Minimax Lower Bounds for Kronecker-Structured Dictionary Learning

Zahra Shakeri, Waheed U. Bajwa, Anand D. Sarwate
Dept. of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey 08854
{zahra.shakeri, waheed.bajwa, anand.sarwate}@rutgers.edu

Abstract—Dictionary learning is the problem of estimating the collection of atomic elements that provide a sparse representation of measured/collected signals or data. This paper reports lower bounds for the minimax risk of dictionary learning for tensor-structured signals, thereby providing fundamental limits on the number of samples needed to estimate dictionaries for tensor-structured signals in terms of dimensions of constituent elements of the underlying dictionaries and parameters of the generative model. The focus of this paper is in particular on second-order tensor data, with the underlying dictionaries constructed by taking the Kronecker product of two smaller dictionaries and the observed data generated by sparse linear combinations of dictionary atoms observed through white Gaussian noise. In this regard, it provides a general lower bound on the minimax risk and also adapts the proof techniques for equivalent results using sparse and Gaussian coefficient models. The reported results suggest that the sample complexity of dictionary learning for tensor-structured data can be significantly lower than that for unstructured data.

A note to reviewers: The current document is a working paper (February 17, 2016) and is made available for review purposes only. Please do not disseminate. Some constants have been altered from the submitted version after further proof reading, but main results are not changed.

I. INTRODUCTION

Dictionary learning has recently received significant attention from researchers due to increased importance of finding sparse representations of signals. In dictionary learning, the goal is to construct an overcomplete basis that represents input signals such that each can be described by a small number of atoms (columns) [1]. Although the existing literature has focused on one-dimensional data, many signals in practice are multi-dimensional and have a tensor structure; examples include 2-dimensional images or 3-dimensional signals produced via magnetic resonance imaging (MRI) or computed tomography (CT) systems. In traditional dictionary learning techniques, multi-dimensional data is processed through vectorization of the signal; this can result in poor sparse representations [2].

In this paper we prove fundamental limits on learning dictionaries for multi-dimensional data representation with tensor structure; we call such dictionaries Kronecker-structured (KS). Several algorithms have been proposed to learn KS dictionaries [2]–[7] but there has been little work on the theoretical guarantees of such algorithms. We prove lower bounds on the minimax risk of learning a KS dictionary; this gives a measure to evaluate the performance of existing algorithms.

Theoretical insights on classical dictionary learning techniques [8]–[16] have either focused on achievability of existing algorithms [8]–[14] or lower bounds on minimax risk for vector problems. [15], [16]. The former provide sample complexity results for accurate dictionary estimation based on the dictionary learning minimization criterion [8]–[14]. More specifically, given a probabilistic model for sparse signals and a finite number of samples, the dictionary is recoverable within some distance of the generating dictionary as the local minimum of some minimization criterion [12]–[14]. Our work is most similar to Jung et al. [15], [16] in that we prove lower bounds for the minimax risk; they provide bounds for several coefficient vector distributions and analyze a regime in which the bounds are tight for some signal-to-noise (SNR) values. Particularly, for a dictionary \( D \in \mathbb{R}^{m \times p} \), the sample size \( N = O(mp^2) \) is sufficient for reliable recovery of the dictionary in its local neighborhood. For sufficient number of samples, the lower bounds do not depend on the local neighborhood and are also applicable to the global dictionary learning problem.

Our main contribution is providing lower bounds for the minimax risk of dictionaries consisting of two coordinate dictionaries that represent 2-dimensional tensors. We describe the case of two dictionaries here: generalizations to higher dimensions are left for future work. To this end, we construct a class of KS dictionaries and consider the multiple hypothesis testing problem to obtain the minimum achievable worst case mean squared error (MSE), similar to Jung et al. [16]. For a dictionary \( D \) consisting of the Kronecker product of two coordinate dictionaries \( A \in \mathbb{R}^{m_1 \times p_1} \) and \( B \in \mathbb{R}^{m_2 \times p_2} \), where \( m = m_1m_2 \) and \( p = p_1p_2 \), we reduce sample complexity from \( O(mp^2) \) in [16] to \( O(mp(p_1 + p_2)) \). Our results hold even when one of the coordinate dictionaries is not overcomplete (note that both \( A \) and \( B \) cannot be undercomplete, otherwise \( D \) won’t be overcomplete). Our analysis is local and our lower bounds depend on the signal distribution.

II. BACKGROUND AND PROBLEM FORMULATION

Notational Convention: Underlined bold upper-case, bold upper-case and lower-case letters are used to denote tensors, matrices and vectors, respectively. Lower-case letters denote scalars. The \( k \)-th column of \( X \) is denoted by \( x_k \). Let \( X_T \) be the matrix consisting of columns of \( X \) with indices \( T \), \( \sum X \) be the sum of all elements of \( X \), \( I_d \) be the \( d \times d \) identity matrix. Norms are given by subscripts, so \( \| v \|_0 \) and \( \| v \|_2 \) are the \( \ell_0 \)
and \( \ell_2 \) norms of \( v \) and \( \| X \|_2 \) and \( \| X \|_F \) are the spectral and Frobenius norms of \( X \). We write \( [K] \) for \( \{1, \ldots, K\} \).

We write \( X_1 \otimes X_2 \) for the Kronecker product of two matrices \( X_1 \in \mathbb{R}^{m \times n} \) and \( X_2 \in \mathbb{R}^{p \times q} \); the result is an \( mp \times nq \) matrix. Given \( X_1 \in \mathbb{R}^{m \times n} \) and \( X_2 \in \mathbb{R}^{p \times n} \), we write \( X_1 \ast X_2 \) for their \( mp \times n \) Khatri-Rao product \([17]\); this is essentially the column-wise Kronecker product of matrices. Given two matrices of the same dimension \( X_1, X_2 \in \mathbb{R}^{m \times n} \), their \( m \times n \) Hadamard product is denoted by \( X_1 \odot X_2 \), which is the element-wise product of \( X_1 \) and \( X_2 \). For two matrices \( X_1 \) and \( X_2 \), we define their distance to be \( \| X_1 - X_2 \|_F \). We use \( f(\epsilon) = O(g(\epsilon)) \) if \( \lim_{\epsilon \to 0} f(\epsilon)/g(\epsilon) = c < \infty \) for some constant \( c \).

