On the Reconstruction Risk of Convolutional Sparse Dictionary Learning

Shashank Singh[1,2], Barnabás Póczos[2], and Jian Ma[3]

Abstract— **Sparse dictionary learning (SDL) has become a popular method for adaptively identifying parsimonious representations of a dataset, a fundamental problem in machine learning and signal processing. While most work on SDL assumes a training dataset of independent and identically distributed samples, a variant known as convolutional sparse dictionary learning (CSDL) relaxes this assumption, allowing more general sequential data sources, such as time series or other dependent data. Although recent work has explored the statistical properties of classical SDL, the statistical properties of CSDL remain unstudied. This paper begins to study this by identifying the minimax convergence rate of CSDL in terms of reconstruction risk, by both upper bounding the risk of an established CSDL estimator and proving a matching information-theoretic lower bound. Our results indicate that consistency in reconstruction risk is possible precisely in the 'ultra-sparse' setting, in which the sparsity (i.e., the number of feature occurrences) is in $o(N)$ in terms of the length N of the training sequence, with precise rates depending the tails of the noise distribution. Notably, our results make very weak assumptions, allowing arbitrary dictionaries and dependent measurement noise. Finally, we verify our theoretical results with numerical experiments on synthetic data.**

I. Introduction

Many problems in machine learning and signal processing can be reduced to, or greatly simplified by, finding a concise representation of a dataset. In recent years, partly inspired by representations within the visual and auditory cortices of the brain [1], [2], [3], the method of *sparse dictionary learning* (SDL) has become a popular way of learning such a concise representation, encoded as a sparse linear combination of learned *dictionary elements*, or patterns recurring throughout the dataset.

SDL has been widely applied in image processing, to such problems as image denoising, demosaicing, and inpainting [4], [5], [6], [7], [8], separation [9], compression [10], object recognition [11], trajectory reconstruction [12], and super-resolution image reconstruction [13], [14]. In speech processing, it has been applied to structured [15] and unstructured [16] denoising, compression [17], speech recognition [18], and speaker separation [19]. SDL has also been applied for feature learning in more general data domains [20].

The vast majority of literature on SDL assumes that the training dataset consists of a large number of independent and identically distributed (IID) samples; typical datasets might be collections of small image patches, pictures of faces, short audio clips of individual utterances, or frames sampled from a video. However, somewhat more recently, there has been work on *convolutional sparse dictionary learning* (CSDL), described in Section III-A), in which a dictionary is learned from a sequential data source, such as a (potentially non-stationary) time series [21], [19], [18]. A number of fast optimization-based CSDL algorithms have been proposed [22], [23], [24], [25].

Also recently, a body of work has formed around understanding the statistical properties of classical (IID) SDL, characterizing the convergence rates of certain algorithms in recovering a true dictionary when it exists [26], [27], [28], [29], as well as proving lower bounds on the intrinsic minimax risk of this problem [30], [31]. However, to the best of our knowledge, the minimax statistical properties of CSDL have not yet been studied. The main contribution of this paper is to begin filling this gap, by analyzing the reconstruction/denoising error of CSDL (i.e., the error of reconstructing a signal from a learned convolutional sparse dictionary decomposition, a process often used for denoising [17], [4]. We do this by (a) upper bounding the reconstruction risk of an established estimator [21] and (b) lower bounding the minimax risk of reconstructing a data source constructed sparsely from a convolutional dictionary.

Paper Organization: Section II defines notation needed to state the CSDL problem and our theoretical results. Section III provides background on the problem of sparse dictionary learning in both the IID and convolutional variants. Section IV reviews recent theoretical work in the IID case, establishing context for our results on the convolutional case. Section V contains statements and some discussion of our main theoretical results, with proofs given in the appendix. In Section VI, we experimentally validate our theoretical results on synthetic data. Finally, in Section VII, we conclude and suggest avenues for future work.

II. Notation

We begin by defining some notation used throughout the remainder of the paper.

Multi-convolution: For two matrices $R \in \mathbb{R}^{(N-n+1)\times K}$ and $D \in \mathbb{R}^{n\times K}$ with an equal number of columns, we define the *multi-convolution* operator $\otimes$ by

$$R \otimes D := \sum_{k=1}^{K} R_k * D_k \in \mathbb{R}^N,$$

where $*$ denotes the standard discrete convolution operator. In the CSDL setting, multi-convolution (rather than standard matrix multiplication, as in IID SDL) is the process by which the data signal is constructed from the weight matrix R and

[1]Department of Statistics, [2]Machine Learning Department, [3]Computational Biology Department, Carnegie Mellon University

the dictionary D. We note that, like matrix multiplication, multi-convolution is a bilinear operation.

Matrix norms: For any matrix $A \in \mathbb{R}^{n \times m}$,

$$\|A\|_{p,q} := \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{n} a_{i,j}^{p} \right)^{q/p} \right)^{1/q}$$

(or the corresponding limit if p or q is ∞) denotes the q-norm of the vector whose entries are the p-norms of the columns of A. Note that the norm $\| \cdot \|_{2,2}$ is precisely the Frobenius norm, in terms of which SDL is often expressed.

Problem Domain: We use

$$\mathcal{S} := \left\{ (R, D) \in [0, \infty)^{(N-n+1) \times K} \times \mathbb{R}^{n \times K} : \|D\|_{2,\infty} \leq 1 \right\}$$

to denote the domain of the dictionary learning problem, (i.e., $(R, D) \in \mathcal{S}$, as described in the next section), and, for any $\lambda \geq 0$, we further use

$$\mathcal{S}_\lambda := \left\{ (R, D) \in \mathcal{S} : \|R\|_{1,1} \leq \lambda \right\}$$

to denote the $\mathcal{L}_1$-constrained version of this domain. Note that both $\mathcal{S}$ and $\mathcal{S}_\lambda$ are convex sets.

III. Background: IID and Convolutional Sparse Dictionary Learning

We now review background on the classical IID and convolutional sparse dictionary learning problems, as well as a standard approach to solving each.

A. IID Sparse Dictionary Learning

The classical SDL problem for IID data considers a dataset $X \in \mathbb{R}^{N \times d}$ of N IID samples with values in $\mathbb{R}^d$. The goal is to find an approximate decomposition $X \approx RD$, where $R \in \mathbb{R}^{N \times K}$ is a sparse weight matrix and $D \in \mathbb{R}^{K \times d}$ is a dictionary of K patterns. Usually, $d < K < N$, so that the dictionary is over-complete (redundant), which enables a sparser R, while ensuring there are enough samples to learn D robustly.

