Minimax Reconstruction Risk of Convolutional Sparse Dictionary Learning

Shashank Singh
sss1@cs.cmu.edu
Machine Learning Department

Barnabás Póczos
bapoczos@cs.cmu.edu
Machine Learning Department
Carnegie Mellon University

Jian Ma
jianma@cs.cmu.edu
Computational Biology Department

Abstract

Sparse dictionary learning (SDL) has become a popular method for learning parsimonious representations of data, a fundamental problem in machine learning and signal processing. While most work on SDL assumes a training dataset of independent and identically distributed (IID) samples, a variant known as convolutional sparse dictionary learning (CSDL) relaxes this assumption to allow dependent, non-stationary sequential data sources. Recent work has explored statistical properties of IID SDL; however, the statistical properties of CSDL remain largely unstudied. This paper identifies minimax rates of CSDL in terms of reconstruction risk, providing both lower and upper bounds in a variety of settings. Our results make minimal assumptions, allowing arbitrary dictionaries and showing that CSDL is robust to dependent noise. We compare our results to similar results for IID SDL and verify our theory with synthetic experiments.

1 Introduction

Many problems in machine learning and signal processing can be reduced to, or greatly simplified by, finding a parsimonious representation of a dataset. In recent years, partly inspired by models of visual and auditory processing in the brain [33, 27, 44], the method of *sparse dictionary learning* (SDL; also called *sparse coding*) has become a popular way of learning such a representation, encoded as a sparse linear combination of learned *dictionary elements*, or patterns recurring throughout the data.

SDL has been widely applied to image processing problems such as denoising, demosaicing, and inpainting [22, 12, 2, 28, 36], separation [37], compression [9], object recognition [21, 47], trajectory reconstruction [53], and super-resolution reconstruction [38, 14]. In audio processing, it has been used for structured [40] and unstructured [17] denoising, compression [13], speech recognition [43], speaker separation [42], and music genre classification [51]. SDL has also been used for both supervised [31, 30] and unsupervised [4] feature learning in general domains.

The vast majority of SDL literature assumes a training dataset consisting of a large number of independent and identically distributed (IID) samples. Additionally, the dimension of these samples must often be limited due to computational constraints. On the other hand, many sources of sequential data, such as images or speech, are neither independent nor identically distributed. Consequently, most of the above SDL applications rely on first segmenting data into small (potentially overlapping) "patches", which are then treated as IID. SDL is then applied to learn a "local" dictionary for sparsely representing patches. This approach suffers from two major drawbacks. First, the learned dictionary needs to be quite large, because the model lacks translation-invariance within patches. Second, the model fails to capture dependencies across distances larger than patch sizes, resulting in a less sparse representation. These factors in turn limit computational and statistical performance of SDL.

To overcome this, recent work has explored *convolutional sparse dictionary learning* (CSDL; also called *convolutional sparse coding*), in which a global dictionary is learned directly from a sequential data source, such as a large image or a long, potentially non-stationary, time series [32, 42, 43]. In the past few years, CSDL algorithms have demonstrated improved performance on several of the above applications, compared to classical (IID) SDL [8, 50, 15, 16, 11].

The **main goal** of this work is to understand the statistical properties of CSDL, in terms of *reconstruction* (or *denoising*) error (i.e., the error of reconstructing a data sequence from a learned convolutional sparse dictionary decomposition, a process widely used for denoising and compression [13, 22, 9, 53, 14]). We do this by (a) upper bounding the reconstruction risk of an established estimator [32]

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

and (b) lower bounding, in a minimax sense, the risk of reconstructing a data source constructed sparsely from a convolutional dictionary. The emphasis in this paper is on proving results under minimal assumptions on the data, in the spirit of recent results on "assumptionless" consistency of the LASSO [10, 39]. As such, we make no assumptions whatsoever on the dictionary or encoding matrices (such as restricted eigenvalue or isometry properties).

Compared even to these "assumptionless" results, we consider dropping yet another assumption, namely that noise is independently distributed. In many of the above applications (such as image demosaicing or inpainting, or structured denoising), noise is strongly correlated across the data, and yet dictionary learning approaches nevertheless appear to perform consistent denoising. To the best of our knowledge, this phenomenon has not been explained theoretically. One likely reason is that this phenomenon does *not* occur in basis pursuit denoising, LASSO, and many related compressed sensing problems, where consistency is usually not possible under arbitrarily dependent noise. In the context of SDL, tolerating dependent noise is especially important in the convolutional setting, where coordinates of the data are explicitly modeled as being spatially or temporally related. However, this phenomenon is also not apparent in the recent work of Papyan et al. [35], which, to the best of our knowledge, is the only work analyzing theoretical guarantees of CSDL, but considers only deterministic bounded noise and studies recovery of the true dictionary, rather than reconstruction error. Under sufficient sparsity, our results imply consistency (in reconstruction error) of CSDL, even under arbitrary noise dependence.

Paper Organization: Section 2 defines notation needed to formalize CSDL and our results. Section 3 provides background on IID and convolutional SDL. Section 4 reviews related theoretical work. Section 5 contains statements of our main theoretical results (proven in the Appendix), which are experimentally validated in Section 6 and discussed further in Section 7, with suggestions for future work.

