A.2  Algorithm running time

Figure 7 compares the running time in flops of the main algorithms.

We obtain the plots in Figure 7 by calculating flops as follows. The Jacobi method performs at each sweep $2dL(dk - \binom{k}{2})$ flops (where $L$ is the number of matrices); the QRJ1 non-orthogonal diagonalization algorithm performs $4d^3L$ flops per sweep. The tensor power method performs a total of $Lkd^3$ flops (where $L$ is the number of restarts), times the number of steps it takes to reach convergence for a given eigenvector. The flop count of Lathauwer’s method is much higher than that of other methods: at one stage, it requires finding the SVD of a $d^4 \times k^2$ matrix. Consequently, we do not include it in our summary.
B Proofs for orthogonal tensor factorization

In this section we prove perturbation bounds for our algorithm in the setting of orthogonal tensors.

Recall that we observe $\hat{T} = T + \epsilon R$ where $T = \sum_{i=1}^{k} \pi_i u_i \otimes u_i$ where $\pi_i$ are factor weights, $u_i \in \mathbb{R}^d$ are orthogonal unit vectors and $R$ is, without loss of generality, symmetric with $\|R\|_{op} = 1$. Our objective is to estimate $\pi$ and $(u_i)$. Algorithm 1 does so by simultaneously diagonalizing a number of projections of $T$; we make use of projections along random vectors and along approximate factors. In this section we will show why both schemes recover $\pi$ and $(u_i)$ with high probability.

**Setup** Let $\mathcal{M} = \{M_1, \ldots, M_L\}$ be the projections of $T$ along vectors $w_1, \ldots, w_L$, and $\hat{\mathcal{M}} = \{\hat{M}_1, \ldots, \hat{M}_L\}$ be the projections of $\hat{T}$ along $w_1, \ldots, w_L$. We have that $M_l = \sum_{i=1}^{d} \pi_i (w_i^\top u_i) u_i$ and that $\hat{M}_l = M_l + \epsilon R_l$, where $R_l = R(I, I, w_l)$. Thus, $M_l$ are a set of simultaneously diagonalizable matrices with factors $U$ and factor weights $\lambda_{il} \triangleq \pi_i (w_i^\top u_i)$. From the discussion in Section 2, let $\hat{U}$ be a full-rank extension of $U$, with columns $u_1, u_2, \ldots u_d$.

Let $\hat{\pi}$ and $\hat{u}$ be a factorization of $\hat{T}$ returned by Algorithm 1. From Lemma 1, we have that

$$\|\hat{u}_j - u_j\|_2 \leq \epsilon \sum_{i=1}^{d} E_{ij}^2 + o(\epsilon),$$

for $j \in [k]$ where $E \in \mathbb{R}^{d \times k}$ has entries

$$E_{ij} = \begin{cases} 0 & \text{for } i = j \\ \frac{-\sum_{l=1}^{L} (\lambda_{il} - \lambda_{lj}) w_i^\top R_l u_i}{\sum_{l=1}^{L} (\lambda_{il} - \lambda_{lj})^2} & \text{for } i \neq j. \end{cases}$$

For notational convenience, let $p_{ij} \triangleq (\pi_i u_i - \pi_j u_j)$ so that $\lambda_{il} - \lambda_{lj} = w_i^\top p_{ij}$. Let $r_{ij} \triangleq R(u_i, u_j, I)$ so that

$$u_j^\top R_l u_i = R(u_i, u_j, w_l) = R(u_i, u_j, I)^\top w_l = r_{ij}^\top w_l.$$

The expression for $E_{ij}$ when $j \neq i$ simplifies to,

$$E_{ij} = \frac{\sum_{l=1}^{L} w_i^\top p_{ij} r_{ij}^\top w_l}{\sum_{l=1}^{L} w_i^\top p_{ij} p_{ij}^\top w_l}. \tag{11}$$

In the rest of this section, we will bound $E_{ij}$ for different choices of $\{w_l\}_{l=1}^{L}$.

**B.1 Plugin projections**

In Section 4 we proposed using approximate factors $\hat{u}_i$ as directions to project the tensor $\hat{T}$ along. In this section, we show that doing so guarantees small errors in $u_i$.

We begin by bounding the terms $E_{ij}$.

**Lemma 3** ($E_{ij}$ with plug-in projections). Let $w_1, \ldots, w_k$ be unit-vectors approximations of the unit vectors $u_1, \ldots, u_k$: $\|w_l - u_l\|_2 \leq \gamma$ (so $L = k$), and let $\hat{\mathcal{M}} = \{\hat{M}_1, \ldots, \hat{M}_L\}$ be constructed via projection of $\hat{T}$ along $w_1, \ldots, w_L$. If the set of matrices $\hat{\mathcal{M}}$ is simultaneously diagonalized, then to a first-order approximation,

$$E_{ij} = \frac{p_{ij}^\top r_{ij}}{||p_{ij}||^2} + O(\gamma).$$

**Proof.** We have that

$$w_i^\top (p_{ij}) = (u_i + (w_i - u_i))^\top (\pi_i u_i - \pi_j u_j)$$

$$= \pi_i \delta_{il} - \pi_j \delta_{lj} + (w_i - u_i)^\top (\pi_i u_i - \pi_j u_j)$$

$$\leq \pi_i \delta_{il} - \pi_j \delta_{lj} + \|w_i - u_i\|_2 \|\pi_i u_i - \pi_j u_j\|_2$$

$$= \pi_i \delta_{il} - \pi_j \delta_{lj} + O(\gamma),$$

where $\delta_{il}$ is the Kronecker delta function.
where \( \delta_{ij} = 1 \) if \( i = j \) and 0 otherwise.

Thus,

\[
E_{ij} = \sum_{l=1}^{L} w_l^T p_{ij}^T w_l
\]

\[
= \sum_{l=1}^{L} \left( \pi_i \delta_{il} - \pi_j \delta_{jl} + O(\gamma) \right) r_{ij}^T w_l
\]

\[
= \frac{\pi_i r_{ij} w_i - \pi_j r_{ij} w_j + O(\gamma)}{\pi_i^2 + \pi_j^2 + O(\gamma)}
\]

\[
= \frac{\pi_i r_{ij} u_i + \pi_i (w_i - u_i)^T r_{ij} - \pi_j r_{ij} u_j - \pi_j (w_j - u_j)^T r_{ij} + O(\gamma)}{\pi_i^2 + \pi_j^2 + O(\gamma)}
\]

Note that \((w_i - u_i)^T r_{ij} = O(\gamma)\) and \((w_j - u_j)^T r_{ij} = O(\gamma)\), and hence both can be included in the \(O(\gamma)\) term.

\[
E_{ij} = \frac{\pi_i r_{ij} (\pi_i u_i - \pi_j u_j) + O(\gamma)}{\pi_i^2 + \pi_j^2 + O(\gamma)}
\]

Finally, recall that \( p_{ij} \triangleq (\pi_i u_i - \pi_j u_j) \) and that \( \|p_{ij}\|^2 = \pi_i^2 + \pi_j^2 \). Combining this with the observation that

\[
\frac{1}{1-x} = 1 + x + o(x)
\]

we obtain

\[
E_{ij} = \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|^2} + O(\gamma).
\]

Next, we use these term-wise bounds to bound the error in \( u_i \).

**Theorem 5** (Tensor factorization with plugin projections). Let \( w_1, \ldots, w_k \) be approximations of \( u_1, \ldots, u_k \) such that \( \|w_l - u_l\|_2 \leq \gamma = O(\epsilon) \), and let \( \hat{M} = \{\hat{M}_1, \ldots, \hat{M}_L\} \) be constructed via projection of \( \hat{T} \) along \( w_1, \ldots, w_L \). Then, for \( j \in [k] \),

\[
\|\tilde{u}_j - u_j\|_2 \leq \left( \frac{2 \sqrt{\|\pi_1 \pi_{\text{max}}\|}}{\pi_i^2} \right) \epsilon + o(\epsilon).
\]

**Proof.** From Equation 9, we have that,

\[
\|\tilde{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{j=1; j \neq i}^{d} E_{ij}^2},
\]

for all \( j \in [k] \). By Lemma 3, we get,

\[
E_{ij} = \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|^2} + O(\epsilon),
\]

and thus,

\[
\|\tilde{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1; i \neq j}^{d} \left( \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|^2} \right)^2} + o(\epsilon).
\]
Now, we must bound $\sum_{i=1; i \neq j}^{d} (p_{ij}^T r_{ij})^2$. We expect this the projection to mostly preserve the norm of $p_{ij}$ because $r_{ij}$ are effectively random vectors. Using Lemma 10 with $\mu = 0$, we get that $\sum_{i=1; i \neq j}^{d} (p_{ij}^T r_{ij})^2 \leq 4\|\pi\|_1 \pi_{\text{max}}$. Finally, $\|p_{ij}\|_2^2 = \pi_i^2 + \pi_j^2 \geq \pi_j^2$.

$$\|\tilde{u}_j - u_j\|_2 \leq \left(\frac{\sqrt{4\|\pi\|_1 \pi_{\text{max}}}}{\pi_j^2}\right) \epsilon + o(\epsilon)$$

$$\leq \left(\frac{2\sqrt{\|\pi\|_1 \pi_{\text{max}}}}{\pi_j^2}\right) \epsilon + o(\epsilon).$$

□

B.2 Random projections

Let us now consider the case when $\{w_l\}_{l=1}^L$ are random Gaussian vectors and present similar bounds. Given Equation 11, we should expect $E_{ij}$ to sharply, and now show that this is indeed the case.

