# SAMPLE COMPLEXITY BOUNDS FOR DICTIONARY LEARNING OF TENSOR DATA

*Zahra Shakeri, Waheed U. Bajwa, Anand D. Sarwate*

Dept. of Electrical and Computer Engineering, Rutgers, The State University of New Jersey

## ABSTRACT

This paper provides bounds on the sample complexity of estimating Kronecker-structured dictionaries for $K$th-order tensor data. The training samples are generated by linear combinations of these structured dictionary atoms and observed through white Gaussian noise. The lower bound follows from a lower bound on the minimax risk for general coefficient distributions and can be further specialized to sparse-Gaussian coefficients. This bound scales linearly with the sum of the product of the dimensions of the (smaller) coordinate dictionaries for tensor data. An explicit dictionary estimation algorithm for 2nd-order tensor data is also provided whose sample complexity matches the lower bound in the scaling sense. Numerical experiments highlight the advantages associated with explicitly accounting for tensor structure of data during dictionary learning.

***Index Terms—*** Kronecker-structured dictionary learning, minimax bounds, sparse representations, tensor data.

## 1. INTRODUCTION

Dictionary learning is a technique for finding sparse representations of signals or data and has applications in various tasks, such as image denoising and inpainting [1], audio processing [2], and classification [3,4]. In data-driven learning of geometric structures, explicitly accounting for the structure of the data has been shown to be advantageous over traditional dictionary learning methods such as $K$-SVD [5–12]. Many real-world signals are high dimensional but may have a simpler latent structure; dictionary learning that uses this structure can yield more efficient representations and subsequent processing. Although there has been prior works that empirically demonstrate the effectiveness of structured learning techniques [6–12], our goal in this paper is to prove that taking advantage of the data's tensor structure can yield more sample-efficient dictionary learning.

In this paper, we focus on the Tucker model [13] for tensor data and provide lower bounds on the minimax risk of estimating Kronecker-structured (KS) dictionaries consisting of $K \geq 2$ coordinate dictionaries that sparsely represent $K$th-order tensor data. Our approach uses the standard procedure for lower bounding the minimax risk in nonparametric estimation by connecting it to the maximum probability of error on a carefully constructed multiple hypothesis testing problem [14, 15]: the technical challenge is in finding the right hypotheses. In particular, consider a dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ consisting of the Kronecker product of $K$ coordinate dictionaries $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, k \in \{1, \dots, K\}$, where $m = \prod_{k=1}^K m_k$ and $p = \prod_{k=1}^K p_k$, that is generated within the radius $r$ neighborhood (taking the Frobenius norm as the distance metric) of a fixed reference dictionary. Then, our analysis shows that given a sufficiently large $r$ and keeping some other parameters constant, a sample complexity of $N = \Omega(\sum_{k=1}^K m_k p_k)$ is necessary for reconstruction of the true dictionary up to a given estimation error. Our second contribution is development and analysis of an algorithm to learn dictionaries formed by the Kronecker product of 2 smaller dictionaries that can be used to represent 2nd-order tensor data. To this end, we show that under certain conditions on the local neighborhood, the proposed algorithm can achieve one of the earlier obtained minimax lower bounds (in terms of scaling). While not a complete converse, this result suggests our lower bounds may be tight in more general settings. Furthermore, we demonstrate the performance of the proposed algorithm via numerical experiments.

Prior theoretical studies of dictionary learning have either focused on existing algorithms for non-KS dictionaries [5, 16–21] or lower bounds on minimax risk of dictionary learning for vector-valued data [22, 23]. In particular, Jung et al. [22, 23] provide minimax lower bounds for dictionary learning from vector-valued data under several coefficient vector distributions and discuss a regime where the bounds are tight for some signal-to-noise (SNR) values. In the case of a given (unstructured) dictionary $\mathbf{D}$ and sufficiently large neighborhood radius $r$, they show that $N = \Omega(mp)$ samples are required for reliable recovery of the dictionary up to a prescribed mean squared error (MSE). This is in contrast to our $N = \Omega(\sum_{k=1}^K m_k p_k)$ bound for learning of structured dictionaries, which matches the essential number of degrees of freedom in a KS dictionary and also generalizes our previous construction [24] for $K = 2$ to general $K$. Further, we provide an algorithm that matches our lower bound for $K = 2$. We conclude by noting that many technical details are omitted in this paper due to space constraints; nonetheless, full details can be found in our journal preprint [25].