2 Notation

Here, we define some notation used throughout 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

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

R_k and D_k denote the k^{th} columns of R and D, respectively, and $*$ 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 data is constructed from the encoding matrix R and the dictionary D. We note that, like matrix multiplication, multi-convolution is a bilinear operation.

Matrix norms: For any matrix $A \in \mathbb{R}^{n\times m}$ and $p, q \in [0, \infty]$, the $\mathcal{L}_{p,q}$ norm[1] of A is

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

(or the corresponding limit if p or q is 0 or ∞) denotes the q-norm of the vector whose entries are p-norms of columns of A. Note that $\|\cdot\|_{2,2}$ is precisely the Frobenius norm.

Problem Domain: For positive integers N, n, and K, we use

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

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 $\lambda \ge 0$, we further use

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

to denote the $\mathcal{L}_{1,1}$-constrained version of this domain. Note that both $\mathcal{S}$ and $\mathcal{S}_\lambda$ are convex sets.

3 Background: IID and Convolutional Sparse Dictionary Learning

We now review the standard formulations of IID and convolutional sparse dictionary learning.

3.1 IID Sparse Dictionary Learning

The IID SDL problem considers a dataset $Y \in \mathbb{R}^{N\times d}$ of N IID samples with values in $\mathbb{R}^d$. The goal is to find an approximate decomposition $Y \approx RD$, where $R \in \mathbb{R}^{N\times K}$ is a sparse encoding matrix and $D \in \mathbb{R}^{K\times d}$ is a dictionary of K patterns. A frequentist starting point for IID SDL is the *linear generative model* [33] (LGM), which supposes there exist R and D as above such that

$$Y = X + \varepsilon \in \mathbb{R}^{N\times d}, \tag{1}$$

where $X = RD$ and 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 $\mathcal{L}_{1,1}$-constrained optimization problem

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

[1] When $\min\{p, q\} < 1$, $\|\cdot\|_{p,q}$ is not a norm (it is not subadditive). Nevertheless, we will say "norm" for simplicity.

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 $\widehat{R}_\lambda$; the $\mathcal{L}_{1,1}$ sparsity constraint can be equivalently expressed as a penalty of $\lambda'\|R\|_{1,1}$ (where $\lambda' \neq \lambda$) on the objective. Inspired by non-negative matrix factorization [26], R is sometimes constrained to be non-negative to promote interpretability – it is often more natural to consider a negative multiple of a feature to be a different feature altogether – but this does not significantly affect theoretical analysis 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 without changing RD, by scaling D correspondingly.

Since matrix multiplication is bilinear, the optimization problem (2) 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 (2) in R and in D. Interestingly, recent work [45, 46] has shown that, despite being non-convex, the SDL problem is often well-behaved such that standard iterative optimization algorithms can sometimes provably converge to global optima, even without multiple random restarts.

3.2 Convolutional Sparse Dictionary Learning

The CSDL problem considers a single data vector $Y \in \mathbb{R}^N$, where N is assumed to be very large. For example, Y might be a speech [42, 43] or music [51] signal over time, or functional genomic data [3] or sequence data [52, 41] over the length of the genome. Simple extensions can consider, for example, large, high-resolution images by letting $Y \in \mathbb{R}^{N_1\times N_2}$ be 2-dimensional [29]. As discussed in Section 7, CSDL can also generalize in multiple ways to multichannel data $Y \in \mathbb{R}^{N\times d}$ (e.g., to handle multiple parallel audio streams or functional genomic signals, or color images). To keep our main results simple, this paper only considers a single-dimensional ($Y \in \mathbb{R}^N$), single-channel ($d = 1$) signal, which already presents interesting questions, (whereas IID SDL with $d = 1$ degenerates to estimating a sparse sequence).

The goal here is to find an approximate decomposition $Y \approx R \otimes D$, where $R \in \mathbb{R}^{(N-n+1)\times K}$ is a sparse encoding matrix and $D \in \mathbb{R}^{n\times K}$ is a dictionary of K patterns.[2] CSDL can also be studied in a frequentist model, the *temporal linear generative model* (TLGM) [32], which supposes there exist R and D as above such that

$$Y = X + \varepsilon \in \mathbb{R}^N, \tag{3}$$

where $X = R \otimes D$, and, again, ε is random noise, though, in this setting, it may make less sense to assume that the rows of ε are independent. Under the assumption that R is sparse, a natural approach to estimating the model parameters R and D is again to solve an $\mathcal{L}_{1,1}$-constrained optimization problem, this time

$$(\widehat{R}_\lambda, \widehat{D}_\lambda) := \operatorname*{argmin}_{(R,D)\in\mathcal{S}_\lambda} \|Y - R \otimes D\|_2^2 \tag{4}$$

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 (4) is again biconvex, and can be approached by alternating minimization. As with the IID case, the constrained optimization problem (4) can equivalently be expressed as a penalized problem, specifically,

$$\left(\widehat{R}_{\lambda'}, \widehat{D}_{\lambda'}\right) = \operatorname*{argmin}_{(R,D)\in\mathcal{S}} \|Y - R \otimes D\|_2^2 + \lambda'\|R\|_{1,1} \tag{5}$$

(where, again, $\lambda \neq \lambda'$). For the remainder of this paper, we will discuss the constrained problem (4), but we show in the Appendix that equivalent results hold for the penalized problem (5).

