1 Sample Complexity Bounds for Dictionary Learning from Vector- and Tensor-valued Data

Zahra Shakeri, Anand D. Sarwate, and Waheed U. Bajwa[a]

Dept. of Electrical and Computer Engineering, Rutgers University–New Brunswick, NJ 08854, USA

Abstract

During the last decade, dictionary learning has emerged as one of the most powerful methods for data-driven extraction of features from data. While the initial focus on dictionary learning had been from an algorithmic perspective, recent years have seen an increasing interest in understanding the theoretical underpinnings of dictionary learning. Many such results rely on the use of information-theoretic analytical tools and help understand the fundamental limitations of different dictionary learning algorithms. This chapter focuses on the theoretical aspects of dictionary learning and summarizes existing results that deal with dictionary learning from both vector-valued data and tensor-valued (i.e., multi-way) data, which are defined as data having multiple modes. These results are primarily stated in terms of lower and upper bounds on the sample complexity of dictionary learning, defined as the number of samples needed to identify or reconstruct the true dictionary underlying data from noiseless or noisy samples, respectively. Many of the analytical tools that help yield these results come from the information theory literature; these include restating the dictionary learning problem as a channel coding problem and connecting the analysis of minimax risk in statistical estimation to Fano's inequality. In addition to highlighting the effects of different parameters on the sample complexity of dictionary learning, this chapter also brings out the potential advantages of dictionary learning from tensor data and concludes with a set of open problems that remain unaddressed for dictionary learning.

1.1 Introduction

Modern machine learning and signal processing relies on finding meaningful and succinct representations of data. Roughly speaking, data representation entails transforming "raw" data from its original domain to another domain in which it can be processed more effectively and efficiently. In particular, the performance

[a] This work is supported in part by the National Science Foundation under awards CCF-1525276 and CCF-1453073, and by the Army Research Office under award W911NF-17-1-0546.

of any information processing algorithm is dependent on the representation it is built on [1]. There are two major approaches to data representation. In *model-based approaches*, a *predetermined* basis is used to transform data. Such a basis can be formed using predefined transforms such as the Fourier transform [2], wavelets [3], and curvelets [4]. The *data-driven approach* infers transforms from the data to yield efficient representations. Prior works on data representation show that data-driven techniques generally outperform model-based techniques as the learned transformations are tuned to the input signals [5, 6].

Since contemporary data are often high-dimensional and high-volume, we need efficient algorithms to manage them. In addition, rapid advances in sensing and data acquisition technologies in recent years have resulted in individual data samples or signals with *multimodal* structures. For example, a single observation may contain measurements from a 2D array over time, leading to a data sample with 3 modes. Such data are often termed *tensors* or multiway arrays [7]. Specialized algorithms can take advantage of this tensor structure to handle multimodal data more efficiently. These algorithms represent tensor data using fewer parameters compared to vector-valued data representation methods by means of tensor decomposition techniques [8–10], resulting in reduced computational complexity and storage costs [11–15].

In this chapter, we focus on data-driven representations. As data collection systems grow and proliferate, we will need efficient data representations for processing, storage, and retrieval. Data-driven representations have successfully been used for signal processing and machine learning tasks such as data compression, recognition, and classification [5, 16, 17]. From a theoretical standpoint, there are several interesting questions surrounding data-driven representations. Assuming there is an unknown generative model forming a "true" representation of data, these questions include: *i)* What algorithms can be used to learn the representation effectively? *ii)* How many data samples are needed to learn the representation? *iii)* What are the fundamental limits on the number of data samples needed to learn the representation? *iv)* How robust are the solutions addressing these questions to parameters such as noise and outliers? In particular, state-of-the-art data representation algorithms have excellent empirical performance but their nonconvex geometry makes analyzing them challenging.

The goal of this chapter is to provide a brief overview of some of the aforementioned questions for a class of data-driven representation methods known as *dictionary learning* (DL). Our focus here will be on both the vector-valued and tensor-valued (i.e., multidimensional/multimodal) data cases.

1.1.1 Dictionary Learning: A Data-driven Approach to Sparse Representations

Data-driven methods have a long history in representation learning and can be divided into two classes. The first class includes linear methods, which involve transforming (typically vector-valued) data using linear functions to exploit the latent structure in data [5, 18, 19]. From a geometrical point of view, these methods effectively learn a low-dimensional subspace and projection of data onto that subspace, given some constraints. Examples of classical linear approaches for vector-valued data include principal component analysis (PCA) [5], linear discriminant analysis (LDA) [18], and independent component analysis (ICA) [19].

The second class consists of nonlinear methods. Despite the fact that historically linear representations have been preferred over nonlinear methods because of ease of computational complexity, recent advances in available analytical tools and computational power have resulted in an increased interest in nonlinear representation learning. These techniques have enhanced performance and interpretability compared to linear techniques. In nonlinear methods, data is transformed into a higher dimensional space, in which it lies on a low dimensional manifold [6, 20–22]. In the world of nonlinear transformations, nonlinearity can take different forms. In manifold-based methods such as diffusion maps, data is projected onto a nonlinear manifold [20]. In kernel (non-linear) PCA, data is projected onto a subspace in a higher dimensional space [21]. Autoencoders encode data based on the desired task [22]. DL uses a union of subspaces as the underlying geometric structure and projects input data onto one of the learned subspaces in the union. This leads to sparse representations of the data, which can be represented in the form of an overdetermined matrix multiplied by a sparse vector [6]. Although nonlinear representation methods result in nonconvex formulations, we can often take advantage of the problem structure to guarantee the existence of a unique solution and hence an optimal representation.

Focusing specifically on DL, it is known to have slightly higher computational complexity in comparison to linear methods, but it surpasses their performance in applications such as image denoising and inpainting [6], audio processing [23], compressed sensing [24], and data classification [17, 25]. More specifically, given input training signals $\mathbf{y} \in \mathbb{R}^m$, the goal in DL is to construct a basis such that $\mathbf{y} \approx \mathbf{D}\mathbf{x}$. Here, $\mathbf{D} \in \mathbb{R}^{m \times p}$ is denoted as the dictionary that has unit-norm columns and $\mathbf{x} \in \mathbb{R}^p$ is the dictionary coefficient vector that has a few nonzero entries. While the initial focus in DL had been on algorithmic development for various problem setups, works in recent years have also provided fundamental analytical results that help us understand the fundamental limits and performance of DL algorithms for both vector-valued [26–33] and tensor-valued [12, 13, 15] data.

There are two paradigms in the DL literature: the dictionary can be assumed to be a complete or an overcomplete basis (effectively, a frame [34]). In both cases, columns of the dictionary span the entire space [27]; in complete dictionaries, the dictionary matrix is square ($m = p$), whereas in overcomplete dictionaries the matrix has more columns than rows ($m < p$). In general, overcomplete representations result in more flexibility to allow both sparse and accurate representations [6].

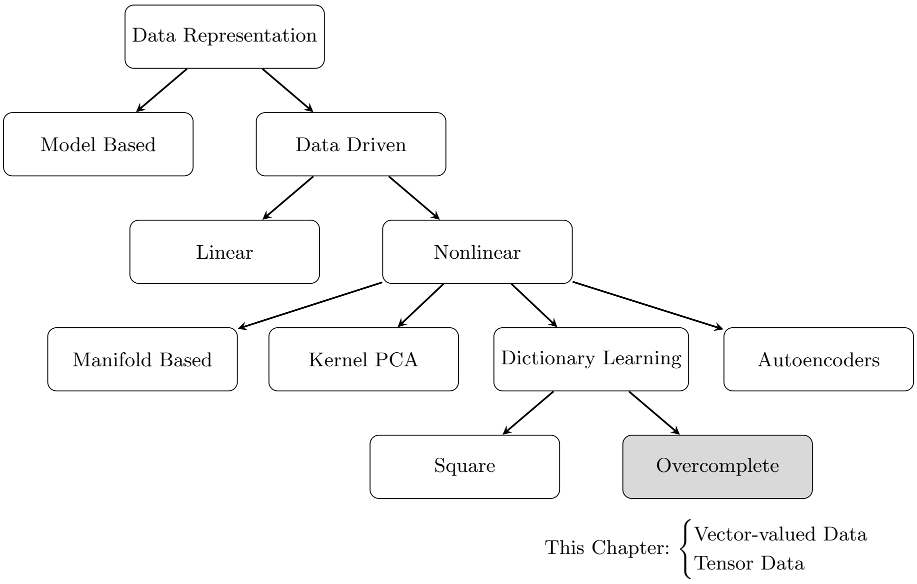

Figure 1.1 A graphical representation of the scope of this chapter in relation to the literature on representation learning.

1.1.2 Chapter Outline

In this chapter, we are interested in summarizing key results in learning of overcomplete dictionaries. We group works based on whether the data is vector-valued (one-dimensional) or tensor-valued (multidimensional). For both these cases, we focus on works that provide fundamental limits on the sample complexity for reliable dictionary estimation, i.e., the number of observations that are necessary to recover the true dictionary that generates the data up to some predefined error. The main information-theoretic tools that are used to derive these results range from reformulating the dictionary learning problem as a channel coding problem and connecting the minimax risk analysis to Fano's inequality. We refer the reader to Fig. 1.1 for a graphical overview of the relationship of this chapter to other themes in representation learning.

We address the DL problem for vector-valued data in Section 1.2, and for tensor data in Section 1.3. Finally, we talk about extensions of these works and some open problems in DL in Section 1.4. We focus here only on the problems of *identifiability* and *fundamental limits*; in particular, we do not survey DL algorithms in depth apart from some brief discussion in Sections 1.2 and 1.3. The monograph of Okoudjou [35] discusses algorithms for vector-valued data. Algorithms for tensor-valued data are relatively more recent and are described in our recent paper [13].

1.2 Dictionary Learning for Vector-valued Data

We first address the problem of reliable estimation of dictionaries underlying data that have a single mode, i.e., are vector valued. In particular, we focus on the subject of the sample complexity of the DL problem from two prospectives: *i)* fundamental limits on the sample complexity of DL using *any* DL algorithm, and *ii)* number of samples that are needed for different DL algorithms to reliably estimate a true underlying dictionary that generates the data.

1.2.1 Mathematical Setup

In the conventional vector-valued dictionary learning setup, we are given a total number N of vector-valued samples, $\{\mathbf{y}^n \in \mathbb{R}^m\}_{n=1}^N$, that are assumed to be generated from a fixed dictionary, $\mathbf{D}^0$, according to the following model:

$$\mathbf{y}^n = \mathbf{D}^0\mathbf{x}^n + \mathbf{w}^n, \quad n = 1, \dots, N. \tag{1.1}$$

Here, $\mathbf{D}^0 \in \mathbb{R}^{m\times p}$ is a (deterministic) unit-norm frame ($m < p$) that belongs to the following compact set[1]:

$$\mathbf{D}^0 \in \mathcal{D} \triangleq \left\{\mathbf{D} \in \mathbb{R}^{m\times p}, \|\mathbf{D}_j\|_2 = 1 \ \forall j \in \{1, \dots, p\}\right\}, \tag{1.2}$$

and is referred to as the *generating*, *true*, or *underlying* dictionary. The vector $\mathbf{x}^n \in \mathbb{R}^p$ is the *coefficient vector* that lies in some set $\mathcal{X} \subseteq \mathbb{R}^p$, and $\mathbf{w}^n \in \mathbb{R}^m$ denotes the observation noise. Concatenating the observations into a matrix $\mathbf{Y} \in \mathbb{R}^{m\times N}$, their corresponding coefficient vectors into $\mathbf{X} \in \mathbb{R}^{p\times N}$, and noise vectors into $\mathbf{W} \in \mathbb{R}^{m\times N}$, we get the following generative model:

$$\mathbf{Y} = \mathbf{D}^0\mathbf{X} + \mathbf{W}. \tag{1.3}$$

Various works in the DL literature impose different conditions on the coefficient vectors $\{\mathbf{x}^n\}$ to define the set $\mathcal{X}$. The most common assumption is that $\mathbf{x}^n$ is sparse with one of several probabilistic models for generating sparse $\mathbf{x}^n$. In contrast to exact sparsity, some works consider approximate sparsity and assume that $\mathbf{x}^n$ satisfies some decay profile [38], while others assume *group sparsity* conditions for $\mathbf{x}^n$ [39]. The latter condition comes up implicitly in DL for tensor data as we discuss in Section 1.3. Similarly, existing works consider a variety of noise models, the most common being Gaussian white noise. Regardless of the assumptions on coefficient and noise vectors, all of these works assume the observations are independent for $n = 1, 2, \dots, N$.

