Almost-lossless compression of a low-rank random tensor

Minh Thanh Vu*

Abstract. In this work, we establish an asymptotic limit of almost-lossless compression of a random, finite alphabet tensor which admits a low-rank canonical polyadic decomposition.

Key words. Information theory, Tensor compression, Tensor decomposition, Information spectrum method, Kruskal’s uniqueness condition.

MSC codes. 68P30, 15A69

1. Introduction. Tensors, or multiway arrays, have been emerging as powerful tools with applications in signal processing, computer vision, and big-data analytics [2, 7]. In these applications tensors of data are approximated using products of lower dimensional structures like factor matrices, core tensors, etc. Different models have been proposed for the approximation purpose such as canonical polyadic decomposition (CPD) [1, 5], Tucker’s decomposition [14], etc. In addition to being an effective analytic tool, tensors also arise naturally from data sources such as color images, videos, etc.

An interesting research question would be how to compress and store multiway arrays of data efficiently. In this paper we put forward this research question by studying the asymptotic limit of tensor compression. For a tractable analysis, we assume in our study that the tensor can be factorized exactly as a sum of multiple rank-one components or a CPD model. This simple assumption is justifiable in practice since many tensors can be well-approximated by a few components. We then assume a probabilistic model for each factor matrix in which elements are drawn from a distribution on a finite alphabet.

In information theory, data are usually modeled in forms of vectors or sequences. Our model hence can be viewed as a generalization of previous data compression models. An achievable tensor compression-reconstruction scheme can be designed based on the conventional typical arguments. Due to the inherent dependence among elements inside a tensor, we need to use novel arguments for the converse proof. Since each factor matrix is randomly generated from a finite alphabet, it is full rank with high probability. We then use the connection between full rank factor matrices and the essential uniqueness of CP decomposition to establish the converse proof.

Our paper is organized as follows. In Section 2 we present assumptions used in our study and recap some definitions from multilinear algebra and information theory. We establish the asymptotic limit of compressing a random rank-one tensor in Section 3. Then we provide two examples which highlight challenges of multiple-component scenarios as well as propose a way to tackle these problems in Section 4. Finally, based on the observations in Section 4 we establish the asymptotic limit of almost-lossless compression of a random low-rank tensor in Section 5.

2. Preliminaries. A tensor $\mathbf{T} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a multiway array indexed by a tuple $(i_1, \ldots, i_N)$ where $i_j \in [1 : I_j]$. The parameter $N$ is called the order of the tensor. We say that $\mathbf{T}$ is a single component tensor if there exist vectors $\mathbf{a}_i \in \mathbb{R}^{I_i \times 1}$, $i \in [1 : N]$, such that

$$T_{i_1, \ldots, i_N} = a_{i_1} \cdots a_{i_N}.$$ 

In other words, $\mathbf{T}$ is the outer product of $\mathbf{a}_i$, and we write $\mathbf{T} = a_1 \circ \cdots \circ a_N$. For such a tensor we say that its rank is one if it is unequal to the all 0 tensor. Assume that $\mathbf{T}$ can be factorized into $R$
single component tensors $T_i$, i.e.,

$$T = \sum_{i=1}^{R} T_i = \sum_{i=1}^{R} a_{i1} \circ \cdots \circ a_{iN}.$$  

Note that the arithmetic in our study is carried out on $\mathbb{R}$. We also write $T = [X_1; \ldots ; X_N]$ where each factor matrix is given by $X_i = [a_{i1}, \ldots , a_{iR}]$. The minimum of such $R$ in such factorizations of $T$ is called the rank of $T$. Determining the rank of a tensor is challenging, in fact NP-hard when $N \geq 3$ [6]. In this paper we assume that each realization of our random tensor $T$ can be decompositions as in (2.1) with $R$ single components, i.e., it does not necessarily mean that $R$ is the rank of the corresponding tensor. In our study we further assume that the parameter $R$ fixed. To put our study into the information theoretic framework a probabilistic postulate of our data needs to be made. For this purpose we assume that factor matrices are independent unless otherwise stated. In each factor matrix $X_i, i \in [1 : N]$, the $(j,r)$-th entry $X_{i,jr}$ is drawn from a distribution $P_{X_i,r}$ on an alphabet $X_i, X_{i,jr} \sim P_{X_i,r}$, for all $j \in [1 : I_i]$ and $r \in [1 : R]$. We further assume that the alphabets $X_i$ are finite and $\max_{x \in X_i} P_{X_i,r}(x) < 1$. With this our main focus would be the set of tensors taking values in the alphabet $T \triangleq X_1 \times \cdots \times X_N$. For simplicity we only consider the case that all sizes are equal to each other

$$I_1 = \cdots = I_N = n, n \rightarrow \infty.$$ 

Now we are ready to define information theoretic quantities of interest.

**Definition 2.1.** For a given $n$, a tensor compression-reconstruction scheme consists of two mappings:

- a compression mapping $\phi_n : T \rightarrow M$ which maps a tensor $T$ to an index $m \in M$, namely $\phi_n(T) = m$, which is stored in a storage medium such as a hard disk,
- and a reconstruction mapping $\psi_n : M \rightarrow T$ which outputs a tensor $\hat{T}$ in the alphabet $T$ from the compressed index $m$, namely $\hat{T} = \psi_n(m)$.

**Definition 2.2.** A compression threshold $C$ is almost-losslessly achievable if there exists a sequence of tensor compression-reconstruction schemes $(\phi_n, \psi_n)$ satisfying

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log |M| \leq C, \lim_{n \rightarrow \infty} \Pr\{\hat{T} \neq T\} \rightarrow 0,$$

where $|M|$ is the cardinality of $M$. We define $C^*_1$ to be the infimum of all almost-losslessly achievable thresholds $C$.

Conventionally we would be interested in the compression rate [3] defined as $\frac{1}{\text{input size}} \log |M|$. In our setting the input size is given by $n^N$ and the compression alphabet size is upper bounded by $(\prod_i |X_i|)^nR$. Therefore the compression rate is zero when $N \geq 2$. However, in practice we are interested in the amount of information that we need to store rather than the rate alone. Therefore the quantity compression threshold is appropriate in this case. Finally to characterize the minimum compression threshold we need to use the entropy. For a distribution $P$ on a finite alphabet $X$ the entropy $H(P)$ is defined as $H(P) = \sum_{x \in X} -P(x) \log P(x)$.

