# On Learning Sparsely Used Dictionaries from Incomplete Samples

**Thanh V. Nguyen** [1] **Akshay Soni** [2] **Chinmay Hegde** [1]

## Abstract

Existing algorithms for dictionary learning assume that the entries of the (high-dimensional) input data are fully observed. However, in several practical applications, only an incomplete fraction of the data entries may be available. For incomplete settings, no provably correct and polynomial-time algorithm has been reported in the dictionary learning literature. In this paper, we provide provable approaches for learning – from incomplete samples – a family of dictionaries whose atoms have sufficiently "spread-out" mass. First, we propose a descent-style iterative algorithm that linearly converges to the true dictionary when provided a sufficiently coarse initial estimate. Second, we propose an initialization algorithm that utilizes a small number of extra fully observed samples to produce such a coarse initial estimate. Finally, we theoretically analyze their performance and provide asymptotic statistical and computational guarantees.

## 1. Introduction

### 1.1. Motivation

In this paper, we consider a variant of the problem of *dictionary learning*, a widely used unsupervised technique for learning compact (sparse) representations of high dimensional data. At its core, the challenge in dictionary learning is to discover a basis (or dictionary) that can sparsely represent a given set of data samples with as little empirical representation error as possible. The study of sparse coding enjoys a rich history in image processing, machine learning, and compressive sensing (Elad & Aharon, 2006; Aharon et al., 2006; Olshausen & Field, 1997; Candes & Tao, 2005; Rubinstein et al., 2010; Gregor & LeCun, 2010; Boureau et al., 2010). While the majority of these aforementioned works involved heuristics, several exciting recent results (Spielman et al., 2012; Agarwal et al., 2013; 2014; Arora et al., 2014; 2015; Sun et al., 2015; Chatterji & Bartlett, 2017; Nguyen et al., 2018) have established rigorous conditions under which their algorithms recover the true dictionary under suitable generative models for the data.

An important underlying assumption that guides the success of all existing dictionary learning algorithms is the availability of (sufficiently many) data samples that are fully observed. Our focus, on the other hand, is on the special case where *the given data points are only partially observed*, that is, we are given access to only a small fraction of the coordinates of the data samples.

Such a setting of incomplete observations is natural in many applications like image-inpainting and demosaicing (Rubinstein et al., 2010). For example, this routinely appears in hyper-spectral imaging (Xing et al., 2012) where entire spectral bands of signals could be missing or unobserved. Moreover, in other applications, collecting fully observed samples can be expensive (or in some cases, even infeasible). Examples include the highly unreliable continuous blood glucose (CBG) monitoring systems that suffer from signal dropouts, where often the task is to learn a dictionary from partially observed signals (Naumova & Schnass, 2017a).

Earlier works that tackle the incomplete variant of the dictionary learning problem only offer heuristic solutions (Xing et al., 2012; Naumova & Schnass, 2017a) or involve constructing intractable statistical estimators (Soni et al., 2016). Indeed, the recovery of the true dictionary involves analyzing an extremely non-convex optimization problem that is, in general, not solvable in polynomial time (Loh & Wainwright, 2011). To our knowledge, our work is the first to give a theoretically sound as well as tractable algorithm to recover the exact dictionary from missing data (provided certain natural assumptions are met).

### 1.2. Our Contributions

In this paper, we make concrete theoretical algorithmic progress to the dictionary learning problem with incomplete samples. Inspired by recent algorithmic advances in dictionary learning (Arora et al., 2014; 2015), we adopt a learning-theoretic setup. Specifically, we assume that each data sample is synthesized from a generative model with an unknown dictionary and a random $k$-*sparse* coefficient

[1]Iowa State University [2]Yahoo! Research. Correspondence to: Thanh V. Nguyen <thanhng@iastate.edu>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. 

vector (or sparse code). Mathematically, the data samples $Y = [y^{(1)}, y^{(2)}, \ldots, y^{(p)}] \in \mathbb{R}^{n \times p}$ are of the form

$$Y = A^* X^* ,$$

where $A^* \in \mathbb{R}^{n \times m}$ denotes the dictionary and $X^* \in \mathbb{R}^{m \times p}$ denotes the (column-wise) $k$-sparse codes.

However, we do not have direct access to the data; instead, each high-dimensional data sample is further subsampled such that only a small fraction of the entries are observed. The assumption we make is that each entry of $Y$ is observed independently with probability $\rho \in (0, 1]$. For reasons that will become clear, we also assume that the ground truth dictionary $A^*$ is both *incoherent* (i.e., the columns of $A^*$ are sufficiently close to orthogonal) and *democratic* (i.e., the energy of each atom is well spread). Both these assumptions are standard in the compressive-sensing literature. We clarify the generative model more precisely in the sequel.

Given a set of such (partially observed) data samples, our goal is to recover the true dictionary $A^*$. Towards this goal, we make the following contributions:

1. Let us assume, for a moment, that we are given a coarse estimate $A^0$ that is sufficiently close to the true dictionary. We devise a descent-style algorithm that leverages the given incomplete data to iteratively refine the dictionary estimate; moreover, we show that it converges rapidly to an estimate within a small ball of the ground truth $A^*$ (whose radius decreases given more samples). Our result can be informally summarized as follows:

**Theorem 1** (Informal, descent)**.** *When given a "sufficiently-close" initial estimate $A^0$, there exists an iterative gradient descent-type algorithm that linearly converges to the true dictionary with $O(mk \operatorname{polylog}(n))$ incomplete samples.*

Our above result mirrors several recent results in non-convex learning that all develop a descent algorithm which succeeds given a good enough initialization (Yuan & Zhang, 2013; Cai et al., 2016; Tu et al., 2016). Indeed, similar guarantees for descent-style algorithms (such as alternating minimization) exist for the related problem of *matrix completion* (Jain et al., 2013), which coincides with our setting if $m \ll n$. However, our setting is distinct, since we are interested in learning *overcomplete* dictionaries, where $m > n$.

