A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli*

Abstract

We consider the problem of learning overcomplete dictionaries in the context of sparse coding, where each sample selects a sparse subset of dictionary elements. Our main result is a strategy to approximately recover the unknown dictionary using an efficient algorithm. Our algorithm is a clustering-style procedure, where each cluster is used to estimate a dictionary element. The resulting solution can often be further cleaned up to obtain a high accuracy estimate, and we provide one simple scenario where ℓ_1-regularized regression can be used for such a second stage.

Keywords: Dictionary learning, sparse coding, overcomplete dictionaries, incoherence, lasso.

1 Introduction

The dictionary learning problem is as follows: given observations Y, the task is to factorize it as

$$Y = AX, \quad Y \in \mathbb{R}^{d\times n},\ A \in \mathbb{R}^{d\times r},\ X \in \mathbb{R}^{r\times n}, \tag{1}$$

where X is referred to as the *coefficient* matrix and the columns of A are referred to as the *dictionary* elements. There are indeed infinite factorizations for (1) unless further constraints are imposed. A natural assumption is that the coefficient matrix X is sparse, and in fact, that each sample y_i selects a sparse subset of dictionary elements from A. This instance of dictionary learning is popularly known as the *sparse coding* problem [30, 25]. It has been argued that sparse coding can provide a succinct representation of the observed data, given only unlabeled samples [25]. Through this lens of unsupervised learning, dictionary learning has received an increased attention from the machine learning community in the last few years; see Section 1.2 for a brief survey.

Although the above problem has been extensively studied, most of the methods are heuristic and lack guarantees. Spielman et. al [31] provide exact recovery results for this problem, when the coefficient matrix has Bernoulli-Gaussian entries and the dictionary matrix $A \in \mathbb{R}^{r\times d}$ has full column rank. This condition entails that the dictionary is *undercomplete*, i.e., the observed dimensionality needs to be greater than the number of dictionary elements ($r \leq d$). However, for

*A. Agarwal is with Microsoft Research, New York. Email: alekha@microsoft.com. A. Anandkumar is with the Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science Dept., University of California, Irvine, USA 92697. Email: a.anandkumar@uci.edu. P. Netrapalli is with Dept. of ECE, The University of Texas at Austin. Email: praneethn@utexas.edu. The work was initiated during the visits of A. Anandkumar and P. Netrapalli to Microsoft Research New England in Summer 2013. A preliminary version of this manuscript containing a subset of these results appears in the proceedings of COLT 2014.

most practical settings, it has been argued that *overcomplete* representations, where $r \gg d$, are far more relevant, and can provide greater flexibility in modeling as well as greater robustness to noise [26, 12]. Moreover, in the context of blind source separation (BSS) of audio, image or video signals, the dictionary learning problem is typically overcomplete, since there are more sources than observations [15]. In this work, we provide guaranteed methods for learning overcomplete dictionaries.

1.1 Summary of Results

In this paper we present a novel algorithm for the estimation of overcomplete dictionaries. The algorithm can be seen as a *clustering style method* followed by a singular value decomposition (SVD) within each cluster resulting in an estimate for each dictionary element. The clusters are formed based on the magnitudes of the correlation between pairs of samples. Under our probabilistic model of generating data as well as assumptions on the coefficients and dictionaries, it can be guaranteed that such a procedure approximately recovers the unknown overcomplete dictionary. Under further conditions, it is often possible to start with this approximate solution and perform additional post-processing on it to obtain arbitrarily good estimates of the dictionary. We present one such set of conditions under which *sparse regression* can be used for this post-processing. More advanced post-processing methods have been developed in subsequent works [1, 8].

We consider a random coefficient matrix, where each column of X has s non-zero entries which are randomly chosen, i.e., each sample y_i selects s dictionary elements uniformly at random. We additionally assume that the dictionary elements are pairwise *incoherent* and that the dictionary matrix satisfies a certain bound on the spectral norm. Under these conditions, we establish that our algorithm estimates the dictionary elements with bounded (constant) error when the number of samples scales as $n = \mathcal{O}(r(\log r + \log d))$, and when the sparsity $s = \mathcal{O}(d^{1/4}, r^{1/4})$. To the best of our knowledge, this is the first result of its kind which analyzes the global recovery properties of a computationally efficient procedure in the setup of overcomplete dictionary learning.

In the special case when the coefficients are $\{-1, 0, 1\}$-valued with zero mean, the resulting solution from the first step can be further plugged into any sparse regression algorithm for estimating the coefficients given this dictionary estimate. Under a more stringent sparsity constraint: $s = \mathcal{O}(d^{1/5}, r^{1/6})$, it can be shown that this second step will recover the coefficients *exactly* even from this approximate dictionary, which then also leads to an exact recovery of the dictionary by solving the linear system. Hence, we provide a simple method for exactly recovering the unknown dictionary in this special case. A natural generalization of this procedure to general weights is analyzed using alternating minimization procedure in a subsequent work [1].

We outline our method as well as our analysis techniques in Section 1.3. This is the first work to provide a tractable method for *guaranteed recovery* of overcomplete dictionaries, and we discuss the previous results below. Finally, concurrently with our work, an approximate recovery result with a similar procedure was recently announced by Arora et al. [8]. A detailed discussion comparing our and their results is presented in Section 1.2.

1.2 Related Works

This work overlaps with and relates to prior works in many different communities and we discuss them below in turn.

Dictionary Learning: Hillar and Sommer [20] consider conditions for identifiability of sparse coding and establish that when the dictionary succeeds in reconstructing a certain set of sparse vectors, there exists a unique sparse coding, up to permutation and scaling. However, the number of samples required to establish identifiability is exponential in r for the general case. In contrast, we show that efficient recovery is possible using $\mathcal{O}(r(\log r + \log d))$ samples, albeit under additional conditions such as *incoherence* among the dictionary elements.

Spielman et. al [31] provide exact recovery results for a ℓ_1 based method in the *undercomplete* setting, where $r \leq d$. In contrast, we allow for the overcomplete setting where $r > d$. There exist a plethora of heuristics for dictionary learning, which work well in practice in many contexts, but lack theoretical guarantees. For instance, Lee et. al. propose an iterative ℓ_1 and ℓ_2 optimization procedures [25]. This is similar to the the method of optimal directions (MOD) proposed in [16]. Another popular method is the so-called K-SVD, which iterates between estimation of X and given an estimate of X, updates the dictionary estimate using a spectral procedure on the residual. Other works consider more sophisticated methods from an optimization viewpoint while still alternating between dictionary and coefficient updates [24, 18]. Geng et al. [18] and Jenatton et al. [23] study the local optimality properties of an alternating minimization procedure. In contrast, our work focuses on global properties of a more combinatorial procedure than several of the above works which are more optimization flavored. The upshot is that our procedure, while still being computationally quite efficient, is able to guarantee global bounds on the quality of the solution obtained.

Recent works [34, 29, 27, 32] provide generalization bounds and algorithmic stability for predictive sparse coding, where the goal of the learned sparse representation is to obtain good performance on some predictive task. This differs from our framework since we do not consider predictive tasks here, but the accuracy in recovering the underlying dictionary elements.

Finally, our results are closely related to the very recent work of Arora et al. [8], carried out independently and concurrently with our work. There are however some important distinctions: we require only $\mathcal{O}(r)$ samples in our analysis, while Arora et al. [8] require $\mathcal{O}(r^2)$ samples in their result. At the same time, their analysis yields milder conditions on the sparsity level s in terms of its dependence on r and d. Following this work, Arora et al. [8] and Agarwal et al. [1] also developed a post-processing techniques which can be thought of as a more advanced variant of the simpler sparse-regression step that we analyze. These subsequent works view the methods developed here as initialization procedures to alternating optimization schemes.

Blind Source Separation/ICA/Topic Models: The problem of dictionary learning is applicable to blind source separation (BSS), where the rows of X are signals from the sources and A represents the linear mixing matrix. The term *blind* implies that the dictionary matrix A is unknown and needs to be jointly estimated with the coefficient matrix X, given samples Y. This problem has been extensively studied and the most popular setting is the independent component analysis (ICA), where the sources are assumed to be independent. In contrast, for the sparse component analysis problem, no assumptions are made on the statistics of the sources. Many works provide guarantees for ICA in the undercomplete setting, where there are fewer sources than observations [21, 9, 4] and some works provide guarantees in the overcomplete setting [14, 19]. However, the techniques are very different since they rely on the independence among the sources. The problem of learning topic models can be cast as a similar factorization problem, where A now corresponds to the topic-word matrix and X corresponds to the proportions of topics in various documents. There are various recent works providing guaranteed methods for learning topic mod-

els, e.g [2, 7, 6, 5]. However, these works make different assumptions on either A or X or both to guarantee recovery. For instance, the work [7] assumes that the topic-word matrix A has rows such that for each column, only the entry corresponding to that column is non-zero. The work [6] assumes expansion conditions on A and provides recovery through ℓ_1-based optimization. We note that the techniques of [6] are related to those employed by Spielman et. al [31] for dictionary learning, but make different assumptions. All these works only deal with the undercomplete setting. The recent work [5] considers topic models in the overcomplete setting, and provides guarantees when A satisfies certain higher order expansion conditions. The techniques are very different from the ones employed here since they involve higher order moments and tensor forms.

Connection to Learning Overlapping Communities: Our initial step for estimating the dictionary elements involves finding large cliques in the sample correlation graph, where the nodes are the samples and the edges represent sufficiently large correlations among the endpoints. The clique finding problem is a special instance of the overlapping community detection problem, which has been studied in various contexts, e.g. [3, 11, 10, 22, 28]. However, the correlation graph here has different kinds of constraints than the ones studied before as follows. In our setting involving noise-free dictionary learning, each community corresponds to a clique and there are no edges across two different communities. In contrast, many works on community detection are concerned about handling *noise* efficiently, where each community is not a full clique, and there are edges across different communities. Here, we need to learn overlapping communities, while many community detection methods limit to learning non-overlapping ones. In our setting, we argue that the overlap across different communities is small under a random coefficient matrix, and thus, we can find the communities efficiently through simple random sampling and neighborhood testing procedures.

1.3 Overview of Techniques

As stated earlier, our main algorithm consists of a clustering procedure which yields an approximate estimate of the dictionary. This estimate can be subsequently post-processed for exact recovery of the dictionary under certain further conditions. Below we give the outline and the main intuition underlying these procedures and their analysis.

Dictionary estimation via clustering: This step first involves construction of the sample correlation graph $G_{\mathrm{corr}(\rho)}$, where the nodes are samples $\{y_1, y_2, \ldots y_n\}$ and an edge $(y_i, y_j) \in G_{\mathrm{corr}(\rho)}$ implies that $|\langle y_i, y_j \rangle| > \rho$, for some $\rho > 0$. We then employ a *clustering* procedure on the graph to obtain a subset of samples, which are then employed to estimate each dictionary element. Roughly, we search for large cliques in the correlation graph and obtain a spectral estimate of each dictionary element using samples from such sets.

Key intuitions for the clustering procedure: The core intuitions can be described in terms of the relationships between the two graphs, viz., the coefficient bipartite graph B_{coeff} and the sample correlation graph G_{corr}, shown in Figures 1a and 1b. As described earlier, the correlation graph G_{corr} consists of edges between well correlated samples. The coefficient bipartite graph B_{coeff} consists of dictionary elements $\{a_i\}$ on one side and the samples $\{y_i\}$ on the other, and the bipartite graph B_{coeff} encodes the sparsity pattern of the coefficient matrix X. In other words, it maps the dictionary elements $\{a_i\}$ to samples $\{y_i\}$ on which they are supported on and $\mathcal{N}_B(y_i)$ denotes the neighborhood of y_i in the graph B_{coeff}.

