

---

# Alternating minimization for dictionary learning with random initialization

---

Niladri S. Chatterji

UC Berkeley

niladri.chatterji@berkeley.edu

Peter L. Bartlett

UC Berkeley

peter@berkeley.edu

## Abstract

We present theoretical guarantees for an alternating minimization algorithm for the dictionary learning/sparse coding problem. The dictionary learning problem is to factorize vector samples  $y^1, y^2, \dots, y^n$  into an appropriate basis (dictionary)  $A^*$  and sparse vectors  $x^{1*}, \dots, x^{n*}$ . Our algorithm is a simple alternating minimization procedure that switches between  $\ell_1$  minimization and gradient descent in alternate steps. Dictionary learning and specifically alternating minimization algorithms for dictionary learning are well studied both theoretically and empirically. However, in contrast to previous theoretical analyses for this problem, we replace the condition on the operator norm (that is, the largest magnitude singular value) of the true underlying dictionary  $A^*$  with a condition on the matrix infinity norm (that is, the largest magnitude term). This not only allows us to get convergence rates for the error of the estimated dictionary measured in the matrix infinity norm, but also ensures that a random initialization will provably converge to the global optimum. Our guarantees are under a reasonable generative model that allows for dictionaries with growing operator norms, and can handle an arbitrary level of overcompleteness, while having sparsity that is information theoretically optimal. We also establish upper bounds on the sample complexity of our algorithm.

## 1 Introduction

In the problem of sparse coding/dictionary learning, given i.i.d. samples  $y^1, y^2, \dots, y^n \in \mathbb{R}^d$  produced from the generative model

$$y^i = A^* x^{i*} \tag{1}$$

for  $i \in \{1, 2, \dots, n\}$ , the goal is to recover a fixed dictionary  $A^* \in \mathbb{R}^{d \times r}$  and  $s$ -sparse vectors  $x^{i*} \in \mathbb{R}^r$ . (An  $s$ -sparse vector has no more than  $s$  non-zero entries.) In many problems of interest, the dictionary is often overcomplete, that is,  $r \geq d$ . This is believed to add flexibility in modeling and robustness. This model was first proposed in neuroscience as an energy minimization heuristic that reproduces features of the V1 portion of the visual cortex [28; 22]. It has also been an extremely successful approach to identifying low dimensional structure in high dimensional data; it is used extensively to find features in images, speech and video (see, for example, references in [13]).

Most formulations of dictionary learning tend to yield non-convex optimization problems. For example, note that if either  $x^{i*}$  or  $A^*$  were known, given  $y^i$ , this would just be a (matrix/sparse) regression problem. However, estimating both  $x^{i*}$  and  $A^*$  simultaneously leads to both computational as well as statistical complications. The heuristic of alternating minimization works well empirically for dictionary learning. At each step, first an estimate of the dictionary is held fixed while the sparse coefficients are estimated; next, using these sparse coefficients the dictionary is updated. Note that in *each step* the sub-problem has a convex formulation, and there is a range of efficient algorithms that can be used. This heuristic has been very successful empirically, and there has also been significant recent theoretical progress in understanding its performance, which we discuss next.

## 1.1 Related Work

A recent line of work theoretically analyzes local linear convergence rates for alternating minimization procedures applied to dictionary learning [1; 4]. Arora et al. [4] present a neurally plausible algorithm that recovers the dictionary exactly for sparsity up to  $s = \mathcal{O}(\sqrt{d}/(\mu \log(d)))$ , where  $\mu$  is the level of incoherence in the dictionary (which is a measure of the similarity of the columns; see Assumption A1 below). Agarwal et al. [1] analyze a least squares/ $\ell_1$  minimization scheme and show that it can tolerate sparsity up to  $s = \mathcal{O}(d^{1/6})$ . Both of these establish local linear convergence guarantees for the maximum column-wise distance. Exact recovery guarantees require a singular-value decomposition (SVD) or clustering based procedure to initialize their dictionary estimates (see also the previous work [5; 2]).

For the undercomplete case (when  $r \leq d$ ), Sun et al. [33] provide a Riemannian trust region method that can tolerate sparsity  $s = \mathcal{O}(d)$ , while earlier work by Spielman et al. [32] provides an algorithm that works in this setting for sparsity  $\mathcal{O}(\sqrt{d})$ .

Local and global optima of non-convex formulations for the problem have also been extensively studied in [36; 17; 18], among others. Apart from alternating minimization, other approaches (without theoretical convergence guarantees) for dictionary learning include K-SVD [3] and MOD [14]. There is also a nice formulation by Barak et al. [7], based on the sum-of-squares hierarchy. Recently, Hazan and Ma [20] provide guarantees for improper dictionary learning, where instead of learning a dictionary, they learn a comparable encoding via convex relaxations. Our work also adds to the recent literature on analyzing alternating minimization algorithms [21; 26; 27; 19; 6].