**Remark 2.3.** In practice, performing exact CP decomposition of $T$, is a challenging problem due to non-linearity. Therefore it is difficult to choose the exact alphabets $X_i$ for modeling. For some data sources such as videos or images, as pixels take values between 0 and 255, we can assume that entries of the tensors $T$ take values on a finite alphabet. Hence the distribution of $T$ is a discrete one. It is then natural to model $X_i$ to be discrete albeit not easy to select.

If we assume that $T$ comes from a continuous distribution then it is also often assumed that entries of factors take values on $\mathbb{R}$. Hence our assumption that $X_i$ is finite, can be seen as a
quantization argument. This naturally introduces distortion into our formulation. Certainly, we want to obtain a compression-distortion trade off. Our study however indicates that obtaining this goal might be formidable.

Despite the above shortcomings our model is practically useful in the following sense. Suppose that a universal (almost lossless) compression algorithm is designed for tensors. Our model can be used as an additional performance benchmark. For example tensors of size $n^N$ can be artificially generated according to our model and provided as the input for the algorithm. If the compression threshold obtained by the algorithm is far from the minimum threshold then there is still room for improvement. It should be noted that algorithms should not be designed specifically for our model due to its simplified assumptions.

**Remark 2.4.** Although it is not the focus of our study, let us consider the case that for all $i \in [1 : N]$ except one $I_i$ are constant while the last dimension grows. This models the case where one single dimension is very large while the others are very small, e.g., in a long video recording session. Tensor decomposition reduces the number of parameters that need to be stored from $\prod_{i=1}^N I_i$ to $\sum_i I_i$, which is not very satisfying. One should not stop there and instead look for new methods that compress the tensor further. We can adapt our theory to this problem but it would be more complicated. Even when $I_i$ are large for all $i \in [1 : N]$, it is important to compress the tensor further since the data might be distributed to a large number of users.

### 3. Rank-one compression.

We begin our study by analyzing the simplest case of rank-one tensor compression. In this we explain our main information theoretic idea for unfamiliar audiences. This section hence serves as a warming up to more complex scenarios in latter sections.

Recall that our random tensor $T$ can be written as

$$T_{i_1, \ldots, i_N} = a_{i_1} \cdots a_{i_N}. \tag{3.1}$$

The following result characterizes the fundamental compression threshold for the single-component, i.e., rank-one, tensor scenario.

**Theorem 3.1.** When $T$ is a single-component tensor, then the minimum almost-lossless compression threshold is given by

$$C^*_t = \sum_{i=1}^N H(P_{X_i}).$$

**Proof.** First we show that for any $\eta > 0$, $\sum_{i=1}^N H(P_{X_i}) + \eta$ is an achievable compression threshold, i.e., $C^*_t \leq \sum_{i=1}^N H(P_{X_i})$ holds.

In information theory to show that a sequence of mappings $(\phi_n, \psi_n)$ exists, one usually uses a typicality argument. Namely we only need to consider a small set of tensors $T$ which takes most of the probability. Our setting is a non-iid setting, as there are correlations between elements of $T$. We therefore need some more work than conventional approaches in $[4, 3]$. Given a $\gamma > 0$, for each $i \in [1 : N]$ we define a typical set

$$\mathcal{A}^n_{i, \gamma} = \{ a_i \mid - \log P^n_{X_i}(a_i) - nH(P_{X_i}) < n\gamma \}.$$

For each $i \in [1 : N]$, to store an element inside $\mathcal{A}^n_{i, \gamma}$ one needs at most $n(H(P_{X_i}) + \gamma)$ nats. By $[3$, Theorem 3.1.2] we also have

$$P^n_{X_i}(\mathcal{A}^n_{i, \gamma}) \geq (1 - \gamma), \text{ for all sufficiently large } n.$$  

We define the Cartesian product set $\mathcal{S}_{\gamma, n} = \times_{i=1}^N \mathcal{A}^n_{i, \gamma}$. Then the typical set used for compression in our setting is defined as

$$\mathcal{T}_{\gamma, n} = \{ a_1 \circ \cdots \circ a_N \mid (a_1, \ldots, a_N) \in \mathcal{S}_{\gamma, n} \}.$$
We observe
\[
\Pr\{ \mathbf{T} \in \mathcal{T}_{\gamma,n} \} \geq \sum_{(a_1, \ldots, a_N) \in \mathcal{S}_{\gamma,n}} \prod_{i=1}^{N} P_{X_i}^{n}(a_i) = \prod_{i=1}^{N} P_{X_i}^{n}(A_{\gamma,n}^{n}) \\
\geq (1 - \gamma)^N.
\]
(3.2)

The first inequality follows since there might exist tuples \((a_1, \ldots, a_N) \in \mathcal{S}_{\gamma,n}\) and \((a_1', \ldots, a_N') \notin \mathcal{S}_{\gamma,n}\) such that
\[
a_1 \circ \cdots \circ a_N = a_1' \circ \cdots \circ a_N'.
\]

Our compression scheme is given as follows. \(\phi_n\) maps each tensor in \(\mathcal{T}_{\gamma,n}\) to a unique index. The reconstruction mapping \(\psi_n\) maps the stored index to the corresponding tensor. For tensors \(\mathbf{T}\) that do not belong to \(\mathcal{T}_{\gamma,n}\), \(\phi_n\) maps these to a fixed index and \(\psi_n\) outputs a fixed tensor. This requires at most \(n(\sum_{i=1}^{N} H(P_{X_i}) + N\gamma)\) nats. Therefore for any given \(\eta > 0\), there exists a \(n_0(\eta)\) such that for all sufficiently large \(n \geq n_0(\eta)\) our tensor compression-reconstruction scheme satisfies
\[
(3.3) \quad \Pr\{ \mathbf{T} \neq \mathbf{T} \} \leq \eta, \quad \frac{1}{n} \log |\mathcal{M}| \leq \sum_{i=1}^{N} H(P_{X_i}) + \eta.
\]

In conclusion we have \(C_t^* \leq \sum_{i=1}^{N} H(P_{X_i})\).