We are interested here in characterizing when it is possible to recover the true dictionary $\mathbf{D}^0$ from observations $\mathbf{Y}$. There is an inherent ambiguity in dictionary recovery: reordering the columns of $\mathbf{D}^0$ or multiplying any column by -1 yields a dictionary which can generate the same $\mathbf{Y}$ (with appropriately modified $\mathbf{X}$).

[1] A frame $\mathbf{F} \in \mathbb{R}^{m\times p}$, $m \leq p$, is defined as a collection of vectors $\{\mathbf{F}_i \in \mathbb{R}^m\}_{i=1}^p$ in some separable Hilbert space $\mathcal{H}$, that satisfy $c_1 \|\mathbf{v}\|_2^2 \leq \sum_{i=1}^p |\langle \mathbf{F}_i, \mathbf{v}\rangle|^2 \leq c_2 \|\mathbf{v}\|_2^2$ for all $\mathbf{v} \in \mathcal{H}$ and for some constants $0 < c_1 \leq c_2 < \infty$. If $c_1 = c_2$, then $\mathbf{F}$ is a tight frame [36, 37].

Thus, each dictionary is equivalent to $2^p p!$ other dictionaries. To measure the distance between dictionaries, we can either define the distance between equivalence classes of dictionaries or consider errors within a local neighborhood of a fixed $\mathbf{D}^0$, where the ambiguity can potentially disappear.

The specific criterion that we focus on is sample complexity, defined as the number of observations necessary to recover the true dictionary up to some predefined error. The measure of closeness of the recovered dictionary and the true dictionary can be defined in several ways. One approach is to compare the *representation error* of these dictionaries. Another measure is the mean squared error (MSE) between the estimated and generating dictionary, defined as

$$\mathbb{E}_{\mathbf{Y}} \left\{ d\left(\widehat{\mathbf{D}}(\mathbf{Y}), \mathbf{D}^0 \right)^2 \right\}, \tag{1.4}$$

where $d(\cdot,\cdot)$ is some distance metric, and $\widehat{\mathbf{D}}(\mathbf{Y})$ is the recovered dictionary according to observations $\mathbf{Y}$. For example, if we restrict the analysis to a local neighborhood of the generating dictionary, then we can use the Frobenius norm as the distance metric.

We now discuss an optimization approach to solving the dictionary recovery problem. Understanding the objective function within this approach is the key to understanding the sample complexity of DL. Recall that solving the DL problem involves using the observations to estimate a dictionary $\widehat{\mathbf{D}}$ such that $\widehat{\mathbf{D}}$ is close to $\mathbf{D}^0$. In the ideal case, the objective function involves solving the *statistical risk minimization* problem as follows:

$$\widehat{\mathbf{D}} \in \arg\min_{\mathbf{D}\in\mathcal{D}} \mathbb{E} \left\{ \inf_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \mathcal{R}(\mathbf{x}) \right\} \right\}. \tag{1.5}$$

Here, $\mathcal{R}(\cdot)$ is a regularization operator that enforces the pre-specified structure, such as sparsity, on the coefficient vectors. Typical choices for this parameter include functions of $\|\mathbf{x}\|_0$ or its convex relaxation, $\|\mathbf{x}\|_1$.[2] However, solving (1.5) requires knowledge of exact distributions of the problem parameters as well as high computational power. Hence, works in the literature resort to algorithms that solve the *empirical risk minimization* (ERM) problem [40]:

$$\widehat{\mathbf{D}} \in \arg\min_{\mathbf{D}\in\mathcal{D}} \left\{ \sum_{n=1}^{N} \inf_{\mathbf{x}^n\in\mathcal{X}} \left\{ \frac{1}{2} \|\mathbf{y}^n - \mathbf{D}\mathbf{x}^n\|_2^2 + \mathcal{R}(\mathbf{x}^n) \right\} \right\}. \tag{1.6}$$

In particular, to provide analytical results, many estimators solve this problem in lieu of (1.5) and then show that the solution of (1.6) is close to (1.5).

There are a number of computational algorithms that have been proposed to solve (1.6) directly for various regularizers, or indirectly using heuristic approaches. One of the most popular heuristic approaches is the K-SVD algorithm, which can be thought of as solving (1.6) with ℓ_0-norm regularization [6]. There are also other methods such as *method of optimal directions* (MOD) [41] and

[2] The so-called ℓ_0-norm counts the number of nonzero entries of a vector; it is not a norm.

online DL [25] that solve (1.6) with convex regularizers. While these algorithms have been known to perform well in practice, attention has shifted in recent years to theoretical studies to *i)* find the fundamental limits of solving the statistical risk minimization problem in (1.5), *ii)* determine conditions on objective functions like (1.6) to ensure recovery of the true dictionary, and *iii)* characterize the number of samples needed for recovery using either (1.5) or (1.6). In this chapter, we are also interested in understanding the sample complexity for the DL statistical risk minimization and ERM problems. We summarize such results in the existing literature for the statistical risk minimization of DL in Subsection 1.2.2 and for the ERM problem in Subsection 1.2.3. Because the measure of closeness or error differs between these theoretical results, the corresponding sample complexity bounds are different.

Remark 1.1 In this section, we assume that the data is available in a batch, centralized setting and the dictionary is deterministic. In the literature, DL algorithms have been proposed for other settings such as streaming data, distributed data, and Bayesian dictionaries [42–45]. Discussions of these scenarios is beyond the scope of this chapter. In addition, some works have looked at ERM problems that are different from (1.6). We briefly discuss these works in Section 1.4.

1.2.2 Minimax Lower Bounds on the Sample Complexity of DL

In this section, we study the fundamental limits on the accuracy of the dictionary recovery problem that is achievable by *any* DL method in the minimax setting. Specifically, we wish to understand the behavior of the *best estimator* that achieves the lowest *worst-case MSE* among all possible estimators. We define the error of such an estimator as the *minimax risk*, which is formally defined as:

$$\varepsilon^* = \inf_{\widehat{\mathbf{D}}(\mathbf{Y})} \sup_{\mathbf{D}\in\widetilde{\mathcal{D}}} \mathbb{E}_{\mathbf{Y}} \left\{ d\left(\widehat{\mathbf{D}}(\mathbf{Y}), \mathbf{D}\right)^2 \right\}. \tag{1.7}$$

Note that the minimax risk does not depend on any specific DL method and provides a lower bound for the error achieved by any estimator.

The first result we present pertains to lower bounds on the minimax risk, i.e., minimax lower bounds, for the DL problem using the Frobenius norm as the distance metric between dictionaries. The result is based on the following assumption:

A1.1 (Local recovery) The true dictionary lies in a neighborhood of a fixed, known reference dictionary,[3] $\mathbf{D}^* \in \mathcal{D}$, i.e., $\mathbf{D}^0 \in \widetilde{\mathcal{D}}$, where

$$\widetilde{\mathcal{D}} = \left\{ \mathbf{D} | \mathbf{D} \in \mathcal{D}, \|\mathbf{D} - \mathbf{D}^*\|_F \le r \right\}. \tag{1.8}$$

The range for the neighborhood radius r in (1.8) is $(0, 2\sqrt{p}]$. This conditioning

[3] The use of a reference dictionary is an artifact of the proof technique and for sufficiently large r, $\mathcal{D} \approx \widetilde{\mathcal{D}}$.

comes from the fact that for any $\mathbf{D}, \mathbf{D}' \in \mathcal{D}$, $\|\mathbf{D} - \mathbf{D}'\|_F \leq \|\mathbf{D}\|_F + \|\mathbf{D}'\|_F = 2\sqrt{p}$. By restricting dictionaries to this class, for small enough r, ambiguities that are a consequence of using the Frobenius norm can be prevented. We also point out that any lower bound on ε^* is also a lower bound on the global DL problem.

THEOREM 1.1 (Minimax lower bounds [33]) *Consider a DL problem for vector-valued data with N i.i.d. observations and true dictionary $\mathbf{D}$ satisfying assumption **A1.1** for some $r \in (0, 2\sqrt{p}]$. Then for any coefficient distribution with mean zero and covariance matrix $\mathbf{\Sigma}_x$, and white Gaussian noise with mean zero and variance σ^2, the minimax risk ε^* is lower bounded as*

$$\varepsilon^* \geq c_1 \min \left\{ r^2, \frac{\sigma^2}{N \|\mathbf{\Sigma}_x\|_2} (c_2 p(m-1) - 1) \right\}, \tag{1.9}$$

for some positive constants c_1 and c_2.

Theorem 1.1 holds for both square and overcomplete dictionaries. To obtain this lower bound on the minimax risk, a standard information-theoretic approach is taken in [33] to reduce the dictionary estimation problem to a multiple hypothesis testing problem. In this technique, given fixed $\mathbf{D}^*$ and r, and $L \in \mathbb{N}$, a packing $\mathcal{D}_L = \{\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^L\} \subseteq \widetilde{\mathcal{D}}$ of $\widetilde{\mathcal{D}}$ is constructed. The distance of the packing is chosen to ensure a tight lower bound on the minimax risk. Given observations $\mathbf{Y} = \mathbf{D}^l \mathbf{X} + \mathbf{W}$, where $\mathbf{D}^l \in \mathcal{D}_L$ and index l is chosen uniformly at random from $\{1, \ldots, L\}$, and any estimation algorithm that recovers a dictionary $\widehat{\mathbf{D}}(\mathbf{Y})$, a minimum distance detector can be used to find the recovered dictionary index $\widehat{l} \in \{1, \ldots, L\}$. Then, Fano's inequality can be used to relate the probability of error, i.e., $\mathbb{P}(\widehat{l}(\mathbf{Y}) \neq l)$, to the mutual information between observations and the dictionary (equivalently, the dictionary index l), i.e., $I(\mathbf{Y}; l)$ [46].

Let us assume that r is sufficiently large such that the minimizer of the left hand side of (1.9) is the second term. In this case, Theorem 1.1 states that to achieve any error $\varepsilon \geq \varepsilon^*$, we need the number of samples to be on the order of $N = \Omega\left(\frac{\sigma^2 mp}{\|\mathbf{\Sigma}_x\|_2 \varepsilon}\right)$.[4] Hence, the lower bound on the minimax risk of DL can be translated to a lower bound on the number of necessary samples, as a function of the desired dictionary error. This can further be interpreted as a lower bound on the sample complexity of the dictionary recovery problem.

We can also specialize this result to sparse coefficient vectors. Assume $\mathbf{x}^n$ has up to s nonzero elements and the random support of the nonzero elements of $\mathbf{x}^n$ is assumed to be uniformly distributed over the set $\{\mathcal{S} \subseteq \{1, \ldots, p\} : |\mathcal{S}| = s\}$, for $n = \{1, \ldots, N\}$. Assuming that the nonzero entries of $\mathbf{x}^n$ are i.i.d. with variance σ_x^2, we get $\mathbf{\Sigma}_x = (s/p)\sigma_x^2 \mathbf{I}_p$. Therefore, for sufficiently large r, the sample complexity scaling to achieve any error ε becomes $\Omega\left(\frac{\sigma^2 mp^2}{\sigma_x^2 s\varepsilon}\right)$. In this special case, it can be seen that in order to achieve a fixed error ε, the sample complex-

[4] We use $f(n) = \Omega(g(n))$ and $f(n) = \mathcal{O}(g(n))$ if for sufficiently large $n \in \mathbb{N}$, $f(n) > c_1 g(n)$ and $f(n) < c_2 g(n)$, respectively, for some positive constants c_1 and c_2.

ity scales with the number of degrees of freedom of the dictionary multiplied by number of dictionary columns, i.e., $N = \Omega(mp^2)$. There is also an inverse dependence on sparsity level s. Defining the signal-to-noise-ratio of the observations as $\text{SNR} = (s\sigma_x^2)/(m\sigma^2)$, this can be interpreted as an inverse relationship with SNR. Moreover, if all parameters except data dimension, m, are fixed, increasing m requires a linear increase in N. Evidently, this linear relation is limited by the fact that $m \leq p$ has to hold to maintain completeness or overcompleteness of the dictionary: increasing m by a large amount requires increasing p also.