## 1.2 Contributions

Our main contribution is to present new conditions under which alternating minimization for dictionary learning converges at a linear rate to the *global optimum*. We impose a condition on the matrix infinity norm (largest magnitude entry) of the underlying dictionary. This allows dictionaries with operator norm growing with dimension  $(d, r)$ . The error rates are measured in the matrix infinity norm, which is sharper than the previous error rates in maximum column-wise error.

We also identify conditions under which a trivial random initialization of the dictionary works, as opposed to the more complex SVD and clustering procedures required in previous work. This is possible as our radius of convergence, again measured in the matrix infinity norm, is larger than that of previous results, which required the initial estimate to be close column-wise. Our results hold for a rather arbitrary level of overcompleteness,  $r = \mathcal{O}(\text{poly}(d))$ . We establish convergence results for sparsity level  $s = \mathcal{O}(\sqrt{d})$ , which is information theoretically optimal for incoherent dictionaries and improves the previously best known results in the overcomplete setting by a logarithmic factor. Our algorithm is simple, involving an  $\ell_1$ -minimization step followed by a gradient update for the dictionary.

A key step in our proofs is an analysis of a robust sparse estimator— $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector—under fixed (worst case) corruption in the dictionary. We prove that this estimator is minimax optimal in this setting, which might be of independent interest.

## 1.3 Organization

In Section 2, we present our alternating minimization algorithm and discuss the sparse regression estimator. In Section 3, we list the assumptions under which our algorithm converges and state the main convergence result. Finally, in Section 4, we prove convergence of our algorithm. We defer technical lemmas, analysis of the sparse regression estimator, and minimax analysis to the appendix.

## Notation

For a vector  $v \in \mathbb{R}^d$ ,  $v_i$  denotes the  $i^{\text{th}}$  component of the vector,  $\|v\|_p$  denotes the  $\ell_p$  norm,  $\text{supp}(v)$  denotes the support of a vector  $v$ , that is, the set of non-zero entries of the vector,  $\text{sgn}(v)$  denotes the sign of the vector  $v$ , that is, a vector with  $\text{sgn}(v)_i = 1$  if  $v_i > 0$ ,  $\text{sgn}(v)_i = -1$  if  $v_i < 0$  and  $\text{sgn}(v)_i = 0$  if  $v_i = 0$ . For a matrix  $W$ ,  $W_i$  denotes the  $i^{\text{th}}$  column,  $W_{ij}$  is the element in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column,  $\|W\|_{op}$  denotes the operator norm, and  $\|W\|_\infty$  denotes the maximum of the magnitudes of the elements of  $W$ . For a set  $J$ , we denote its cardinality by  $|J|$ . Throughout the paper,

---

**Algorithm 1:** Alternating Minimization for Dictionary Learning

---

**Input** : Step size  $\eta$ , samples  $\{y^k\}_{k=1}^n$ , initial estimate  $A^{(0)}$ , number of steps  $T$ , thresholds  $\{\tau^{(t)}\}_{t=1}^T$ , initial radius  $R^{(0)}$  and parameters  $\{\gamma^{(t)}\}_{t=1}^T, \{\lambda^{(t)}\}_{t=1}^T$  and  $\{\nu^{(t)}\}_{t=1}^T$ .

```

1 for  $t = 1, 2, \dots, T$  do
2   for  $k = 1, 2, \dots, n$  do
3      $w^{k,(t)} = MUS_{\gamma^{(t)}, \lambda^{(t)}, \nu^{(t)}}(y^k, A^{(t-1)}, R^{(t-1)})$ 
4     for  $l = 1, 2, 3, \dots, r$  do
5        $x_l^{k,(t)} = w_l^{k,(t)} \mathbb{I}(|w_l^{k,(t)}| > \tau^{(t)})$ , ( $x^{k,(t)}$  is the sparse estimate)
6     end
7   end
8   for  $i = 1, 2, \dots, d$  do
9     for  $j = 1, 2, \dots, r$  do
10       $A_{ij}^{(t)} = A_{ij}^{(t-1)} - \frac{\eta}{n} \sum_{k=1}^n \left[ \sum_{p=1}^r \left( A_{ip}^{(t-1)} x_p^{k,(t)} x_j^{k,(t)} - y_i^k x_j^{k,(t)} \right) \right]$ 
11    end
12  end
13   $R^{(t)} = \frac{3}{4} R^{(t-1)}$ .
14 end
```

---

we use  $C$  multiple times to denote global constants that are independent of the problem parameters and dimension. We denote the indicator function by  $\mathbb{I}(\cdot)$ .