Now we show the reverse direction, i.e., \(C_t^* \geq \sum_{i=1}^{N} H(P_{X_i})\). For a given achievable compression threshold \(C\) let \(\{(\phi_n, \psi_n)\}_{n=1}^{\infty}\) be a given sequence of tensor compression-reconstruction schemes satisfying
\[
\limsup_{n \to \infty} \frac{1}{n} \log |\mathcal{M}| \leq C, \quad \lim_{n \to \infty} \Pr\{ \psi_n(\phi_n(\mathbf{T})) \neq \mathbf{T} \} \to 0.
\]

Then for a given \(\epsilon > 0\) there exists a \(n_\epsilon\) such that for all \(n \geq n_\epsilon\) we have
\[
\log |\mathcal{M}| \leq n(C + \epsilon), \quad \text{and} \quad \Pr\{ \psi_n(\phi_n(\mathbf{T})) \neq \mathbf{T} \} \leq \epsilon.
\]

For a given \(n\) we define the following correct decodable set of tensors
\[
\mathcal{D}_n = \{ \mathbf{T} \mid \psi_n(\phi_n(\mathbf{T})) = \mathbf{T} \}.
\]

We define the expanding set of factors
\[
\mathcal{S}_n = \{ (a_1, \ldots, a_N) \mid a_1 \circ \cdots \circ a_N \in \mathcal{D}_n \}.
\]

We have
\[
(3.4) \quad \Pr\{ (X_1, \ldots, X_N) \in \mathcal{S}_n \} = \Pr\{ \mathbf{T} \in \mathcal{D}_n \}.
\]

We observe that for a given tensor \(\mathbf{T}\), once \(a_i\) are given for all but one \(i \in [1 : N]\) we can deduce the remaining elements in vectors \(\{a_j\}_{j=1}^{N}\). Therefore we have \(|\mathcal{S}_n| \leq \min_i \prod_{j \neq i} |A_j| |\mathcal{D}_n|\). Additionally since \(\phi_n\) can only take \(|\mathcal{M}|\) values so does the composite mapping \(\psi_n(\phi_n(\cdot))\), which implies that \(|\mathcal{D}_n| \leq |\mathcal{M}|\). In conclusion we have \(|\mathcal{S}_n| \leq \min_i \prod_{j \neq i} |A_j| |\mathcal{M}|\).

Since the distribution of \(\mathbf{T}\) is no longer a product of identical components in our case, standard arguments using Fano’s inequality as in [3] are no longer applicable. In order to show the converse we define the following atypical set, parameterized by \(n\) and \(\eta\),
\[
(3.5) \quad \mathcal{T}_{n,\eta} = \{ (a_1, \ldots, a_N) \mid -\sum_{i=1}^{N} \log P_{X_i}^{n}(a_i) \geq \log |\mathcal{M}| + \eta \}.
\]
For all sufficiently large \( n \geq n_\epsilon \), using the information-spectrum arguments \([4]\) we then have

\[
\Pr\{(X_1, \ldots, X_N) \in T_{n, \eta}\} = \Pr\{(X_1, \ldots, X_N) \in S_n \cap T_{n, \eta}\} + \Pr\{(X_1, \ldots, X_N) \in S_n^c \cap T_{n, \eta}\} \\
\leq \Pr\{(X_1, \ldots, X_N) \in S_n^c\} + \Pr\{(X_1, \ldots, X_N) \in S_n \cap T_{n, \eta}\} \\
\overset{(3.4)}{=} \Pr\{\psi_n(\phi_n(T)) \neq T\} + \sum_{(a_1, \ldots, a_N) \in S_n \cap T_{n, \eta}} \prod_{i=1}^{N} P_{X_i}^n(a_i) \\
\overset{(3.5)}{\leq} \epsilon + |S_n \cap T_{n, \eta}| e^{-\eta} |M| \leq \epsilon + |S_n| e^{-\eta} |M| \\
\leq \epsilon + e^{-\eta} \min_i |X_i|.
\]

We take \( \eta = n \gamma \). If for all \( n_0 \geq n_\epsilon \) there exists a \( n > n_0 \) such that \( \log |M| + \eta n < n(\sum_i H(P_{X_i}) - N \gamma) \), then we have

\[
\lim_{n \to \infty} \sup \Pr\{(X_1, \ldots, X_N) \in T_{n, \eta}\} = 1,
\]
due to the weak law of large numbers, which violates that last inequality. We must have \( \log |M| + n \gamma \geq n(\sum_i H(P_{X_i}) - N \gamma) \), which in turn implies that \( n(C + \epsilon + \gamma) \geq n(\sum_i H(P_{X_i}) - N \gamma) \), for all \( n \geq n_\epsilon \). Since \( \gamma \) and \( \epsilon \) are arbitrary we have \( C^* \geq \sum_i H(P_{X_i}) \).

**Remark 3.2.** In the converse direction, we have carefully controlled the contribution of atypical tuples of factors, those who are in \( T_{n, \eta} \). In the single-component scenario, as we have seen, for each tensor the number of such tuples of factors is always bounded by a constant. This no longer holds in the multi-component scenarios.

**Remark 3.3.** Due to our assumption on finiteness of \( \{X_i\}_{i=1}^{N} \) and the relation (3.1), performing tensor decomposition of the rank-one model is relatively easy. The output of a decomposition rule \( \{a_i\}_{i=1}^{N} \) is however not necessarily the same as the underlying \( \{a_i\}_{i=1}^{N} \). Straightforward usage of compression schemes designed for \( P_{X_i} \) on \( a_i \) is not recommended due to a distribution mismatch. In other words, a two-step algorithm involving a tensor decomposition in the first step and compression of factors in the second step might not be optimal. Our scheme indicates that to ensure optimality at least some additional constraints, such as typicality in our model, need to be imposed on top of the tensor decomposition.

