JMLR: Workshop and Conference Proceedings vol 35:1–15, 2014

Learning Sparsely Used Overcomplete Dictionaries

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research, New York NY USA

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Dept of EECS, UC Irvine, Irvine, CA USA

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research, Bangalore, India

Praneeth Netrapalli PRANEETHN@UTEXAS.EDU
Dept of ECE, UT Austin, Austin, TX USA

Rashish Tandon RASHISH@CS.UTEXAS.EDU
Dept of CS, UT Austin, Austin, TX USA

Abstract

We consider the problem of learning sparsely used overcomplete dictionaries, where each observation is a sparse combination of elements from an unknown overcomplete dictionary. We establish exact recovery when the dictionary elements are mutually incoherent. Our method consists of a clustering-based initialization step, which provides an approximate estimate of the true dictionary with guaranteed accuracy. This estimate is then refined via an iterative algorithm with the following alternating steps: 1) estimation of the dictionary coefficients for each observation through ℓ_1 minimization, given the dictionary estimate, and 2) estimation of the dictionary elements through least squares, given the coefficient estimates. We establish that, under a set of sufficient conditions, our method converges at a linear rate to the true dictionary as well as the true coefficients for each observation.

Keywords: Dictionary learning, sparse coding, overcomplete dictionaries, alternating minimization, lasso.

1. Introduction

The problem of dictionary learning can be stated as follows: given observations $Y \in \mathbb{R}^{d\times n}$, the task is to decompose it as

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

A^* is referred to as the *dictionary* matrix and X^* is the *coefficient* matrix, and both are unknown. We consider the challenging case when the number of dictionary elements $r \geq d$. Without further constraints, the solution to (1) is not unique. A popular framework is to assume that the coefficient matrix X^* is sparse, and that each observation $Y_i \in \mathbb{R}^d$ is a sparse combination of the dictionary elements (i.e. columns of the dictionary matrix). This problem is known as *sparse coding* and it has been argued that sparse coding can provide a succinct representation of the observed data, given only unlabeled samples, see (Olshausen and Field, 1997; Lee et al., 2006). Through this lens of unsupervised learning, dictionary learning has recently received increased attention from the learning community, e.g. (Mehta and Gray, 2013; Balasubramanian et al., 2013; Maurer et al., 2013).

Although several methods exist for sparse coding, most of them lack guarantees. Spielman et al. (2012) recently provided a method for guaranteed recovery when the dictionary matrix $A^* \in \mathbb{R}^{d\times r}$ is a basis. This implies that the number of dictionary elements $r \leq d$, where d is the observed dimension. However, in most settings, the dictionary is *overcomplete* $(r \gg d)$ as overcomplete representations can provide greater flexibility in modeling as well as better robustness to noise, see (Lewicki and Sejnowski, 2000; Bengio et al., 2012; Elad, 2010) for details. In this paper, we establish *exact* recovery of sparsely used overcomplete dictionaries.

Summary of Results: We establish exact recovery of the dictionary and the coefficient matrix under a set of natural conditions, viz., the dictionary satisfies a mutual incoherence condition, each observation consists of s dictionary elements, and the coefficients are generated from a probabilistic model with a uniformly random sparsity pattern. Our method for dictionary learning that consists of two phases. The initialization phase is a clustering-based procedure for recovering the dictionary with bounded error. In particular, we establish that the recovery error of the initialization procedure is bounded by a constant (dependent only on s), as long as the sparsity satisfies $s = O\left(d^{1/4}, r^{1/4}\right)$. The number of samples needed for this initialization procedure scales as $n = O\left(r(\log r + \log d)\right)$.

The second stage of our method consists of an alternating minimization scheme which outputs successively improved estimates of the coefficients and the dictionary through lasso and least-squares steps respectively. We establish convergence to the global optimum when the alternating minimization is initialized with an approximate dictionary, with an error of at most $O\left(1/s^2\right)$. Further, when $s = O\left(d^{1/6}\right)$ and the number of samples satisfies $n = O\left(r^2\right)$, we establish a linear rate of convergence for the alternating minimization procedure to the true dictionary.

Thus, taken together, the two stages of our method yield exact recovery of both the dictionary and the coefficient matrix, as long as the sparsity level satisfies $s = O\left(d^{1/9}, r^{1/8}\right)$, and the number of samples is $n = O\left(r^2\right)$. We believe that this is the first exact recovery result for dictionary learning in the overcomplete setting. Note that our alternating minimization guarantee is independent of the initialization procedure, and it is entirely possible to use other initialization procedures for the alternating minimization algorithm. Indeed, the recent and concurrent work of Arora et al. (2013) can be seen as another initialization procedure for alternating minimization, and we discuss these implications in related work below.

Finally, we present numerical simulations confirming the linear rate of convergence for the alternating minimization procedure, and thereby demonstrating the extent of gains beyond the initialization step. We also empirically test the recovery performance of the procedure, and find that it succeeds with $n = O\left(r\right)$ samples, and hence suggesting room for tightening our analysis in future work.