While the tightness of this result remains an open problem, Jung et al. [33] have shown that for a special class of square dictionaries that are perturbations of the identity matrix, and for sparse coefficients following a specific distribution, this result is order-wise tight. In other words, a square dictionary that is perturbed from the identity matrix can be recovered from this sample size order. Although this result does not extend to overcomplete dictionaries, it suggests that the lower bounds may be tight.

Finally, while distance metrics that are invariant to dictionary ambiguities have been used for achievable overcomplete dictionary recovery results [30, 31], obtaining minimax lower bounds for DL using these distance metrics remains an open problem.

In this section, we discussed the number of *necessary* samples for reliable dictionary recovery (sample complexity lower bound). In the next subsection, we focus on achievability results, i.e., the number of *sufficient* samples for reliable dictionary recovery (sample complexity upper bound).

1.2.3 Achievability Results

The preceding lower bounds on minimax risk hold for any estimator or computational algorithm. However, the proofs do not provide an understanding of how to construct effective estimators and provide little intuition about the potential performance of practical estimation techniques. In this section, we direct our attention to explicit reconstruction methods and their sample complexities that ensure reliable recovery of the underlying dictionary. Since these *achievability* results are tied to specific algorithms that are guaranteed to recover the true dictionary, the sample complexity bounds from these results can also be used to derive upper bounds on the minimax risk. As we will see later, there remains a gap between the lower bound and the upper bound on the minimax risk. Alternatively, one can interpret the sample complexity lower bound and upper bound as the number of necessary samples and sufficient samples for reliable dictionary recovery, respectively. In the following, we only focus on *identifiability* results: the estimation procedures are not required to be computationally efficient.

One of the first achievability results for DL were derived in [27, 28] for square matrices. Since then, a number of works have been carried out for overcomplete DL involving vector-valued data [26, 29–32, 38]. These works differ from each other in terms of their assumptions on the true underlying dictionary, the dictionary coefficients, presence or absence of noise and outliers, reconstruction objective function, the distance metric used to measure the accuracy of the solution, and the local or global analysis of the solution. In this section, we summarize a few of these results based on various assumptions on the noise and outliers and provide a brief overview of the landscape of these results in Table 1.1. We begin our discussion with achievability results for DL for the case where $\mathbf{Y}$ is exactly given by $\mathbf{Y} = \mathbf{D}^0\mathbf{X}$, i.e., the noiseless setting.

Before proceeding, we provide a definition and an assumption that will be used for the rest of this section. We note that the constants that are used in the presented theorems change from one result to another.

(Worst-case coherence) For any dictionary $\mathbf{D} \in \mathcal{D}$, its worst-case coherence is defined as $\mu(\mathbf{D}) = \max_{i \neq j} |\langle \mathbf{D}_i, \mathbf{D}_j \rangle|$, where $\mu(\mathbf{D}) \in (0, 1)$ [36].

(Random support of sparse coefficient vectors) For any $\mathbf{x}^n$ that has up to s nonzero elements, the support of the nonzero elements of $\mathbf{x}^n$ is assumed to be distributed uniformly at random over the set $\{\mathcal{S} \subseteq \{1, \dots, p\} : |\mathcal{S}| = s\}$, for $n = \{1, \dots, N\}$.

Noiseless Recovery

We begin by discussing the first work that proves local identifiability of the overcomplete DL problem. The objective function that is considered in that work is

$$\left(\widehat{\mathbf{X}}, \widehat{\mathbf{D}}\right) = \underset{\mathbf{D} \in \mathcal{D}, \mathbf{X}}{\arg\min} \|\mathbf{X}\|_1 \text{ subject to } \mathbf{Y} = \mathbf{D}\mathbf{X}, \tag{1.10}$$

where $\|\mathbf{X}\|_1 \triangleq \sum_{i,j} |\mathbf{X}_{i,j}|$ denotes the sum of absolute values of $\mathbf{X}$.

This result is based on the following set of assumptions:

A2.1 (Gaussian random coefficients). The values of the nonzero entries of $\mathbf{x}^n$'s are independent Gaussian random variables with zero mean and common standard deviation $\sigma_x = \sqrt{p/sN}$.

A2.2 (Sparsity level). The sparsity level satisfies $s \leq \min\left\{c_1/\mu(\mathbf{D}^0), c_2 p\right\}$ for some constants c_1 and c_2.

THEOREM 1.2 (Noiseless, local recovery [29]) *There exist positive constants* c_1, c_2 *such that if assumptions **A2.1**–**A2.2** are satisfied for true* $(\mathbf{X}, \mathbf{D}^0)$, *then* $(\mathbf{X}, \mathbf{D}^0)$ *is a local minimum of* (1.10) *with high probability.*

The probability in this theorem depends on various problem parameters and implies that $N = \Omega\left(sp^3\right)$ samples are sufficient for the desired solution, i.e., true dictionary and coefficient matrix, to be locally recoverable. The proof of this theorem relies on studying the local properties of (1.10) around its optimal solution and does not require defining a distance metric.

We now present a result that is based on the use of a combinatorial algorithm, which can provably and exactly recover the true dictionary. The proposed algorithm solves the objective function is (1.6) with $\mathcal{R}(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$, where λ is the regularization parameter and the distance metric that is used is the column-wise distance. Specifically, for two dictionaries $\mathbf{D}^1$ and $\mathbf{D}^2$, their column-wise distance is defined as

$$d(\mathbf{D}_j^1, \mathbf{D}_j^2) = \min_{l \in \{-1,1\}} \left\| \mathbf{D}_j^1 - l\mathbf{D}_j^2 \right\|_2, \quad j \in \{1, \dots, p\}, \tag{1.11}$$

where $\mathbf{D}_j^1$ and $\mathbf{D}_j^2$ are the jth column of $\mathbf{D}^1$ and $\mathbf{D}^2$, respectively. This distance metric avoids the sign ambiguity among dictionaries belonging to the same equivalence class. To solve (1.6), Agarwal et al. provide a novel DL algorithm that consists of an initial dictionary estimation stage and an alternating minimization stage to update the dictionary and coefficient vectors [30]. The provided guarantees are based on using this algorithm to update the dictionary and coefficients. The forthcoming result is based on the following set of assumptions:

A3.1 (Bounded random coefficients). The nonzero entries of $\mathbf{x}^n$'s are drawn from a zero-mean unit-variance distribution and their magnitude satisfies $x_{\min} \leq |\mathbf{x}_i^n| \leq x_{\max}$.

A3.2 (Sparsity level). The sparsity level satisfies $s \leq \min\left\{c_1/\sqrt{\mu(\mathbf{D}^0)},\ c_2 m^{1/9}, c_3 p^{1/8}\right\}$ for some positive constants c_1, c_2, c_3 that depend on $x_{\min}$, $x_{\max}$, and the spectral norm of $\mathbf{D}^0$.

A3.3 (Dictionary assumptions). The true dictionary has bounded spectral norm, i.e., $\left\|\mathbf{D}^0\right\|_2 \leq c_4\sqrt{p/m}$, for some positive c_4.

THEOREM 1.3 (Noiseless, exact recovery [30]) *Consider a DL problem with N i.i.d. observations and assume that assumptions **A3.1**–**A3.3** are satisfied. Then, there exists a universal constant c such that for given $\eta > 0$, if*

$$N \geq c \left(\frac{x_{\max}}{x_{\min}}\right)^2 p^2 \log \frac{2p}{\eta}, \tag{1.12}$$

there exists a procedure consisting of an initial dictionary estimation stage and an alternating minimization stage such that after $T = \mathcal{O}(\log(\frac{1}{\varepsilon}))$ iterations of the second stage, with probability at least $1 - 2\eta - 2\eta N^2$, $d(\widehat{\mathbf{D}}_j, \mathbf{D}_j^0) \leq \varepsilon, \forall \varepsilon > 0, \forall j \in \{1, \dots, p\}$.

This theorem guarantees that the true dictionary can be recovered to an arbitrary precision given $N = \Omega(p^2 \log p)$ samples. This result is based on two steps. The first step is guaranteeing an error bound for the initial dictionary estimation step. This step involves using a clustering-style algorithm to approximate the dictionary columns. The second step is proving a local convergence result for the alternating minimization stage. This step involves improving estimates of the coefficient vectors and the dictionary through Lasso [47] and least-square steps, respectively. More details for this work can be found in [30].

While works in [29, 30] study the sample complexity of the overcomplete DL problem, they do not take noise into account. Next, we present works that obtain sample complexity results for noisy reconstruction of dictionaries.

Noisy Reconstruction

The next result we discuss is based on the following objective function:

$$\max_{\mathbf{D}\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N} \max_{|\mathcal{S}|=s} \|\mathbf{P}_{\mathcal{S}}(\mathbf{D})\mathbf{y}^n\|_2^2\,, \tag{1.13}$$

where $\mathbf{P}_{\mathcal{S}}(\mathbf{D})$ denotes projection of $\mathbf{D}$ onto the span of $\mathbf{D}_{\mathcal{S}} = \{\mathbf{D}_j\}_{j\in\mathcal{S}}$.[5] Here, the distance metric that is used is $d(\mathbf{D}^1, \mathbf{D}^2) = \max_{j\in\{1,\dots,p\}} \left\|\mathbf{D}_j^1 - \mathbf{D}_j^2\right\|_2$. In addition, the results are based on the following set of assumptions:

A4.1 (Unit-norm tight frame). The true dictionary is a unit-norm tight frame, i.e., for all $\mathbf{v}\in\mathbb{R}^m$ we have $\sum_{j=1}^{p}\left|\langle\mathbf{D}_j^0, \mathbf{v}\rangle\right|^2 = \frac{p\|\mathbf{v}\|_2^2}{m}$.

A4.2 (Lower isometry constant). The lower isometry constant of $\mathbf{D}^0$, defined as $\delta_s(\mathbf{D}^0) \triangleq \max_{|\mathcal{S}|\le s}\delta_{\mathcal{S}}(\mathbf{D}^0)$ with $1-\delta_{\mathcal{S}}(\mathbf{D}^0)$ denoting the minimal eigenvalue of ${\mathbf{D}_{\mathcal{S}}^0}^*\mathbf{D}_{\mathcal{S}}^0$, satisfies $\delta_s(\mathbf{D}^0)\le 1-\frac{s}{m}$.

A4.3 (Decaying random coefficients). The coefficient vector $\mathbf{x}^n$ is drawn from a symmetric decaying probability distribution ν on the unit sphere S^{p-1}.[6]

A4.4 (Bounded random noise). The vector $\mathbf{w}^n$ is a bounded random white noise vector satisfying $\|\mathbf{w}^n\|_2 \le M_w$ almost surely, $\mathbb{E}\{\mathbf{w}^n\} = \mathbf{0}$ and $\mathbb{E}\{\mathbf{w}^n\mathbf{w}^{n*}\} = \rho^2\mathbf{I}_m$.