Now given this bipartite graph B_{coeff}, for each dictionary element a_i, consider a set of samples[1] which (pairwise) have only one dictionary element a_i in common, and denote such a set by $\mathcal{C}_i$ i.e.

$$\mathcal{C}_i := \{y_k, k \in S : \mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l) = a_i, \ \forall\, k, l \in S\}. \tag{2}$$

For a random coefficient matrix (resulting in a random bipartite graph), we argue that there exists (large) sets $\mathcal{C}_i$, for each $i \in [r]$, which consists of a large fraction of $\mathcal{N}_B(a_i)$, and no two elements a_i and a_j have a large fraction of samples in common. In other words, for random coefficient matrices, we see a diversity in the dictionary elements among the samples, and this can be viewed as an *expansion* property from the dictionary elements to the set of samples. We exploit this property to establish success for our method.

Our subsequent analysis is broadly divided into two parts, viz., establishing that (large) sets $\{\mathcal{C}_i\}$ can be found efficiently, and that the dictionary elements can be estimated accurately once such sets $\{\mathcal{C}_i\}$ are found. We establish that the sets $\{\mathcal{C}_i\}$ are cliques in the correlation graph when the dictionary elements are incoherent, as shown in Figure 1b. Combined with the previous argument that the different sets $\mathcal{C}_i$'s have only a small amount of overlap for random coefficient matrices, we argue that these sets can be found efficiently through simple random sampling and neighborhood testing on the correlation graph. Once a large enough set $\mathcal{C}_i$ is found, we argue that under incoherence, the dictionary element a_i can be estimated accurately through SVD over the samples in $\mathcal{C}_i$.

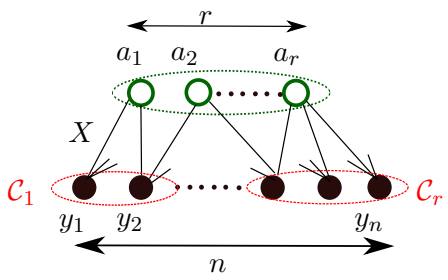
Sparse regression for post-processing: This is a relatively straightforward procedure. Once an initial estimate of the dictionary matrix is obtained, we estimate the coefficient matrix X through any sparse regression procedure (such as Lasso) and then perform thresholding on the recovered coefficients. Now, we re-estimate the dictionary, given this coefficient matrix, by solving another linear system. This provides us with a final estimate of both the dictionary as well as the coefficient matrix.

Since we only have a noisy estimate of the dictionary, our analysis here is slightly different from the usual analysis for a sparse linear system. The noise in our system is dependent on the approximate dictionary employed, which differs from the typical statistical setting, where noise is assumed to be independent. We exploit the known guarantees available for Lasso under deterministic noise [13] for our setting. Combining Lasso with a simple thresholding procedure, we guarantee exact recovery of the coefficient matrix, albeit under a more stringent condition on the sparsity and the coefficient values (namely zero mean and $\{-1, 0, 1\}$-valued). The dictionary is then re-estimated by solving another linear system, which is of course correct owing to the exact estimation of the coefficient matrix.

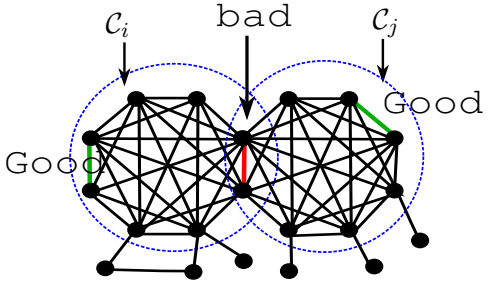
2 Method and Guarantees

Notation: Let $[n] := \{1, 2, \ldots, n\}$ and for a vector w, let $\text{Supp}(w)$ denote the support of w, i.e. the set of indices where w is non-zero. Let $\|w\|$ denote the ℓ_2 norm of vector w, and similarly for a matrix W, $\|W\|$ denotes its spectral norm. Let $A = [a_1|a_2|\ldots|a_r]$, where a_i denotes the i^{th} column, and similarly for $Y = [y_1|y_2|\ldots y_n]$ and $X = [x_1|\ldots|x_n]$. For a graph $G = (V, E)$, let $\mathcal{N}_G(i)$ denote set of neighbors for node i in G.

[1] Note that such a set need not be unique.



(a) Coefficient bipartite graph B mapping dictionary elements $a_1, a_2, \dots a_r$ to samples $y_1, \dots y_n$: $y_i = \sum_{j \in [r]} x_{ji} a_j$. See (2) for definition of $\mathcal{C}_i$.



(b) Sample correlation graph G_{corr} with nodes $\{y_k\}$ and edge (y_i, y_j) s.t. $|\langle y_i, y_j \rangle| > \rho$. $\mathcal{C}_i$, defined in (2), is a clique in the correlation graph. See Lemma 3.1.

Figure 1: Coefficient bipartite graph and the sample correlation graph.

2.1 Clustering procedure and its analysis

We start with presenting the main algorithm of our work and bound the recovery error under certain assumptions.

2.1.1 Algorithm

Our main algorithm is presented in Algorithm 1. Given samples Y, we first construct the correlation graph $G_{\text{corr}(\rho)}$, where the nodes are samples $\{y_1, y_2, \dots y_n\}$ and an edge $(y_i, y_j) \in G_{\text{corr}(\rho)}$ implies that $|\langle y_i, y_j \rangle| > \rho$, for some threshold $\rho > 0$. We then determine a good subset of samples via a *clustering* procedure on the graph as follows: we first randomly sample an edge $(y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}$ and then consider the intersection of their neighborhoods, denoted by $\widehat{S}$. We then employ UniqueIntersection routine in Procedure 1 to determine if $\widehat{S}$ is a "good set" for estimating a dictionary element, and this is done by ensuring that the set $\widehat{S}$ has sufficient number of mutual neighbors[2] in the correlation graph. Once $\widehat{S}$ is determined to be a good set, we then proceed by estimating the matrix $\widehat{M}$ using samples in $\widehat{S}$ and output its top singular vector as the estimate of a dictionary element. The method is repeated over all edges in the correlation graph to ensure that all the dictionary elements get estimated with high probability.

2.1.2 Assumptions and Main Result

Assumptions: We now provide guarantees for the proposed method under the following assumptions on A and X.

($A1$) **Unit-norm Dictionary Elements:** All the elements are normalized: $\|a_i\| = 1$, for $i \in [r]$.

($A2$) **Incoherent Dictionary Elements:** We assume pairwise incoherence condition on the dictionary elements, for some constant $\mu_0 > 0$,

$$|\langle a_i, a_j \rangle| < \frac{\mu_0}{\sqrt{d}}. \tag{3}$$

[2]For convenience to avoid dependency issues, in Procedure 1, we partition $\widehat{S}$ into sets consisting of node pairs and determine if there are sufficient number of node pairs which are neighbors.

Algorithm 1 DictionaryLearn$(Y, \epsilon_{\text{dict}}, \rho)$: Clustering approach for estimating dictionary elements.

Input: Samples $Y = [y_1|\dots|y_n]$. Correlation threshold ρ. Desired separation parameter ϵ between recovered dictionary elements.
Output: Initial Dictionary Estimate $\bar{A}$.
Construct correlation graph $G_{\text{corr}(\rho)}$ s.t. $(y_i, y_j) \in G_{\text{corr}(\rho)}$ when $|\langle y_i, y_j \rangle| > \rho$.
Set $\bar{A} \leftarrow \emptyset$.
for each edge $(y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}$ **do**
  $\widehat{S} \leftarrow \mathcal{N}_{G_{\text{corr}(\rho)}}(y_{i^*}) \cap \mathcal{N}_{G_{\text{corr}(\rho)}}(y_{j^*})$.
  if UniqueIntersection$(\widehat{S}, G_{\text{corr}(\rho)})$ **then**
    $\widehat{L} \leftarrow \sum_{y \in \widehat{S}} yy^\top$ and $\bar{a} \leftarrow u_1$, where u_1 is top singular vector of $\widehat{L}$.
    if $\min_{b \in \bar{A}} \|\bar{a} - b\| > 2\epsilon_{\text{dict}}$ **then**
      $\bar{A} \leftarrow \bar{A} \cup \bar{a}$
    end if
  end if
end for
Return $\bar{A}$

Procedure 1 UniqueIntersection(S, G): Determine if samples in S have a unique intersection.

Input: Set S with $2\overline{n}$ vectors $y_1, \dots y_{2\overline{n}}$ and graph G with $y_1, \dots, y_{2\overline{n}}$ as nodes.
Output: Indicator variable UNIQUE_INT
Partition S into sets $S_1, \dots, S_{\overline{n}}$ such that each $|S_t| = 2$.
if $|\{t|S_t \in G\}| > \frac{61\overline{n}}{64}$ **then**
  UNIQUE_INT $\leftarrow 1$
else
  UNIQUE_INT $\leftarrow 0$
end if
Return UNIQUE_INT

(*A*3) **Spectral Condition on Dictionary Elements:** The dictionary matrix has bounded spectral norm, for some constant $\mu_1 > 0$,

$$\|A\| < \mu_1 \sqrt{\frac{r}{d}}. \tag{4}$$

(*A*4) **Entries in Coefficient Matrix:** We assume that the non-zero entries of X are drawn from a zero-mean distribution supported on $[-M, -m] \cup [m, M]$ for some fixed constants m and M.

(*A*5) **Sparse Coefficient Matrix:** The columns of coefficient matrix have bounded number of non-zero entries s which are selected randomly, i.e.

$$|\operatorname{Supp}(x_i)| = s, \quad \forall\, i \in [n]. \tag{5}$$

We require s to be

$$s < c \min\left(\sqrt{\frac{m^2\sqrt{d}}{2M^2\mu_0}}, \sqrt[3]{\frac{r}{1536}}\right), \tag{6}$$

for some small enough constant c.

($A6$) **Sample Complexity:** Given a parameter $\alpha \in (0, 1/20)$ (which is related to the error in recovery of dictionary, see Theorem 2.1), and a universal constant $c > 0$, choose $\delta > 0$ and the number of samples n such that

$$n := n(d, r, s, \delta, \alpha) = \frac{cr}{\alpha^2 s} \log \frac{d}{\delta}, \quad n^2\delta < 1.$$

($A7$) **Choice of Threshold for Correlation Graph:** The correlation graph $G_{\mathrm{corr}(\rho)}$ is constructed using threshold ρ such that

$$\rho = \frac{m^2}{2} - \frac{s^2M^2\mu_0}{\sqrt{d}} > 0. \tag{7}$$

($A8$) **Choice of Separation Parameter ϵ_{dict} between Estimated Dictionary Elements:** This is the desired accuracy of the estimated dictionary elements to the true dictionary elements using just the initialization step. It can be chosen to be:

$$\frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right) < \epsilon_{\mathrm{dict}}^2 < \frac{1}{4}. \tag{8}$$

The assumption ($A1$) on normalization is without loss of generality since we can always rescale the dictionary elements and the corresponding coefficients and obtain the same observations. The assumption ($A2$) on incoherence is crucial to our analysis. In particular, incoherence also leads to a bound on the RIP constant; see Lemma A.5 in Appendix A.6. The assumption ($A3$) provides a bound on the spectral norm of A.