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

1. In CSDL, we seek a decomposition $Y \approx R \otimes D$, whereas, in IID SDL, we seek a decomposition $Y \approx RD$. 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 related features.
2. In CSDL, the noise ε may have arbitrary dependencies, whereas, in IID SDL, it typically has independent rows.
3. CSDL involves an additional (potentially unknown) parameter n controlling the length of the dictionary entries,[3] whereas, in IID SDL, $n = d$ is known.

4 Related Work

There has been some work theoretically analyzing the non-convex optimization problem (2) in terms of which IID SDL is typically cast [29, 46, 45], with a consensus that despite being non-convex, this problem is often efficiently solvable in practice. Our work focuses on the statistical aspects of CSDL, and we assume, for simplicity, that the global optimum of the CSDL optimization problem (4) can be computed to high accuracy; more work is needed to link the parameters of the CSDL problem to efficient and accurate computability of this optimum.

There has been some work analyzing statistical properties of IID SDL. For some algorithms, upper bounds have been

[2] This choice of notation implies some coupling between parameters n and N (namely $n \leq N$), but we usually have $n \ll N$, and our discussion involves the length of $R \otimes D$ (the sample size) more than the length of R, so it is convenient to use N for the former.

[3] 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.

shown on the risk (in Frobenius norm, up to permutation of the dictionary elements) of estimating the true dictionary D [1, 5]. Vainsencher et al. [49] studied generalizability of dictionary learning in terms of representation error of a learned dictionary on an independent test set from the same distribution as the training set. Recently, Jung et al. [19, 20] proved the first minimax lower bounds for IID SDL, in several settings, including a general dense model, a sparse model, and a sparse Gaussian model.

The work most closely related to ours is that of Papyan et al. [35], who recently began studying the numerical properties of the CSDL. Importantly, they show that, although CSDL (4) can be expressed as a constrained version of IID SDL (2), a novel, direct analysis of CSDL, in terms of more refined problem parameters leads to stronger guarantees than does applying analysis from IID SDL. Their results are complementary to ours, for several reasons:

1. We study error of reconstructing $X = R \otimes D$, whereas they studied recovery of the dictionary D, which requires strong assumptions on D (see below).
2. They consider worst-case deterministic noise of bounded $\mathcal{L}_2$ norm, whereas we consider random noise under several different statistical assumptions. Correspondingly, we give (tighter) bounds on expected, rather than worst-case, error.
3. They study the $\mathcal{L}_{0,0}$-"norm" version of the problem, while we study the relaxed $\mathcal{L}_{1,1}$-norm version. By comparison to analysis for best subset selection and the LASSO in linear regression, we might expect solutions to the $\mathcal{L}_{0,0}$ problem to have superior statistical performance in terms of reconstruction error, and that stronger assumptions on the dictionary D may be necessary to recover D via the $\mathcal{L}_{1,1}$ approach than via the $\mathcal{L}_{0,0}$ approach. Conversely, the $\mathcal{L}_{0,0}$ problem is NP-hard, whereas the $\mathcal{L}_{1,1}$ problem can be solved via standard optimization approaches, and is hence used in practice. [4]

Notably, these previous results all require strong restrictions on the structure of the dictionary. These restrictions have been stated in several forms, from incoherence assumptions [18, 1] and restricted isometry conditions [19, 20, 35] to bounds on the Babel function [48, 49] of the dictionary, but all essentially boil down to requiring that the dictionary elements are not too correlated. As discussed in Papyan et al. [35], the assumptions needed in the convolutional case are even stronger than in the IID case: no *translations* of the dictionary elements can be too correlated. Furthermore, these conditions are not verifiable in practice. 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 reconstruction error, rather than to the error of learning the dictionary itself. As noted above, this can be compared to the fact that, in sparse linear regression, bounds on prediction error can be derived with essentially no assumptions on the covariates [10], 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, as noted earlier, we make minimal assumptions on the structure of the noise; in particular, though we require the noise to have light tails (in either a sub-Gaussian or finite-moment sense), we allow arbitrary dependence across the data sequence. This is important because, in many applications of CSDL, errors are likely to be correlated with those in nearby portions of the sequence. In the vast compressed sensing literature, there is relatively little work under these very general conditions, likely because most problems in this area, such as basis pursuit denoising or the LASSO, are clearly not consistently solvable under such general conditions. All the previously mentioned works on guarantees for dictionary learning [49, 1, 5, 19, 20] also assume no or independent noise.

5 Theoretical Results

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

$$M(N, n, \lambda, \sigma) := \inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N}\mathbb{E}\left[\left\|\widehat{X} - X\right\|_2^2\right], \qquad (6)$$

where the infimum is taken over all estimators $\widehat{X}$ of X (i.e., all functions of the observation Y). The quantity (6) 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 compression and denoising applications, rather than for learning an interpretable dictionary.

5.1 Upper Bounds for Constrained CSDL

In this section, we present our upper bounds on the reconstruction risk of the $\mathcal{L}_{1,1}$-constrained CSDL estimator (4), thereby upper bounding the minimax risk (6). We begin by noting a simple oracle bound, which serves as the starting point for our remaining upper bound results.