## 2 Algorithm

Given an initial estimate of the dictionary  $A^{(0)}$  we alternate between an  $\ell_1$  minimization procedure (specifically the  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector— $MUS_{\gamma, \lambda, \nu}$  in the algorithm—followed by a thresholding step) and a gradient step, under sample  $\ell_2$  loss, to update the dictionary. We analyze this algorithm and demand linear convergence at a rate of  $3/4$ ; convergence analysis for other rates follows in the same vein with altered constants. In subsequent sections, we also establish conditions under which the initial estimate for the dictionary  $A^{(0)}$  can be chosen randomly. Below we state the permitted range for the various parameters in the algorithm above.

1. Step size: We need to set the step size in the range  $3r/4s < \eta < r/s$ .
2. Threshold: At each step set the threshold at  $\tau^{(t)} = 16R^{(t-1)}M(R^{(t-1)}(s+1) + s/\sqrt{d})$ .
3. Tuning parameters: We need to pick  $\lambda^{(t)}$  and  $\nu^{(t)}$  such that the assumption (D5) is satisfied. A choice that is suitable that satisfies this assumption is

$$128s \left( R^{(t-1)} \right)^2 \leq \nu^{(t)} \leq 3,$$

$$32 \left( s^{3/2} \left( R^{(t-1)} \right)^2 + \frac{s^{3/2} R^{(t-1)}}{d^{1/2}} \right) \left( 4 + \frac{6}{\sqrt{s}} \right) \leq \lambda^{(t)} \leq 3.$$

We need to set  $\gamma^{(t)}$  as specified by Theorem 16,

$$\gamma^{(t)} = \sqrt{s} \left( R^{(t-1)} \right)^2 + \sqrt{\frac{s}{d}} R^{(t-1)}.$$

### 2.1 Sparse Regression Estimator

Our proof of convergence for Algorithm 1 also goes through with a different choices of robust sparse regression estimators, however, we can establish the tightest guarantees when the  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector is used in the sparse regression step. The  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector [8] was established as a modification of the Dantzig selector to handle uncertainty in the dictionary. There is a beautiful line of work that precedes this that includes [30; 31; 9]. There are also modified non-convex LASSO

programs that have been studied [23] and Orthogonal Matching Pursuit algorithms under in-variable errors [11]. However these estimators require the error in the dictionary to be stochastic and zero mean which makes them less suitable in this setting. Also note that standard  $\ell_1$  minimization estimators like the LASSO and Dantzig selector are highly unstable under errors in the dictionary and would lead to much worse guarantees in terms of radius of convergence (as studied in [1]). We establish the error guarantees for a robust sparse estimator  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector under fixed corruption in the dictionary. We also establish that this estimator is minimax optimal when the error in the sparse estimate is measured in infinity norm  $\|\hat{\theta} - \theta^*\|_\infty$  and the dictionary is corrupted.

### The $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector

Define the estimator  $\hat{\theta}$  such that  $(\hat{\theta}, \hat{t}, \hat{u}) \in \mathbb{R}^r \times \mathbb{R}_+ \times \mathbb{R}_+$  is the solution to the convex minimization problem

$$\min_{\theta, t, u} \left\{ \|\theta\|_1 + \lambda t + \nu u \mid \theta \in \Theta, \left\| \frac{1}{d} A^\top (y - A\theta) \right\|_\infty \leq \gamma t + R^2 u, \|\theta\|_2 \leq t, \|\theta\|_\infty \leq u \right\} \quad (2)$$

where,  $\gamma, \lambda$  and  $\nu$  are tuning parameters that are chosen appropriately.  $R$  is an upper bound on the error in our dictionary measured in matrix infinity norm. Henceforth the first coordinate ( $\hat{\theta}$ ) of this estimator is called  $MUS_{\gamma, \lambda, \nu}(y, A, R)$ , where the first argument is the sample, the second is the matrix, and the third is the value of the upper bound on the error of the dictionary measured in infinity norm. We will see that under our assumptions we will be able to establish an upper bound on the error on the estimator,  $\|\hat{\theta} - \theta^*\|_\infty \leq 16RM \left( R(s+1) + s/\sqrt{d} \right)$ , where  $|\theta_j^*| \leq M \forall j$ . We define a threshold at each step  $\tau = 16RM(R(s+1) + s/\sqrt{d})$ . The thresholded estimate  $\tilde{\theta}$  is defined as

$$\tilde{\theta}_i = \hat{\theta}_i \mathbb{I}[|\hat{\theta}_i| > \tau] \quad \forall i \in \{1, 2, \dots, r\}. \quad (3)$$

Our assumptions will ensure that we have the guarantee  $\text{sgn}(\tilde{\theta}) = \text{sgn}(\theta^*)$ . This will be crucial in our proof of convergence. The analysis of this estimator is presented in Appendix B.