The assumption ($A4$) assumes that the non-zero entries of X are drawn from a zero-mean distribution with natural upper and lower bounds on the coefficients. Note that a similar assumption is made in the work of Arora et.al [8].

The assumption ($A5$) on sparsity in the coefficient matrix is crucial for identifiability of dictionary learning problem. We require for the sparsity to be not too large for recovery.

The assumption ($A6$) provides a bound on sample complexity. We subsequently establish that in order to have decaying error for recovery of dictionary elements, we require $n = \omega(r)$ samples for recovery. Thus, we obtain a nearly linear sample complexity for our method.

Assumption ($A7$) specifies the threshold for the construction of the correlation graph. Intuitively, we require a threshold such that we can distinguish pairs of samples which share a dictionary element from those which do not.

Main Result: We now present our main result which bounds the error in the estimates of Algorithm 1.

Algorithm 2 RecoverCoeff($Y, \bar{A}, \epsilon_{\text{coeff}}$): Exact Recovery through lasso

input Samples Y, approximate dictionary $\bar{A}$ and accuracy parameter ϵ_{coeff}. sign(X) returns a matrix with signs of the entries of X.

1: **for** samples $i = 1, 2, \ldots, n$ **do**

2: Estimate

$$\widehat{x}_i = \arg\min_{x \in \mathbb{R}^r} \|x\|_1, \quad \text{subject to} \quad \|y_i - \bar{A}x\|_2 \leq \epsilon_{\text{coeff}}. \tag{10}$$

3: Threshold: $\widehat{X} \leftarrow \text{sign}(\widehat{X})$.

4: **end for**

5: Estimate $\widehat{A} = Y\widehat{X}^T(\widehat{X}\widehat{X}^T)^{-1}$

6: Normalize: $\widehat{a}_i = \frac{\widehat{a}_i}{\|\widehat{a}_i\|_2}$

output $\widehat{A}$

Theorem 2.1 (Approximate recovery of dictionary)**.** *Suppose the output of Algorithm 1 is $\bar{A}$. Then with probability greater than $1 - 2n^2\delta$, there exists a permutation matrix P such that:*

$$\epsilon_A^2 := \min_{i \in [r]} \min_{z \in \{-1,+1\}} \left\| z a_i - (P\bar{A})_i \right\|_2^2 < \frac{32sM^2}{m^2} \left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}} \right). \tag{9}$$

Remark: Note that we have a sign ambiguity in recovery of the dictionary elements, since we can exchange the signs of the dictionary elements and the coefficients to obtain the same observations. The assumption on sparsity in ($A4$) implies that the first two terms in (9) decay. For the third term in (9) to decay, we require $s = o(r^{1/4})$ instead of $s = \mathcal{O}(r^{1/3})$ as in ($A4$). Moreover, we require that $\alpha^2 s = o(1)$. Since the sample complexity in ($A7$) scales as $n = \Omega\left(\frac{r}{\alpha^2 s}\right)$, we require $n = \omega(r)$ samples for recovery of dictionary with decaying error. Thus, we obtain a near linear sample complexity for our method. We observe that the error in our estimation depends inversely on dimension-related quantities such as d and r and not on the number of samples n. This is because the errors in our estimates arise from errors in SVD step, specifically from the discrepancy between the SVD vector and the dictionary element responsible for a cluster. Even the population SVD will suffer from an approximation error here, which is responsible for our error bound, but the probability in the error bound improves with the number of samples as we get closer and closer to the population SVD estimate.

2.2 Post-processing for binary coefficients

We now present the post-processing step which will be analyzed under a more stringent condition on the coefficients.

2.2.1 Algorithm

Once we obtain an estimate of the dictionary elements, we proceed to estimate the coefficient matrix. The main observation at this step is that the coefficient vector x_i for each sample y_i is a s-sparse vector in r-dimensions. Hence, recovering the coefficients would be a standard sparse linear problem if we knew the dictionary A exactly. Our analysis will show that even an approximately correct dictionary $\bar{A}$ from Algorithm 1 suffices to provide guarantees for this recovery.

Once the coefficients are estimated, the dictionary can be re-estimated by solving another linear system. The procedure is formally described in Algorithm 2. We do not prescribe any particular choice of computational procedure to solve the optimization problem (10), but there are many algorithms available in standard literature. As a concrete example, the GraDeS algorithm of Garg and Khandekar [17] or OMP of Tropp and Gilbert [33]works in our setting.

2.2.2 Exact recovery for bernoulli coefficients

Our second result is that under stronger conditions than before, it is possible to exactly recover the unknown dictionary A with high probability. This result will be obtained by initializing Algorithm 2 with the output of Algorithm 1. We start with the additional assumptions, putting restrictions on the allowed sparsity level s as a function of r and d.

Assumption B1 (Conditions for exact recovery)**.** *The non-zero coefficients in coefficient matrix X are zero-mean Bernoulli$\{-1, 1\}$. This corresponds to setting $M = m = 1$ in Assumption (A4).*

The sparsity level s, and the number of dictionary elements r and the observed dimension d satisfy

$$32s\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d}\right) \leq \frac{1}{1200s^2}, \quad \text{and} \quad \frac{32s^4}{r} \leq \frac{1}{1200s^2}.$$

The constant α in Theorem 2.1 satisfies

$$32s\left(\alpha^2 + \frac{\alpha}{\sqrt{s}}\right) \leq \frac{1}{1200s^2}.$$

The number of samples n, in addition to assumption (A6), satisfies

$$n \geq \frac{4r}{c_0} \log \frac{d}{\delta},$$

where c_0 is a universal constant.

The accuracy parameter ϵ_{coeff} in Algorithm 2 is chosen as $\epsilon_{\text{coeff}} = s\epsilon_A$, where ϵ_A is the error in estimating the dictionary elements in (9).

Theorem 2.2 (Exact recovery for bernoulli coefficients)**.** *Under the conditions of Theorem 2.1, and suppose, in addition Assumption B1 holds, then the output $\widehat{A}$ of Algorithm 2 initialized with Algorithm 1 satisfies $\widehat{A} = A$ up to permutation of columns, with probability at least $1 - 3n^2\delta$.*

Remark: Assumption B1 for exact recovery places more stringent conditions on the distribution of the coefficients and the sparsity level s, compared to (A4) for approximate recovery. While for approximate recovery, we require $s = \mathcal{O}(d^{1/4}, r^{1/4})$, in Assumption B1, we require $s = \mathcal{O}(d^{1/5}, r^{1/6})$ for exact recovery. Note that the additional constraint on sample complexity n in Assumption B1 still has the same scaling, and thus, $n = \mathcal{O}(r(\log r + \log d))$ suffices both for approximate and exact recovery.

We also observe that the result of Theorem 2.2 relies on Algorithm 1 as the initialization procedure, but in principle we can also use a different approximate recovery procedure to initialize Algorithm 2. In particular, a different initialization procedure with a better error guarantee would also directly translate to better recovery properties in the second step, in terms of the assumptions relating s to r and d. Understanding these issues appears to be an interesting direction for future research.

3 Proofs of main results

In this section we will present the proofs of our main results, Theorems 2.1 and 2.2. We will start by presenting a host of useful lemmas, and sketch out how they fit together to yield the main results before moving on to the proofs.

3.1 Correlation graph properties

In this section we will present some useful properties of the correlation graph $G_{\text{corr}(\rho)}$ described in Section 1.3. Recall that $G_{\text{corr}(\rho)}$, where the nodes are samples $\{y_1, y_2, \dots y_n\}$ and an edge $(y_i, y_j) \in G_{\text{corr}(\rho)}$ implies that $|\langle y_i, y_j \rangle| > \rho$, for some $\rho > 0$. This is employed by Algorithm 1 as a proxy for identifying samples which have common dictionary elements. We now make this connection concrete in the next few lemmas. For this we also recall our notation $\mathcal{N}_B(y)$ which is the neighborhood of a sample y in the coefficient bipartite graph (see Figure 1a), that is, the set of dictionary elements that combine to yield y.

Lemma 3.1 (Correlation graph)**.** *Under the incoherence assumption* $(A2)$ *and the threshold* ρ *in assumption* $(A7)$*, the following is true for the edges in the correlation graph* $G_{\text{corr}(\rho)}$*:*

$$|\mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l)| = 1 \Rightarrow (y_k, y_l) \in G_{\text{corr}(\rho)}, \quad \forall\, i \in [r], \tag{11}$$

$$(y_k, y_l) \in G_{\text{corr}(\rho)} \Rightarrow |\mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l)| \geq 1, \tag{12}$$

for all $k, l \in \{1, 2, , \dots, n\}, k \neq l$*.*

Lemma 3.1 suggests that nodes which intersect in *exactly one* dictionary element are special, in that they are guaranteed to have an edge between them in $G_{\text{corr}(\rho)}$. Our next lemma works towards establishing something even stronger. We will next establish that there are large cliques in the correlation graph where any two samples in the clique intersect in the same unique dictionary element. In order to state the lemma, we need some additional notation.

For each dictionary element a_i, consider a set of samples[3] $\{y_k, k \in S\}$, for some $S \subset \{1, 2, \dots, n\}$, such that they only have a_i in common, and denote such a set by $\mathcal{C}_i$ i.e.

$$\mathcal{C}_i := \{y_k, k \in S : \mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l) = \{a_i, \}\ \forall\, k, l \in S\}. \tag{13}$$

Lemma 3.1 implies that in the correlation graph, the set of nodes in $\mathcal{C}_i$ form a clique (not necessarily maximal), for each $i \in \{1, 2, \dots, r\}$, as shown in Figure 1b. The above implication can be exploited for recovery of dictionary elements: if we find the set $\mathcal{C}_i$, then we can hope to recover the element a_i, since that is the only element in common to the samples in $\mathcal{C}_i$.

For ease of stating the next lemma, we further define two shorthand notations.

$$\text{Uniq-intersect}(y_i, y_j) := \{(y_i, y_j) \in G_{\text{corr}(\rho)} \quad \text{and} \quad |\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j)| = 1\}, \tag{14}$$

Intuitively, the samples satisfying Uniq-intersect(y_i, y_j) are guaranteed to have an edge between them by Lemma 3.1. In order to guarantee large cliques, we will also need to measure the number of triangles in $G_{\text{corr}(\rho)}$.

[3]Note that such a set need not be unique.

In order to do this, given anchor samples y_{i^*} and y_{j^*} have a unique intersection, we now bound the probability that a randomly chosen sample y_i, among the neighborhood set of y_{i^*} and y_{j^*} in the correlation graph also has a unique intersection. Now define unique intersection event for a new sample y_i with respect to anchor samples y_{i^*} and y_{j^*} as follows

$$\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) := \{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{i^*}) = \mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{j^*}) = \{a_k\}\}, \tag{15}$$

where $a_k = \mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})\}$ is the unique intersection of the anchor samples y_{i^*} and y_{j^*}. In other words, Uniq-intersect$(y_i; y_{i^*}, y_{j^*})$ indicates the event that the pairwise intersections of the new sample y_i with each of the anchors y_{i^*} and y_{j^*} is unique and equal to the unique intersection of y_{i^*} and y_{j^*}.

