Journal of Machine Learning Research 20 (2019) 1-43 Submitted 12/17; Revised 5/19; Published 9/19

Provably Accurate Double-Sparse Coding

Thanh V. Nguyen THANHNG@IASTATE.EDU
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011, USA

Raymond K. W. Wong RAYWONG@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Chinmay Hegde CHINMAY@IASTATE.EDU
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011, USA

Editor: Animashree Anandkumar

Abstract

Sparse coding is a crucial subroutine in algorithms for various signal processing, deep learning, and other machine learning applications. The central goal is to learn an overcomplete dictionary that can sparsely represent a given input dataset. However, a key challenge is that storage, transmission, and processing of the learned dictionary can be untenably high if the data dimension is high. In this paper, we consider the double-sparsity model introduced by Rubinstein et al. (2010b) where the dictionary itself is the product of a fixed, known basis and a data-adaptive sparse component. First, we introduce a simple algorithm for double-sparse coding that can be amenable to efficient implementation via neural architectures. Second, we theoretically analyze its performance and demonstrate asymptotic sample complexity and running time benefits over existing (provable) approaches for sparse coding. To our knowledge, our work introduces the first computationally efficient algorithm for double-sparse coding that enjoys rigorous statistical guarantees. Finally, we corroborate our theory with several numerical experiments on simulated data, suggesting that our method may be useful for problem sizes encountered in practice.

Keywords: Sparse coding, provable algorithms, unsupervised learning

1. Introduction

1.1 Motivation

Representing signals as sparse linear combinations of atoms from a dictionary is a popular approach in many domains. In this paper, we study the problem of *dictionary learning* (also known as sparse coding), where the goal is to learn an efficient basis (dictionary) that represents the underlying class of signals well. In the typical sparse coding setup, the dictionary is *overcomplete* (i.e., the cardinality of the dictionary exceeds the ambient signal dimension) while the representation is *sparse* (i.e., each signal is encoded by a combination of only a few dictionary atoms.)

Sparse coding has a rich history in diverse fields such as signal processing, machine learning, and computational neuroscience. Discovering optimal basis representations of data is a central focus of image analysis (Krim et al., 1999; Elad and Aharon, 2006; Rubinstein et al., 2010a), and dictionary learning has proven widely successful in imaging problems such as denoising, deconvolution, inpainting, and compressive sensing (Elad and Aharon, 2006; Candes and Tao, 2005; Rubinstein et al., 2010a). Sparse coding approaches have also been used as a core building block of deep learning systems for prediction (Gregor and LeCun, 2010; Boureau et al., 2010) and associative memory (Mazumdar and Rawat, 2017). Interestingly, the seminal work by Olshausen and Field (1997) has shown intimate connections between sparse coding and neuroscience: the dictionaries learned from image patches of natural scenes bear remarkable resemblance to spatial receptive fields observed in the mammalian primary visual cortex.

From a mathematical standpoint, the sparse coding problem is formulated as follows. Given p data samples $Y = [y^{(1)}, y^{(2)}, \ldots, y^{(p)}] \in \mathbb{R}^{n \times p}$, the goal is to find a dictionary $D \in \mathbb{R}^{n \times m}$ $(m > n)$ and corresponding sparse code vectors $X = [x^{(1)}, x^{(2)}, \ldots, x^{(p)}] \in \mathbb{R}^{m \times p}$ such that the representation DX fits the data samples as well as possible. Typically, one obtains the dictionary and the code vectors as the solution to the following optimization problem:

$$\begin{aligned} \min_{D,X} \mathcal{L}(D, X) &= \frac{1}{2} \sum_{j=1}^{p} \|y^{(j)} - Dx^{(j)}\|_2^2, \\ \text{s.t. } & \sum_{j=1}^{p} \mathcal{S}(x^{(j)}) \leq S \end{aligned} \tag{1}$$

where $\mathcal{S}(\cdot)$ is some sparsity-inducing penalty function on the code vectors, such as the ℓ_1-norm. The objective function $\mathcal{L}$ controls the reconstruction error while the constraint enforces the sparsity of the representation. However, even a cursory attempt at solving the optimization problem (1) reveals the following obstacles:

The constrained optimization problem (1) involves a non-convex (in fact, bilinear) objective function, as well as potentially non-convex constraints depending on the choice of the sparsity-promoting function $\mathcal{S}$ (for example, the ℓ_0 function.) Hence, obtaining *provably* correct algorithms for this problem can be challenging. Indeed, the vast majority of practical approaches for sparse coding have been heuristics (Engan et al., 1999; Aharon et al., 2006; Mairal et al., 2009). Recent works in the theoretical machine learning community have bucked this trend, providing provably accurate algorithms if certain assumptions are satisfied (Spielman et al., 2012; Agarwal et al., 2014; Arora et al., 2014a, 2015; Sun et al., 2015; Błasiok and Nelson, 2016; law Adamczak, 2016; Chatterji and Bartlett, 2017). However, relatively few of these newer methods have been shown to provide good empirical performance in actual sparse coding problems.

Even if theoretical correctness issues were to be set aside, and we are somehow able to efficiently learn sparse codes of the input data, we often find that applications using such learned sparse codes encounter *memory* and *running-time* issues. Indeed, in the overcomplete case, the storage of the learned dictionary D incurs $mn = \Omega(n^2)$ memory cost, which is prohibitive when n is large. Therefore, in practical applications (such as image analysis)

one typically resorts to chop the data into smaller blocks (e.g., partitioning image data into patches) to make the problem manageable.

A related line of research has been devoted to learning dictionaries that obey some type of *structure*. Such structural information can be leveraged to incorporate prior knowledge of underlying signals as well as to resolve computational challenges due to the data dimension. For instance, the dictionary is assumed to be separable, or obey a convolutional structure. One such variant is the *double-sparse* coding problem (Rubinstein et al., 2010b; Sulam et al., 2016) where the dictionary D *itself* exhibits a sparse structure. To be specific, the dictionary is expressed as:

$$D = \Phi A,$$

i.e., it is composed of a known "base dictionary" $\Phi \in \mathbb{R}^{n\times n}$, and a learned "synthesis" matrix $A \in \mathbb{R}^{n\times m}$ whose columns are sparse. The base dictionary Φ is typically any fixed basis chosen according to domain knowledge, while the synthesis matrix A is column-wise sparse and is to be learned from the data. The basis Φ is typically orthonormal (such as the canonical or wavelet basis); however, there are cases where the base dictionary Φ is overcomplete (Rubinstein et al., 2010b; Sulam et al., 2016).

There are several reasons why such the double-sparsity model can be useful. First, the double-sparsity assumption is rather appealing from a conceptual standpoint, since it lets us combine the knowledge of decades of modeling efforts in harmonic analysis with the flexibility of learning new representations tailored to specific data families. Moreover, such a double-sparsity model has computational benefits. If the columns of A are (say) r-sparse (i.e., each column contains no more than $r \ll n$ non-zeroes) then the overall burden of storing, transmitting, and computing with A is much lower than that for general unstructured dictionaries. Finally, such a model lends itself well to *interpretable* learned features if the atoms of the base dictionary are semantically meaningful.

All the above reasons have spurred researchers to develop a series of algorithms to learn doubly-sparse codes (Rubinstein et al., 2010b; Sulam et al., 2016). However, despite their empirical promise, no theoretical analysis of their performance have been reported in the literature and to date, we are unaware of a provably accurate, polynomial-time algorithm for the double-sparse coding problem. Our goal in this paper is precisely to fill this gap.

1.2 Our Contributions

In this paper, we provide a new framework for double-sparse coding. To the best of our knowledge, our approach is the first method that enjoys *provable* statistical and algorithmic guarantees for this problem. In addition, our approach enjoys two benefits: we demonstrate that the method is *neurally plausible* (i.e., its execution can plausibly be achieved using a neural network architecture) and *robust* to noise.

Inspired by the aforementioned recent theoretical advances in sparse coding, we assume a learning-theoretic setup where the data samples arise from a ground-truth generative model. Informally, suppose there exists a true (but unknown) synthesis matrix A^* that is column-wise r-sparse, and the i^{th} data sample is generated as:

$$y^{(i)} = \Phi A^* x^{*(i)} + \text{ noise}, \quad i = 1, 2, \ldots, p$$

where the code vector $x^{*(i)}$ is independently drawn from a distribution supported on the set of k-sparse vectors. We desire to learn the underlying matrix A^*. Informally, suppose that the synthesis matrix A^* is *incoherent* (the columns of A^* are sufficiently close to orthogonal) and has bounded spectral norm. Finally, suppose that the number of dictionary elements, m, is at most a constant multiple of n. All of these assumptions are standard[1].

We will demonstrate that the true synthesis matrix A^* can be recovered (with small error) in a tractable manner as sufficiently many samples are provided. Specifically, we make the following novel contributions:

1. We propose a new algorithm that produces a coarse estimate of the synthesis matrix that is sufficiently close to the ground truth A^*. Our algorithm builds upon spectral initialization-based ideas that have recently gained popularity in non-convex machine learning (Zhang et al., 2016; Wang et al., 2016).

2. Given the above coarse estimate of the synthesis matrix A^*, we propose a descent-style algorithm to refine the above estimate of A^*. This algorithm is simpler than previously studied double-sparse coding algorithms (such as the Trainlets approach of Sulam et al. (2016)), while still giving good statistical performance. Moreover, this algorithm can be realized in a manner amenable to neural implementations.

3. We provide a rigorous analysis of both algorithms. Put together, our analysis produces the first provably polynomial-time algorithm for double-sparse coding. We show that the algorithm provably returns a good estimate of the ground-truth; in particular, in the absence of noise we prove that $\Omega(mr \text{ polylog } n)$ samples are sufficient for a good enough initialization in the first algorithm, as well as guaranteed linear convergence of the descent phase up to a precise error parameter that can be interpreted as the radius of convergence.

   Indeed, our analysis shows that employing the double-sparsity model helps in this context, and leads to a strict improvement in sample complexity, as well as running time over previous rigorous methods for (regular) sparse coding such as Arora et al. (2015).

4. We also analyze our approach in a more realistic setting with the presence of additive noise and demonstrate its stability. We prove that $\Omega(mr \text{ polylog } n)$ samples are sufficient to obtain a good enough estimate in the initialization, and also to obtain guaranteed linear convergence during descent to provably recover A^*.

5. We underline the benefit of the double-sparse structure over the regular model by analyzing the algorithms in Arora et al. (2015) under the noisy setting. As a result, we obtain the sample complexity $O\big((mk + \sigma_\varepsilon^2 \frac{mn^2}{k})\text{polylog } n\big)$, which demonstrates a negative effect of noise on this approach.

6. We rigorously develop a hard thresholding intialization that extends the spectral scheme in Arora et al. (2015). Additionally, we provide more results for the case where A is orthonormal, sparse dictionary to relax the condition on r, which may be of independent interest.