**Lemma 4** (Concentration of error $E_{ij}$). Let $w_1, \ldots, w_L$ be i.i.d. random Gaussian vectors $w_l \sim \mathcal{N}(0, I)$, and let $\tilde{M} = \{\tilde{M}_1, \ldots, \tilde{M}_L\}$ be constructed via projection of $\tilde{T}$ along $w_1, \ldots, w_L$. If the set of matrices $\tilde{M}$ is simultaneously diagonalized, then the first-order error $E_{ij}$ is sharply concentrated. If $L \geq 16 \log(2/\delta)$, then with probability at least $1 - \delta$,

$$E_{ij} \leq \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|_2^2} + \frac{10 \log(2/\delta)}{\sqrt{L}} \|r_{ij}\|_2.$$

**Proof.** The numerator and denominator of Equation 11 are both distributed as the sum of $\chi^2$ variables; we show below that they respectively concentrate about $p_{ij}^T r_{ij}$ and $\|p_{ij}\|_2^2$.

From Lemma 14, we have that the following hold independently with probability at least $1 - \delta/2$,

$$\frac{1}{L} \sum_{l=1}^{L} w_l^T p_{ij}^T r_{ij}^T u_l \leq \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|_2^2} + \|p_{ij}\| \|r_{ij}\| \left(3\frac{\log(2/\delta)}{L}\right)$$

$$\frac{1}{L} \sum_{l=1}^{L} w_l^T p_{ij} p_{ij}^T u_l \geq \|p_{ij}\|^2 \left(1 - \frac{2 \log(2/\delta)}{\sqrt{L}}\right).$$

Applying a union bound on both these events, we get that with probability at least $1 - \delta$,

$$E_{ij} = \frac{\sum_{l=1}^{L} w_l^T p_{ij}^T r_{ij}^T u_l}{\sum_{m=1}^{L} \|w_m^T p_{ij}\|_2^2}$$

$$\leq \frac{p_{ij}^T r_{ij} + \|p_{ij}\| \|r_{ij}\| \left(3\frac{\log(2/\delta)}{L}\right)}{\|p_{ij}\|_2^2 \left(1 - \frac{2 \log(2/\delta)}{\sqrt{L}}\right)}.$$

Note that with the given condition on $L$, $\frac{2 \log(2/\delta)}{\sqrt{L}} < \frac{1}{2}$. Using the property that when $x \leq \frac{1}{2}$, $\frac{1}{1-x} \leq 1 + 2x$, we have that

$$\frac{1}{1 - \frac{2 \log(2/\delta)}{\sqrt{L}}} \leq 1 + \frac{4 \log(2/\delta)}{\sqrt{L}}.$$
Consequently,
\[
E_{ij} \leq \frac{1}{\|p_{ij}\|_2^2} \left( p_{ij}^T r_{ij} + \|p_{ij}\|_2 \|r_{ij}\|_2 \left( 3 \sqrt{\frac{\log(2/\delta)}{L}} \right) \right) \left( 1 + \frac{4 \log(2/\delta)}{\sqrt{L}} \right) \\
\leq \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|_2^2} \left( 1 + \frac{4 \log(2/\delta)}{\sqrt{L}} \right) + \frac{6 \|r_{ij}\|_2}{\|p_{ij}\|_2} \sqrt{\frac{\log(2/\delta)}{L}} \\
\leq \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|_2^2} + \frac{10 \log(2/\delta)}{\sqrt{L}} \frac{\|r_{ij}\|_2}{\|p_{ij}\|_2}.
\]

With this term-wise bound, we can again proceed to bounding the error \(u_j\).

**Theorem 6** (Tensor factorization with random projections). Let \(w_1, \ldots, w_L\) be i.i.d. random Gaussian vectors, \(w_i \sim \mathcal{N}(0, I)\), and let \(\tilde{M} = \{\tilde{M}_1, \ldots, \tilde{M}_L\}\) be constructed via projection of \(\tilde{T}\) along \(w_1, \ldots, w_L\). Furthermore, let \(L \geq 16 \log(2(d(k-1)/\delta)^2\), then, with probability at least 1 − \(\delta\),
\[
\|\tilde{u}_j - u_j\|_2 \leq \left( \frac{2\sqrt{2}\|\pi\|_{1\pi_{\max}}}{\pi_i^2} \right) \epsilon + \left( 20\sqrt{2} \log(2(d(k-1)/\delta) \frac{\sqrt{d/L}}{\pi_i} \right) \epsilon + o(\epsilon).
\]
for all \(j \in [k]\).

**Proof.** From Equation 9, we have that,
\[
\|\tilde{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1; i \neq j}^d E_{ij}^2} + o(\epsilon).
\]
By Lemma 4, with probability at least 1 − \(\delta/(d(k-1))\),
\[
E_{ij} \leq \frac{|p_{ij}^T r_{ij}|}{\|p_{ij}\|_2^2} + \frac{10 \log(2(d(k-1)/\delta)}{\sqrt{L}} \frac{\|r_{ij}\|_2}{\|p_{ij}\|_2}.
\]
Applying a union bound over \((E_{ij})_{j \neq i}^d\), we have that with probability at least 1 − \(\delta\),
\[
\|\tilde{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1; i \neq j}^d \left( \frac{p_{ij}^T r_{ij}}{\|p_{ij}\|_2^2} \right)^2} + \epsilon \frac{10 \log(2(d(k-1)/\delta)}{\sqrt{L}} \sum_{i=1; i \neq j}^d \left( \frac{\|r_{ij}\|_2}{\|p_{ij}\|_2} \right)^2 + o(\epsilon),
\]
for all \(j \in [k]\). We have used the fact that for \(a, b \geq 0\), \((a+b)^2 = a^2 + 2ab + b^2 \leq a^2 + (a^2 + b^2) + b^2 = 2a^2 + 2b^2\) and \(\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}\).

Note that \(\|p_{ij}\|_2 = \sqrt{\pi_i^2 + \pi_j^2} \geq |\pi_i|\). In Lemma 10, we show that \(\sum_{i=1; i \neq j}^d (p_{ij}^T r_{ij})^2 \leq 4\|\pi\|_{1\pi_{\max}}\). Furthermore, \(\|r_{ij}\|_1 \leq 1\) by the operator norm bound on \(R\). Thus, we get,
\[
\|\tilde{u}_j - u_j\|_2 \leq \left( \frac{2\sqrt{2}\|\pi\|_{1\pi_{\max}}}{\pi_i^2} \right) \epsilon + \left( 20\sqrt{2} \log(2(d(k-1)/\delta) \frac{\sqrt{d/L}}{\pi_i} \right) \epsilon + o(\epsilon).
\]

**C** Proofs for non-orthogonal tensor factorization

In this section we extend our previous analysis to non-orthogonal tensor decomposition.
Setup  As before, let \( M = \{M_1, \ldots, M_L\} \) be the projections of \( T \) along vectors \( w_1, \ldots, w_L \), and \( \hat{M} = \{\hat{M}_1, \ldots, \hat{M}_L\} \) be the projections of \( \hat{T} \) along \( \hat{w}_1, \ldots, \hat{w}_L \). We have that \( M_l = \sum_{i=1}^d \pi_i(w_i^\top u_i)u_i \otimes u_i \) and that \( \hat{M}_l = M_l + \epsilon R_l \), where \( R_l = R(I, I, w_l) \). Thus, \( M_l \) are a set of simultaneously diagonalizable matrices with factors \( U \) and factor weights \( \lambda_{il} \). Let \( \hat{U} \) be the full-rank extension of \( U \) with unit-norm columns \( v_1, v_2, \ldots, v_d \). In this setting, however, the factor \( U \) is not orthogonal. Let \( V = U^{-1} \), with rows \( v_1, v_2, \ldots, v_d \). Note that we place our incoherence assumption on the columns of \( U \) and present results in terms of the 2-norm of \( V^\top \). When \( U \) is incoherent, it can be shown that \( \|V^\top\|_2 \leq 1 + O(\mu) \). Finally, note that in the orthogonal case, when \( \mu = 0 \), the rows \( (v_i) \) and columns \( (u_i) \) are identical, and no distinction between the two need be made.