Lemma 1. *Let $Y = X + \varepsilon \in \mathbb{R}^N$. Then, for any $(R, D) \in \mathcal{S}_\lambda$,*

$$\begin{aligned}&\|X - \widehat{R}_\lambda \otimes \widehat{D}_\lambda\|_2^2 \\ &\leq \|X - R \otimes D\|_2^2 + 2\langle \varepsilon, \widehat{R}_\lambda \otimes \widehat{D}_\lambda - R \otimes D\rangle. \end{aligned} \qquad (7)$$

This result decomposes the error of constrained CSDL into error due to model misspecification:

$$\|X - R \otimes D\|_2^2 \qquad (8)$$

[4] Recent algorithmic advances in mixed integer optimization have rendered best subset selection computable for moderately large problem sizes [6]. It remains to be seen how effectively these methods can be leveraged for SDL.

and statistical error:

$$2\langle \varepsilon, \widehat{R}_\lambda \otimes \widehat{D}_\lambda - R \otimes D \rangle. \tag{9}$$

For simplicity, our remaining results assume the TLGM (3) holds, so that (8) is 0. We therefore focus on bounding the statistical error (9), uniformly over $(R, D) \in \mathcal{S}_\lambda$. However, the reader should keep in mind that our upper bounds hold, with an additional $\inf_{(R,D)\in\mathcal{S}_\lambda} \|X - R \otimes D\|_2^2$ term, without the TLGM. Equivalently, our results can be considered bounds on excess risk relative to the optimal sparse convolutional approximation of X. In particular, this suggests robustness of constrained CSDL to model misspecification.

Our main upper bounds apply under sub-Gaussian noise assumptions. We distinguish two notions of multivariate sub-Gaussianity, defined as follows:

Definition 2 (Componentwise Sub-Gaussianity)**.** *An $\mathbb{R}^N$-valued random variable ε is said to be componentwise sub-Gaussian with constant $\sigma > 0$ if*

$$\sup_{i\in\{1,\ldots,N\}} \mathbb{E}\left[e^{t\varepsilon_i}\right] \leq e^{t^2\sigma^2/2}, \quad \text{for all } t \in \mathbb{R}.$$

Definition 3 (Joint Sub-Gaussianity)**.** *An $\mathbb{R}^N$-valued random variable ε is said to be jointly sub-Gaussian with constant $\sigma > 0$ if*

$$\mathbb{E}\left[e^{\langle t,\varepsilon\rangle}\right] \leq e^{\|t\|_2^2\sigma^2/2}, \quad \text{for all } t \in \mathbb{R}^N.$$

Componentwise sub-Gaussianity is a much weaker condition than joint sub-Gaussianity, and, in real data, it is also often more intuitive or measurable. However, the two definitions coincide when the components of ε are independent, since the moment generating function $\mathbb{E}\left[e^{\langle t,\varepsilon\rangle}\right]$ factors, and, for this reason, joint sub-Gaussianity is commonly assumed in high-dimensional statistical problems [25, 7, 39]. As we will show, these two conditions lead to different minimax error rates. For sake of generality, we avoid making any independence assumptions, but our results for jointly sub-Gaussian noise can be thought of as equivalent to results for componentwise sub-Gaussian noise, under the additional assumption that the noise has independent components.

Consider first the case of componentwise sub-Gaussian noise. Then, constrained CSDL satisfies the following:

Theorem 4 (Upper Bound for Componentwise Sub-Gaussian Noise)**.** *Assume the TLGM holds, suppose the noise ε is componentwise sub-Gaussian with constant σ, and let the constrained CSDL tuning parameter λ satisfy $\lambda \geq \|R\|_{1,1}$. Then, the reconstruction estimate $\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda$ satisfies*

$$\frac{1}{N}\mathbb{E}\left[\|\widehat{X}_\lambda - X\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}. \tag{10}$$

Consider now the case of jointly sub-Gaussian noise. Then, an appropriately tuned constrained CSDL estimate satisfies the following tighter bound:

Theorem 5 (Upper Bound for Jointly Sub-Gaussian Noise)**.** *Assume the TLGM holds, suppose the noise ε is jointly sub-Gaussian with constant σ, and let the constrained CSDL tuning parameter λ satisfy $\lambda \geq \|R\|_{1,1}$. Then, the reconstruction estimate $\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda$ satisfies*

$$\frac{1}{N}\mathbb{E}\left[\|\widehat{X}_\lambda - X\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2\log(2(N-n+1))}}{N}. \tag{11}$$

The main difference between the bounds (10) and (11) is the presence of a $\sqrt{n}$ factor in the former. In Section 5.2, we will show that both upper bounds are essentially tight, and the minimax rates under these two noise conditions are indeed separated by a factor of $\sqrt{n}$. We also empirically verify this phenomenon in Section 6, and, in Section 7, we provide some intuition for why this occurs. Also note that, in the Appendix, we provide related results bounding the error of the penalized form (5) of CSDL and also bounding error under weaker finite-moment noise conditions.

5.2 Lower Bounds

We now present minimax lower bounds showing that the upper bound rates in Theorems 4 and 5 are essentially tight in terms of the sparsity λ, noise level σ, sequence length N, and dictionary length n.