A typical frequentist starting point for IID sparse dictionary learning is the *linear generative model* of [1], which supposes there exist (deterministic) R and D as above such that

$$X = RD + \varepsilon \in \mathbb{R}^{N \times d},$$

where ε is a random noise matrix with independent rows. Under the assumption that R is sparse, a natural approach to estimating the model parameters R and D is to solve the $\mathcal{L}_1$-constrained optimization problem

$$\left(\widehat{R}, \widehat{D} \right) = \underset{(R,D)}{\operatorname{argmin}} \|X - RD\|_{2,2}^2 \qquad (1)$$
$$\text{subject to} \quad \|R\|_{1,1} \leq \lambda \quad \text{and} \quad \|D\|_{2,\infty} \leq 1.$$

where the minimization is over all $R \in [0, \infty)^{N \times K}$ and $D \in \mathbb{R}^{K \times d}$. Here, $\lambda \geq 0$ is a tuning parameter controlling the sparsity of the estimate $\widehat{R}$; the sparsity constraint is more often expressed as a penalty of $\lambda' \|R\|_{1,1}$ on the objective, but the equivalent constrained formulation above is more convenient for stating our theoretical results. Inspired by non-negative matrix factorization [32], the constraint that R is non-negative is included primarily for interpretability – it is often preferable to consider a negative multiple of a feature to be a different feature altogether – but this does not otherwise significantly affect the difficulty of the problem. The constraint $\|D\|_{2,\infty} \leq 1$ normalizes the size of the dictionary entries; without this, $\|R\|_{1,1}$ could become arbitrarily small by scaling D correspondingly.

Since matrix multiplication is bilinear, the optimization problem (1) is not jointly convex in R and D, but it is *biconvex*, i.e., convex in R when D is fixed and convex in D when R is fixed. This enables, in practice, a number of iterative optimization algorithms, typically based on alternating minimization, i.e., alternating between minimizing (1) in R and in D. Interestingly, recent work [33], [34] has shown that, despite being non-convex, the SDL problem is often well-behaved such that standard iterative optimization algorithms provably converge to global optima, even without multiple random restarts.

B. Convolutional Sparse Dictionary Learning

The CSDL problem considers a single data vector $X \in \mathbb{R}^N$.[1] The goal here is to find an approximate decomposition $X \approx R \otimes D$, where $R \in \mathbb{R}^{(N-n+1) \times K}$ is a sparse weight matrix and $D \in \mathbb{R}^{n \times K}$ is a dictionary of K patterns.

CSDL also begins with a frequentist model, the *temporal linear generative model*, proposed by [21], which instead supposes there exist (deterministic) R and D as above such that

$$X = R \otimes D + \varepsilon \in \mathbb{R}^N, \qquad (2)$$

where, again, ε is random noise. In this setting, it makes less sense to assume that the rows of ε are independent, an assumption we will avoid in our results. Under the assumption that R is sparse, a natural approach to estimating the model parameters R and D is to again solve an $\mathcal{L}_1$-constrained optimization problem, this time

$$(\widehat{R}, \widehat{D}) := \underset{(R,D)}{\operatorname{argmin}} \|X - R \otimes D\|_2^2 \qquad (3)$$
$$\text{subject to} \quad \|R\|_{1,1} \leq \lambda \quad \text{and} \quad \|D\|_{2,\infty} \leq 1,$$

where the minimization is over all $R \in [0, \infty)^{(N-n+1) \times K}$ and $D \in \mathbb{R}^{n \times K}$. Since multi-convolution is bilinear, the optimization problem (3) is again biconvex, and can, in practice, be solved by alternating minimization.

To summarize, the key differences between the IID and convolutional SDL problems setups are:

1) In the convolutional case, we seek a decomposition $X \approx R \otimes D$, whereas, in the IID case, we seek a decomposition $X \approx RD$. Unlike matrix multiplication, by which each row of R corresponds to a single row of X, multi-convolution allows each row of R to contribute to up to n consecutive rows of X, modeling, for example, temporally or spatially dependent features.

[1] As described in Section VII, this can be generalized in different ways to a multidimensional signal $X \in \mathbb{R}^{N \times d}$. For simplicity, in this paper, we only consider the case $d = 1$, which already presents interesting questions, whereas, the IID case with $d = 1$ becomes trivial.

2) In the convolutional case, the noise ε may have arbitrary dependencies, whereas, in the IID case, it must typically have independent rows.
3) The convolutional case involves an extra parameter n controlling the length of the dictionary entries,[2] whereas, in the IID case, $n = d$.

IV. Related Work

One body of work has been devoted to theoretical analysis of the non-convex optimization problem (1) in terms of which SDL is typically cast [35], [34], [33], with some consensus that despite being non-convex, this problem is often efficiently solvable in practice. Although, to the best of our knowledge, the CSDL optimization problem (3) has not been theoretically analyzed, our focus is on the statistical properties of the problem, and our upper bound implicitly assumes that the optimization problem (3) can be solved accurately.

More similar to our work, is the body of work analyzing the statistical properties of IID SDL. Here, [26] studied the generalizability of dictionary learning in terms of bounding the representation error of a learned dictionary on an independent test set from the same distribution as the training set. For certain algorithms, upper bounds have been shown on the risk (in Frobenius norm) of estimating the true dictionary D, up to permutation of the dictionary elements [28], [29]. More recently, [30], [31] proved what we believe are the first minimax lower bounds for SDL, in a variety settings, including a very general model without sparsity, a sparse model, and a sparse Gaussian model.

In addition to assuming the data are IID, these previous results essentially all also require restrictions on the structure of the dictionary. These restrictions have been stated in several forms, from incoherence assumptions [36], [28] and restricted isometry conditions [30], [31] to bounds on the Babel function [37], [26] of the dictionary, but all essentially boil down to requiring that the dictionary elements are not too similar or correlated. A notable feature of our upper bounds is that they make no assumptions whatsoever on the dictionary D; this is possible because our bounds apply to the *reconstruction* or *denoising* error of dictionary learning, rather than to the error of learning the dictionary itself. This can be compared to the fact that, in linear regression, bounds on prediction error can be derived with essentially no assumptions on the covariates [38], whereas, much stronger assumptions, to the effect that the covariates are not too strongly correlated, are needed to derive bounds for estimating the linear regression coefficients.

Finally, our results make minimal assumptions on the structure of the measurement noise; in particular, though we require the noise to have light tails (either sub-Gaussian or having some number of finite moments), we allow arbitrary dependence across the signal. This is important in practice because, in many applications, measurement errors are likely to be correlated with those in nearby portions of the signal.

V. Theoretical Results

We now present our theoretical results on the minimax average $\mathcal{L}_2$-risk of reconstructing $R \otimes D$ from $\widehat{R} \otimes \widehat{D}$, i.e., the quantity

$$\inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N}\mathbb{E}\left[\left\|\widehat{X} - R\otimes D\right\|_2^2\right], \tag{4}$$

where the infimum is taken over all estimators $\widehat{X}$ of $R\otimes D$ (i.e., all functions of the observation X). The quantity (4) characterizes the worst-case mean squared error of the average coordinate of $\widehat{X}$, for the best possible estimator $\widehat{X}$. Since it bounds within-sample reconstruction error, these results are primarily relevant for the applications of compression and denoising, rather than for learning an interpretable dictionary.