In the conventional dictionary learning setup, it is assumed that the observation \( y \in \mathbb{R}^m \) is generated via a fixed dictionary \( y = Dx + n \), (1) in which the dictionary \( D \in \mathbb{R}^{m \times p} \) is an overcomplete basis \((m < p)\) with unit norm columns, \( x \in \mathbb{R}^p \) is the coefficient vector, and \( n \in \mathbb{R}^m \) is the underlying noise vector. In this paper we focus on 2-dimensional tensors. Consider the 2-dimensional observation \( Y \in \mathbb{R}^{m_1 \times m_2} \). Using any separable transform \( Y \) can be written as
\[
Y = (T_1^{-1})^T \tilde{X} T_2^{-1},
\]
where \( \tilde{X} \in \mathbb{R}^{p_1 \times p_2} \) is the matrix of coefficients and \( T_1 \in \mathbb{R}^{p_1 \times m_1} \) and \( T_2 \in \mathbb{R}^{p_2 \times m_2} \) are non-singular matrices transforming the columns and rows, respectively. Defining \( A \triangleq (T_2^{-1})^T \) and \( B \triangleq (T_1^{-1})^T \), we can use a property of Kronecker products \([18]\), vec\((B^T X A^T) = (A \otimes B) \text{vec}(\tilde{X})\), to get the following expression for \( y \triangleq \text{vec}(Y)\):
\[
y = (A \otimes B)x + n.
\]
for coefficient vector \( x \triangleq \text{vec}(\tilde{X}) \in \mathbb{R}^p \), and noise vector \( n \in \mathbb{R}^m \), where \( p \triangleq p_1 p_2 \) and \( m \triangleq m_1 m_2 \). In this work, we assume \( N \) i.i.d. noisy observations \( y_k \) which are realizations according to model (3). Concatenating the observations in \( Y \in \mathbb{R}^{m_1 \times N} \), we have
\[
Y = DX + N,
\]
where \( D \triangleq A \otimes B \) is the unknown KS dictionary, \( X \in \mathbb{R}^{p \times N} \) is the coefficient matrix consisting of zero-mean random coefficient vectors with known distribution and covariance matrix \( \Sigma_x \), and \( N \in \mathbb{R}^{m \times N} \) is assumed to be additive white Gaussian noise (AWGN) with zero mean and variance \( \sigma^2 \).

Our analysis is local, meaning there is a reference KS dictionary \( D_0 \) with unit norm columns, i.e. \( D_0 \in D_0 \):
\[
D_0 \triangleq \{ D_0' \in \mathbb{R}^{m \times p} : \| d_{0,j} \|_2 = 1 \ \forall j \in [p] \}, \quad D_0' = A_0' \otimes B_0', \quad A_0' \in \mathbb{R}^{m_1 \times p_1}, \quad B_0' \in \mathbb{R}^{m_2 \times p_2}.
\]
We assume the KS dictionary \( D \) consists of unit norm columns and belongs to a small neighborhood around a fixed \( D_0 \):
\[
D \triangleq D' \in \mathbb{R}^{m \times p} : \| d_{j} \|_2 = 1 \ \forall j \in [p] \}, \quad D' = A' \otimes B', \quad A' \in \mathbb{R}^{m_1 \times p_1}, B' \in \mathbb{R}^{m_2 \times p_2}.
\]
and
\[
D \in \mathcal{X}(D_{0}, r) \triangleq \{ D' \in D : \| D' - D_0 \|_F^2 < r \}
\]
where the radius \( r \) is known. Similar to \([16]\), our analysis is applicable to the global KS dictionary learning problem.

The dictionary \( D \) satisfies the restricted isometry property (RIP) \([s, \delta_s]\) of order \( s \) with constant \( \delta_s \) if
\[
(1 - \delta_s)\|x\|_2^2 \leq \|Dx\|_2^2 \leq (1 + \delta_s)\|x\|_2^2.
\]
If \( D = A \otimes B \), where \( A \) and \( B \) have unit norm columns and satisfy RIP\((s, \delta_A^A)\) and RIP\((s, \delta_B^B)\), respectively, then \( D \) satisfies RIP\((s, \delta_s)\) with \( \delta_s \geq \max(\delta_A^A, \delta_B^B) \) \([19]\).

We are interested in the minimax risk for the problem of estimating \( D \) based on the observations, which is measured using the MSE \( \mathbb{E}_Y \{ \| \hat{D}(Y) - D \|_F^2 \} \). Here, \( \hat{D}(Y) \) is the estimated dictionary. More concretely, we want to lower bound
\[
e^* = \inf_D \sup_{D \in \mathcal{X}(D_{0}, r)} \mathbb{E}_Y \{ \| \hat{D}(Y) - D \|_F^2 \}.
\]
The minimax risk depends on various parameters such as the number of observations \( N \), the noise variance \( \sigma^2 \), dimension of the underlying dictionary and the coefficient distribution.

To address this problem, we utilize the multiple hypothesis testing problem \([20]\). Given \( D_0 \) and \( r \), we construct a class of dictionaries \( D_L = \{ D_1, \ldots, D_L \} \subset \mathcal{X}(D_{0}, r) \) for some \( L \geq 2 \) such that the distance between any two of these dictionaries in the neighborhood is sufficiently large and the hypothesis testing problem is sufficiently hard, i.e., distinct dictionaries result in similar observations. Specifically, for \( l, l' \in [L] \), we desire a construction such that
\[
\| D_l - D_{l'} \|_F^2 \geq \alpha_L
\]
\[
D_{KL} \left( f(D_l(Y)) || f(D_{l'}(Y)) \right) \leq \beta_L
\]
where \( D_{KL} \left( f(D_l(Y)) || f(D_{l'}(Y)) \right) \) denotes the Kullback-Leibler divergence between the distribution of observations based on \( D_l \neq D_{l'} \) selected from \( D_L \), and \( \alpha_L, \beta_L \) are non-negative. Then, given observations from dictionary \( D_l \in D_L \), the minimum-distance detector is able to detect the true dictionary if the recovered dictionary \( \hat{D}(Y) \) satisfies \( \| \hat{D}(Y) - D_l \|_F < \gamma_L \), where \( 2\gamma_L < \alpha_L \). According to Fano’s inequality \([20]\),
\[
P(l' \neq l) \geq 1 - \frac{I(Y; l) + 1}{\log_2 L}.
\]
Hence, the lower bound for the minimax risk is obtained by connecting the probability of error to the mutual information (MI) \( I(Y; l) \). For this purpose, we use KL-divergence arguments to upper bound the MI \([16]\). To approximate the upper bound for the KL-divergence, we assume access to some side information \( T(X) \) on the coefficient distribution; we choose \( T(X) \) such that the observations have a multivariate Gaussian distribution which allows us to approximate the upper bound on the KL-divergence \([21]\). Any lower bound for the minimax risk given side information \( T(X) \) is also the lower bound for the general case \([16]\); we consider different \( T(X) \) depending on the coefficient distribution.
A. Coefficient distribution

We will obtain the lower bound of the minimax risk for various coefficient distributions. We follow the same generative model as in [16]. First, we consider any coefficient distribution and only assume that the coefficient covariance matrix exists. We then specialize to sparse coefficient vectors and by adding additional conditions on the reference dictionary $\mathbf{D}_0$, we obtain a tighter lower bound for the minimax risk for some SNR regimes.

1) General coefficients: First, we consider the general case, where $x$ is a zero-mean random coefficient vector with covariance matrix $\Sigma_x$. We make no additional assumption on the distribution of $x$. We assume side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ to obtain a lower bound on the minimax risk.