First consider the componentwise sub-Gaussian case, analogous to that considered in Theorem 4:

Theorem 6 (Lower Bound for Componentwise sub-Gaussian Noise)**.** *Assume the TLGM holds. Then, there exists a (Gaussian) noise pattern ε that is componentwise sub-Gaussian with constant σ such that the following lower bound on the minimax average $\mathcal{L}_2$ reconstruction risk holds:*

$$M(\lambda, N, n, \sigma) \geq \frac{\lambda}{8N}\min\left\{\lambda, \sigma\sqrt{n\log(N-n+1)}\right\} \tag{12}$$

The min here reflects the fact that, in the extremely sparse or noisy regime $\lambda \leq \sigma\sqrt{n\log(N-n+1)}$, the trivial estimator $\widehat{\theta} = 0$ becomes optimal, with average $\mathcal{L}_2$-risk at most λ^2/N. Except in this extremely sparse/noisy case, this minimax bound essentially matches the upper bound provided by Theorem 4.

Rather than directly utilizing information theoretic bounds such as Fano's inequality, our lower bounds are based on reducing the classical $\mathcal{L}_1$-constrained Gaussian sequence estimation problem (see, e.g., Section 2.3 of Rigollet [39]) to CSDL (i.e., showing that an estimator for the prescribed CSDL problem can be used to construct a estimator for the

mean of an $\mathcal{L}_1$-constrained Gaussian sequence such that the error of the Gaussian sequence estimator is bounded in terms of that of the CSDL estimator). Standard lower bounds for the $\mathcal{L}_1$-constrained Gaussian sequence problem (e.g., Corollary 5.16 of Rigollet [39]) then directly imply a lower bound for the CSDL estimator. Again, detailed proofs are provided in the Appendix.

Now consider the jointly sub-Gaussian case, analogous to that considered in Theorem 5:

Theorem 7. *(Lower Bound for Jointly sub-Gaussian Noise): Assume the TLGM holds, and suppose that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, so that ε is jointly sub-Gaussian with constant σ. Then, the following lower bound on the minimax average $\mathcal{L}_2$ reconstruction risk holds:*

$$M(\lambda, N, n, \sigma) \geq \frac{\lambda}{8N} \min\left\{\lambda, \sigma\sqrt{\log(N-n+1)}\right\} \tag{13}$$

Again, the min here reflects the fact that, in the extremely sparse or noisy regime $\lambda \leq \sigma\sqrt{\log(N-n+1)}$, the trivial estimator $\widehat{\theta} = 0$ becomes optimal. Except in this extreme case, this minimax bound essentially matches the upper bound provided by Theorem 5.

5.3 Comparison to IID SDL

As mentioned previously, the IID SDL algorithm (2) has historically been applied to problems with spatial or temporal structure better modeled by a TLGM (3) than by an LGM (1), by means of partitioning the data into patches. In this section, we consider how the statistical performance of IID SDL compares with that of CSDL, under the TLGM. For simplicity, we consider the case of componentwise sub-Gaussian noise; conclusions under joint sub-Gaussianity are similar. For clarity, in this section, notation used according to the LGM/IID SDL setting will be denoted with the prime mark $'$, while other quantities should be interpreted as in the TLGM/CSDL setting.[5]

The next result is an analogue of Theorem 4 for IID SDL under the LGM; the proof is similar, with Young's inequality for convolutions replaced by a simple linear bound.

Theorem 8 (Upper Bound for IID SDL)**.** *Assume the LGM holds, suppose the noise ε is componentwise sub-Gaussian with constant σ (i.e., for each dimension $j \in [d']$, $\varepsilon_j \in \mathbb{R}^{N'}$ is componentwise sub-Gaussian with constant σ), and let the constrained IID SDL parameter λ' satisfy $\lambda' \geq \|R'\|_{1,1}$. Then, the reconstruction estimate $\widehat{X}'_{\lambda'} = \widehat{R}'_{\lambda'} \widehat{D}'_{\lambda'}$ satisfies*

$$\frac{1}{N'd'} \|X - \widehat{X}'_{\lambda'}\|_{2,2}^2 \leq \frac{4\lambda'\sigma\sqrt{2d'\log(2N'd)}}{N'd'} \tag{14}$$

[5]In this section, λ' should not be confused with the tuning parameter of penalized CSDL, also denoted λ'.

The natural conversion of parameters from the TLGM to the LGM sets $d' = n$ and $N' = N/n$. At first glance, plugging these into (14) appears to give the same rate as (10). The catch is the condition that $\lambda' \geq \|R'\|_{1,1}$. When converting data from the TLGM to the format of the LGM, the sparsity level can grow by a factor of up to n; that is, in the worst case, we can have $\|R'\|_{1,1} = n\|R\|_{1,1}$. Therefore, the bound can increase by a factor of n, relative to the bound for CSDL. From a statistical perspective, this may explain the superior performance of CSDL over IID SDL, when data is better modeled by the TLGM than by the LGM, especially when the patterns being modeled in the data are relatively large.