As with previous analyses for alternating minimization for dictionary learning [1; 4], we rely on identifying the sign of the sparse covariates correctly at each step. For the sparse recovery step, we restrict ourselves to a class of two step estimators, where we first estimate a vector  $\hat{\theta}$  with some error guarantee in infinity norm  $\|\hat{\theta} - \theta^*\|_\infty \leq \tau$  and then have an element-wise thresholding step ( $\tilde{\theta}_i = \hat{\theta}_i \mathbb{I}[|\hat{\theta}_i| > \tau]$ ). To identify the sign correctly using this class of thresholded estimators we would like in the first step to use an estimator that is *optimal*, as this would lead to tighter control over the radius of convergence. This makes the choice of  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector natural, as we will show it is minimax optimal under certain settings.

**Theorem 1 (informal).** *Define the sets of matrices  $\mathcal{A} = \{B \in \mathbb{R}^{d \times r} \mid \|B_i\|_2 \leq 1, \forall i \in \{1, \dots, r\}\}$  and  $\mathcal{W} = \{P \in \mathbb{R}^{d \times r} \mid \|P\|_\infty \leq R\}$  with  $R = \mathcal{O}(1/\sqrt{s})$ . Then there exists an  $A^* \in \mathcal{A}$  and  $W \in \mathcal{W}$  with  $A \triangleq A^* + W$  such that*

$$\inf_{\hat{T}} \sup_{\theta^*} \|\hat{T} - \theta^*\|_\infty \geq CRL \left( \sqrt{1 - \frac{\log(s)}{\log(r)}} \right), \quad (4)$$

where the  $\inf_{\hat{T}}$  is over all measurable estimators  $\hat{T}$  with input  $(A^* \theta^*, A, R)$ , and the  $\sup$  is over  $s$ -sparse vectors  $\theta^*$  with 2-norm  $L > 0$ .

**Remark 2.** *Note that when  $R = \mathcal{O}(1/\sqrt{s})$  and  $s = \mathcal{O}(\sqrt{d})$ , this lower bound matches the upper bound we have for Theorem 16 (up to logarithmic factors) and hence the  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU Selector is minimax optimal.*

The proof of this theorem follows by Fano's method and is relegated to Appendix C.

## 2.2 Gradient Update for the dictionary

We note that the update to the dictionary at each step in Algorithm 1 is as follows

$$A_{ij}^{(t)} = A_{ij}^{(t-1)} - \underbrace{\eta \left( \frac{1}{n} \sum_{k=1}^n \left[ \sum_{p=1}^r \left( A_{ip}^{(t-1)} x_p^{k,(t)} x_j^{k,(t)} - y_i^k x_j^{k,(t)} \right) \right] \right)}_{\triangleq \hat{g}_{ij}^{(t)}},$$

for  $i \in \{1, \dots, d\}$ ,  $j \in \{1, \dots, r\}$  and  $t \in \{1, \dots, T\}$ . If we consider the loss function at time step  $t$  built using the vector samples  $y^1, \dots, y^n$  and sparse estimates  $x^{1,(t)}, \dots, x^{n,(t)}$ ,

$$\mathcal{L}_n(A) = \frac{1}{2n} \sum_{k=1}^n \|y^k - Ax^{k,(t)}\|_2^2, \quad A \in \mathbb{R}^{d \times r},$$

we can identify the update to the dictionary  $\hat{g}^{(t)}$  as the gradient of this loss function evaluated at  $A^{(t-1)}$ ,

$$\hat{g}^{(t)} = \left. \frac{\partial \mathcal{L}_n(A)}{\partial A} \right|_{A^{(t-1)}}.$$

## 3 Main Results and Assumptions

In this section we state our convergence result and state the assumptions under which our results are valid.

### 3.1 Assumptions

**Assumptions on  $A^*$**

(A1)  **$\mu$  - incoherence:** We assume the true underlying dictionary is  $\mu$ -incoherent

$$|\langle A_i^*, A_j^* \rangle| \leq \mu \quad \forall i, j \in \{1, \dots, r\} \text{ such that } i \neq j.$$

This is a standard assumption in the sparse regression literature when support recovery is of interest. It was introduced in [15; 34] in signal processing and independently in [38; 25] in statistics. We can also establish guarantees under the strictly weaker  $\ell_\infty$ -sensitivity condition (cf. [16]) used in analyzing sparse estimators under in-variable uncertainty in [9; 31]. The  $\{\ell_1, \ell_2, \ell_\infty\}$ -MU selector that we use for our sparse recovery step also works with this more general assumption, however for ease of exposition we assume  $A^*$  to be  $\mu$ -incoherent.