A. Upper Bound

Our first upper bound applies under the following assumptions:

A1) The TLGM model (2) holds for some $(R,D) \in \mathcal{S}_\lambda$; that is, $X = R\otimes D + \varepsilon \in \mathbb{R}^N$.

A2) The true sparseness $\|R\|_{1,1}$, or at least an upper bound on $\|R\|_{1,1}$, is known.

A3) The measurement noise is sub-Gaussian with parameter σ; that is, for each $i \in [N]$, $\mathbb{E}[\varepsilon_i] = 0$ and

$$\mathbb{E}\left[e^{t\varepsilon_i}\right] \le e^{t^2\sigma^2/2}, \qquad \text{for all } t\in\mathbb{R}.$$

The assumption that λ is known is likely unrealistic but, in practice, it may be feasible to upper bound λ, which is sufficient. The sub-Gaussian assumption precludes the noise having heavy tails, but is otherwise quite a mild assumption, allowing the elements of ε to exhibit a wide range of distributions and arbitrary dependencies.

Theorem 1: Under assumptions A1), A2), and A3) above, consider the $\mathcal{L}_1$-constrained dictionary estimator

$$\left(\widehat{R},\widehat{D}\right) = \operatorname*{argmin}_{(\widetilde{R},\widetilde{D})\in\mathcal{S}} \left\|X - \widetilde{R}\otimes\widetilde{D}\right\|_2 \quad \text{s.t.} \quad \|\widetilde{R}\|_{1,1} \le \lambda,$$

with any parameter value $\lambda \ge \|R\|_{1,1}$. Then, the reconstruction/denoising estimate $\widehat{R}\otimes\widehat{D}$ satisfies,

$$\frac{1}{N}\mathbb{E}\left[\|\widehat{R}\otimes\widehat{D} - R\otimes D\|_2^2\right] \le \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}. \tag{5}$$

We also consider, in the next theorem, how the bound differs when we replace the sub-Gaussian noise assumption with the following, somewhat broader, moment assumption:

A4) For some $p\in[1,\infty]$, the measurement noise has bounded p^{th} moment $\mu_p := \sup_{i\in[N]}\left(\mathbb{E}\left[\varepsilon_i^p\right]\right)^{1/p} < \infty$.

Theorem 2: Under assumptions A1), A2), and A4) above, consider the $\mathcal{L}_1$-constrained dictionary estimator

$$\left(\widehat{R},\widehat{D}\right) = \operatorname*{argmin}_{(\widetilde{R},\widetilde{D})\in\mathcal{S}} \left\|X - \widetilde{R}\otimes\widetilde{D}\right\|_2 \quad \text{s.t.} \quad \|\widetilde{R}\|_{1,1} \le \lambda,$$

[2] In fact, n can be distinct for each of the K features, suggesting a natural approach to learning *multi-scale* convolutional dictionaries, which are useful in many contexts. We leave this avenue for future work.

with any parameter value $\lambda \geq \|R\|_{1,1}$. Then, the reconstruction/denoising estimate $\widehat{R} \otimes \widehat{D}$ satisfies,

$$\frac{1}{N}\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2 \leq 4\lambda\mu_p N^{\frac{1-p}{p}} n^{\max\left\{0, \frac{p-2}{2p}\right\}}. \qquad (6)$$

Note that, while assumptions A3) and A4) are closely related, neither case covers the other. For $p \in [1, \infty)$, the assumption $\mu_p < \infty$ is strictly weaker than the sub-Gaussian assumption, and the convergence rate is correspondingly slower (polynomially in N). On the other hand, the assumption $\mu_\infty < \infty$ (i.e., that the ε_i's are bounded almost surely) is strictly stronger than the sub-Gaussian assumption, and the rate is correspondingly faster (by a multiplicative factor of $\sqrt{\log N}$).

The bounds (5) and (6) may be quite weak, since we assume $\lambda \geq \|R\|_{1,1}$, which may scale linearly with N. Thus, they are most useful in the 'ultra-sparse' settings $\|R\|_{1,1} \in o\left(\frac{N}{\sqrt{\log N}}\right)$ (under sub-Gaussian assumptions) or $\|R\|_{1,1} \in o\left(N^{\frac{1-p}{p}}\right)$ (under finite-moment assumptions). Here, assuming σ and n are fixed, the bounds imply $\mathcal{L}_2$-consistency in the average coordinate of the denoising estimator as $N \to \infty$. The strength of these results is that they make only very mild assumptions; specifically,

1) no independence or stationarity assumptions are made on the additive noise ε; it need only have somewhat bounded tails. This generality is especially desirable in the convolutional setting, as we may often expect experimental noise to correlate across nearby regions of the input, and may expect different distributions of noise in different regions of the input.
2) no assumptions whatsoever are made on R and D. While recovering R and D requires restrictions on the dictionary (e.g., that entries not be too correlated), our result suggests that sparsity of R is sufficient for effective compression and denoising. In this sense, Theorems 1 and 2 might be compared to so-called 'slow-rate' bounds for the LASSO [39], [38], which show that the LASSO with n samples and p covariates has mean squared prediction loss of order $O\left(\sqrt{\frac{\log p}{n}}\right)$, even with no assumptions on the design of the covariates.

B. Lower Bound

We now present a minimax lower bound showing that the upper bound rate in Theorem 1 is tight in terms of the sparsity λ, noise level σ, and signal length N. For simplicity, in this section, we only consider a single dictionary element ($K = 1$). Our lower bound is based on the following standard information-theoretic lower bound for estimating the mean of the $\mathcal{L}_1$-constrained Gaussian sequence model:

Lemma 3: (Corollary 5.16 of [40] Consider the $\mathcal{L}_1$-constrained Gaussian sequence model, in which we observe $Y = \theta + \varepsilon \in \mathbb{R}^d$, where $\varepsilon \sim \mathcal{N}(0_d, \sigma^2 I_d)$ and we know that $\|\theta\|_1 \leq \lambda$. Then, we have the minimax lower bound [3]

$$\inf_{\widehat{\theta}} \sup_{\|\theta\|_1 \leq \lambda} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \geq \frac{\lambda}{8}\min\left\{\lambda, \sigma\sqrt{\log d}\right\}$$

for estimating the model parameter θ.
The min here reflects the fact that, in the extremely sparse or noisy regime $\lambda \leq \sigma\sqrt{\log d}$, the trivial estimator $\widehat{\theta} = 0$ becomes optimal, with worst-case $\mathcal{L}_2$-risk at most λ^2.