2. Having established the efficiency of the above refinement procedure, we then address the challenge of actually coming up with a coarse estimate of $A^*$. We do not know of a provable procedure that produces a good enough initial estimate using partial samples. To circumvent this issue, we assume availability of $O(m)$ *fully* observed samples along with the partial samples[1]. Given this setting, we show that we can provide a "sufficiently close" initial estimate in polynomial time. Our result can be summarized as follows:

**Theorem 2** (Informal, initialization)**.** *There exists an initialization algorithm that, given $O(m \operatorname{polylog}(n))$ fully observed samples and an additional $O(mk \operatorname{polylog}(n))$ partially observed samples, returns an initial estimate $A^0$ that is sufficiently close to $A^*$ in a column-wise sense.*

### 1.3. Techniques

The majority of our theoretical contributions are fairly technical, so for clarity, we provide some non-rigorous intuition.

At a high level, our approach merges ideas from two main themes in the algorithmic learning theory literature. We build upon recent seminal, theoretically-sound algorithms for sparse coding (specifically, the framework of Arora et al. (2015)). Their approach consists of a descent-based algorithm performed over the surface of a suitably defined loss function of the dictionary parameters. The descent is achieved by alternating between updating the dictionary estimate and updating the sparse codes of the data samples. The authors prove that this algorithm succeeds provided that the codes are sparse enough, the columns of $A^*$ are incoherent, and that we are given sufficiently many samples.

However, a direct application of the above framework to the partially observed setting does not seem to succeed. To resolve this, we leverage a specific property that is commonly assumed in the matrix completion literature: we suppose that the dictionaries are not "spiky" and that the energy of each atom is spread out among its coordinates; specifically, the *sub*-dictionaries formed by randomly sub-selecting rows are still incoherent. We call such dictionaries *democratic*, following the terminology of Davenport et al. (2009). (In matrix completion papers, this property is also sometimes referred to incoherence, but we avoid doing so since that overloads the term.) Our main contribution is to show that democratic, incoherent dictionaries can be learned via a similar alternating descent scheme if only a small fraction of the data entries are available. Our analysis is novel and distinct than that provided in (Arora et al., 2015).

Of course, the above analysis is somewhat local in nature since we are using a descent-style method. In order to get global guarantees for recovery of $A^*$, we need to initialize carefully. Here too, the spectral initialization strategies suggested in earlier dictionary learning papers (Arora et al., 2014; 2015) do not succeed. To resolve this, we again appeal to the democracy property of $A^*$. We also need

[1] While this might be a limitation of our analysis, we emphasize that the number of full samples needed by our method is relatively small. Indeed, the state-of-the-art approach for dictionary learning (Arora et al., 2015) requires $O(mk \operatorname{polylog}(n))$ fully observed samples, while our method needs only $O(m \operatorname{polylog}(n))$ samples, which represents a polynomial improvement since $k$ can be as large as $\sqrt{n}$.

to assume that provided a small hold-out set of additional, *fully* observed samples is available[2]. Using this hold-out set (which can be construed as additional prior information or "side" information) together with the available samples gives us a spectral initialization strategy that provably gives a good enough initial estimate.

Putting the above two pieces together: if we are provided $O(mk/\rho^4$ polylog $n)$ partially observed samples from the generative model, together with an additional $O(m$ polylog $n)$ full samples, then we can guarantee a fast, provable algorithm for learning $A^*$. See Table 1 for a summary of our results, and comparison with existing work. We remark that while our algorithms only succeed up to sparsity level $k \leq O(\rho\sqrt{n})$, we obtain a running time improvement over the best available dictionary learning approaches.

### 1.4. Relation to Prior Work

The literature on dictionary learning (or sparse coding) is very vast and hence our references to prior work will necessarily be incomplete; we refer to the seminal work of Rubinstein et al. (2010) for a list of applications. Dictionary learning with incompletely observed data, however, is far less well-understood. Initial attempts in this direction (Xing et al., 2012) involve Bayesian-style techniques; more recent attempts have focused on alternating minimization techniques, along with incoherence- and democracy-type assumptions akin to our framework (Naumova & Schnass, 2017b;a). However, none of these methods provide rigorous polynomial-time algorithms that provably succeed in recovering the dictionary parameters.

Our setup can also be viewed as an instance of matrix completion, which has been a source of intense interest in the machine learning community over the last decade (Candès & Recht, 2009; Keshavan et al., 2010). The typical assumption in such approaches is that the data matrix $Y = A^*X^*$ is low-rank (i.e., $A^*$ typically spans a low-dimensional subspace). This assumption leads to either feasible convex relaxations, or a bilinear form that can be solved approximately via alternating minimization. However, our work differs significantly from this setup, since we are interested in the case where $A^*$ is over-complete; moreover, our guarantees are not in terms of estimating the missing entries of $Y$, but rather obtaining the atoms in $A^*$. Note that our generative model also differs from the setup of *high-rank* matrix completion (Eriksson et al., 2012), where the data is sampled randomly from a finite union-of-subspaces. In contrast, our data samples are synthesized via sparse linear combinations of a given dictionary.

[2]We do not know how to remove this assumption, and it appears that techniques stronger than spectral initialization (e.g., involving higher-order moments) are required.

In the context of matrix-completion, perhaps the most related work to ours is the statistical analysis of matrix-completion under the *sparse-factor model* of Soni et al. (2016), which employs a similar generative data model to ours. (Similar sparse-factor models have been studied in the work of Lan et al. (2014), but no complexity guarantees are provided.) For this model, Soni et al. (2016) propose a highly non-convex statistical estimator for estimate $Y$ and provide error bounds for this estimator under various noise models. However, they do not discuss an efficient algorithm to realize that estimator. In contrast, we provide rigorous polynomial time algorithms, together with error bounds on the estimation quality of $A^*$. Overall, we anticipate that our work can shed some light on the design of provable algorithms for matrix-completion in such more general settings.