1. We clarify both the data and the noise model more concretely in Section 2 below.

Setting	Reference	Sample (w/o noise)	Sample (w/ noise)	Time	Expt
Regular	MOD (Engan et al., 1999)	✗	✗	✗	✓
	K-SVD (Aharon et al., 2006)	✗	✗	✗	✓
	Spielman et al. (2012)	$O(n^2 \log n)$	✗	$\widetilde{\Omega}(n^4)$	✓
	Arora et al. (2014b)	$\widetilde{O}(m^2/k^2)$	✗	$\widetilde{O}(np^2)$	✗
	Gribonval et al. (2015a)	$O(nm^3)$	$O(nm^3)$	✗	✗
	Arora et al. (2015)	$\widetilde{O}(mk)$	✗	$\widetilde{O}(mn^2p)$	✗
Double Sparse	Double Sparsity (Rubinstein et al., 2010b)	✗	✗	✗	✓
	Gribonval et al. (2015b)	$\widetilde{O}(mr)$	$\widetilde{O}(mr)$	✗	✗
	Trainlets (Sulam et al., 2016)	✗	✗	✗	✓
	This paper	$\widetilde{O}(mr)$	$\widetilde{O}(mr + \sigma_\varepsilon^2 \frac{mnr}{k})$	$\widetilde{O}(mnp)$	✓

Table 1: *Comparison of various sparse coding techniques. Expt: whether numerical experiments have been conducted.* ✗ *in all other columns indicates no provable guarantees. Here,* n *is the signal dimension, and* m *is the number of atoms. The sparsity levels for* A *and* x *are* r *and* k *respectively, and* p *is the sample size.*

7. While our analysis mainly consists of sufficiency results and involves unknown constants hidden in big-O notation, we demonstrate our findings by reporting a suite of numerical experiments on synthetic test datasets.

Overall, our approach results in strict improvement in sample complexity, as well as running time, over previous rigorously analyzed methods for (regular) sparse coding, such as Arora et al. (2015). See Table 1 for a detailed comparison.

1.3 Techniques

At a high level, our method is an adaptation of the seminal approach of Arora et al. (2015). As is common in the statistical learning literature, we assume a "ground-truth" generative model for the observed data samples, and attempt to estimate the parameters of the generative model given a sufficient number of samples. In our case, the parameters correspond to the synthesis matrix A^*, which is column-wise r-sparse. The natural approach is to formulate a loss function in terms of A such as Equation (1), and perform gradient descent with respect to the surface of the loss function to learn A^*.

The key challenge in sparse coding is that the gradient is inherently coupled with the *codes* of the training samples (i.e., the columns of X^*), which are unknown *a priori*. However, the main insight of Arora et al. (2015) is that within a small enough neighborhood of A^*, a noisy version of X^* can be estimated, and therefore the overall method is similar to performing *approximate gradient descent.* Formulating the actual algorithm as a noisy variant of approximate gradient descent allows us to overcome the finite-sample variability

of the loss, and obtain a descent property directly related to (the population parameter) A^*.

The second stage of our approach (i.e., our descent-style algorithm) leverages this intuition. However, instead of standard gradient descent, we perform approximate *projected* gradient descent, such that the column-wise r-sparsity property is enforced in each new estimate of A^*. Indeed, such an extra projection step is critical in showing a sample complexity improvement over the existing approach of Arora et al. (2015).The key novelty is in figuring out how to perform the projection in each gradient iteration. For this purpose, we develop a novel initialization algorithm that identifies the locations of the non-zeroes in A^* even before commencing the descent phase. This is nontrivially different from initialization schemes used in previous rigorous methods for sparse coding, and the analysis is somewhat more involved.

In Arora et al. (2015), (the principal eigenvector of) a weighted covariance matrix of y (estimated by the weighted average of outer products $y_i y_i^T$) is shown to provide a coarse estimate of a dictionary atom. We extend this idea and rigoriously show that the diagonal of the weighted covariance matrix serves as a good indicator of the support of a column in A^*. The success relies on the concentration of the diagonal vector with dimension n, instead of the covariance matrix with dimensions $n \times n$. With the support selected, our scheme only utilizes a reduced weighted covariance matrix with dimensions at most $r \times r$. This initialization scheme enables us to effectively reduce the dimension of the problem, and therefore leads to significant improvement in sample complexity and running time over previous (provable) sparse coding methods when the data representation sparsity k is much smaller than m.

Further, we rigorously analyze the proposed algorithms in the presence of noise with a bounded expected norm. Our analysis shows that our method is stable, and in the case of i.i.d. Gaussian noise with bounded expected ℓ_2-norms, is at least a polynomial factor better than previous polynomial time algorithms for sparse coding.

The empirical performance of our proposed method is demonstrated by a suite of numerical experiments on synthetic datasets. In particular, we show that our proposed methods are simple and practical, and improve upon previous provable algorithms for sparse coding.

1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 introduces notation, key model assumptions, and informal statements of our main theoretical results. Section 3 outlines our initialization algorithm (along with supporting theoretical results) while Section 4 presents our descent algorithm (along with supporting theoretical results). Section 5 provides a numerical study of the efficiency of our proposed algorithms, and compares it with previously proposed methods. Finally, Section 6 concludes with a short discussion. All technical proofs are relegated to the appendix.

2. Setup and Main Results

2.1 Notation

We define $[m] \triangleq \{1, \ldots, m\}$ for any integer $m > 1$. For any vector $x = [x_1, x_2, \ldots, x_m]^T \in \mathbb{R}^m$, we write $\mathrm{supp}(x) \triangleq \{i \in [m] : x_i \neq 0\}$ as the support set of x. Given any subset $S \subseteq [m]$, x_S corresponds to the sub-vector of x indexed by the elements of S. For any matrix $A \in \mathbb{R}^{n \times m}$, we use $A_{\bullet i}$ and $A_{j\bullet}^T$ to represent the i-th column and the j-th row respectively. For some appropriate sets R and S, let $A_{R\bullet}$ (respectively, $A_{\bullet S}$) be the submatrix of A with rows (respectively columns) indexed by the elements in R (respectively S). In addition, for the i-th column $A_{\bullet i}$, we use $A_{R,i}$ to denote the sub-vector indexed by the elements of R. For notational simplicity, we use $A_{R\bullet}^T$ to indicate $(A_{R\bullet})^T$, the tranpose of A after a row selection. Besides, we use $\circ$ and $\mathrm{sgn}(\cdot)$ to represent the element-wise Hadamard operator and the element-wise sign function respectively. Further, $\mathrm{threshold}_K(x)$ is a thresholding operator that replaces any elements of x with magnitude less than K by zero.

The ℓ_2-norm $\|x\|$ for a vector x and the spectral norm $\|A\|$ for a matrix A appear several times. In some cases, we also utilize the Frobenius norm $\|A\|_F$ and the operator norm $\|A\|_{1,2} \triangleq \max_{\|x\|_1 \leq 1} \|Ax\|$. The norm $\|A\|_{1,2}$ is essentially the maximal Euclidean norm of any column of A.

For clarity, we adopt asymptotic notations extensively. We write $f(n) = O(g(n))$ (or $f(n) = \Omega(g(n))$) if $f(n)$ is upper bounded (respectively, lower bounded) by $g(n)$ up to some positive constant. Next, $f(n) = \Theta(g(n))$ if and only if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. Also $\widetilde{\Omega}$ and $\widetilde{O}$ represent Ω and O up to a multiplicative poly-logarithmic factor respectively. Finally $f(n) = o(g(n))$ (or $f(n) = \omega(g(n))$) if $\lim_{n\to\infty} |f(n)/g(n)| = 0$ ($\lim_{n\to\infty} |f(n)/g(n)| = \infty$).

Throughout the paper, we use the phrase "with high probability" (abbreviated to w.h.p.) to describe an event with failure probability of order at most $n^{-\omega(1)}$. In addition, $g(n) = O^*(f(n))$ means $g(n) \leq Kf(n)$ for some small enough constant K.

2.2 Generative Model of Data

Suppose that the observed samples are given by

$$y^{(i)} = Dx^{*(i)} + \varepsilon, \quad i = 1, \ldots, p,$$

i.e., we are given p samples of y generated from a fixed (but unknown) dictionary D where the sparse code x^* and the error ε are drawn from a joint distribution $\mathcal{D}$ specified below. In the double-sparse setting, the dictionary is assumed to follow a decomposition $D = \Phi A^*$, where $\Phi \in \mathbb{R}^{n \times n}$ is a known *orthonormal* basis matrix and A^* is an unknown, ground truth synthesis matrix. An alternative (and interesting) setting is an overcomplete Φ with a square A^*, which our analysis below does not cover; we defer this to future work. Our approach relies upon the following assumptions on the synthesis dictionary A^*:

- **A1** A^* is overcomplete (i.e., $m \geq n$) with $m = O(n)$.
- **A2** A^* is μ-incoherent, i.e., for all $i \neq j$, $|\langle A_{\bullet i}^*, A_{\bullet j}^* \rangle| \leq \mu/\sqrt{n}$.
- **A3** $A_{\bullet i}^*$ has at most r non-zero elements, and is normalized such that $\|A_{\bullet i}^*\| = 1$ for all i. Moreover, $|A_{ij}^*| \geq \tau$ for $A_{ij}^* \neq 0$ and $\tau = \Omega(1/\sqrt{r})$.

A4 A^* has bounded spectral norm such that $\|A^*\| \leq O(\sqrt{m/n})$.

All these assumptions are standard. In Assumption **A2**, the incoherence μ is typically of order $O(\log n)$ with high probability for a normal random matrix (Arora et al., 2014b). Assumption **A3** is a common assumption in sparse signal recovery. The bounded spectral norm assumption is also standard (Arora et al., 2015). In addition to Assumptions **A1-A4**, we make the following distributional assumptions on $\mathcal{D}$:

B1 Support $S = \text{supp}(x^*)$ is of size at most k and uniformly drawn without replacement from $[m]$ such that $\mathbb{P}[i \in S] = \Theta(k/m)$ and $\mathbb{P}[i, j \in S] = \Theta(k^2/m^2)$ for some $i, j \in [m]$ and $i \neq j$.

B2 The nonzero entries x^*_S are pairwise independent and sub-Gaussian given the support S with $\mathbb{E}[x^*_i | i \in S] = 0$ and $\mathbb{E}[x^{*2}_i | i \in S] = 1$.

B3 For $i \in S$, $|x^*_i| \geq C$ where $0 < C \leq 1$.

B4 The additive noise ε has i.i.d. Gaussian entries with variance σ^2_ε with $\sigma_\varepsilon = O(1/\sqrt{n})$.

For the rest of the paper, we set $\Phi = I_n$, the identity matrix of size n. This only simplifies the arguments but does not change the problem because one can study an equivalent model:

$$y' = Ax^* + \varepsilon',$$

where $y' = \Phi^T y$ and $\varepsilon' = \Phi^T \varepsilon$, as $\Phi^T \Phi = I_n$. Due to the Gaussianity of ε, ε' also has independent entries. Although this property is specific to Gaussian noise, all the analysis carried out below can be extended to sub-Gaussian noise with minor (but rather tedious) changes in concentration arguments.

Our goal is to devise an algorithm that produces a provably "good" estimate of A^*. For this, we need to define a suitable measure of "goodness". We use the following notion of distance that measures the maximal column-wise difference in ℓ_2-norm under some suitable transformation.