2) Sparse coefficients: In the case where the coefficient vector is sparse, additional assumptions on the non-zero entries can be made to obtain a lower bound on the minimax risk using less side information, i.e., $\text{supp}(x)$, which denotes the support of $x$. We assume the support of $x$ is randomly distributed uniformly over $\mathcal{E} = \{S \subseteq [p] : |S| = s\}$:

$$\Pr(\text{supp}(x) = S) = \frac{1}{\binom{p}{s}} \quad \text{for any } S \in \mathcal{E}. \quad (13)$$

For a signal $x$ with sparsity pattern $\text{supp}(x)$, we model the non-zero entries of $x$, i.e., $x_S$, as drawn independently and identically from a probability distribution with known variance $\sigma_x^2$:

$$\mathbb{E}_x \{x_Sx_S^T | S\} = \sigma_x^2 \mathbf{I}_s. \quad (14)$$

Any $x$ with sparsity model (13) and nonzero entries satisfying (14) has covariance matrix $\Sigma_x = \frac{s}{p} \sigma_x^2 \mathbf{I}_p$. \quad (15)

We define SNR to be $\text{SNR} = \frac{\sigma_x^2}{m \sigma_x^2}$. III. Lower Bound for General Distribution

We now provide our main result for the lower bound for the minimax risk of the KS dictionary learning problem for the general coefficient distribution.

**Theorem 1.** Consider a KS dictionary learning problem with $N$ i.i.d observations according to model (3) and the true dictionary satisfies (7) for $r \leq 2 \sqrt{p}$. Assuming that $\Sigma_x$ exists for the zero-mean random coefficient vector $x$, then the minimax risk is lower bounded by

$$\epsilon^* \geq \frac{0.0012(0.16(p_1(m_1 - 1) + p_2(m_2 - 1)) - 3) \sigma_x^2}{N\|\Sigma_x\|_2}. \quad (16)$$

**Outline of Proof:** The idea of the proof is that we construct a set of $L$ distinct KS dictionaries that satisfy:

- $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subset \mathcal{X}(\mathbf{D}_0, r)$. Note that the proof of this Theorem differs from the proof of Theorem 1 in [16], as the construction of the KS dictionary class $\mathcal{D}_L$ is not a simple extension of Theorem 1 in [16].

- Any two distinct dictionaries have some minimum distance in the neighborhood, i.e., for any $l, l' \in [L]$:

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \geq C(L). \quad (17)$$

- If the true dictionary, $\mathbf{D}_l$, is selected uniformly at random from $\mathcal{D}_L$, then we have the following bounds for the conditional MI:

$$\eta_2 \leq I(\mathbf{Y}; l | \mathbf{T}(\mathbf{X})) \leq \eta_1, \quad (18)$$

where $l \in [L]$. It can be shown that the bounds on the conditional MI depend on parameters $\epsilon^*, N, m, p, \Sigma_x, s$, and $\sigma_x^2$. The lower bound for the conditional MI is obtained using Fano’s inequality. Given side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, observations have a multivariate Gaussian distribution and an upper bound for the conditional MI is attained by approximating the upper bound for KL-divergence of multivariate Gaussian distribution. Using (18), the lower bound for the minimax risk $\epsilon^*$ is attained.

- We prove that $C(L) = \frac{1}{6\epsilon}$ and the recovered dictionary $\mathbf{D}(\mathbf{Y})$ satisfies $\|\mathbf{D}(\mathbf{Y}) - \mathbf{D}_l\|_F \leq 2\epsilon$, the minimum distance detector can recover the true dictionary $\mathbf{D}_l$, where $\epsilon < \epsilon^*$.

The technical proof relies on the following lemmas, which are proved in the Appendix.

**Lemma 1.** There exists a set of $L = 2^{c_1(mp) - \frac{1}{2}}$ matrices $\mathbf{A}_l \in \mathbb{R}^{m \times p}$, where elements of $\mathbf{A}_l$ take values $\pm \alpha$ for some $\alpha > 0$, such that for $l, l' \in [L], l \neq l'$, and $t > 0$, the following relation is satisfied:

$$\left| \sum_l (\mathbf{A}_l \otimes \mathbf{A}_{l'}) \right| \leq t \quad (19)$$

for $c_1 < \frac{1}{2 \log 2} \left( \frac{p^4 t}{2^{c_1 - 1} m p^2} \right)^2$.

**Lemma 2.** Considering the generative model in (3) such that (7) holds for some $r \leq 2 \sqrt{p}$, there exists a set $\mathcal{D}_l \subseteq \mathcal{D}$ of cardinality $L = 2^{c_2((m_1 - 1)p_1 + (m_2 - 1)p_2) - 1}$ such that for any $l, l' \in [L]$, with $l \neq l'$,

$$c_2 \epsilon' \leq \|\mathbf{D}_l - \mathbf{D}_{l'}\|_F^2 \leq \epsilon' \quad (20)$$

is satisfied with any $\epsilon' > 0$ such that

$$\epsilon' < \frac{r^2}{2} \quad (21)$$

for any $c_1 > 0$ satisfying $c_1 < \frac{r^2}{2 \log 2}$ and $c_2 = \left( \frac{1}{4} - \epsilon^2 \right)$ and any $t > 0$ satisfying $\frac{1}{2} - 2t^2 < 1 - 2t$. Furthermore, considering the general coefficient model for $\mathbf{X}$ and assuming side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, we have,

$$I(\mathbf{Y}; l | \mathbf{T}(\mathbf{X})) \leq \frac{N\|\Sigma_x\|_2^2 \epsilon'}{\sigma_x^2}. \quad (22)$$

**Lemma 3.** Consider the generative model in (3) with minimax risk $\epsilon^* < \epsilon$ for some $\epsilon > 0$. Assume there exists a finite set $\mathcal{D}_l \subseteq \mathcal{D}$ with $L$ dictionaries satisfying

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \geq 8\epsilon \quad (23)$$

for $l \neq l'$. Then for any side information $\mathbf{T}(\mathbf{X})$, we have

$$I(\mathbf{Y}; l | \mathbf{T}(\mathbf{X})) \geq \frac{1}{2} \log_2 (L) - 1. \quad (24)$$
Proof of Theorem 1. According to Lemma 2 and Lemma 3, for any \( c' \leq r^2 \), there exists a set \( D_L \subseteq \mathcal{H}(D_0, r) \) of cardinality \( L = 2^{c_1((m_1-1)p_1 + (m_2-1)p_2) - 1} \) that satisfies (22) and (23) if we set
\[
8\epsilon = c_2 c'
\] (25)
and \( c_1 > 0 \) and \( t > 0 \) satisfy constraints in Lemma 2. Combining (22) and (24) we get
\[
\frac{1}{2} \log_2(L) - 1 \leq I(Y; l|T(X)) \leq \frac{N\|\Sigma_x\|_2c'}{\sigma^2},
\] (26)
which implies
\[
\epsilon \geq C_1 \frac{\sigma^2}{N\|\Sigma_x\|_2} (c_1(p_1(m_1-1) + p_2(m_2-1)) - 3),
\] (27)
where \( C_1 = \frac{2}{16} \). Setting \( c_1 = 0.16 \), \( t = 0.48 \), \( c_2 = 0.0196 \) and \( C_1 = 1.2 \times 10^{-3} \), we obtain (16).

\[\square\]

IV. LOWER BOUND FOR SPARSE DISTRIBUTIONS

Since we are interested in sparse coefficient vectors, we now consider the sparse coefficient model (13) and obtain lower bounds for the corresponding minimax risk. We first state a corollary of Theorem 1, corresponding to \( T(X) = X \).