A4.5 (Maximal projection constraint). Define $\mathbf{c}(\mathbf{x}^n)$ to be the non-increasing rearrangement of the absolute values of $\mathbf{x}^n$. Given a sign sequence $\mathbf{l}\in\{-1,1\}^p$ and a permutation operator $\pi : \{1,\dots,p\}\to\{1,\dots,p\}$, define $\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}^n)$ whose ith element is equal to $\mathbf{l}_i\mathbf{c}(\mathbf{x}^n)_{\pi(i)}$ for $i\in\{1,\dots,p\}$. There exists $\kappa > 0$ such that for $\mathbf{c}(\mathbf{x}^n)$ and $\mathcal{S}_\pi \triangleq \pi^{-1}(\{1,\dots,s\})$, we have

$$\nu\Bigg(\min_{\pi,\mathbf{l}}\left(\left\|\mathbf{P}_{\mathcal{S}_\pi}(\mathbf{D}^0)\mathbf{D}^0\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}^n)\right\|_2 - \max_{|\mathcal{S}|=s,\mathcal{S}\neq\mathcal{S}_\pi}\left\|\mathbf{P}_{\mathcal{S}}(\mathbf{D}^0)\mathbf{D}^0\mathbf{c}_{\pi,\mathbf{l}}(\mathbf{x}^n)\right\|_2\right) \ge 2\kappa + 2M_w\Bigg) = 1. \tag{1.15}$$

THEOREM 1.4 (Noisy, local recovery [38]) *Consider a DL problem with N i.i.d. observations and assume that assumptions **A4.1**–**A4.5** are satisfied. If for some $0 < q < 1/4$, the number of samples satisfies:*

$$2N^{-q} + N^{-2q} \le \frac{c_1\sqrt{1-\delta_s(\mathbf{D}^0)}}{\sqrt{s}\left(1 + c_2\sqrt{\log\left(\frac{c_3 p\binom{p}{s}}{c_4 s(1-\delta_s(\mathbf{D}^0))}\right)}\right)}, \tag{1.16}$$

[5] This objective function can be thought of as a manipulation of (1.6) with the ℓ_0-norm regularizer for the coefficient vectors. See [38, Equation 2] for more details.

[6] A probability measure ν on the unit sphere S^{p-1} is called symmetric if for all measurable sets $\mathcal{X}\subseteq S^{p-1}$, for all sign sequences $\mathbf{l}\in\{-1,1\}^p$ and all permutations $\pi : \{1,\dots,p\}\to\{1,\dots,p\}$, we have

$$\nu(\mathbf{l}\mathcal{X}) = \nu(\mathcal{X}),\ \text{where } \mathbf{l}\mathcal{X} = \{(\mathbf{l}_1\mathbf{x}_1,\dots,\mathbf{l}_p\mathbf{x}_p) : \mathbf{x}\in\mathcal{X}\}\,, \text{and}$$
$$\nu(\pi(\mathcal{X})) = \nu(\mathcal{X}),\ \text{where } \pi(\mathcal{X}) = \left\{\left(\mathbf{x}_{\pi(1)},\dots,\mathbf{x}_{\pi(p)}\right) : \mathbf{x}\in\mathcal{X}\right\}. \tag{1.14}$$

then, with high probability, there is a local maximum of (1.13) *within distance at most* $2N^{-q}$ *of* $\mathbf{D}^0$.

The constants c_1, c_2, c_3 and c_4 in Theorem 1.4 depend on the underlying dictionary, coefficient vectors, and the underlying noise. The proof of this theorem relies on the fact that for the true dictionary and its perturbations, the maximal response, i.e., $\left\|\mathbf{P}_{\mathcal{S}}(\widetilde{\mathbf{D}})\mathbf{D}^0\mathbf{x}^n\right\|_2$,[7] is attained for the set $\mathcal{S} = \mathcal{S}_\pi$ for most signals. A detailed explanation of the theorem and its proof can be found in the paper of Schnass [38].

In order to understand Theorem 1.4, let us set $q \approx \frac{1}{4} - \frac{\log p}{\log N}$. We can then understand this theorem as follows. Given $N/\log N = \Omega(mp^3)$, except with probability $\mathcal{O}(N^{-mp})$, there is a local minimum of (1.13) within distance $\mathcal{O}(pN^{-1/4})$ of the true dictionary. Moreover, since the objective function that is considered in this work is also solved for the K-SVD algorithm, this result gives an understanding of the performance of the K-SVD algorithm. Compared to results with $\mathcal{R}(\mathbf{x})$ being a function of the ℓ_1-norm [29,30], this result requires the true dictionary to be a tight frame. On the flip side, the coefficient vector in Theorem 1.4 is not necessarily sparse; instead, it only has to satisfy a decaying condition.

Next, we present a result obtained by Arora et al. [31] that is similar to that of Theorem 1.3 in the sense that it uses a combinatorial algorithm that can provably recover the true dictionary given noiseless observations. It further obtains dictionary reconstruction results for the case of noisy observations. The objective function considered in this work is similar to that of the K-SVD algorithm and can be thought of as (1.6) with $\mathcal{R}(\mathbf{x}) = \lambda \left\|\mathbf{x}\right\|_0$, where λ is the regularization parameter.

Similar to Agarwal et al. [30], Arora et al. [31] define two dictionaries $\mathbf{D}^1$ and $\mathbf{D}^2$ to be *column-wise* ε *close* if there exists a permutation π and $l \in \{-1, 1\}$ such that $\left\|\mathbf{D}^1_j - l\mathbf{D}^2_{\pi(j)}\right\|_2 \leq \varepsilon$. This distance metric captures the distance between equivalent classes of dictionaries and avoids the sign-permutation ambiguity. They propose a DL algorithm that first uses combinatorial techniques to recover the support of coefficient vectors, by clustering observations into overlapping clusters that use the same dictionary columns. To find these large clusters, a clustering algorithm is provided. Then, the dictionary is roughly estimated given the clusters, and the solution is further refined. The provided guarantees are based on using the proposed DL algorithm. In addition, the results are based on the following set of assumptions:

A5.1 (Bounded coefficient distribution). Nonzero entries of $\mathbf{x}^n$ are drawn from a zero-mean distribution and lie in $[-x_{\max}, -1] \cup [1, x_{\max}]$, where $x_{\max} = \mathcal{O}(1)$. Moreover, conditioned on any subset of coordinates in $\mathbf{x}^n$ being nonzero, nonzero values of $\mathbf{x}^n_i$ are independent from each other. Finally, the distribution has bounded 3-wise moments, i.e., the probability that $\mathbf{x}^n$ is nonzero

[7] $\widetilde{\mathbf{D}}$ can be $\mathbf{D}^0$ itself or some perturbation of $\mathbf{D}^0$.

in any subset $\mathcal{S}$ of 3 coordinates is at most c^3 times $\prod_{i\in\mathcal{S}} \mathbb{P}\left\{\mathbf{x}_i^n \neq 0\right\}$, where $c = \mathcal{O}(1)$.[8]

A5.2 (Gaussian noise). The $\mathbf{w}^n$'s are independent and follow a spherical Gaussian distribution with standard deviation $\sigma = o(\sqrt{m})$.

A5.3 (Dictionary coherence). The true dictionary is $\widetilde{\mu}$-incoherent, that is, for all $i \neq j$, $\langle \mathbf{D}_i^0, \mathbf{D}_j^0 \rangle \leq \widetilde{\mu}(\mathbf{D}^0)/\sqrt{m}$ and $\widetilde{\mu}(\mathbf{D}^0) = \mathcal{O}(\log(m))$.

A5.4 (Sparsity level). The sparsity level satisfies $s \leq c_1 \min\left\{p^{2/5}, \frac{\sqrt{m}}{\widetilde{\mu}(\mathbf{D}^0)\log m}\right\}$, for some positive constant c_1.

THEOREM 1.5 (Noisy, exact recovery [31]) *Consider a DL problem with N i.i.d. observations and assume that assumptions **A5.1**–**A5.4** are satisfied. Provided that*

$$N = \Omega\left(\sigma^2\varepsilon^{-2} p \log p \left(\frac{p}{s^2} + s^2 + \log\frac{1}{\varepsilon}\right)\right), \tag{1.17}$$

there is a universal constant c_1 and a polynomial-time algorithm that learns the underlying dictionary. With high probability, this algorithm returns $\widehat{\mathbf{D}}$ that is column-wise ε close to $\mathbf{D}^0$.

For desired error ε, the run time of the algorithm and the sample complexity depend on $\log\frac{1}{\varepsilon}$. With the addition of noise, there is also a dependency on ε^{-2} for N, which is inevitable for noisy reconstruction of the true dictionary [31,38]. In the noiseless setting, this result translates into $N = \Omega\left(p\log p\left(\frac{p}{s^2} + s^2 + \log\frac{1}{\varepsilon}\right)\right)$.

Noisy Reconstruction with Outliers

In some scenarios, in addition to observations $\mathbf{Y}$ drawn from $\mathbf{D}^0$, we encounter observations $\mathbf{Y}_{out}$ that are not generated according to $\mathbf{D}^0$. We call such observations outliers (as opposed to inliers). In this case, the observation matrix is $\mathbf{Y}_{obs} = [\mathbf{Y}, \mathbf{Y}_{out}]$, where $\mathbf{Y}$ is the inlier matrix and $\mathbf{Y}_{out}$ is the outlier matrix. In this part, we study the robustness of dictionary identification in the presence of noise and outliers. The following result studies (1.6) with $\mathcal{R}(\mathbf{x}) = \lambda\left\|\mathbf{x}\right\|_1$, where λ is the regularization parameter. Here, the Frobenius norm is considered as the distance metric. In addition, the result is based on the following set of assumptions:

A6.1 (Cumulative coherence). The cumulative coherence of the true dictionary $\mathbf{D}^0$, which is defined as

$$\mu_s(\mathbf{D}^0) \triangleq \sup_{|\mathcal{S}|\leq s}\sup_{j\notin\mathcal{S}} \left\|{\mathbf{D}_{\mathcal{S}}^0}^T \mathbf{D}_j^0\right\|_1, \tag{1.18}$$

satisfies $\mu_s(\mathbf{D}^0) \leq 1/4$.

[8] This condition is trivially satisfied if the set of the locations of nonzero entries of $\mathbf{x}^n$ is a random subset of size s.

A6.2 (Bounded random coefficients). Assume nonzero entries of $\mathbf{x}^n$ are drawn i.i.d. from a distribution with absolute mean $\mathbb{E}\{|x|\}$ and variance $\mathbb{E}\{x^2\}$. We denote $\mathbf{l}^n = \text{sign}(\mathbf{x}^n)$.[9] Dropping the index of $\mathbf{x}^n$ and $\mathbf{l}^n$ for simplicity of notations, the following assumptions are satisfied for the coefficient vector: $\mathbb{E}\{\mathbf{x}_\mathcal{S}\mathbf{x}_\mathcal{S}^T|\mathcal{S}\} = \mathbb{E}\{x^2\}\mathbf{I}_s$, $\mathbb{E}\{\mathbf{x}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\} = \mathbb{E}\{|x|\}\mathbf{I}_s$, $\mathbb{E}\{\mathbf{l}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\} = \mathbf{I}_s$, $\|\mathbf{x}\|_2 \le M_x$, and $\min_{i\in\mathcal{S}} |\mathbf{x}_i| \ge x_{\min}$. We define $\kappa_x \triangleq \frac{\mathbb{E}\{|x|\}}{\sqrt{\mathbb{E}\{x^2\}}}$ as a measure of the flatness of $\mathbf{x}$. Moreover,the following inequality is satisfied:

$$\frac{\mathbb{E}\{x^2\}}{M_x\mathbb{E}\{|x|\}} > \frac{cs}{(1-2\mu_s(\mathbf{D}^0))p}\left(\left\|\mathbf{D}^0\right\|_2+1\right)\left\|{\mathbf{D}^0}^T\mathbf{D}^0-\mathbf{I}\right\|_F, \tag{1.19}$$

where c is a positive constant.

A6.3 (Regularization parameter). The Regularization parameter satisfies $\lambda \le x_{\min}/4$.