Definition 1 ((δ, κ)-nearness) *A is said to be δ-close to A^* if there is a permutation $\pi : [m] \to [m]$ and a sign flip $\sigma : [m] : \{\pm 1\}$ such that $\|\sigma(i) A_{\bullet \pi(i)} - A^*_{\bullet i}\| \leq \delta$ for every i. In addition, A is said to be (δ, κ)-near to A^* if $\|A_{\bullet \pi} - A^*\| \leq \kappa \|A^*\|$ also holds.*

For notational simplicity, in our theorems we simply replace π and σ in Definition 1 with the identity permutation $\pi(i) = i$ and the positive sign $\sigma(\cdot) = +1$ while keeping in mind that in reality we are referring to one element of the equivalence class of all permutations and sign flip transforms of A^*.

We will also need some technical tools from Arora et al. (2015) to analyze our gradient descent-style method. Consider any iterative algorithm that looks for a desired solution $z^* \in \mathbb{R}^n$ to optimize some function $f(z)$. Suppose that the algorithm produces a sequence of estimates $z^1, \ldots, z^s$ via the update rule:

$$z^{s+1} = z^s - \eta g^s,$$

for some vector g^s and scalar step size η. The goal is to characterize "good" directions g^s such that the sequence converges to z^* under the Euclidean distance. The following gives one such sufficient condition for g^s.

Definition 2 *A vector g^s at the s^{th} iteration is $(\alpha, \beta, \gamma_s)$-correlated with a desired solution z^* if*

$$\langle g^s, z^s - z^* \rangle \geq \alpha \|z^s - z^*\|^2 + \beta \|g^s\|^2 - \gamma_s.$$

We know from convex optimization that if f is 2α-strongly convex and $1/2\beta$-smooth, and g^s is chosen as the gradient $\nabla_z f(z)$, then g^s is $(\alpha, \beta, 0)$-correlated with z^*. In our setting, the desired solution corresponds to A^*, the ground-truth synthesis matrix. In Arora et al. (2015), it is shown that $g^s = \mathbb{E}_y[(A^s x - y)\mathrm{sgn}(x)^T]$, where $x = \mathrm{threshold}_{C/2}((A^s)^T y)$ indeed satisfies Definition 2. This g^s is a population quantity and not explicitly available, but one can estimate such g^s using an empirical average. The corresponding estimator $\widehat{g}^s$ is a random variable, so we also need a related *correlated-with-high-probability* condition:

Definition 3 *A direction $\widehat{g}^s$ at the s^{th} iteration is $(\alpha, \beta, \gamma_s)$-correlated-w.h.p. with a desired solution z^* if, w.h.p.,*

$$\langle \widehat{g}^s, z^s - z^* \rangle \geq \alpha \|z^s - z^*\|^2 + \beta \|\widehat{g}^s\|^2 - \gamma_s.$$

From Definition 2, one can establish a form of descent property in each update step, as shown in Theorem 1.

Theorem 1 *Suppose that g^s satisfies the condition described in Definition 2 for $s = 1, 2, \ldots, T$. Moreover, $0 < \eta \leq 2\beta$ and $\gamma = \max_{s=1}^T \gamma_s$. Then, the following holds for all s:*

$$\|z^{s+1} - z^*\|^2 \leq (1 - 2\alpha\eta)\|z^s - z^*\|^2 + 2\eta\gamma_s.$$

In particular, the above update converges geometrically to z^* with an error γ/α. That is,

$$\|z^{s+1} - z^*\|^2 \leq (1 - 2\alpha\eta)^s \|z^0 - z^*\|^2 + 2\gamma/\alpha.$$

We can obtain a similar result for Definition 3 except that $\|z^{s+1} - z^*\|^2$ is replaced with its expectation.

Armed with the above tools, we now state some informal versions of our main results:

Theorem 2 (Provably correct initialization, informal) *There exists a neurally plausible algorithm to produce an initial estimate A^0 that has the correct support and is $(\delta, 2)$-near to A^* with high probability. Its running time and sample complexity are $\widetilde{O}(mnp)$ and $\widetilde{O}(mr)$ respectively. This algorithm works when the sparsity level satisfies $r = O^*(\log n)$.*

Our algorithm can be regarded as an extension of Arora et al. (2015) to the double-sparse setting. It reconstructs the support of one single column and then estimates its direction in the subspace defined by the support. Our proposed algorithm enjoys neural plausibility by implementing a thresholding non-linearity and Oja's update rule. We provide a neural implementation of our algorithm in Appendix G. The adaption to the sparse structure results in a strict improvement upon the original algorithm both in running time and sample complexity. However, our algorithm is limited to the sparsity level $r = O^*(\log n)$, which is rather small but plausible from the modeling standpoint. For comparison, we analyze a natural extension of the algorithm of Arora et al. (2015) with an extra hard-thresholding

step for every learned atom. We obtain the same order restriction on r, but somewhat worse bounds on sample complexity and running time. The details are found in Appendix F.

We hypothesize that a stronger incoherence assumption can lead to provably correct initialization for a much wider range of r. For purposes of theoretical analysis, we consider the special case of a *perfectly incoherent* synthesis matrix A^* such that $\mu = 0$ and $m = n$. In this case, we can indeed improve the sparsity parameter to $r = O^*\big(\min(\frac{\sqrt{n}}{\log^2 n}, \frac{n}{k^2 \log^2 n})\big)$, which is an exponential improvement. This analysis is given in Appendix E.

The next theorem summarizes our result for the descent algorithm:

Theorem 3 (Provably correct descent, informal) *There exists a neurally plausible algorithm for double-sparse coding that converges to A^* with geometric rate when the initial estimate A^0 has the correct support and $(\delta, 2)$-near to A^*. The running time per iteration is $O(mkp + mrp)$ and the sample complexity is $\widetilde{O}(m + \sigma_\varepsilon^2 \frac{mnr}{k})$.*

Similar to Arora et al. (2015), our proposed algorithm enjoys neural plausibility. Moreover, we can achieve a better running time and sample complexity per iteration than previous methods, particularly in the noisy case. We show in Appendix F that in this regime the sample complexity of Arora et al. (2015) is $\widetilde{O}(m + \sigma_\varepsilon^2 \frac{mn^2}{k})$. For instance, when $\sigma_\varepsilon \asymp n^{-1/2}$, the sample complexity bound is significantly worse than $\widetilde{O}(m)$ in the noiseless case. In contrast, our proposed method leverages the sparse structure to overcome this problem and obtain improved results.

We are now ready to introduce our methods in detail. As discussed above, our approach consists of two stages: an initialization algorithm that produces a coarse estimate of A^*, and a descent-style algorithm that refines this estimate to accurately recover A^*.

3. Stage 1: Initialization

In this section, we present a neurally plausible algorithm that can produce a coarse initial estimate of the ground truth A^*. We give a neural implementation of the algorithm in Appendix G.

Our algorithm is an adaptation from the algorithm in Arora et al. (2015). The idea is to estimate dictionary atoms in a greedy fashion by iteratively re-weighting the given samples. The samples are re-scaled in a way that the weighted (sample) covariance matrix has the dominant first singular value, and its corresponding eigenvector is close to one particular atom with high probability. However, while this algorithm is conceptually very appealing, it incurs severe computational costs in practice. More precisely, the overall running time is $\widetilde{O}(mn^2p)$ in expectation, which is unrealistic for large-scale problems.

To overcome this burden, we leverage the double-sparsity assumption in our generative model to obtain a more efficient approach. The high-level idea is to first estimate the support of each column in the synthesis matrix A^*, and then obtain a coarse estimate of the nonzero coefficients of each column based on knowledge of its support. The key ingredient of our method is a novel spectral procedure that gives us an estimate of the column supports purely from the observed samples. The full algorithm, that we call *Truncated Pairwise Reweighting*, is listed in pseudocode form as Algorithm 1.

Algorithm 1 Truncated Pairwise Reweighting

Initialize $L = \emptyset$

Randomly divide p samples into two disjoint sets $\mathcal{P}_1$ and $\mathcal{P}_2$ of sizes p_1 and p_2 respectively

While $|L| < m$. Pick u and v from $\mathcal{P}_1$ at random

For every $l = 1, 2, \ldots, n$; compute

$$\widehat{e}_l = \frac{1}{p_2} \sum_{i=1}^{p_2} \langle y^{(i)}, u \rangle \langle y^{(i)}, v \rangle (y_l^{(i)})^2$$

Sort $(\widehat{e}_1, \widehat{e}_2, \ldots, \widehat{e}_n)$ in descending order

If $r' \leq r$ s.t $\widehat{e}_{(r')} \geq O(k/mr)$ and $\widehat{e}_{(r'+1)}/\widehat{e}_{(r')} < O^*(r/\log^2 n)$

Let $\widehat{R}$ be set of the r' largest entries of $\widehat{e}$

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

$\delta_1, \delta_2 \leftarrow$ top singular values of $\widehat{M}_{u,v}$

$z_{\widehat{R}} \leftarrow$ top singular vector of $\widehat{M}_{u,v}$

If $\delta_1 \geq \Omega(k/m)$ and $\delta_2 < O^*(k/m \log n)$

If $\text{dist}(\pm z, l) > 1/\log n$ for any $l \in L$

Update $L = L \cup \{z\}$

Return $A^0 = (L_1, \ldots, L_m)$

Let us provide some intuition of our algorithm. Fix a sample $y = A^* x^* + \varepsilon$ from the available training set, and consider samples

$$u = A^* \alpha + \varepsilon_u, v = A^* \alpha' + \varepsilon_v.$$

Now, consider the (very coarse) estimate for the sparse code of u with respect to A^*:

$$\beta = A^{*T} u = A^{*T} A^* \alpha + A^{*T} \varepsilon_u.$$

As long as A^* is incoherent enough and ε_u is small, the estimate β behaves just like α, in the sense that for each sample y:

$$\langle y, u \rangle \approx \langle x^*, \beta \rangle \approx \langle x^*, \alpha \rangle.$$

Moreover, the above inner products are large only if α and x^* share some elements in their supports; else, they are likely to be small. Likewise, the weight $\langle y, u \rangle \langle y, v \rangle$ depends on whether or not x^* shares the support with both α and α'.

Now, suppose that we have a mechanism to isolate pairs u and v who share exactly one atom among their sparse representations. Then by scaling each sample y with an increasing function of $\langle y, u \rangle \langle y, v \rangle$ and linearly adding the samples, we magnify the importance of the samples that are aligned with that atom, and diminish the rest. The final direction can be obtained via the top *principal component* of the reweighted samples and hence can be used as a coarse estimate of the atom. This is exactly the approach adopted in Arora et al. (2015). However, in our double-sparse coding setting, we know that the estimated atom

should be sparse as well. Therefore, we can naturally perform an extra "sparsification" step of the output. An extended algorithm and its correctness are provided in Appendix F. However, as we discussed above, the computational complexity of the re-weighting step still remains.