## 2. Preliminaries

**Notation.** Given a vector $x \in \mathbb{R}^m$ and a subset $S \subseteq [m]$, we denote $x_S \in \mathbb{R}^m$ as a vector which equals $x$ in indices belonging to $S$ and equals zero elsewhere. We use $A_{\bullet i}$ and $A^T_{j\bullet}$ respectively to denote the $i^{\text{th}}$ column and the $j^{\text{th}}$ row of matrix $A \in \mathbb{R}^{n\times m}$. We use $A_{\bullet S}$ as the submatrix of $A$ with columns in $S$. In contrast, we use $A_{\Gamma\bullet}$ to indicate the submatrix of $A$ with rows not in $\Gamma$ set to zero. Let $\text{supp}(x)$ and $\text{sgn}(x)$ be the support and element-wise sign of $x$. Let $\text{threshold}_K(x)$ be the *hard-thresholding* operator that sets all entries of $x$ with magnitude less than $K$ to zero. The symbol $\|\cdot\|$ refers to the $\ell_2$-norm, unless otherwise specified.

For asymptotic analysis, we use $\widetilde{\Omega}(\cdot)$ and $\widetilde{O}(\cdot)$ to represent $\Omega(\cdot)$ and $O(\cdot)$ up to (unspecified) poly-logarithmic factors depending on $n$. Besides, $g(n) = O^*(f(n))$ denotes $g(n) \leq Kf(n)$ for some sufficiently small constant $K$. Finally, the terms "with high probability" (abbreviated to w.h.p.) is used to indicate an event with failure probability $O(n^{-\omega(1)})$. We make use of the following definitions.

**Definition 1** (Incoherence)**.** *The matrix $A$ is* incoherent *with parameter $\mu$ if the following holds for all columns $i \neq j$:*

$$\frac{|\langle A_{\bullet i}, A_{\bullet j}\rangle|}{\|A_{\bullet i}\|\|A_{\bullet j}\|} \leq \frac{\mu}{\sqrt{n}}.$$

The incoherence property requires the columns of $A$ to be approximately orthogonal, and is a canonical property to resolve identifiability issues in dictionary learning and sparse recovery. We distinguish this from the conventional notion of "incoherence" widely used in the matrix completion literature. This notion is related to a notion that we call *democracy*, which we define next.

**Definition 2** (Democracy)**.** *Suppose that the matrix $A$ is $\mu$-incoherent. $A$ is further said to be* democratic *if the submatrix $A_{\Gamma\bullet}$ is $\mu$-incoherent for any subset $\Gamma \subset [n]$ of size $\sqrt{n} \leq |\Gamma| \leq n$.*

*Table 1.* Comparisons between different approaches.

| Setting | Reference | Sample complexity w/o noise | Running time | Sparsity | Incomplete samples |
|---|---|---|---|---|---|
| Regular | (Spielman et al., 2012) | $O(n^2 \log n)$ | $\widetilde{\Omega}(n^4)$ | $O(\sqrt{n})$ | ✗ |
| | (Arora et al., 2014) | $\widetilde{O}(m^2/k^2)$ | $\widetilde{O}(np^2)$ | $O(\sqrt{n})$ | ✗ |
| | (Arora et al., 2015) | $\widetilde{O}(mk)$ | $\widetilde{O}(mn^2p)$ | $O(\sqrt{n})$ | ✗ |
| Incomplete | (Xing et al., 2012) | ✗ | ✗ | ✗ | ✓ |
| | (Naumova & Schnass, 2017a) | ✗ | ✗ | ✗ | ✓ |
| | This paper | $\widetilde{O}(mk/\rho^4)$ partial samples<br>$\widetilde{O}(m)$ full samples | $\widetilde{O}(\rho mn^2p)$ | $O(\rho\sqrt{n})$ | ✓ |

✗ *indicates no complexity guarantees. Here,* $n$ *is the data dimension;* $m$ *is the size of dictionary;* $k$ *is the sparsity of* $x$*;* $p$ *is the number of observed samples;* $\rho$ *is the subsampling probability.*

This property tells us that the rows of $A$ have roughly the same amount of "information", and that the submatrix of $A$ restricted to any subset of rows $\Gamma$ is also incoherent. A similar concept (stated in terms of the restricted isometry property) is well-known in the compressive sensing literature (Davenport et al., 2009). Several probabilistic constructions of dictionaries satisfy this property; typical examples include random matrices drawn from i.i.d. Gaussian or Rademacher distributions. The $\sqrt{n}$ lower bound on $|\Gamma|$ is to ensure that the submatrix of $A$ including only the rows in $\Gamma$ is balanced in terms of dimensions.

We seek an algorithm that provides a provably "good" estimate of $A^*$. For this, we need a suitable measure of "goodness". The following notion of distance records the maximal column-wise difference between any estimate $A$ and $A^*$ in $\ell_2$-norm under a suitable permutation and sign flip.

**Definition 3** ($(\delta, \kappa)$-nearness)**.** *The matrix $A$ is said to be $\delta$-close to $A^*$ if $\|\sigma(i)A_{\bullet\pi(i)} - A^*_{\bullet i}\| \leq \delta$ holds for every $i = 1, 2, \ldots, m$ and some permutation $\pi : [m] \to [m]$ and sign flip $\sigma : [m] : \{\pm 1\}$. In addition, if $\|A_{\bullet\pi} - A^*\| \leq \kappa\|A^*\|$ holds, then $A$ is said to be $(\delta, \kappa)$-near to $A^*$.*

To keep notation simple, in our convergence theorems below, whenever we discuss nearness, we simply replace the transformations $\pi$ and $\sigma$ in the above definition with the identity mapping $\pi(i) = i$ and the positive sign $\sigma(\cdot) = +1$ while keeping in mind that in reality, we are referring to finding one element in the equivalence class of all permutations and sign flips of $A^*$.