Related Work: There have been many works on dictionary learning both from a theoretical and empirical viewpoint. Hillar and Sommer (2011) consider conditions for identifiability of sparse coding. However, the number of samples required to establish identifiability is exponential in r for the general case. Most closely related to our work, Spielman et al. (2012) provide exact recovery results for an ℓ_1 based method, but they focus on the *undercomplete* setting, where $r \leq d$. We consider the overcomplete setting where $r > d$.

There exist many heuristics for dictionary learning, which work well in practice in many contexts, but lack theoretical guarantees. For instance, Lee et al. (2006) propose an iterative ℓ_1 and ℓ_2 optimization procedure similar to the the method of optimal directions (Engan et al., 1999). Another popular method is 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 establish local optimality of the true solution (A^*, X^*) for certain non-convex programs in the noiseless (Jenatton et al., 2010; Geng et al., 2011) as well as noisy (Jenatton et al., 2012; Gribonval and Schnass, 2010), but do not prescribe algorithms which can reach the true solution (A^*, X^*). Recent works (Vainsencher et al., 2011; Mehta and Gray, 2013; Maurer et al., 2012; Thiagarajan et al., 2013) provide generalization bounds for predictive sparse coding, without computational considerations.

Finally, our results are closely related to the recent work of Arora et al. (2013), carried out independently and concurrently with our work. They provide an approximate recovery result followed by an alternating minimization procedure. A key distinction between our alternating minimization procedure as compared to theirs is that we use the *same* samples in each iteration, while they require fresh samples for each iteration of alternating minimization. This enables us to obtain *exact* recovery of the dictionary when $n = \Omega(r^2)$, whereas the error in their method is only below $\exp\left(-O\left(n/r^2\right)\right)$. Our algorithm is also robust in the sense that we do not expect to recover the complete support in the first iteration – we gradually recover more and more elements of the support as our dictionary estimate gets better. On the other hand, Arora et al. (2013) employ different probabilistic arguments allowing them to handle larger levels of sparsity, in terms of r and d. Overall, we believe the techniques of two papers can be combined to have a better sample complexity with respect to both sparsity s and the desired accuracy parameter ϵ.

The remainder of the paper is organized as follows. We present our algorithms next, followed by our assumptions and the recovery results. We provide proof sketches in Section 3, and simulation results are described in Section 4. Detailed proofs can be found in the longer versions of the paper, with the initialization technique in Agarwal et al. (2013b) and the alternating minimization analysis in Agarwal et al. (2013a).

2. Algorithm

Notation: Let $[n] := \{1, 2, \ldots, n\}$. For a vector v or a matrix W, we will use the shorthand $\mathrm{Supp}(v)$ and $\mathrm{Supp}(W)$ to denote the set of non-zero entries of v and W respectively. Let $\|w\|$ denote the ℓ_2 norm of vector w, and similarly for a matrix W, $\|W\|$ denotes its spectral norm. For a matrix X, X^i, X_i and X^i_j denote the i^{th} row, i^{th} column and $(i, j)^{\text{th}}$ element of X respectively. For a graph $G = (V, E)$, let $\mathcal{N}_G(i)$ denote set of neighbors for node i in G.

2.1. Initial Estimate of Dictionary Matrix

The first step is to obtain an initial estimate $\widehat{A}$ of the dictionary elements, and is given in Algorithm 1. The estimate $\widehat{A}$ is then employed in alternating steps to estimate the coefficient matrix and re-estimate the dictionary matrix respectively.

Given samples Y, we first construct the 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 threshold $\rho > 0$ (Figure 1 shows an example of a typical correlation graph under our assumptions). We then determine a good subset of samples via a *clustering* procedure on the graph as follows: we first randomly sample an edge $(Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}$ and consider the intersection of the neighborhoods of Y_{i^*} and Y_{j^*}, denoted by $\widehat{S}$. We further employ UniqueIntersection routine in Procedure 1 to determine if $\widehat{S}$ is a "good set" for estimating a dictionary element. This is done by ensuring that the set

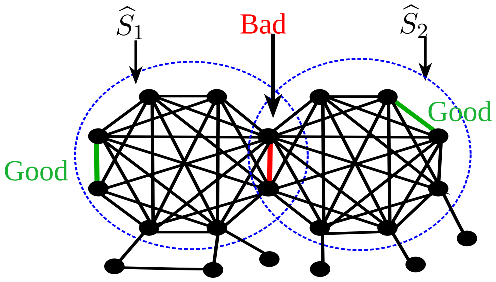


Figure 1: Sample correlation graph G_{corr} with nodes $\{Y_k\}$ and edge (Y_i, Y_j) s.t. $|\langle Y_i, Y_j\rangle| > \rho$. $\widehat{S}_1, \widehat{S}_2$ are the sets returned as true from UniqueIntersection procedure. The edges labeled "good" above refers to good anchor pairs which satisfy unique intersection in Algorithm 1,while the bad anchor pair does not satisfy the unique intersection. Good anchor pairs lead to formation of sets $\widehat{S}_1$ and $\widehat{S}_2$.

$\widehat{S}$ has a sufficient number of edges[1] in the correlation graph. For instance, the procedure will return true when evaluated on the green edges labeled *Good*, but false on the red edges labeled *Bad*. Once $\widehat{S}$ is determined to be a good set, we then proceed by computing the empirical covariance matrix $\widehat{L}$ of the 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.