We overcome this obstacle by first identifying the locations of the nonzero entries in each atom. Specifically, define the matrix:

$$M_{u,v} = \frac{1}{p_2}\sum_{i=1}^{p_2}\langle y^{(i)}, u\rangle\langle y^{(i)}, v\rangle y^{(i)}y^{(i)T}.$$

Then, the diagonal entries of $M_{u,v}$ reveals the support of the atom of A^* shared among u and v: the r-largest entries of $M_{u,v}$ will correspond to the support we seek. Since the desired direction remains unchanged in the r-dimensional subspace of its nonzero elements, we can restrict our attention to this subspace, construct a reduced covariance matrix $\widehat{M}_{u,v}$, and proceed as before. This truncation step alleviates the computational burden by a significant amount; the running time is now $\widetilde{O}(mnp)$, which improves the original by a factor of n.

The success of the above procedure relies upon whether or not we can isolate pairs u and v that share one dictionary atom. Fortunately, this can be done via checking the decay of the singular values of the (reduced) covariance matrix. Here too, we show via our analysis that the truncation step plays an important role. Overall, our proposed algorithm not only accelerates the initialization in terms of running time, but also improves the sample complexity over Arora et al. (2015). The performance of Algorithm 1 is described in the following theorem, whose formal proof is deferred to Appendix B.

Theorem 4 *Suppose that Assumptions* **B1-B4** *hold and Assumptions* **A1-A3** *satify with* $\mu = O^*\big(\frac{\sqrt{n}}{k\log^3 n}\big)$ *and* $r = O^*(\log n)$. *When* $p_1 = \widetilde{\Omega}(m)$ *and* $p_2 = \widetilde{\Omega}(mr)$, *then with high probability Algorithm 1 returns an initial estimate* A^0 *whose columns share the same support as* A^* *and with* $(\delta, 2)$*-nearness to* A^* *with* $\delta = O^*(1/\log n)$.

The limit on r arises from the minimum non-zero coefficient τ of A^*. Since the columns of A^* are standardized, τ should degenerate as r grows. In other words, it is getting harder to distinguish the "signal" coefficients from zero as r grows with n. However, this limitation can be relaxed when a better incoherence available, for example the orthonormal case. We study this in Appendix E.

To provide some intuition about the working of the algorithm (and its proof), let us analyze it in the case where we have access to infinite number of samples. This setting, of course, is unrealistic. However, the analysis is much simpler and more transparent since we can focus on expected values rather than empirical averages. Moreover, the analysis reveals several key lemmas, which we will reuse extensively for proving Theorem 4. First, we give some intuition behind the definition of the "scores", $\widehat{e}_l$.

Lemma 1 *Fix samples* u *and* v *and suppose that* $y = A^*x^* + \varepsilon$ *is a random sample independent of* u, v. *The expected value of the score for the* ℓ^{th} *component of* y *is given by:*

$$e_l \triangleq \mathbb{E}[\langle y, u\rangle\langle y, v\rangle y_l^2] = \sum_{i\in U\cap V} q_i c_i \beta_i \beta_i' A_{li}^{*2} + \text{ perturbation terms}$$

where $q_i = \mathbb{P}[i \in S]$, $q_{ij} = \mathbb{P}[i, j \in S]$ *and* $c_i = \mathbb{E}[x_i^4 | i \in S]$. *Moreover, the perturbation terms have absolute value at most* $O^*(k/m \log n)$.

From Assumption **B1**, we know that $q_i = \Theta(k/m)$, $q_{ij} = \Theta(k^2/m^2)$ and $c_i = \Theta(1)$. Besides, we will show later that $|\beta_i| \approx |\alpha_i| = \Omega(1)$ for $i \in U$, and $|\beta_i| = o(1)$ for $i \notin U$. Consider the first term $E_0 = \sum_{i \in U \cap V} q_i c_i \beta_i \beta_i' A_{li}^{*2}$. Clearly, $E_0 = 0$ if $U \cap V = \emptyset$ or that l does not belong to support of any atom in $U \cap V$. On the contrary, as $E_0 \neq 0$ and $U \cap V = \{i\}$, then $E_0 = |q_i c_i \beta_i \beta_i' A_{li}^{*2}| \geq \Omega(\tau^2 k/m) = \Omega(k/mr)$ since $|q_i c_i \beta_i \beta_i'| \geq \Omega(k/m)$ and $|A_{li}^*| \geq \tau$.

Therefore, Lemma 1 suggests that if u and v share a unique atom among their sparse representations, and r is not too large, then we can indeed recover the correct support of the shared atom. When this is the case, the expected scores corresponding to the nonzero elements of the shared atom will dominate the remaining of the scores.

Now, given that we can isolate the support R of the corresponding atom, the remaining questions are how best we can estimate its non-zero coefficients, and when u and v share a unique elements in their supports. These issues are handled in the following lemmas.

Lemma 2 *Suppose that* $u = A^*\alpha + \varepsilon_u$ *and* $v = A^*\alpha' + \varepsilon_v$ *are two random samples. Let* U *and* V *denote the supports of* α *and* α' *respectively.* R *is the support of some atom of interest. The truncated re-weighting matrix is formulated as*

$$M_{u,v} \triangleq \mathbb{E}[\langle y, u\rangle \langle y, v\rangle y_R y_R^T] = \sum_{i \in U \cap V} q_i c_i \beta_i \beta_i' A_{R,i}^* A_{R,i}^{*T} + \text{ perturbation terms}$$

where the perturbation terms have norms at most $O^*(k/m \log n)$.

Using the same argument for bounding E_0 in Lemma 1, we can see that $M_0 \triangleq q_i c_i \beta_i \beta_i' A_{R,i}^* A_{R,i}^{*T}$ has norm at least $\Omega(k/m)$ when u and v share a unique element i ($\|A_{R,i}^*\| = 1$). According to this lemma, the spectral norm of M_0 dominates those of the other perturbation terms. Thus, given R we can use the first singular vector of $M_{u,v}$ as an estimate of $A_{\bullet i}^*$.

Lemma 3 *Under the setup of Theorem 4, suppose* $u = A^*\alpha + \varepsilon_u$ *and* $v = A^*\alpha' + \varepsilon_v$ *are two random samples with supports* U *and* V *respectively.* $R = \text{supp}(A_i^*)$. *If* u *and* v *share the unique atom* i, *the first* r *largest entries of* e_l *is at least* $\Omega(k/mr)$ *and belong to* R. *Moreover, the top singular vector of* $M_{u,v}$ *is* δ*-close to* $A_{R,i}^*$ *for* $O^*(1/\log n)$.

Proof The recovery of $A_{\bullet i}^*$'s support directly follows Lemma 1. For the latter part, recall from Lemma 2 that

$$M_{u,v} = q_i c_i \beta_i \beta_i' A_{R,i}^* A_{R,i}^{*T} + \text{ perturbation terms}$$

The perturbation terms have norms bounded by $O^*(k/m \log n)$. On the other hand, the first term is has norm at least $\Omega(k/m)$ since $\|A_{R,i}^*\| = 1$ for the correct support R and $|q_i c_i \beta_i \beta_i'| \geq \Omega(k/m)$. Then using Wedin's Theorem to $M_{u,v}$, we can conclude that the top singular vector must be $O^*(k/m \log n)/\Omega(k/m) = O^*(1/\log n)$ -close to $A_{R,i}^*$. ■

Lemma 4 *Under the setup of Theorem 4, suppose $u = A^*\alpha + \varepsilon_u$ and $v = A^*\alpha' + \varepsilon_v$ are two random samples with supports U and V respectively. If 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.*

Proof The proof follows from that of Lemma 37 in Arora et al. (2015). The main idea is to separate the possible cases of how u and v share support and to use Lemma 2 with the bounded perturbation terms to conclude when u and v share exactly one. We note that due to the condition where $\widehat{e}_{(s)} \geq \Omega(k/mr)$ and $\widehat{e}_{(s+1)}/\widehat{e}_{(s)} \leq O^*(r/\log n)$, it must be the case that u and v share only one atom or share more than one atoms with the same support. When their supports overlap more than one, then the first singular value cannot dominate the second one, and hence it must not be the case. ■

Similar to (Arora et al., 2015), our initialization algorithm requires $\widetilde{O}(m)$ iterations in expectation to estimate all the atoms, hence the expected running time is $\widetilde{O}(mnp)$. All the proofs of Lemma 1 and 2 are deferred to Appendix B.

4. Stage 2: Descent

We now adapt the neural sparse coding approach of Arora et al. (2015) to obtain an improved estimate of A^*. As mentioned earlier, at a high level the algorithm is akin to performing approximate gradient descent. The insight is that within a small enough neighborhood (in the sense of δ-closeness) of the true A^*, an estimate of the ground-truth code vectors, X^*, can be constructed using a neurally plausible algorithm.

The innovation, in our case, is the double-sparsity model since we know *a priori* that A^* is itself sparse. Under sufficiently many samples, the support of A^* can be deduced from the initialization stage; therefore we perform an extra *projection* step in each iteration of gradient descent. In this sense, our method is non-trivially different from Arora et al. (2015). The full algorithm is presented as Algorithm 2.

As discussed in Section 2, convergence of noisy approximate gradient descent can be achieved as long as $\widehat{g}^s$ is correlated-w.h.p. with the true solution. However, an analogous convergence result for *projected* gradient descent does not exist in the literature. We fill this gap via a careful analysis. Due to the projection, we only require the correlated-w.h.p. property for *part* of $\widehat{g}^s$ (i.e., when it is restricted to some support set) with A^*. The descent property is still achieved via Theorem 5. Due to various perturbation terms, $\widehat{g}$ is only a biased estimate of $\nabla_A \mathcal{L}(A, X)$; therefore, we can only refine the estimate of A^* until the column-wise error is of order $O(\sqrt{k/n})$. The performance of Algorithm 2 can be characterized via the following theorem.

Theorem 5 *Suppose that the initial estimate A^0 has the correct column supports and is $(\delta, 2)$-near to A^* with $\delta = O^*(1/\log n)$. If Algorithm 2 is provided with $p = \widetilde{\Omega}(mr)$ fresh samples at each step and $\eta = \Theta(m/k)$, then*

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

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

Algorithm 2 Double-Sparse Coding Descent Algorithm

Initialize A^0 is $(\delta, 2)$-near to A^*. $H = (h_{ij})_{n\times m}$ where $h_{ij} = 1$ if $i \in \text{supp}(A^0_{\bullet j})$ and 0 otherwise.

Repeat for $s = 0, 1, \ldots, T$

Decode: $x^{(i)} = \text{threshold}_{C/2}((A^s)^T y^{(i)}) \quad$ for $i = 1, 2, \ldots, p$

Update: $A^{s+1} = \mathcal{P}_H(A^s - \eta \widehat{g}^s) = A^s - \eta \mathcal{P}_H(\widehat{g}^s)$

where $\widehat{g}^s = \frac{1}{p}\sum_{i=1}^p (A^s x^{(i)} - y^{(i)})\text{sgn}(x^{(i)})^T$ and $\mathcal{P}_H(G) = H \circ G$

We defer the full proof of Theorem 5 to Section D. In this section, we take a step towards understanding the algorithm by analyzing $\widehat{g}^s$ in the infinite sample case, which is equivalent to its expectation $g^s \triangleq \mathbb{E}[(A^s x - y)\text{sgn}(x)^T]$. We establish the $(\alpha, \beta, \gamma_s)$-correlation of a truncated version of $g^s_{\bullet i}$ with $A^*_{\bullet i}$ to obtain the descent in Theorem 6 for the infinite sample case.