A6.4 (Bounded random noise). Assume nonzero entries of $\mathbf{w}^n$ are drawn i.i.d. from a distribution with mean 0 and variance $\mathbb{E}\{w^2\}$. Dropping the index of vectors for simplicity, $\mathbf{w}$ is a bounded random white noise vector satisfying $\mathbb{E}\{\mathbf{w}\mathbf{w}^T|\mathcal{S}\} = \mathbb{E}\{w^2\}\mathbf{I}_m$, $\mathbb{E}\{\mathbf{w}\mathbf{x}^T|\mathcal{S}\} = \mathbb{E}\{\mathbf{w}\mathbf{l}^T|\mathcal{S}\} = \mathbf{0}$, and $\|\mathbf{w}\|_2 \le M_w$. Furthermore, denoting $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$:

$$\frac{M_w}{M_x} \le \frac{7}{2}\left(c_{\max}-c_{\min}\right)\bar{\lambda}, \tag{1.20}$$

where $c_{\min}$ and $c_{\max}$ depend on problem parameters such as s, coefficient distribution, and $\mathbf{D}^0$.

A6.5 (Sparsity level) The sparsity level satisfies $s \le \frac{p}{16\left(\|\mathbf{D}^0\|_2+1\right)^2}$.

A6.6 (Radius range) The error radius $\varepsilon > 0$ satisfies $\varepsilon \in \left(\bar{\lambda}c_{\min}, \bar{\lambda}c_{\max}\right)$.

A6.7 (Outlier energy). Given inlier matrix $\mathbf{Y} = \{\mathbf{y}^n\}_{n=1}^N$ and outlier matrix $\mathbf{Y}_{out} = \{\mathbf{y'}^n\}_{n=1}^{N_{out}}$, the energy of $\mathbf{Y}_{out}$ satisfies

$$\frac{\|\mathbf{Y}_{out}\|_{1,2}}{N} \le \frac{c_1\varepsilon\sqrt{s}\mathbb{E}\left\{\|\mathbf{x}\|_2^2\right\}}{\bar{\lambda}\mathbb{E}\{|x|\}}\left(\frac{A^0}{p}\right)^{3/2}\left[\frac{1}{p}\left(1-\frac{c_{\min}\bar{\lambda}}{\varepsilon}\right)-c_2\sqrt{\frac{mp+\eta}{N}}\right], \tag{1.21}$$

where $\|\mathbf{Y}_{out}\|_{1,2}$ denotes the sum of the ℓ_2-norms of the columns of $\mathbf{Y}_{out}$, c_1 and c_2 are positive constants, independent of parameters, and A^0 is the lower frame bound of $\mathbf{D}^0$, i.e., $A^0\|\mathbf{v}\|_2^2 \le \left\|{\mathbf{D}^0}^T\mathbf{v}\right\|_2^2$ for any $\mathbf{v} \in \mathbb{R}^m$.

THEOREM 1.6 (Noisy with outliers, local recovery [32]) *Consider a DL problem with N i.i.d. observations and assume that assumptions* ***A6.1–A6.6*** *are satisfied. Suppose*

$$N > c_0(mp+\eta)p^2\left(\frac{M_x^2}{\mathbb{E}\left\{\|\mathbf{x}\|_2^2\right\}}\right)^2\left(\frac{\varepsilon+\left(\frac{M_w}{M_x}+\bar{\lambda}\right)+\left(\frac{M_w}{M_x}+\bar{\lambda}\right)^2}{\varepsilon-c_{\min}\bar{\lambda}}\right), \tag{1.22}$$

[9] The sign of the vector $\mathbf{v}$ is defined as $\mathbf{l} = \text{sign}(\mathbf{v})$, whose elements are $\mathbf{l}_i = \frac{\mathbf{v}_i}{|\mathbf{v}_i|}$ for $\mathbf{v}_i \ne 0$ and $\mathbf{l}_i = 0$ for $\mathbf{v}_i = 0$, where i denotes any index of the elements of $\mathbf{v}$.

then with probability at least $1 - 2^{-\eta}$, (1.6) *admits a local minimum within distance* ε *of* $\mathbf{D}^0$. *In addition, this result is robust to the addition of outlier matrix* $\mathbf{Y}_{out}$, *provided that the assumption in* ***A6.7*** *is satisfied.*

The proof of this theorem relies on using the Lipschitz continuity property of the objective function in (1.6) with respect to the dictionary and sample complexity analysis using Rademacher averages and Slepian's Lemma [48]. Theorem 1.6 implies that

$$N = \Omega\left(\left(mp^3 + \eta p^2\right)\left(\frac{M_w}{M_x \varepsilon}\right)^2\right) \tag{1.23}$$

samples are sufficient for the existence of a local minimum within distance ε of true dictionary $\mathbf{D}^0$, with high probability. In the noiseless setting, this result translates into $N = \Omega\left(mp^3\right)$, and sample complexity becomes independent of the radius ε. Furthermore, this result applies to overcomplete dictionaries with dimensions $p = \mathcal{O}(m^2)$.

1.2.4 Summary of Results

In this section, we have discussed DL minimax risk lower bounds [33] and achievability results [29–32,38]. These results differ in terms of the distance metric they use. An interesting question that rises here is: Can these results be unified so that the bounds can be directly compared with one another? Unfortunately, the answer to this question is not as straightforward as it seems and the inability to unify them is a limitation that we discuss in Section 1.4. A summary of the general scaling of the discussed results for sample complexity of (overcomplete) dictionary learning are provided in Table 1.1. We note that these are general scalings that ignore other technicalities. Here, the provided sample complexity results depend on the present or absence of noise and outliers. All the presented results require the underlying dictionary satisfies incoherence conditions in some way. For a one-to-one comparison of these results, the bounds for the case of absence of noise and outliers can be compared. A detailed comparison of the noiseless recovery for square and overcomplete dictionaries can be found in [32, Table I].

Table 1.1 Summary of the sample complexity results of various works

Reference	Jung et al. [33]	Geng et al. [29]	Agarwal et al. [30]	Schnass et al. [38]	Arora et al. [31]	Gribonval et al. [32]
Distance Metric	$\left\|\mathbf{D}^1 - \mathbf{D}^2\right\|_F$	–	$\min\limits_{l \in \{\pm 1\}} \left\|\mathbf{D}_j^1 - l\mathbf{D}_j^2\right\|_2$	$\max\limits_{j} \left\|\mathbf{D}_j^1 - \mathbf{D}_j^2\right\|_2$	$\min\limits_{l \in \{\pm 1\}, \pi} \left\|\mathbf{D}_j^1 - l\mathbf{D}_{\pi(j)}^2\right\|_2$	$\left\|\mathbf{D}^1 - \mathbf{D}^2\right\|_F$
Regularizer	ℓ_0	ℓ_1	ℓ_1	ℓ_1	ℓ_0	ℓ_1
Sparse Coefficient Distribution	nonzero i.i.d zero-mean, variance σ_x^2	nonzero i.i.d. $\sim$ $\mathcal{N}(0, \sigma_x)$	nonzero zero-mean unit-variance $x_{\min} \leq \|\mathbf{x}_i\| \leq x_{\max}$	symmetric decaying (non-sparse)	nonzero zero-mean $\mathbf{x}_i \in \pm[1, x_{\max}]$	nonzero $\|\mathbf{x}_i\| > x_{\min}$, $\left\|\mathbf{x}\right\|_2 \leq M_x$
Sparsity Level	–	$\mathcal{O}(\min$ $\{1/\mu, p\})$	$\mathcal{O}(\min\{1/\sqrt{\mu},$ $m^{1/9}, p^{1/8}\})$	$\mathcal{O}(1/\mu)$	$\mathcal{O}(\min\{1/(\mu \log m),$ $p^{2/5}\})$	$\mathcal{O}(m)$
Noise Distribution	i.i.d.$\sim$ $\mathcal{N}(0, \sigma)$	–	–	$\mathbb{E}\{\mathbf{w}\} = \mathbf{0}$ $\mathbb{E}\{\mathbf{w}\mathbf{w}^*\} =$ $\rho^2 \mathbf{I}_m$ $\left\|\mathbf{w}\right\|_2 \leq M_w$	i.i.d.$\sim \mathcal{N}(0, \sigma)$	$\mathbb{E}\left\{\mathbf{w}\mathbf{w}^T \vert \mathcal{S}\right\} =$ $\mathbb{E}\left\{w^2\right\} \mathbf{I}_m$ $\mathbb{E}\left\{\mathbf{w}\mathbf{x}^T \vert \mathcal{S}\right\} =$ $\mathbf{0}$, $\left\|\mathbf{w}\right\|_2 \leq M_w$
Outlier	–	–	–	–	–	Robust
Local-Global	Local	Local	Global	Local	Global	Local
Sample Complexity	$\dfrac{mp^2}{\varepsilon^2}$	sp^3	$p^2 \log p$	mp^3	$\dfrac{p}{\varepsilon^2} \log p(p/s^2 + s^2 +$ $\log \frac{1}{\varepsilon})$	$\dfrac{mp^3}{\varepsilon^2}$

1.3 Dictionary Learning for Tensors

Many of today's data are collected using various sensors and tend to have a multidimensional/tensor structure (cf. Fig. 1.2). Examples of tensor data include: 1) hyperspectral images that have three modes; two spatial and one spectral, 2) colored videos that have four modes; two spacial, one depth, and one temporal, and 3) dynamic magnetic resonance imaging in a clinical trial that has five modes; three spatial, one temporal, and one subject. To find representations of tensor data using DL, one can follow two paths. A naive approach is to vectorize tensor data and use traditional vectorized representation learning techniques. A better approach is to take advantage of the multidimensional structure of data to learn representations that are specific to tensor data. While the main focus of the literature on representation learning has been on the former approach, recent works have shifted focus to the latter approaches [8–11]. These works use various tensor decompositions to decompose tensor data into smaller components. The representation learning problem can then be reduced to learning the components that represent different modes of the tensor. This results in reduction in the number of degrees of freedom in the learning problem, due to the fact that the dimensions of the representations learned for each mode are significantly smaller than the dimensions of the representation learned for the vectorized tensor. Consequently, this approach gives rise to compact and efficient representation of tensors.

To understand the fundamental limits of dictionary learning for tensor data, one can use the sample complexity results in Section 1.2, which are a function of the underlying dictionary dimensions. However, considering the reduced number of degrees of freedom in the tensor DL problem compared to vectorized DL, this problem should be solvable with a smaller number of samples. In this section, we formalize this intuition and address the problem of reliable estimation of dictionaries underlying tensor data. Similar to the previous section, we will focus on the subject of sample complexity of the DL problem from two prospectives; *i)* fundamental limits on the sample complexity of DL for tensor data using any DL algorithm, and *ii)* number of samples that are needed for different DL algorithms to reliably estimate the true dictionary from which the tensor data is generated.

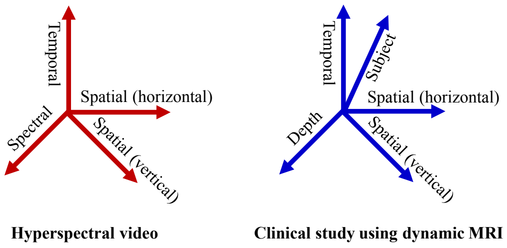

Figure 1.2 Two of countless examples of tensor data in today's sensor-rich world.

1.3.1 Tensor Terminology

A tensor is defined as a multiway array and the tensor order is defined as the number of components of the array. For instance, $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is a Kth-order tensor. For $K = 1$ and $K = 2$, the tensor is effectively a vector and a matrix, respectively. In order to better understand the results reported in this section, we first need to define some tensor notation that will be useful throughout this section.

Tensor Unfolding: Elements of tensors can be rearranged to form matrices. Given a Kth-order tensor, $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, its mode-$k$ unfolding is denoted as $\mathbf{X}_{(k)} \in \mathbb{R}^{p_k \times \prod_{i \neq k} p_i}$. The columns of $\mathbf{X}_{(k)}$ are formed by fixing all the indices, except one in the kth mode.