At a high level, the above procedure aims to find large cliques in the correlation graph. For instance, in Figure 1, the sets $\widehat{S}_1, \widehat{S}_2$ are the sets which are returned as true by the UniqueIntersection Procedure, when the node pairs labeled as "good" in the figure are used as anchor samples Y_{i^*} and Y_{j^*}. On the other hand, note that a bad anchor pair which sits at the overlap of multiple cliques is not returned as true by the UniqueIntersection Procedure. Thus, this procedure yields subsets of samples which correspond to large cliques in the correlation graph. Once, such a subset is found, Algorithm 1 computes SVD over the samples in such sets. As our proofs will demonstrate, any such clique $\widehat{S}_i$ involves samples that all contain a *unique* dictionary element in common, which can then be recovered approximately by the subsequent SVD step.

2.2. Alternating Minimization

Once an initial estimate of the dictionary is obtained, we alternate between two procedures, viz., a sparse recovery step for estimating the coefficients given a dictionary, and a least squares step for a dictionary given the estimates of the coefficients (details are presented in Algorithm 2).

The sparse recovery step of Algorithm 2 is based on ℓ_1-regularization, followed by thresholding. The thresholding is required for us to guarantee that the support set of our coefficient estimate $X(t)$ is a *subset* of the true support with high probability. Once we have an estimate of the coefficients, the dictionary is re-estimated through least squares. The overall algorithmic scheme is popular for dictionary learning, and there are a number of variants of the basic method. For instance, the ℓ_1-regularized problem in step 3 can also be replaced by other robust sparse recovery

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

```
**Algorithm 1** InitDictionaryLearn($Y, \epsilon_{\text{dict}}, \rho$): Initial step for estimating dictionary elements.
**Input:** Samples $Y = [Y_1 | \dots | Y_n]$. Correlation threshold $\rho$. Desired separation parameter $\epsilon_{\text{dict}}$
    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**
            $\widehat{L} \leftarrow \sum_{Y_i \in \widehat{S}} Y_i Y_i^\top$ 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
```

procedures such as OMP (Tropp and Gilbert, 2007) or GraDeS (Garg and Khandekar, 2009). More generally the exact lasso and least-squares steps may be replaced with other optimization methods for computational efficiency, e.g. (Jenatton et al., 2010).

3. Guarantees

In this section, we provide our exact recovery result and also clearly specify all the required assumptions on A^* and X^*. We then provide guarantees for each of the individual steps (initialization step and alternating minimization steps) in Section 3.2 and Section 3.3, respectively. We provide a brief sketch of our proof for each of the steps in Section 3.4.

3.1. Assumptions and exact recovery result

We start by formally describing the assumptions needed for the main recovery result of this paper.

Assumptions on the dictionary:

Algorithm 2 AltMinDict$(Y, A(0), \epsilon_0)$: Alternating minimization for dictionary learning

```
Input: Samples $Y$, initial dictionary estimate $A(0)$, accuracy sequence $\epsilon_t$ and sparsity level $s$.
    Thresholding function $\mathcal{T}_\rho(a) = a$ if $|a| > \rho$ and 0 o.w.
for iterations $t = 0, 1, 2, \ldots, T-1$ do
    for samples $i = 1, 2, \ldots, n$ do
        $X(t+1)_i = \arg\min_{x \in \mathbb{R}^r} \|x\|_1$
        such that, $\|Y_i - A(t)x\|_2 \le \epsilon_t$.
    end for
    Threshold: $X(t+1) = \mathcal{T}_{9s\epsilon_t}(X(t+1))$.
    Estimate $A(t+1) = YX(t+1)^+$
    Normalize: $A(t+1)_i = \frac{A(t+1)_i}{\|A(t+1)_i\|_2}$
end for
Output: $A(T)$
```

$(A1)$ **Mutual Incoherence:** Without loss of generality, assume that all the elements are normalized: $\|A_i^*\| = 1$, for $i \in [r]$. We assume pairwise incoherence condition on the dictionary elements, i.e., for some $\mu_0 > 0$, we have $|\langle A_i^*, A_j^* \rangle| < \frac{\mu_0}{\sqrt{d}}$ for all $i, j \in [r]$.

$(A2)$ **Bound on the Spectral Norm:** The dictionary matrix has bounded spectral norm, i.e., for some $\mu_1 > 0$, we have $\|A^*\| < \mu_1 \sqrt{\frac{r}{d}}$.

Assumptions on the coefficients:

$(B1)$ **Non-zero Entries in Coefficient Matrix:** We assume that the non-zero entries of X^* are drawn i.i.d. from a zero-mean unit-variance distribution, and satisfy the following a.s.: $m \le |X^{*i}_{\ j}| \le M, \forall i, j$.

$(B2)$ **Sparse Coefficient Matrix:** The columns of coefficient matrix have s non-zero entries which are selected uniformly at random from the set of all s-sized subsets of $[r]$, i.e. $|\operatorname{Supp}(X_i^*)| = s, \forall\, i \in [n]$.We require s to satisfy

$$s < c_1 \min\left(\frac{m}{M} \frac{d^{1/4}}{\sqrt{\mu_0}}, \left(\frac{d}{\mu_1^2} \frac{m^4}{M^4} \right)^{1/9}, r^{1/8} \left(\frac{m}{M} \right)^{1/4} \right),$$

for some universal constant $c_1 > 0$. Constants m, M are as specified above.