Theorem 6 *Suppose that the initial estimate A^0 has the correct column supports and is $(\delta, 2)$-near to A^*. If Algorithm 2 is provided with infinite number of samples at each step and $\eta = \Theta(m/k)$, then*

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

for some $0 < \rho < 1/2$ and for $s = 1, 2, \ldots, T$. Consequently, it converges to A^* geometrically until column-wise error is $O(k/n)$.

Note that the better error $O(k^2/n^2)$ is due to the fact that infinitely many samples are given. The term $O(\sqrt{k/n})$ in Theorem 5 is a trade-off between the accuracy and the sample complexity of the algorithm. The proof of this theorem composes of two steps with two main results: 1) an explicit form of g^s (Lemma 5); 2) $(\alpha, \beta, \gamma_s)$-correlation of column-wise g^s with A^* (Lemma 6). The proof of those lemmas are deferred to Appendix C. Since the correlation primarily relies on the $(\delta, 2)$-nearness of A^s to A^* that is provided initially and maintained at each step, then we need to argue that the nearness is preserved after each step.

Lemma 5 *Suppose that the initial estimate A^0 has the correct column supports and is $(\delta, 2)$-near to A^*. The column-wise update has the form $g^s_{R,i} = p_i q_i(\lambda^s_i A^s_{R,i} - A^*_{R,i} + \xi^s_i \pm \zeta)$ where $R = \text{supp}(A^s_{\bullet i})$, $\lambda^s_i = \langle A^s_{\bullet i}, A^*_{\bullet i}\rangle$ and*

$$\xi^s_i = A^s_{R,-i}\text{diag}(q_{ij})(A^s_{\bullet -i})^T A^*_{\bullet i}/q_i.$$

Moreover, ξ_i has norm bounded by $O(k/n)$ for $\delta = O^*(1/\log n)$ and ζ is negligible.

We underline that the correct support of A^s allows us to obtain the closed-form expression of $g^s_{R_i,i}$ in terms of $A^s_{\bullet i}$ and $A^*_{\bullet i}$. Likewise, the gradient form suggests that $g^s_{\bullet i}$ is almost equal to $p_i q_i(A^s_{\bullet i} - A^*_{\bullet i})$ (since $\lambda^s_i \approx 1$), which directs to the desired solution $A^*_{\bullet i}$. With Lemma 5, we will prove the $(\alpha, \beta, \gamma_s)$-correlation of the approximate gradient to each column $A^*_{\bullet i}$ and the nearness of each new update to the true solution A^*.

4.1 $(\alpha, \beta, \gamma_s)$-Correlation

Lemma 6 *Suppose that A^s to be $(\delta, 2)$-near to A^* and $R = \text{supp}(A^*_{\bullet i})$, then $2g^s_{R,i}$ is $(\alpha, 1/2\alpha, \epsilon^2/\alpha)$-correlated with $A^*_{R,i}$; that is*

$$\langle 2g^s_{R,i}, A^s_{R,i} - A^*_{R,i}\rangle \geq \alpha\|A^s_{R,i} - A^*_{R,i}\|^2 + 1/(2\alpha)\|g^s_{R,i}\|^2 - \epsilon^2/\alpha.$$

where $\delta = O^*(1/\log n)$ and $\epsilon = O\big(\frac{k^2}{mn}\big)$. Futhermore, the descent is achieved by

$$\|A^{s+1}_{\bullet i} - A^*_{\bullet i}\|^2 \leq (1 - 2\alpha\eta)^s\|A^0_{\bullet i} - A^*_{\bullet i}\|^2 + \eta\epsilon^2/\alpha.$$

Proof Throughout the proof, we omit the superscript s for simplicity and denote $2\alpha = p_i q_i$. First, we rewrite $g^s_{\bullet i}$ as a combination of the true direction $A^s_{\bullet i} - A^*_{\bullet i}$ and a term with small norm:

$$g_{R,i} = 2\alpha(A_{R,i} - A^*_{R,i}) + v, \tag{2}$$

where $v = 2\alpha[(\lambda_i - 1)A_{\bullet i} + \epsilon_i]$ with norm bounded. In fact, since $A_{\bullet i}$ is δ-close to $A^*_{\bullet i}$, and both have unit norm, then $\|2\alpha(\lambda_i - 1)A_{\bullet i}\| = \alpha\|A_{\bullet i} - A^*_{\bullet i}\|^2 \leq \alpha\|A_{\bullet i} - A^*_{\bullet i}\|$ and $\|\xi_i\| \leq O(k/n)$ from the inequality (9). Therefore,

$$\|v\| = \|2\alpha(\lambda_i - 1)A_{R,i} + 2\alpha\xi_i\| \leq \alpha\|A_{R,i} - A^*_{R,i}\| + \epsilon$$

where $\epsilon = O(k^2/mn)$. Now, we make use of (2) to show the first part of Lemma 6:

$$\langle 2g_{R,i}, A_{R,i} - A^*_{R,i}\rangle = 4\alpha\|A_{R,i} - A^*_{R,i}\|^2 + \langle 2v, A_{R,i} - A^*_{R,i}\rangle. \tag{3}$$

We want to lower bound the inner product term with respect to $\|g_{R_i,i}\|^2$ and $\|A_{R,i} - A^*_{R,i}\|^2$. Effectively, from (2)

$$\begin{aligned} 4\alpha\langle v, A_{\bullet i} - A^*_{\bullet i}\rangle &= \|g_{R,i}\|^2 - 4\alpha^2\|A_{R,i} - A^*_{R,i}\|^2 - \|v\|^2 \\ &\geq \|g_{R,i}\|^2 - 6\alpha^2\|A_{R,i} - A^*_{R,i}\|^2 - 2\epsilon^2, \end{aligned} \tag{4}$$

where the last step is due to Cauchy-Schwarz inequality: $\|v\|^2 \leq 2(\alpha^2\|A_{R,i} - A^*_{R,i}\|^2 + \epsilon^2)$.

Substitute $2\langle v, A_{\bullet i} - A^*_{\bullet i}\rangle$ in (3) for the right hand side of (4), we get the first result:

$$\langle 2g_{R,i}, A_{R,i} - A^*_{R,i}\rangle \geq \alpha\|A_{R,i} - A^*_{R,i}\|^2 + \frac{1}{2\alpha}\|g_{R,i}\|^2 - \frac{\epsilon^2}{\alpha}.$$

The second part is directly followed from Theorem 1. Moreover, we have $p_i = \Theta(k/m)$ and $q_i = \Theta(1)$, then $\alpha = \Theta(k/m)$, $\beta = \Theta(m/k)$ and $\gamma_s = O(k^3/mn^2)$. Then $g^s_{R,i}$ is $(\Omega(k/m), \Omega(m/k), O(k^3/mn^2))$-correlated with the true solution $A^*_{R,i}$. □

Proof [Proof of Theorem 6] The descent in Theorem 6 directly follows from the above lemma. Next, we will establish the nearness for the update at step s:

4.2 Nearness

Lemma 7 *Suppose that A^s is $(\delta, 2)$-near to A^*, then $\|A^{s+1} - A^*\| \leq 2\|A^*\|$*

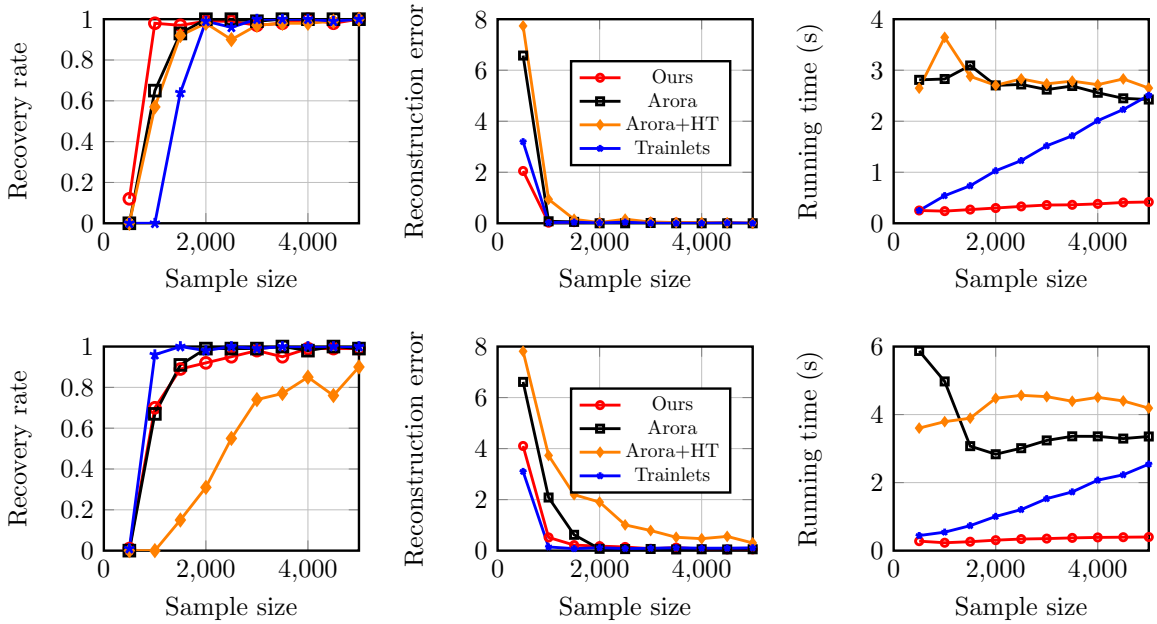

Figure 1: *(top) The performance of four methods on three metrics (recovery rate, reconstruction error and running time (in seconds)) in sample size in the noiseless case. (bottom) The same metrics are measured for the noisy case.*

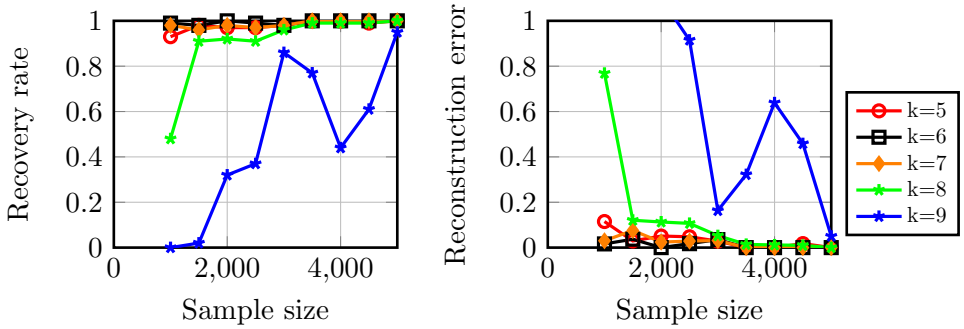

Figure 2: The performance of our method in the noiseless case as the sparsity k varies.