Let \( \hat{\pi} \) and \( \hat{u} \) be a factorization of \( \hat{T} \) returned by Algorithm 1. From Lemma 2, we have that

\[
\|\hat{u}_j - u_j\|_2 = \sqrt{\sum_{i=1}^d E_{ij}},
\]

where the entries of \( E \in \mathbb{R}^{d \times k} \) are bounded by Lemma 17:

\[
|E_{ij}| \leq \frac{1}{1 - \rho_{ij}} \left( \frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2} \right) \left( \|\sum_{l=1}^L v_i^\top R_l v_j \lambda_{jl}\|_2 + \|\sum_{l=1}^L v_i^\top R_l v_j \lambda_{il}\|_2 \right),
\]

where \( \lambda_i \in \mathbb{R}^L \) is the vector of \( i \)-th factor values of \( M_l \), i.e. \( \lambda_{il} \) is the \( i \)-th factor value of matrix \( M_l \) (i.e. \( \lambda_{il} = (\Lambda_i)_i \)) and \( \rho_{ij} = \|\lambda_i\|_2 \|\lambda_j\|_2 \), the modulus of uniqueness, is a measure of the singularity of the problem.

When \( \lambda_{il} \) is generated by projections, \( \lambda_{il} = \pi_i w_i^\top u_i \). Let \( r_{ij} \triangleq R(v_i, v_j, I) \) so that

\[
v_i^\top R_i v_j = R(v_i, v_j, w_l) = R(v_i, v_j, I)^\top w_l = r_{ij} w_l.
\]

Note that \( \|r_{ij}\|_2 \leq \|v_i\|_2 \|v_j\|_2 \leq \|V^\top\|_2 \).

Equation 12 then simplifies to,

\[
|E_{ij}| \leq \frac{1}{1 - \rho_{ij}} \left( \frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2} \right) \left( |\pi_j| \|\sum_{l=1}^L w_i^\top u_j r_{ij}^\top w_l\|_2 + |\pi_i| \|\sum_{l=1}^L w_i^\top u_i r_{ij}^\top w_l\|_2 \right),
\]

where \( \|\lambda_i\|_2 = \sum_{l=1}^L w_l^\top u_i u_i^\top w_l \), and \( \rho_{ij} \) has the following expression,

\[
\rho_{ij} = \frac{\|\lambda_i\|_2 \|\lambda_j\|_2}{\sum_{l=1}^L w_l^\top u_i u_i^\top w_l} = \frac{-\sum_{l=1}^L w_l^\top u_i u_j^\top w_l}{\sqrt{\sum_{l=1}^L w_i^\top u_i u_i^\top w_l} (\sum_{l=1}^L w_i^\top u_j^\top u_j^\top w_l)}.
\]

Observe that the terms \( u_i \) interact with the factor weights \( \lambda_{il} \), while the terms \( v_i \) interact only with the noise terms \( R_l \).

In the rest of this section, we will bound \( E_{ij} \) and \( \rho_{ij} \) with different choices of \( \{w_l\}_{l=1}^L \).

### C.1 Plugin projections

We now assume we have plugin estimates \( (w_l) \) that are close to the inverse factors \( (v_i) \): \( \|w_l - v_l\|_2 \leq O(\gamma) \) for \( l \in [k] \). Then,

\[
\begin{align*}
    w_l^\top u_i &= (v_i + (w_l - v_l))^\top u_i \\
    &= v_i^\top u_i + \|w_l - v_l\|_2 \cdot \frac{(w_l - v_l)^\top u_i}{\|w_l - v_l\|_2} \\
    &= v_i^\top u_i + O(\gamma).
\end{align*}
\]

Recall that \( V = U^{-1} \), so \( v_i^\top u_i = \delta_{il} \).
It will be useful to keep track of $\|\lambda_i\|_2^2$,

$$
\|\lambda_i\|_2^2 = \sum_{l=1}^{L} \pi_i^2 (w^\top_l u_i)^2 \\
= \pi_i^2 \sum_{l=1}^{k} (v^\top_l u_i + O(\gamma))^2 \\
= \pi_i^2 + O(\gamma).
$$

(15)

**Lemma 5** (Modulus of uniqueness for plugin projections). Let $w_1,\ldots,w_k$ be approximations of $v_1,\ldots,v_k$: $\|w_l - v_l\|_2 \leq O(\gamma)$ for $l \in [k]$, and let $\hat{\mathcal{M}} = \{\hat{M}_1,\ldots,\hat{M}_L\}$ be constructed via projection of $\hat{T}$ along $w_1,\ldots,w_L$. Then, for $i \neq j$,

$$
\rho_{ij}^2 \leq O(\gamma),
$$

**Proof.** Let us first bound the numerator of Equation 14.

$$
(\lambda_i^\top \lambda_j)^2 = \pi_i^2 \pi_j^2 \left( \sum_{l=1}^{L} w^\top_l u_j^\top u_i^\top w_l \right)^2 \\
= \pi_i^2 \pi_j^2 \left( \sum_{l=1}^{L} v^\top_l u_j^\top u_i^\top v_l + O(\gamma) \right)^2 \\
= \pi_i^2 \pi_j^2 \delta_{ij} + O(\gamma) \\
= O(\gamma).
$$

Using Equation 15, we get that

$$
\rho_{ij}^2 = \frac{O(\gamma)}{(1 + O(\gamma))(1 + O(\gamma))} \\
= O(\gamma).
$$

where in the last line we used the fact that $\frac{1}{1+x} = 1 + x + o(x)$.

**Lemma 6** (Bound on $E_{ij}$ for non-orthogonal plugin projections). Let $w_1,\ldots,w_k$ be approximations of $v_1,\ldots,v_k$: $\|w_l - v_l\|_2 \leq O(\gamma)$ for $l \in [k]$, and let $\hat{\mathcal{M}} = \{\hat{M}_1,\ldots,\hat{M}_L\}$ be constructed via projection of $\hat{T}$ along $w_1,\ldots,w_L$.

$$
|E_{ij}| \leq \left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right) \|V\|_2 \rho_{ij}^r + O(\gamma),
$$

where $p_{ij} \triangleq \frac{\pi_i}{\|v_i\|_2} + \frac{\pi_j}{\|v_j\|_2}$.

**Proof.** Let us bound each term within our expression for $E_{ij}$ (Equation (13)).

$$
\sum_{l=1}^{k} w^\top_l u_j r^\top_{ij} w_l = \sum_{l=1}^{k} v^\top_l u_j r^\top_{ij} v_l + O(\gamma) \\
\leq r^\top_{ij} v_j + O(\gamma).
$$

Similarly,

$$
\sum_{l=1}^{k} w^\top_l u_i r^\top_{ij} w_l \leq r^\top_{ij} v_i + O(\gamma),
$$
From Equation (15), we have
\[ \|\lambda_i\|^2_2 = |\pi_i|^2 + O(\gamma) \]
\[ \|\lambda_j\|^2_2 = |\pi_j|^2 + O(\gamma). \]

From Lemma 5 we have that
\[ \rho_{ij}^2 \leq O(\gamma) \]
\[ \frac{1}{1 - \rho_{ij}^2} \leq \frac{1}{1 - O(\gamma)} + O(\gamma) \]
\[ \leq 1 + O(\gamma). \]

Finally,
\[ |E_{ij}| \leq \left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right) \left( (|\pi_i|v_i + |\pi_j|v_j)^\top r_{ij} \right) + O(\gamma) \]
\[ \leq \left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right) \|V^\top\|_2 p_{ij}^\top r_{ij} + O(\gamma). \]

Note that the error terms depend not on \(u_i\) but rather \(v_i\). This is because the projections \((w_l)\) are chosen to be close to the \(v_i\). Now, let us bound the error in \(u_i\).

**Theorem 7 (Non-orthogonal tensor factorization with plug-in projections).** Let \(w_1,\ldots,w_k\) be approximations of \(v_1,\ldots,v_k\): \(\|w_l - v_l\|_2 \leq O(\epsilon)\) for \(l \in [k]\) and let \(\hat{\mathcal{M}} = \{\hat{M}_1,\ldots,\hat{M}_L\}\) be constructed via projection of \(\hat{T}\) along \(w_1,\ldots,w_L\). Then, for all \(j \in [k]\),
\[ \|\hat{u}_j - u_j\|_2 \leq 8\epsilon \sqrt{\frac{\|\pi\|_1}{\pi_{\max}^2}} \|V^\top\|_2^3 + o(\epsilon). \]

**Proof.** From Lemma 16 we have that
\[ \|\hat{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon), \]
for \(j \in [k]\), where \(E_{ij}\) is bounded in Lemma 6 as follows:
\[ |E_{ij}| \leq \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \|V^\top\|_2 p_{ij}^\top r_{ij} + O(\epsilon) \]
\[ \leq \frac{2}{\pi_{\min}} \|V^\top\|_2 p_{ij}^\top r_{ij} + O(\epsilon). \]

Consequently,
\[ \|\hat{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i \neq j}^{d} E_{ij}^2} \]
\[ \leq \frac{2\epsilon}{\pi_{\min}^2} \sqrt{\sum_{i \neq j}^{d} (\|V^\top\|_2 p_{ij}^\top r_{ij} + O(\epsilon))^2} + o(\epsilon) \]
\[ \leq \frac{4\epsilon}{\pi_{\min}^2} \left( \sqrt{\sum_{i \neq j}^{d} (\|V^\top\|_2 p_{ij}^\top r_{ij})^2} + o(\epsilon) \right). \]
where we have used the fact that \((a + b)^2 \leq 2(a^2 + b^2)\) and that \(\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}\).