Armed with the above concepts, we now posit a generative model for our observed data. Suppose that the data samples $Y = [y^{(1)}, y^{(2)}, \ldots, y^{(p)}]$ are such that each column is generated according to the rule:

$$y = \mathcal{P}_\Gamma(A^*x^*), \tag{1}$$

where $A^*$ is an unknown, ground truth dictionary; $x^*$ and $\Gamma$ are drawn from some distribution $\mathcal{D}$ and $\mathcal{P}_\Gamma$ is the sampling operator that keeps entries in $\Gamma$ untouched and zeroes out everything else. We emphasize that $\Gamma$ is independently chosen for each $y^{(i)}$, so more precisely, $y^{(i)} = y^{(i)}_{\Gamma^{(i)}} \in \mathbb{R}^n$. We ignore the superscript to keep the notation simple. We also make the following assumptions:

**Assumption 1.** *The true dictionary $A^*$ is over-complete with $m \leq Kn$ for some constant $K > 1$, and democratic with parameter $\mu$. All columns of $A^*$ have unit norms.*

**Assumption 2.** *The true dictionary $A^*$ has bounded spectral and max ($\ell_\infty$-) norms such that $\|A^*\| \leq O(\sqrt{m/n})$ and $\|A^*\|_{\max} \leq O(1/\sqrt{n})$.*

**Assumption 3.** *The code vector $x^*$ is $k$-sparse random with uniform support $S$. The nonzero entries of $x^*$ are pairwise independent sub-Gaussian with variance 1, and bounded below by some known constant $C$.*

**Assumption 4.** *Each entry of the sample $A^*x^*$ is independently observed with constant probability $\rho \in (0, 1]$.*

The incoherence and spectral bound are ubiquitous in the dictionary learning literature (Arora et al., 2014; 2015). For the incomplete setting, we further require the democracy and max-norm bounds to control the spread of energy of the entries of $A^*$, so that $A^*$ is not "spiky". Such conditions are often encountered in the matrix completion literature (Candès & Recht, 2009; Keshavan et al., 2010). The distributional assumptions on the code vectors $x^*$ are standard in theoretical dictionary learning (Agarwal et al., 2014; Arora et al., 2014; Gribonval et al., 2015; Arora et al., 2015). Finally, we also require the sparsity $k \leq O^*(\rho\sqrt{n}/\log n)$ throughout the paper.

## 3. A Descent-Style Learning Algorithm

We now design and analyze an algorithm for learning the dictionary $A^*$ given incomplete samples of the form (1). Our strategy will be to use a descent-like scheme to construct a sequence of estimates $A$ which successively gets closer to

$A^*$ in the sense of $(\delta, \kappa)$-nearness.

Let us first provide some intuition. The natural approach to solve this problem is to perform gradient descent over an appropriate empirical loss of the dictionary parameters. More precisely, we consider the squared loss between observed entries of $Y$ and their estimates (which is the typical loss function used in the incomplete observations setting (Jain et al., 2013)):

$$\mathcal{L}(A) = \frac{1}{2} \sum_{i,j \in \Omega} (Y_{ij} - (AX)_{ij})^2, \tag{2}$$

where $\Omega$ is the set of locations of observed entries in the samples $Y$. However, straightforward gradient descent over $A$ is not possible for several reasons: (i) the gradient depends on the finite sample variability of $Y$; (ii) the gradient with respect to $A$ depends on the optimal code vectors of the data samples, $x_i^*$, which are unknown *a priori*; (iii) since we are working in the overcomplete setting, care has to be taken to ensure that the code vectors (i.e., columns of $X$) obey the sparsity model (as specified in Assumption 2).

The *neurally-plausible sparse coding* algorithm of Arora et al. (2015) provides a crucial insight into the understanding of the loss surface of $\mathcal{L}_A$ in the fully observed setting. Basically, within a small ball around the ground truth $A^*$, the surface is well behaved such that a *noisy* version of $X^*$ is sufficient to construct a good enough approximation to the gradient of $\mathcal{L}$. Moreover, given an estimate within a small ball around $A^*$, a noisy (but good enough) estimate of $X^*$ can be quickly computed using a thresholding operation.

We extend this understanding to the (much more challenging) setting of incomplete observations. Specifically, we show the loss surface in (2) behaves well even with missing data. This enables us to devise an algorithm similar to that of Arora et al. (2015) and obtain a descent property directly related to (the population parameter) $A^*$. The full procedure is detailed as Algorithm 1.

We now analyze our proposed algorithm. Specifically, we can show that if initialized properly and with proper choice of step size, Algorithm 1 exhibits *linear* convergence to a ball of radius $O(\sqrt{k/n})$ around $A^*$. Formally, we have:

**Theorem 3.** *Suppose that the initial estimate $A^0$ is $(\delta, 2)$-near to $A^*$ with $\delta = O^*(1/\log n)$ and the sampling probability satisfies $\rho \geq 1/(k+1)$. If Algorithm 1 is given $p = \widetilde{\Omega}(mk)$ fresh partial samples at each step and uses learning rate $\eta = \Theta(m/\rho k)$, then*

$$\mathbb{E}[\|A^s_{\bullet i} - A^*_{\bullet i}\|^2] \leq (1-\tau)^s \|A^0_{\bullet i} - A^*_{\bullet i}\|^2 + O(\sqrt{k/n})$$

*for some $0 < \tau < 1/2$ and $s = 1, 2, \ldots, T$. As a corollary, $A^s$ converges geometrically to $A^*$ until column-wise $O(\sqrt{k/n})$ error.*

**Algorithm 1** Gradient descent-style algorithm

**Input:** Partial samples $Y$ with observed entry set $\Gamma^{(i)}$
Initial $A^0$ that is $(\delta, 2)$-near to $A^*$
**for** $s = 0, 1, \ldots, T$ **do**
  /* Encoding step */
  **for** $i = 1, 2, \ldots, p$ **do**
    $x^{(i)} \leftarrow \text{threshold}_{C/2}(\frac{1}{\rho}(A^s)^T y^{(i)})$
  **end**
  /* Update step */
  $\widehat{g}^s \leftarrow \frac{1}{p} \sum_{i=1}^{p} (\mathcal{P}_{\Gamma^{(i)}}(A^s x^{(i)}) - y^{(i)}) \text{sgn}(x^{(i)})^T$
  $A^{s+1} \leftarrow A^s - \eta \widehat{g}^s$
**end**
**Output:** $A \leftarrow A^T$ as a learned dictionary

We defer the full proof of Theorem 3 to Appendix C. To understand the working of the algorithm and its correctness, let us consider the setting where we have access to infinitely many samples. This setting is, of course, fictional; however, expectations are easier to analyze than empirical averages, and moreover, this exercise reveals several key elements for proving Theorem 3. More precisely, we first provide bounds on the expected value of $\widehat{g}^s$, denoted as

$$g^s \triangleq \mathbb{E}_y[(\mathcal{P}_\Gamma(A^s x) - y)\text{sgn}(x)^T],$$

to establish the descent property for the infinite sample case. The sample complexity argument emerges when we control the concentration of $\widehat{g}^s$, detailed in Appendix C. Here, we separately discuss the encoding and update steps in Algorithm 1.