Assumption $(A1)$ on normalization of dictionary elements is without loss of generality since we can always rescale the dictionary elements and the corresponding coefficients and obtain the same observations. However, the incoherence assumption is crucial in establishing our guarantees. In particular, incoherence also leads to a bound on the restricted isometry property (RIP) constant (Rauhut, 2010). The assumption $(A2)$ provides a bound on the spectral norm of A^*. Note that the incoherence and spectral assumptions are satisfied with high probability (w.h.p.) when the dictionary elements are randomly drawn from a mean-zero sub-gaussian distribution.

Assumption $(B1)$ imposes lower and upper bounds on the non-zero entries of X^*. We use the lower bound assumption on $X^*(i, j)$ for simplicity of exposition, as explained in Section 3.4, we can remove this assumption as the thresholding coefficient in Algorithm 2 decreases with each iteration. Assumption$(B2)$ on sparsity in the coefficient matrix is crucial for identifiability of the dictionary learning problem.

We now give the main result of this paper.

Theorem 1 (Exact recovery) *Suppose assumptions $(A1) - (A2)$ and $(B1) - (B2)$ are satisfied. Then there exists a universal constant c_3 such that, if*

1. **Sample Complexity:** $n \geq c_3 \, r^2 \frac{M^2}{m^2} \log \frac{2r}{\delta}$,
2. **Choice of Parameters for Initial Dictionary Estimation:** *inputs ρ and ϵ_{dict} to Algorithm 1 are chosen such that*
$$\rho = \frac{m^2}{2} - \frac{s^2 M^2 \mu_0}{\sqrt{d}} > 0, \quad \text{and} \quad \frac{1}{2}\left(\frac{1}{2592 s^2}\right)^2 < \epsilon_{\text{dict}}^2 < \frac{1}{4},$$
3. **Choice of Parameters for Alternating Minimization:** *Algorithm 2 uses a sequence of accuracy parameters $\epsilon_0 = 1/2592 s^2$ and*
$$\epsilon_{t+1} = \frac{25050 \mu_1 s^3}{\sqrt{d}} \epsilon_t \leq \frac{\epsilon_t}{2}. \tag{2}$$

Then the alternating minimization procedure (Algorithm 2) when seeded with Algorithm 1, outputs $A(t)$ at the t-th step ($t \geq 1$) that satisfies the following with probability at least $1 - 2\delta - 2n^2\delta$:

$$\min_{z \in \{-1,1\}} \|z A_i(t) - A_i^*\|_2 \leq \sqrt{2}\epsilon_t, \ \forall 1 \leq i \leq r,$$

where ϵ_t is as given in hypothesis (3) above. In particular, after $T = O(\log(\frac{\epsilon_0}{\epsilon}))$ steps of Algorithm 2, we obtain:

$$\min_{z \in \{-1,1\}} \|z A_i(t) - A_i^*\|_2 \leq \epsilon, \ \forall 1 \leq i \leq r, \forall \epsilon > 0.$$

Remarks: Note the 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.

Note that Theorem 1 guarantees that we can recover the dictionary A^* to an arbitrary precision ϵ (based on the number of iterations T of Algorithm 2), given $n = O\left(r^2\right)$ samples. We contrast this with the results of Arora et al. (2013), who also provide recovery guarantees to an arbitrary accuracy ϵ, but only if the number of samples is allowed to increase as $O\left(r^2 \log \frac{1}{\epsilon}\right)$.

Establishing the above result requires two main ingredients, viz., guaranteeing an error bound for the initial dictionary estimation step, and proving a local convergence result for the alternating minimization step, and obtaining a bound on the *basin of attraction* for the solution consisting of the true dictionary and coefficient matrices. Below, we provide these individual results explicitly.

3.2. Guarantees for the Initialization Step

We now give the result for approximate recovery of the dictionary in the initialization step.

Theorem 2 (Approximate recovery of dictionary) *Suppose the output of Algorithm 1 is $A(0)$. Fix $\alpha \in (0, 1/20)$. Under assumptions $(A1) - (A2)$ and $(B1) - (B2)$, and if*

1. **Sample Complexity:** $n \geq c_3 \frac{r}{\alpha^2 s} \log \frac{d}{\delta}$, *for a large enough constant* c_3, *and* $n^2\delta < 1$,
2. **Choice of Parameters for Initial Dictionary Estimation:** *inputs* ρ *and* ϵ_{dict} *to Algorithm 1 are chosen such that*
$$\rho = \frac{m^2}{2} - \frac{s^2 M^2 \mu_0}{\sqrt{d}} > 0, \text{ and } \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right) < \epsilon_{\text{dict}}^2 < \frac{1}{4},$$

then, with probability greater than $1 - 2n^2\delta$, *there exists a permutation matrix* P *such that:*
$$\epsilon_A^2 := \max_{i\in[r]} \min_{z\in\{-1,+1\}} \|zA_i^* - (PA(0))_i\|_2^2 < 32s\frac{M^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right).$$

Remarks: We note that the error in Theorem 2 does not go down with the number of samples n, since it depends on geometric properties of the dictionary, that are determined by the dimension dependent factors such as s, r and d. However, the error probability does go down with the number of samples, since the sample correlation graph becomes an increasingly accurate representative of the population version.