(A2) **Bounded max-norm:** We assume that  $A^*$  is bounded in matrix infinity norm

$$\|A^*\|_\infty \leq \frac{C_b}{s}.$$

This is in contrast with previous work that imposes conditions on the operator norm of  $A^*$  [4; 1; 5]. Our assumptions help provide guarantees under alternate assumptions and it also allows the operator norm to grow with dimension, whereas earlier work requires  $A^*$  to be such that  $\|A^*\|_{op} \leq C \left( \sqrt{r/d} \right)$ . In general the infinity norm and operator norm balls are hard to compare. However, one situation where a comparison is possible is if we assume the entries of the dictionary to be drawn iid from a Gaussian distribution  $\mathcal{N}(0, \sigma^2)$ . Then by standard concentration theorems, for the operator norm condition to be satisfied we would need the variance ( $\sigma^2$ ) of the distribution to scale as  $\mathcal{O}(1/d)$  while, for the infinity norm condition to be satisfied we need the variance to be  $\tilde{\mathcal{O}}(1/s^2)$ . This means that modulo constants the variance can be much larger for the infinity norm condition to be satisfied than for the operator norm condition.

(A3) **Normalized Columns:** We assume that all the columns of  $A^*$  are normalized to 1,

$$\|A_i^*\|_2 = 1 \quad \forall i \in \{1, \dots, r\}.$$

Note that the samples  $\{y^i\}_{i=1}^n$  are invariant when we scale the columns of  $A^*$  or under permutations of its columns. Thus we restrict ourselves to dictionaries with normalized columns and label the entire equivalence class of dictionaries with permuted columns and varying signs as  $A^*$ . We will converge linearly to the dictionary in this equivalence class that is closest to our initial estimate  $A^{(0)}$ .

### Assumption on the initial estimate and initialization

(B1) We require an initial estimate for the dictionary  $A^{(0)}$  such that,

$$\|A^{(0)} - A^*\|_\infty \leq \frac{C_R}{s}.$$

with  $2C_b \leq C_R$ ; where  $C_R = 1/2000M^2$ . Assuming  $2C_b \leq C_R$  allows us have a fast random initialization where we draw each entry of the initial estimate from the uniform distribution (on the interval  $[-C_b/s, C_b/s]$ ). This allows us to circumvent the computationally heavy SVD/clustering step involved in getting an initial dictionary estimate required in previous work [4; 1; 5]. Note that we need a random initialization and cannot start off with  $A^{(0)} = 0$  as this will be equally close to the entire equivalence class of dictionaries  $A^*$  (with varying signs of columns) and cause our sparse estimator to fail. A random initialization perturbs the initial solution to be closest to one of the dictionaries in the equivalence class and then we linearly converge to that dictionary in the class.

### Assumptions on $x^*$

Next we assume a generative model on the  $s$ -sparse covariates  $x^*$ . Here are the assumptions we make about the (unknown) distribution of  $x^*$ .

(C1) **Conditional Independence:** We assume that distribution of non-zero entries of  $x^*$  is conditionally independent and identically distributed. That is,  $x_i^* \perp\!\!\!\perp x_j^* | x_i^*, x_j^* \neq 0$ .

(C2) **Sparsity Level:** We assume that the level of sparsity  $s$  is bounded

$$2 \leq s \leq \min(2\sqrt{d}, C_b\sqrt{d}, C\sqrt{d}/\mu),$$

where  $C$  is an appropriate global constant such that  $A^*$  satisfies assumption (D3), see Remark 15. For  $\mu$ -incoherent dictionaries this upper bound is tight up to constant factors for sparse recovery to be feasible [12; 18].

(C3) **Boundedness:** Conditioned on the event that  $i$  is in the subset of non-zero entries, we have

$$m \leq |x_i^*| \leq M,$$

with  $m \geq 32R^{(0)}M(R^{(0)}(s+1) + s/\sqrt{d})$  and  $M > 1$ . This is needed for the thresholded sparse estimator to correctly predict the sign of the true covariate ( $\text{sgn}(x) = \text{sgn}(x^*)$ ). We can also relax the boundedness assumption: it suffices for the  $x_i^*$  to have sub-Gaussian distributions.

(C4) **Probability of support:** The probability of  $i$  being in the support of  $x^*$  is uniform over all  $i \in \{1, 2, \dots, r\}$ . This translates to

$$\Pr(x_i^* \neq 0) = \frac{s}{r} \quad \forall i \in \{1, \dots, r\}, \quad \Pr(x_i^*, x_j^* \neq 0) = \frac{s(s-1)}{r(r-1)} \quad \forall i \neq j \in \{1, \dots, r\}.$$

(C5) **Mean and variance of variables in the support:** We assume that the non-zero random variables in the support of  $x^*$  are centered and are normalized

$$\mathbb{E}(x_i^* | x_i^* \neq 0) = 0, \quad \mathbb{E}(x_i^{*2} | x_i^* \neq 0) = 1.$$

We note that these assumptions (A1), (A3) and (C1) - (C5) are similar to those made in [4; 1]. Agarwal et al. [1] require the matrices to satisfy the restricted isometry property, which is strictly weaker than  $\mu$ -incoherence, however they can tolerate a much lower level of sparsity ( $d^{1/6}$ ).