Tensor Multiplication: The mode-k product between the Kth-order tensor, $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, and a matrix, $\mathbf{A} \in \mathbb{R}^{m_k \times p_k}$, is defined as

$$(\underline{\mathbf{X}} \times_k \mathbf{A})_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} = \sum_{i_k=1}^{p_k} \underline{\mathbf{X}}_{i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_K} \mathbf{A}_{j,i_k}. \tag{1.24}$$

Tucker Decomposition [49]: Given a Kth-order tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ satisfying rank $\left(\underline{\mathbf{Y}}_{(k)}\right) \leq p_k$, the Tucker decomposition decomposes $\underline{\mathbf{Y}}$ into a *core* tensor $\underline{\mathbf{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ multiplied by *factor matrices* $\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}$ along each mode, i.e.,

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \cdots \times_K \mathbf{D}_K. \tag{1.25}$$

This can be restated as:

$$\operatorname{vec}\left(\underline{\mathbf{Y}}_{(1)}\right) = (\mathbf{D}_K \otimes \mathbf{D}_{K-1} \otimes \cdots \otimes \mathbf{D}_1) \operatorname{vec}\left(\underline{\mathbf{X}}_{(1)}\right), \tag{1.26}$$

where $\otimes$ denotes the matrix Kronecker product [50] and vec(.) denotes stacking of the columns of a matrix into one column. We will use the shorthand notation $\operatorname{vec}(\underline{\mathbf{Y}})$ to denote $\operatorname{vec}\left(\underline{\mathbf{Y}}_{(1)}\right)$ and $\bigotimes_k \mathbf{D}_k$ to denote $\mathbf{D}_1 \otimes \cdots \otimes \mathbf{D}_K$.

1.3.2 Mathematical Setup

To exploit the structure of tensors in DL, we can model tensors using various tensor decomposition techniques. These include Tucker decomposition, CANDECOMP/PARAFAC (CP) decomposition [51], and the t-product tensor factorization [52]. While the Tucker decomposition can be restated as the Kronecker product of matrices multiplied by a vector, other decompositions result in different formulations. In this chapter, we consider the Tucker decomposition due to the following reasons: *i)* it represents a sequence of independent transformations, i.e., factor matrices, for different data modes, and *ii)* Kronecker-structured matrices have successfully been used for data representation in applications such as magnetic resonance imaging, hyperspectral imaging, video acquisition, and distributed sensing [8, 9].

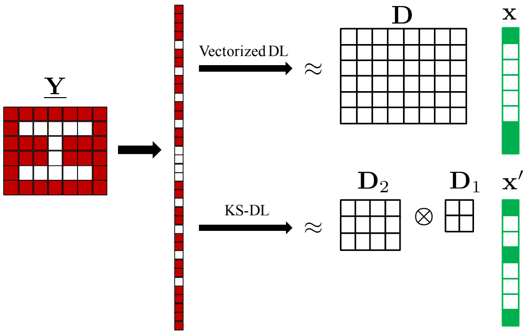

Figure 1.3 Illustration of the distinctions of KS-DL versus vectorized DL for a 2nd-order tensor: both vectorize the observation tensor, but the structure of the tensor is exploited in the KS dictionary, leading to the learning of two coordinate dictionaries with reduced number of parameters compared to the dictionary learned in vectorized DL.

Kronecker-structured Dictionary Learning (KS-DL)

In order to state the main results of this section, we begin with a generative model for tensor data based on Tucker decomposition. Specifically, we assume we have access to a total number of N tensor observations, $\underline{\mathbf{Y}}^n \in \mathbb{R}^{m_1 \times \cdots \times m_K}$, that are generated according to the following model:[10]

$$\mathrm{vec}(\underline{\mathbf{Y}}^n) = \big(\mathbf{D}_1^0 \otimes \mathbf{D}_2^0 \otimes \cdots \otimes \mathbf{D}_K^0\big)\, \mathrm{vec}(\underline{\mathbf{X}}^n) + \mathrm{vec}(\underline{\mathbf{W}}^n), \quad n = 1, \ldots, N. \tag{1.27}$$

Here, $\{\mathbf{D}_k^0 \in \mathbb{R}^{m_k \times p_k}\}_{k=1}^K$ are the true fixed *coordinate dictionaries*, $\underline{\mathbf{X}}^n \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the coefficient tensor, and $\underline{\mathbf{W}}^n \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ is the underlying noise tensor. In this case, the true dictionary $\mathbf{D}^0 \in \mathbb{R}^{m \times p}$ is Kronecker-structured (KS) and has the form

$$\mathbf{D}^0 = \bigotimes_k \mathbf{D}_k^0, \quad m = \prod_{k=1}^K m_k \quad \text{and} \quad p = \prod_{k=1}^K p_k,$$
$$\text{where} \quad \mathbf{D}_k^0 \in \mathcal{D}_k = \big\{\mathbf{D}_k \in \mathbb{R}^{m_k \times p_k}, \|\mathbf{D}_{k,j}\|_2 = 1 \ \forall j \in \{1, \ldots, p_k\}\big\}. \tag{1.28}$$

We define the set of KS dictionaries as

$$\mathcal{D}_{KS} = \Big\{\mathbf{D} \in \mathbb{R}^{m \times p} : \mathbf{D} = \bigotimes_k \mathbf{D}_k, \mathbf{D}_k \in \mathcal{D}_k \ \forall k \in \{1, \ldots, K\}\Big\}. \tag{1.29}$$

Comparing (1.27) to the traditional formulation in (1.1), it can be seen that KS-DL also involves vectorizing the observation tensor, but it has the main difference that the structure of the tensor is captured in the underlying KS dictionary. An illustration of this for a 2nd-order tensor is shown in Figure 1.3. Similar to (1.3), we can stack the vectorized observations, $\mathbf{y}^n = \mathrm{vec}(\underline{\mathbf{Y}}^n)$, vectorized coefficient tensors, $\mathbf{x}^n = \mathrm{vec}(\underline{\mathbf{X}}^n)$, and vectorized noise tensors, $\mathbf{w}^n = \mathrm{vec}(\underline{\mathbf{W}}^n)$, in columns of $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$, respectively. We now discuss the role of sparsity in

[10] We have reindexed $\mathbf{D}_k$'s here for simplicity of notation.

coefficient tensors for dictionary learning. While in vectorized DL it is usually assumed that the random support of nonzero entries of $\mathbf{x}^n$ is uniformly distributed, there are two different definitions of the random support of $\underline{\mathbf{X}}^n$ for tensor data:

1) Random sparsity: The random support of $\mathbf{x}^n$ is uniformly distributed over the set $\{\mathcal{S} \subseteq \{1, \dots, p\} : |\mathcal{S}| = s\}$.
2) Separable sparsity: The random support of $\mathbf{x}^n$ is uniformly distributed over the set $\mathcal{S}$ that is related to $\{\mathcal{S}_1 \times \dots \mathcal{S}_K : \mathcal{S}_k \subseteq \{1, \dots, p_k\}, |\mathcal{S}_k| = s_k\}$ via lexicographic indexing. Here, $s = \prod_k s_k$.

Separable sparsity requires nonzero entries of the coefficient tensor to be grouped in blocks. This model also implies that the columns of $\mathbf{Y}_{(k)}$ have s_k-sparse representations with respect to $\mathbf{D}_k^0$ [53].

The aim in KS-DL is to estimate coordinate dictionaries, $\widehat{\mathbf{D}}_k$'s, such that they are close to $\mathbf{D}_k^0$'s. In this scenario, the statistical risk minimization problem has the form:

$$\left(\widehat{\mathbf{D}}_1, \dots, \widehat{\mathbf{D}}_K\right) \in \underset{\{\mathbf{D}_k \in \mathcal{D}_k\}_{k=1}^K}{\arg\min} \ \mathbb{E}\left\{ \inf_{\mathbf{x} \in \mathcal{X}} \left\{ \frac{1}{2} \left\| \mathbf{y} - \left(\bigotimes_k \mathbf{D}_k \right) \mathbf{x} \right\|_2^2 + \mathcal{R}(\mathbf{x}) \right\} \right\}, \tag{1.30}$$

and the ERM problem is formulated as:

$$\left(\widehat{\mathbf{D}}_1, \dots, \widehat{\mathbf{D}}_K\right) \in \underset{\{\mathbf{D}_k \in \mathcal{D}_k\}_{k=1}^K}{\arg\min} \left\{ \sum_{n=1}^N \inf_{\mathbf{x}^n \in \mathcal{X}} \left\{ \frac{1}{2} \left\| \mathbf{y}^n - \left(\bigotimes_k \mathbf{D}_k \right) \mathbf{x}^n \right\|_2^2 + \mathcal{R}(\mathbf{x}^n) \right\} \right\}, \tag{1.31}$$

where $\mathcal{R}(.)$ is a regularization operator on the coefficient vectors. Various KS-DL algorithms have been proposed that solve (1.31) heuristically by means of optimization tools such as alternative minimization [9] and tensor rank minimization [54], and by taking advantage of techniques in tensor algebra such as the higher-order SVD for tensors [55]. In particular, an algorithm called "STARK" is proposed in [11] that shows that the Kronecker product of any number of matrices can be rearranged to form a rank-1 tensor. In order to solve (1.31), therefore, a regularizer is added in [11] to the objective function that enforces this low rankness on the rearrangement tensor. The dictionary update stage of this algorithm involves learning the rank-1 tensor and rearranging it to form the KS dictionary. This is in contrast to learning the individual coordinate dictionaries by means of alternating minimization [9].

In the case of theory for KS-DL, the notion of closeness can have two interpretations. One is the distance between the true KS dictionary and the recovered KS dictionary, i.e., $d\left(\widehat{\mathbf{D}}(\mathbf{Y}), \mathbf{D}^0\right)$. The other is the distance between each true coordinate dictionary and the corresponding recovered coordinate dictionary, i.e., $d\left(\widehat{\mathbf{D}}_k(\mathbf{Y}), \mathbf{D}_k^0\right)$. While small recovery errors for coordinate dictionaries imply a small recovery error for the KS dictionary, the other side of the statement does not necessarily hold. Hence, the latter notion is of importance when we are interested in recovering the structure of the KS dictionary.

In this section, we focus on the sample complexity of the KS-DL problem. The questions that we address in this section are *i)* What are the fundamental limits of solving the statistical risk minimization problem in (1.30)? *ii)* Under what kind of conditions do objective functions like (1.31) recover the true coordinate dictionaries and how many samples do they need for this purpose? *iii)* How do these limits compare to their vectorized DL counterparts? Addressing these question will help in understanding the benefits of KS-DL for tensor data.

1.3.3 Fundamental Limits on the Minimax Risk of KS-DL

Below, we present a result that obtains lower bounds on the minimax risk of the KS-DL problem. This result can be considered as an extension of Theorem 1.1 for the KS-DL problem for tensor data. Here, the Frobenius norm is considered as the distance metric and the result is based on the following assumption:

A7.1 (Local recovery). The true KS dictionary lies in a neighborhood of some reference dictionary, $\mathbf{D}^* \in \mathcal{D}_{KS}$, i.e., $\mathbf{D}^0 \in \widetilde{\mathcal{D}}_{KS}$, where

$$\widetilde{\mathcal{D}}_{KS} = \{\mathbf{D}|\mathbf{D} \in \mathcal{D}_{KS}, \|\mathbf{D}-\mathbf{D}^*\|_F \leq r\}. \tag{1.32}$$

THEOREM 1.7 (KS-DL minimax lower bounds [13]) *Consider a KS-DL problem with N i.i.d. observations and true KS dictionary $\mathbf{D}^0$ satisfying assumption* ***A7.1*** *for some $r \in (0, 2\sqrt{p}]$. Then, for any coefficient distribution with mean zero and covariance matrix $\mathbf{\Sigma}_x$, and white Gaussian noise with mean zero and variance σ^2, the minimax risk ε^* is lower bounded as*

$$\varepsilon^* \geq \frac{t}{4}\min\left\{p, \frac{r^2}{2K}, \frac{\sigma^2}{4NK\|\mathbf{\Sigma}_x\|_2}\left(c_1\left(\sum_{k=1}^{K}(m_k-1)p_k\right) - \frac{K}{2}\log_2 2K - 2\right)\right\}, \tag{1.33}$$

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

Similar to Theorem 1.1, the proof of this theorem relies on using the standard procedure for lower bounding the minimax risk by connecting it to the maximum probability of error of a multiple hypothesis testing problem. Here, since the constructed hypothesis testing class consists of KS dictionaries, the construction procedure and the minimax risk analysis are different from that in [33].