For the approximate recovery of dictionary elements, it turns out that a less stringent requirement on the sparsity level and the sample complexity suffices. Specifically, we can replace assumption $(B2)$ with the weaker condition $s < c_1 \min\left(\frac{m}{M}\frac{d^{1/4}}{\sqrt{\mu_0}}, \frac{dm^4}{\mu_1^2 M^4}, r^{1/4}\sqrt{\frac{m}{M}}\right)$, which suffices for the error in Theorem 2 to be $o(1)$. The more stringent requirement on sparsity arises in Theorem 1 since we need the error from Theorem 2 to be at most $O\left(1/s^2\right)$ for the subsequent alternating minimization steps to succeed. Note that the initialization step also has a milder requirement on the number of samples, and does not need the condition $n = O\left(r^2\log(1/\delta)\right)$. Thus, we obtain a near linear sample complexity for our initialization method.

3.3. Guarantees for Alternating Minimization

We now prove a local convergence result for alternating minimization. We assume that we have access to a good initial estimate of the dictionary:

$(C1)$ **Initial dictionary with guaranteed error bound:** We assume that we have access to an initial dictionary estimate $A(0)$ such that
$$\widehat{\epsilon_0} := \max_{i\in[r]} \min_{z\in\{-1,+1\}} \|zA_i^* - A(0)_i\|_2 < \frac{1}{2592s^2}.$$

Theorem 3 (Local linear convergence) *Under assumptions* $(A1)$-$(A2)$, $(B1)$-$(B2)$ *and* $(C1)$, *if*

1. **Sample Complexity:** $n \geq c_3 \max\left(r^2, rM^2 s\right) \log \frac{2r}{\delta}$,
2. **Choice of Parameters for Alternating Minimization:** *Algorithm 2 uses a sequence of accuracy parameters* $\epsilon_0 = 1/2592s^2$ *and*
$$\epsilon_{t+1} = \frac{25050\mu_1 s^3}{\sqrt{d}}\epsilon_t.$$

Then, with probability at least $1 - 2\delta$ *the iterate* $A(t)$ *of Algorithm 2 satisfies for all* $t \geq 1$*:*
$$\min_{z\in\{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \sqrt{2}\epsilon_t, \ \forall\, 1 \leq i \leq r.$$

Remarks: The consequences of Theorem 3 are powerful combined with our Assumption $(B2)$ and the recurrence 2 (since $(B2)$ ensures that ϵ_t forms a decreasing sequence). In particular, it is implied that with high probability we obtain,

$$\min_{z\in\{-1,1\}} \|zA(t)_i - A^*{}_i\|_2 \leq \epsilon_0 2^{-t}.$$

Given the above bound, we need at most $O\left(\log_2 \frac{\epsilon_0}{\epsilon}\right)$ in order to ensure $\|zA(T)_i - A^*{}_i\|_2 \leq \epsilon$ for all the dictionary elements $i = 1, 2, \ldots, r$. In the convex optimization parlance, the result demonstrates a local linear convergence of Algorithm 2 to the globally optimal solution under an initialization condition. Another way of interpreting our result is that the global optimum has a *basin of attraction* of size $\Omega\left(1/s^2\right)$ for our alternating minimization procedure under these assumptions (since we require $\widehat{\epsilon_0} = O\left(1/s^2\right)$).

We note that Theorem 3 does not crucially rely on initialization specifically by the output of Algorithm 1, and admits any other initialization satisfying Assumption $(C1)$. In particular, some of the assumptions in $(B1) - (B2)$ are not essential for Theorem 3, but are only made for the overall result of Theorem 1. Indeed, it suffices to have a sparsity level satisfying $s < \frac{d^{1/6}}{c_2\mu_1^{1/3}}$ for a universal constant $c_2 > 0$ (without any dependence on r). The theorem also does not rely on lower bounded entries, and only needs $\|X^*\|_\infty \leq M$. We also recall that the lasso step in Algorithm 2 can be replaced with a different robust sparse recovery procedure, with qualitatively similar results.

As remarked earlier, the recent work of Arora et al. (2013) provides an alternative initialization strategy for our alternating minimization procedure. Indeed, under our sample complexity assumption, their OVERLAPPINGAVERAGE method provides a solution with $\widehat{\epsilon_0} = O\left(s/\sqrt{r}\right)$ assuming $s = O\left(\max(r^{2/5}, \sqrt{d})\right)$.

3.4. Overview of Proof

In this section, we first provide a proof for Theorem 1 using Theorems 2 and 3. We then outline the key steps in proving Theorems 2 and 3.