Lemma 3.2 (Formation of clique under good anchor samples)**.**

$$\mathbb{P}\left[\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*}),\ \textit{and}\ (y_i, y_{i^*}), (y_i, y_{j^*}) \in G_{\text{corr}(\rho)}\right] \geq 1 - \frac{s^3}{r}.$$

Lemma 3.2 is crucial for our algorithm. It guarantees that given a pair of good anchor elements—one satisfying unique intersection property—a large fraction of their neighrbors also contain this common dictionary element. Some further arguments can then be made to establish that a large fraction of the neighbors of y_{i^*} and y_{j^*} also have edges amongst themselves and hence form cliques as defined in Equation 13.

3.2 Correctness of Procedure 1

A key component in our analysis is the correctness of Procedure 1. As we saw in the previous lemmas, it is crucial for a chosen pair of anchor elements to have a unique intersection in order to use them for identifying large cliques $\mathcal{C}_i$ in $G_{\text{corr}(\rho)}$. Procedure 1 plays a crucial role by providing a verifiable test for whether a pair of anchor elements have a unique intersection or not. Our next two lemmas help us establish that this test is sound with high probability. We first show that two neighbors of a bad anchor pair do not have an edge amongst them with high probability.

Denote the event

$$\Delta(y_i, y_j, y_k) := \{(y_i, y_j), (y_j, y_k), (y_i, y_k) \in G_{\text{corr}(\rho)}\},$$

i.e., the samples y_i, y_j, y_k form a triangle in the correlation graph.

Lemma 3.3 (Detection of bad anchor samples)**.** *For randomly chosen samples y_i, y_j*

$$\mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \neg\, \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] > \frac{1}{16}.$$

Intuitively, this means that the number of sets S_i which will be edges in $G_{\text{corr}(\rho)}$ is rather small for an anchor pair with multiple dictionary elements in common. In order for correctness of the procedure, we will in fact need this number to be substantially smaller than that for a good anchor pair. This is indeed the case as we next establish.

Lemma 3.4 (Detection of good anchor samples)**.** *For randomly chosen samples y_i, y_j*

$$\mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \leq \frac{24s^3}{r}.$$

Combining the above two lemmas, the correctness of Procedure 1 naturally follows.

Proposition 3.1 (Correctness of Procedure 1)**.** *Suppose $(y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}$. Suppose that $s^3 \leq r/1536$ and $\gamma \leq 1/64$. Then Algorithm 1 returns the value of* Uniq-intersect(y_{i^*}, y_{j^*}) *correctly with probability greater than $1 - 2\exp(-\gamma^2\overline{n})$.*

3.3 Proof of Theorem 2.1

In this section we will put all the pieces together and establish Theorem 2.1. We start by establishing that given a pair of good anchor elements, the SVD step in Algorithm 1 approximately recovers the unique dictionary element in the intersection of the two anchors.

Proposition 3.2 (Accuracy of SVD)**.** *Consider anchor samples y_{i^*} and y_{j^*} such that* Uniq-intersect(y_{i^*}, y_{j^*}) *is satisfied, and wlog, let $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*}) = \{a_1\}$. Recall the definition of $\widehat{S}$ (25), and further define $\widehat{L} := \sum_{i \in \widehat{S}} y_i y_i^\top$ and $\widehat{n} = |\widehat{S}|$. If $\widehat{a}$ is the top singular vector of $\widehat{L}$, then there exists a universal constant c such that we have:*

$$\min_{z \in \{-1,1\}} \|\widehat{a} - z a_1\|_2^2 < \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right),$$

with probability greater than $1 - d\exp\left(-c\alpha^2\widehat{n}\right)$ for $\alpha < 1/20$.

Given the above proposition, the proof of Theorem 2.1 is relatively straightforward. Indeed, the key missing piece is the dependence on the random quantity $|\widehat{S}|$ in the error probability in Proposition 3.2. We now present the proof.

Proof of Theorem 2.1:

Consider a particular iteration of Algorithm 1. Procedure 1 returns Uniq-intersect(y_{i^*}, y_{j^*}) with probability greater than $1 - 2\exp(-\gamma^2|\widehat{S}|/2)$. If $\neg$ Uniq-intersect(y_{i^*}, y_{j^*}), then Algorithm 1 proceeds to the next iteration. Consider the case of Uniq-intersect(y_{i^*}, y_{j^*}) and suppose $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*}) = \{a_l\}$. Using Proposition 3.2, with probability greater than $1 - d\exp\left(-c\alpha^2|\widehat{S}|\right)$, we have:

$$\|a_l - \widehat{a}\|_2^2 < \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right).$$

Using Lemma A.4 and Lemma 3.1, we see that $|\widehat{S}| \geq \frac{ns}{4r}$ with probability greater than $1 - \exp\left(\frac{-ns}{16r}\right)$. Using a union bound over all the iterations (which are at most n^2), the above claims hold for all iterations with probability greater than $1 - n^2 d\exp\left(\frac{-c\alpha^2 ns}{r}\right) - 2n^2\exp\left(\frac{-\gamma^2 ns}{8r}\right) - n^2\exp\left(\frac{-ns}{16r}\right)$.

Using Lemma A.4 and Lemma 3.1, with probability greater than $1 - r\exp\left(\frac{-ns}{64r}\right)$, for every $l \in [r]$, there are at least $\frac{ns}{8r}$ pairs (i^*, j^*) such that $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*}) = \{a_l\}$ and $(i^*, j^*) \in G_{\text{corr}(\rho)}$. Lines 9-11 of the algorithm then ensure that there is a unique copy of the approximation to a_l dictionary element. Using a union bound now gives the result.

□

3.4 Analysis of post-processing step

In this section, we will show how to clean up the approximate recovery of the previous section and obtain exact recovery of the dictionary under Assumption B1. We start by setting up the problem as that of sparse estimation with deterministic noise and describing some guarantees in a general setup. We then specialize these to the assumptions of our problem and present the proof of Theorem 2.2.

3.4.1 Lasso with determinstic noise

Recalling the model (1), we see that each observation y_i is generated according to the linear model

$$y_i = Ax_i, \quad \text{for } i = 1, 2, \ldots, n,$$

where x_i is a s-sparse vector in r dimensions. If we knew the dictionary A, then this is the usual sparse linear system. Given the knowledge of an approximate dictionary $\bar{A}$ however, we can rewrite the system as

$$y_i = \bar{A}x_i + \underbrace{(A - \bar{A})x_i}_{w_i}, \tag{16}$$

where $W \in \mathbb{R}^{d \times n}$ is the error matrix. Note that the errors in W are not zero mean, or even independent of $\bar{A}$ unlike typical statistical settings. Under our initialization, however, they are bounded, which we establish subsequently. For the remainder of this section, we assume the following facts about $\bar{A}$. Note that this is not an assumption about the model, but a condition on the output of Algorithm 1, which will be proved in the next section.

Assumption C1 (Approximate initialization)**.** *Assume that $\bar{A}$ is an approximately correct initialization for A, meaning the following hold:*

RIP: *The $2s$-RIP constant of the matrix $\bar{A}$, $\delta_{2s} < \frac{1}{7}$. That is, for every $S \subseteq \{1, 2, \ldots, r\}$ with $|S| \leq 2s$, the smallest and largest singular values, $\sigma_{\min}$ and $\sigma_{\max}$ respectively of the $d \times |S|$ matrix $\bar{A}_S$ satisfy:*

$$\frac{6}{7} < \sigma_{\min} < \sigma_{\max} < \frac{8}{7}.$$

Bounded error: *$\|\bar{a}_i - a_i\|_2 \leq \epsilon_A$ for all $i = 1, 2, \ldots, r$.*

Under these general assumptions, we can provide a guarantee on the error incurred in (10) in step (2) of Algorithm 2. While this result has been obtained in many contexts by various authors, we use the following precise form from Candes [13].

Theorem 3.1 (Theorem 1.2 from Candes [13])**.** *Suppose y_i is generated according to the linear model* (16)*, where x_i is s-sparse and assume that $\delta_{2s} \leq \sqrt{2} - 1$. Then the solution to Equation* (10) *obeys the following, for a universal constant C_1,*

$$\|\widehat{x}_i - x_i\|_2 \leq C_1 \|w_i\|_2.$$

In particular, $C_1 = 8.5$ suffices for $\delta_{2s} \leq 0.2$.

3.4.2 Proof of Theorem 2.2

In order to prove Theorem 2.2, we first establish that under our assumptions, the coefficients x_i are exactly recovered in Equation (10). Once this is established, Theorem 2.2 follows in a straightforward manner. We start with a useful proposition.

Proposition 3.3. *Under conditions of Theorem 2.1, assume further that $\epsilon_A \leq 1/(20s)$ for the dictionary returned in Algorithm 1. Then Algorithm 2 guarantees that $\widehat{x}_i = x_i$ for all $i = 1, 2, \ldots, n$.*

Proof: We would like to use Theorem 3.1 to show that we recover the coefficients x_i correctly in the lasso step (10) of Algorithm 2. In order to do this, we first need to verify Assumption C1 for the dictionary returned by Algorithm 1, and then obtain bounds on the quantity $\|w_1\|_2$. We start with the former.

Consider any $2s$-sparse subset S of $[r]$. We have:

$$\sigma_{\min}(\bar{A}_S) \geq \sigma_{\min}(A_S) - \|A_S - \bar{A}_S\|_2 \overset{(\zeta_1)}{\geq} 1 - \frac{2\mu_0 s}{\sqrt{d}} - \left\|A_S - \bar{A}_S\right\|_F \quad \text{and,}$$
$$\sigma_{\max}(\bar{A}_S) \leq \sigma_{\max}(A_S) + \|A_S - \bar{A}_S\|_2 \overset{(\zeta_2)}{\leq} 1 + \frac{2\mu_0 s}{\sqrt{d}} + \left\|A_S - \bar{A}_S\right\|_F,$$

where ζ_1 and ζ_2 follow from Lemma A.5 in Appendix A.6. Since A_S is a $d \times 2s$ matrix, it satisfies that $\left\|A_S - \bar{A}_S\right\|_F \leq \sqrt{2s\epsilon_A}$. Given the assumption $\epsilon_A \leq 1/(20s)$, it immediately follows that the minimum and maximum singular values of $\bar{A}_s$ are at least $6/7$ and $8/7$ respectively, so that we obtain $\delta_{2s} = 1/7 < 0.2$.

This shows that $\bar{A}$ satisfies Assumption C1. Next we bound the ℓ_2 norm of the noise vector w_i. Again bounding the frobenius norm of the error in the dictionary in the same way as above, we obtain

$$\|w_i\|_2 \leq \|(A - \bar{A})_{S_i}\|_2 \|x_i\|_2 \leq \left\|(A - \bar{A})_{S_i}\right\|_F \sqrt{s} \leq s\epsilon_A,$$

where S_i is the support of x_i. Consequently, we obtain from Theorem 3.1 that the output $\widehat{x}_i$ of Equation 10 satisfies

$$\|\widehat{x}_i - x_i\|_2 \leq C_1\, s\epsilon_A \leq 9s\epsilon_A \leq 9/20. \tag{17}$$

We now observe that an ℓ_2 error guarantee is also an ℓ_∞ error guarantee. Recall that by the model assumption, each non-zero coefficient of X has an absolute value of 1. Since Equation (17) guarantees that the ℓ_2 error guarantee is no larger than $1/2$, all the coefficients will be uniquely recovered and hence $\widehat{x}_i = x_i$.