From Lemma 10 we have, \(p_{ij}^T r_{ij} \leq 4\|\pi\|_1 \pi_{\text{max}} \|V^T\|_2^4\),

\[
\|\tilde{u}_j - u_j\|_2 \leq \frac{4\epsilon}{\pi_{\text{min}}} \left( \sqrt{4\|\pi\|_1 \pi_{\text{max}} \|V^T\|_2^6} + o(\epsilon) \right) \\
\leq 8\epsilon \frac{\sqrt{\|\pi\|_1 \pi_{\text{max}}}}{\pi_{\text{min}}} \|V^T\|_2^3 + o(\epsilon).
\]

\[\square\]

C.2 Random projections

We now study the case where the random projections, \((w_l)\), are drawn from a standard Gaussian distribution.

First let us show that the modulus of uniqueness \(\rho_{ij}\) sharply concentrates around \(u_i^T u_j\).

**Lemma 7** (Modulus of Uniqueness with random projections). Let \(w_1, \cdots, w_L \in \mathbb{R}^d\) be entries drawn i.i.d. from the standard Normal distribution. Let \(L > 16 \log(3/\delta)^2\), then, with probability at least \(1 - \delta\),

\[\rho_{ij} \leq u_i^T u_j + \frac{10 \log(3/\delta)}{\sqrt{L}}.\]

**Proof.** Observe from Equation 14 that the numerator and the denominator of \(\rho_{ij}\) are essentially distributed as a \(\chi^2\) distribution (Lemma 14). Thus, with probability at least \(1 - \delta/3\) each, the following hold,

\[
\frac{1}{L} \sum_{l=1}^{L} (u_i^T u_j) w_l \leq u_i^T u_j + \|u_i\|_2^2 \|u_j\|_2 \left( 3 \sqrt{\frac{\log(3/\delta)}{L}} \right) \\
\frac{1}{L} \sum_{l=1}^{L} (u_i^T u_j)^2 \geq \|u_i\|_2^2 \left( 1 - \frac{2 \log(3/\delta)}{\sqrt{L}} \right) \\
\frac{1}{L} \sum_{l=1}^{L} (u_i^T u_j)^2 \geq \|u_j\|_2^2 \left( 1 - \frac{2 \log(3/\delta)}{\sqrt{L}} \right).
\]

Noting that \(\|u_i\|_2 = \|u_j\|_2 = 1\) and applying a union bound on the above three events, we get that with probability at least \(1 - \delta\),

\[\rho_{ij} \leq u_i^T u_j + 3 \sqrt{\frac{\log(3/\delta)}{L}} + \frac{10 \log(3/\delta)}{\sqrt{L}}.\]

Under the conditions on \(L\), \(\frac{2 \log(3/\delta)}{\sqrt{L}} \leq \frac{1}{2}\). Applying the property that when \(x < \frac{1}{2}\), \(\frac{1}{1 - \frac{2x}{\sqrt{L}}} \leq 1 + \frac{4 \log(3/\delta)}{\sqrt{L}} < 2\),

Finally,

\[\rho_{ij} \leq u_i^T u_j + \frac{10 \log(3/\delta)}{\sqrt{L}}.\]

\[\square\]
Let’s now bound the inverse modulus of uniqueness.

**Lemma 8** (Bounding inverse modulus of uniqueness). Let \( w_1, \ldots, w_L \in \mathbb{R}^d \) be entries drawn i.i.d. from the standard Normal distribution. Assume incoherence \( \mu \) for that the \((u_i)\): \( u_i^\top u_j \leq \mu \) for \( i \neq j \). Let \( L_0 \triangleq \left( \frac{50}{(1 - \mu^2)^2} \right)^2 \). Let \( L \geq L_0 \log(3/\delta)^2 \). Then, with probability at least \( 1 - \delta \),

\[
\frac{1}{1 - \rho_{ij}^2} \leq \frac{1}{1 - (u_i^\top u_j)^2} \left( 1 + \sqrt{\frac{L_0}{L}} \log(3/\delta) \right).
\]

**Proof.** From Lemma 7, we have that with probability at least \( 1 - \delta \),

\[
\rho_{ij} \leq u_i^\top u_j + \frac{10 \log(3/\delta)}{\sqrt{L}}.
\]

Then,

\[
\rho_{ij}^2 \leq (u_i^\top u_j)^2 + 2u_i^\top u_j \left( \frac{10 \log(3/\delta)}{\sqrt{L}} \right) + \left( \frac{10 \log(3/\delta)}{\sqrt{L}} \right)^2.
\]

Given the assumptions on \( L \), we have that \( L \geq L_0 \log(3/\delta)^2 \geq 50 \log(3/\delta)^2 \) and thus \( \frac{10 \log(3/\delta)}{\sqrt{L}} \leq \frac{1}{2} \):

\[
\rho_{ij}^2 \leq (u_i^\top u_j)^2 + 2 \left( \frac{10 \log(3/\delta)}{\sqrt{L}} \right) + \frac{10 \log(3/\delta)}{\sqrt{L}} = (u_i^\top u_j)^2 + \frac{25 \log(3/\delta)}{\sqrt{L}}.
\]

Now, we bound \( \frac{1}{1 - \rho_{ij}^2} \),

\[
\frac{1}{1 - \rho_{ij}^2} \leq \frac{1}{1 - (u_i^\top u_j)^2} - \frac{25 \log(3/\delta)}{\sqrt{L}} \leq \frac{1}{1 - (u_i^\top u_j)^2} - \frac{10 \log(3/\delta)}{\sqrt{L}} \leq \frac{1}{1 - (u_i^\top u_j)^2} - \frac{1}{1 - \frac{1}{2} \log(3/\delta)} \sqrt{\frac{L_0}{L}}.
\]

Again, given assumptions on \( L \), \( \frac{1}{2} \log(3/\delta) \sqrt{\frac{L_0}{L}} \leq \frac{1}{2} \). Using the identity that if \( x < \frac{1}{2} \), \( \frac{1}{1 - x} \leq 1 + 2x \),

\[
\frac{1}{1 - \rho_{ij}^2} \leq \frac{1}{1 - (u_i^\top u_j)^2} \left( 1 + \log(3/\delta) \sqrt{\frac{L_0}{L}} \right).
\]

\[ \square \]

We are now ready to bound the termwise entries of \( E \).

**Lemma 9** (Concentration of \( E_{ij} \)). Let \( w_1, \ldots, w_L \) be i.i.d. random Gaussian vectors \( w_l \sim \mathcal{N}(0, I) \), and let \( \hat{M} = \{\hat{M}_1, \ldots, \hat{M}_L\} \) be constructed via projection of \( \hat{T} \) along \( w_1, \ldots, w_L \). Assume incoherence \( \mu \) for that the \((u_i)\): \( u_i^\top u_j \leq \mu \) for \( i \neq j \). Furthermore, let \( L \geq L_0 \log(15/\delta)^2 \). Then, with probability at least \( 1 - \delta \),

\[
|E_{ij}| \leq \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \left( \bar{\pi}_{ij} r_{ij} + \frac{\bar{\pi}_{ij} \|r_{ij}\|_2}{1 - (u_i^\top u_j)^2} \left( 20 + \sqrt{L_0} \right) \log(15/\delta) \right),
\]

where \( \bar{\pi}_{ij} \triangleq |\pi_i|u_i + |\pi_j|u_j \) and \( \bar{\pi}_{ij} \triangleq |\pi_i| + |\pi_j| \).
Proof. Each term in Equation 13 concentrates sharply about its mean value. We bound each in turn.