**Encoding step**. The first main result is to show that the hard-thresholding (or pooling)-based rule for estimating the sparse code vectors is sufficiently accurate. This rule adapts the encoding step of the dictionary learning algorithm proposed in (Arora et al., 2015), with an additional scaling factor $1/\rho$. This scaling is necessary to avoid biases arising due to the presence of incomplete information.

The primary novelty is in our analysis. Specifically, we prove that the estimate of $X$ obtained via the encoding step (even under partial observations) enables a good enough identification of the *support* of the true $X^*$. The key, here, is to leverage the fact that $A^*$ is *democratic* and that $A^s$ is near $A^*$. We call this property *support consistency* and establish it as follows.

**Lemma 1.** *Suppose that $A^s$ is $(\delta, 2)$-near to $A^*$ with $\delta = O^*(1/\log n)$. With high probability over $y = \mathcal{P}_\Gamma(A^* x^*)$, the estimate $x$ obtained by the encoding step of Algorithm 1 has the same sign as the true $x^*$; that is,*

$$\text{sgn}\big(\text{threshold}_{C/2}\big(\frac{1}{\rho}(A^s)^T y\big)\big) = \text{sgn}(x^*), \tag{3}$$

*This holds true for incoherence parameter $\mu \leq \frac{\sqrt{n}}{2k}$, sparsity parameter $k \geq \Omega(\log m)$ and subsampling probability $\rho \geq 1/(k+1)$.*

Lemma 1 implies that when the "mass" of $A^*$ is spread out across entries, within a small neighborhood of $A^*$ the estimate $x$ is reliable *even* if $y$ is incompletely observed. This lemma is the main ingredient for bounding the behavior of the update rule.

**Update step**. The support consistency property of the estimated $x$ arising in the encoding step is key to rigorously analyzing the expected gradient $g^s$. This relatively 'simple' encoding enables an explicit form of the update rule, and gives an intuitive reasoning on how the descent property can be achieved. In fact, we will see that

$$g_i^s = \rho p_i q_i (\lambda_i^s A_{\bullet i}^s - A_{\bullet i}^*) + o(\rho p_i q_i)$$

for $p_i = \mathbb{E}[|x_i^*||i \in S]$, $q_i = \mathbb{P}[i \in S]$ and $\lambda_i^s = \langle A_{\bullet i}, A_{\bullet i}^* \rangle$. Since we assume that the current estimate $A^s$ is (column-wise) sufficiently close to $A^*$, each $\lambda_i^s$ is approximately equal to 1, and hence $g_i^s \approx \rho p_i q_i (A_{\bullet i}^s - A_{\bullet i}^*)$, i.e., the gradient points in the desired direction. Combining this with standard analysis of gradient descent, we can prove that the overall algorithm geometrically decreases the error in each step $s$ as long as the learning rate $\eta$ is properly chosen. Specifically, we get the following theoretical result.

**Theorem 4.** *Suppose that $A^0$ is $(\delta, 2)$-near to $A^*$ with $\delta = O^*(1/\log n)$ and the sampling probability satisfies $\rho \geq 1/(k+1)$. Assuming infinitely many partial samples at each step, Algorithm 1 geometrically converges to $A^*$ until column-wise error $O(k/\rho n)$. More precisely,*

$$\|A_{\bullet i}^{s+1} - A_{\bullet i}^*\|^2 \leq (1-\tau)\|A_{\bullet i}^s - A_{\bullet i}^*\|^2 + O\big(k^2/\rho^2 n^2\big)$$

*for some $0 < \tau < 1/2$ and for $s = 1, 2, \ldots, T$ provided the learning rate obeys $\eta = \Theta(m/\rho k)$.*

We provide the mathematical proof for the form of $g^s$ as well as the descent in Appendix A.2. We also argue that the $(\delta, 2)$-nearness of $A^{s+1}$ and $A^*$ is maintained after each update. This is studied in Lemma 7 in Appendix A.

## 4. An Initialization Algorithm

In the previous section, we provided an algorithm that (accurately) recovers $A^*$ in an iterative descent-style approach. In order to establish correctness guarantees, the algorithm requires a coarse estimate $A^0$ that is $\delta$-close to the ground truth with closeness parameter $\delta = O^*(1/\log n)$. This section presents an initialization strategy to obtain such a good starting point for $A^*$.

Again, we begin with some intuition. At a high level, our algorithm mimics the spectral initialization strategy for dictionary learning proposed by (Arora et al., 2015). In essence, the idea is to re-weight the data samples (which are fully observed) appropriately. When this is the case, analyzing the spectral properties of the covariance matrix of the new re-weighted samples gives us the desired initialization. The re-weighting itself relies upon the computation of pairwise correlations between the samples with two fixed samples (say, $u$ and $v$) chosen from an independent *hold-out set*. This strategy is appealing in both from the standpoint of statistical efficiency as well as computational ease.

Unfortunately, a straightforward application of this strategy to our setting of incomplete observations does not work. The major issue, of course, is that pairwise correlation (the inner product) of two high dimensional vectors is highly uninformative if each vector is only partially observed. We circumvent this issue via the following simple (but key) observation: *provided the underlying dictionary is democratic* and *the representation is sufficiently sparse*, the correlation between a partially observed data sample $y$ with a *fully* observed sample $u$ is indeed proportional to the actual correlation between $y$ and $u$. Therefore, assuming that we are given a hold-out set *that is fully observed*, an adaptation of the spectral approach of Arora et al. (2015) provably succeeds. Moreover, the size of the hold-out set need not be large; in particular, we need only $O(m \operatorname{polylog}(n))$ fully-observed samples, as opposed to the $O(mk \operatorname{polylog}(n))$ samples required by the analysis of Arora et al. (2015). The parameter $k$ can be as big as $\sqrt{n}$, so in fact we require polynomially fewer fully-observed samples.