### 4. Examples of 2-component compression.

In this section we present some examples involving random tensors admitting two-component decompositions. In these examples we show that there exist tensors for which the number of tuples of factors increases exponentially with \( n \). This implies that a straightforward application of previous arguments in Theorem 3.1 is no longer possible. However the probability of the set of these tensors are negligible. Hence the converse arguments can be fixed.

**Example 1:** In the first example we consider a supersymmetric scenario where the order of the tensor is 3, \( N = 3 \), and

\[
X_1 = X_2 = X_3 = [a_1, a_2] \in \mathcal{X}^{n \times 2}.
\]

We further assume that the underlying alphabet \( \mathcal{X} \) is \( \mathcal{X} = \{-1, 1\} \). Assume that \( a_{1i} \sim P \) and \( a_{2i} \sim Q \) where \( P \) and \( Q \) are two distributions on \( \mathcal{X} \) satisfying \( P(a) \neq 0 \) and \( Q(a) \neq 0 \) for all \( a \in \mathcal{X} \). To design a compression mapping one only needs to look at the set

\[
\mathcal{A}_n^0 = \{X \mid || - \log [P^n(a_1)Q^n(a_2)] - n[H(P) + H(Q)]|| < n\gamma\}.
\]

We similarly form a set

\[
T_{n, \gamma} = \{[X; X; X] \mid X \in \mathcal{A}_n^0\},
\]

and index all of its elements. If \( T \) belongs to \( T_{n, \gamma} \) we store the corresponding index. Otherwise we store a special index for all tensors that are not in \( T_{n, \gamma} \). This requires \( n[H(P) + H(Q) + \gamma] \) nats.
In the converse direction given a realization $T$ we bound the number of pairs of factors $(a_1, a_2)$ resulting in $T$. Each element of the tensor $T$ is given by

$$T_{i_1i_2i_3} = a_{1i_1}a_{1i_2}a_{1i_3} + a_{2i_1}a_{2i_2}a_{2i_3}.$$ 

Since the alphabet is $\{-1, 1\}$ we have $a^3 = a$ for any $a \in \mathcal{X}$. To recover $(a_1, a_2)$ from $T$ we therefore only need to consider $n + \binom{n}{2}$ expressions of the forms $a_{1i} + a_{2i} = T_{i}$ where $i \in [1 : n]$, and $a_{1i_1}a_{1i_2}a_{1i_3} + a_{2i_1}a_{2i_2}a_{2i_3} = T_{i_1i_2i_3}$ where $(i_1, i_2, i_3) \in [1 : n]^3$ are mutually different.

- Let us consider the first case when $a_{1i} + a_{2i} = 0$ for all $i \in [1 : n]$. This implies that $a_{1i} = -a_{2i}$ for all $i \in [1 : n]$. For any triple $(i_1, i_2, i_3)$ we then have $a_{1i_1}a_{1i_2}a_{1i_3} = -a_{2i_1}a_{2i_2}a_{2i_3}$ which leads to $a_{1i_1}a_{1i_2}a_{1i_3} + a_{2i_1}a_{2i_2}a_{2i_3} = 0$. Therefore when $T = 0$, the number of decompositions is controlled by the system of equations $\{a_{1i} + a_{2i} = 0\}_{i=1}^n$. The number of pairs of factors that result in this particular tensor is hence $2^n$. In more detail each solution factor matrix has the form

$$X = [a, -a], \ a \in \mathcal{X}^n,$$

i.e., a rank-deficient matrix. Next, we will calculate the probability of the event $T = 0$. Then we have

$$\Pr\{T = 0\} = \sum_{a \in \mathcal{X}^n} P^n(a)Q^n(-a) = \left[ \sum_{a \in \mathcal{X}} P(a)Q(a) \right]^n.$$ 

As

$$\sum_{a \in \mathcal{X}} P(x)Q(-x) = P(1)Q(-1) + P(-1)Q(1) < \max\{Q(-1), Q(1)\} < 1$$

we obtain that $\Pr\{T = 0\} \to 0$ as $n \to \infty$.

- Consider a tensor $T$ for which we have, without the loss of generality,

$$a_{1n} + a_{2n} = 2,$$

and $a_{1i} + a_{2i} = 0$, $\forall i \in [1 : n - 1]$.

Then for all $i \in [1 : n - 1]$ since $a_{1i} = -a_{2i}$ holds we have

$$T_{1i} = a_{11}a_{1i}a_{1n} + a_{21}a_{2i}a_{2n} = 2a_{11}a_{1i}. $$

This implies that for a given choice of $a_{11}$ and a given $T$ we can infer the other values of $a_{1i}$ and $a_{2i}$ uniquely. The number of pairs of factors for a given $T$ in this case is 2. Having more constraints of the form $a_{1i} + a_{2i} \neq 0$ does not increase the number of decompositions for a similar reason.

- When $a_{1i} + a_{2i} \neq 0$ for all $i \in [1 : n]$ then there is a unique pair of factors $(a_1, a_2)$ resulting in $T$.

By excluding the all 0 tensor, $T = 0$, from the decodable set $\mathcal{D}_n$, since the event has a vanishing probability, we can apply a similar argument as in the converse proof of Theorem 3.1, for example the atypical set $\mathcal{T}_{n, \eta}$ can be defined accordingly as

$$\mathcal{T}_{n, \eta} = \{ (a_1, a_2) \ | -\log P^n(a_1) + \log Q^n(a_2) \geq \log |\mathcal{M}| + \eta \},$$

to conclude that in this case the minimum compression threshold is given by $C^*_t = H(P) + H(Q)$.

**Example 2:** Let us consider the compression of an order-2 tensor admitting a two single component decomposition

$$T = X_1X_2^T = xy^T + uv^T,$$

where $(\cdot)^T$ is the transpose operation and for notation brevity we have abbreviated $X_1 = [x, u]$ as well as $X_2 = [y, v]$. We assume that all random variables take values in the set $\{-1, 1\}$ as well.
Presenting the entire details is rather complex and unnecessary. We consider two representative scenarios occurring when \( k = n \). Let \( m \) the number of rows indexed by \( \{j_1, \ldots, j_m\} \) such that \( T_{ji} \neq 0 \) for all \( i \in [1 : n] \) and \( j \in \{j_1, \ldots, j_m\} \).