First, consider \( \|\lambda_i\|_2^2/L = \frac{1}{L} |\pi_i|^2 \sum_{l=1}^{L} (u_l^T u_i)^2 \). With probability at least \( 1 - \delta/5 \) each, the following hold,

\[
\frac{1}{L} \|\lambda_i\|_2^2 \geq \pi_i^2 \|u_i\|_2^2 \left(1 - \frac{2 \log(5/\delta)}{\sqrt{L}}\right)
\]

\[
\frac{1}{L} \|\lambda_j\|_2^2 \geq \pi_j^2 \|u_j\|_2^2 \left(1 - \frac{2 \log(5/\delta)}{\sqrt{L}}\right).
\]

Thus, using the fact that \( \|u_i\|_2^2 = 1 \),

\[
L \left(\frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2}\right) \leq \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \left(1 - \frac{2 \log(5/\delta)}{\sqrt{L}}\right).
\]

Given our assumption on \( L \), it follows that \( \frac{2 \log(5/\delta)}{\sqrt{L}} \leq \frac{1}{2} \). Thus we can use the fact that \( \frac{1}{1-x} \leq 1 + 2x \) when \( x \leq \frac{1}{2} \) to obtain the following bound:

\[
L \left(\frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2}\right) \leq \left(\frac{1}{\pi_i^2} + \frac{1}{\pi_j^2}\right) \left(1 + 4 \log(5/\delta)\right).
\]

Next, we bound \( \frac{1}{L} \sum_{l=1}^{L} w_l^T u_i r_{ij}^T w_l \) and \( \frac{1}{L} \sum_{l=1}^{L} w_l^T u_j r_{ij}^T w_l \). From Lemma 14, we have with probability at least \( 1 - \delta/5 \) each,

\[
\frac{1}{L} \sum_{l=1}^{L} w_l^T u_i r_{ij}^T w_l \leq \pi_i \|r_{ij}\|_2 \|u_j\|_2 \left(3 \sqrt{\frac{\log(5/\delta)}{L}}\right)
\]

\[
\frac{1}{L} \sum_{l=1}^{L} w_l^T u_j r_{ij}^T w_l \leq \pi_j \|r_{ij}\|_2 \|u_i\|_2 \left(3 \sqrt{\frac{\log(5/\delta)}{L}}\right).
\]

Note that by definition, \( \|u_i\|_2 = 1 \).

Using Lemma 8, we have that with probability at least \( 1 - \delta/5 \),

\[
\frac{1}{1 - \rho_{ij}} \leq \frac{1}{1 - (u_i^T u_j)^2} \left(1 + \sqrt{\frac{L_0}{L}} \log(15/\delta)\right).
\]

Putting it all together, we get that with probability at least \( 1 - \delta \),

\[
|E_{ij}| \leq \frac{1}{1 - (u_i^T u_j)^2} \left(1 + \sqrt{\frac{L_0}{L}} \log(15/\delta)\right) \left(\frac{1}{\pi_i^2} + \frac{1}{\pi_j^2}\right) \left(1 + \frac{4 \log(5/\delta)}{\sqrt{L}}\right)
\]

\[
\left(|\pi_i| r_{ij}^T u_i + |\pi_j| r_{ij}^T u_j + (|\pi_i| + |\pi_j|) \|r_{ij}\|_2 \left(3 \sqrt{\frac{\log(5/\delta)}{L}}\right)\right).
\]

Let us define \( \bar{\rho}_{ij} \triangleq |\pi_i| u_i + |\pi_j| u_j \) and \( \bar{\pi}_{ij} \triangleq |\pi_i| + |\pi_j| \):

\[
|E_{ij}| \leq \frac{1}{1 - (u_i^T u_j)^2} \left(1 + \sqrt{\frac{L_0}{L}} \log(15/\delta)\right) \left(\frac{1}{\pi_i^2} + \frac{1}{\pi_j^2}\right) \left(1 + \frac{4 \log(5/\delta)}{\sqrt{L}}\right)
\]

\[
\left(\bar{\rho}_{ij} r_{ij} + \bar{\pi}_{ij} \|r_{ij}\|_2 \left(3 \sqrt{\frac{\log(5/\delta)}{L}}\right)\right).
\]

Given that \( L \geq L_0 \log(15/\delta)^2 \), we have that \( \sqrt{\frac{L_0}{L}} \log(15/\delta) \leq 1 \) and \( \frac{4 \log(5/\delta)}{\sqrt{L}} \leq 1 \), thus

\[
\left(1 + \sqrt{\frac{L_0}{L}} \log(15/\delta)\right) \left(1 + \frac{4 \log(5/\delta)}{\sqrt{L}}\right) \leq 2 \times 2 \leq 4.
\]
Finally, note that $|\pi_i r_{ij}^T u_i + |\pi_j r_{ij}^T u_j | \leq (|\pi_i| + |\pi_j|) \| r_{ij} \|_2$, giving us,

\[
|E_{ij}| \leq \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \left( \frac{\bar{p}_{ij}^T r_{ij}}{1 - (u_i^T u_j)^2} \right) + \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \frac{\pi_{ij} \| r_{ij} \|_2}{1 - (u_i^T u_j)^2} \left( \sqrt{\frac{L_0}{L}} \log(15/\delta) + 2 \frac{4(\log(5/\delta))}{\sqrt{L}} + 4 \left( \frac{3}{\log(5/\delta)/L} \right) \right)
\]

\[
\leq \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \left( \frac{\bar{p}_{ij}^T r_{ij}}{1 - (u_i^T u_j)^2} \right) + \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \frac{\pi_{ij} \| r_{ij} \|_2}{1 - (u_i^T u_j)^2} \left( 20 + \sqrt{L_0} \log(15/\delta) \right).
\]

\[
\square
\]

Finally, we bound the error in estimating $u_j$.

**Theorem 8** (Non-orthogonal tensor factorization with random projections). Let $w_1, \ldots, w_L$ be i.i.d. random Gaussian vectors, $w_l \sim \mathcal{N}(0, I)$, and let $\mathcal{M} = \{ \bar{M}_1, \ldots, \bar{M}_L \}$ be constructed via projection of $\bar{T}$ along $w_1, \ldots, w_L$. Assume incoherence $\mu \leq \frac{1}{d^2}$ for both $(u_i)$ and $(v_i)$: $u_i^T u_j \leq \mu$ and $v_i^T v_j \leq \mu$ for $i \neq j$. Let $L_0 \doteq \left( \frac{50}{1-\mu^2} \right)^2$. Let $L \geq L_0 \log(15d(k-1)/\delta)^2$. Then, with probability at least $1 - \delta$ and for $\epsilon$ small enough,

\[
\| \tilde{u}_j - u_j \|_2 \leq \frac{8\epsilon}{1 - \mu^2} \sqrt{\frac{\| \pi_{1, \pi_{ij}}^T \| \| v_{1, v_j}^T \|}{\pi_{\min}^2}} \left( 1 + C(\delta) \sqrt{d} \right),
\]

where $C(\delta) \doteq \frac{20 + \sqrt{L_0}}{\sqrt{L}} \log(15(d(k-1))/\delta)$.

**Proof.** From **Lemma 16** we have that

\[
\| \tilde{u}_j - u_j \|_2 \leq \epsilon \sqrt{d} \sum_{i=1}^{d} E_{ij}^2 + o(\epsilon),
\]

for $j \in [k]$.

Using **Lemma 9**, we have that with probability at least $1 - \delta/(d(k-1))$,

\[
|E_{ij}| \leq \left( \frac{1}{\pi_i^2} + \frac{1}{\pi_j^2} \right) \left( \frac{\bar{p}_{ij}^T r_{ij}}{1 - (u_i^T u_j)^2} \right) + \frac{\pi_{ij} \| r_{ij} \|_2}{1 - (u_i^T u_j)^2} \left( 20 + \sqrt{L_0} \log(15(d(k-1))/\delta) \right)
\]

\[
\leq \left( \frac{1}{\pi_{\min}^2} \right) \left( \frac{1}{1 - \mu^2} \right) \left( \frac{\bar{p}_{ij}^T r_{ij} + 2\pi_{\min} \| v_{1, v_j}^T \|}{1 - (u_i^T u_j)^2} \left( 20 + \sqrt{L_0} \log(15(d(k-1))/\delta) \right) \right)
\]

\[
\leq \left( \frac{2}{\pi_{\min}} \right) \left( \frac{1}{1 - \mu^2} \right) \left( \bar{p}_{ij}^T r_{ij} + 2\pi_{\min} \| V^T \|_2^2 \right) C(\delta),
\]

where we have defined $C(\delta) \doteq \frac{20 + \sqrt{L_0}}{\sqrt{L}} \log(15(d(k-1))/\delta)$ and are using the fact that $u_i^T u_j \leq \mu$ and $\pi_{ij} = |\pi_i| + |\pi_j| \leq 2|\pi_{\max}|$. 
Applying a union bound on all the entries of $E_{ij}$, we arrive at the following bound for all $j$.

$$\|u_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i \neq j} E_{ij}^2}$$

$$\leq \frac{2\epsilon}{\pi^2_{\min}(1 - \mu^2)} \sqrt{\left( \sum_{i \neq j} p_{ij}^r r_{ij}^2 + 2\pi_{\max}\|V^T\|_2^2 C(\delta) \right)^2}$$

$$\leq \frac{4\epsilon}{\pi^2_{\min}(1 - \mu^2)} \left( \sqrt{\left( \sum_{i \neq j} (p_{ij}^r r_{ij})^2 + 2\pi_{\max}\|V^T\|_2^2 C(\delta) \right)^2} \sum_{i \neq j} 1 \right)^{1/2}$$

$$\leq \frac{4\epsilon}{\pi^2_{\min}(1 - \mu^2)} \left( \sum_{i \neq j} (p_{ij}^r r_{ij})^2 + 2\pi_{\max}\|V^T\|_2^2 C(\delta) \sqrt{d} \right).$$

where we use the fact that $(a + b)^2 \leq 2(a^2 + b^2)$.