## 3.2 Main Result

**Theorem 3.** Suppose that true dictionary  $A^*$  and the distribution of the  $s$ -sparse samples  $x^*$  satisfy the assumptions stated in Section 3.1 and we are given an estimate  $A^{(0)}$  such that  $\|A^{(0)} - A^*\|_\infty \leq R^{(0)} \leq C_R/s$ . If we are given  $n$  i.i.d. samples in every iteration with  $n = \Omega(rsM^2 \log(dr/\delta))$  then Algorithm 1 with parameters  $(\{\tau^{(t)}\}_{t=1}^T, \{\gamma^{(t)}\}_{t=1}^T, \{\lambda^{(t)}\}_{t=1}^T, \{\nu^{(t)}\}_{t=1}^T, \eta)$  chosen as specified in Section 3.1 after  $T$  iterations returns a dictionary  $A^{(T)}$  such that,

$$\|A^{(T)} - A^*\|_\infty \leq \left(\frac{3}{4}\right)^T R^{(0)} + 4\eta|\epsilon_n|$$

where  $4\eta|\epsilon_n| \leq R^{(0)}/4$  with probability  $1 - \delta$ .

We note that the  $\epsilon_n$  can be driven to be smaller with high probability at the cost of more samples.

## 4 Proof of Convergence

In this section we will prove the main convergence results stated as Theorem 3. To prove this result we will analyze the gradient update to the dictionary at each step. We will decompose this gradient update (which is a random variable) into a first term which is its expected value and a second term which is its deviation from expectation. We will prove a deterministic convergence result by working with the expected value of the gradient and then appeal to standard concentration arguments to control the deviation of the gradient from its expected value with high probability.

By Lemma 8, Algorithm 1 is guaranteed to estimate the sign pattern correctly at every round of the algorithm,  $\text{sgn}(x) = \text{sgn}(x^*)$  (see proof in Appendix A.1).

To un-clutter notation let,  $A_{ij}^* = a_{ij}^*$ ,  $A_{ij}^{(t)} = a_{ij}$ ,  $A_{ij}^{(t+1)} = a_{ij}'$ . The  $k^{\text{th}}$  coordinate of the  $m^{\text{th}}$  covariate is written as  $x_k^{m*}$ . Similarly let  $x_k^m$  be the  $k^{\text{th}}$  coordinate of the estimate of the  $m^{\text{th}}$  covariate at step  $t$ . Finally let  $R^{(t)} = R$  and  $\hat{g}_{ij}$  be the  $(i, j)^{\text{th}}$  element of the gradient with  $n$  samples at step  $t$ . Unwrapping the expression for  $\hat{g}_{ij}$  we get,

$$\begin{aligned}\hat{g}_{ij} &= \frac{1}{n} \sum_{m=1}^n \left[ \sum_{k=1}^r (a_{ik} x_k^m x_j^m) - y_i^m x_j^m \right] \\ &= \frac{1}{n} \sum_{m=1}^n \left[ \sum_{k=1}^r (a_{ik} x_k^m - a_{ik}^* x_k^{m*}) x_j^m \right] \\ &= \mathbb{E} \left[ \sum_{k=1}^r (a_{ik} x_k - a_{ik}^* x_k^*) x_j \right] \\ &\quad + \left[ \frac{1}{n} \sum_{m=1}^n \left[ \sum_{k=1}^r (a_{ik} x_k^m - a_{ik}^* x_k^{m*}) x_j^m \right] - \mathbb{E} \left[ \sum_{k=1}^r (a_{ik} x_k - a_{ik}^* x_k^*) x_j \right] \right] \\ &= g_{ij} + \underbrace{\hat{g}_{ij} - g_{ij}}_{\triangleq \epsilon_n},\end{aligned}$$

where  $g_{ij}$  denotes  $(i, j)^{\text{th}}$  element of the expected value (infinite samples) of the gradient. The second term  $\epsilon_n$  is the deviation of the gradient from its expected value. By Theorem 10 we can control the deviation of the sample gradient from its mean via an application of McDiarmid's inequality. With this notation in place we are now ready to prove Theorem 3.

**Proof** [Proof of Theorem 3] First we analyze the structure of the expected value of the gradient.

*Step 1:* Unwrapping the expected value of the gradient we find it decomposes into three terms

$$\begin{aligned}g_{ij} &= \mathbb{E} (a_{ij} x_j^2 - a_{ij}^* x_j^* x_j) + \mathbb{E} \left[ \sum_{k \neq j} a_{ik} x_k x_j - a_{ik}^* x_k^* x_j \right] \\ &= \underbrace{(a_{ij} - a_{ij}^*) \frac{s}{r} \mathbb{E} [x_j^2 | x_j^* \neq 0]}_{\triangleq g_{ij}^c} \\ &\quad + \underbrace{a_{ij}^* \frac{s}{r} \mathbb{E} [(x_j - x_j^*) x_j | x_j^* \neq 0]}_{\triangleq \Xi_1} + \underbrace{\mathbb{E} \left[ \sum_{k \neq j} a_{ik} x_k x_j - a_{ik}^* x_k^* x_j \right]}_{\triangleq \Xi_2}.\end{aligned}$$

The first term  $g_{ij}^c$  points in the *correct direction* and will be useful in converging to the right answer. The other terms could be in a bad direction and we will control their magnitude with Lemma 5 such that  $|\Xi_1| + |\Xi_2| \leq \frac{s}{3r} R$ . The proof of Lemma 5 is the main technical challenge in the convergence analysis to control the error in the gradient. Its proof is deferred to the appendix.