Corollary 1. Consider a KS dictionary learning problem with \( N \) i.i.d. observations according to model (3). Assuming the true dictionary satisfies (7) for \( r \leq 2\sqrt{p} \) and the reference dictionary \( D_0 \) satisfies RIP\((s, \frac{1}{2})\), if the random coefficient vector \( x \) is selected according to (13), the minimax risk is lower bounded by
\[
\epsilon^* \geq \frac{\sigma^2}{\sigma_a^2} \frac{0.0012p \left( 0.16(p_1(m_1-1) + p_2(m_2-1)) - 3 \right)}{NS}.
\] (28)

This result is a direct consequence of Theorem 1, by substituting the covariance matrix of the sparse coefficient matrix given in (15) in (16).

A. Sparse Gaussian Coefficients

In this section, we make an additional assumption on the coefficient vector generated according to (13), assuming non-zero elements of the vector follow a Gaussian distribution. By additionally assuming the non-zero entries of \( x \) are i.i.d Gaussian distributed, we can instead write \( x_S \) as
\[
x_S \sim \mathcal{N}(0, \sigma_a^2 I_s).
\] (29)

As a result, given side information \( T(x_k) = \text{supp}(x_k) \), observations \( y_k \) follow a multivariate Gaussian distribution. We provide a theorem for the lower bound attained for this coefficient distribution:

Theorem 2. Consider the KS dictionary learning problem with \( N \) i.i.d. observations according to model (3). Assuming the true dictionary satisfies (7) for \( r \leq 2\sqrt{p} \) and the reference dictionary \( D_0 \) satisfies RIP\((s, \frac{1}{2})\), if the random coefficient vector \( x \) is selected according to (13) and (29), the minimax risk is lower bounded by
\[
\epsilon^* \geq \frac{\sigma^2}{\sigma_a^2} \frac{Cp \left( 0.16(p_1(m_1-1) + p_2(m_2-1)) - 3 \right)}{Ns^2},
\] (30)
where \( C = 7.25 \times 10^{-8} \).

Outline of Proof: The constructed dictionary class \( \mathcal{D}_L \) in Theorem 2 is similar to that in Theorem 1. The upper bound for the conditional MI \( I(Y; l|\text{supp}(X)) \) differs from that of Theorem 1 as the side information is different.

Given the true dictionary \( D_l \) and support \( S_k \) for the coefficient vector \( x_k \), \( D_{l,s_k} \) denotes the true dictionary restricted to columns corresponding to the non-zeros elements of \( x_k \). In this case
\[
y_k = D_{l,s_k} x_{s_k} + n_k.
\] (31)

We can write the dictionary \( D_{l,s_k} \) in terms of the Khatri-Rao product of matrices:
\[
D_{l,s_k} = A_{l,s_k} B_{l,k},
\] (32)
where \( S_k = \{i_k, k \in [p] \} \) and \( S_{k,s} = \{i_k', k \in [p] \} \) are multisets with the following relationship with \( S_k = \{i_k'\}_{k=1}^s \) and \( i_k' \in [p] \):
\[
i_k' = (i_k - 1)p_2 + i', \quad k \in [s].
\] (33)

Note that \( A_{l,s_k} \) and \( B_{l,k} \) are not submatrices of \( A_{l,a} \) and \( B_{l,s} \), as \( S_{ka} \) and \( S_{kb} \) are multisets. Figure 1 demonstrates an example of (32). Therefore, the covariance matrix of the coefficient matrix can be written as
\[
\Sigma_{k,l} = \sigma_a^2 (A_{l,s_k} B_{l,k}) (A_{l,s_k} B_{l,k})^T + \sigma^2 I_s.
\] (34)

We state a variation of Lemma 2 necessary for the proof. The proof of the lemma is provided in the Appendix.

Lemma 4. Consider the KS dictionary learning problem such that (7) holds for some \( r \leq 2\sqrt{p} \). Then there exists a set of dictionaries \( D_L \subseteq D \) of cardinality \( L = 2^{c_1((m_1-1)p_1 + (m_2-1)p_2)} \) such that for some \( c_1 > 0 \) and \( c_2 > 0 \) satisfying conditions in Lemma 1 and any \( 0 < c' \leq r^2 \):
\[
c_2 c' \leq \|D_L - D_l\|_F^2 \leq c', \quad \text{for} \quad D_l, D'_l \in D_0, \quad l \neq l'.
\] (35)
Furthermore, assuming the reference dictionary $\mathbf{D}_0$ satisfies RIP$(s, \frac{1}{2})$ and the coefficient matrix $\mathbf{X}$ is selected according to (13) and (29), considering side information $\mathbf{T}(\mathbf{X}) = \text{supp}(\mathbf{X})$, for some $c_3 > 0$ and any $c' > 0$ satisfying

$$c' \leq \frac{r^2}{s},$$

we have the following upper bound for $I(Y; l|\mathbf{T}(\mathbf{X}))$:

$$I(Y; l|\mathbf{T}(\mathbf{X})) \leq c_3 N s^2 \left( \frac{\sigma_a}{\sigma} \right)^4 \frac{c'}{p}.$$  \hfill (36)

**Proof of Theorem 2.** According to Lemma 4, for any $c' \leq \frac{c_2}{r^2}$, there exists a set $\mathcal{D}_L \subseteq \mathcal{X}(\mathbf{D}_0, r)$ of cardinality $L = 2^{c_1((m_1-1)p_1 + (m_2-1)p_2) - 1}$ that satisfies (23) and (37) if (25) is satisfied and $c_1 > 0$ and $t > 0$ satisfy constraints in Lemma 2. Consequently,

$$\frac{1}{2} \log_2(L) - 1 \leq I(Y; l|\mathbf{T}(\mathbf{X})) \leq c_3 N s^2 \left( \frac{\sigma_a}{\sigma} \right)^4 \frac{c'}{p}$$

which implies

$$\epsilon \geq C_2 \left( \frac{\sigma_a}{\sigma} \right)^4 \frac{p}{N s^2} (c_1 (p_1 (m_1 - 1) + p_2 (m_2 - 1)) - 3),$$

\hfill (38)

where $C_2 = \frac{c_2}{16c_3}$. Setting $c_1 = 0.16$, $t = 0.48$, $c_2 = 0.0196$, $c_3 = 16900$, and $C_2 = 7.25 \times 10^{-8}$, we obtain (30).

\[\Box\]

### V. Discussion and Conclusion

In this paper we follow an information-theoretic approach to provide lower bounds for the worst-case MSE of KS dictionaries that generate 2-dimensional tensor data. Table I demonstrates the dependence of the minmax rate on various parameters of the dictionary learning problem. Our bounds capture the dependence on the SNR. In high SNR regimes, the lower bound in (28) is tighter, while (30) results in a tighter lower bound in low SNR regimes. Compared to results in [16] for the unstructured dictionary, we are able to decrease the lower bound in all cases by reducing scaling $O(\sigma^2 m)$ to $O(p(m_1 p_1 + m_2 p_2))$ for KS dictionaries.