- When \( T = 0 \), for a given \( x_1, u_1 \in \{-1, 1\} \) we have

\[
\begin{align*}
x_1 y_1 + u_1 v_1 &= 0 \implies v_i = -u_1 x_1 y_i, \quad \forall i \in [1 : n], \\
x_i y_1 + u_i v_1 &= 0 \implies u_i = x_1 u_1 x_i, \quad \forall i \in [2 : n].
\end{align*}
\]

The factor matrices have the following form \( X_1 = [x, ax] \) and \( X_2 = [y, -ay] \) where \( a \in \{-1, 1\}, x \in \{-1, 1\}^n, y \in \{-1, 1\}^n \). They are rank-deficient. The number of tuples of factors for the all 0 tensor is \( 2 \times 2^n \times 2^n = 2^{2n+1} \). The probability of this event is given by

\[
\Pr\{T = 0\} = \sum_{x,y,a} \sum_{x,y,a} P_X^n(x) P_Y^n(y) P_U^n(ax) P_V^n(-ay)
\]

\[
= \sum_{a} \sum_{x} P_X^n(x) P_Y^n(ax) \sum_{y} P_Y^n(y) P_V^n(-ay)
\]

\[
= \sum_{a} \left[ \sum_{x} P_X^n(x) P_U^n(ax) \right]^{n} \left[ \sum_{y} P_Y^n(y) P_V^n(-ay) \right]^{n}
\]

\[
+ \left[ \sum_{x} P_X^n(x) P_V^n(-x) \right]^{n} \left[ \sum_{y} P_Y^n(y) P_V^n(y) \right]^{n}.
\]

We have

\[
\Pr\{T = 0\} \to 0 \text{ as } n \to \infty.
\]

- When \( m \geq 1 \) holds, then given \((y_1, v_1), (y_2, \ldots, y_n)\) and \((v_2, \ldots, v_n)\) are determined through

\[
x_{j_1} = \sign(T_{j_1}) y_i, \quad u_{j_1} = \sign(T_{j_1}) v_i, \quad \forall i \in [1 : n].
\]

For \( j \notin \{j_1, \ldots, j_m\} \) we can select \( x_j \) freely. Therefore for each of these tensors we can find \( 8 \times 2^{n-1-m} = 2^{n-m+2} \) tuples of factors \((x, y, u, v)\) resulting in it. Furthermore due to the structure we also have \((T_{j_1}, \ldots, T_{j_n}) = \pm(T_{j_1}, \ldots, T_{j_n})\) for all \( i \in [2 : m] \). A representative tuple of factors has the following form

\[
\begin{align*}
x &= (x_1, \beta_1 x_1, \ldots, \beta_m x_1, x_m + 2, \ldots, x_n)^T, \\
u &= (-u_1, \beta_1 u_1, \ldots, \beta_m u_1, -x_1 u_1 x_m + 2, \ldots, -x_1 u_1 x_n)^T, \\
y &= (\alpha_1 x_1, \ldots, \alpha_n x_1)^T, \\
v &= (\alpha_1 u_1, \ldots, \alpha_n u_1)^T.
\end{align*}
\]
where \( \beta_j \in \{-1, 1\} \) for all \( j \in [1 : m] \) and \( \alpha_i \in \{-1, 1\} \) for all \( i \in [1 : n] \). We observe that the second factor matrix \( X_2 = [y, v] \) is also rank-deficient. We calculate the contribution of this event in the following. The probability of the given tuple of factors is

\[
P_X(x_1)P_U((-u_1)\prod_{\beta} P^n_U(x_1\beta)P^n_U(u_1\beta)}
\]

\[
\times P^n_{X}^{-1}(x_{m+2})P^n_{U}^{-1}(-x_1u_1x_{m+2})P^n_U(x_1\alpha)P^n_U(u_1\alpha),
\]

where \( x^n_{m+2} = (x_{m+2}, \ldots, x_n) \). Summing over \( \alpha \) we obtain

\[
\sum_{\alpha} P^n_U(x_1\alpha)P^n_U(u_1\alpha) = \sum_{\alpha} P_Y(x_1\alpha)P_Y(u_1\alpha)]^n.
\]

Summing over \( \beta \) we obtain

\[
\sum_{\beta} P^n_X(x_1\beta)P^n_U(u_1\beta) = \sum_{\beta} P_X(x_1\beta)P_U(u_1\beta)]^m.
\]

Summing over \( x^n_{m+2} \) we obtain

\[
\sum_{x^n_{m+2}} P^n_{X}^{-1}(x_{m+2})P^n_{U}^{-1}(-x_1u_1x_{m+2}) = \sum_{x} P_X(x)P_U(-x_1u_1x)]^{n-m-1}.
\]

Finally by summing over the all possible choices of \( \{j_1, \ldots, j_m\} \), the number of rows \( m \), \( x_1 \) and \( u_1 \), the total probability of this event is hence

\[
\sum_{x_1, u_1} P_X(x_1)P_U((-u_1)\sum_{\alpha} P_Y(x_1\alpha)P_Y(u_1\alpha)]^n
\]

\[
\times \sum_{m=1}^{n-1} \left( \frac{n-1}{m} \right)[\sum_{\beta} P_X(x_1\beta)P_U(u_1\beta)]^m \sum_{x} P_X(x)P_U(-x_1u_1x)]^{n-m-1
\]

\[
= \left[ \sum_{x} P_X(x)P_U(-x)\right][\sum_{\alpha} P_Y(\alpha)P_Y(\alpha)]^n
\]

\[
\times \sum_{m=1}^{n-1} \left( \frac{n-1}{m} \right)[\sum_{\beta} P_X(\beta)P_U(\beta)]^m \sum_{x} P_X(x)P_U(-x)]^{n-m-1}
\]

\[
+ \left[ \sum_{x} P_X(x)P_U(x)\right][\sum_{\alpha} P_Y(\alpha)P_Y(-\alpha)]^n
\]

\[
\times \sum_{m=1}^{n-1} \left( \frac{n-1}{m} \right)[\sum_{\beta} P_X(\beta)P_U(-\beta)]^m \sum_{x} P_X(x)P_U(x)]^{n-m-1
\]

\[
\leq \left[ \sum_{x} P_X(x)P_U(-x)\right][\sum_{\alpha} P_Y(\alpha)P_Y(\alpha)]^n
\]

\[
+ \left[ \sum_{x} P_X(x)P_U(x)\right][\sum_{\alpha} P_Y(\alpha)P_Y(-\alpha)]^n \rightarrow 0.
\]