Step 2: Given this bound, we analyze the gradient update,

$$\begin{aligned} a'_{ij} &= a_{ij} - \eta \hat{g}_{ij} \\ &= a_{ij} - \eta(g_{ij} + \epsilon_n) \\ &= a_{ij} - \eta [g_{ij}^c + (\Xi_1 + \Xi_2) + \epsilon_n]. \end{aligned}$$

So if we look at the distance to the optimum  $a_{ij}^*$  we have the relation,

$$a'_{ij} - a_{ij}^* = a_{ij} - a_{ij}^* - \eta(a_{ij} - a_{ij}^*) \frac{s}{r} \mathbb{E}[x_j^2 | x_j^* \neq 0] - \eta \{(\Xi_1 + \Xi_2) + \epsilon_n\}.$$

Taking absolute values, we get

$$\begin{aligned} |a'_{ij} - a_{ij}^*| &\stackrel{(i)}{\leq} \left(1 - \eta \frac{s}{r} \mathbb{E}[x_j^2 | x_j^* \neq 0]\right) |a_{ij} - a_{ij}^*| + \eta \{|\Xi_1| + |\Xi_2| + |\epsilon_n|\} \\ &\stackrel{(ii)}{\leq} \left(1 - \eta \frac{s}{r} \mathbb{E}[x_j^2 | x_j^* \neq 0]\right) |a_{ij} - a_{ij}^*| + \eta \left(\frac{s}{3r} R\right) + \eta |\epsilon_n| \\ &\leq \left(1 - \eta \frac{s}{r} \left\{\mathbb{E}[x_j^2 | x_j^* \neq 0] - \frac{1}{3}\right\}\right) R + \eta |\epsilon_n|, \end{aligned}$$

provided the first term is at non-negative. Here, (i) follows by triangle inequality and (ii) is by Lemma 5. Next we give an upper and lower bound on  $\mathbb{E}[x_j^2 | x_j^* \neq 0]$ . We would expect that as  $R$  gets smaller this variance term approaches  $\mathbb{E}[x_j^{*2} | x_j^* \neq 0] = 1$ . By invoking Lemma 6 we can bound this term to be  $\frac{2}{3} \leq \mathbb{E}[x_j^2 | x_j^* \neq 0] \leq \frac{4}{3}$ . So if we want converge at a rate  $3/4$  then it suffices to have

$$0 \leq \left(1 - \eta \frac{s}{r} \left\{\mathbb{E}[x_j^2 | x_j^* \neq 0] - \frac{1}{3}\right\}\right) \stackrel{(ii)}{\leq} \frac{3}{4}.$$

Coupled with Lemma 6, Inequality (i) follows from  $\eta \leq \frac{r}{s}$  while inequality (ii) follows from  $\eta \geq \frac{3r}{4s}$ . So if we unroll the bound for  $t$  steps we have,

$$\begin{aligned} |a_{ij}^{(t)} - a_{ij}^*| &\leq \frac{3}{4} |R^{(t-1)}| + \eta |\epsilon_n| \\ &\leq \frac{3}{4} \left(\frac{3}{4} |R^{(t-2)}| + \eta |\epsilon_n|\right) + \eta |\epsilon_n| \\ &\leq \left(\frac{3}{4}\right)^t |R^{(0)}| + \left\{\sum_{q=0}^{t-1} \left(\frac{3}{4}\right)^q\right\} (\eta |\epsilon_n|) \\ &\leq \left(\frac{3}{4}\right)^t |R^{(0)}| + 4\eta |\epsilon_n| \quad \text{as } \left(\sum_{q=0}^{\infty} (3/4)^q = 4\right). \end{aligned}$$

By Theorem 10, we have that  $4\eta |\epsilon_n| \leq R^{(0)}/4$  with probability  $1 - \delta$ , thus we are guaranteed to remain in our initial ball of radius  $R^{(0)}$  with high probability, completing the proof. ■

## 5 Conclusion

An interesting question would be to further explore and analyze the range of algorithms for which alternating minimization works and identifying the conditions under which they provably converge. There also seem to be many open questions for improper dictionary learning and developing provably faster algorithms there. Going beyond sparsity  $\sqrt{d}$  still remains challenging, and as noted in previous work alternating minimization also appears to break down experimentally and new algorithms are required in this regime. Also all theoretical work on analyzing alternating minimization for dictionary learning seems to rely on identifying the signs of the samples ( $x^*$ ) correctly at every step. It would be an interesting theoretical question to analyze if this is a necessary condition or if an alternate proof strategy and consequently a bigger radius of convergence are possible. Lastly, it is not known what the optimal sample complexity for this problem is and lower bounds there could be useful in designing more sample efficient algorithms.