Our bounds depend on the signal distribution and imply necessary sample complexity scaling $N = O(p(m_1 p_1 + m_2 p_2))$. Future work includes extending the lower bounds for more than 2-dimensional tensors and also finding a learning scheme that achieves these lower bounds.

### Table I

<table>
<thead>
<tr>
<th>Coefficient Distribution</th>
<th>Unstructured Dictionary [16]</th>
<th>KS Dictionary [this paper]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. General Case</td>
<td>$\frac{\sigma^4 m p}{N|\Sigma_x|^2}$</td>
<td>$\frac{\sigma^4(m_1 p_1 + m_2 p_2)}{N|\Sigma_x|^2}$</td>
</tr>
<tr>
<td>2. Sparse</td>
<td>$(\frac{\sigma}{\sigma_a})^4 \frac{m p^2}{N s}$</td>
<td>$(\frac{\sigma}{\sigma_a})^4 \frac{p(m_1 p_1 + m_2 p_2)}{N s}$</td>
</tr>
<tr>
<td>3. Gaussian Sparse</td>
<td>$(\frac{\sigma}{\sigma_a})^4 \frac{m p^2}{N s^2}$</td>
<td>$(\frac{\sigma}{\sigma_a})^4 \frac{p(m_1 p_1 + m_2 p_2)}{N s^2}$</td>
</tr>
</tbody>
</table>

**Acknowledgement**

The work of the authors was supported in part by the National Science Foundation under awards CCF-1525276 and CCF-1453073, and by the Army Research Office under award W911NF-14-1-0295.

### References


Therefore, for any $c_1 > 0$ such that
\[ c_1 < \frac{1}{2\log 2} \left( \frac{t}{2\alpha^2 mp} \right)^2, \] (45)
we can take $L = 2^{c_1(m_p + \frac{1}{2})}$.
\[ \square \]

**Proof of Lemma 2.** We define the reference directory, $D_0 \in D_0$, to be
\[ D_0 = A_0 \otimes B_0, \] (46)
where $A_0 \in \mathbb{R}^{m_1 \times p_1}$ and $B_0 \in \mathbb{R}^{m_2 \times p_2}$. Let $\{U_j\}_{j=1}^{p_1}$ be arbitrary unitary matrices satisfying
\[ a_{0,j} = U_j e_1, \quad b_{0,j} = U_j e_1. \] (47)

To construct the dictionary class $D_L \subseteq X(D_0, r)$, we follow several steps. According to Lemma 1, there is a simultaneous existence of a set of $L_a$ matrices $A_{1,l_a} \in \left\{ \frac{1}{\sqrt{2(m_1-1)p_1}}, \frac{1}{\sqrt{2(m_1-1)p_1}} \right\}, l_a \in [L_a]$ with $L_a = 2^{c_1(m_1-1)p_1-1}$ and a set of $L_b$ matrices $B_{1,l_b} \in \left\{ \frac{1}{\sqrt{2(m_2-1)p_2}}, \frac{1}{\sqrt{2(m_2-1)p_2}} \right\}, l_b \in [L_b]$ with $L_b = 2^{c_1(m_2-1)p_2-1}$ such that the following inequalities are satisfied:
\[ \sum (A_{1,l_a} \otimes B_{1,l_b}) \leq t \] (48)
for any $t > 0$ satisfying
\[ \frac{1}{2} - 2t^2 \leq 1 - 2t \] (49)
and $l_a \neq l'_a, l_b \neq l'_b$. Note that as we need the existence of two sets of matrices that simultaneously satisfy (48), the probability that (48) is not satisfied for any set has to be less than $\frac{1}{2}$ instead of 1. This results in smaller set cardinalities relative to the one obtained in Lemma 1.

We now define sets of matrices $\{A_{2,l_a} \in \mathbb{R}^{m_1 \times p_1}, l_a \in [L_a]\}$ and $\{B_{2,l_b} \in \mathbb{R}^{m_2 \times p_2}, l_b \in [L_b]\}$ that satisfy
\[ \|A_{2,l_a} - A_{2,l'_a}\|_F \geq 1 - 2t \] (50)
\[ \|B_{2,l_b} - B_{2,l'_b}\|_F \geq 1 - 2t \]
for any $l_a, l'_a \in [L_a], l_b \neq l'_b$ and $l_b, l'_b \in [L_b], l_b \neq l'_b$. To show (50), we construct $A_{2,l_a}$ and $B_{2,l_b}$ column-wise. Let the $j$-th columns of $A_{2,l_a}$ and $B_{2,l_b}$ be given by
\[ a_{2,l_a,j} = U_j \begin{pmatrix} 0 \\ a_{1,l_a,j} \end{pmatrix} \] (51)
\[ b_{2,l_b,j} = U_j' \begin{pmatrix} 0 \\ b_{1,l_b,j} \end{pmatrix} \]
for any $l_a \in [L_a]$ and $l_b \in [L_b]$. It is apparent that $\|A_{2,l_a}\|_F = \|A_{1,l_a}\|_F$ and $\|B_{2,l_b}\|_F = \|B_{1,l_b}\|_F$. For distinct indices $l_a$ and $l'_a$,
\[ \sum (A_{2,l_a} \otimes A_{2,l'_a}) = \sum_{j=1}^{p_1} (a_{2,l_a,j} a_{2,l'_a,j}) \]
where (a) follows from the fact that \( \mathbf{U}_j \) is unitary. Hence

\[
\| \mathbf{A}_{2,l_a} - \mathbf{A}_{2,l'_a} \|^2_F
= \| \mathbf{A}_{2,l_a} \|^2_F + \| \mathbf{A}_{2,l'_a} \|^2_F - 2 \sum_{i=1}^{p_1} \langle \mathbf{a}_{1,l_a,i}, \mathbf{a}_{1,l'_a,i} \rangle
\geq \| \mathbf{A}_{2,l_a} \|^2_F - 2 \sum_{i=1}^{p_1} \langle \mathbf{a}_{1,l_a,i}, \mathbf{a}_{1,l'_a,i} \rangle
= \| \mathbf{A}_{2,l_a} \|^2_F - 2 \sum_{i=1}^{p_1} \langle \mathbf{a}_{1,l_a,i}, \mathbf{a}_{1,l'_a,i} \rangle
= 1/2 + 1 - 2t
= 1 - 2t. \tag{53}
\]

Similarly, for distinct \( l_b, l'_b \), we have

\[
\sum \langle \mathbf{b}_{2,l_b,i}, \mathbf{b}_{2,l'_b,i} \rangle = \sum \langle \mathbf{b}_{1,l_b,i}, \mathbf{b}_{1,l'_b,i} \rangle
\geq 1 - 2t. \tag{54}
\]