Proof of Theorem 1 In order to establish the theorem, we just need to verify that all the preconditions of Theorems 2 and 3 are satisfied. We start by checking the preconditions of Theorem 2, for which we need to specify a value of the constant α. We will choose $\alpha = cm^2/(s^{-9/2}M^2)$ for a small enough universal constant c. This imposes the requirement that $n \geq c_3 r/(\alpha^2 s)\log(d/\delta)$. Note that we have

$$\frac{r}{\alpha^2 s} = \frac{rs^8}{c^2}\frac{M^4}{m^4} \leq \frac{r}{c^2}\frac{M^2}{m^2} \leq \frac{r^2}{c^2}\frac{M^2}{m^2},$$

where the first equality follows from the setting of α and the first inequality comes from Assumption (B2) on the sparsity level. This establishes the sample complexity requirement in Theorem 2. Based on this setting of α, we observe that $(\alpha\sqrt{s} + \alpha^2 s)M^2/m^2 = O(s^{-4})$. Similarly, based on the assumption (B2), it can be verified that all the remaining terms in the error bound ϵ_A^2 of Theorem 2 are $O(s^{-4})$, yielding $\epsilon_A^2 = O(s^{-4})$.

Specifically, this ensures that Theorem 2 supplies a dictionary $A(0)$ satisfying Assumption (C1) in Theorem 3. It is easily checked that using Assumption (B2), $rM^2 s \leq r^2$, so that the sample complexity assumption in Theorem 3 is also met. Consequently the result of Theorem 3 will guarantee local linear convergence, establishing Theorem 1. As one final remark, it is also using Assumption (B2) that we can verify $25050\mu_1 s^3/\sqrt{d} \leq 1/2$, so that $\epsilon_{t+1} \leq \epsilon_t/2$, ensuring that we always reduce our error by a factor of 2. This completes the proof.

Analysis of initial dictionary estimation: The core intuitions for this step 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 2 and 1 respectively. B_{coeff} consists of dictionary elements $\{A_i^*\}$ on one side and the samples $\{Y_i\}$ on the other. There is an edge between Y_i and A_j^* iff ${X^*}_j^i \neq 0$, 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[2] which (pairwise) have only one dictionary element A_i^* in common, and denote such a set by $\mathcal{C}_i$ i.e. $\mathcal{C}_i := \{Y_k, k \in S : \mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l) = A_i^*,\ \forall\, k, l \in S\}$. Intuitively, the sets $\widehat{S}$ constructed in Algorithm 1 are our proxies for $\mathcal{C}_i$. Indeed, the first part of the proof is to demonstrate that for a random coefficient matrix X^*, adequately large cliques $\mathcal{C}_i$ exist in the graph B_{coeff}.

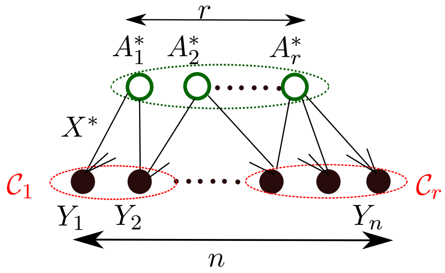


Figure 2: Bipartite graph B mapping dictionary elements $A_1^*, \ldots A_r^*$ to samples $Y_1, \ldots Y_n$. See text for definition of $\mathcal{C}_i$.

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 start with a proposition that demonstrates the correctness of Procedure 1 at identifying these cliques. We use the notation $\text{Uniq-intersect}(Y_i, Y_j)$ to denote that Y_i and Y_j have exactly one dictionary element in common.

Proposition 4 (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* $\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})$ *correctly with probability at least* $1 - 2\exp(-\gamma^2 \overline{n})$.

Given a large sample of elements with a unique dictionary element (say A_1^*) in common ($\widehat{S}$ in Algorithm 1), we next show that the subsequent SVD step recovers this dictionary element approximately. Intuitively this happens since each sample $Y_i \in \widehat{S}$ contains A_1^* with a coefficient at least m (in absolute value). Hence the covariance matrix $\widehat{L}$ has a larger component along A_1^* than other dictionary elements, which leads to approximate recovery via the top singular vector.

Proposition 5 (Accuracy of SVD) *Consider anchor samples* Y_{i^*} *and* Y_{j^*} *such that* $\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})$ *in Algorithm 1 is satisfied, and wlog, let* $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A_1^*\}$. *Recall the definition of* $\widehat{S}$

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

in Algorithm 1, and further define $\widehat{L} := \sum_{i\in\widehat{S}} Y_iY_i^\top$ 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 for any $0 < \alpha < 1/20$ we have:

$$\min_{z\in\{-1,1\}} \|z\widehat{a} - A_1^*\|_2^2 < \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right),$$

with probability at least $1 - d\exp\left(-c\alpha^2\widehat{n}\right)$.

Note the ambiguity in signs above, since SVD cannot recover the sign of the top singular vector. The proposition essentially yields the Theorem since the error bound above is identical to the bound in Theorem 2, and the result follows by lower bounding the number of samples $\widehat{n}$ in the above proposition.

Analysis of alternating minimization: Given an approximate estimate of the dictionary, we then establish a local convergence result for alternating minimization.