By showing that CSDL estimators can be used to construct estimators for the $\mathcal{L}_1$-constrained Gaussian sequence problem, we are able to use Lemma 3 to prove the following lower bound on CSDL:

Theorem 4: (Minimax Lower Bound on CSDL Denoising): Consider the additive noise model $X = R \otimes D + \varepsilon \in \mathbb{R}^N$. Then, there exists an identically (but possibly dependently) distributed Gaussian noise pattern ε, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i \in [N]$, such that the following minimax lower bound on the average $\mathcal{L}_2$ reconstruction/denoising risk holds:

$$\begin{aligned}\inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N}\mathbb{E}\left[\left\|\widehat{X} - R \otimes D\right\|_2^2\right] \\ \geq \frac{\lambda}{8N}\min\left\{\lambda n, \sigma\sqrt{\log(N-n+1)}\right\} \qquad (7)\end{aligned}$$

Notably, this lower bound holds even if we assume the dictionary D is known, showing that the difficulty of the CSDL reconstruction problem is dominated by the difficulty of estimating R. Indeed, the $\sqrt{n}$ factor in the upper bound stems not from the need to estimate D but rather from the fact that $\|D\|_1$, which appears as a scaling factor when estimating R, may be as large as $\sqrt{n}\|D\|_2 = \sqrt{n}$. Again, in the extremely sparse or noisy setting $\lambda\sqrt{n} \leq \sigma\sqrt{\log(N-n+1)}$, the trivial estimator $\widehat{R} = 0$ achieves the optimal risk of order $\lambda^2 n/N$.

Proof Sketch: The full proof of Theorem 4 is given in the appendix, but a briefly sketch is as follows. To prove a lower bound, we can suppose D is known. Specifically (as suggested by the upper bound proof, which relies on the fact that $\|D\|_1 \leq \sqrt{n}$, with equality if and only if D is uniform) we set D to be the uniform unit vector

$$D = \frac{1}{\sqrt{n}}[1, 1, 1, ..., 1]^T \in \mathbb{R}^n.$$

We show that, if D is known, then any estimator $\widehat{X}$ of $R \otimes D$ given $R \otimes D + \varepsilon$ can be used to construct an estimator $\widehat{R}$ for R given $R + \zeta$, where $\zeta \sim \mathcal{N}\left(0, \sigma^2 I\right)$. This latter problem is precisely equivalent to estimating the parameter of a Gaussian sequence model, so that, for any estimator $\widehat{R}$ of R, Lemma 3 lower bounds

$$\sup_{\|R\|_{1,1} \leq \lambda} \mathbb{E}\left[\left\|\widehat{R} - R\right\|_2^2\right].$$

[3] Although Corollary 5.16 is written for a sequence taking values in all of $\mathbb{R}^d$, its proof involves only points in non-negative multiples of the hypercube vertices $\{0, 1\}^d$. Hence, the result will apply despite our additional constraint that R is non-negative.

The proof then consists of upper bounding $\mathbb{E}\left[\left\|\widehat{R}-R\right\|_2^2\right]$ in terms of $\mathbb{E}\left[\left\|\widehat{X}-R\otimes D\right\|_2^2\right]$ and verifying that the sub-Gaussian noise condition is satisfied.

VI. Empirical Results

In this section, we present numerical experiments on synthetic data, with the goal of verifying the convergence rates derived in the previous section.

Optimization Algorithm: Since the focus of this paper is on the *statistical* properties of CSDL, we assume the estimator $(\widehat{R}_\lambda, \widehat{D}_\lambda)$ defined by the optimization problem 3 can be computed to high precision using a very simple alternating projected gradient descent algorithm, which iteratively performs the following four steps: 1) gradient step with respect to D, 2) project the columns of D onto the unit sphere, 3) gradient step with respect to R, and 4) project R (with respect to Frobenius norm) into the intersection of the non-negative orthant and the $\mathcal{L}_{1,1}$ ball of radius λ. The parameters of this algorithm are as follows:

- Number of iterations: 200
- Step size: $\frac{1}{100\sqrt{i}}$, where $i \in [200]$ is the iteration number

We then use the reconstruction estimate $\widehat{X}_\lambda := \widehat{R}_\lambda \otimes \widehat{D}_\lambda$ to estimate $R\otimes D$.

Comparisons: We compare the error of the optimal CSDL estimator $\widehat{X}_s$ to the following:

- Our theoretical upper bound (Inequality (5)).
- Our lower upper bound (Inequality (7)).
- Error of the trivial estimator $\widehat{X}_0 = 0$.
- Error of the original signal $\widehat{X}_\infty = X$.

Experimental Setup: In all experiments, unless noted otherwise, the data are generated using the following parameter settings:

- Signal length $N = 1000$
- $\mathcal{L}_1$-Sparsity $s = \|R\|_{1,1} = 100$
- Dictionary element length $n = 10$
- Dictionary size $K = 2$
- Noise level $\sigma = 0.1$

All results presented are averaged over 100 IID trials.[4] In each trial, the K dictionary elements (columns of D) are sampled independently and uniformly from the $\mathcal{L}_2$ unit sphere in $\mathbb{R}^n$. R is generated by initializing $R = 0_{(N-n+1)\times K}$ and then repeatedly adding 1 to uniformly random coordinates of R until the desired value of $\|R\|_{1,1}$ is achieved. MATLAB code for reproducing our experiments is available at `https://github.com/sss1/convolutional_dictionary`.

Experiment 1: Our first experiment studies the relationship between the length N of the signal and the true $\mathcal{L}_1$-sparsity $\|R\|_{1,1}$ of the data. Figure 1 shows error as a function of N for logarithmically spaced values between 10^2 and 10^4, with $\|R\|_{1,1}$ scaling as constant $\|R\|_{1,1} = 5$, square-root $\|R\|_{1,1} = \left\lfloor\sqrt{N}\right\rfloor$, and linearly $\|R\|_{1,1} = \lfloor N/2\rfloor$. The results confirm the main prediction of our theoretical results, namely that the error of the CSDL estimator using the optimal tuning parameter $\lambda = \|R\|_{1,1}$ lies between the lower and upper bounds, converging at a rate of order s/N (ignoring log factors). As a result, the estimator is inconsistent when s grows linearly with N, in which case there is no benefit to applying CSDL to denoise the signal over using the original signal $\widehat{X}_\infty = X$, even though the latter is never consistent.[5] On the other hand, if s scales sub-linearly, CSDL is consistent and outperforms both trivial estimator (although, of course, the trivial estimator $\widehat{X}_0 = 0$ is also consistent in this setting).