Moreover, defining

\[
\mathcal{D}_2 \triangleq \{ \mathbf{A}_{2,l_a} \otimes \mathbf{B}_{2,l_b} : l_a \in [L_a], l_b \in [L_b] \} \tag{55}
\]

it is clear that \( \mathcal{D}_2 \) has \( L \triangleq L_a L_b \) elements. Any element of \( \mathcal{D}_2 \) can be written as

\[
\mathbf{D}_{2,l} = \mathbf{A}_{2,l_a} \otimes \mathbf{B}_{2,l_b}. \tag{56}
\]

A direct calculation gives us that \( \mathbf{a}_{0,j} \) is orthogonal to \( \mathbf{a}_{2,l_a,j} \) and \( \mathbf{b}_{0,j} \) is orthogonal to \( \mathbf{b}_{2,l_a,j} \). Hence, we have

\[
diag \left( D_0^T D_2 \right) = diag \left( (A_0^T \otimes B_0^T)(A_{2,l_a} \otimes B_{2,l_b}) \right)
= diag \left( A_0^T A_{2,l_a} \right) \otimes diag \left( B_0^T B_{2,l_b} \right)
= 0. \tag{57}
\]

where (b) follows from properties of Kronecker products \([23]\) and \( 0 \in \mathbb{R}^{p} \) is the zero vector. Also, for any \( j \in [p] \), we have

\[
\| \mathbf{d}_{2,l,j} \|^2_F = \| A_{2,l_a,j} \|^2_F \geq 1 - 2t. \tag{58}
\]

where (c) follows from norm properties of the Kronecker product and \( j_a \in [p_1], j_b \in [p_2] \). Therefore for \( \mathbf{D}_{2,l} \), if \( l_a \neq l'_a \) and \( l_b \neq l'_b \), we have

\[
\| \mathbf{D}_{2,l} - \mathbf{D}_{2,l'} \|^2_F
= \| \mathbf{A}_{2,l_a} \otimes \mathbf{B}_{2,l_b} - \mathbf{A}_{2,l'_a} \otimes \mathbf{B}_{2,l'_b} \|^2_F
\geq \| \mathbf{A}_{2,l_a} \|^2_F + \| \mathbf{A}_{2,l'_a} \|^2_F - 2 \| \mathbf{A}_{2,l_a} \otimes \mathbf{A}_{2,l'_a} \|_F^2
- 2 \sum \langle \mathbf{A}_{2,l_a} \otimes \mathbf{A}_{2,l'_a} \rangle \langle \mathbf{B}_{2,l_b} \otimes \mathbf{B}_{2,l'_b} \rangle
= 1/2 + 1 - 2t
= 1/2 - 2t^2. \tag{59}
\]

If \( l_a = l'_a \) and \( l_b \neq l'_b \), we have

\[
\| \mathbf{D}_{2,l} - \mathbf{D}_{2,l'} \|^2_F
\geq \| B_{2,l_b} - B_{2,l'_b} \|^2_F
\geq \frac{1}{2} (1 - 2t). \tag{60}
\]

where (d) follows from (49). We obtain the same lower bound for the case where \( l_a \neq l'_a \) and \( l_b = l'_b \). Therefore, for any distinct \( l, l' \in [L] \), we have

\[
\| \mathbf{D}_{2,l} - \mathbf{D}_{2,l'} \|^2_F \geq \frac{1}{4} - t^2. \tag{61}
\]

We are ready to define \( \mathcal{D}_L \). The conditions in the lemma ensure that any \( \epsilon' \) satisfies

\[
0 < \epsilon' < r^2 < 4p. \tag{62}
\]

We define

\[
\mathbf{A}_{l_a} = \sqrt{1 - \frac{\epsilon'}{4p}} \mathbf{A}_0 + \sqrt{\frac{\epsilon'}{2p}} \mathbf{A}_{2,l_a},
\]

\[
\mathbf{B}_{l_b} = \sqrt{1 - \frac{\epsilon'}{4p}} \mathbf{B}_0 + \sqrt{\frac{\epsilon'}{2p}} \mathbf{B}_{2,l_b}, \tag{63}
\]

and we define the dictionary class \( \mathcal{D}_L \) to be

\[
\mathcal{D}_L \triangleq \{ \mathbf{A}_{l_a} \otimes \mathbf{B}_{l_b} : l_a \in [L_a], l_b \in [L_b] \}, \tag{64}
\]

which has cardinality \( |\mathcal{D}_L| = L \) and its elements can be written as

\[
\mathbf{D}_l = \mathbf{A}_{l_a} \otimes \mathbf{B}_{l_b}
= \left( 1 - \frac{\epsilon'}{4p} \right) (\mathbf{A}_0 \otimes \mathbf{B}_0) + \sqrt{\frac{\epsilon'}{2p}} \left( 1 - \frac{\epsilon'}{4p} \right) (\mathbf{A}_{2,l_a} \otimes \mathbf{B}_{2,l_b})
+ \sqrt{\frac{\epsilon'}{4p}} (\mathbf{A}_{2,l_a} \otimes \mathbf{B}_{2,l_b}) \tag{65}
\]

To show \( \mathbf{D}_l \in \mathcal{A}(\mathbf{D}_0, r) \), we first show it has unit norm.
columns:
\[
\|d_{i,j}\|^2 = \|\mathbf{a}_{i,j} \otimes \mathbf{b}_{i,j}\|^2_F \\
= \|\mathbf{a}_{i,j}\|^2_F \|\mathbf{b}_{i,j}\|^2_F \\
= \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{a}_{i,j}\|^2_F + \frac{\epsilon'}{2p} \|\mathbf{a}_{i,j}\|^2_F \\
+ \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{b}_{i,j}\|^2_F + \frac{\epsilon'}{2p} \|\mathbf{b}_{i,j}\|^2_F \\
= \left(1 - \frac{\epsilon'}{4p}\right) + \frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right) \\
+ \left(1 - \frac{\epsilon'}{4p}\right) + \frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right) \\
= 1. \tag{66}
\]

Furthermore, we have
\[
\|D_0 - D_l\|^2_F = \left\|\left(1 - \left(1 - \frac{\epsilon'}{4p}\right)\right) \mathbf{D}_0 \right. \\
- \sqrt{\frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right)} (\mathbf{A}_0 \otimes \mathbf{B}_{2,l_0}) \\
- \sqrt{\frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right)} (\mathbf{A}_{2,l_0} \otimes \mathbf{B}_0) - \sqrt{\frac{\epsilon'}{4p} \mathbf{D}_{2,l}} \right\|_F^2 \\
= \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{D}_0\|^2_F + \frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_0\|^2_F \|\mathbf{B}_{2,l_0}\|^2_F \\
+ \frac{\epsilon'}{2p} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{B}_0\|^2_F \|\mathbf{A}_{2,l_0}\|^2_F + \frac{\epsilon'}{2p} \mathbf{D}_{2,l} \|\mathbf{D}_{2,l}\|^2_F \\
= \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{D}_0\|^2_F + \frac{\epsilon'}{4p} \left(1 - \frac{\epsilon'}{4p}\right) + \frac{\epsilon'}{4p} \left(1 - \frac{\epsilon'}{4p}\right) + \frac{\epsilon'}{16p} \\
= \frac{\epsilon'}{2} \left(1 - \frac{\epsilon'}{2} - \frac{\epsilon'}{8} \right), \tag{67}
\]
where (e) and (f) follow from (57) and (62), respectively. Thus, we have shown \(D_L \subset \mathcal{A}(D_0, r)\). We now obtain bounds on the distance between any two distinct dictionaries, \(D_l, D_{l'} \in D_L\).