Proof [Proof] From Lemma 5 we have $g^s_{\bullet i} = p_i q_i(\lambda_i A^s_{\bullet i} - A^*_{\bullet i}) + A_{\bullet -i}\mathrm{diag}(q_{ij})A^T_{\bullet -i}A^*_{\bullet i} \pm \zeta$. Denote $\bar{R} = [n]\backslash R$, then it is obvious that $g^s_{\bar{R},i} = A_{\bar{R},-i}\mathrm{diag}(q_{ij})A^T_{\bullet -i}A^*_{\bullet i} \pm \zeta$ is bounded by $O(k^2/m^2)$. Then we follows the proof of Lemma 24 in (Arora et al., 2015) for the nearness with full $g^s = g^s_{R,i} + g^s_{\bar{R},i}$ to finish the proof for this lemma. ■

In sum, we have shown the descent property of Algorithm 2 in the infinite sample case. The study of the concentration of $\widehat{g}^s$ around its mean to the sample complexity is provided in Section D. In the next section, we corroborate our theory by some numerical results on synthetic data.

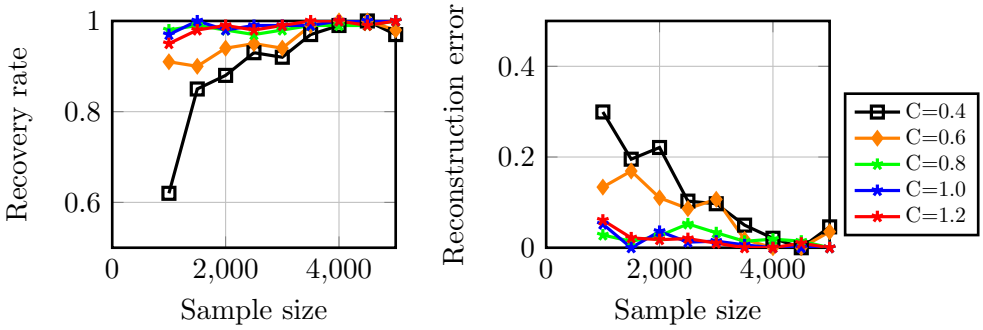

Figure 3: The performance of our method in the noiseless case as the thresholding parameter C varies.

5. Empirical Study

We compare our method with three different methods for both standard sparse and double-sparse coding. For a baseline, we implement the algorithm proposed in Arora et al. (2015), which currently is the most theoretically sound approach for provable sparse coding. However, since their approach is not directly designed for the double-sparsity model, we implement a modified version that performs a hard thresholding (HT)-based post-processing step in the initialization and learning procedures (which we dub *Arora + HT*). The final comparison is with *Trainlets*, the proposed approach by Sulam et al. (2016).

We generate a synthetic training dataset according to the model described in Section 2. The base dictionary Φ is the identity matrix of size $n = 64$, and the square synthesis matrix A^* has a special block structure with 32 blocks. Each block is of size 2×2 and of form $[1\ 1; 1\ -1]$ (i.e., the column sparsity of A^* is $r = 2$). The support of x^* is drawn uniformly over all 6-dimensional subsets of $[m]$, and the nonzero coefficients are randomly set to ± 1 with equal probability. In our simulations with noise, we add Gaussian noise ε with entrywise variance $\sigma_\varepsilon^2 = 0.01$ to each of those above samples. For all the approaches except Trainlets, we use $T = 2000$ iterations for the initialization procedure, and set the number of steps in the descent stage to 25. Since Trainlets does not have a specified initialization procedure, we initialize it with a random Gaussian matrix upon which column-wise sparse thresholding is then performed. The learning step of Trainlets[2] is executed for 50 iterations, which tolerates its initialization deficiency. For each Monte Carlo trial, we uniformly draw p samples, feed these samples to the four different algorithms, and observe their ability to reconstruct A^*. Matlab implementation of our algorithms is available online[3].

5.1 Comparison with Other Approaches

We evaluate these approaches on three metrics as a function of the number of available samples: (i) fraction of trials in which each algorithm successfully recovers the ground truth A^*; (ii) reconstruction error; and (iii) running time (in seconds). The synthesis matrix is

2. We utilize Trainlets's implementation provided at http://jsulam.cswp.cs.technion.ac.il/home/software/.
3. https://github.com/thanh-isu/double-sparse-coding

said to be "successfully recovered" if the Frobenius norm of the difference between the estimate $\widehat{A}$ and the ground truth A^* is smaller than a threshold which is set to 10^{-4} in the noiseless case, and to 0.5 in the other. All three metrics are averaged over 100 Monte Carlo simulations. As discussed above, the Frobenius norm is only meaningful under a suitable permutation and sign flip transformation linking $\widehat{A}$ and A^*. We estimate this transformation using a simple maximum weight matching algorithm. Specifically, we construct a weighted bipartite graph with nodes representing columns of A^* and $\widehat{A}$ and adjacency matrix defined as $G = |A^{*T}\widehat{A}|$, where $|\cdot|$ is taken element-wise. We compute the optimal matching using the Hungarian algorithm, and then estimate the sign flips by looking at the sign of the inner products between the matched columns.

The results of our experiments are shown in Figure 1 with the top and bottom rows respectively for the noiseless and noisy cases. The two leftmost figures suggest that all algorithms exhibit a "phase transition" in sample complexity that occurs in the range of 500-2000 samples. In the noiseless case, our method achieves the phase transition with the fewest number of samples. In the noisy case, our method nearly matches the best sample complexity performance (next to Trainlets, which is a heuristic and computationally expensive). Our method achieves the best performance in terms of (wall-clock) running time in all cases.

5.2 Robustness to Data Assumptions

In this last experiement, we show that our approach is robust to the data assumptions. We numerically study how the initialization and descent algorithms behave when the sparsity k and the thresholding parameter C slightly vary around the groundtruth values. Since our focus is on the recovery property of our approach, we assume that the dictionary size m and sparsity r are known a priori and do not experiement on them.

The results are shown in Figures 2 and 3. When the sparsity and the minimum coefficient are around the true setting, $k_{\text{model}} = 6$ and $C_{\text{model}} = 1.0$, our algorithm is still able recover the dictionary perfectly. When these parameters are set more extreme, the phase transition is not obvious but is gradually achieved with more and more samples.

6. Conclusion

In this paper, we have addressed an open theoretical question on learning sparse dictionaries under a special type of generative model. Our proposed algorithm consists of a novel initialization step followed by a descent-style step, both are able to take advantage of the sparse structure. We rigorously demonstrate its efficacy in both sample- and computation-complexity over existing heuristics as well as provable approaches for double-sparse and regular sparse coding. This results in *the first known provable approach for double-sparse coding problem* with statistical and algorithmic guarantees. Besides, we also show three benefits of our approach: neural plausibility, robustness to noise and practical usefulness via the numerical experiments.

Nevertheless, several fundamental questions regarding our approach remain. First, our initialization method (in the overcomplete case) achieves its theoretical guarantees under fairly stringent limitations on the sparsity level r. This arises due to our reweighted spectral initialization strategy, and it is an open question whether a better initialization strategy

exists (or whether these types of initialization are required at all). Second, our analysis holds for complete (fixed) bases Φ, and it remains open to study the setting where Φ is over-complete. Finally, understanding the reasons behind the very promising practical performance of methods based on heuristics, such as Trainlets, on real-world data remains a very challenging open problem.

Acknowledgments

We would like to thank the anonymous reviewers for the constructive feedback and suggestions. This work is supported in part by the National Science Foundation under the grants CCF-1566281 and DMS-1612985.

Appendix Organization We organize the appendix as follows: we prove the two key lemmas for Theorem 4 of the initialization algorithm 1 in Appendix B. In Appendix C, we prove the result stated in Theorem 5 for the infinite-sample case. The sample complexity results for both stages are proved in Appendix D.

Additionally, we prove some extended results from Arora et al. (2015) and for some special cases in Appendices E and F. The final section details the neural implementation of our approach.

Appendix A. Useful Result

We start our proof with the following claim, which we will use throughout.

Claim 1 (Maximal row ℓ_1-norm) *Given that $\|A^*\|_F^2 = m$ and $\|A^*\| = O(\sqrt{m/n})$, then $\|A^{*T}\|_{1,2} = \Theta(\sqrt{m/n})$.*

Proof Recall the definition of the operator norm:

$$\|A^{*T}\|_{1,2} = \sup_{x \neq 0} \frac{\|A^T x\|}{\|x\|_1} \leq \sup_{x \neq 0} \frac{\|A^T x\|}{\|x\|} = \|A^{*T}\| = O(\sqrt{m/n}).$$

Since $\|A^*\|_F^2 = m$, $\|A^{*T}\|_{1,2} \geq \|A^*\|_F/\sqrt{n} = \sqrt{m/n}$. Combining with the above, we have $\|A^{*T}\|_{1,2} = \Theta(\sqrt{m/n})$. ■

Along with Assumptions **A1** and **A3**, the above claim implies the number of nonzero entries in each row is $O(r)$. This Claim is an important ingredient in our analysis of our initialization algorithm shown in Section 3.

Appendix B. Analysis of Initialization Algorithm

B.1 Proof of Lemma 1

Recall some important notations: $y = A^* x^* + \varepsilon$ and two samples

$$u = A^* \alpha + \varepsilon_u, v = A^* \alpha' + \varepsilon_v.$$

Also, recall the very coarse estimate for the sparse code of u with respect to A^*:

$$\beta = A^{*T} u = A^{*T} A^* \alpha + A^{*T} \varepsilon_u.$$

We split the proof of Lemma 1 into three steps: 1) we first establish useful properties of β with respect to α; 2) we then explicitly derive e_l in terms of the generative model parameters and β; and 3) we finally bound the error terms in E based on the first result and appropriate assumptions.

Claim 2 *In the generative model, $\|x^*\| \leq \widetilde{O}(\sqrt{k})$ and $\|\varepsilon\| \leq \widetilde{O}(\sigma_\varepsilon \sqrt{n})$ with high probability.*

Proof The claim directly follows from the fact that x^* is a k-sparse random vector whose nonzero entries are independent sub-Gaussian with variance 1. Meanwhile, ε has n independent Gaussian entries of variance σ_ε^2. □

Despite its simplicity, this claim will be used in many proofs throughout the paper. Note also that in this section we will calculate the expectation over y and often refer probabilistic bounds (w.h.p.) under the randomness of u and v.

Claim 3 *Suppose that $u = A^*\alpha + \varepsilon_u$ is a random sample and $U = \mathrm{supp}(\alpha)$. Let $\beta = A^{*T}u$, then, w.h.p., we have (a) $|\beta_i - \alpha_i| \le \frac{\mu k \log n}{\sqrt{n}} + \sigma_\varepsilon \log n$ for each i and (b) $\|\beta\| \le \widetilde{O}(\sqrt{k} + \sigma_\varepsilon\sqrt{n})$.*

Proof The proof mostly follows from Claim 36 of Arora et al. (2015), with an additional consideration of the error ε_u. Write $W = U\backslash\{i\}$ and observe that

$$|\beta_i - \alpha_i| = |A^{*T}_{\bullet i} A^*_{\bullet W}\alpha_W + A^{*T}_{\bullet i}\varepsilon_u| \le |\langle A^{*T}_{\bullet W} A^*_{\bullet i}, \alpha_W\rangle| + |\langle A^*_{\bullet i}, \varepsilon_u\rangle|$$

Since A^* is μ-incoherence, then $\|A^{*T}_{\bullet i} A^*_{\bullet W}\| \le \mu\sqrt{k/n}$. Moreover, α_W has $k-1$ independent sub-Gaussian entries of variance 1, therefore $|\langle A^{*T}_{\bullet W} A^*_{\bullet i}, \alpha_W\rangle| \le \frac{\mu k \log n}{\sqrt{n}}$ with high probability. Also recall that ε_u has independent Gaussian entries of variance σ_ε^2, then $A^{*T}_{\bullet i}\varepsilon_u$ is Gaussian with the same variance ($\|A^*_{\bullet i}\| = 1$). Hence $|A^{*T}_{\bullet i}\varepsilon| \le \sigma_\varepsilon \log n$ with high probability. Consequently, $|\beta_i - \alpha_i| \le \frac{\mu k \log n}{\sqrt{n}} + \sigma_\varepsilon \log n$, which is the first part of the claim.