For ease of notation, let us consider just one iteration of Algorithm 2 and denote $X(t+1)$ as X, $A(t+1)$ as A and $A(t)$ as $\widetilde{A}$. Then we have the least-squares update

$$A - A^* = YX^+ - A^* = A^*X^*X^+ - A^*XX^+ = A^*\triangle XX^+,$$

where $\triangle X = X^* - X$. This means that we can understand the error in dictionary recovery by the error in the least squares operator $\triangle XX^+$. In particular, we can further expand the error in a column p as: $A_p - A^*{}_p = A^*{}_p(\triangle XX^+)_p^p + A^*{}_{\backslash p}(\triangle XX^+)_p^{\backslash p}$, where the notation $\backslash p$ represents the collection of all indices apart from p. Hence we see two sources of error in our dictionary estimate. The element $(\triangle XX^+)_p^p$ causes the rescaling of A_p relative to $A^*{}_p$. However, this is a minor issue since the renormalization would correct it.

More serious is the contribution from the off-diagonal terms $(\triangle XX^+)_p^{\backslash p}$, which corrupt our estimate A_p with other dictionary elements beyond $A^*{}_p$. Indeed, a crucial argument in our proof is controlling the contribution of these terms at an appropriately small level. In order to do that, we start by controlling the magnitude of $\triangle X$.

Lemma 6 (Error in sparse recovery) *Let $\triangle X := X(t) - X^*$. Assume that $2\mu_0 s/\sqrt{d} \leq 0.1$ and $\sqrt{s\epsilon_t} \leq 0.1$ Then, we have $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X^*)$ and the error bound $\|\triangle X\|_\infty \leq 9s\epsilon_t$.*

This lemma is very uesful in our error analysis, since we establish that any matrix W satisfying $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$ has a good bound on its spectral norm (even if the entries depend on A^*, X^*).

Lemma 7 *With probability at least $1 - r\exp\left(-\frac{Cn}{rs}\right)$, for every $r \times n$ matrix W s.t. $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$, we have $\|W\|_2 \leq 2\|W\|_\infty\sqrt{\frac{s^2n}{r}}$.*

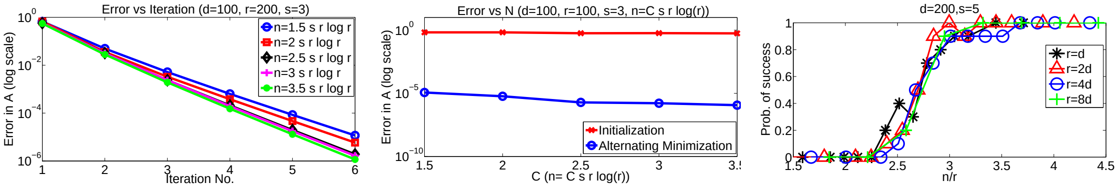
A particular consequence of this lemma is that it guarantees the invertibility of the matrix $XX^\top$, so that the pseudo-inverse X^+ is well-defined for subsequent least squares updates. Next, we present the most crucial step which is controlling the off-diagonal terms $(\triangle XX^+)_p^{\backslash p}$.

Lemma 8 (Off-diagonal error bound) *With probability at least $1-r\exp\left(-\frac{Cn}{rM^2s}\right)-r\exp\left(-\frac{Cn}{r^2}\right)$, we have uniformly for every $p \in [r]$ and every $\triangle X$ such that $\|\triangle X\|_\infty < \frac{1}{288s}$.*

$$\left\|\left(\triangle XX^+\right)_p^{\backslash p}\right\|_2 = \left\|\left(X^*X^+\right)_p^{\backslash p}\right\|_2 \leq \frac{1968s^2\|\triangle X\|_\infty}{\sqrt{r}}.$$

The lemma uses the earlier two lemmas along with some other auxilliary results. Given these lemmas, the proof of the main theorem follows with some algebra. Specifically, for any unit vector w such that $w \perp {A^*}_p$, we can bound the normalized inner product $\langle w, A_p \rangle / \|A_p\|_2$ which suffices to obtain the result of the theorem.

4. Experiments



(a) (b) (c)

Figure 3: (a): Average error after each step alternating minimization step of Algorithm 2 on log-scale. (b): Average error after the initialization procedure (Algorithm 1) and after 5 alternating minimization steps of Algorithm 2. (c): Sample complexity requirement of the alternating minimization algorithm. For ease of experiments, we initialize the dictionary using a random perturbation of the true dictionary rather than using Algorithm 1 which should in fact give better initial point with smaller error.

Alternating minimization/descent approaches have been widely used for dictionary learning and several existing works show effectiveness of these methods on real-world/synthetic datasets (Balasubramanian et al., 2013; Thiagarajan et al., 2013). Hence, instead of replicating those results, in this section we focus on illustrating the following three key properties of our algorithms via experiments in a controlled setting: a) Advantage of alternating minimization over one-shot initialization, b) linear convergence of alternating minimization, c) sample complexity of alternating minimization.

Data generation model: Each entry of the dictionary matrix A is chosen i.i.d. from $\mathcal{N}(0, 1)$. Note that, random Gaussian matrices are known to satisfy incoherence and the spectral norm bound (Vershynin, 2010). The support of each column of X was chosen independently and uniformly from the set of all s-subsets of $[r]$. Similarly, each non-zero element of X was chosen independently from the uniform distribution on $[-2, -1] \cup [1, 2]$. We use the GraDeS algorithm of Garg and Khandekar (2009) to solve the sparse recovery step, as it is faster than lasso. We measure error in the recovery of dictionary by $error(A) = \max_i \sqrt{1 - \frac{\langle A_i, A_i^* \rangle^2}{\|A_i\|_2^2 \|A_i^*\|_2^2}}$. The first two plots are for a typical run and the third plot averages over 10 runs. The implementation is in Matlab.