a) Lower bounding \(\|D_l - D_{l'}\|^2_F\): For any \(l, l' \in [L]\) such that \(l_a \neq l'_a\) and \(l_i \neq l'_i\), we have,
\[
\|D_l - D_{l'}\|^2_F = \left\|\sqrt{\frac{\epsilon'}{2p_1}} \left(1 - \frac{\epsilon'}{4p}\right) (\mathbf{B}_{2,l_0} - \mathbf{B}_{2,l'_0}) \right. \\
+ \sqrt{\frac{\epsilon'}{2p_2}} \left(1 - \frac{\epsilon'}{4p}\right) ((\mathbf{A}_{2,l_0} - \mathbf{A}_{2,l'_0}) \otimes \mathbf{B}_0) \\
+ \sqrt{\frac{\epsilon'}{2p_1}} \left(1 - \frac{\epsilon'}{4p}\right) (\mathbf{A}_{2,l_0} \otimes \mathbf{B}_{2,l_0}) \\
= \frac{\epsilon'}{2p_1} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l_0} - \mathbf{B}_{2,l'_0}\|^2_F \\
+ \frac{\epsilon'}{2p_2} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_{2,l_0} - \mathbf{A}_{2,l'_0}\|^2_F + \frac{\epsilon'}{2p_1} \|\mathbf{A}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l_0}\|^2_F \\
+ \frac{\epsilon'}{4p} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{B}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l'_0}\|^2_F \\
\leq \frac{\epsilon'}{2p_1} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l_0} - \mathbf{B}_{2,l'_0}\|^2_F \\
+ \frac{\epsilon'}{2p_2} \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_{2,l_0} - \mathbf{A}_{2,l'_0}\|^2_F + \frac{\epsilon'}{2p_1} \|\mathbf{A}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l_0}\|^2_F \\
+ \frac{\epsilon'}{4p} \|\mathbf{B}_{2,l_0}\|^2_F \|\mathbf{B}_{2,l'_0}\|^2_F \\
\geq \left(1 - \frac{\epsilon'}{4p}\right) \left(1 - 2t\right) + \frac{\epsilon'}{4p} \left(1 - 2t^2\right) + \frac{\epsilon'}{8p} \left(1 - 2t^2\right) \\
\geq \left(1 - \frac{\epsilon'}{2} - \frac{\epsilon'}{8} \right) \left(1 - 2t\right), \tag{70}
\]

b) Upper bounding \(\|D_l - D_{l'}\|^2_F\):
\[
\|D_l - D_{l'}\|^2_F = \epsilon' \left(1 - \frac{\epsilon'}{4p}\right) \|\mathbf{A}_{2,l_0} - \mathbf{A}_{2,l'_0}\|^2_F \\
+ \frac{\epsilon'}{4p} \|\mathbf{D}_{2,l} - \mathbf{D}_{2,l'}\|^2_F \\
\leq \epsilon' \left(1 - \frac{\epsilon'}{4p}\right) \left(\|\mathbf{A}_{2,l_0}\|^2_F + \|\mathbf{A}_{2,l'_0}\|^2_F\right)^2 + \frac{\epsilon'}{4p} \left(\|\mathbf{D}_{2,l}\|^2_F + \|\mathbf{D}_{2,l'}\|^2_F\right)^2 \\
\leq \epsilon'. \tag{71}
\]

We now obtain the upper bound for \(I(Y; l|T(X))\) for the dictionary set \(D_L\) according to the general coefficient model.
and side information $T(X) = X$.

Assuming side information $T(X) = X$, conditioned on the coefficients $x_k$, the observations $y_k$ follow a multivariate Gaussian distribution with covariance matrix $\sigma^2 I$ and mean vector $Dx_k$. From the convexity of the KL divergence [24], following similar arguments as in Jung et al. [16] based on the Fano method [20], we have

$$ I(Y; l|T(X)) = I(Y; l|X) $$

where $f_{DL}(Y|X)$ is the probability distribution of the observations $Y$, given the coefficient matrix $X$ and $D_l$. From [25], we have

$$ D_{KL}(f_{DL}(Y|X)||f_{DL'}(Y|X)) = \sum_{k \in [N]} \frac{1}{2\sigma^2} \| (D_l - D_{l'}) x_k \|^2_2 $$

Substituting (73) in (72) results in

$$ I(Y; l|T(X)) \leq \sum_{k \in [N]} \frac{1}{2\sigma^2} \| \Sigma_x \|_2 \| (D_l - D_{l'}) x_k \|^2_2 $$

where (i) follows from $\text{Tr}(A^T A \Sigma_x) \leq \| \Sigma_x \|_2 \| A \|_F^2$ and (j) follows from (35).

Proof of Lemma 3. The proof is similar to the proof of Lemma 4 in [16].

Proof of Lemma 4. The dictionary class $D_L$ constructed in Lemma 2 is considered here. Note that (36) implies $\ell' < r^2$ as $s \geq 1$. In this case the coefficient vector is assumed to be sparse according to (13), hence conditioned on $S_k = \text{supp}(x_k)$, observations $y_k$'s are zero-mean independent multivariate Gaussian random vectors with covariances

$$ \Sigma_{k,l} = \sigma^2 D_l, k, D_l^T, k, I_k, $$

where $D_l$ is the true dictionary. According to [21] and [16], this conditional MI has the following upper bound:

$$ I(Y; l|T(X)) \leq \text{rank}(\Sigma_{k,l} - \Sigma_{k,l'}) $$

Conditioned on the support $S_k$, the locations of non-zeros in $x_k$ are known. Denoting $x_{S_k}$ as the elements of $x_k$ with indices $S_k$, we have

$$ y_k = D_l, s, x_{S_k} + n_k. $$

When non-zero elements of the coefficient vector are selected according to (29), we can write the dictionary $D_l, s, k$ in terms of the Khatri-Rao product of matrices:

$$ D_l, s, k = A_{l,s, s}, k * B_{l,s, s}, k, $$

where

$$ S_k = \{ i_k \}_{k=1}^l, i_k \in [p_1], $$

$$ S_{kb} = \{ j_k \}_{k=1}^l, j_k \in [p_2], $$

and $A_{l,s, s}, k * B_{l,s, s}, k \in \mathbb{R}^{m_1 m_2 \times s}$. Note that $S_k$ and $S_{kb}$ are multisets. Therefore, (75) can be written as

$$ \Sigma_{k,l} = \gamma^2 \sigma^2 (A_{l,s, s} * B_{l,s, s})(A_{l,s, s} * B_{l,s, s})^T + \sigma^2 I_s. $$