***Notation Conventions:*** 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 $\mathbf{X}$ is denoted by $\mathbf{x}_k$ and $\mathbf{x}_\mathcal{I}$ denotes the vector consisting of the elements of $\mathbf{x}$ with indices $\mathcal{I}$. Let $\mathbf{I}_d$ be the $d \times d$ identity matrix. Norms are given by subscripts, so $\|\mathbf{v}\|_0$ and $\|\mathbf{v}\|_2$ are the $\ell_0$ and $\ell_2$ norms of $\mathbf{v}$, while $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ are the spectral and Frobenius norms of $\mathbf{X}$. We write $[K]$ for $\{1, \dots, K\}$. We write $\mathbf{X}_1 \otimes \mathbf{X}_2$ for the *Kronecker product* of two matrices $\mathbf{X}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{p \times q}$: the result is an $mp \times nq$ matrix. For matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, we define their distance to be $\|\mathbf{X}_1 - \mathbf{X}_2\|_F$. We write $\text{vec}(\underline{\mathbf{Y}})$ for the vectorization of a tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$. We use the standard "big-$\mathcal{O}$" notation for asymptotic scaling.

## 2. PROBLEM FORMULATION

We model our multidimensional (training) signals as $K$th-order tensors $\underline{\mathbf{Y}}_j \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$. According to the Tucker model [13], given *coordinate dictionaries* $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}$, a *coefficient tensor* $\underline{\mathbf{X}}_j \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_K}$, and a *noise tensor* $\underline{\mathbf{N}}_j = \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$,

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.

we can write $\mathbf{y}_j \triangleq \mathrm{vec}(\underline{\mathbf{Y}}_j)$ as

$$\mathbf{y}_j = \Big(\bigotimes_{k\in[K]} \mathbf{D}_k\Big)\mathbf{x}_j + \mathbf{n}_j, \tag{1}$$

where $\mathbf{x}_j \triangleq \mathrm{vec}(\underline{\mathbf{X}}_j)$ and $\mathbf{n}_j \triangleq \mathrm{vec}(\underline{\mathbf{N}}_j)$. Let $m = \prod_{k\in[K]} m_k$ and $p = \prod_{k\in[K]} p_k$. Concatenating $N$ independent and identically distributed (i.i.d) noisy observations $\{\mathbf{y}_j\}_{j=1}^N$, which are realizations according to the model (1), into $\mathbf{Y} \in \mathbb{R}^{m\times N}$, we obtain

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N}, \tag{2}$$

where $\mathbf{D} \triangleq \bigotimes_{k\in[K]} \mathbf{D}_k$ is the unknown KS dictionary, $\mathbf{X} \in \mathbb{R}^{p\times N}$ is a coefficient matrix consisting of i.i.d random coefficient vectors with known distribution that has zero-mean and covariance matrix $\mathbf{\Sigma}_x$, and $\mathbf{N} \in \mathbb{R}^{m\times N}$ is assumed to be additive white Gaussian noise (AWGN) with zero mean and variance $\sigma^2$.

We assume the true KS dictionary $\mathbf{D}$ consists of unit norm columns and we carry out local analysis around a *reference dictionary* $\mathbf{D}_0$. Specifically, let $\mathbf{D}_0 = \bigotimes_{k\in[K]} \mathbf{D}_{(0,k)} \in \mathcal{D}$ be a KS dictionary with $\|\mathbf{d}_{(0,k),i}\|_2 = 1$ for all $k \in [K]$ and $i \in [p_k]$, where

$$\mathcal{D} \triangleq \Big\{\mathbf{D}' \in \mathbb{R}^{m\times p} : \|\mathbf{d}'_i\|_2 = 1 \ \forall i \in [p], \mathbf{D}' = \bigotimes_{k\in[K]} \mathbf{D}'_k, \\ \mathbf{D}'_k \in \mathbb{R}^{m_k\times p_k} \ \forall k \in [K]\Big\} \tag{3}$$

and we assume the true generating KS dictionary $\mathbf{D}$ belongs to a neighborhood around $\mathbf{D}_0$:

$$\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r) \triangleq \big\{\mathbf{D}' \in \mathcal{D} : \big\|\mathbf{D}' - \mathbf{D}_0\big\|_F < r\big\} \tag{4}$$

for some fixed radius $r$. Note that the reference dictionary $\mathbf{D}_0$ appears in the analysis as an artifact of our proof technique to construct the dictionary class. In particular, if $r$ is sufficiently large, then $\mathcal{X}(\mathbf{D}_0, r) \approx \mathcal{D}$.