□

Proof of Theorem 2.2:

We are now ready to provide our proof of exact recovery. Based on Proposition 3.3, we only need to verify two things. First is that the initialization $\bar{A}$ satisfies $\epsilon_A < 1/(20s)$ and the second is that the linear system $Y = AX$ is well-posed when we solve for A. In order to verify the former, we observe that our additional conditions in Assumption B1 guarantee that

$$
\begin{aligned}
32s\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}\right) &\leq \frac{1}{1200s^2},\\
\frac{32s^4}{r} &\leq \frac{1}{1200s^2}, \quad \text{and}\\
32s\left(\alpha^2+\frac{\alpha}{\sqrt{s}}\right) &\leq \frac{1}{1200s^2}.
\end{aligned}
$$

Hence we obtain from Theorem 2.1 that with probability at least $1-n^2d\exp\left(\frac{-c\alpha^2 ns}{r}\right)-4\max(n^2,r)\exp\left(\frac{-ns}{32768r}\right)$, $\epsilon_A < 1/(20s)$. Hence, it only remains to verify that the linear system is well-posed.

According to Lemma A.7 in Appendix A.6, the matrix $\mathbb{E}\left[XX^T\right] = \frac{s}{r}I_{r\times r}$ so that all of its singular values are equal to s/r. We now appeal to Theorem A.1 with $W = X$, $d = r$ and $u = \sqrt{s}$. Then we obtain for any $t > 0$ with probability at least $1 - r\exp(-ct^2)$

$$
\sigma_{\min}(XX^T) \geq \frac{ns}{r} - n\max\left\{\sqrt{\frac{s}{r}}\delta, \delta^2\right\},
$$

where $\delta = t\sqrt{s/n}$. Substituting the value of δ, we obtain the lower bound

$$
\begin{aligned}
\sigma_{\min}(XX^T) &\geq \frac{ns}{r} - n\max\left\{\sqrt{\frac{s}{r}}t\sqrt{\frac{s}{n}}, t^2\frac{s}{n}\right\}\\
&\geq \frac{ns}{r}\left(1 - t\sqrt{\frac{r}{n}} - \frac{t^2 r}{n}\right)\\
&= \frac{ns}{4r},
\end{aligned}
$$

for $t = \sqrt{n/(4r)}$. This means that the linear system is well-posed with probability at least $1 - r\exp(-cn/(4r))$. Choosing c_0 to now be $\min(c, 1/32768)$ finishes the proof. □

4 Discussion and Conclusion

In this paper, we proposed simple and tractable methods for dictionary learning. We present a novel clustering-based approach which can approximately recover the uknown overcomplete dictionary from samples. We also analyzed a simple denoising strategy based on sparse recovery algorithms for reconstructing the dictionary exactly under some simplifying assumptions on the model. In particular, the second step is not tied to the first step in any critical way, and more sophisticated post-processing procedures have since been developed. There is of course, also room for developing better approximate recovery schemes, building on our work.

In the analysis of the clustering step, we provide guarantees when the coefficient matrix is sparse and randomly drawn. In principle, our analysis can be extended to general sparse coefficient matrices and can be cast as a higher-order expansion condition on the coefficient bipartite graph. Similar (and yet not the same) expansion conditions have appeared in other contexts involving

learning of overcomplete models. For instance, in [5], Anandkumar et. al. establish that under an expansion condition on the topic-word matrix, unsupervised learning of the model is possible. Here, the hidden topics correspond to dictionary elements, and the observed words correspond to the samples in the dictionary setting.

Finally, our work suggests some natural and interesting directions for future research. While both the steps of our algorithm seem inherently robust to noise, it remains important to quantify the recovery properties when the observations are noisy in future work. Another natural question is raised by the fact that we use only one step of lasso and least squares for exact recovery. Indeed, the subsequent work [1] analyzes a generalization where we perform multiple iterations of lasso followed by subsequent dictionary estimation, and is able to exactly recover the dictionary under a much broader set of conditions. Since our study was motivated by natural applications of dictionary learning in signal processing and machine learning, it would also be interesting to investigate how our provably correct procedures perform compared to the popular heuristic methods.

Acknowledgements

A. Agarwal thanks Yonina Eldar for suggesting the problem to him. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, and ARO YIP Award W911NF-13-1-0084. P. Netrapalli thanks Yash Deshpande for helpful discussions. The authors thank Matus Telgarsky for suggesting Lemma A.6 and thank Sham Kakade and Dean Foster for initial discussions.

A Proofs for clustering analysis

In this section we will provide the proofs of many of the Lemmas along with some auxilliary results in Sections 3.1- 3.3. Some of the more technical results that are required will be deferred to Appendix A.6.

A.1 Proofs of correlation graph properties

We start by proving Lemmas 3.1 and 3.2 in Section 3.1.

Proof of Lemma 3.1:

We first prove (12) via contradiction. Suppose $\mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l) = \emptyset$, we then have

$$\begin{aligned}|\langle y_k, y_l\rangle| &= |\sum_{i,j} x_{ik}x_{jl}\langle a_i, a_j\rangle| \leq \sum_{i,j} |x_{ik}x_{jl}\langle a_i, a_j\rangle| \\ &\leq |\mathcal{N}_B(y_k)| \cdot |\mathcal{N}_B(y_l)| \cdot \max_{i,j,k,l} |x_{ik}x_{jl}| \cdot \max_{i\neq j} |\langle a_i, a_j\rangle| \leq \frac{s^2M^2\mu_0}{\sqrt{d}}\end{aligned}$$

For (11), let $\{a_{i^*}\} = \mathcal{N}_B(y_k) \cap \mathcal{N}_B(y_l)$

$$\begin{aligned}|\langle y_k, y_l\rangle| &= |\sum_{i,j} x_{ik}x_{jl}\langle a_i, a_j\rangle| \geq |x_{i^*k}x_{i^*l}|\langle a_{i^*}, a_{i^*}\rangle - \sum_{i\neq j} |x_{ik}x_{jl}\langle a_i, a_j\rangle| \\ &\geq m^2 - \frac{s^2M^2\mu_0}{\sqrt{d}},\end{aligned}$$

using the above analysis. The claims now follow from the setting of ρ. □

We next establish Lemma 3.2.

Proof of Lemma 3.2: Define the event

$$\mathcal{A} := \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{i^*})| \geq 1\} \cap \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{j^*})| \geq 1\}.$$

From Lemma 3.1, we have that

$$\begin{aligned}&\mathbb{P}\left[\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*}),\ \text{and } (y_i, y_{i^*}), (y_i, y_{j^*}) \in G_{\text{corr}(\rho)}\right] \\ &\geq \mathbb{P}\left[\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*}), \mathcal{A}\right]\end{aligned}$$

In order to lower bound $\mathbb{P}\left[\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*}), \mathcal{A}\right]$, we instead upper bound the probability of the complementary event $\mathbb{P}\left[\neg\,\text{Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*}), \mathcal{A}\right]$

In order to do so, we first bound the following

$$\mathbb{P}\left[\mathcal{A} \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \geq \frac{s}{r}, \tag{18}$$

since $\mathcal{A}$ holds when the unique element in $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})$ is chosen and its probability is s/r. We also have

$$\mathbb{P}\left[\neg \text{ Uniq-intersect}(y_i; y_{i^*}, y_{j^*}) \cap \mathcal{A} \,\middle|\, \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \leq \frac{(s-1)^2\binom{r-3}{s-2}}{\binom{r}{s}},$$

since for $\neg$ Uniq-intersect$(y_i; y_{i^*}, y_{j^*})$ to hold, we need to choose at least one of the $s-1$ elements in $\mathcal{N}_B(y_{i^*})/\mathcal{N}_B(y_{j^*})$, and similarly one from the $s-1$ elements of $\mathcal{N}_B(y_{j^*})/\mathcal{N}_B(y_{i^*})$. The rest of the $s-2$ elements can be picked arbitrarily from the $r-3$ dictionary atoms that remain after excluding the two already picked and the unique intersection $\mathcal{N}_B(y_{j^*}) \cap \mathcal{N}_B(y_{i^*})$.

It is easy to check that

$$\begin{aligned}\frac{(s-1)^2\binom{r-3}{s-2}}{\binom{r}{s}} &= \frac{(s-1)^2(r-s)s(s-1)}{r(r-1)(r-2)} \\ &\leq \frac{s^4}{r^2}. \end{aligned} \tag{19}$$

Taking the ratio of the two bounds in (18) and (19) completes the proof. □

A.2 Proofs of Lemmas 3.3 and 3.4

We now prove the two lemmas that are crucial to establishing the correctness of Prcedure 1.

Proof of Lemma 3.3: Let $\mathcal{A}_1$ and $\mathcal{A}_2$ denote the following events:

$$\begin{aligned}\mathcal{A}_1 :=& \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{i^*})| \geq 1\} \cap \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{j^*})| \geq 1\} \\ & \cap \{|\mathcal{N}_B(y_j) \cap \mathcal{N}_B(y_{i^*})| \geq 1\} \cap \{|\mathcal{N}_B(y_j) \cap \mathcal{N}_B(y_{j^*})| \geq 1\} \\ \mathcal{A}_2 :=& \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{i^*})| = 1\} \cap \{|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{j^*})| = 1\} \\ & \cap \{|\mathcal{N}_B(y_j) \cap \mathcal{N}_B(y_{i^*})| = 1\} \cap \{|\mathcal{N}_B(y_j) \cap \mathcal{N}_B(y_{j^*})| = 1\} \end{aligned} \tag{20}$$

In words, both y_i and y_j have at least dictionary element in common with each of y_{i^*} and y_{j^*} under the event $\mathcal{A}_1$, while the number of common elements is *exactly one* under the event $\mathcal{A}_2$. We have

$$\begin{aligned}&\mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\ &\overset{(a)}{=} \mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)} \mid \mathcal{A}_1, \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\ &= \mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)}, \Delta(y_j, y_{i^*}, y_{j^*}) \mid \mathcal{A}_1, \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\ &\geq \mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)}, \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}) \mid \mathcal{A}_1, \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*}), (y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}\right] \\ &\overset{(b)}{\geq} \mathbb{P}\left[(y_i, y_j) \notin G_{\text{corr}(\rho)}, \mathcal{A}_2 \mid \mathcal{A}_1, \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*}), (y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}\right] \\ &\overset{(c)}{\geq} \mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \mathcal{A}_1, \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*}), (y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}\right], \end{aligned} \tag{21}$$

where the inequalities (a), (b) and (c) follow from Lemma 3.1. We will now work on lower bounding this resulting probability.