In summary: in order to initialize our descent procedure, we assume the availability of a small (but fully observed) hold-out set. In practice, we can imagine expending some amount of effort in the beginning to collect all the entries of a small subset of the available data samples. The availability of such additional information (or "side-information") has been made in the literature on matrix completion (Natarajan & Dhillon, 2014).

The full procedure is described in pseudocode form as Algorithm 2. Our main theoretical result (Theorem 5) summarizes its performance.

**Theorem 5.** *Suppose that the available training dataset consists of $p_1$ fully observed samples, together with $p_2$ incompletely observed samples according to the observation model (1). Suppose $\mu = O^*\big(\frac{\sqrt{n}}{k \log^3 n}\big)$, $\frac{1}{\rho} - 1 \leq k \leq O^*(\frac{\rho\sqrt{n}}{\log n})$. When $p_1 = \widetilde{\Omega}(m)$ and $p_2 = \widetilde{\Omega}(mk/\rho^4)$, then with high probability, Algorithm 2 returns an initial estimate $A^0$ whose columns share the same support as $A^*$ and is $(\delta, 2)$-near to $A^*$ with $\delta = O^*(1/\log n)$.*

The full proof is provided in Appendix B. To provide some intuition about the working of the algorithm and its proof, let us again consider the setting where we have access to infinitely many samples. These analyses result in key lemmas,

**Algorithm 2** Spectral initialization algorithm

**Input:** $\mathcal{P}_1$: $p_1$ fully observed samples
$\mathcal{P}_2$: $p_2$ partially observed samples
Set $L = \emptyset$
**while** $|L| < m$ **do**
  Pick $u$ and $v$ from $\mathcal{P}_1$ at random
  Construct the weighted covariance matrix $\widehat{M}_{u,v}$ using samples $y^{(i)}$ from $\mathcal{P}_2$

$$\widehat{M}_{u,v} \leftarrow \frac{1}{p_2\rho^4}\sum_{i=1}^{p_2}\langle y^{(i)}, u\rangle\langle y^{(i)}, v\rangle y^{(i)}(y^{(i)})^T$$

  $\delta_1, \delta_2 \leftarrow$ top singular values
  **if** $\delta_1 \geq \Omega(k/m)$ *and* $\delta_2 < O^*(k/m\log n)$ **then**
    $z \leftarrow$ top singular vector
    **if** $z$ *is not within distance* $1/\log n$ *of vectors in* $L$ *even with sign flip* **then**
      $L \leftarrow L \cup \{z\}$
    **end**
  **end**
**end**
**Output:** $A^0 \leftarrow \text{Proj}_{\mathcal{B}}(\tilde{A})$ where $\tilde{A}$ is the matrix whose columns in $L$ and $\mathcal{B} = \{A : \|A\| \leq 2\|A^*\|\}$

which we will reuse extensively for proving Theorem 5.

First, consider two *fully observed* data samples $u = A^*\alpha$ and $v = A^*\alpha'$ drawn from the hold-out set. (Here, $A^*, \alpha, \alpha'$ are unknown.) Consider also a partially observed sample $y = A^*_{\Gamma\bullet}x^*$ under a random subset $\Gamma \subseteq [n]$. Define:

$$\beta = \frac{1}{\rho}A^{*T}_{\Gamma\bullet}u, \text{ and } \beta' = \frac{1}{\rho}A^{*T}_{\Gamma\bullet}v$$

respectively as (crude) estimates of $\alpha$ and $\alpha'$, simply obtained by applying a (scaled) adjoint of $A_{\Gamma\bullet}$ to $u$ and $v$ respectively. It follows from the above definition that:

$$\beta = \frac{1}{\rho}A^{*T}_{\Gamma\bullet}A^*\alpha, \text{ and } \langle y, u\rangle = \rho\langle\beta, x^*\rangle.$$

Our main claim is that since $A^*$ is assumed to satisfy the democracy property, $\frac{1}{\rho}A^{*T}_{\Gamma\bullet}A^*$ resembles the identity, and hence $\beta$ "looks" like the true code vector $\alpha$. In particular, we have the following lemma.

**Lemma 2.** *With high probability over the randomness in* $u$ *and* $\Gamma$*, we have: (a)* $|\beta_i - \alpha_i| \leq \frac{\mu k\log n}{\sqrt{n}} + \sqrt{\frac{1-\rho}{\rho n^{1/2}}}$ *for each* $i = 1, 2, \ldots, m$ *and (b)* $\|\beta\| \leq \frac{\sqrt{k}\log n}{\rho}$.

*Proof.* Denote $U = \text{supp}(\alpha)$ and $W = U\backslash\{i\}$, then

$$\begin{aligned}|\beta_i - \alpha_i| &= \Big|\frac{1}{\rho}A^{*T}_{\Gamma,i}A^*_{\bullet W}\alpha_W + \big(\frac{1}{\rho}\langle A^*_{\Gamma,i}, A^*_{\bullet i}\rangle - 1\big)\alpha_i\Big| \\ &\leq \frac{1}{\rho}\big|A^{*T}_{\Gamma,i}A^*_{\bullet W}\alpha_W\big| + \Big|\big(\frac{1}{\rho}A^{*T}_{\Gamma,i}A^*_{\bullet i} - 1\big)\alpha_i\Big|. \end{aligned} \tag{4}$$

We will bound these terms on the right hand side of (4) using the properties of $A^*$ and $\alpha$. First, we notice that for any $\Gamma \subset [n]$:

$$\|A^{*T}_{\Gamma,i}A^*_{\bullet W}\|^2 = \sum_{j\in W}\langle A^*_{\Gamma,i}, A^*_{\bullet j}\rangle^2 \leq \frac{\mu^2}{n}\sum_{j\in W}\|A^*_{\Gamma,i}\|^2\|A^*_{\Gamma,j}\|^2,$$

where we have used the democracy of $A^*$ with respect to $\Gamma$. Moreover, using the Chernoff bound for $\|A^*_{\Gamma,i}\|^2 = \sum_{i=1}^n A^{*2}_{li}\mathbf{1}[l \in \Gamma]$, we have $\|A^*_{\Gamma,i}\|^2 \leq \rho + o(\rho)$ w.h.p. Hence, $\|A^{*T}_{\Gamma,i}A^*_{\bullet W}\|^2 \leq \rho^2\mu^2 k/n$ with high probability. In addition, $\|\alpha_W\| \leq \sqrt{k}\log n$ w.h.p. because $\alpha_W$ is $k$-sparse sub-Gaussian. Therefore, the first term in (4) gives $\frac{1}{\rho}|A^{*T}_{\Gamma,i}A^*_{\bullet W}\alpha_W| \leq \frac{\mu k\log n}{\sqrt{n}}$ with high probability.

