# Minimax Lower Bounds for Dictionary Learning from Tensor Data

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

***Abstract*—This paper provides lower bounds on the sample complexity of estimating Kronecker-structured dictionaries for $K$th-order tensor data. The results suggest the sample complexity of dictionary learning for tensor data can be significantly lower than that for unstructured data.**

## I. Introduction

Dictionary learning (DL) is a powerful feature learning technique that enables sparse representations of data and facilitates subsequent processing [1], [2]. DL involves construction of an overcomplete basis, $\mathbf{D}$, from input training data such that each data sample can be described by a small number of columns of $\mathbf{D}$. In the modern era of cheap and ubiquitous sensing, much of today's real-world data is being collected in a multi-modal manner. The collected data in this case has a *tensor* structure, defined as a multiway array [3]. Because of the inherent correlation across multiple dimensions of tensor data, it is important to explicitly account for the tensor structure within data-driven feature learning. The traditional DL literature, however, often ignores the tensor structure and resorts to conversion of multidimensional data into one-dimensional samples through vectorization of training data. Since such approaches ignore the tensor data structure, they result in sub-optimal sparse representations.

In this paper, we consider *Kronecker-structured* (KS) dictionaries that account for the tensor structure of data and provide lower bounds on the minimax risk of estimating KS dictionaries from tensor data using any estimator. These bounds not only help us understand potential advantages of explicitly accounting for the tensor structure of data in DL algorithms, but they also help quantify the performance of existing KS-DL algorithms [4], [5]. In particular, we show that reliable estimation of a dictionary that is the Kronecker product of $K$ coordinate dictionaries (and thus represents $K$th-order tensor data) requires the number of samples to scale linearly with the sum of the product of the dimensions of the coordinate dictionaries.

## II. Problem Formulation

According to the Tucker decomposition [6], we can model a tensor observation $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ as $\mathrm{vec}(\underline{\mathbf{Y}}) = \left(\bigotimes_{k=1}^{K} \mathbf{D}_k\right) \mathrm{vec}(\underline{\mathbf{X}}) + \mathrm{vec}(\underline{\mathbf{N}})$, where $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ denotes the random coefficient tensor with a known zero-mean distribution with covariance matrix $\mathbf{\Sigma}_x$, $\{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}\}_{k=1}^{K}$ are the coordinate dictionaries (factor matrices), $\underline{\mathbf{N}} = \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_K}$ denotes the additive white Gaussian noise tensor with zero mean and variance $\sigma^2$, and $\bigotimes$ denotes the Kronecker product [7]. Note that while the tensor observations are being vectorized in this setup, their structure is being preserved through the KS dictionary. We assume that the unknown KS dictionary, $\mathbf{D} \in \mathbb{R}^{m \times p} = \bigotimes_{k=1}^{K} \mathbf{D}_k$, $m = \prod_{k=1}^{K} m_k$, $p = \prod_{k=1}^{K} p_k$, has unit norm columns and belongs to a local neighborhood around a reference KS dictionary $\mathbf{D}_0$ with unit norm columns, i.e., $\|\mathbf{D} - \mathbf{D}_0\|_F < r$. Given $N$ tensor observations, concatenating the vectorized observations and coefficient tensors into matrices $\mathbf{Y}$ and $\mathbf{X}$, our goal is to lower bound the minimax risk of estimating $\mathbf{D}$ based on $\mathbf{Y}$, defined as the worst-case mean squared error that can be obtained by the best KS dictionary estimator $\widehat{\mathbf{D}}(\mathbf{Y})$:

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}} \sup_{\|\mathbf{D}-\mathbf{D}_0\|<r} \mathbb{E}_{\mathbf{Y}} \left\{ \left\| \widehat{\mathbf{D}}(\mathbf{Y}) - \mathbf{D} \right\|_F^2 \right\}. \tag{1}$$

For this purpose, we use a standard reduction to the multiple hypothesis testing problem [8]. We construct a set of $L$ distinct KS dictionaries with unit-norm columns such that any two distinct dictionaries are separated by some distance that is a function of the desired error. We then use Fano's inequality [9], which requires an upper bound on the mutual information between $\mathbf{Y}$ and the true dictionary index $l \in \{1, \ldots, L\}$. Since evaluating $I(\mathbf{Y}; l)$ is challenging, we assume the decoder/estimator has access to some side information (SI) [10] $\mathbf{T}(\mathbf{X})$ such that the conditional distribution of $\mathbf{Y}$ becomes multivariate Gaussian and we then upper bound the conditional mutual information $I(\mathbf{Y}; l|\mathbf{T}(\mathbf{X}))$ by upper bounding the Kullback–Leibler (KL) divergence between multivariate Gaussians.