To understand this theorem, let us assume that r and p are sufficiently large such that the minimizer of the left hand side of (1.33) is the third term. In this case, Theorem 1.7 states that to achieve any error ε for the Kth-order tensor dictionary recovery problem, we need the number of samples to be on the order of $N = \Omega\left(\dfrac{\sigma^2\sum_k m_k p_k}{K\|\mathbf{\Sigma}_x\|_2\,\varepsilon}\right)$. Comparing this scaling to the results for the unstructured dictionary learning problem provided in Theorem 1.1, the lower bound here is decreased from the scaling $\Omega\left(mp\right)$ to $\Omega\left(\sum_k m_k p_k/K\right)$. This reduction can be attributed to the fact that the average number of degrees of freedom in a KS-DL problem is $\sum_k m_k p_k/K$, compared to the number of degrees of the vectorized DL problem, which is mp. For the case of $K = 2$ and $m_1 = m_2 = \sqrt{m}$ and $p_1 = p_2 = \sqrt{p}$, the sample complexity lower bound scales with $\Omega(mp)$ for vectorized DL, and with $\Omega(\sqrt{mp})$ for KS-DL. On the other hand, when $m_1 = \alpha m, m_2 = 1/\alpha$ and $p_1 = \alpha m_1, p_2 = 1/\alpha$, where $\alpha < 1, 1/\alpha \in \mathbb{N}$, the sample complexity lower bound scales with $\Omega(mp)$ for KS-DL, which is similar to the scaling for vectorized DL.

Specializing this result to random sparse coefficient vectors and assuming that the nonzero entries of $\mathbf{x}^n$ are i.i.d. with variance σ_x^2, we get $\mathbf{\Sigma}_x = (s/p)\sigma_x^2\mathbf{I}_p$. Therefore, for sufficiently large r, the sample complexity scaling to achieve any error ε for strictly sparse representations becomes $\Omega\left(\frac{\sigma^2 p \sum_k m_k p_k}{\sigma_x^2 s K \varepsilon}\right)$.

A very simple KS-DL algorithm is also provided in [13] that can recover a square KS dictionary that consists of the Kronecker product of 2 smaller dictionaries and is a perturbation of the identity matrix. It is shown that in this case, the lower bound provided in (1.33) is order-wise achievable for the case of sparse coefficient vectors. This suggests that the obtained sample complexity lower bounds for overcomplete KS-DL are not too loose.

In the next subsection, we focus on achievability results for the KS dictionary recovery problem, i.e., upper bounds on the sample complexity of KS-DL.

1.3.4 Achievability results

While the results in the previous section provide us with a lower bound on the sample complexity of the KS-DL problem, we are further interested in the sample complexity of specific KS-DL algorithms that solve (1.31). Below, we present a KS-DL achievability result that can be interpreted as an extension of Theorem 1.6 to the KS-DL problem.

Noisy Recovery

We present a result that states conditions that ensure reliable recovery of the coordinate dictionaries from noisy observations using (1.31). Shakeri et al. [15] solve (1.31) with $\mathcal{R}(\mathbf{x}) = \lambda \left\|\mathbf{x}\right\|_1$, where λ is a regularization parameter. Here, the coordinate dictionary error is defined as

$$\varepsilon_k = \left\|\widehat{\mathbf{D}}_k - \mathbf{D}_k^0\right\|_F, k \in \{1, \ldots, K\}, \tag{1.34}$$

where $\widehat{\mathbf{D}}_k$ is the recovered coordinate dictionary. The result is based on the following set of assumptions:

A8.1 (Cumulative coherence). The cumulative coherences of the true coordinate dictionaries satisfy $\mu_{s_k}(\mathbf{D}_k^0) \leq 1/4$ and the cumulative coherences of the true dictionary satisfies $\mu_s(\mathbf{D}^0) \leq 1/2$.[11]

A8.2 (Bounded random coefficients). The random support of $\mathbf{x}^n$ is generated from the separable sparsity model. Assume nonzero entries of $\mathbf{x}^n$ are drawn i.i.d. from a distribution with absolute mean $\mathbb{E}\left\{|x|\right\}$ and variance $\mathbb{E}\left\{x^2\right\}$. Denoting $\mathbf{l}^n = \text{sign}(\mathbf{x}^n)$, and dropping the index of $\mathbf{x}^n$ and $\mathbf{l}^n$ for simplicity of notation, the following assumptions are satisfied for the coefficient vector: $\mathbb{E}\left\{\mathbf{x}_\mathcal{S}\mathbf{x}_\mathcal{S}^T|\mathcal{S}\right\} = \mathbb{E}\left\{x^2\right\}\mathbf{I}_s$, $\mathbb{E}\left\{\mathbf{x}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\right\} = \mathbb{E}\left\{|x|\right\}\mathbf{I}_s$, $\mathbb{E}\left\{\mathbf{l}_\mathcal{S}\mathbf{l}_\mathcal{S}^T|\mathcal{S}\right\} = \mathbf{I}_s$, $\|\mathbf{x}\|_2 \leq M_x$, and $\min_{i\in\mathcal{S}}|\mathbf{x}_i| \geq x_{\min}$. Moreover, defining $\kappa_x \triangleq \frac{\mathbb{E}\{|x|\}}{\sqrt{\mathbb{E}\{x^2\}}}$ as a measure of the flatness of $\mathbf{x}$, the following inequality is satisfied:

$$\frac{\mathbb{E}\left\{x^2\right\}}{M_x\mathbb{E}\left\{|x|\right\}} > \frac{c_1}{1-2\mu_s(\mathbf{D}^0)} \max_{k\in\{1,\dots,K\}} \left(\frac{s_k}{p_k}\left(\left\|\mathbf{D}_k^0\right\|_2+1\right)\left\|{\mathbf{D}_k^0}^T\mathbf{D}_k^0-\mathbf{I}\right\|_F\right), \tag{1.35}$$

where c_1 is a positive constant that is an exponential function of K.

A8.3 (Regularization parameter). The Regularization parameter satisfies $\lambda \leq x_{\min}$ $/c_2$, where c_2 is a positive constant that is an exponential function of K.

A8.4 (Bounded random noise). Assume nonzero entries of $\mathbf{w}^n$ are drawn i.i.d. from a distribution with mean 0 and variance $\mathbb{E}\left\{w^2\right\}$. Dropping the index of vectors for simplicity of notation, $\mathbf{w}$ is a bounded random white noise vector satisfying $\mathbb{E}\left\{\mathbf{w}\mathbf{w}^T|\mathcal{S}\right\} = \mathbb{E}\left\{w^2\right\}\mathbf{I}_m$, $\mathbb{E}\left\{\mathbf{w}\mathbf{x}^T|\mathcal{S}\right\} = \mathbb{E}\left\{\mathbf{w}\mathbf{l}^T|\mathcal{S}\right\} = \mathbf{0}$, and $\|\mathbf{w}\|_2 \leq M_w$. Furthermore, denoting $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\{|x|\}}$, we have

$$\frac{M_w}{M_x} \leq c_3\left(\bar{\lambda}Kc_{\max} - \sum_{k=1}^{K}\varepsilon_k\right), \tag{1.36}$$

where c_3 is a positive constant that is an exponential function of K and $c_{\max}$ depends on the coefficient distribution, $\mathbf{D}^0$, and K.

A8.5 (Sparsity level). The sparsity levels for each mode satisfy $s_k \leq \frac{p_k}{8\left(\left\|\mathbf{D}_k^0\right\|_2+1\right)^2}$ for $k \in \{1,\dots,K\}$.

A8.6 (Radii range). The error radii $\varepsilon_k > 0$ satisfy $\varepsilon_k \in \left(\bar{\lambda}c_{k,\min}, \bar{\lambda}c_{\max}\right)$ for $k \in \{1,\dots,K\}$, where $c_{k,\min}$ depends on s, the coefficient distribution, $\mathbf{D}^0$, and K.

THEOREM 1.8 (Noisy KS-DL, local recovery [15]) *Consider a KS-DL problem with N i.i.d. observations and suppose that assumptions **A8.1**–**A8.6** are satisfied. Assume*

$$N \geq \max_{k\in[K]}\Omega\left(\frac{p_k^2(\eta+m_kp_k)}{(\varepsilon_k-\varepsilon_{k,\min}(\bar{\lambda}))^2}\left(\frac{2^K(1+\bar{\lambda}^2)M_x^2}{s^2\mathbb{E}\{x^2\}^2}+\left(\frac{M_w}{s\mathbb{E}\{x^2\}}\right)^2\right)\right), \tag{1.37}$$

where $\varepsilon_{k,\min}(\bar{\lambda})$ is a function of K, $\bar{\lambda}$, and $c_{k,\min}$. Then, with probability at least $1-e^{-\eta}$, there exists a local minimum of (1.31), $\widehat{\mathbf{D}} = \bigotimes\widehat{\mathbf{D}}_k$, *such that $d(\widehat{\mathbf{D}}_k, \mathbf{D}_k^0) \leq \varepsilon_k$, for all $k \in \{1,\dots,K\}$.*

[11] The cumulative coherence is defined in (1.18).

Table 1.2 Comparison of the scaling of vectorized DL sample complexity bounds with KS-DL, given fixed SNR.

	Vectorized DL	KS-DL
Minimax Lower Bound	$\frac{mp^2}{\varepsilon^2}$ [33]	$\frac{p\sum_k m_k p_k}{K\varepsilon^2}$ [13]
Achievability Bound	$\frac{mp^3}{\varepsilon^2}$ [32]	$\max_k \frac{m_k p_k^3}{\varepsilon_k^2}$ [15]

The proof of this theorem relies on coordinate-wise Lipschitz continuity of the objective function in (1.31) with respect to coordinate dictionaries and using similar sample complexity analysis arguments as in [32]. Theorem 1.8 implies that for fixed K and SNR, $N = \max_{k \in \{1,\dots,K\}} \Omega\left(m_k p_k^3 \varepsilon_k^{-2}\right)$ is sufficient for existence of a local minimum within distance ε_k of true coordinate dictionaries, with high probability. This result holds for coefficients that are generated according to the separable sparsity model. The case of coefficients generated according to the random sparsity model requires a different analysis technique that is not explored in [15].

We compare this result to the scaling in the vectorized DL problem in Theorem 1.6, which stated that $N = \Omega\left(mp^3\varepsilon^{-2}\right) = \Omega\left(\prod_k m_k p_k^3 \varepsilon^{-2}\right)$ is sufficient for existence of $\mathbf{D}^0$ as a local minimum of (1.6) up to the predefined error ε. In contrast, $N = \max_k \Omega\left(m_k p_k^3 \varepsilon_k^{-2}\right)$ is sufficient in the case of tensor data for existence of $\mathbf{D}_k^0$'s as local minimums of (1.31) upto predefined errors ε_k. This reduction in the scaling can be attributed to the reduction in the number of degrees of freedom of the KS-DL problem.

We can also compare this result to the sample complexity lower bound scaling obtained in 1.7 for KS-DL, which stated that given sufficiently large r and p, $N = \Omega\left(p\sum_k m_k p_k \varepsilon^{-2}/K\right)$ is necessary to recover true KS dictionary $\mathbf{D}^0$ up to error ε. We can relate ε to ε_k's using the relation $\varepsilon \leq \sqrt{p}\sum_k \varepsilon_k$ [15]. Assuming all ε_k's are equal to each other, this implies that $\varepsilon \leq \sqrt{p}K\varepsilon_k$ and we have $N = \max_k \Omega\left(2^K K^2 p (m_k p_k^3)\varepsilon^{-2}\right)$. It can be seen from Theorem 1.7 that the sample complexity lower bound depends on the average dimension of coordinate dictionaries; in contrast, the sample complexity upper bound reported in this section depends on the maximum dimension of coordinate dictionaries. There is also a gap between the lower bound and the upper bound of order $\max_k p_k^2$. This suggests that the obtained bounds may be loose.