## Acknowledgments

We gratefully acknowledge the support of the NSF through grant IIS-1619362, and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the ARC Centre of Excellence for Mathematical and Statistical Frontiers. Thanks also to the Simons Institute for the Theory of Computing Spring 2017 Program on Foundations of Machine Learning.

## References

- [1] Agarwal, A., A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon (2014). Learning sparsely used overcomplete dictionaries. In *COLT*, pp. 123–137.
- [2] Agarwal, A., A. Anandkumar, and P. Netrapalli (2013). A clustering approach to learn sparsely-used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*.
- [3] Aharon, M., M. Elad, and A. Bruckstein (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing* 54(11), 4311–4322.
- [4] Arora, S., R. Ge, T. Ma, and A. Moitra (2015). Simple, efficient, and neural algorithms for sparse coding. In *COLT*, pp. 113–149.
- [5] Arora, S., R. Ge, and A. Moitra (2013). New algorithms for learning incoherent and overcomplete dictionaries.
- [6] Balakrishnan, S., M. J. Wainwright, B. Yu, et al. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* 45(1), 77–120.
- [7] Barak, B., J. A. Kelner, and D. Steurer (2015). Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 143–151. ACM.
- [8] Belloni, A., M. Rosenbaum, and A. B. Tsybakov (2014). An  $\{\ell_1, \ell_2, \ell_\infty\}$ -Regularization Approach to High-Dimensional Errors-in-variables Models. *arXiv preprint arXiv:1412.7216*.
- [9] Belloni, A., M. Rosenbaum, and A. B. Tsybakov (2016). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- [10] Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.
- [11] Chen, Y. and C. Caramanis (2013). Noisy and Missing Data Regression: Distribution-Oblivious Support Recovery.
- [12] Donoho, D. L. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory* 47(7), 2845–2862.
- [13] Elad, M. and M. Aharon (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing* 15(12), 3736–3745.
- [14] Engan, K., S. O. Aase, and J. H. Husoy (1999). Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, Volume 5, pp. 2443–2446. IEEE.
- [15] Fuchs, J.-J. (2004). Recovery of exact sparse representations in the presence of noise. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Volume 2, pp. ii–533. IEEE.
- [16] Gautier, E. and A. Tsybakov (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.
- [17] Gribonval, R., R. Jenatton, and F. Bach (2015). Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory* 61(11), 6298–6319.

- [18] Gribonval, R. and M. Nielsen (2003). Sparse representations in unions of bases. *IEEE transactions on Information theory* 49(12), 3320–3325.
- [19] Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660. IEEE.
- [20] Hazan, E. and T. Ma (2016). A Non-generative Framework and Convex Relaxations for Unsupervised Learning. In *Advances in Neural Information Processing Systems*, pp. 3306–3314.
- [21] Jain, P., P. Netrapalli, and S. Sanghavi (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM.
- [22] Lewicki, M. S. and T. J. Sejnowski (2000). Learning overcomplete representations. *Neural computation* 12(2), 337–365.
- [23] Loh, P.-L. and M. J. Wainwright (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pp. 2726–2734.
- [24] McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics* 14(1), 148–188.
- [25] Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The annals of statistics*, 1436–1462.
- [26] Netrapalli, P., P. Jain, and S. Sanghavi (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804.
- [27] Netrapalli, P., U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain (2014). Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pp. 1107–1115.
- [28] Olshausen, B. A. and D. J. Field (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research* 37(23), 3311–3325.
- [29] Rigollet, P. and A. Tsybakov (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 731–771.
- [30] Rosenbaum, M., A. B. Tsybakov, et al. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics* 38(5), 2620–2651.
- [31] Rosenbaum, M., A. B. Tsybakov, et al. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes—A Festschrift in Honor of Jon A. Wellner*, pp. 276–290. Institute of Mathematical Statistics.
- [32] Spielman, D. A., H. Wang, and J. Wright (2012). Exact Recovery of Sparsely-Used Dictionaries. In *COLT*, pp. 37–1.
- [33] Sun, J., Q. Qu, and J. Wright (2017). Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory* 63(2), 853–884.
- [34] Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory* 52(3), 1030–1051.
- [35] Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats.
- [36] Wu, S. and B. Yu (2015). Local identifiability of  $\ell_1$ -minimization dictionary learning: a sufficient and almost necessary condition. *arXiv preprint arXiv:1505.04363*.
- [37] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.
- [38] Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine learning research* 7(Nov), 2541–2563.