For the second term in (4), consider a random variable $T = (\frac{1}{\rho}A^{*T}_{\Gamma,i}A^*_{\bullet i} - 1)\alpha_i$ over $\Gamma$ and $\alpha_i$. We first observe for any vector $w \in \mathbb{R}^n$ that:

$$\begin{aligned}\mathbb{E}[(w_\Gamma^T w)^2] &= \sum_{i=1}^n \mathbb{E}[w_i^4\mathbf{1}_{i\in\Gamma}] + \sum_{i\neq j}^n \mathbb{E}[w_i^2 w_j^2\mathbf{1}_{i,j\in\Gamma}] \\ &= \rho(1-\rho)\sum_{i=1}^n w_i^4 + \rho^2.\end{aligned}$$

Hence, $T$ has mean 0 and variance $\sigma_T^2 = (1-\rho)/\rho\sum_{j=1}^n A_{ji}^4$, which is bounded by $O(\frac{1-\rho}{\rho n})$ because $\|A^*\|_{\max} \leq O(1/\sqrt{n})$. By Chebyshev's inequality, we have $|T| \leq \sqrt{\frac{1-\rho}{\rho n^{1/2}}}$ with failure probability $1/\sqrt{n}$. Combining everything, we get

$$|\beta_i - \alpha_i| \leq \frac{\mu k\log n}{\sqrt{n}} + \sqrt{\frac{1-\rho}{\rho n^{1/2}}},$$

w.h.p., which is the first part of the claim.

For the second part, we bound $\|\beta\|$ by expanding it as:

$$\|\beta\| = \frac{1}{\rho}\|A^{*T}_{\Gamma\bullet}A^*_{\bullet U}\alpha_U\| \leq \frac{1}{\rho}\|A^*_{\Gamma\bullet}\|\|A^*_{\bullet U}\|\|\alpha_U\|,$$

and again, if we use $\|\alpha_U\| \leq \sqrt{k}\log n$ w.h.p.and $\|A^*\| \leq O(1)$, then $\|\beta\| \leq \sqrt{k}\log n/\rho$. □

We briefly compare the above result with that of Arora et al. (2015). Our upper bounds are more general, and are stated in terms of the incompleteness factor $\rho$. Indeed,

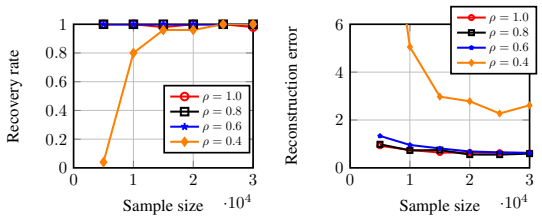


*Figure 1. (top) The performance of our approach on recovery rate and reconstruction error in sample size and sampling probability.*

our results match the previous bounds when $\rho = 1$. The above lemma suggests the following interesting regime of parameters. Specifically, for $\mu = O^*\big(\frac{\sqrt{n}}{k \log^3 n}\big)$ and $\frac{1}{\rho} - 1 \leq k \leq O^*(\frac{\rho\sqrt{n}}{\log n})$, one can see that $|\beta_i - \alpha_i| \leq O^*(1/\log^2 n)$ w.h.p., which implies that $\beta$ is a good estimate of $\alpha$ even when a subset of rows in $A^*$ is given.

In the next lemma, we show that that the pairwise correlation of $u$ and any sample $y$ is sufficiently informative for the same re-weighted spectral estimation strategy of Arora et al. (2015) to succeed in the incomplete setting.

**Lemma 3.** *Suppose that $u, v$ are a pair of fully observed samples and $y$ is an incomplete sample independent of $u, v$. The weighted covariance matrix $M_{u,v}$ has the form:*

$$M_{u,v} \triangleq \frac{1}{\rho^4}\mathbb{E}_y[\langle y,u\rangle\langle y,v\rangle yy^T] = \sum_{i\in U\cap V} q_i c_i \beta_i \beta_i' A^*_{\bullet i} A^{*T}_{\bullet i} + O^*(k/m \log n),$$

*where $c_i = \mathbb{E}[x_i^{*4} | i \in S]$ and $q_i = \mathbb{P}[i \in S]$.*

The complete proof is relegated to Appendix B. We will instead discuss some implications of this Lemma. Recall that $c_i$ is a constant with $0 < c < 1$ and $q_i = \Theta(k/m)$.

Suppose, for a moment, that the sparse representations of $u$ and $v$ share exactly one common dictionary element, say $A^*_{\bullet i}$ (i.e., if $U = \text{supp}(u)$ and $V = \text{supp}(v)$ then $U \cap V = \{i\}$.) The first term, $q_i c_i \beta_i \beta_i' A^*_{\bullet i} A^{*T}_{\bullet i}$, has norm $|q_i c_i \beta_i \beta_i'|$. From Claim 2, $|\beta_i| \geq |\alpha_i| - |\beta_i - \alpha_i| \geq C - o(1)$. Therefore, $q_i c_i \beta_i \beta_i' A^*_{\bullet i} A^{*T}_{\bullet i}$ has norm at least $\Omega(k/m)$ whereas the perturbation terms are at most $O^*(k/m \log n)$. According to Wedin's theorem, we conclude that the top singular vector of $M_{u,v}$ must be $O^*(k/m \log n)/\Omega(k/m) = O^*(1/\log n)$ -close to $A^*_{\bullet i}$. This gives us a coarse estimate of $A^*_{\bullet i}$.