6 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. MATLAB code for reproducing these experiments and figures is available at `https://github.com/sss1/convolutional-dictionary`. Since the focus of this paper is on the statistical, rather than algorithmic, properties of CSDL, we assume the estimator $(\widehat{R}_\lambda, \widehat{D}_\lambda)$ defined by the optimization problem (4) can be computed to high precision using a simple alternating projected gradient descent algorithm, the details of which are provided in the appendix. We then use the reconstruction estimate $\widehat{X}_\lambda := \widehat{R}_\lambda \otimes \widehat{D}_\lambda$ to estimate $X = R \otimes D$. All results presented are averaged over 1000 IID trials, between which R and S were regenerated randomly.[6] 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 $R = 0 \in \mathbb{R}^{(N-n+1)\times K}$ and then repeatedly adding 1 to uniformly random coordinates of R until the desired value of $\|R\|_{1,1}$ is achieved. Further details of the experiment setup and implementation are given in the Appendix.

Comparisons: We compare the error of the optimal CSDL estimator $\widehat{X}_\lambda$ to the theoretical upper bounds (Inequalities (10) and (11)) and lower bounds (Inequalities (13) and (12)), as well as the trivial estimators $\widehat{X}_0 = 0$ and $\widehat{X}_\infty = Y$.

Experiment 1. Our first experiment studies the relationship between the length N of the sequence and the true $\mathcal{L}_{1,1}$-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 linear $\|R\|_{1,1} = \lfloor N/10 \rfloor$ functions of N. The results are consistent with the main predictions of Theorems 5 and 7 for jointly sub-Gaussian noise, namely that the error of the CSDL estimator using the optimal tuning parameter $\lambda = \|R\|_{1,1}$ lies between the lower and upper

[6]We initially plotted 95% confidence intervals for each point based on asymptotic normality of the empirical $\mathcal{L}_2$ error. Since intervals were typically smaller than markers sizes, we removed them to avoid clutter.

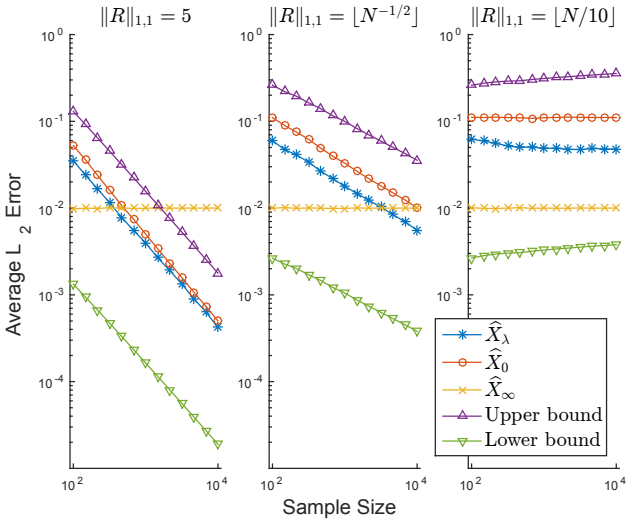


Figure 1: Experiment 1: Average $\mathcal{L}_2$-error as a function of sequence length N, with sparsity scaling as $\|R\|_{1,1} = 5$ (first panel), $\|R\|_{1,1} = \left\lfloor \sqrt{N} \right\rfloor$ (second panel), and $\|R\|_{1,1} = \lfloor N/10 \rfloor$ (third panel).

bounds, and converges at a rate of order $\|R\|_{1,1}/N$, up to log factors). As a result, the estimator is inconsistent when $\|R\|_{1,1}$ grows linearly with N, in which case there is no benefit to applying CSDL to denoise the sequence over using the original sequence $\widehat{X}_\infty = X$, even though the latter is never consistent.[7] On the other hand, if $\|R\|_{1,1}$ scales sublinearly with N, CSDL is consistent and outperforms both trivial estimators (although, of course, the trivial estimator $\widehat{X}_0 = 0$ is also consistent in this setting). In the appendix, we present an analogous experiment in the heavy-tailed case, where the noise ε has only 2 finite moments.

Experiment 2. Our second experiment studies the dependence of the error on the length n of the dictionary elements, and how this varies with the dependence structure of the noise ε. Specifically, we considered two ways of generating the noise ε: (1) IID Gaussian entries (as in other experiments), and (2) perfectly correlated Gaussian entries (i.e., a single Gaussian sample was drawn and added to every entry of $R \otimes D$. In the former case, Theorems 5 and 7 suggest a convergence rate independent of n, whereas, in the latter case, Theorems 4 and 6 suggest a rate scaling as quickly as $\sqrt{n}$. To allow a larger range of values for n, we fixed a larger value of $N = 5000$ for this experiment. Figure 2 shows error as a function of n for logarithmically spaced values between $10^{1/2}$ and 10^3, in the cases of independent noise and perfectly correlated noise. As predicted, the error of the CSDL estimator using the optimal tuning parameter $\lambda = \|R\|_{1,1}$ lies between the lower and upper bounds, and exhibits worse scaling in the case of dependent noise.

[7]Even if $\|R\|_{1,1}$ grows linearly with N, as long as $\|R\|_{1,1}/N$ is small, CSDL may still be useful for compression, if a constant-factor loss is acceptable.