Since rank($\Sigma_{k,l}$) $\leq s$, rank($\Sigma_{k,l} - \Sigma_{k,l'}$) $\leq 2s$. Defining the constants

$$ \alpha = \sqrt{\frac{\ell'}{2p_1}}, $$

$$ \beta = \frac{\ell'}{2p_1}, $$

$$ \gamma = \sqrt{1 - \frac{\ell'}{4p_1}}, $$

we have

$$ \frac{1}{\sigma^2} (\Sigma_{k,l} - \Sigma_{k,l'}) $$

$$ = (A_{l,s, s} * B_{l,s, s})(A_{l,s, s} * B_{l,s, s})^T $$

$$ - (A_{l,s, s} * B_{l,s, s})(A_{l,s, s} * B_{l,s, s})^T $$

$$ = (\gamma A_{l,s, s} + \alpha A_{2,l,s, s}) * (\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ (\gamma A_{l,s, s} + \alpha A_{2,l,s, s}) * (\gamma B_{l,s, s} + \beta B_{2,l,s, s})^T $$

$$ = \gamma^2 (A_{l,s, s} * B_{l,s, s})(\gamma (B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ + (\alpha A_{2,l,s, s} * (\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ = (A_{l,s, s} * B_{l,s, s})(\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ + (\alpha A_{2,l,s, s} * (\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ = (A_{l,s, s} * B_{l,s, s})(\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

$$ + (\alpha A_{2,l,s, s} * (\gamma B_{l,s, s} + \beta B_{2,l,s, s}) $$

where $D_l$ is the true dictionary. According to [21] and [16], this conditional MI has the following upper bound:

$$ I(Y; l|T(X)) $$

$$ \leq \text{rank}(\Sigma_{k,l} - \Sigma_{k,l'}) $$

$$ \mathbb{E}_{T(X)} \left\{ \sum_{k=1}^N \frac{1}{L^2} \sum_{l,l'} \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2 \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2 \right\}. $$

(76)
Using
\[ \| A_1 * A_2 \|_2 = \| (A_1 \otimes A_2)J \|_2 \]
\[ \leq \| (A_1 \otimes A_2) \|_2 \| J \|_2 \]
\[ = \| A_1 \|_2 \| A_2 \|_2 \| J \|_2 \]
\[ = \| A_1 \|_2 \| A_2 \|_2, \]
where \( J \in \mathbb{R}^{p \times s} \) is a selection matrix. Assuming \( s_k = \{ i_k^s \}_{k=1}^1, j_k = e_{i_k}, \) for \( k \in [s], \) and \( J^T J = I_s. \) We have
\[ \frac{1}{\sigma_a^2} \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2 \]
\[ \leq 2 \gamma^2 \| A_{0,s_{ka}} * B_{0,s_{kb}} \|_2 (\gamma \| A_{0,s_{ka}} * B_{2,t_0,s_{kb}} \|_2)
+ \alpha \| A_{2,t_0,s_{ka}} * (\gamma B_{0,s_{kb}} + B_{2,t_0,s_{kb}}) \|_2\]
\[ + 2(\gamma \| A_{0,s_{ka}} * B_{2,t_0,s_{kb}} \|_2)
+ \alpha \| A_{2,t_0,s_{ka}} * (\gamma B_{0,s_{kb}} + B_{2,t_0,s_{kb}}) \|_2\]
\[ \| (\gamma A_{0,s_{ka}} + \alpha A_{2,t_0,s_{ka}}) * (\gamma B_{0,s_{kb}} + B_{2,t_0,s_{kb}}) \|_2 \]
\[ \leq 2(2) \gamma^3 \| A_{0,s_{ka}} \|_2 \| B_{0,s_{kb}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ 2\gamma^3 \alpha \| A_{0,s_{ka}} \|_2 \| B_{0,s_{kb}} \|_2 \| A_{2,t_0,s_{ka}} \|_2
+ 4\gamma^2 \alpha \| A_{0,s_{ka}} \|_2 \| B_{0,s_{kb}} \|_2 \| A_{2,t_0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ \gamma^2 \| A_{0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ 2\gamma \| A_{0,s_{ka}} \|_2 \| B_{0,s_{kb}} \|_2 \| A_{2,t_0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ \gamma \| A_{0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ 2\gamma \| A_{0,s_{ka}} \|_2 \| B_{0,s_{kb}} \|_2 \| A_{2,t_0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2
+ \alpha \| A_{2,t_0,s_{ka}} \|_2 \| B_{2,t_0,s_{kb}} \|_2. \]
(84)

From (36) it is apparent that
\[ \frac{se'}{4p} \leq \sqrt{\frac{se'}{4p}} \leq 1. \]
(85)

Furthermore,
\[ \| A_{0,s_{ka}} \|_2 \leq \sqrt{\frac{3}{2}} < 2, \quad \| B_{0,s_{kb}} \|_2 \leq \sqrt{\frac{3}{2}} < 2, \]
(86)

and
\[ \| A_{2,t_0,s_{ka}} \|_2 \leq \sqrt{\frac{s}{2p_1}}, \quad \| B_{2,t_0,s_{kb}} \|_2 \leq \sqrt{\frac{s}{2p_2}}, \]
(87)

where (86) follows from the RIP condition for \( D_0 \) and (87) follows from basic properties of matrix norms that \( \| A_{2,t_0,s_{ka}} \|_2 \leq \| A_{2,t_0,s_{ka}} \|_F \) [26]. Hence (84) can be written as
\[ \frac{1}{\sigma_a^2} \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2 \]
\[ \leq 2 \left( 16 \sqrt{\frac{se'}{4p}} + 4 \frac{se}{4p} + 16 \sqrt{\frac{se'}{4p}} + 16 \frac{se}{4p} + 4 \sqrt{\frac{se'}{4p}} \frac{se'}{4p} \right) \]
\[ + 4 \frac{se}{4p} + 4 \sqrt{\frac{se'}{4p}} \frac{se'}{4p} \]
\[ \leq 130 \sqrt{\frac{se'}{4p}}, \] (88)

where (a) follows from (85). Since \( \lambda_{\min}(\Sigma_{k,l}) \geq \sigma^2 [16], \)
\[ \| \Sigma_{k,l}^{-1} - \Sigma_{k,l'}^{-1} \|_2 \leq 2 \| \Sigma_{k,l}^{-1} \|_2 \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2 \]
\[ \leq \frac{2}{\sigma^4} \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2. \] (89)

Now (76) can be stated as
\[ I(\mathbf{Y}; l||T(\mathbf{X})) \leq \frac{4N_a}{\sigma^4 L^2} \sum_l \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2^2 \]
\[ \leq \frac{4N_a}{\sigma^4} \| \Sigma_{k,l} - \Sigma_{k,l'} \|_2^2 \]
\[ \leq (130)^2 N_a^2 \left( \frac{\sigma_a}{\sigma} \right)^4 \frac{4e'}{p} \]
\[ = 16900 N_a^2 \left( \frac{\sigma_a}{\sigma} \right)^4 \frac{4e'}{p}, \] (90)

where (b) follow from (88). Thus, the proof is complete. \( \square \)