Linear convergence: In the first set of experiments, we fixed $d = 100$, $r = 200$ and measured error after each step of our algorithm for increasing values of n. Figure 3 (a) plots error observed after each iteration of alternating minimization; the first data point refers to the error incurred by the initialization method. As expected due to Theorem 3, we observe a geometric decay in the error.

One-shot vs iterative algorithm: It is conceivable that the initialization procedure of Algorithm 1 itself is sufficient to obtain an estimate of the dictionary upto reasonable accuracy. of Algorithm 2. Figure 3(b) shows that this is not the case. The figure plots the error in recovery vs the number of samples used for both Algorithm 1 and Algorithm 2. It is clear that the recovery error of the alternating minimization procedure is significantly smaller than that of the initialization procedure. For example, for $n = 2.5sr\log r$ with $s = 3, r = 200, d = 100$, initialization incurs error of $.56$ while alternating minimization incurs error of 10^{-6}. Note however that the recovery accuracy of the initialization procedure is non-trivial and also crucial to the success of alternating minimization- a random vector in $\mathbb{R}^d$ would give an error of $1 - \frac{1}{d} = 0.99$, where as the error after initialization procedure is ≈ 0.55.

Sample complexity: Finally, we study sample complexity requirement of the alternating minimization algorithm which is $n = O\left(r^2 \log r\right)$ according to Theorem 3, assuming good enough initialization. Figure 3(c) suggests that in fact only $O\left(r\right)$ samples are sufficient for success of alternating minimization. The figure plots the probability of success with respect to $\frac{n}{r}$ for various values of r. A trial is said to succeed if at the end of 25 iterations, the error is smaller than 10^{-6}. Since we focus only on the sample complexity of alternating minimization, we use a faster initialization procedure: we initialize the dictionary by randomly perturbing the true dictionary as $A(0) = A^* + Z$, where each element of Z is an $\mathcal{N}(0, 0.5)$ random variable. Figure 3 (c) shows that the success probability transitions at nearly the same value for various values of r, suggesting that the sample complexity of the alternating minimization procedure in this regime of $r = O\left(d\right)$ is just $O(r)$.

5. Conclusion

In this paper we present an exact recovery result for learning incoherent and overcomplete dictionaries with sparse coefficients. The first part of our result uses a novel initialization procedure, which uses a clustering-style algorithm to approximately recover the dictionary elements. The second step of our approach is an alternating minimization procedure which is quite widely used by practitioners for this problem already. We believe that our results are an important and timely advance in the understanding of this problem. There is an increasing interest on supervised and unsupervised feature learning methods in machine learning. However, we have an extremely rudimentary theoretical understanding of these problems as compared to standard classification of regression problems. A systematic understanding of dictionary learning and related models (both supervised and unsupervised) can help bridge this gap. Moreover, the applications of dictionary learning in other areas such as signal processing and coding make these results of broader interest beyond machine learning.

We believe that our work suggests several avenues for future research. We focus on the unsupervised setting in this paper, but extensions to supervised setting would be interesting for future work. Our theory also suggests room for strengthening the lasso step with further constraints on the global structure of the iterates $X(t)$, which might lead to better recovery properties with milder assumptions. Our simulations hint at the possibility of a better sample complexity, at least in certain regimes of parameters. Understanding these issues, as well as others such as noise robustness remain important questions for further research in this area.

Acknowledgments

Part of this work was done when A. Anandkumar, P. Netrapalli and R. Tandon were visiting Microsoft Research. We thank the anonymous reviewers for their helpful suggestions. 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.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013a.

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013b.

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

Krishnakumar Balasubramanian, Kai Yu, and Guy Lebanon. Smooth sparse coding via marginal regression for learning sparse representations. In *ICML*, 2013.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

Kjersti Engan, Sven Ole 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.

Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

Quan Geng, Huan Wang, and John 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.

Rémi Gribonval and Karin Schnass. Dictionary Identification - Sparse Matrix-Factorisation via ℓ_1-Minimisation. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Christopher J Hillar and Friedrich T Sommer. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *arXiv preprint arXiv:1106.3616*, 2011.

Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume 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.

Rodolphe Jenatton, Rémi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. Technical report, 2012. URL http://hal.inria.fr/hal-00737152.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. *arXiv preprint arXiv:1209.0738*, 2012.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013.

Nishant Mehta and Alexander 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.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Holger Rauhut. Compressive sensing, structured random matrices and recovery of functions in high dimensions. In *Oberwolfach Reports*, volume 7, pages 1990–1993, 2010.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proc. of Conf. on Learning Theory*, 2012.

Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. Learning stable multilevel dictionaries for sparse representation of images. *ArXiv 1303.0448*, 2013.

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

Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. The sample complexity of dictionary learning. *The Journal of Machine Learning Research*, 12:3259–3281, 2011.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.