Experiment 2: Our second experiment studies the dependence of the error on the length n of the dictionary elements. For this experiment, we considered two ways of generating D: (1) with normalized Gaussian columns (as in other experiments), and (2) fixed to be the uniform $\mathcal{L}_2$-unit vector

$$D = \frac{1}{\sqrt{n}}[1,1,1,...,1]^T \in \mathbb{R}^n.$$

This latter choice was chosen because the proofs of our lower and upper bounds both suggest this may elicit worst-case results. Figure 2 shows error as a function of n for logarithmically spaced values between $10^{1/2}$ and 10^2, with D Gaussian in the first panel and D uniform in the second panel. As predicted by our theoretical results, the error of the CSDL estimator using the optimal tuning parameter $\lambda = \|R\|_{1,1}$ lies between the lower and upper bounds and scales as $\sqrt{n}$.

Experiment 3: Our third experiment studies the sensitivity of the estimator $\widehat{X}_\lambda$ to its tuning parameter. Figure 3 shows error as a function of λ for logarithmically spaced valued between 10^{-2} and 10^4.

VII. Conclusions and Future Work

Theorems 1 and 4 together have several interesting consequences. Firstly, in a fairly broad setting, (ignoring dependence on K) the minimax average $\mathcal{L}_2$ risk for CSDL reconstruction is of order

$$\frac{\lambda}{N}\min\left\{\lambda n, \sigma\sqrt{n\log N}\right\}.$$

Hence, for a constant noise level σ and dictionary element size n, consistent reconstruction is possible if and only if the sparsity $\lambda \in o\left(\frac{N}{\log N}\right)$. Second, this holds regardless of whether D is known or unknown; a similar phenomenon has been observed in classic SDL, where [29] showed that upper bounds are of the same rate as lower bounds for the case where the dictionary is known beforehand (a classic problem known as *sparse recovery*).

A. Future Work

There remain several natural open questions about the statistical properties of CSDL.

[4] We initially plotted 95% confidence intervals for each point based on asymptotic normality of the empirical $\mathcal{L}_2$ error. However, intervals were consistently smaller than markers sizes, and so omitted them to avoid clutter.

[5] Even if s grows linearly with N, as long as s/N is small, CSDL may still be useful for compression, if a constant-factor loss is acceptable.

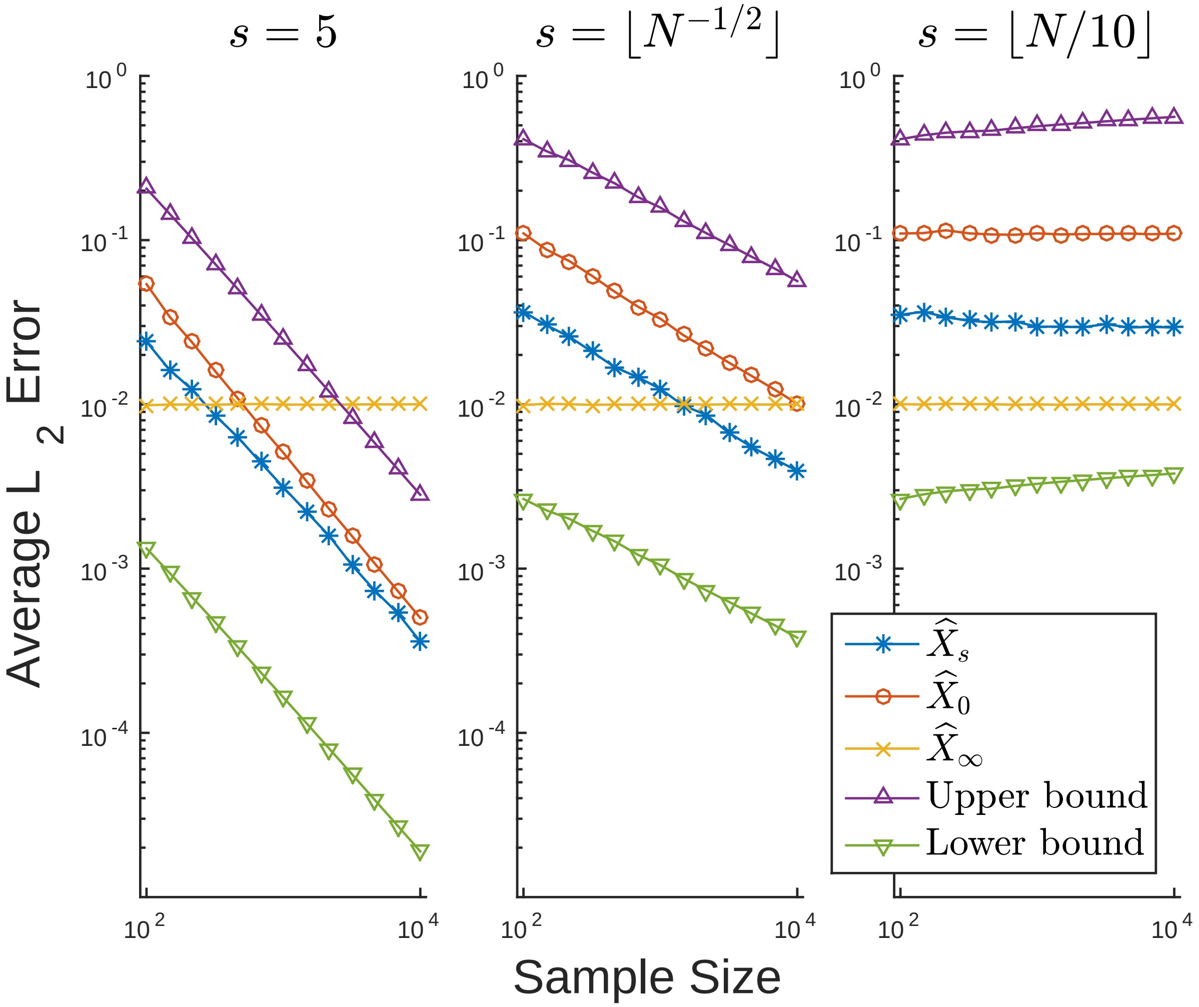

Fig. 1. Experiment 1: Average $\mathcal{L}_2$-error as a function of signal length N, with sparsity scaling as $\|R\|_{1,1} = 5$ (first panel), $\|R\|_{1,1} = \left\lfloor \sqrt{N} \right\rfloor$ (second panel), and $\|R\|_{1,1} = \lfloor N/2 \rfloor$ (third panel).

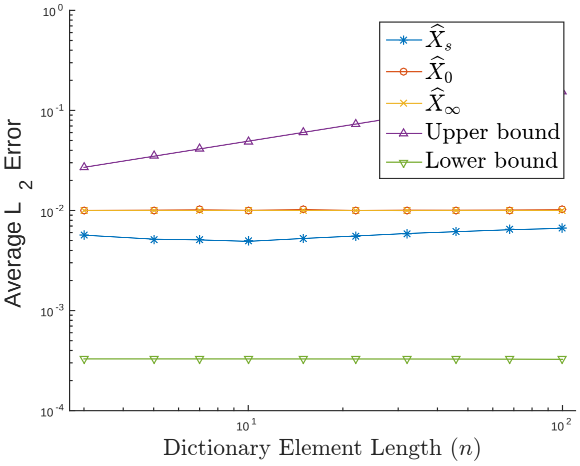

Fig. 2. Experiment 2: Average $\mathcal{L}_2$-error as a function of dictionary element length n.