The question remains when and how whether we can *a priori* certify whether $u, v$ share a unique dictionary atom among their sparse representations. Fortunately, the following Lemma provides a simple test for this via examining the decay of the singular vectors of the cross-covariance matrix $M_{u,v}$. The proof follows directly from that of Lemma 37 in (Arora et al., 2015).

**Lemma 4.** *When the top singular value of $M_{u,v}$ is at least $\Omega(k/m)$ and the second largest one is at most $O^*(k/m \log n)$, then $u$ and $v$ share a unique dictionary element with high probability.*

The above discussion isolates one of the columns of $A^*$. We can repeat this procedure several times by randomly choosing pairs of samples $u$ and $v$ from the hold-out set. Using the result of Arora et al. (2015), if $|\mathcal{P}_1|$ is $p_1 = \widetilde{O}(m)$, then we can estimate all the $m$ dictionary atoms. Overall, the sample complexity of Algorithm 2 is dominated by $p_2 = \widetilde{O}(mk/\rho^4)$.

## 5. Experiments

We corroborate our theory by demonstrating some representative numerical benefits of our proposed algorithms. We generate a synthetic dataset based on the generative model described in Section 2. The ground truth dictionary $A^*$ is of size $256 \times 256$ with independent standard Gaussian entries. We normalize columns of $A^*$ to be unit norm. Then, we generate 6-sparse code vectors $x^*$ with support drawn uniformly at random. Entries in the support are sampled from $\pm 1$ with equal probability. We generate all full samples, and isolate 5000 samples as "side information" for the initialization step. The remaining are then subsampled with different parameters $\rho$.

We set the number of iterations to $T = 3000$ in the initialization procedure and the number of descent steps $T = 50$ for the descent scheme. Besides, we slightly modify the thresholding operator in the encoding step of Algorithm 1. We use another operator that keeps $k$ largest entries of the input untouched and sets everything else to zero due to its stability. For each Monte Carlo trial, we uniformly draw $p$ partial samples. The task, for our algorithm, is to learn $A^*$. An implementation of our method is available online[3].

We evaluate our algorithm on two metrics against $p$ and $\rho$: (i) recovery rate, i.e., the fraction of trials in which each algorithm successfully recovers the ground truth $A^*$; and (ii) reconstruction error. All the metrics are averaged over 50 Monte Carlo simulations. "Successful recovery" is defined according to a threshold $\tau = 6$ on the Frobenius norm of the difference between the estimate $\widehat{A}$ and the ground truth $A^*$. (Since we can only estimate $\widehat{A}$ modulo a permutation and sign flip, the optimal column and sign matching is computed using the Hungarian algorithm.)

Figure 1 shows our experimental results. Here, sample size refers to the number of incomplete samples. Our algorithms are able to recover the dictionary for $\rho = 0.6, 0.8, 1.0$. For $\rho = 0.4$, we can observe a "phase transition" in sample complexity of successful recovery around $p = 10,000$ samples.

[3] https://github.com/thanh-isu

**Acknowledgements** The authors thank the anonymous reviewers for many insightful comments and suggestions during the review process. This work was supported in part by the National Science Foundation under grants CCF-1566281 and CCF-1750920, and in part by a Faculty Fellowship from the Black and Veatch Foundation.

## References

Agarwal, A., Anandkumar, A., and Netrapalli, P. Exact recovery of sparsely used overcomplete dictionaries. *IEEE Transactions on Information Theory*, 1050:8, 2013.

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

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

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

Arora, S., Ge, R., Ma, T., and Moitra, A. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pp. 113–149, 2015.

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2559–2566. IEEE, 2010.

Cai, T. T., Li, X., Ma, Z., et al. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Candes, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Chatterji, N. and Bartlett, P. Alternating minimization for dictionary learning with random initialization. 2017. arXiv:1711.03634v1.

Davenport, M. A., Laska, J. N., Boufounos, P. T., and Baraniuk, R. G. A simple proof that random matrices are democratic. *arXiv preprint arXiv:0911.0736*, 2009.

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

Eriksson, B., Balzano, L., and Nowak, R. High-rank matrix completion. In *Artificial Intelligence and Statistics*, pp. 373–381, 2012.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *International Conference on Machine Learning (ICML)*, pp. 399–406, 2010.

Gribonval, R., Jenatton, R., Bach, F., Kleinsteuber, M., and Seibert, M. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *ACM Symposium on Theory of Computing*, pp. 665–674. ACM, 2013.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15(1):1959–2008, 2014.

Loh, P.-L. and Wainwright, M. J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Neural Information Processing Systems*, pp. 2726–2734, 2011.

Natarajan, N. and Dhillon, I. S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

Naumova, V. and Schnass, K. Dictionary learning from incomplete data. *arXiv preprint arXiv:1701.03655*, 2017a.

Naumova, V. and Schnass, K. Dictionary learning from incomplete data for efficient image restoration. In *European Signal Processing Conference (EUSIPCO)*, pp. 1425–1429, Aug 2017b.

Nguyen, T. V., Wong, R. K. W., and Hegde, C. A provable approach for double-sparse coding. In *Proc. Conf. American Assoc. Artificial Intelligence (AAAI)*, Feb. 2018.

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

Rubinstein, R., Bruckstein, A. M., and Elad, M. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

Soni, A., Jain, S., Haupt, J., and Gonella, S. Noisy matrix completion under sparse factor models. *IEEE Transactions on Information Theory*, 62(6):3636–3661, June 2016.

Spielman, D. A., Wang, H., and Wright, J. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pp. 37–1, 2012.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery using nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pp. 2351–2360, 2015.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973, 2016.

Xing, Z., Zhou, M., Castrodad, A., Sapiro, G., and Carin, L. Dictionary learning for noisy and incomplete hyperspectral images. *SIAM Journal on Imaging Sciences*, 5 (1):33–56, 2012.

Yuan, X.-T. and Zhang, T. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(Apr):899–925, 2013.