We first lower bound the numerator in writing the above conditional probability as the ratio of a joint to marginal probability. We begin by noting that

$$\mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*}, (y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)})\right]$$
$$= \mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*}), (y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}\right]$$

Let us define $\widehat{l} = |\mathcal{N}_B(y_{i^*}) \cup \mathcal{N}_B(y_{j^*})| \in [s, 2s]$ and $l = |\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})| \geq 2^4$. The event in the probability above, that is $\mathcal{A}_2$ holds while y_i and y_j do not share a dictionary element, can be arranged by choosing two of the l elements, and assigning a unique element to each y_i and y_j. Similarly the remaining elements can be chosen outside $\mathcal{N}_B(y_{i^*}) \cup \mathcal{N}_B(y_{j^*})$ in a non-overlapping manner: for y_i assign $s-1$ elements among $r - \widehat{l}$ elements, and then for y_j assign from remaining $r - \widehat{l} - s + 1$ elements. This logic yields the following lower bound on the probability

$$\mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*})\right]$$
$$\geq \frac{2\binom{l}{2}\binom{r-\widehat{l}}{s-1}\binom{r-\widehat{l}-s+1}{s-1}}{\binom{r}{s}^2} \geq \frac{2\binom{l}{2}\binom{r-2s}{s-1}\binom{r-3s+1}{s-1}}{\binom{r}{s}^2},$$

where the second inequality uses $\widehat{l} \leq 2s$. Now with some straightforward algebra, we can further lower bound this expression as

$$\mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*})\right]$$
$$\geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{3s-3}{r-s}\right)^{s-1}\left(1 - \frac{2s-1}{r-s}\right)^{s-1}$$
$$\geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{3s}{r-s}\right)^{s}\left(1 - \frac{2s}{r-s}\right)^{s}.$$

Now we invoke Lemma A.6 to further lower bound the RHS and obtain

$$\mathbb{P}\left[\{\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*})\right]$$
$$\geq \frac{s^2(l-1)^2}{r^2}\exp\left(-\frac{3s^2}{r-s}\right)\exp\left(-\frac{2s^2}{r-s}\right) \geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{10s^2}{r-s}\right)$$
$$\geq \frac{s^2(l-1)^2}{2r^2},$$

where the final inequality holds since $s^2 \leq r/40$.

In order to lower bound the conditional probability in Equation 21, we need to further upper bound the marginal probability in the denominator. To this end, we observe that we have to upper bound $\mathbb{P}\left[\mathcal{A}_1 | \neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*})\right]$. Now conditioned on $\neg\,\text{Uniq-intersect}(y_{i^*}, y_{j^*})$, for each y_i and y_j, $\mathcal{A}_1$ can be satisfied in two ways: choose at least one element from l elements in $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})$ or choose at least two elements from $m - l$ elements in $\mathcal{N}_B(y_{i^*}) \cup \mathcal{N}_B(y_{j^*})$. Making this precise, we obtain

[4] the intersection is at least 1 by Lemma 3.1

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_1 \mid \neg \text{ Uniq-intersect}(y_{i^*}, y_{j^*})\right] &\leq \left(\frac{ls}{r} + \frac{(\widehat{l} - l)^2 \binom{r-2}{s-2}}{\binom{r}{s}}\right)^2 \\
&\leq \left(\frac{ls}{r} + \frac{s^2(\widehat{l} - l)^2}{(r-1)^2}\right)^2 \\
&\leq \left(\frac{ls}{r} + \frac{s^2(2s-2)^2}{(r-1)^2}\right)^2 \\
&\leq \frac{2l^2 s^2}{r^2}, \text{ (since } 4s^3 < r-1)
\end{aligned}$$

The result follows by using the fact that $l \geq 2$. □

The proof of Lemma 3.4 is similar, but involves controlling slightly different events.

Proof of Lemma 3.4:

We will establish the lemma by lower bounding the probability of the complementary event. We recall the events $\mathcal{A}_1$ and $\mathcal{A}_2$ defined in Equation 20 in the proof of Lemma 3.3. We can mimick the initial arguments in the proof of Lemma 3.3 to conclude that

$$\begin{aligned}
&\mathbb{P}\left[(y_i, y_j) \in G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&\geq \mathbb{P}\left[\text{Uniq-intersect}(y_i, y_j) \cap \mathcal{A}_2 \mid \mathcal{A}_1, \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right],
\end{aligned}$$

and we provide a lower bound for this. Once again, we express the conditional probability as the ratio of a joint to a marginal and then lower bound the numerator and upper bound the denominator. In the numerator, we have the event

We have

$$\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(y_i, y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&= \mathbb{P}\left[\text{Uniq-intersect}(y_i, y_j) \cap \mathcal{A}_2 \mid \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right]
\end{aligned}$$

The event Uniq-intersect$(y_i, y_j) \cap \mathcal{A}_2$ is guaranteed to occur if we choose y_i and y_j so that they have the only element in $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})$ in common. This yields the lower bound

$$\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(y_i, y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&\geq \frac{\binom{r-2s+1}{s-1}\binom{r-3s+2}{s-1}}{\binom{r}{s}^2}.
\end{aligned}$$

It is easy to further conclude that

$$
\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(y_i, y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&\geq \frac{s^2}{r^2}\left(1 - \frac{3s-3}{r-s+1}\right)^{(s-1)}\left(1 - \frac{2s-2}{r-s+1}\right)^{s-1} \\
&\geq \frac{s^2}{r^2}\exp(-5(s-1)^2/(r-s+1)) \\
&\geq \frac{s^2}{r^2}\left(1 - \frac{10s^2}{r-s}\right) \geq \frac{s^2}{r^2}\left(1 - \frac{20s^2}{r}\right),
\end{aligned}
$$

where we again invoked Lemma A.6 as well as the fact that $s \leq r/2$. As for the marginal probability in the denominator, we need to upper bound

$$
\begin{aligned}
\mathbb{P}\left[\mathcal{A}_1 \mid \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] &\leq \left(\frac{s}{r} + \frac{(2s-1)^2\binom{r-2}{s-2}}{\binom{r}{s}}\right)^2 \\
&\leq \left(\frac{s}{r} + \frac{(2s-1)^2(s-1)^2}{(r-1)^2}\right)^2 \leq \frac{s^2}{r^2}\left(1 + \frac{4s^3}{r}\right)^2,
\end{aligned}
$$

since for each y_i and y_j, $\mathcal{A}_1$ can be satisfied in two ways: choose the unique element from $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})$ or choose at least two elements from $2s-1$ elements in $\mathcal{N}_B(y_{i^*}) \cup \mathcal{N}_B(y_{j^*})$.

Using the above two inequalities, we have:

$$
\begin{aligned}
&\mathbb{P}\left[(y_i, y_j) \in G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&\geq \frac{1 - \frac{20s^2}{r}}{\left(1 + \frac{4s^3}{r}\right)^2}.
\end{aligned}
$$

It is easy to verify that $1/(1+x)^2 \leq 1 - x$ for $0 \leq x \leq (\sqrt{2}-1)/2$. Since $s^3 \leq r/5$, we obtain

$$
\begin{aligned}
&\mathbb{P}\left[(y_i, y_j) \in G_{\text{corr}(\rho)} \mid \Delta(y_i, y_{i^*}, y_{j^*}), \Delta(y_j, y_{i^*}, y_{j^*}), \text{Uniq-intersect}(y_{i^*}, y_{j^*})\right] \\
&\geq \left(1 - \frac{20s^2}{r}\right)\left(1 - \frac{4s^3}{r}\right) \\
&\geq 1 - \frac{24s^3}{r}.
\end{aligned}
$$

□

A.3 Proof of Proposition 3.1

Let us start with the case when $\text{Uniq-intersect}(y_{i^*}, y_{j^*}) = 1$. For any pair (y_i, y_j) where y_i and y_j are taken from $\mathcal{N}_{G_{\text{corr}(\rho)}}(y_{i^*}) \cap \mathcal{N}_{G_{\text{corr}(\rho)}}(y_{j^*})$, let E_{ij} be the random variable which is 1 if $(y_i, y_j) \in G_{\text{corr}(\rho)}$. Then Lemma 3.4 guarantees $\mathbb{P}(E_{ij} = 1) \geq 1 - 24s^3/r$. The size of the set being checked in Procedure 1 is equal to $\sum_t E_{S_t}$. Hoeffding's inequality guarantees that with probability at least $1 - 2\exp(-2\overline{n}\gamma^2)$

$$\left|\frac{1}{\overline{n}}\sum_{t=1}^{\overline{n}}(E_{S_t}-\mathbb{P}(E_{S_t}=1))\right|\leq\gamma.$$

Combining with the lower bound on $\mathbb{P}(E_{ij}=1)$, we obtain that with probability at least $1-2\exp(-2\overline{n}\gamma^2)$,

$$\sum_t E_{S_t}\geq\overline{n}\left(1-24\frac{s^3}{r}\right)-\overline{n}\gamma. \tag{22}$$

Using $\gamma\leq 1/64$, we see that this quantity is at least $62\overline{n}/64$ under the conditions of the lemma, which means that Procedure 1 returns 1.

Now let us consider the case when Uniq-intersect$(y_{i^*},y_{j^*})=0$. Defining E_{ij} the same way as above, we see that by Lemma 3.3, $\mathbb{P}(E_{ij}=1)\leq 15/16$. Then, a similar application of Hoeffding's inequality yields this time

$$\sum_t E_{S_t}\leq\frac{\overline{n}}{16}+\overline{n}\gamma, \tag{23}$$

which is at most $61\overline{n}/64$ for $\gamma\leq 1/64$. Hence Procedure 1 returns 0 in this case.

A.4 Proof of Proposition 3.2

We now prove Proposition 3.2. We need a couple of auxilliary results for the proof. We first restate a theorem from [35], which we will heavily use in the sequel.

Theorem A.1 (Restatement of Theorem 5.44 from [35])**.** *Consider a $d\times n$ matrix W where each column w_i of W is an independent random vector with covariance matrix Σ. Suppose further that $\|w_i\|_2\leq\sqrt{u}$ a.s. for all i. Then for any $t\geq 0$, the following inequality holds with probability at least $1-d\exp\left(-ct^2\right)$:*

$$\left\|\frac{1}{n}WW^T-\Sigma\right\|_2\leq\max\left(\|\Sigma\|_2^{1/2}\,\delta,\delta^2\right)\ \text{ where }\delta=t\sqrt{\frac{u}{n}}.$$

Here $c>0$ is an absolute numerical constant. In particular, this inequality yields:

$$\|W\|_2\leq\|\Sigma\|_2^{\frac{1}{2}}\sqrt{n}+t\sqrt{u}.$$

In order to bound the errors made in Algorithm 1, we need some additional notation and auxilliary results. For now, let us consider a fixed pair of anchor samples y_{i^*} and y_{j^*} such that Uniq-intersect(y_{i^*},y_{j^*}) is satisfied, and wlog, let $\mathcal{N}_B(y_{i^*})\cap\mathcal{N}_B(y_{j^*})=\{a_1\}$. We define the following sets of interest

$$\widehat{S}=\mathcal{N}_{\text{corr}}(y_{i^*})\cap\mathcal{N}_{\text{corr}}(y_{j^*}),$$
$$S=\{y_i\in\widehat{S}:\mathcal{N}_B(y_i)\cap\mathcal{N}_B(y_{i^*})=\mathcal{N}_B(y_i)\cap\mathcal{N}_B(y_{j^*})=\{a_1\}\},\text{ and} \tag{24}$$
$$\widetilde{S}=\widehat{S}\setminus S. \tag{25}$$

For the purposes of understanding the errors in Algorithm 1, it would be helpful to decompose each vector $y_i \in S$ as

$$\breve{y}_i := y_i - x_{1i} a_1, \tag{26}$$

and accordingly define $\breve{Y}_S$ to be the $d \times |S|$ matrix of all such vectors in S. Intuitively, if all the vectors $\breve{y}$ were 0, then Algorithm 1 can recover a_1 via SVD in a relatively straightforward manner. We start by controlling the norm of the vectors y_i and $\breve{y}_i$.