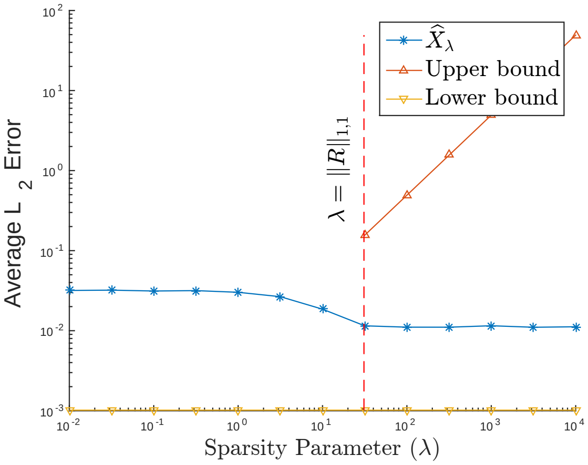

Fig. 3. Experiment 3: Average $\mathcal{L}_2$-error as a function of the tuning parameter λ of $\widehat{X}_\lambda$. The dashed line indicates the $\mathcal{L}_1$-sparsity $\|R\|_{1,1} = \left\lfloor \sqrt{N} \right\rfloor = 33$. Note that the upper bound only applies when $\lambda \geq \|R\|_{1,1}$.

First, can our bounds be tightened to have matching dependence on the number K of dictionary elements? We believe this can likely be achieved via a straightforward refinement of our lower bound proof.

Second, how do rates extend to the case of a multidimensional signal $X \in \mathbb{R}^{N \times d}$? There are multiple possible extensions of CSDL to this case. For example, the simplest is to make $R \in \mathbb{R}^{(N-n+1) \times K \times d}$ and $D \in \mathbb{R}^{n \times K \times d}$ each become 3-tensors and learn separate dictionary and weight matrices in each dimension, but another interesting approach may be to keep $R \in \mathbb{R}^{(N-n+1) \times K}$ as a matrix and to make $D \in \mathbb{R}^{n \times K \times d}$ a 3-tensor (and to generalize the multi-convolution operator appropriately), such that the positions encoded by R are shared across dimensions, while different dictionary elements are learned in each dimension. Although a somewhat more restrictive model, this latter approach would likely have the advantage that statistical risk would *decrease* with d, as data from multiple dimensions could contribute to the difficult problem of estimating R.

A third direction may be to consider a model with secondary spatial structure, such as correlations between occurrences of dictionary elements; for example, in speech data, consecutive phonemes are likely to be highly dependent. This might be better modeled in a Bayesian framework, where R is itself randomly generated with a certain (unknown) dependence structure between its columns.

Finally, under what assumptions can we bound the risk of estimating a true dictionary D (which may be the main goal in many scientific applications)? In classic SDL, this is understood to require incoherence or similar assumptions on D that can be somewhat unrealistic, especially for large overcomplete dictionaries. In the convolutional case, this assumption may need to be even stronger, due to the potential for interactions between different alignments of the rows of D in $R \otimes D$.

Acknowledgements: This research is supported in part by an NSF Graduate Research Fellowship to S.S. under Grant No. DGE-1252522. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[2] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[3] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

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

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[6] M. Aharon, M. Elad, and A. Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[7] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on image processing*, vol. 17, no. 1, pp. 53–69, 2008.

[8] G. Peyré, "Sparse modeling of textures," *Journal of Mathematical Imaging and Vision*, vol. 34, no. 1, pp. 17–31, 2009.

[9] G. Peyré, J. M. Fadili, and J.-L. Starck, "Learning adapted dictionaries for geometry and texture separation," in *SPIE Wavelets XII*, vol. 6701. SPIE, 2007, p. 67011T.

[10] O. Bryt and M. Elad, "Compression of facial images using the k-svd algorithm," *Journal of Visual Communication and Image Representation*, vol. 19, no. 4, pp. 270–282, 2008.

[11] K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," *arXiv preprint arXiv:1010.3467*, 2010.

[12] Y. Zhu and S. Lucey, "Convolutional sparse coding for trajectory reconstruction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 529–540, 2015.

[13] M. Protter and M. Elad, "Super resolution with probabilistic motion estimation," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1899–1904, 2009.

[14] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1823–1831.

[15] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Machine Learning for Signal Processing, 2007 IEEE Workshop on*. IEEE, 2007, pp. 431–436.

[16] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1025–1031, 2011.

[17] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5. IEEE, 1999, pp. 2443–2446.

[18] W. Smit and E. Barnard, "Continuous speech recognition with sparse coding," *Computer Speech & Language*, vol. 23, no. 2, pp. 200–219, 2009.

[19] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[20] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in neural information processing systems*, 2007, pp. 41–48.

[21] B. A. Olshausen, "Sparse codes and spikes," *Probabilistic Models of the Brain*, p. 257, 2002.

[22] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 391–398.

[23] B. Wohlberg, "Efficient convolutional sparse coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7173–7177.

[24] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5135–5143.

[25] F. Huang and A. Anandkumar, "Convolutional dictionary learning through tensor factorization," in *Feature Extraction: Modern Questions and Challenges*, 2015, pp. 116–129.

[26] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *Journal of Machine Learning Research*, vol. 12, no. Nov, pp. 3259–3281, 2011.

[27] Q. Geng and J. Wright, "On the local correctness of âĎŞ 1-minimization for dictionary learning," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 3180–3184.

[28] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Conference on Learning Theory*, 2014, pp. 123–137.

[29] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Conference on Learning Theory*, 2014, pp. 779–806.

[30] A. Jung, Y. C. Eldar, and N. Görtz, "Performance limits of dictionary learning for sparse coding," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 765–769.

[31] A. Jung, Y. C. Eldar, and N. Görtz, "On the minimax risk of dictionary

learning," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1501–1515, 2016.

[32] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[33] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," in *Sampling Theory and Applications (SampTA), 2015 International Conference on*. IEEE, 2015, pp. 407–410.

[34] ——, "When are nonconvex problems not scary?" *arXiv preprint arXiv:1510.06096*, 2015.

[35] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[36] R. Jenatton, R. Gribonval, and F. Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," *arXiv preprint arXiv:1210.0685*, 2012.

[37] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[38] S. Chatterjee, "Assumptionless consistency of the lasso," *arXiv preprint arXiv:1303.5817*, 2013.

[39] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[40] P. Rigollet, "18. s997: High dimensional statistics," *Lecture Notes), Cambridge, MA, USA: MIT OpenCourseWare*, 2015.