**Minimax Risk.** We are interested in lower bounding the minimax risk of estimating $\mathbf{D}$ based on observations $\mathbf{Y}$, which is defined as the worst-case mean squared error (MSE) that can be obtained by the best KS dictionary estimator $\widehat{\mathbf{D}}(\mathbf{Y})$. That is,

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}} \sup_{\mathbf{D}\in\mathcal{X}(\mathbf{D}_0,r)} \mathbb{E}_{\mathbf{Y}}\Big\{\big\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\big\|_F^2\Big\}, \tag{5}$$

where $\widehat{\mathbf{D}}(\mathbf{Y})$ can be estimated using any KS dictionary learning algorithm. In order to lower bound this minimax risk $\varepsilon^*$, we employ a standard reduction to the multiple hypothesis testing used in the literature on nonparametric estimation [14, 15]. This approach is equivalent to generating a KS dictionary $\mathbf{D}_l$ uniformly at random from a carefully constructed class $\mathcal{D}_L = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\} \subseteq \mathcal{X}(\mathbf{D}_0, r), L \geq 2$, for a given $(\mathbf{D}_0, r)$. A lower bound on the minimax risk in this setting depends not only on problem parameters such as the number of observations $N$, noise variance $\sigma^2$, dimensions $\{m_k\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$ of the true KS dictionary, neighborhood radius $r$, and coefficient covariance $\mathbf{\Sigma}_x$, but also on various aspects of the constructed class $\mathcal{D}_L$ [14]. To ensure a tight lower bound, we must construct $\mathcal{D}_L$ such that the distance between any two dictionaries in $\mathcal{D}_L$ is large but the hypothesis testing problem is hard; that is, two distinct dictionaries $\mathbf{D}_l$ and $\mathbf{D}_{l'}$ should produce similar observations. Specifically, for $l, l' \in [L]$, and given error $\varepsilon \geq \varepsilon^*$, we desire a construction such that for all $l \neq l'$

$$\|\mathbf{D}_l - \mathbf{D}_{l'}\|_F \geq 2\sqrt{\gamma\varepsilon}, \text{ and } D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}) || f_{\mathbf{D}_{l'}}(\mathbf{Y})\right) \leq \alpha_L,$$

where $D_{KL}\left(f_{\mathbf{D}_l}(\mathbf{Y}) || f_{\mathbf{D}_{l'}}(\mathbf{Y})\right)$ denotes the Kullback-Leibler (KL) divergence between the distributions of observations based on $\mathbf{D}_l \in \mathcal{D}_L$ and $\mathbf{D}_{l'} \in \mathcal{D}_L$, while $\gamma$, $\alpha_L$, and $\varepsilon$ are non-negative parameters. Observations $\mathbf{Y} = \mathbf{D}_l\mathbf{X} + \mathbf{N}$ in this setting can be interpreted as channel outputs that are used to estimate the input $\mathbf{D}_l$ using an arbitrary KS dictionary algorithm that is assumed to achieve the error $\varepsilon$. Our goal is to detect the correct generating KS dictionary index $l$ from estimated dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$. For this purpose, we use a minimum distance detector, $\widehat{l} = \min_{l'\in[L]} \big\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_{l'}\big\|_F$. We can then derive a lower bound on the probability of error using Fano's inequality [15], which involves the mutual information $I(\mathbf{Y}; l)$ between the observations $\mathbf{Y}$ and the dictionary $\mathbf{D}_l$. Since evaluating $I(\mathbf{Y}; l)$ is challenging, we adopt the same approach as Jung et al. [23] by assuming the decoder/estimator has access to some side information $\mathbf{T}(\mathbf{X})$ such that the conditional distribution of $\mathbf{Y}$ becomes multivariate Gaussian (recall that $I(\mathbf{Y}; l) \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$). Our final results then follow from the fact that any lower bound for $\varepsilon^*$ given the side information $\mathbf{T}(\mathbf{X})$ is also a lower bound for the general case [23]. Note that our final results are applicable to the global KS dictionary learning problem, since the minimax lower bounds that are obtained for any $\mathbf{D} \in \mathcal{X}(\mathbf{D}_0, r)$ are also trivially lower bounds for $\mathbf{D} \in \mathcal{D}$.

**Coefficient Distribution.** Our lower bounds hold under a generative model for the tensor data. We provide a bound for general coefficient distributions and a tighter bound (in some regimes) for the special case of sparse Gaussian coefficients.

*1. General Coefficients:* First, we consider the general case, where $\mathbf{x}$ is a zero-mean random coefficient vector with covariance matrix $\mathbf{\Sigma}_x = \mathbb{E}_{\mathbf{x}}\left\{\mathbf{x}\mathbf{x}^\top\right\}$. We make no additional assumption on the distribution of $\mathbf{x}$. We assume side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ to obtain a lower bound on the minimax risk in this case.