Next, in order to bound $\|\beta\|$, we express β as

$$\|\beta\| = \|A^{*T} A^*_{\bullet U}\alpha_U + A^{*T}\varepsilon_u\| \le \|A^*\|\|A^*_{\bullet U}\|\|\alpha_U\| + \|A^*\|\|\varepsilon_u\|$$

Using Claim 2 to get $\|\alpha_U\| \le \widetilde{O}(\sqrt{k})$ and $\|\varepsilon_u\| \le \widetilde{O}(\sigma_\varepsilon\sqrt{n})$ w.h.p., and further noticing that $\|A^*_{\bullet U}\| \le \|A^*\| \le O(1)$, we complete the proof for the second part. ■

Claim 3 suggests that the difference between β_i and α_i is bounded above by $O^*(1/\log^2 n)$ w.h.p. if $\mu = O^*(\frac{\sqrt{n}}{k\log^3 n})$. Therefore, w.h.p., $C - o(1) \le |\beta_i| \le |\alpha_i| + o(1) \le O(\log m)$ for $i \in U$ and $|\beta_i| \le O^*(1/\log^2 n)$ otherwise. On the other hand, under Assumption **B4**, $\|\beta\| \le \widetilde{O}(\sqrt{k})$ w.h.p. We will use these results multiple times in the next few proofs.

Proof [Proof of Lemma 1] We decompose d_l into small parts so that the stochastic model $\mathcal{D}$ is made use.

$$\begin{aligned} e_l &= \mathbb{E}[\langle y,u\rangle\langle y,v\rangle y_l^2] = \mathbb{E}[\langle A^*x^* + \varepsilon, u\rangle\langle A^*x^* + \varepsilon, v\rangle(\langle A^*_{l\cdot}, x^*\rangle + \varepsilon_l)^2] \\ &= \mathbb{E}\big[\big\{\langle x^*,\beta\rangle\langle x^*,\beta'\rangle + x^{*T}(\beta v^T + \beta' u^T)\varepsilon + u^T\varepsilon\varepsilon^T v\big\}\big\{\langle A^*_{l\bullet}, x^*\rangle^2 + 2\langle A^*_{l\bullet}, x^*\rangle\varepsilon_l + \varepsilon_l^2\big\}\big] \\ &= E_1 + E_2 + \cdots + E_9 \end{aligned}$$

where the terms are

$$\begin{aligned} E_1 &= \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\langle A^*_{l\bullet}, x^*\rangle^2] \\ E_2 &= 2\mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\langle A^*_{l\bullet}, x^*\rangle\varepsilon_l] \\ E_3 &= \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\varepsilon_l^2] \\ E_4 &= \mathbb{E}\big[\langle A^*_{l\cdot}, x^*\rangle^2 x^{*T}(\beta v^T + \beta' u^T)\varepsilon\big] \\ E_5 &= \mathbb{E}\big[\langle A^*_{l\cdot}, x^*\rangle x^{*T}(\beta v^T + \beta' u^T)\varepsilon\varepsilon_l\big] \\ E_6 &= \mathbb{E}\big[(\beta v^T + \beta' u^T)\varepsilon\varepsilon_l^2\big] \\ E_7 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v\langle A^*_{l\bullet}, x^*\rangle^2] \\ E_8 &= 2\mathbb{E}[u^T\varepsilon\varepsilon^T v\langle A^*_{l\bullet}, x^*\rangle\varepsilon_l] \\ E_9 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_l^2] \end{aligned} \tag{5}$$

Because x^* and ε are independent and zero-mean, E_2 and E_4 are clearly zero. Moreover,

$$E_6 = (\beta v^T + \beta' u^T)\mathbb{E}[\varepsilon\varepsilon_l^2] = 0$$

due to the fact that $\mathbb{E}[\varepsilon_j\varepsilon_l^2] = 0$, for $j \neq l$, and $\mathbb{E}[\varepsilon_l^3] = 0$. Also,

$$E_8 = A_{l\bullet}^{*T}\mathbb{E}[x^*]\mathbb{E}\big[u^T\varepsilon\varepsilon^T v\varepsilon_l\big] = 0.$$

We bound the remaining terms separately in the following claims.

Claim 4 *In the decomposition* (5)*, E_1 is of the form*

$$E_1 = \sum_{i\in U\cap V} q_i c_i \beta_i\beta_i' A_{li}^{*2} + \sum_{i\notin U\cap V} q_i c_i \beta_i\beta_i' A_{li}^{*2} + \sum_{j\neq i} q_{ij}(\beta_i\beta_i' A_{lj}^{*2} + 2\beta_i\beta_j' A_{li}^* A_{lj}^*)$$

where all those terms except $\sum_{i\in U\cap V} q_i c_i\beta_i\beta_i' A_{li}^{*2}$ have magnitude at most $O^*(k/m\log^2 n)$ w.h.p.

Proof Using the generative model in Assumptions **B1**-**B4**, we have

$$\begin{aligned}
E_1 &= \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\langle A_{l\bullet}^*, x^*\rangle^2] \\
&= \mathbb{E}_S\big[\mathbb{E}_{x^*|S}[\sum_{i\in S}\beta_i x_i^*\sum_{i\in S}\beta_i' x_i^*\big(\sum_{i\in S} A_{li}^* x_i^*\big)^2]\big] \\
&= \sum_{i\in[m]} q_i c_i\beta_i\beta_i' A_{li}^{*2} + \sum_{i,j\in[m],j\neq i} q_{ij}(\beta_i\beta_i' A_{lj}^{*2} + 2\beta_i\beta_j' A_{li}^* A_{lj}^*) \\
&= \sum_{i\in U\cap V} q_i c_i\beta_i\beta_i' A_{li}^{*2} + \sum_{i\notin U\cap V} q_i c_i\beta_i\beta_i' A_{li}^{*2} + \sum_{j\neq i} q_{ij}(\beta_i\beta_i' A_{lj}^{*2} + 2\beta_i\beta_j' A_{li}^* A_{lj}^*),
\end{aligned}$$

where we have used the $q_i = \mathbb{P}[i\in S]$, $q_{ij} = \mathbb{P}[i,j\in S]$ and $c_i = \mathbb{E}[x_i^4|i\in S]$ and Assumptions **B1**-**B4**. We now prove that the last three terms are upper bounded by $O^*(k/m\log n)$. The key observation is that all these terms typically involve a quadratic form of the l-th row $A_{l\bullet}^*$ whose norm is bounded by $O(1)$ (by Claim 1 and Assumption **A4**). Moreover, $|\beta_i\beta_i'|$ is relatively small for $i\notin U\cap V$ while $q_{ij} = \Theta(k^2/m^2)$. For the second term, we apply the Claim 3 for $i\in[m]\backslash(U\cap V)$ to bound $|\beta_i\beta_i'|$. Assume $\alpha_i = 0$ and $\alpha_i'\neq 0$, then with high probability

$$|\beta_i\beta_i'| \leq |(\beta_i-\alpha_i)(\beta_i'-\alpha_i')| + |\beta_i\alpha_i'| \leq O^*(1/\log n)$$

Using the bound $q_i c_i = \Theta(k/m)$, we have w.h.p.,

$$\Big|\sum_{i\notin U\cap V} q_i c_i\beta_i\beta_i' A_{li}^{*2}\Big| \leq \max_i |q_i c_i\beta_i\beta_i'| \sum_{i\notin U\cap V} A_{li}^{*2} \leq \max_i |q_i c_i\beta_i\beta_i'| \|A^*\|_{1,2}^2 \leq O^*(k/m\log n).$$

For the third term, we make use of the bounds on $\|\beta\|$ and $\|\beta'\|$ from the previous claim where $\|\beta\|\|\beta'\| \leq \widetilde{O}(k)$ w.h.p., and on $q_{ij} = \Theta(k^2/m^2)$. More precisely, w.h.p.,

$$\begin{aligned}
\Big|\sum_{j\neq i} q_{ij}\beta_i\beta_i' A_{lj}^{*2}\Big| &= \Big|\sum_i \beta_i\beta_i'\sum_{j\neq i} q_{ij} A_{lj}^{*2}\Big| \leq \sum_i |\beta_i\beta_i'|\big(\sum_{j\neq i} q_{ij} A_{lj}^{*2}\big) \\
&\leq (\max_{i\neq j} q_{ij})\sum_i |\beta_i\beta_i'|\Big(\sum_j A_{lj}^{*2}\Big) \leq (\max_{i\neq j} q_{ij})\|\beta\|\|\beta'\|\|A^*\|_{1,2}^2 \leq \widetilde{O}(k^3/m^2),
\end{aligned}$$

where the second last inequality follows from the Cauchy-Schwarz inequality. For the last term, we write it in a matrix form as $\sum_{j\neq i} q_{ij}\beta_i\beta'_j A^*_{li}A^*_{lj} = A^{*T}_{l\bullet}Q_\beta A^*_{l\bullet}$ where $(Q_\beta)_{ij} = q_{ij}\beta_i\beta'_j$ for $i \neq j$ and $(Q_\beta)_{ij} = 0$ for $i = j$. Then

$$|A^{*T}_{l\bullet}Q_\beta A^*_{l\bullet}| \leq \|Q_\beta\|\|A^*_{l\bullet}\|^2 \leq \|Q_\beta\|_F\|A^*\|^2_{1,2},$$

where $\|Q_\beta\|_F^2 = \sum_{i\neq j} q_{ij}^2\beta_i^2(\beta'_j)^2 \leq (\max_{i\neq j} q_{ij}^2)\sum_i \beta_i^2\sum_j(\beta'_j)^2 \leq (\max_{i\neq j} q_{ij}^2)\|\beta\|^2\|\beta'\|^2$. Ultimately,

$$\Big|\sum_{j\neq i} q_{ij}\beta_i\beta'_j A^*_{li}A^*_{lj}\Big| \leq (\max_{i\neq j} q_{ij})\|\beta\|\|\beta'\|\|A^*\|^2_{1,2} \leq \widetilde{O}(k^3/m^2).$$

Under Assumption $k = O^*(\frac{\sqrt{n}}{\log n})$, then $\widetilde{O}(k^3/m^2) \leq O^*(k/m\log^2 n)$. As a result, the two terms above are bounded by the same amount $O^*(k/m\log n)$ w.h.p., so we complete the proof of the claim. ■

Claim 5 *In the decomposition (5), $|E_3|$, $|E_5|$, $|E_7|$ and $|E_9|$ are at most $O^*(k/m\log^2 n)$.*