Lemma A.1. *Under the model 1 and given assumptions 3, 5 and 6 we have for all $i = 1, 2, \ldots, n$*

$$\|y_i\|_2 \leq \sqrt{2s} M \quad \textit{and} \quad \|\breve{y}_i\|_2 \leq 2M\sqrt{s}.$$

Proof:

The proof is relatively straightforward consequence of our model and the assumptions. The model allows us to write

$$\begin{aligned}
\|y_i\|_2^2 = \langle y_i, y_i \rangle &= \sum_{a_p, a_q \in \mathcal{N}_B(y_i)} x_{pi} x_{qi} \langle a_p, a_q \rangle \\
&\leq \sum_{a_p, a_q \in \mathcal{N}_B(y_i)} |x_{pi} x_{qi}| |\langle a_p, a_q \rangle| \\
&= \sum_{a_p \in \mathcal{N}_B(y_i)} x_{pi}^2 \|a_p\|_2^2 + \sum_{a_p \neq a_q \in \mathcal{N}_B(y_i)} |x_{pi} x_{qi}| |\langle a_p, a_q \rangle| \\
&\leq M^2 \left(s + s^2 \frac{\mu_0}{\sqrt{d}} \right) \\
&\leq M^2 \left(s + \frac{1}{2} \right) \leq \frac{3sM^2}{2}.
\end{aligned}$$

Finally, by triangle inequality we further have that $\|\breve{y}_i\|_2 \leq \|y_i\|_2 + M$.

□

Given this result, we would next like to control the amount of contribution the $\breve{y}_i$ directions can have in the SVD step of Algorithm 1. Our next result shows that while these vectors are not zero, their random support along with the incoherence of our dictionary elements ensures that these vectors are not strongly aligned with any one direction. We do so by bounding the spectral norm of the matrix $\breve{Y}_S$.

Lemma A.2. *With the vectors $\breve{y}_i$ defined in Equation 26, we have the following bound with probability greater than $1 - d \exp\left(-c\alpha^2 |S|\right)$ for any $\alpha > 0$*

$$\left\| \breve{Y}_S \right\|_2 \leq M \sqrt{s|S|} \left(\frac{\mu_1}{\sqrt{d}} + 2\alpha \right),$$

where c is a universal constant.

Proof:

In order to prove the lemma, we first calculate the spectral norm of the covariance matrix of $\breve{y}_i$ and then use Theorem A.1. Note that from Lemma A.1, we have $\|\breve{y}_i\|_2 \leq 2M\sqrt{s}$. We first bound the spectral norm of the covariance matrix of $\breve{y}_i \in S$ i.e., we bound $\left\|\mathbb{E}\left(\breve{y}_i\breve{y}_i^T\right)\right\|_2$. In order to do this, we first fix $w \in \mathbb{R}^d$ and calculate:

$$w^T\mathbb{E}\left[\breve{y}_i\breve{y}_i^T\right]w = \mathbb{E}\left[\left(w^T\breve{y}_i\right)^2\right] = \mathbb{E}\left[\left(w^TA\breve{x}_i\right)^2\right] = \mathbb{E}\left[\left(z^T\breve{x}_i\right)^2\right],$$

where we use the notation $z := A^Tw$ and $\breve{x}_i$ is the same as x_i but with x_{i1} set to 0. We further simplify as

$$\begin{aligned}
w^T\mathbb{E}\left[\breve{y}_i\breve{y}_i^T\right]w &\leq \mathbb{E}\left[\left(\sum_{p=1}^r z_p\breve{x}_{pi}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{p=1}^r z_p^2\breve{x}_{pi}^2\right] + \mathbb{E}\left[\sum_{p\neq q=1}^r z_pz_q\breve{x}_{pi}\breve{x}_{qi}\right] \\
&\leq \sum_{p=1}^r z_p^2\mathbb{E}\left[\breve{x}_{pi}^2\right] + \sum_{p\neq q=1}^r |z_pz_q||\mathbb{E}\left[\breve{x}_{pi}\breve{x}_{qi}\right]| \\
&\leq \sum_{p=1}^r z_p^2M^2\frac{s}{r} + 0,
\end{aligned}$$

where the last inequality uses the fact that the values of $\mathbb{E}[x_{pi}x_{qi}] = 0$, since both of them are independent mean zero random variables.

Then we can further simplify the upper bound to obtain

$$w^T\mathbb{E}\left[\breve{y}_i\breve{y}_i^T\right]w \leq \frac{sM^2}{r}\|z\|_2^2 \overset{(\zeta)}{\leq} \frac{sM^2}{r}\cdot\frac{\mu_1^2r}{d} = \frac{\mu_1^2M^2s}{d},$$

where (ζ) follows from Assumption $(A3)$, since

$$\|z\|_2 = \left\|A^Tw\right\|_2 \leq \left\|A^T\right\|_2\|w\|_2 = \left\|AA^T\right\|_2^{\frac{1}{2}}\|w\|_2 \leq \sqrt{\frac{\mu_1^2r}{d}}.$$

Recalling that w was an arbitrary unit vector, this immediately yields a spectral norm bound on the expected covariance

$$\left\|\mathbb{E}\left[\breve{y}_i\breve{y}_i^T\right]\right\|_2 \leq \frac{\mu_1^2M^2s}{d}.$$

We are now in a position to apply Theorem A.1 with the matrix $W = \breve{Y}_S$ of size $d \times |S|$, where $u = \left(2M\sqrt{s}\right)^2$ and $t = \alpha\sqrt{|S|}$ for some $\alpha > 0$. Doing so yields the inequality

$$\left\|\breve{Y}_S\right\|_2 \leq \sqrt{\frac{\mu_1^2 M^2 s}{d}} \cdot \sqrt{|S|} + \alpha\sqrt{|S|} \cdot 2M\sqrt{s}$$
$$\leq M\sqrt{s|S|}\left(\sqrt{\frac{\mu_1^2}{d}} + 2\alpha\right),$$

with probability greater than $1 - d\exp\left(-c\alpha^2|S|\right)$. □

Finally we are in a position to establish a bound on the accuracy of the SVD step in Algorithm 1. Having bounded the contribution from from the directions apart from a_1 in the previous lemma, we will now lower bound the contribution of the a_1 direction, which will ensure that the largest singular vector is close to a_1.

Proof of Proposition 3.2: Recall the definitions of the sets S and $\widetilde{S}$ (25). In order for a vector $y_i \in \widehat{S}$ to end up in $\widetilde{S}$, the event in Lemma 3.2 has to fail. Hence, if we define E_i to be the random variable which is 1 if $y_i \in \widetilde{S}$, then we have from Hoeffding's inequality

$$|\frac{1}{\widehat{n}}\sum_{i=1}^{\widehat{n}}(E_i - \mathbb{P}[E_i = 1])| \leq \sqrt{\frac{2\log(2/\delta)}{\widehat{n}}},$$

with probability at least $1 - \delta/2$. From Lemma 3.2 we know that $\mathbb{P}[E_i = 1] \leq s^3/r$ so that

$$|\widetilde{S}| \leq \frac{\widehat{n}s^3}{r} + \alpha\widehat{n}, \tag{27}$$

with probability at least $1 - \exp(-2\alpha^2\widehat{n})$. As a consequence, the size of S is at least

$$|S| \geq \widehat{n}(1 - s^3/r - \alpha) \geq 9\widehat{n}/10 \tag{28}$$

for $\alpha < 1/20$ by our assumption that $s^3 < r/384$.

In order to understand the singular vector $\hat{a}$, we now write the matrix $\widehat{L}$ as the sum of two matrices L and $\widetilde{L}$ as follows:

$$\widehat{L} = L + \widetilde{L}, \text{ where,}$$
$$L := \sum_{y_i \in S} y_i y_i^T \text{ and } \widetilde{L} := \sum_{y_i \in \widetilde{S}} y_i y_i^T.$$

Recalling our earlier notation $\breve{y}_i$ (26), we expand L as follows:

$$L = \sum_{y_i \in S} y_i y_i^T = \sum_{i:y_i \in S} x_{1i}^2 a_1 {a_1}^T + \sum_{i:y_i \in S} x_{1i}\left(a_1 \breve{y}_i^T + \breve{y}_i {a_1}^T\right) + \sum_{i:y_i \in S} \breve{y}_i \breve{y}_i^T$$

We wish to show that a_1 is close to the top singular vector of $\widehat{L}$. In order to show this, we bound the spectral norms of the following matrices: $\sum_{i:y_i \in S} x_{1i}\left(a_1 \breve{y}_i^T + \breve{y}_i {a_1}^T\right)$, $\sum_{i:y_i \in S} \breve{y}_i \breve{y}_i^T$ and $\widetilde{L}$.

Using Lemma A.2, we first obtain:

$$\left\|\sum_{i:y_i\in S} x_{1i}a_1\breve{y}_i^T\right\|_2 \le \|a_1\|_2 \left\|\breve{Y}_S\right\|_2 \|x_{S1}\|_2$$
$$\le M\sqrt{s|S|}\left(\sqrt{\frac{\mu_1^2}{d}}+2\alpha\right)\cdot M\sqrt{|S|}$$
$$= M^2 s|S|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{2\alpha}{\sqrt{s}}\right) \text{ and,} \tag{29}$$
$$\left\|\sum_{i:y_i\in S}\breve{y}_i\breve{y}_i^T\right\|_2 = \left\|\breve{Y}_S\breve{Y}_S^T\right\|_2 \le 2M^2 s|S|\left(\frac{\mu_1^2}{d}+4\alpha^2\right). \tag{30}$$

Finally, we have the following bound on the spectral norm of $\widetilde{L}$:

$$\left\|\widetilde{L}\right\|_2 = \left\|\sum_{y_i\in\widetilde{S}} y_i {y_i}^T\right\|_2 \le |\widetilde{S}|\,\|y_i\|_2^2 \le |\widetilde{S}|2sM^2. \tag{31}$$

Using (29), (30) and (31), we now prove the statement of the lemma. Denote $\theta = |\langle a_1, \widehat{a}\rangle|$ and $Z = \frac{1}{|S|}\sum_{i:y_i\in S} x_{1i}^2$. On one hand, we have:

$$\left\|\widehat{a}^T\widehat{L}\widehat{a}\right\|_2 \le \theta^2 Z|S| + 2\left\|\sum_{i:y_i\in S} x_{1i}a_1\breve{y}_i^T\right\|_2 + \left\|\sum_{i:y_i\in S}\breve{y}_i\breve{y}_i^T\right\|_2 + \left\|\widetilde{L}\right\|_2$$
$$\le \theta^2 Z|S| + 2M^2 s|S|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{2\alpha}{\sqrt{s}}\right) + 2M^2 s|S|\left(\frac{\mu_1^2}{d}+4\alpha^2\right) + |\widetilde{S}|2sM^2$$
$$\le M^2|S|\left[\frac{Z}{M^2}\theta^2 + 8s\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\alpha^2+\frac{\alpha}{\sqrt{s}}+\left(\frac{s^3}{r}+\alpha\right)\right)\right],$$

where the last step uses the bounds (27) and (28). On the other hand, we have

$$\left\|\widehat{a}^T\widehat{L}\widehat{a}\right\|_2 = \left\|\widehat{L}\right\|_2 \ge Z|S|\cdot\|a_1\|_2^2 - 2\left\|\sum_{i:y_i\in S} x_{1i}a_1\breve{y}_i^T\right\|_2 - \left\|\sum_{i:y_i\in S}\breve{y}_i\breve{y}_i^T\right\|_2 - \left\|\widetilde{L}\right\|_2$$
$$\ge Z|S| - 2M^2 s|S|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\alpha}{\sqrt{s}}\right) - 2M^2 s|S|\left(\frac{\mu_1^2}{d}+4\alpha^2\right) - |\widetilde{S}|2M^2 s$$
$$\ge M^2|S|\left[\frac{Z}{M^2} - 8s\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\alpha^2+\frac{\alpha}{\sqrt{s}}+\left(\frac{s^3}{r}+\alpha\right)\right)\right].$$

Using the above two inequalities, and the fact that $Z \ge m^2$, we obtain

$$\theta^2 \ge 1 - \frac{16sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\frac{s^3}{r}\right) - \frac{16sM^2}{m^2}\left(\alpha^2+\frac{\alpha}{\sqrt{s}}\right).$$

Now we observe that since $\|a_1\|_2 = \|\hat{a}\|_2 = 1$, we have

$$\|\hat{a} - a_1\|_2^2 = 2(1-\theta) \leq 2(1-\theta^2),$$

for $0 \leq \theta \leq 1$, which completes the proof.