7 Discussion

Theorems 4,5,6, and 7 together have several interesting consequences. Firstly, in the fairly broad setting of componentwise sub-Gaussian noise, for fixed K, the minimax average $\mathcal{L}_2$ risk for CSDL reconstruction in the case of sub-Gaussian noise is, up to constant factors, of order

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

Under a stronger assumption of joint sub-Gaussianity (e.g., independent noise), the minimax risk becomes

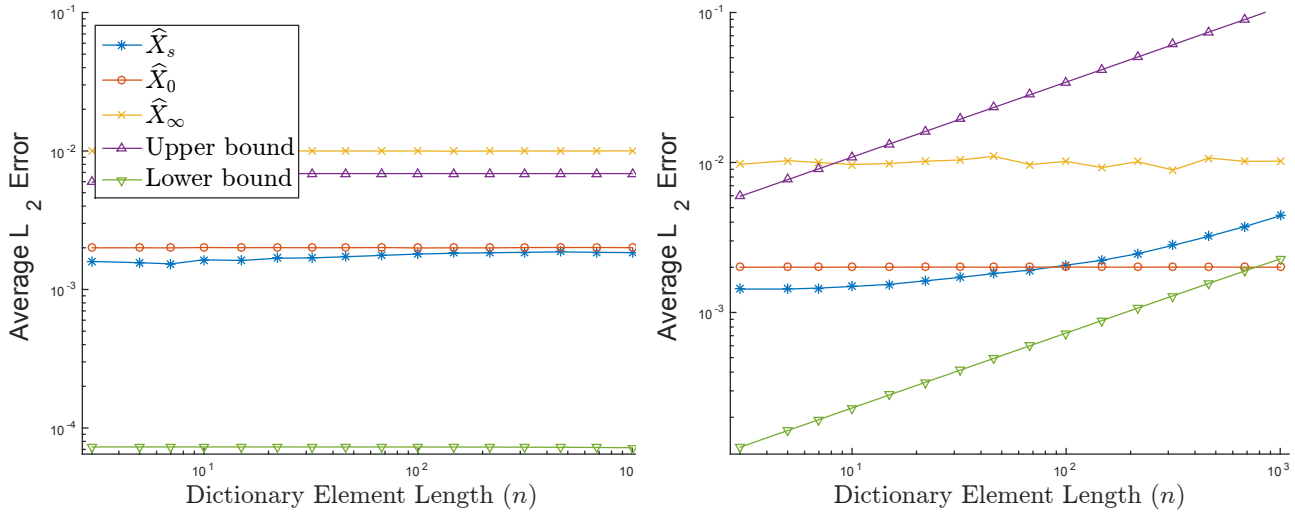
$$M(\lambda, N, n, \sigma) \asymp \frac{\lambda}{N} \min\left\{\lambda, \sigma\sqrt{\log N}\right\}.$$

In retrospect, the presence of an additional $\sqrt{n}$ factor in error under componentwise sub-Gaussianity is quite intuitive. Dictionary learning consists, approximately, of performing compressed sensing in an (unknown) basis of greatest sparsity. Componentwise sub-Gaussianity bounds the noise level in the original (data) basis, whereas joint sub-Gaussianity bounds the noise level in *all* bases. Hence, the componentwise noise level σ may be amplified (by factor of up to $\sqrt{n}$) when transforming to the basis in which the data are sparsest. Perhaps surprisingly, the fact that the dictionary D is unknown does not affect these rates; our lower bounds are both derived for fixed choices of D. A similar phenomenon has been observed in IID SDL, where Arora et al. [5] showed upper bounds that are of the same rate as lower bounds for the case where the dictionary is known beforehand (the classic problem of *sparse recovery*).

For $n \ll N$, an important consequence is that CSDL (and, similarly, IID SDL when $d \ll N$) is robust to arbitrary noise dependencies. This aspect of SDL is relatively unique amongst compressed sensing problems, and has important practical implications, explaining why these algorithms perform well for image, speech, or other highly dependent data.

On the negative side, our lower bounds imply that a high degree of sparsity is required to guarantee consistency. Specifically, under sub-Gaussian noise, for fixed n and σ, guaranteeing consistency requires $\lambda \in o\left(\frac{N}{\log N}\right)$. Our additional results in the Appendix suggest that an even higher degree (polynomially in n) of sparsity is required when the noise has tails that are heavier than Gaussian.

Our results focused on error measured in $\mathcal{L}_2$-loss, but they also have some consequences for rates under different losses. In particular, for $p \geq 2$, since the $\mathcal{L}_2$-norm on $\mathbb{R}^N$ dominates the $\mathcal{L}_p$-norm, our upper bounds extend to error measured in $\mathcal{L}_p$-loss. Similarly, for $p \in [1, 2]$, since the $\mathcal{L}_p$-norm dominates the $\mathcal{L}_2$-norm, our lower bounds extend to error measured in $\mathcal{L}_p$-loss. On the other hand, further



(a) (b)

Figure 2: Experiment 2: Average $\mathcal{L}_2$-error as a function of dictionary element length n, when 2a noise is independent across the input and 2b noise is perfectly correlated across the input.

work is needed to obtain tight lower bounds for $p > 1$ and to obtain tight upper bounds for $p \in [1, 2)$.