The sample complexity scaling results in Theorems 1.1, 1.6, 1.7, and 1.8 are demonstrated in Table 1.2 for sparse coefficient vectors.

1.4 Extensions and Open Problems

In Sections 1.2 and 1.3, we summarized some of the key results of dictionary identification for vectorized and tensor data. In this section, we look at extensions of these works and discuss related open problems.

1.4.1 DL for vector-valued Data

Extensions to alternative objective functions. The works discussed in Section 1.2 all analyze variants of (1.5) and (1.6), which minimizes the representation error of the dictionary. However, there do exist other works that look for a dictionary that optimizes different criteria. Schnass [56] proposed a new DL objective function called the "response maximization criterion" that extends the K-means objective function to the following:

$$\max_{\mathbf{D}\in\mathcal{D}} \sum_{n=1}^{N} \max_{|\mathcal{S}|=s} \left\|\mathbf{D}_{\mathcal{S}}^{*}\mathbf{y}^{n}\right\|_{1}. \tag{1.38}$$

Given distance metric $d(\mathbf{D}^1, \mathbf{D}^2) = \max_j \left\|\mathbf{D}_j^1 - \mathbf{D}_j^2\right\|_2$, Schnass shows the sample complexity needed to recover a true generating dictionary up to precision ε scales as $\mathcal{O}\left(mp^3\varepsilon^{-2}\right)$ using this objective. This sample complexity is achieved by a novel DL algorithm, ITKM (Iterative Thresholding and K-Means), that solves (1.38) under certain conditions on coefficient distribution, noise, and the underlying dictionary.

Efficient representations can help improve the complexity and performance of machine learning tasks such as prediction. This means that a DL algorithm could explicitly tune the representation to optimize prediction performance. For example, some works learn dictionaries to improve classification performance [17, 25]. These works add terms to the objective function that measure the prediction performance and minimize this loss. While these DL algorithms can yield improved performance for their desired prediction task, proving sample complexity bounds for these algorithms remains an open problem.

Tightness guarantees. While dictionary identifiability has been well studied for vector-valued data, there remains a gap between the upper and lower bounds on the sample complexity. The lower bound presented in Theorem 1.1 is for the case of a particular distance metric, i.e., the Frobenius norm, whereas the presented achievability results in Theorems 1.2–1.6 are based on a variety of distance metrics. Restricting the distance metric to the Frobenius norm, we still observe a gap of order p between the sample complexity lower bound in Theorem 1.1 and upper bound in Theorem 1.6. The partial converse result for square dictionaries that is provided in [33] shows that the lower bound is achievable for square dictionaries close to the identity matrix. For more general square matrices, however, the gap may be significant: either improved algorithms can achieve the lower bounds or the lower bounds may be further tightened. For overcomplete dictionaries the question of whether the upper bound or lower bound is tight remains open. For metrics other than the Frobenius norm, the bounds are incomparable, making it challenging to assess the tightness of many achievability results.

Finally, the works reported in Table 1.1 differ significantly in terms of the mathematical tools they use. Each approach yields a different insight into the structure of the DL problem. However, there is no unified analytical framework encompassing all of these perspectives. This gives rise to the question: is there a unified mathematical tool that can be used to generalize existing results on DL?

1.4.2 DL for Tensor Data

Extensions of sample complexity bounds for KS-DL. In terms of theoretical results, there are many aspects of KS-DL that have not been addressed in the literature so far. The results that are obtained in Theorems 1.7 and 1.8 are based on the Frobenius norm distance metric and only provide local recovery guarantees. Open questions include corresponding abounds for other distance metrics and global recovery guarantees. In particular, getting global recovery guarantees requires using a distance metric that can handle the inherent permutation and sign ambiguities in the dictionary. Moreover, the results of Theorem 1.8 are based on the fact that the coefficient tensors are generated according to the separable sparsity model. Extensions to coefficient tensors with arbitrary sparsity patterns, i.e., the random sparsity model, have not been explored.

Algorithmic open problems. Unlike vectorized DL problems whose sample complexity is explicitly tied to the actual algorithmic objective functions, the results in [13, 15] are not tied to an explicit KS-DL algorithm. While there exist KS-DL algorithms in the literature, none of them explicitly solve the problem in these papers. Empirically, KS-DL algorithms can outperform vectorized DL algorithms for a variety of real-world data sets [10, 11, 57–59]. However, these algorithms lack theoretical analysis in terms of sample complexity, leaving open the question how many samples are needed to learn a KS dictionary using practical algorithms.

Parameter selection in KS-DL. In some cases we may not know a priori the parameters for which a KS dictionary yields a good model for the data. In particular, given dimension p, the problem of selecting the p_k's for coordinate dictionaries such that $p = \prod_k p_k$ has not been studied. For instance, in case of RGB images, selection of p_k's for the spatial modes is somewhat intuitive, as each column in the separable transform represents a pattern in each mode. However, selecting the number of columns for the depth mode, which has 3 dimensions (red, green, and blue), is less obvious. Given a fixed number of overall columns for the KS dictionary, how should we divide it between the number of columns for each coordinate dictionary?

Alternative structures on dictionary. In terms of DL for tensor data, extensions of identifiability results to structures other than the Kronecker product is an open problem. The main assumption in KS-DL is that the transforms for different modes of the tensor are separable from one another, which can be a limiting assumption for real-world datasets. Other structures can be enforced on the underlying dictionary to reduce sample complexity while applying to a wider range of datasets. Examples include DL using the CP decomposition [60] and the tensor t-product [61]. Characterizing the DL problem and understanding the practical benefits of these models remains an interesting question for future work.

References

[1] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications.* McGraw-Hill New York, 1986, vol. 31999.

[3] I. Daubechies, *Ten lectures on wavelets.* SIAM, 1992, vol. 61.

[4] E. J. Candes and D. L. Donoho, “Curvelets: A surprisingly effective nonadaptive representation for objects with edges,” *Curves and Surfaces*, pp. 105–120, 2000.

[5] I. T. Jolliffe, “Principal component analysis and factor analysis,” in *Principal Component Analysis.* Springer, 1986, pp. 115–128.

[6] 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, Nov. 2006.

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

[8] M. F. Duarte and R. G. Baraniuk, “Kronecker compressive sensing,” *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Feb. 2012.

[9] C. F. Caiafa and A. Cichocki, “Multidimensional compressed sensing and their applications,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 355–380, Nov./Dec. 2013.

[10] S. Hawe, M. Seibert, and M. Kleinsteuber, “Separable dictionary learning,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, Jun. 2013, pp. 438–445.

[11] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, “STARK: Structured dictionary learning through rank-one tensor recovery,” in *Proc. IEEE 7th Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing*, Dec. 2017, pp. 1–5.

[12] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, “Minimax lower bounds for Kronecker-structured dictionary learning,” in *Proc. 2016 IEEE Int. Symp. Inf. Theory*, Jul. 2016, pp. 1148–1152.

[13] ——, “Minimax lower bounds on dictionary learning for tensor data,” *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, Apr. 2018.

[14] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, “Identification of Kronecker-structured dictionaries: An asymptotic analysis,” in *Proc. IEEE 7th Int. Workshop Computational Advances in Multi-Sensor Adaptive Processing*, Dec. 2017, pp. 1–5.

[15] ——, "Identifiability of Kronecker-structured dictionaries for tensor data," *IEEE J. Sel. Topics Signal Processing (early access)*, May 2018.

[16] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, 2005.

[17] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Machine learning.* ACM, 2007, pp. 759–766.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis.* John Wiley & Sons, 2004, vol. 46.

[20] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[21] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artificial Neural Networks.* Springer, 1997, pp. 583–588.

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[23] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," in *Proc. 23rd Conf. Uncertainty in Artificial Intelligence*, Jul. 2007, pp. 149–158.

[24] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, 2009.

[25] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Analys. and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[26] M. Aharon, M. Elad, and A. 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, Jul. 2006.

[27] R. Remi and K. Schnass, "Dictionary identification–sparse matrix-factorization via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[28] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conf. Learning Theory*, 2012, pp. 37–1.

[29] Q. Geng and J. Wright, "On the local correctness of ℓ_1-minimization for dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory.* IEEE, Jun. 2014, pp. 3180–3184.

[30] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Proc. 27th Annu. Conf. Learning Theory*, ser. JMLR: Workshop and Conf. Proc., vol. 35, no. 1, 2014, pp. 1–15.

[31] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Proc. 25th Annu. Conf. Learning Theory*, ser. JMLR: Workshop and Conf. Proc., vol. 35, 2014, pp. 1–28.

[32] R. Gribonval, R. Jenatton, and F. Bach, "Sparse and spurious: Dictionary learning with noise and outliers," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6298–6319, Nov. 2015.

[33] 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, Mar. 2015.

[34] O. Christensen, *An introduction to frames and Riesz bases.* Springer, 2016.

[35] K. A. Okoudjou, *Finite frame theory: a complete introduction to overcompleteness.* American Mathematical Soc., 2016, vol. 93.

[36] W. U. Bajwa, R. Calderbank, and D. G. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Applied and Computational Harmonic Analysis*, vol. 33, no. 1, pp. 58–78, 2012.

[37] W. U. Bajwa and A. Pezeshki, "Finite frames for sparse signal processing," in *Finite Frames*, P. Casazza and G. Kutyniok, Eds. Cambridge, MA: Birkhäuser Boston, 2012, ch. 10, pp. 303–335.

[38] K. 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, Nov. 2014.

[39] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[40] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Advances in Neural Information Processing Systems*, 1992, pp. 831–838.

[41] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, vol. 5. IEEE, Mar. 1999, pp. 2443–2446.

[42] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, no. Jan., pp. 19–60, 2010.

[43] H. Raja and W. U. Bajwa, "Cloud K-SVD: A collaborative dictionary learning algorithm for big, distributed data," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 173–188, 2016.

[44] Z. Shakeri, H. Raja, and W. U. Bajwa, "Dictionary learning based nonlinear classifier training from distributed data," in *Proc. 2nd IEEE Global Conf. Signal and Information Processing*, Dec. 2014, pp. 759–763.

[45] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images," *IEEE Trans. Image Process,*, vol. 21, no. 1, pp. 130–144, 2012.

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

[47] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[48] P. Massart, *Concentration inequalities and model selection.* Springer, 2007, vol. 6.

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

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

[51] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, Dec. 1970.

[52] M. E. Kilmer, C. D. Martin, and L. Perrone, “A third-order generalization of the matrix svd as a product of third-order tensors,” *Tufts University, Department of Computer Science, Tech. Rep. TR-2008-4*, 2008.

[53] C. F. Caiafa and A. Cichocki, “Computing sparse representations of multidimensional signals using Kronecker bases,” *Neural Computation*, vol. 25, no. 1, pp. 186–220, Jan. 2013.

[54] S. Gandy, B. Recht, and I. Yamada, “Tensor completion and low-n-rank tensor recovery via convex optimization,” *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[55] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A multilinear singular value decomposition,” *SIAM J. Matrix Analy. and Applicat.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[56] K. Schnass, “Local identification of overcomplete dictionaries,” *J. Machine Learning Research*, vol. 16, pp. 1211–1242, Jun. 2015.

[57] S. Zubair and W. Wang, “Tensor dictionary learning with sparse Tucker decomposition,” in *Proc. IEEE 18th Int. Conf. Digital Signal Process.*, Jul. 2013, pp. 1–6.

[58] 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.*, May 2014, pp. 3963–3967.

[59] C. F. Dantas, M. N. da Costa, and R. da Rocha Lopes, “Learning dictionaries as a sum of Kronecker products,” *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 559–563, Mar. 2017.

[60] Y. Zhang, X. Mou, G. Wang, and H. Yu, “Tensor-based dictionary learning for spectral CT reconstruction,” *IEEE Trans. Medical Imaging*, vol. 36, no. 1, pp. 142–154, 2017.

[61] S. Soltani, M. E. Kilmer, and P. C. Hansen, “A tensor-based dictionary learning approach to tomographic image reconstruction,” *BIT Numerical Mathematics*, pp. 1–30, 2015.

Index