*2. Sparse Gaussian Coefficients:* We also consider sparse coefficient vectors and obtain a bound which only uses the support $\mathrm{supp}(\mathbf{x})$ (indices of nonzero entries of $\mathbf{x}$) as side information. In this case, the random support of $\mathbf{x}$ is assumed to be distributed uniformly over $\mathcal{E} = \{\mathcal{S} \subseteq [p] : |\mathcal{S}| = s\}$:

$$\mathbb{P}(\mathrm{supp}(\mathbf{x}) = \mathcal{S}) = \frac{1}{\binom{p}{s}}, \quad \text{for any } \mathcal{S} \in \mathcal{E}. \tag{6}$$

We further assume that the coefficient vectors generated according to (6) are i.i.d. Gaussian: $\mathbf{x}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2\mathbf{I}_s)$. As a result, conditioned on side information $\mathbf{T}(\mathbf{x}_k) = \mathrm{supp}(\mathbf{x}_k)$, observations $\mathbf{y}_k$ follow a multivariate Gaussian distribution. Parts of our forthcoming results also rely on the *restricted isometry property* (RIP) of a matrix:

*Restricted Isometry Property (*RIP*):* A matrix $\widehat{\mathbf{D}}$ with unit $\ell_2$ norm columns satisfies the RIP of order $s$ with constant $\delta_s$ if $(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|\widetilde{\mathbf{D}}\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2$ for all $\mathbf{x}$ such that $\|\mathbf{x}\|_0 \leq s$.

## 3. LOWER BOUND FOR GENERAL DISTRIBUTION

**Theorem 1.** *Consider a KS dictionary learning problem with $N$ i.i.d observations generated according to model* (1). *Suppose the true dictionary satisfies* (4) *for some $r$ and fixed reference dictionary $\mathbf{D}_0$. Then for any coefficient distribution with mean zero and covariance $\mathbf{\Sigma}_x$, we have the following lower bound on $\varepsilon^*$:*

$$\varepsilon^* \geq t \times \min\bigg\{\frac{p}{4}, \frac{r^2}{8K}, \frac{\sigma^2}{16NK\|\mathbf{\Sigma}_x\|_2}\bigg(c_1\Big(\sum_{k\in[K]} p_k(m_k - 1)\Big) \\ - \frac{K}{2}\log_2 2K - 2\bigg)\bigg\}, \tag{7}$$

*for any* $0 < t < 1$ *and any* $0 < c_1 < \frac{1-t}{8\log 2}$.

*Outline of Proof:* The idea of the proof is to construct a set of $L$ distinct KS dictionaries $\mathcal{D}_L = \{\mathbf{D}_1, \dots, \mathbf{D}_L\} \subset \mathcal{X}(\mathbf{D}_0, r)$ such that for any pair $l, l' \in [L]$ and any positive desired error $\varepsilon < \frac{tp}{4}\min\left\{r^2, \frac{r^4}{2Kp}\right\}$,

$$\|\mathbf{D}_l - \mathbf{D}'_l\|_F \geq 2\sqrt{2\varepsilon}, \text{ for } l \neq l'. \tag{8}$$

We select $\mathbf{D}_l \in \mathcal{D}_L$ uniformly from $\mathcal{D}_L$ and generate data according to (2). Given side information $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, the entries of $\mathbf{Y}$ are multivariate Gaussian; we can then upper bound the conditional mutual information $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ by upper bounding the KL divergence between multivariate Gaussians. Assuming (8) holds for $\mathcal{D}_L$, if there exists an estimator achieving the minimax risk $\varepsilon^* \leq \varepsilon$ and the recovered dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$ satisfies $\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F < \sqrt{2\varepsilon}$, the minimum distance detector can recover $\mathbf{D}_l$. Since $\varepsilon^*$ is bounded, using Markov's inequality, the probability of error $\mathbb{P}(\widehat{\mathbf{D}}(\mathbf{Y}) \neq \mathbf{D}_l) \leq \mathbb{P}(\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}_l\|_F \geq \sqrt{2\varepsilon})$ can be upper bounded by $\frac{1}{2}$. Then from Fano's inequality, $(1 - \mathbb{P}(\widehat{\mathbf{D}}(\mathbf{Y}) \neq \mathbf{D}_l))\log_2 L - 1 \leq I(\mathbf{Y}; l) \leq I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$, we find a lower bound on $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$. Using our bounds on the conditional MI, we can finally give a bound on $\varepsilon^*$ in terms of $L$. We use the constraint in (8) in the proof of Theorem 1 for simplicity: the constant $2\sqrt{2}$ can be replaced with any arbitrary $\gamma > 0$. The complete technical proof of Theorem 1 is given in [25]. Although the similarity of our model to that of [23] suggests that the proof should be a simple extension of the proof of Theorem 1 in Jung et al. [23], the hypotheses construction for KS dictionaries is more complex and its analysis requires a different approach.