Our results rely on choosing the tuning parameter λ well ($\lambda \approx \|R\|_{1,1}$). In the Appendix, we show that similar rates hold for the penalized form (5) of CSDL, but that the tuning parameter λ' should be chosen proportional to the noise level σ (independent of $\|R\|_{1,1}$). Depending on the application and domain knowledge, it may be easier to estimate σ or $\|R\|_{1,1}$, which may influence the choice of whether to use the constrained or penalized estimator (in conjunction with possible computational advantages of either form).

Finally, we reiterate that, while precise quantitative bounds are more difficult to study, our results extend beyond the TLGM; in general, the $\mathcal{L}_1$-constrained estimator converges to the optimal λ-sparse convolutional representation of X.

7.1 Future Work

There remain several natural open questions about the statistical properties of CSDL. First, how do rates extend to the case of multi-channel data $X \in \mathbb{R}^{N \times d}$? There are multiple possible extensions of CSDL to this case; 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 3-tensors and learn separate dictionary and encoding 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. Though a somewhat more restrictive model, this latter approach would have the advantage that statistical risk would *decrease* with d, as data from multiple dimensions could contribute to the difficult problem of estimating R.

Another 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 words 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, while this work contributes to understanding the statistical properties of convolutional models, more work is needed to relate these results on sparse convolutional dictionaries to the hierarchical convolutional models that underlie state-of-the-art neural network methods for a variety of natural data, ranging from images to language and genomics[23, 24, 52, 41]. In this direction, Papyan et al. [34] very recently began making progress by extending the theoretical results of Papyan et al. [35] to multilayer convolutional neural networks.

Acknowledgements

This material is based upon work supported by the NSF Graduate Research Fellowship DGE-1252522 (to S.S.). The work was also partly supported by NSF IIS-1563887, Darpa D3M program, and AFRL (to B.P.), and NSF IIS-1717205 and NIH HG007352 (to J.M.). 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 NSF and the NIH.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014.

[2] M. Aharon, M. Elad, and A. Bruckstein. k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[3] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

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

[5] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014.

[6] D. Bertsimas, A. King, R. Mazumder, et al. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

[9] O. Bryt and M. Elad. Compression of facial images using the k-svd algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282, 2008.

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

[11] I. Y. Chun and J. A. Fessler. Convolutional dictionary learning: Acceleration and convergence. *arXiv preprint arXiv:1707.00389*, 2017.

[12] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12): 3736–3745, 2006.

[13] 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*, volume 5, pages 2443–2446. IEEE, 1999.

[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*, pages 1823–1831, 2015.

[15] 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*, pages 5135–5143, 2015.

[16] F. Huang and A. Anandkumar. Convolutional dictionary learning through tensor factorization. In *Feature Extraction: Modern Questions and Challenges*, pages 116–129, 2015.

[17] M. G. Jafari and M. D. Plumbley. Fast dictionary learning for sparse representations of speech signals. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1025–1031, 2011.

[18] 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.

[19] 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*, pages 765–769. IEEE, 2014.

[20] A. Jung, Y. C. Eldar, and N. Görtz. On the minimax risk of dictionary learning. *IEEE Transactions on Information Theory*, 62(3):1501–1515, 2016.

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

[22] 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*, 15(2):349–396, 2003.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[25] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

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

[27] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2): 337–365, 2000.

[28] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.

[29] 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*, pages 689–696. ACM, 2009.

[30] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 343–351, 2013.

[31] N. Mehta and A. Gray. Sparsity-based generalization bounds for predictive sparse coding. In *International Conference on Machine Learning*, pages 36–44, 2013.

[32] B. A. Olshausen. Sparse codes and spikes. *Probabilistic Models of the Brain*, page 257, 2002.

[33] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[34] V. Papyan, Y. Romano, and M. Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

[35] V. Papyan, J. Sulam, and M. Elad. Working locally thinking globally: Theoretical guarantees for convolutional sparse coding. *IEEE Transactions on Signal Processing*, 2017.

[36] G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

[37] G. Peyré, J. M. Fadili, and J.-L. Starck. Learning adapted dictionaries for geometry and texture separation. In *SPIE Wavelets XII*, volume 6701, page 67011T. SPIE, 2007.

[38] M. Protter and M. Elad. Super resolution with probabilistic motion estimation. *IEEE Transactions on Image Processing*, 18(8):1899–1904, 2009.

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

[40] 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*, pages 431–436. IEEE, 2007.

[41] S. Singh, Y. Yang, B. Poczos, and J. Ma. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *bioRxiv*, page 085241, 2016.

[42] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, 2007.

[43] W. Smit and E. Barnard. Continuous speech recognition with sparse coding. *Computer Speech & Language*, 23(2):200–219, 2009.

[44] E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

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

[46] J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[47] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias. Multiple kernel sparse representations for supervised and unsupervised learning. *IEEE transactions on Image Processing*, 23(7):2905–2915, 2014.

[48] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

[49] D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(Nov):3259–3281, 2011.

[50] B. Wohlberg. Efficient convolutional sparse coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7173–7177. IEEE, 2014.

[51] C.-C. M. Yeh and Y.-H. Yang. Supervised dictionary learning for music genre classification. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 55. ACM, 2012.

[52] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931, 2015.

[53] Y. Zhu and S. Lucey. Convolutional sparse coding for trajectory reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):529–540, 2015.