The last inequality is valid since the following reduction holds

\[
\sum_{m=1}^{n-1} \left( \frac{n-1}{m} \right)[\sum_{\beta} P_X(\beta)P_U(\beta)]^m \sum_{x} P_X(x)P_U(-x)]^{n-m-1
\]

\[
= \left( \sum_{\beta} P_X(\beta)P_U(\beta) + \sum_{x} P_X(x)P_U(-x)\right)^{n-1} - (\sum_{x} P_X(x)P_U(-x)]^{n-1}
\]

\[
= 1 - (\sum_{x} P_X(x)P_U(-x)]^{n-1}.
\]

Using similar lines of arguments we can show that in this case the minimum compression threshold is \( C^*_1 = H(P_X) + H(P_Y) + H(P_U) + H(P_Y). \)
5. Compression of multi-component tensor. Recall that a multi-component tensor has the following form
\[
\mathbf{T} = [\mathbf{X}_1; \cdots; \mathbf{X}_N] = \sum_{r=1}^{R} \mathbf{a}_{r1} \otimes \cdots \otimes \mathbf{a}_{rN},
\]
where \( R \geq 2 \) holds. In the last section we have seen that for some tensors such as the all 0 tensor, the number of tuples of factors grows exponentially with \( n \). We also observe that some factor matrices in these cases are rank-deficient. Therefore, a workaround idea for the general scenario would be restricting our attention to the set of tuples of full rank factor matrices, i.e., all factor matrices in a given tuple are full rank. This does not immediately guarantee that there would not exist a set of tensors \( \mathbf{T} \) admitting a growing number of factorizations which perhaps has non-vanishing probability. Fortunately, our analysis in the following shows that the case does not occur. In the following we first show that a random factor matrix is full rank with high probability.

**Lemma 5.1.** For each \( i \in [1 : N] \) we have
\[
\Pr\{\text{rank}(\mathbf{X}_i) = R\} \to 1, \text{ as } n \to \infty.
\]

When \( P_{X_i,r_1} = P_r \) for all \( r \in [1 : R] \) the result can be deduced from the fact that the probability of a square random matrix with iid elements being singular is vanishing [8]. For our setting we use arguments in [13].

**Proof.** For a given \( i \in [1 : N] \), \( \mathbf{X}_i \) is rank-deficient, i.e., \( \text{rank}(\mathbf{X}_i) < R \), implies that one of the following events happens
\[
\mathcal{A}_{i0} = \{\mathbf{X}_{i1} = \mathbf{0}\}, \quad \mathcal{A}_{ir} = \{\mathbf{X}_{i(r+1)} \in \text{span}(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ir})\}, \quad r = 1, \ldots, R - 1.
\]

First of all we have
\[
\Pr\{\mathcal{A}_{i0}\} = P_{X_i,1}(0)^n < \rho_i^n,
\]
where \( \rho_i = \max_{x,k} P_{X_i,k}(x) < 1 \) by our assumption. The inequality also holds when \( 0 \notin X_i \) occurs. For each \( r = 1, \ldots, R - 1 \), by summing over all possible vector subspaces \( V \), which is finite, we have
\[
\Pr\{\mathcal{A}_{ir} \cap \mathcal{A}_{i0}\} = \sum_V \Pr\{\text{span}(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ir}) = V, \mathbf{X}_{i1} \neq \mathbf{0}\} \Pr\{\mathbf{X}_{i(r+1)} \in V\}.
\]

Given a vector subspace \( V \) with dimension \( t = \dim(V) \in [1 : r] \), we can determine each vector \( \mathbf{v} \in V \) completely based on \( t \) coordinates, for instance we can calculate \( (v_{t+1}, \ldots, v_n) \) based on \( (v_1, \ldots, v_t) \). Without the loss of generality we denote this relation by \( (v_{t+1}, \ldots, v_n) = f_V(v_1, \ldots, v_t) \). Therefore in this case
\[
\Pr\{\mathbf{X}_{i(r+1)} \in V\} \leq \sum_{(x_1, \ldots, x_t) \in \mathcal{X}_i^t} \Pr\{(X_{i(r+1),1}, \ldots, X_{i(r+1),t}) = (x_1, \ldots, x_t)\}
\]
\[
\times \Pr\{(X_{i(r+1),t+1}, \ldots, X_{i(r+1),n}) = f_V(x_1, \ldots, x_t)\}
\]
\[
\leq \sum_{(x_1, \ldots, x_t) \in \mathcal{X}_i^t} \Pr\{(X_{i(r+1),1}, \ldots, X_{i(r+1),t}) = (x_1, \ldots, x_t)\} \times \rho_i^{-t}
\]
\[
= \rho_i^{-t} \leq \rho_i^{n-r}
\]
(5.1)
as \( \rho_i < 1 \) holds. In summary we obtain
\[
\Pr\{\text{rank}(\mathbf{X}_i) \leq R\} \leq \Pr\{\mathcal{A}_{i0}\} + \sum_{r=1}^{R-1} \Pr\{\mathcal{A}_{ir} \cap \mathcal{A}_{i0}\}
\]
\[
\leq \sum_{r=0}^{R-1} \rho_i^{n-r} = \zeta_i \to 0, \text{ as } n \to \infty.
\]
(5.2)
We need a link between the full rank property of each factor matrix and the number of tuples of full rank factor matrices resulting in a given tensor. In case $N \geq 3$ this link is established through the Kruskal’s rank (or k-rank) introduced in [9].

**Definition 5.2.** The k-rank of a matrix $A$, $k_A$, is the largest value of $t$ such that every subset of $t$ columns of $A$ is linearly independent.