□

A.5 Approximate recovery guarantee for Algorithm 1

Building on all our work so far, this section presents the main guarantee for Algorithm 1. So far, we have established that the sub-procedure in Algorithm 1 correctly detects good anchor pairs with high probability. Conditioned on this, Proposition 3.2 shows that we can recover the dictionary element in this intersection to a bounded error with high probability. The next theorem, which puts everything together shows that in an appropriate number of iterations, Algorithm 1 will approximately receover *all* the dictionary elements with high probability.

Lemma A.3 (Number of good anchor pairs)**.** *Suppose we have n examples. Then, we have:*

$$\mathbb{P}\left\{\cup_{l\in[r]}|\{(i,j) : \mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \{a_l\}\}| > \frac{ns}{8r}\right\} \geq 1 - r\exp\left(\frac{-ns}{64r}\right).$$

Proof: Fix $l \in [r]$. Define the set $S \subseteq [n]$ as follows:

$$S := \{i : a_l \in \mathcal{N}_B(y_i)\}.$$

Since for every $i \in [n]$, the probability of $i \in S$ is $\frac{s}{r}$, using standard Chernoff bounds, we see that:

$$\mathbb{P}\left[|S| < \frac{ns}{2r}\right] < \exp\left(\frac{-ns}{8r}\right). \tag{32}$$

Consider any two examples $y_i, y_j \in S$. Then,

$$\mathbb{P}\left[\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \{a_l\}\right] \geq 1 - \frac{s^2}{r}.$$

Dividing the set S into $\frac{|S|}{2}$ disjoint pairs and using Chernoff bounds, we see that

$$\mathbb{P}\left[|\{(i,j) : \mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \{a_l\}\}| < \frac{|S|}{4}\right] \leq \exp\left(\frac{-\left(1-\frac{s^2}{r}\right)|S|}{16}\right) \leq \exp\left(\frac{-|S|}{32}\right). \tag{33}$$

Using (32) and (33), we have:

$$\mathbb{P}\left[|\{(i,j) : \mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \{a_l\}\}| > \frac{ns}{8r}\right] \geq 1 - \exp\left(\frac{-ns}{64r}\right).$$

Using a union bound over different dictionary elements, we have:

$$\mathbb{P}\left[|\{(i,j)|\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_j) = \{a_l\}\}| > \frac{ns}{8r}\ \forall\, l \in [r]\right] \geq 1 - r\exp\left(\frac{-ns}{64r}\right).$$

□

Lemma A.4. *In each iteration of Algorithm 1, the size of the set $\widehat{S}$ satisfies:*

$$|\widehat{S}| \geq \frac{ns}{4r},$$

with probability greater than $1 - \exp\left(\frac{-ns}{16r}\right)$.

Proof: Since $(y_{i^*}, y_{j^*}) \in G_{\text{corr}(\rho)}$, from Lemma 3.1, we know that $\mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*}) \neq \emptyset$. Wlog let $a_1 \in \mathcal{N}_B(y_{i^*}) \cap \mathcal{N}_B(y_{j^*})$. Since each sample y_i has probability of at least

$$\frac{s}{r} \cdot \frac{\binom{r-2s+1}{s-1}}{\binom{r-1}{s-1}} \geq \frac{s}{r} \cdot \left(\frac{r-3s}{r-s}\right)^s \geq \frac{s}{r} \cdot \left(1 - \frac{2s}{r-s}\right)^s \geq \frac{s}{r} \cdot \left(1 - \frac{2s^2}{r-s}\right) \geq \frac{s}{2r},$$

of satisfying $\mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{i^*}) = \mathcal{N}_B(y_i) \cap \mathcal{N}_B(y_{j^*}) = \{a_1\}$, using Chernoff bounds, we have:

$$\mathbb{P}\left[|i : \text{Uniq-intersect}\,(y_i, y_{i^*})\,\&\,\text{Uniq-intersect}\,(y_i, y_{j^*})| < \frac{ns}{4r}\right] \leq \exp\left(\frac{-ns}{16r}\right).$$

Using Lemma 3.1 now finishes the proof. □

A.6 Auxiliary Results

Below, we establish that the incoherence assumption on the dictionary elements leads to a bound on the RIP constant.

Lemma A.5. *The $2s$-RIP constant of A, δ_{2s} satisfies $\delta_{2s} < \frac{2\mu_0 s}{\sqrt{d}}$.*

Proof: Consider a $2s$-sparse unit vector $w \in \mathbb{R}^r$ with $\text{Supp}(w) = S$. We have:

$$\begin{aligned}
\|Aw\|^2 = \left(\sum_{j \in S} w_j a_j\right)^2 &= \sum_j w_j^2 \|a_j\|^2 + \sum_{j,l \in S, j \neq l} w_j w_l \langle a_j, a_l \rangle \\
&\geq 1 - \sum_{j,l \in S, j \neq l} |w_j w_l| |\langle a_j, a_l \rangle| \\
&\geq 1 - \sum_{j,l \in S, j \neq l} |w_j w_l| \frac{\mu_0}{\sqrt{d}} \\
&\geq 1 - \frac{\mu_0}{\sqrt{d}} {\|w\|_1}^2 \\
&\geq 1 - \frac{\mu_0}{\sqrt{d}} 2s \cdot \|w\|^2 = 1 - \frac{2\mu_0 s}{\sqrt{d}}.
\end{aligned}$$

Similarly, we have:

$$\|Aw\|^2 \leq 1 + \frac{2\mu_0 s}{\sqrt{d}}.$$

This proves the lemma. □

Lemma A.6. *For $r > 2, c > 0$, let $0 \leq x \leq r/(2c+1)$. Then $(1 - cx/(r-x))^x \geq \exp(-cx^2/(r-x)) \geq 1 - \frac{2x^2}{r-x}$.*

Proof:

We start by observing that $x/(r-x)$ is an increasing function of x for $x < r$, so that $x < r/(2c+1)$ implies that $cx/(r-x) < 1/2$. Additionally, we have the following fact for any $\theta > 0$

$$1 - \theta \leq e^{-\theta} \leq 1 - \theta + \frac{\theta^2}{2}. \tag{34}$$

The first inequality is a consequence of the convexity of $e^{-\theta}$ while the second one follows since the second derivative of $e^{-\theta}$ is at most 1 when $\theta > 0$. Since we have $x/(r-x) \leq 1/2$, it is easy to see that

$$1 - \frac{cx}{r-x} \geq 1 - 2\frac{cx}{r-x} + 2\frac{c^2x^2}{(r-x)^2}.$$

Now applying the inequalities (34) with $\theta = 2cx/(r-x)$, we obtain

$$\begin{aligned}\left(1 - \frac{cx}{r-x}\right)^x &\geq \left(1 - 2\frac{cx}{r-x} + 2\frac{cx^2}{(r-x)^2}\right)^x \\ &\geq (\exp(-2cx/(r-x)))^x = \exp(-2cx^2/(r-x) \\ &\geq 1 - \frac{2cx^2}{r-x},\end{aligned}$$

where the second inequality follows from again using (34), this time with $\theta = 2cx^2/(r-x)$. □

Lemma A.7. *We have:*

$$\mathbb{E}\left[XX^T\right] = \frac{s}{r} I_{r\times r},$$

where $I_{r\times r}$ is the $r \times r$ identity matrix.

Proof: Let $\Sigma := \mathbb{E}\left[XX^T\right]$. We will first calculate the diagonal elements of Σ:

$$\Sigma_{jj} = \mathbb{E}\left[x_{ji}^2\right] = \frac{s}{r}.$$

On the other hand, any off diagonal element can be calculated as follows:

$$\Sigma_{jk} = \mathbb{E}\left[x_{ji}x_{ki}\right] = \mathbb{E}\left[x_{ji}\right]\mathbb{E}\left[x_{ki}\right] = 0.$$

This proves the lemma. □

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[2] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. A Spectral Algorithm for Latent Dirichlet Allocation. In *Proc. of Neural Information Processing (NIPS)*, Dec. 2012.

[3] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Conference on Learning Theory (COLT)*, June 2013.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor Methods for Learning Latent Variable Models. *ArXiv:1210.7559*, Oct. 2012.

[5] A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity. *ArXiv 1308.2853*, Aug. 2013.

[6] A. Anandkumar, D. Hsu, and A. J. S. M. Kakade. Learning Topic Models and Latent Bayesian Networks Under Expansion Constraints. *Preprint. ArXiv:1209.5350*, Sept. 2012.

[7] S. Arora, R. Ge, Y. Halpern, D. M. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. *ArXiv 1212.4777*, 2012.

[8] S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *ArXiv e-prints*, Aug. 2013.

[9] S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ica with unknown gaussian noise, and implications for gaussian mixtures and autoencoders. *arXiv preprint arXiv:1206.5349*, 2012.

[10] S. Arora, R. Ge, S. Sachdeva, and G. Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012.

[11] M.-F. Balcan, C. Borgs, M. Braverman, J. T. Chayes, and S.-H. Teng. I like her more than you: Self-determined communities. *CoRR*, abs/1201.4899, 2012.

[12] Y. Bengio, A. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

[13] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589 – 592, 2008.

[14] L. De Lathauwer, J. Castaing, and J.-F. Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007.

[15] M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[16] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[17] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[18] Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ_1 minimization for dictionary learning. *arXiv preprint arXiv:1101:5672*, 2011. Preprint, URL:http://arxiv.org/abs/1101.5672.

[19] N. Goyal, S. Vempala, and Y. Xiao. Fourier pca. *ArXiv 1306.5825*, 2013.

[20] C. J. Hillar and F. T. Sommer. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *arXiv preprint arXiv:1106.3616*, 2011.

[21] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[22] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *arXiv preprint arXiv:1104.4803*, 2011.

[23] 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.

[24] R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, 2010.

[25] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

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

[27] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. *arXiv preprint arXiv:1209.0738*, 2012.

[28] F. McSherry. Spectral partitioning of random graphs. In *FOCS*, 2001.

[29] N. Mehta and A. G. Gray. Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 36–44, 2013.

[30] 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.

[31] D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proc. of Conf. on Learning Theory*, 2012.

[32] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias. Learning stable multilevel dictionaries for sparse representation of images. *ArXiv 1303.0448*, 2013.

[33] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

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

[35] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.