By Lemma 10 we also have, $\sum_{i \neq j} (p_{ij}^r r_{ij})^2 \leq 4\pi_{\max}\|1\|_{\max}\|V^T\|_2^2$. Finally, note that $\pi_{\max} \leq \sqrt{\pi_{\max}\|1\|_{\max}}$:

$$\|u_j - u_j\|_2 \leq \frac{4\epsilon}{\pi^2_{\min}(1 - \mu^2)} \left( \sqrt{4\pi_{\max}\|1\|_{\max}\|V^T\|_2^2} + 2\pi_{\max}\|V^T\|_2^2 C(\delta) \sqrt{d} \right)$$

$$\leq \frac{8\epsilon}{1 - \mu^2} \frac{\sqrt{\|V^T\|_2^2}}{\pi^2_{\min}} \left( 1 + C(\delta) \sqrt{d} \right).$$

$\square$

### D Proofs of auxiliary lemmas

In this section, we prove some auxiliary results that appear as intermediate steps in the main lemmas above.

**Lemma 10** (Bounding $p_{ij}^r r_{ij}$). Let $p_{ij} \triangleq \pi_i u_i - \pi_j u_j \in \mathbb{R}^d$ and $r_{ij} \triangleq R(v_i, v_j, I) \in \mathbb{R}^d$, where $R$ is a tensor with unit operator norm and where $(u_i) \in \mathbb{R}^d$ are unit vectors and $(v_i) \in \mathbb{R}^d$ form the columns of the matrix $V$ with bounded 2 norm. Then,

$$\sum_{i \neq j} (p_{ij}^r r_{ij})^2 \leq 4\pi_{\max}\|1\|_{\max}\|V\|_2^4.$$ 

**Proof.** Firstly, note that it is trivial to bound the sum as follows,

$$\sum_{i \neq j} (p_{ij}^r r_{ij})^2 \leq \sum_{i \neq j} \|p_{ij}\|_2^2 \|r_{ij}\|_2^2$$

$$\leq 4(d - 1)\pi_{\max}^2\|V\|_2^4.$$ 

using the properties that $p_{ij} \triangleq \pi_i u_i - \pi_j u_j$ and that $R$ has unit operator norm and thus $\|p_{ij}\|_2 \leq 2\pi_{\max}$ and $\|r_{ij}\|_2 = \|R(v_i, v_j, I)\|_2 \leq \|V\|_2^2$.

However, we would like a tighter bound with a lower-order dependence on $k$. To do so, let us expand $p_{ij}$,

$$\sum_{i \neq j} (p_{ij}^r r_{ij})^2 = \sum_{i \neq j} ((\pi_i u_i - \pi_j u_j)^\top r_{ij})^2$$

$$= \sum_{i \neq j} (\pi_i R(v_i, v_j, u_i) - \pi_j R(v_i, v_j, u_j))^2$$

$$= \sum_{i \neq j} \pi_i^2 R(v_i, v_j, u_j)^2 + \sum_{i \neq j} \pi_j^2 R(v_i, v_j, u_i)^2 - \sum_{i \neq j} 2\pi_i \pi_j R(v_i, v_j, u_i)R(v_i, v_j, u_j).$$
Using the assumption that $R$ has unit norm, the latter two terms can be bounded by $\|\pi\|_2^2\|V\|_2$ and $2\,\pi_j\|\pi\|_1\|V\|_2^4$ respectively.

We now focus on the first term, $\pi_j^2 \sum_{i\neq j}^d R(v_i, v_j, u_j)^2$. Note that $R(v_i, v_j, u_j) = R(I, v_j, u_j)^\top v_i = \tilde{r}_j^\top v_i$, where $\tilde{r}_j \triangleq R(I, v_j, u_j)$ and $\|\tilde{r}_j\|_2 \leq \|V\|_2$ by the operator norm condition on $R$.

$$
\sum_{i=1}^d (\tilde{r}_j^\top v_i)^2 = \|V\|_2^2 \\
\leq \|V\|_2^2\|\tilde{r}_j\|_2^2 \\
= \|V\|_2^2
$$

Put together, we get that,

$$
\sum_{i\neq j}^d (p_{ij}^\top r_{ij})^2 \leq \pi_j^2 \|V\|_2^4 + \|\pi\|_2^2\|V\|_2^4 + 2\pi_j\|\pi\|_1\|V\|_2^4.
$$

Finally, $\pi_j^2 \leq \pi_{\text{max}}\|\pi\|_1$ and, by Hölder’s inequality, $\|\pi\|_2^2 \leq \pi_{\text{max}}\|\pi\|_1$, giving us,

$$
\sum_{i\neq j}^d (p_{ij}^\top r_{ij})^2 \leq 4\pi_{\text{max}}\|\pi\|_1\|V\|_2^4.
$$

\[\square\]

E  Incoherence lemmas

In this section, we prove a couple of useful facts when dealing with incoherent vectors.

**Lemma 11 (Projection on incoherent vectors).** Let $u_1, \ldots, u_k \in \mathbb{R}^d$ be a set of incoherent unit vectors: $|u_i^\top u_j| \leq \mu$ for $i \neq j$. If $\mu \leq 1/(k-1)$, then, for any vector $r \in \mathbb{R}^d$,

$$
\sum_{l=1}^k |r^\top u_l| \leq \frac{1 + (k - 1)\mu}{\sqrt{1 - (k - 1)\mu}} \sqrt{k}\|r\|_2 \\
\sum_{l=1}^k (r^\top u_l)^2 \leq \frac{(1 + (k - 1)\mu)^2}{1 - (k - 1)\mu} \|r\|_2^2.
$$

Furthermore, if $\mu \leq 1/2(k - 1)$, then the bounds simplify to,

$$
\sum_{l=1}^k |r^\top u_l| \leq (1 + (k - 1)\mu) \sqrt{2k}\|r\|_2 \\
\sum_{l=1}^k (r^\top u_l)^2 \leq (1 + 7(k - 1)\mu) \|r\|_2^2.
$$

**Proof.** Without loss of generality, let $\|r\|_2 = 1$. Let $r = \sum_{l=1}^k \rho_l u_l \approx \rho \perp u_\perp$, where $u_\perp$ is a unit vector orthogonal to $u_1, \ldots, u_k$. Let $\rho = (\rho_1, \ldots, \rho_k) \in \mathbb{R}^k$ be the vector of coefficients.

Firstly, note that

$$
\|r\|_2^2 = \left(\sum_{l=1}^k \rho_l u_l \approx \rho \perp u_\perp\right)^2 \\
= \sum_{l=1}^k \rho_l^2 + \sum_{l=1}^k \sum_{l' \neq l}^k \rho_l \rho_{l'} u_l^\top u_{l'} + \rho_\perp^2 \\
\geq (1 + \mu) \sum_{l=1}^k \rho_l^2 - \mu \sum_{l=1}^k \sum_{l' \neq l}^k |\rho_l \rho_{l'}| \\
\geq (1 + \mu) \left(\sum_{l=1}^k \rho_l^2\right) - \mu \left(\sum_{l=1}^k |\rho_l|\right)^2 \\
\geq (1 + \mu)\|\rho\|_2^2 - \mu\|\rho\|_1^2.
$$
Next, we use the fact that \( \|\rho\|_2^2 \leq \|\rho\|_1^2 \leq k\|\rho\|_2^2 \) to get the bounds \( \|\rho\|_2 \leq \|r\|_2/\sqrt{1 - (k - 1)\mu} \) and \( \|\rho\|_1 \leq \sqrt{k}\|\rho\|_2 \).