We always have $k_A \leq \text{rank}(A)$. For a full-column rank matrix $A$, we have $k_A = \text{rank}(A)$. Lemma 5.1 implies the following result.

**Corollary 5.3.** For each $i \in [1 : N]$ we have

$$\Pr\{k_{X_i} = R\} \geq 1 - \zeta_i \to 1, \text{ as } n \to \infty.$$ 

A sufficient condition based on k-rank that enables us to bound the number of tuples of full rank factor matrices yielding the same tensor is given below.

**Theorem 5.4 ([11, Theorem 3]).** Assume that the order of tensor $N$ satisfies $N \geq 3$. Given deterministic matrices $(X_i)^N_{i=1}$ of size $I_i \times R$ satisfying $Z = [X_1; \ldots; X_N]$ such that $R$ is the rank of tensor $Z$, i.e., the minimum number of rank-one decompositions of $Z$, if

$$\sum_{i=1}^{N} k_{X_i} \geq 2R + (N - 1),$$

Then the matrices $X_i$ are essentially uniquely determined.

The essential uniqueness means that if $(X'_i)^N_{i=1}$ is another tuple of factor matrices satisfying $Z = [X'_1; \ldots; X'_N]$ then for all $i \in [1 : N]$, $X'_i = X_iPA_i$ holds where $P \in \{0,1\}^{R \times R}$ is a unique permutation matrix and $A_i$ are unique diagonal matrices of size $R$ satisfying $\prod_{i=1}^{N} A_i = I_R \times R$.

We then have the following upper bound on the number of full rank factors yielding the same tensor.

**Lemma 5.5.** For given $N \geq 2$, $R \geq 2$, $n \geq R$, and finite alphabets $(X_i)^N_{i=1}$, there exists a number $\Gamma_N$ that does not depend on $n$ such that the number of tuples of full rank matrices $(X'_i)^N_{i=1}$, $X_i \in X_i^{n \times R}$ for all $i \in [1 : N]$, satisfying $T = [X_1; \ldots; X_N]$ is upper bounded by $\Gamma_N$.

**Proof.** Assume that $N \geq 3$ holds. Note that when $X_i$ are full rank for all $i \in [1 : N]$, then rank of $T = [X_1; \ldots; X_N]$ is $R$. Assume otherwise that $T = [X'_1; \ldots; X'_N]$ where $X'_i$ is of size $I_i \times (R - 1)$ for all $i \in [1 : N]$. Unfolding the tensor $T$ according to the first dimension we obtain

$$\sum_{i=1}^{N} k_{X_i} \geq 2R + (N - 1),$$

where $\odot$ is the Khatri-Rao product. We need the following property of the Khatri-Rao product. Since $X_3$ and $X_2$ are full rank, $k_{X_3} = k_{X_2} = R$, we have rank$(X_3 \odot X_2) \geq k_{X_3} + k_{X_2} - 1 \geq R$ by [12, Lemma 3.3]. By applying this inequality consecutively we obtain

$$k_{X_{i-1} \odot X_2} \geq \min\{k_{X_i} + k_{X_{i-1} \odot X_2} - 1, R\} \geq R,$$

since $k_{X_{i-1} \odot X_2} \geq R$ for all $i = 4, \ldots, N$. Therefore we have

$$\text{rank}(X_N \odot \cdots \odot X_2) \geq \text{rank}(X_N \odot \cdots \odot X_2) \geq R.$$
Since the rank of $T$ is $R$, the condition of Theorem 5.4 is satisfied as $R \geq 1 + 1/(N - 2)$ holds. This implies that if $(X_i)_{i=1}^N$, $X_i' \in X_i^{n \times R}$, $\forall i \in [1 : N]$, is another tuple of full rank factor matrices satisfying $T = [X_i' : \cdots : X_N']$ we then have $X_i' = X_i P \Lambda_i$, for all $i \in [1 : N]$. Since our alphabets are discrete, for each $i \in [1 : N]$ the number of such $\Lambda_i$ is finite and does not depend on $n$. We denote the upper bound on the number of different matrices $P \Lambda_i$ by $\Gamma_N$. Hence the conclusion holds in this case.

When $N = 2$, i.e., $T$ is a square matrix, the essential uniqueness does not hold in general. We recall that when $N = 2$ a tensor $T$ can be written as

$$T = X_1 X_2^T.$$  

When $X_1$ and $X_2$ are of rank $R$ then the above expression is a full rank factorization of $T$. If $T = X_1' X_2'^T$ is another full rank factorization of $T$, then there exists an invertible matrix $W \in \mathbb{R}^{R \times R}$ such that $X_1' = X_1 W$ and $X_2' = X_2 (W^{-1})^T$ due to [10, Theorem 2]. Since the alphabets in our study are finite, the number of realizations of the principal minor $\{X_i\}_{1:R,1:R}$ is finite. Therefore the number of such matrices $W$ is finite and does not depend on $n$. We denote the corresponding upper bound on the number of matrices $W$ by $\Gamma_2$. 

By combining Lemma 5.1 and Lemma 5.5 our result when the order of tensor $N \geq 2$ and the number of components $R \geq 2$ is given in the following.

**Theorem 5.6.** The minimum almost-lossless compression threshold is given by

$$C^*_t = \sum_{i,r} H(P_{X_i,r}).$$

We now describe a full compression scheme similar to the one in Theorem 3.1.

**Proof.** For each mode $i \in [1 : N]$ we define a typical set

$$A_{i,\gamma}^n = \{X_i \mid - \log P(X_i) - n \sum_r H(P_{X_i,r}) < n \gamma\},$$

where for each $i \in [1 : N]$ the probability of a realization of factor matrix $X_i$, $P(X_i)$, is given by $P(X_i) = \prod_{r=1}^R P_{X_i,r}(a_{ir})$. Then we define two sets $S_{n,\gamma} = \bigcup_{i=1}^N A_{i,\gamma}^n$ and

$$T_{n,\gamma} = \{[X_1; \cdots : X_N] \mid (X_1, \ldots, X_N) \in S_{n,\gamma}\}.$$

We index all elements in $T_{n,\gamma}$. If $T \in T_{n,\gamma}$ we store the corresponding index, otherwise we store a given index. We need at most $n (\sum_r H(P_{X_i,r}) + \eta)$ nats for the indexing scheme. Hence we have $C^*_t \leq \sum_{i,r} H(P_{X_i,r})$.