[41] W. Beckner, "Inequalities in Fourier analysis," *Annals of Mathematics*, pp. 159–182, 1975.

Appendix

Here, we present the proofs of our main theoretical results.

A. Upper Bound Proof

Theorem 1: Suppose the following hold:

A1) The TLGM model (2) holds for some $(R, D) \in \mathcal{S}_\lambda$; that is, $X = R \otimes D + \varepsilon \in \mathbb{R}^N$.

A3) The measurement noise is sub-Gaussian with parameter σ; that is, for each $i \in [N]$, $\mathbb{E}[\varepsilon_i] = 0$ and

$$\mathbb{E}\left[e^{t\varepsilon_i}\right] \leq e^{t^2\sigma^2/2}, \qquad \text{for all } t \in \mathbb{R}.$$

Consider the $\mathcal{L}_1$-constrained dictionary estimator

$$\left(\widehat{R}, \widehat{D}\right) = \operatorname*{argmin}_{(\widetilde{R}, \widetilde{D}) \in \mathcal{S}} \left\| X - \widetilde{R} \otimes \widetilde{D} \right\|_2 \quad \text{s.t.} \quad \|\widetilde{R}\|_{1,1} \leq \lambda,$$

with any parameter value $\lambda \geq \|R\|_{1,1}$. Then, the reconstruction/denoising estimate $\widehat{R} \otimes \widehat{D}$ satisfies,

$$\frac{1}{N}\mathbb{E}\left[\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}.$$

Proof: Since $\|R\|_{1,1} \leq \lambda$, (R, D) is a feasible point for the optimization problem, and so

$$\|X - \widehat{R} \otimes \widehat{D}\|_2 \leq \|X - R \otimes D\|_2.$$

Rearranging this and using the fact that $X = R \otimes D + \varepsilon$, we have

$$\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2 \leq 2\langle \varepsilon, \widehat{R} \otimes \widehat{D} - R \otimes D\rangle. \tag{8}$$

By Hölder's inequality,

$$\langle \varepsilon, \widehat{R} \otimes \widehat{D} - R \otimes D\rangle \leq \|\varepsilon\|_\infty \|\widehat{R} \otimes \widehat{D} - R \otimes D\|_1.$$

Then, by the triangle inequality and Young's inequality for convolutions (see Theorem 1 of [41]),

$$\begin{aligned}
\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_1 &= \left\| \sum_{k=1}^K \widehat{R}_k * \widehat{D}_k - R_k * D_k \right\|_1 \\
&\leq \sum_{k=1}^K \|\widehat{R}_k * \widehat{D}_k\|_1 + \|R_k * D_k\|_1 \\
&\leq \sum_{k=1}^K \|\widehat{R}_k\|_1 \|\widehat{D}_k\|_1 + \|R_k\|_1 \|D_k\|_1 \\
&\leq \sqrt{n} \sum_{k=1}^K \|\widehat{R}_k\|_1 + \|R_k\|_1 \\
&= \sqrt{n}\left(\|\widehat{R}\|_{1,1} + \|R_k\|_{1,1}\right) \leq 2\lambda\sqrt{n},
\end{aligned}$$

where we used that facts that $\|\widehat{D}_k\|_2 = \|D_k\|_2 = 1$ and $\widehat{D}_k, D_k \in \mathbb{R}^n$, so that $\|\widehat{D}_k\|_1, \|D_k\|_1 \leq \sqrt{n}$. Combining this series of inequalities with inequality (8) gives

$$\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2 \leq 4\lambda\|\varepsilon\|_\infty\sqrt{n}.$$

Theorem 1 now follows by a standard bound on the expected supremum of a sub-Gaussian process (see, e.g., Lemma 4 of [38]), which implies

$$\mathbb{E}\left[\|\varepsilon\|_\infty\right] \leq \sigma\sqrt{2\log(2N)}.$$

■

Theorem 2: Suppose the following hold:

A1) The TLGM model (2) holds for some $(R, D) \in \mathcal{S}_\lambda$; that is, $X = R \otimes D + \varepsilon \in \mathbb{R}^N$.

A4) For some $p \in [1, \infty]$, the measurement noise has bounded p^{th} moment $\mu_p := \sup_{i \in [N]} \mathbb{E}\left[\varepsilon_i^p\right] < \infty$.

Consider the $\mathcal{L}_1$-constrained dictionary estimator

$$\left(\widehat{R}, \widehat{D}\right) = \operatorname*{argmin}_{(\widetilde{R}, \widetilde{D}) \in \mathcal{S}} \left\| X - \widetilde{R} \otimes \widetilde{D} \right\|_2 \quad \text{s.t.} \quad \|\widetilde{R}\|_{1,1} \leq \lambda,$$

with any parameter value $\lambda \geq \|R\|_{1,1}$. Then, the reconstruction/denoising estimate $\widehat{R} \otimes \widehat{D}$ satisfies,

$$\frac{1}{N}\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2 \leq 4\lambda\mu_p N^{\frac{1-p}{p}} n^{\max\left\{0, \frac{p-2}{2p}\right\}}. \tag{9}$$

Proof: As in the sub-Gaussian case, the proof relies on the "basic inequality" (8), which follows from the construction of the $\mathcal{L}_1$-constrained dictionary estimator and the assumption that $\|R\|_{1,1} \leq \lambda$. By Hölder's inequality,

$$\langle \varepsilon, \widehat{R} \otimes \widehat{D} - R \otimes D\rangle \leq \|\varepsilon\|_p \|\widehat{R} \otimes \widehat{D} - R \otimes D\|_q,$$

where $q = \frac{p}{p-1} \geq 1$.

If $q \geq 2$ (i.e., if $p \leq 2$), then $\|D_k\|_q \leq \|D_k\|_2 = 1$ and $\|\widehat{D}_k\|_q \leq \|\widehat{D}_k\|_2 = 1$. Otherwise

$$\|D_k\|_q \leq n^{1/q - 1/2}\|D_k\|_2 = n^{1/q-1/2} = n^{\frac{p-2}{2p}},$$

and, similarly, $\|\widehat{D}_k\|_q \leq n^{\frac{p-2}{2p}}$. In short,

$$\|D_k\|_q, \|\widehat{D}_k\|_q \leq n^{\max\left\{0, \frac{p-2}{2p}\right\}}.$$