## 4. LOWER BOUND FOR SPARSE GAUSSIAN DISTRIBUTION

**Theorem 2.** *Consider a KS dictionary learning problem with $N$ i.i.d observations generated according to model* (1). *Suppose the true dictionary satisfies* (4) *for some $r$ and fixed reference dictionary $\mathbf{D}_0$. If the reference coordinate dictionaries $\{\mathbf{D}_{0,k}, k \in [K]\}$ satisfy* $\mathsf{RIP}(s, \frac{1}{2})$ *and the random coefficient vector $\mathbf{x}$ is selected according to* (6) *with $\mathbf{x}_\mathcal{S} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2\mathbf{I}_s)$, we have the following bound on $\varepsilon^*$:*

$$\varepsilon^* \geq t \times \min\left\{\frac{p}{4s}, \frac{r^2}{8K}, \frac{\sigma^4 p}{144(3^{4K})\sigma_a^4 N s^2}\left(c_1\Big(\sum_{k\in[K]} p_k(m_k - 1)\Big) - \frac{K}{2}\log_2 2K - 2\right)\right\}, \tag{9}$$

*for any* $0 < t < 1$ *and any* $0 < c_1 < \frac{1-t}{8\log 2}$.

Note that in Theorem 2, $\mathbf{D}$ (or its coordinate dictionaries) need not satisfy the RIP condition. Rather, the RIP is only needed for the coordinate reference dictionaries, $\{\mathbf{D}_{0,k}, k \in [K]\}$, which is a significantly weaker (and possibly trivial to satisfy) condition. The proof of Theorem 2 is provided in [25].

## 5. PARTIAL CONVERSE

We now study a special case and introduce an algorithm that achieves the lower bound in Theorem 1 for 2nd-order tensors.

**Theorem 3.** *Consider a dictionary learning problem with $N$ i.i.d observations according to model* (1) *for $K = 2$ and let the true dictionary satisfy* (4) *for $\mathbf{D}_0 = \mathbf{I}_p$ and some $r > 0$. Further, assume the random coefficient vector $\mathbf{x}$ is selected according to* (6), *$\mathbf{x} \in \{-1, 0, 1\}^p$ and nonzero entries of $\mathbf{x}$ can have any distribution. Next, assume noise standard deviation $\sigma$ and express the KS dictionary as $\mathbf{D} = (\mathbf{I}_{p_1} + \mathbf{\Delta}_1) \otimes (\mathbf{I}_{p_2} + \mathbf{\Delta}_2)$, where $p = p_1p_2$, $\|\mathbf{\Delta}_1\|_F \leq r_1$ and $\|\mathbf{\Delta}_2\|_F \leq r_2$. Then, if the following inequalities are satisfied:*

$$\begin{aligned} r_1\sqrt{p_2} + r_2\sqrt{p_1} + r_1r_2 \leq r, \quad (r_1 + r_2 + r_1r_2)\sqrt{s} \leq 0.1 \\ \max\left\{\frac{r_1^2}{p_2}, \frac{r_2^2}{p_1}\right\} \leq \frac{1}{2N}, \quad \sigma \leq 0.4, \end{aligned} \tag{10}$$

*there exists a dictionary learning scheme whose MSE satisfies*

$$\mathbb{E}_\mathbf{Y}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_F^2\} \leq \frac{8p}{N}\left(\frac{\sigma^2(p_1m_1 + p_2m_2)}{s} + 3(p_1 + p_2)\right) + 8p\exp\left(-\frac{0.08pN}{\sigma^2}\right).$$

**KS Dictionary Learning Algorithm:** To prove Theorem 3, we describe an estimator that thresholds observations and applies an alternating update rule to learn the coordinate dictionaries. The analysis of this estimator (provided in [25]) is the proof of Theorem 3.

*Coefficient Update:* We utilize a simple thresholding technique for this purpose. Specifically, for all $j \in [N]$ we have:

$$\widehat{\mathbf{x}}_j = (\widehat{x}_{j,1}, \dots, \widehat{x}_{j,p})^\top, \quad \widehat{x}_{j,l} = \begin{cases} 1 & \text{if } y_{j,l} > 0.5, \\ -1 & \text{if } y_{j,l} < -0.5, \\ 0 & \text{otherwise.} \end{cases}$$

*Dictionary Update:* Denoting $\mathbf{A} \triangleq \mathbf{I}_{p_1} + \mathbf{\Delta}_1$ and $\mathbf{B} \triangleq \mathbf{I}_{p_2} + \mathbf{\Delta}_2$, we can write $\mathbf{D} = \mathbf{A} \otimes \mathbf{B}$. We update the columns of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ separately. To learn $\widehat{\mathbf{A}}$, we take advantage of the Kronecker structure of the dictionary and divide each observation $\mathbf{y}_j \in \mathbb{R}^{p_1p_2}$ into $p_2$ observations $\mathbf{y}'_{(j,k_2)} \in \mathbb{R}^{p_1}$:

$$\mathbf{y}'_{(j,k_2)} = \{y_{j,p_2i+k_2}\}_{i=0}^{p_1-1}, \ k_2 = [p_2], \ j = [N]. \tag{11}$$