In the converse direction, let $(\phi_n, \psi_n)$ be a sequence of tensor compression-reconstruction mappings such that the compression threshold $C$ is almost-losslessly achievable. For a given $\epsilon > 0$ there exists a sufficiently large $n_0(\epsilon)$ such that

$$\Pr\{\psi_n(\phi_n(T)) \neq T\} \leq \epsilon, \ \forall n \geq n_0(\epsilon).$$

Similarly we define the decodable set of tensors

$$D_n = \{T \mid \phi_n(\psi_n(T)) = T\},$$

and the expanding set of corresponding tuples of factor matrices by

$$S_n = \{(X_i)_{i=1}^N \mid [X_1; \cdots : X_N] \in D_n\}.$$

We also define the set of tuples of full rank factors

$$E_n = \{(X_i)_{i=1}^N \mid \text{rank}(X_i) = R, \ \forall i \in [1 : N]\}.$$
By our previous analysis we have $|\mathcal{D}_n| \leq |\mathcal{M}|$ and $|\mathcal{S}_n \cap \mathcal{E}_n| \leq \Gamma_N |\mathcal{D}_n| \leq \Gamma_N |\mathcal{M}|$ for some large enough constant $\Gamma_N$. We define an atypical set

$$
\mathcal{T}_{n,\eta} = \{(X_i)_{i=1}^N \mid -\sum_{i=1}^N \log P(X_i) \geq \log |\mathcal{M}| + \eta \}. 
$$

We then have

$$
\Pr\{\sum_{i=1}^N \log P(X_i) \geq \log |\mathcal{M}| + \eta \} = \Pr\{\sum_{i=1}^N \log P(X_i) \geq \log |\mathcal{M}| + \eta \} 
\leq \Pr\{\sum_{i=1}^N \log P(X_i) \geq \log |\mathcal{M}| + \eta \} = \epsilon + \sum_{i=1}^N \Pr\{\text{rank}(X_i) < R\} + \prod_{i=1}^N P(X_i) 
\leq \epsilon + \sum_{i=1}^N \Pr\{\text{rank}(X_i) < R\} + |\mathcal{S}_n \cap \mathcal{E}_n \cap \mathcal{T}_{n,\eta}| e^{-\eta} / |\mathcal{M}| 
\leq \epsilon + \sum_{i=1}^N \Pr\{\text{rank}(X_i) < R\} + e^{-\eta} \Gamma_N.
$$

(5.5) \hspace{1cm} \text{Lemma 5.5}

We select $\eta = n\gamma$ where $\gamma > 0$ is an arbitrary number and apply a similar line of reasoning as in the converse proof of Theorem 3.1. If for all $n_1 \geq n_0(\epsilon)$ there exists a $n > n_1$ such that $\log |\mathcal{M}| + \eta < n(\sum_{i=1}^n \log |\mathcal{M}| - N \gamma)$, then we have

$$
\lim_{n \to \infty} \Pr\{\sum_{i=1}^N \log P(X_i) \geq \log |\mathcal{M}| + \eta \} = 1,
$$

due to the weak law of large numbers. The last inequality is violated since by Lemma 5.1 we have

$$
\sum_{i=1}^N \Pr\{\text{rank}(X_i) < R\} \to 0, \text{ as } n \to \infty. 
$$

Therefore we must have $\log |\mathcal{M}| + n\gamma \geq n(\sum_{i=1}^n \log |\mathcal{M}| - N \gamma)$ for all $n \geq n_\epsilon$. Since $\gamma$ and $\epsilon$ are arbitrary we have $C_\epsilon \geq \sum_{i=1}^n \log |\mathcal{M}|$.

\section*{REFERENCES}

[1] J. D. Carroll and J.-J. Chang, \textit{Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition}, Psychometrika, 35 (1970), pp. 283–319.

[2] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan, \textit{Tensor decompositions for signal processing applications: From two-way to multiway component analysis}, IEEE Signal Processing Magazine, 32 (2015), pp. 145–163.

[3] T. M. Cover, \textit{Elements of information theory}, John Wiley & Sons, 1999.

[4] T. S. Han, \textit{Information-Spectrum Methods in Information Theory}, Springer-Verlag Berlin Heidelberg, 2003.

[5] R. A. Harshman et al., \textit{Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis}, (1970).

[6] C. J. Hillar and L.-H. Lim, \textit{Most tensor problems are NP-hard}, Journal of the ACM (JACM), 60 (2013), pp. 1–39.

[7] T. S. Han, \textit{Information-Spectrum Methods in Information Theory}, Springer-Verlag Berlin Heidelberg, 2003.

[8] J. Komlós, \textit{On the determinant of random matrices}, Studia Scientiarum Mathematicarum Hungarica, 3 (1968), pp. 387–399.

[9] J. B. Kruskal, \textit{Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics}, Linear algebra and its applications, 18 (1977), pp. 95–138.

[10] R. Piziak and P. Odell, \textit{Full rank factorization of matrices}, Mathematics Magazine, 72 (1999), pp. 193–201.

[11] N. D. Sidiropoulos and R. Bro, \textit{On the uniqueness of multilinear decomposition of N-way arrays}, Journal of Chemometrics: A Journal of the Chemometrics Society, 14 (2000), pp. 229–239.
[12] A. Stegeman and N. D. Sidiropoulos, On Kruskal’s uniqueness condition for the Candecomp/Parafac decomposition, Linear Algebra and its Applications, 420 (2007), pp. 540–552.

[13] T. Tao and V. Vu, On random ±1 matrices: Singularity and determinant, Random Structures & Algorithms, 28 (2006), pp. 1–23.

[14] L. R. Tucker, Implications of factor analysis of three way matrices for measurements of change, In Harris, CW (Editor), Problems in measuring change, (1963), pp. 122–137.