## III. Results and Discussion

Table I compares lower bounds on the minimax rates for various coefficient distributions when one ignores the tensor structure [10] and this work. The bounds are given in terms of tensor order $K$, coordinate dictionary size parameters ($m_k$'s and $p_k$'s), number of samples $N$, and SNR, which is defined as $\mathrm{SNR} = \frac{\mathbb{E}_{\mathbf{x}}\{\|\mathbf{x}\|_2^2\}}{\mathbb{E}_{\mathbf{n}}\{\|\mathbf{n}\|_2^2\}} = \frac{\mathrm{Tr}(\mathbf{\Sigma}_x)}{m\sigma^2}$. These scaling results hold for sufficiently large $p$ and neighborhood radius $r$. Compared to the results for the unstructured dictionary learning problem [10], we decrease the lower bound for various coefficient distributions by reducing the scaling $\Omega(\prod_{k\in[K]} m_k p_k)$ 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 larger $N$ and $K$; in particular, larger $K$ for fixed $mp$ means more structure, which simplifies the estimation problem. The results for "general coefficient" distribution in the first row of Table I are obtained using SI $\mathbf{T}(\mathbf{X}) = \mathbf{X}$ and show that the minimax risk scales like $1/\mathrm{SNR}$. In the second row, we assume the coefficients are strictly sparse and the non-zero entries follow a Gaussian distribution. In this case, we can obtain the minimax lower bound using less SI, $\mathbf{T}(\mathbf{X}) = \mathrm{supp}(\mathbf{X})$, where $\mathrm{supp}(\mathbf{X})$ denotes the indices of non-zero entries of $\mathbf{X}$; nonetheless, we do require here that the reference coordinate dictionaries satisfy the restricted isometry property [11], $\mathsf{RIP}(s, 1/2)$, where $s$ denotes the sparsity level of coefficients. This additional assumption of $\mathsf{RIP}$ introduces the factor of $1/3^{4K}$ in the minimax lower bound. Nevertheless, the minimax lower bound is tighter for sparse Gaussian coefficients than for general coefficients in some $\mathrm{SNR}$ regimes.

We conclude by noting that while our analysis is local, our derived lower bounds for the minimax risk effectively become independent of this constraint for sufficiently large neighborhood radius. We refer the readers to [12] for full version of this work.

This work is supported in part by the NSF under awards CCF-1525276 and CCF-1453073, and by the ARO under award W911NF-14-1-0295.

| Dictionary Structure / Coefficient Distribution | Side Information $\mathbf{T}(\mathbf{X})$ | Unstructured Dictionary [10] | KS Dictionary (this work) |
|---|---|---|---|
| 1. General Coefficients | $\mathbf{X}$ | $\frac{\sigma^2 mp}{N\|\mathbf{\Sigma}_x\|_2}$ | $\frac{\sigma^2(\sum_{k\in[K]} m_k p_k)}{NK\|\mathbf{\Sigma}_x\|_2}$ |
| 2. Gaussian Sparse Coefficients | $\mathrm{supp}(\mathbf{X})$ | $\frac{p^2}{Nm\,\mathrm{SNR}^2}$ | $\frac{p(\sum_{k\in[K]} m_k p_k)}{3^{4K} Nm^2\,\mathrm{SNR}^2}$ |

TABLE I: Order-wise lower bounds on the minimax risk for various coefficient distributions.

## REFERENCES


[1] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, February 2003.

[2] M. Aharon, M. Elad, and A. 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.

[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[4] F. Roemer, G. Del Galdo, and M. 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.

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

[6] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *Problems in measuring change*, pp. 122–137, 1963.

[7] C. F. Van Loan, "The ubiquitous Kronecker product," *J. Computational and Appl. Mathematics*, vol. 123, no. 1, pp. 85–100, November 2000.

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

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

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

[11] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[12] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *arXiv preprint arXiv:1608.02792*, August 2016.