This increases the number of observations to $Np_2$. We also divide the original and estimated coefficient vectors and the noise vectors:

$$\begin{aligned} &\mathbf{x}'_{(j,k_2)} = \{x_{j,p_2i+k_2}\}_{i=0}^{p_1-1}, \ \widehat{\mathbf{x}}'_{(j,k_2)} = \{\widehat{x}_{j,p_2i+k_2}\}_{i=0}^{p_1-1}, \\ &\mathbf{n}'_{(j,k_2)} = \{n_{j,p_2i+k_2}\}_{i=0}^{p_1-1}, \ k_2 = [p_2], \ j = [N]. \end{aligned} \tag{12}$$

To update columns of $\widehat{\mathbf{B}}$, we follow a different procedure to divide the observations. Specifically, we divide each observation $\mathbf{y}_j \in \mathbb{R}^{p_1p_2}$ into $p_1$ observations $\mathbf{y}''_{(j,k_1)} \in \mathbb{R}^{p_2}$:

$$\mathbf{y}''_{(j,k_1)} = \left\{y_{j,i+p_1(k_1-1)}\right\}_{i=1}^{p_2}, \ k_1 = [p_1], \ j = [N]. \tag{13}$$

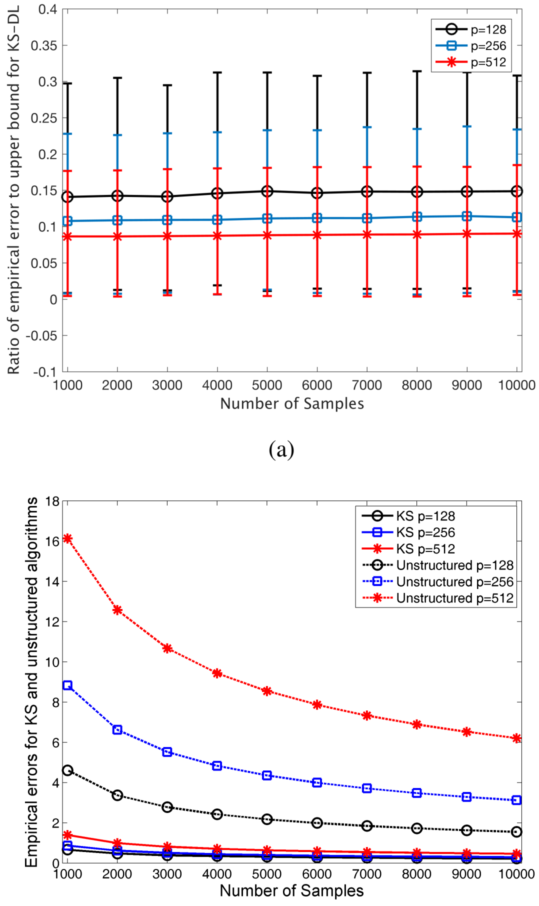
This increases the number of observations to $Np_1$. The coefficient vectors and noise vectors are also divided similarly:

$$\begin{aligned} &\mathbf{x}''_{(j,k_1)} = \left\{x_{j,i+p_1(k_1-1)}\right\}_{i=0}^{p_1-1}, \ \widehat{\mathbf{x}}''_{(j,k_1)} = \left\{\widehat{x}_{j,i+p_1(k_1-1)}\right\}_{i=0}^{p_1-1}, \\ &\mathbf{n}''_{(j,k_1)} = \left\{n_{j,i+p_1(k_1-1)}\right\}_{i=1}^{p_2}, \ k_1 = [p_1], \ j = [N]. \end{aligned} \tag{14}$$

The recovered dictionary in this case is $\widehat{\mathbf{D}} = \widehat{\mathbf{A}} \otimes \widehat{\mathbf{B}}$ and the update rules for columns of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are:

$$\widetilde{\mathbf{a}}_{l_1} = \frac{p_1}{Ns}\sum_{j=1}^{N}\sum_{k_2=1}^{p_2} \widehat{x}'_{(j,k_2),l_1}\mathbf{y}'_{(j,k_2)}, \quad \widehat{\mathbf{a}}_{l_1} = P_{\mathcal{B}_1}(\widetilde{\mathbf{a}}_{l_1}), \ l_1 \in [p_1],$$

$$\widetilde{\mathbf{b}}_{l_2} = \frac{p_2}{Ns}\sum_{j=1}^{N}\sum_{k_1=1}^{p_1} \widehat{x}''_{(j,k_1),l_2}\mathbf{y}''_{(j,k_1)}, \quad \widehat{\mathbf{b}}_{l_2} = P_{\mathcal{B}_1}(\widetilde{\mathbf{b}}_{l_2}), \ l_2 \in [p_2],$$