Now,

\[
\sum_{l=1}^{k} |r^\top u_l| \leq \sum_{l=1}^{k} \sum_{l'=1}^{k} |\rho_{l'} u_{l'}^\top u_l| \\
\leq (1 - \mu) \sum_{l'=1}^{k} |\rho_{l'}| + \mu \sum_{l=1}^{k} \sum_{l'=1}^{k} |\rho_l| \\
\leq (1 - \mu) \|\rho\|_1 + \mu k\|\rho\|_1 \\
\leq \frac{1 + (k - 1)\mu}{\sqrt{1 - (k - 1)\mu}} \sqrt{k}\|r\|_2.
\]

In the 2-norm case,

\[
\sum_{l=1}^{k} (r^\top u_l)^2 = \sum_{l=1}^{k} \left( \sum_{l'=1}^{k} \rho_{l'} u_{l'}^\top u_l \right)^2 \\
\leq \sum_{l=1}^{k} ((1 - \mu)|\rho_l| + \mu\|\rho\|_1)^2 \\
= (1 - \mu)^2 \sum_{l=1}^{k} \rho_l^2 + 2(1 - \mu)\mu\|\rho\|_1 \sum_{l=1}^{k} |\rho_l| + \mu^2 \sum_{l=1}^{k} \|\rho_l\|_1^2 \\
\leq ((1 - \mu)^2 + 2k(1 - \mu)\mu + k^2\mu^2) \|\rho\|_2^2 \\
\leq (1 - \mu + k\mu)^2 \|\rho\|_2^2 \\
\leq (1 + (k - 1)\mu)^2 \|\rho\|_2^2 \\
\leq \frac{(1 + (k - 1)\mu)^2}{1 - (k - 1)\mu} \|r\|_2^2.
\]

Let us simplify these expressions when \( \mu \leq 1/2(k - 1) \).

\[
\sum_{l=1}^{k} |r^\top u_l| \leq \frac{1 + (k - 1)\mu}{\sqrt{1 - (k - 1)\mu}} \sqrt{k}\|r\|_2 \\
\leq \frac{1 + (k - 1)\mu}{\sqrt{1/2}} \sqrt{k}\|r\|_2 \\
\leq (1 + (k - 1)\mu) \sqrt{2k}\|r\|_2.
\]

Finally,

\[
\sum_{l=1}^{k} (r^\top u_l)^2 \leq \frac{(1 + (k - 1)\mu)^2}{1 - (k - 1)\mu} \|r\|_2^2 \\
= \left( 1 + \frac{(1 + (k - 1)\mu)^2 - (1 - (k - 1)\mu)}{1 - (k - 1)\mu} \right) \|r\|_2^2 \\
= \left( 1 + \frac{3(k - 1)\mu + (k - 1)^2\mu^2}{1 - (k - 1)\mu} \right) \|r\|_2^2 \\
\leq \left( 1 + \frac{3(k - 1)\mu + \frac{1}{2}(k - 1)^2\mu}{1/2} \right) \|r\|_2^2 \\
\leq (1 + 7(k - 1)\mu) \|r\|_2^2.
\]
In this section, we present several concentration results that are key to our results. The $\chi^2$ tail bounds presented in Laurent and Massart [34] play a key role and are reproduced below.

**Lemma 12 ($\chi^2_k$ tail inequality).** Let $q \sim \chi^2_k$ be distributed as a chi-squared variable with $k$ degrees of freedom. Then, for any $t > 0$,

$$
P(q - k > 2\sqrt{kt} + 2t) \leq e^{-t}
$$

$$
P(k - q > 2\sqrt{kt}) \leq e^{-t}.
$$

Alternatively, we have that with probability at least $1 - \delta$,

$$
q \geq k \left( 1 - \frac{2\log(1/\delta)}{\sqrt{k}} \right),
$$

(16)

and similarly, with probability at least $1 - \delta$,

$$
q \leq k \left( 1 + 2\sqrt{\frac{\log(1/\delta)}{k}} + \frac{2\log(1/\delta)}{k} \right).
$$

(17)

*Proof.* See Laurent and Massart [34, Lemma 1].

**Lemma 13 (Gaussian quadratic forms).** Let $x \sim \mathcal{N}(0, I) \in \mathbb{R}^d$ be a random Gaussian vector. If $A$ is symmetric, $x^\top Ax$ is distributed as the sum of $d$ independent $\chi^2$ variables, $\sum_{i=1}^d \lambda_i(A)\chi^2_1$, where $\lambda_i$ are the eigenvalues of $A$.

*Proof.* Let $A = \sum_{i=1}^d \lambda_i u_i u_i^\top$ be the eigendecomposition of $A$. Then, $x^\top Ax = \sum_{i=1}^d \lambda_i ||u_i^\top x||^2$. However, $u_i^\top x_i$ is distributed as independent $\chi^2_1$ random variables. Thus, $x^\top Ax = \sum_{i=1}^d \lambda_i \chi^2_1$. □

**Lemma 14 (Gaussian products).** Let $x_i \sim \mathcal{N}(0, I) \in \mathbb{R}^d$ for $i = 1, \ldots, L$ be random Gaussian vectors. Let $L \geq 4\log(1/\delta)$. Then,

1. $\sum_{i=1}^L (x_i a)^2$ where $a \in \mathbb{R}^d$ is distributed as $||a||_2^2 \chi^2_L$. Consequently, with probability at least $1 - \delta$,

$$
\frac{1}{L} \sum_{i=1}^L (x_i a)^2 \leq ||a||_2^2 \left( 1 + 2\sqrt{\frac{\log(1/\delta)}{L}} + \frac{2\log(1/\delta)}{L} \right)
$$

$$
\leq ||a||_2^2 \left( 1 - 2\log(1/\delta) \frac{1}{\sqrt{L}} \right).
$$

2. $\sum_{i=1}^L x_i a^\top b; a, b \in \mathbb{R}^d$ and $a \neq b$ is sharply concentrated around $a^\top b$: with probability at least $1 - \delta$,

$$
\frac{1}{L} \sum_{i=1}^L x_i a^\top b \leq a^\top b + ||a||_2 ||b||_2 \left( 2\sqrt{\frac{\log(1/\delta)}{L}} + \frac{2\log(1/\delta)}{L} \right)
$$

$$
\leq a^\top b + ||a||_2 ||b||_2 \left( 3\sqrt{\frac{\log(1/\delta)}{L}} \right).
$$
Proof. The first part follows directly from Lemma 13 and the \( \chi^2 \) tail bound, Lemma 12.

For the second part, let \( A = \frac{ab^T+ba^T}{2} \). Note that \( x_i^T ab^T x_i = x_i^T A x_i \). Then, by Lemma 13, \( x_i^T A x_i = \lambda_1 \chi_1^2 + \lambda_2 \chi_2^2 \), where \( \lambda_1 \) and \( \lambda_2 \) are the eigenvalues of \( A \). Furthermore, because \( A = \frac{ab^T+ba^T}{2} \), one of \( \lambda_1 \) or \( \lambda_2 \) is negative, and the other is positive. Without loss of generality, let \( \lambda_1 > 0 > \lambda_2 \).

Applying the \( \chi^2 \) tail bound, Lemma 12, we get that with probability at least \( 1 - \delta \),

\[
\lambda_1 \chi_1^2 \leq \lambda_1(1 + 2 \sqrt{\frac{\log(2/\delta)}{L}} + 2 \frac{\log(2/\delta)}{L})
\]

\[
|\lambda_2| \chi_2^2 \geq |\lambda_2|(1 - \frac{2 \log(2/\delta)}{\sqrt{L}}).
\]

Applying a union bound, we get,

\[
\frac{1}{L} \sum_{i=1}^L x_i^T ab^T x_i \leq \lambda_1(1 + 2 \sqrt{\frac{\log(2/\delta)}{L}} + 2 \frac{\log(2/\delta)}{L}) + \lambda_2(1 - \frac{2 \log(2/\delta)}{\sqrt{L}})
\]

\[
\leq (\lambda_1 + \lambda_2) + |\lambda_1| \left(2 \sqrt{\frac{\log(2/\delta)}{L}} + \frac{2 \log(2/\delta)}{L}\right) + |\lambda_2| \frac{2 \log(2/\delta)}{\sqrt{L}}
\]

\[
\leq (\lambda_1 + \lambda_2) + (|\lambda_1| + |\lambda_2|) \left(2 \sqrt{\frac{\log(2/\delta)}{L}} + \frac{2 \log(2/\delta)}{L}\right).
\]

Observe that \( \lambda_1 + \lambda_2 = \text{tr}(A) = a^T b \). Similarly, \( |\lambda_1| + |\lambda_2| = \|A\|_* = 2(\frac{1}{2}\|a\|_2\|b\|_2) \). Thus, we finally have that with probability at least \( 1 - \delta \),

\[
\frac{1}{L} \sum_{i=1}^L x_i^T ab^T x_i \leq a^T b + \|a\|_2\|b\|_2 \left(2 \sqrt{\frac{\log(2/\delta)}{L}} + \frac{2 \log(2/\delta)}{L}\right).
\]

\[
\square
\]

G Perturbation bounds for joint diagonalization

In this section, we present minor extensions to the perturbation bounds of Cardoso [28] and Afsari [24] so that they apply in the low-rank setting.