Hence, by the triangle inequality and Young's inequality for convolutions (see Theorem 1 of [41]),

$$\begin{aligned}
\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_q &= \left\| \sum_{k=1}^{K} \widehat{R}_k * \widehat{D}_k - R_k * D_k \right\|_q \\
&\leq \sum_{k=1}^{K} \|\widehat{R}_k * \widehat{D}_k\|_q + \|R_k * D_k\|_q \\
&\leq \sum_{k=1}^{K} \|\widehat{R}_k\|_1 \|\widehat{D}_k\|_q + \|R_k\|_1 \|D_k\|_q \\
&\leq n^{\max\left\{0, \frac{p-2}{2p}\right\}} \sum_{k=1}^{K} \|\widehat{R}_k\|_1 + \|R_k\|_1 \\
&= \left(\|\widehat{R}\|_{1,1} + \|R_k\|_{1,1} \right) n^{\max\left\{0, \frac{p-2}{2p}\right\}} \\
&\leq 2\lambda n^{\max\left\{0, \frac{p-2}{2p}\right\}},
\end{aligned}$$

Combining this series of inequalities with inequality (8) gives

$$\|\widehat{R} \otimes \widehat{D} - R \otimes D\|_2^2 \leq 4\lambda \|\varepsilon\|_2 n^{\max\left\{0, \frac{p-2}{2p}\right\}}.$$

Theorem 2 now follows by observing that

$$\mathbb{E}\left[\|\varepsilon\|_p\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|_p^p\right]\right)^{1/p} \leq \mu_p N^{1/p}.$$

■

B. Lower Bound Proof

Theorem 4: (Minimax Lower Bound on CSDL Denoising): Consider the additive noise model $X = R \otimes D + \varepsilon \in \mathbb{R}^N$. Then, there exists an identically (but possibly dependently) distributed Gaussian noise pattern ε, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i \in [N]$, such that the following minimax lower bound on the average $\mathcal{L}_2$ reconstruction/denoising risk holds:

$$\begin{aligned}
&\inf_{\widehat{X}} \sup_{(R,D) \in \mathcal{S}_\lambda} \frac{1}{N} \mathbb{E}\left[\left\| \widehat{X} - R \otimes D \right\|_2^2 \right] \\
&\geq \frac{\lambda}{8N} \min\left\{ \lambda n, \sigma \sqrt{n \log(N - n + 1)} \right\}
\end{aligned}$$

Proof: We first introduce a linear operator, denoted T_D, which is central to the proof, and show its relevant properties. Let

$$D = \frac{1}{\sqrt{n}} [1, 1, 1, \ldots, 1]^T \in \mathbb{R}^n$$

be the non-negative uniform $\mathcal{L}_2$-unit vector in $\mathbb{R}^n$, and let

$$T_D = \begin{bmatrix}
D_1 & 0 & 0 & \cdots & 0 & 0 \\
D_2 & D_1 & 0 & \cdots & 0 & 0 \\
D_3 & D_2 & D_1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & D_n & D_{n-1} \\
0 & 0 & 0 & \cdots & 0 & D_n
\end{bmatrix} \in \mathbb{R}^{N \times (N-n+1)},$$

denote its convolution matrix, so that

$$r \otimes D = T_D r \in \mathbb{R}^N, \quad \forall r \in \mathbb{R}^{N-n+1}.$$

Let $\mathcal{I} := T_D\left([0, \infty)^{N-n+1}\right) \subseteq \mathbb{R}^N$ denote the image of $[0, \infty)^{N-n+1}$ under T_D. It is easy to check that T_D has full rank $N - n + 1$. Hence, it has a left inverse $T_D^{-1} : \mathcal{I} \to \mathbb{R}^{N-n+1}$ such that, $\forall x \in \mathcal{I}$, $T_D^{-1} T_D x = x$. Since R and D are both non-negative, one can also check that

$$\|T_D \widetilde{R}\|_2 \geq \|\widetilde{R}\|_2.$$

Hence, for all $x \in \mathcal{I}$,

$$\|T_D^{-1} x\|_2 \leq \|x\|_2. \tag{10}$$

Noting that $\mathcal{I}$ is a convex set, let $\Pi_{\mathcal{I}} : \mathbb{R}^N \to \mathcal{I}$ denote the $\mathcal{L}_2$-projection operator onto $\mathcal{I}$; i.e.,

$$\Pi_{\mathcal{I}}(x) = \operatorname*{argmin}_{y \in \mathcal{I}} \|y - x\|_2, \quad \forall x \in \mathbb{R}^N.$$

Having defined the necessary quantities, we now proceed to the main proof.

Suppose we have an estimator $\widehat{X}$ of $R \otimes D$ given $R \otimes D + \varepsilon$ (so that $\widehat{X}$ is a function from $\mathbb{R}^N$ to $\mathbb{R}^N$). Then, given an observation $Y = R + \zeta \in \mathbb{R}^{N-n+1}$, where $\zeta \sim \mathcal{N}(0, \sigma^2 I)$, define the estimator

$$\widehat{R} = T_D^{-1}\left(\Pi\left(\widehat{X}(Y \otimes D)\right)\right)$$

of R. Then, by inequality (10) and the fact that $T_D R \in \mathcal{I}$,

$$\begin{aligned}
\left\|\widehat{R} - R\right\|_2 &= \left\|T_D^{-1}\left(\Pi_{\mathcal{I}}\left(\widehat{X}((R + \zeta) \otimes D)\right) - T_D R\right)\right\|_2 \\
&\leq \left\|\Pi_{\mathcal{I}}\left(\widehat{X}((R + \zeta) \otimes D)\right) - T_D R\right\|_2 \\
&\leq \left\|\widehat{X}((R + \zeta) \otimes D) - T_D R\right\|_2 \qquad (11) \\
&= \left\|\widehat{X}(R \otimes D + \zeta \otimes D) - R \otimes D\right\|_2 \qquad (12)
\end{aligned}$$

Since $\zeta \sim \mathcal{N}(0, \sigma^2 I)$, for any $t \in \mathbb{R}$, each $\mathbb{E}[e^{t\zeta_j}] = e^{\sigma^2 t^2/(2n)}$. By construction of D and ζ, each coordinate of $T_D \zeta$ is the sum of at most n independent centered normal random variables with variance $\frac{\sigma^2}{n}$. Thus, each

$$\mathbb{E}\left[e^{t(T_D \zeta)_j}\right] = e^{\sigma^2 t^2/2},$$

and so the noise $\zeta \otimes D = T_D \zeta$ satisfies the sub-Gaussian assumption with parameter σ. Therefore, since the estimator $\widehat{X}$ is arbitrary, to lower bound the minimax rate in question, it suffices to lower bound

$$\frac{1}{N} \mathbb{E}\left[\left\|\widehat{X}(R \otimes D + \zeta \otimes D) - R \otimes D\right\|_2^2\right].$$

Indeed, combining Lemma 3 with inequality (12) gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{X}(R \otimes D + \zeta \otimes D) - R \otimes D\right\|_2^2\right] \\
&\geq \mathbb{E}\left[\left\|\widehat{R} - R\right\|_2^2\right] \\
&\geq \frac{\lambda}{8} \min\left\{\lambda n, \sigma \sqrt{\log(N - n + 1)}\right\}.
\end{aligned}$$

The theorem now follows by dividing through by N. ■