(a)

(b)

**Fig. 1**: Performance summary of KS dictionary learning algorithm for $p = \{128, 256, 512\}$, $s = 5$ and $r = 0.1$. (a) plots the ratio of the empirical error of our KS dictionary learning algorithm to the obtained error upper bound along with error bars and (b) shows the performance of our KS dictionary learning algorithm compared to the unstructured learning algorithm proposed in [23].

where $P_{\mathcal{B}_1}(\cdot)$ denotes projection on the closed unit ball and ensures that $\|\widehat{\mathbf{a}}_{l_1}\|_2 \leq 1$ and $\|\widehat{\mathbf{b}}_{l_2}\|_2 \leq 1$. Note that although projection onto the closed unit ball does not ensure the columns of $\widehat{\mathbf{D}}$ will have unit norms, our analysis only imposes this condition on the generating dictionary and the reference dictionary, and not on the final recovered dictionary.

## 6. NUMERICAL EXPERIMENTS

We implemented the preceding estimation algorithm for 2nd-order tensor data. Figure 1a shows the ratio of the empirical error of the proposed KS dictionary learning algorithm to the obtained upper bound in Theorem 3 for 50 Monte Carlo experiments. This ratio is plotted as a function of the sample size for three choices of the number of columns $p$: 128, 256, and 512. The experiment shows that the ratio is approximately constant as a function of sample size, verifying the theoretical result that the estimator meets the minimax bound in terms of error scaling as a function of sample size. Figure 1b shows the performance of our KS dictionary learning algorithm compared to the unstructured dictionary learning algorithm provided in [23]. It is evident that the error of our algorithm is significantly less than that for the unstructured algorithm for all choices of $p$. This verifies that taking the structure of data into consideration can indeed lead to lower dictionary identification error.

## 7. DISCUSSION AND CONCLUSION

**Table 1**: Order-wise lower bounds on the minimax risk

| **Dictionary**<br>**Distribution** | Unstructured | KS (this paper) |
|---|---|---|
| 1. General | $\dfrac{\sigma^2 mp}{N\|\mathbf{\Sigma}_x\|_2}$ | $\dfrac{\sigma^2(\sum_{k\in[K]} m_k p_k)}{NK\|\mathbf{\Sigma}_x\|_2}$ |
| 3. Gaussian Sparse | $\dfrac{p^2}{Nm\,\mathrm{SNR}^2}$ | $\dfrac{p(\sum_{k\in[K]} m_k p_k)}{3^{4K} Nm^2\,\mathrm{SNR}^2}$ |

In this paper we first gave a lower bound for the worst-case mean-squared error (MSE) in learning Kronecker-structured (KS) dictionaries that generate $K$th-order tensor data. Table 1 summarizes the lower bounds on the minimax rates from [23] and this work. The bounds are given in terms of the number of coordinate dictionaries $K$, the dictionary size parameters ($m_k$'s and $p_k$'s), the coefficient distribution parameters, the number of samples $N$, and SNR, which is defined as $\mathrm{SNR} = \dfrac{\mathbb{E}_{\mathbf{x}}\left\{\|\mathbf{x}\|_2^2\right\}}{\mathbb{E}_{\mathbf{n}}\left\{\|\mathbf{n}\|_2^2\right\}} = \dfrac{\mathrm{Tr}(\mathbf{\Sigma}_x)}{m\sigma^2}$. These scaling results hold for sufficiently large $p$ and neighborhood radius $r$. Compared to the results for the unstructured dictionary learning problem [23], we are able to decrease the lower bound for various coefficient distributions by reducing the scaling $\Omega(mp)$ to $\Omega(\sum_{k\in[K]} m_k p_k)$, which is the number of degrees of freedom in a KS dictionary. The risk decreases with the number of samples $N$ and the tensor order $K$; larger $K$ for fixed $mp$ means assuming more structure, thereby simplifying the problem.

The general coefficient results in the first row of Table 1 show that the minimax risk scales like $1/\,\mathrm{SNR}$ (since $\|\mathbf{\Sigma}_x\|_2 \leq \mathrm{Tr}(\mathbf{\Sigma}_x)$). For the sparse Gaussian coefficient results in the second row, we assume less side information for the lower bound but require that the reference coordinate dictionaries satisfy $\mathsf{RIP}(s, 1/2)$. This additional assumption has two implications: (1) it introduces the factor of $1/3^{4K}$ in the minimax lower bound, and (2) it imposes the following condition on the sparsity model: $s \leq \min_{k\in[K]}\{p_k\}$. Nonetheless, the minimax lower bound is tighter for sparse Gaussian coefficients than general coefficients for some SNR regimes.