**Notation** Let \( M_l = U_l \Lambda_l U_l^T + \epsilon R_l \) for \( l = 1, 2, \ldots, L \) be a set of \( d \times d \) matrices to be jointly diagonalized. \( \Lambda_l \in \mathbb{R}^{k \times k} \) is a diagonal matrix, \( R_l \in \mathbb{R}^{d \times d} \) is an arbitrary unit operator norm matrix and \( \epsilon \) is a scalar. In the orthogonal setting, \( U \in \mathbb{R}^{d \times k} \) is orthogonal, while in the non-orthogonal setting \( U \in \mathbb{R}^{d \times k} \) is an arbitrary matrix with unit operator norm. Let \( \lambda_{il} \triangleq \Lambda_{il} \) be the \( i \)-th factor weight of matrix \( M_l \). Finally, we say that a set of matrices \( \{M_1, \cdots, M_L\} \), \( M_l = \sum_{i=1}^d \lambda_{il} u_i v_i^T \) has joint rank \( k \) if \( \left|\{i \mid \sum_{l=1}^L \lambda_{il} > 0\}\right| = k \).

**Lemma 15** (Cardoso [28]). Let \( M_l = U_l \Lambda_l U_l^T + \epsilon R_l \), \( l \in [L] \), be matrices with common factors \( U \in \mathbb{R}^{d \times k} \) and diagonal \( \Lambda_l \in \mathbb{R}^{k \times k} \). Let \( U \in \mathbb{R}^{d \times d} \) be a full-rank extension of \( U \) with columns \( u_1, u_2, \ldots, u_d \) and let \( \tilde{U} \in \mathbb{R}^{d \times d} \) be the orthogonal minimizer of the joint diagonalization objective \( F(\cdot) \). Then, for all \( u_j, j \in [k] \), there exists a column \( \tilde{u}_j \) of \( \tilde{U} \) such that

\[
\|\tilde{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1}^d E_{ij}^2} + o(\epsilon), \tag{18}
\]

where \( E \in \mathbb{R}^{d \times k} \) is

\[
E_{ij} \triangleq \frac{\sum_{l=1}^L (\lambda_{il} - \lambda_{jl})u_i^T R_l u_i}{\sum_{l=1}^L (\lambda_{il} - \lambda_{jl})^2} \tag{19}
\]

when \( i \neq j \) and \( i \leq k \) or \( j \leq k \). We define \( E_{ij} = 0 \) when \( i = j \) and \( \lambda_{il} = 0 \) when \( i > k \).
Proof. See Cardoso [28, Proposition 1]. Note that in the low rank setting, the entries of $E_{ij}$ (Cardoso [28, Equation 15]) where $i, j > k$ are not defined, however, these terms only effect the last $d - k$ columns of $\bar{U}$. The bounds for vectors $u_1, \ldots, u_d$ only depend on $E_{ij}$ where $i \in [d]$ and $j \in [k]$, and these are derived in the low-rank setting in the same way as they are derived in the full-rank proof of Cardoso [28].

We now present the corresponding perturbation bounds in Afsari [24] to the low rank setting.

**Lemma 16** (Afsari [24]). Let $M_l = U A_l U^T + c R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $A_l \in \mathbb{R}^{k \times k}$. Let $\bar{U} \in \mathbb{R}^{d \times d}$ be a full-rank extension of $U$ with columns $u_1, u_2, \ldots, u_d$ and let $\bar{V} = \bar{U}^{-1}$, with rows $v_1, v_2, \ldots, v_d$. Let $\bar{V} \in \mathbb{R}^{d \times d}$ be the minimizer of the joint diagonalization objective $F(\cdot)$ and let $\bar{U} = \bar{V}^{-1}$.

Then, for all $u_j, j \in [k]$, there exists a column $\bar{u}_j$ of $\bar{U}$ such that

$$
\|\bar{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon),
$$

where the entries of $E \in \mathbb{R}^{d \times k}$ satisfy the equation

$$
\begin{bmatrix}
E_{ij} \\
E_{ji}
\end{bmatrix} = \frac{-1}{\gamma_{ij}(1 - \rho_{ij})} \begin{bmatrix}
\eta_{ij} & -\rho_{ij} \\
-\rho_{ij} & \eta_{ij}
\end{bmatrix} \begin{bmatrix}
T_{ij} \\
T_{ji}
\end{bmatrix},
$$

when $i \neq j$ and either $i \leq k$ or $j \leq k$. When $i = j$, $E_{ij} = 0$. The matrix $T$ has zero on-diagonal elements, and is defined as

$$
T_{ij} = \sum_{l} v_i^T R_l v_j \lambda_{lj},
$$

for $1 \leq j \neq i \leq d$

and the other parameters are

$$
\gamma_{ij} = \|\lambda_i\|_2 \|\lambda_j\|_2, \quad \eta_{ij} = \frac{\|\lambda_i\|_2}{\|\lambda_j\|_2}, \quad \rho_{ij} = \frac{\lambda_i^T \lambda_j}{\|\lambda_i\|_2 \|\lambda_j\|_2}, \quad (\lambda_i)_k = \lambda_{ik}.
$$

We define $\lambda_{il} = 0$ when $i > k$.

**Proof.** In Afsari [24, Theorem 3] it is shown that $\bar{V} = (I + \epsilon E)V + o(\epsilon)$, where $E_{ij}$ is defined for $i, j \in [d]$ (Afsari [24, Equation 36]). Then,

$$
\bar{U} = \bar{U}(I + \epsilon E)^{-1} + o(\epsilon)
$$

$$
= \bar{U}(I - \epsilon E) + o(\epsilon).
$$

Note that, once again, in the low rank setting, the entries of $E_{ij}$ when $i, j > k$ are not characterized by Afsari’s results; however, these terms only effect the last $d - k$ columns of $\bar{U}$.

**Lemma 17.** Let $M_l = U A_l U^T + c R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $A_l \in \mathbb{R}^{k \times k}$. Let $\bar{U} \in \mathbb{R}^{d \times d}$ be a full-rank extension of $U$ with columns $u_1, u_2, \ldots, u_d$ and let $\bar{V} = \bar{U}^{-1}$, with rows $v_1, v_2, \ldots, v_d$. Let $\bar{V} \in \mathbb{R}^{d \times d}$ be the minimizer of the joint diagonalization objective $F(\cdot)$ and let $\bar{U} = \bar{V}^{-1}$.

Then, for all $u_j, j \in [k]$, there exists a column $\bar{u}_j$ of $\bar{U}$ such that

$$
\|\bar{u}_j - u_j\|_2 \leq \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon),
$$

where the entries of $E \in \mathbb{R}^{d \times k}$ are bounded by

$$
|E_{ij}| \leq \frac{1}{1 - \rho_{ij}^2} \left( \frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2} \right) \left( \sum_{l=1}^{L} v_i^T R_l v_j \lambda_{lj} \right) + \sum_{l=1}^{L} v_i^T R_l v_j \lambda_{il},
$$

when $i \neq j$ and $E_{ij} = 0$ when $i = j$ and $\lambda_{il} = 0$ when $i > k$. Here $\lambda_l = (\lambda_{l1}, \lambda_{l2}, \ldots, \lambda_{Ll}) \in \mathbb{R}^L$ and $\rho_{ij} = \frac{\lambda_i^T \lambda_j}{\|\lambda_i\|_2 \|\lambda_j\|_2}$ is the modulus of uniqueness, a measure of how ill-conditioned the problem is.
Proof. From Lemma 16, we have that
\[
\left\| \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} \right\| \leq \frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}(1 - \rho_{ij}^2)} \left\| \begin{bmatrix} T_{ij} \\ T_{ji} \end{bmatrix} \right\|
\]
where
\[
\gamma_{ij} = \|\lambda_i\|_2 \|\lambda_j\|_2, \quad \eta_{ij} = \frac{\|\lambda_i\|_2}{\|\lambda_j\|_2}, \quad \rho_{ij} = \frac{\lambda_i^\top \lambda_j}{\|\lambda_j\|_2 \|\lambda_i\|_2},
\]
and the matrix \(T\) is defined to be zero on the diagonal and for \(i \neq j\) defined as
\[
T_{ij} = \sum_{l=1}^L v_i^\top R_l v_j \lambda_{jl}, \quad \text{for } 1 \leq j \neq i \leq d
\]
Taking \(\|\cdot\|\) to be the \(l_1\)-norm in the above expression, we have that
\[
|E_{ij}| \leq |E_{ij}| + |E_{ji}| \leq \frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}(1 - \rho_{ij}^2)} \left( |T_{ij}| + |T_{ji}| \right).
\]
Since
\[
\frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}} = \frac{\|\lambda_i\|_2^2 + \|\lambda_j\|_2^2}{\|\lambda_i\|_2^2 \|\lambda_j\|_2^2} = \frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2}
\]
and
\[
T_{ij} = \sum_{l=1}^L v_i^\top R_l v_j \lambda_{jl},
\]
the claim follows. \(\square\)