We also provided a simple KS dictionary learning algorithm in Section 5 for $K = 2$ and analyzed its MSE $\mathbb{E}\{\|\widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D}\|_F^2\}$. In terms of scaling, the upper bound obtained for the MSE in Theorem 3 matches the lower bound in Theorem 1 provided $p_1 + p_2 < \dfrac{m_1 p_1 + m_2 p_2}{m\,\mathrm{SNR}}$ holds. This result suggests that more general KS dictionary learning algorithms may be developed to achieve the lower bounds reported in this paper.

Finally, while our analysis is local in the sense that we assume the true dictionary belongs in a local neighborhood with known radius around a fixed reference dictionary, the derived minimax risk lower bounds effectively become independent of this radius for sufficiently neighborhood radius. The full version of this work can be found in our journal preprint [25].

## 8. REFERENCES

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein, "$K$-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[2] Roger Grosse Rajat Raina, Helen Kwong, and Andrew Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. 23rd Conf. Uncertainty in Artificial Intelligence (UAI2007)*, July 2007.

[3] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng, "Self-taught learning:ttransfer learning from unlabeled data," in *Proc. 24th Int. Conf. Machine Learning*. ACM, June 2007, pp. 759–766.

[4] Julien Mairal, Francis Bach, and Jean Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Analys. and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, April 2012.

[5] Michal Aharon, Michael Elad, and Alfred M Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear Algebra and its Applicat.*, vol. 416, no. 1, pp. 48–67, July 2006.

[6] Zemin Zhang and Shuchin Aeron, "Denoising and completion of 3D data via multidimensional dictionary learning," *arXiv preprint arXiv:1512.09227*, December 2015.

[7] Guifang Duan, Hongcui Wang, Zhenyu Liu, Junping Deng, and Yen-Wei Chen, "K-CPD: Learning of overcomplete dictionaries for tensor sparse coding," in *Proc. IEEE 21st Int. Conf. Pattern Recognition (ICPR)*, November 2012, pp. 493–496.

[8] Simon Hawe, Matthias Seibert, and Martin Kleinsteuber, "Separable dictionary learning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, June 2013, pp. 438–445.

[9] Syed Zubair and Wenwu Wang, "Tensor dictionary learning with sparse Tucker decomposition," in *Proc. IEEE 18th Int. Conf. Digital Signal Process. (DSP)*, July 2013, pp. 1–6.

[10] Florian Roemer, Giovanni Del Galdo, and Martin Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, May 2014, pp. 3963–3967.

[11] Yi Peng, Deyu Meng, Zongben Xu, Chenqiang Gao, Yi Yang, and Biao Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, June 2014, pp. 2949–2956.

[12] Sara Soltani, Misha E Kilmer, and Per Christian Hansen, "A tensor-based dictionary learning approach to tomographic image reconstruction," *BIT Numerical Mathematics*, pp. 1–30, 2015.

[13] Ledyard R Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *Problems in Measuring Change*, pp. 122–137, 1963.

[14] Alexandre B Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, NJ USA, 2009.

[15] Bin Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

[16] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. 27th Annu. Conf. Learning Theory*, 2014, vol. 35 of *JMLR: Workshop and Conf. Proc.*, pp. 1–15.

[17] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli, "A clustering approach to learn sparsely-used overcomplete dictionaries," *arXiv preprint arXiv:1309.1952.*, July 2014.

[18] Sanjeev Arora, Rong Ge, and Ankur Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 25th Annu. Conf. Learning Theory*, 2014, vol. 35 of *JMLR: Workshop and Conf. Proc.*, pp. 1–28.

[19] Karin Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Appl. and Computational Harmonic Anal.*, vol. 37, no. 3, pp. 464–491, November 2014.

[20] Karin Schnass, "Local identification of overcomplete dictionaries," *J. Machine Learning Research*, vol. 16, pp. 1211–1242, June 2015.

[21] Rémi Gribonval, Rodolphe Jenatton, and Francis Bach, "Sparse and spurious: Dictionary learning with noise and outliers," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6298–6319, November 2015.

[22] Alexandra Jung, Yonina C Eldar, and Norbert Görtz, "Performance limits of dictionary learning for sparse coding," in *Proc. IEEE 22nd European Signal Process. Conf. (EUSIPCO)*, September 2014, pp. 765–769.

[23] Alexander Jung, Yonina C Eldar, and Norbert Görtz, "On the minimax risk of dictionary learning," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1501–1515, March 2015.

[24] Zahra Shakeri, Waheed U. Bajwa, and Anand D. Sarwate, "Minimax lower bounds for Kronecker-structured dictionary learning," in *Proc. 2016 IEEE Int. Symp. Inf. Theory*, July 2016, pp. 1148–1152.

[25] Zahra Shakeri, Waheed U Bajwa, and Anand D Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *arXiv preprint arXiv:1608.02792*, August 2016.