Proof Recall that $\mathbb{E}[x_i^2|S] = 1$ and $q_i = \mathbb{P}[i \in S] = \Theta(k/m)$ for $S = \text{supp}(x^*)$, then

$$\begin{aligned} E_3 = \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\varepsilon_l^2] &= \sigma_\varepsilon^2\mathbb{E}_S\big[\mathbb{E}_{x^*|S}[\sum_{i,j\in S}\beta_i\beta'_j x_i^*x_j^*]\big] \\ &= \sigma_\varepsilon^2\mathbb{E}_S[\sum_{i\in S}\beta_i\beta'_i] = \sum_i \sigma_\varepsilon^2 q_i\beta_i\beta'_i \end{aligned}$$

Denote $Q = \text{diag}(q_1, q_2, \ldots, q_m)$, then $|E_3| = |\sigma_\varepsilon^2\langle Q\beta,\beta'\rangle| \leq \sigma_\varepsilon^2\|Q\|\|\beta\|\|\beta'\| \leq \widetilde{O}(\sigma_\varepsilon^2 k^2/m) = \widetilde{O}(k^3/mn)$ where we have used $\|\beta\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. and $\sigma_\varepsilon \leq O(1/\sqrt{n})$. For convenience, we handle the seventh term before E_5:

$$E_7 = \mathbb{E}[u^T\varepsilon\varepsilon^T v\langle A^*_{l\bullet}, x^*\rangle^2] = \mathbb{E}[\langle A^*_{l\bullet}, x^*\rangle^2]u^T\mathbb{E}[\varepsilon\varepsilon^T]v = \sum_i \sigma_\varepsilon^2\langle u,v\rangle q_i A^2_{li} = \sigma_\varepsilon^2\langle u,v\rangle A^T_{l\bullet}QA_{l\bullet}$$

To bound this term, we use Claim 9 in Appendix D to have $\|u\| = \|A^*\alpha + \varepsilon_u\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. and $\langle u,v\rangle \leq \widetilde{O}(\sqrt{k})$ w.h.p. Consequently, $|E_7| \leq \sigma_\varepsilon^2\|Q\|\|A_{l\bullet}\|^2|\langle u,v\rangle| \leq \widetilde{O}(k^2/mn)$ because $\|A_{l\bullet}\|^2 \leq O(m/n)$ and $\sigma_\varepsilon \leq O(1/\sqrt{n})$. Now, the firth term E_5 is expressed as follows

$$\begin{aligned} E_5 &= \mathbb{E}\big[\langle A^*_{l\cdot}, x^*\rangle x^{*T}(\beta v^T + \beta' u^T)\varepsilon\varepsilon_l\big] \\ &= A^{*T}_{l\bullet}\mathbb{E}\big[x^*x^{*T}\big](\beta v^T + \beta' u^T)\mathbb{E}[\varepsilon\varepsilon_l] \\ &= \sigma_\varepsilon^2 A^{*T}_{l\bullet}Q(v_l\beta + u_l\beta') \end{aligned}$$

Observe that $|E_5| \leq \sigma_\varepsilon^2\|A^{*T}_{l\bullet}\|\|Q(v_l\beta + u_l\beta')\| \leq \sigma_\varepsilon^2\|A^{*T}_{l\bullet}\|\|Q\|\|v_l\beta + u_l\beta'\|$ and that $\|v_l\beta + u_l\beta'\| \leq 2\|u\|\|\beta\| \leq \widetilde{O}(k)$ w.h.p. using the result $\|u\| \leq \widetilde{O}(k)$ and $\|\beta\| \leq \widetilde{O}(k)$ from Claim 3, then E_5 bounded by $\widetilde{O}(k^2/mn)$.

The last term

$$E_9 = \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_l^2] = u^T\mathbb{E}\big[\varepsilon\varepsilon^T\varepsilon_l^2\big]v = 9\sigma_\varepsilon^4\langle u,v\rangle$$

because the independent entries of ε and $\mathbb{E}[\varepsilon_l^4] = 9\sigma_\varepsilon^4$. Therefore, $|E_9| \leq 9\sigma_\varepsilon^4\|u\|\|v\| \leq \widetilde{O}(k^2/n^2)$. Since $m = O(n)$ and $k \leq O^*(\frac{\sqrt{n}}{\log n})$, we obtain the same bound $O^*(k/m\log^2 n)$ for $|E_3|$, $|E_5|$, $|E_7|$ and $|E_9|$, and conclude the proof of the claim. ■

Combining the bounds from Claim 4, 5 for every single term in (5), we finish the proof for Lemma 1. □

B.2 Proof of Lemma 2

We prove this lemma by using the same strategy used to prove Lemma 1.

$$\begin{aligned}
M_{u,v} &\triangleq \mathbb{E}[\langle y,u\rangle\langle y,v\rangle y_R y_R^T] \\
&= \mathbb{E}[\langle A^*x^*+\varepsilon,u\rangle\langle A^*x^*+\varepsilon,v\rangle(A^*_{R\bullet}x^*+\varepsilon_R)(A^*_{R\bullet}x^*+\varepsilon_R)^T] \\
&= \mathbb{E}\big[\big\{\langle x^*,\beta\rangle\langle x^*,\beta'\rangle + x^{*T}(\beta v^T+\beta' u^T)\varepsilon + u^T\varepsilon\varepsilon^T v\big\}\big\{A^*_{R\bullet}x^*x^{*T}A^{*T}_{R\bullet} + A^*_{R\bullet}x^*\varepsilon_R^T + \varepsilon_R x^{*T}A^{*T}_{R\bullet} + \varepsilon_R\varepsilon_R^T\big\}\big] \\
&= M_1+\cdots+M_8,
\end{aligned}$$

in which only nontrivial terms are kept in place, including

$$\begin{aligned}
M_1 &= \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle A^*_{R\bullet}x^*x^{*T}A^{*T}_{R\bullet}] \\
M_2 &= \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\varepsilon_R\varepsilon_R^T] \\
M_3 &= \mathbb{E}[x^{*T}(\beta v^T+\beta' u^T)\varepsilon A^*_{R\bullet}x^*\varepsilon_R^T] \\
M_4 &= \mathbb{E}[x^{*T}(\beta v^T+\beta' u^T)\varepsilon\varepsilon_R x^{*T}A^{*T}_{R\bullet}] \\
M_5 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v A^*_{R\bullet}x^*x^{*T}A^{*T}_{R\bullet}] \\
M_6 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v A^*_{R\bullet}x^*\varepsilon_R^T] \\
M_7 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_R^T x^{*T}A^{*T}_{R\bullet}] \\
M_8 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_R\varepsilon_R^T]
\end{aligned} \tag{6}$$

By swapping inner product terms and taking advantage of the independence, we can show that $M_6 = \mathbb{E}[A^*_{R\bullet}x^*u^T\varepsilon\varepsilon^T v\varepsilon_R^T] = 0$ and $M_7 = \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_R^T x^{*T}A^{*T}_{R\bullet}] = 0$. The remaining are bounded in the next claims.

Claim 6 *In the decomposition* (6)*,*

$$M_1 = \sum_{i\in U\cap V} q_i c_i\beta_i\beta_i' A^*_{R,i}A^{*T}_{R,i} + E_1' + E_2' + E_3'$$

where $E_1' = \sum_{i\notin U\cap V} q_i c_i\beta_i\beta_i' A^*_{R,i}A^{*T}_{R,i}$*,* $E_2' = \sum_{i\neq j} q_{ij}\beta_i\beta_i' A^*_{R,j}A^{*T}_{R,j}$ *and* $E_3' = \sum_{i\neq j} q_{ij}(\beta_i A^*_{R,i}\beta_j' A^{*T}_{R,j} + \beta_i' A^*_{R,i}\beta_j A^{*T}_{R,j})$ *have norms bounded by* $O^*(k/m\log n)$*.*

Proof The expression of M_1 is obtained in the same way as E_1 is derived in the proof of Lemma 1. To prove the claim, we bound all the terms with respect to the spectral norm of $A^*_{R\bullet}$ and make use of Assumption **A4** to find the exact upper bound.

For the first term E_1', rewrite $E_1' = A^*_{R,S}D_1A^{*T}_{R,S}$ where $S = [m]\backslash(U\cap V)$ and D_1 is a diagonal matrix whose entries are $q_ic_i\beta_i\beta_i'$. Clearly, $\|D_1\| \le \max_{i\in S}|q_ic_i\beta_i\beta_i'| \le O^*(k/m\log n)$ as shown in Claim 4, then

$$\|E_1'\| \le \max_{i\in S}|q_ic_i\beta_i\beta_i'|\|A^*_{R,S}\|^2 \le \max_{i\in S}|q_ic_i\beta_i\beta_i'|\|A^*_{R\bullet}\|^2 \le O^*(k/m\log n)$$

where $\|A^*_{R,S}\| \le \|A^*_{R\bullet}\| \le O(1)$. The second term E_2' is a sum of positive semidefinite matrices, and $\|\beta\| \le O(k\log n)$, then

$$E_2' = \sum_{i\neq j} q_{ij}\beta_i\beta_i' A^*_{R,j}A^{*T}_{R,j} \preceq \max_{i\neq j} q_{ij}\Big(\sum_i \beta_i\beta_i'\Big)\Big(\sum_j A^*_{R,j}A^{*T}_{R,j}\Big) \preceq (\max_{i\neq j} q_{ij})\|\beta\|\|\beta'\|A^*_{R\bullet}A^{*T}_{R\bullet}$$

which implies that $\|E_2'\| \leq (\max_{i\neq j} q_{ij})\|\beta\|\|\beta'\|\|A^*_{R\bullet}\|^2 \leq \widetilde{O}(k^3/m^2)$. Observe that E_3' has the same form as the last term in Claim 4, which is $E_3' = A^{*T}_{R\bullet} Q_\beta A^*_{R\bullet}$. Then

$$\|E_3'\| \leq \|Q_\beta\|\|A^*_{R\bullet}\|^2 \leq (\max_{i\neq j} q_{ij})\|\beta\|\|\beta'\|\|A^*_{R\bullet}\|^2 \leq \widetilde{O}(k^3/m^2)$$

By Claim 3, we have $\|\beta\|$ and $\|\beta'\|$ are bounded by $O(\sqrt{k}\log n)$, and note that $k \leq O^*(\sqrt{n}/\log n)$, then we complete the proof for Lemma 6. ■

Claim 7 *In the decomposition* (6)*, M_2, M_3, M_4, M_5 and M_8 have norms bounded by $O^*(k/m\log n)$.*

Proof Recall the definition of Q in Claim 5 and use the fact that $\mathbb{E}[x^*x^{*T}] = Q$, we can get $M_2 = \mathbb{E}[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle\varepsilon_R\varepsilon_R^T] = \sum_i \sigma_\varepsilon^2 q_i\beta_i\beta_i' I_r$. Then, $\|M_2\| \leq \sigma_\varepsilon^2 \max_i q_i\|\beta\|\|\beta'\| \leq O(\sigma_\varepsilon^2 k^2\log^2 n/m)$.

The next three terms all involve $A^*_{R\bullet}$ whose norm is bounded according to Assumption **A4**. Specifically,

$$\begin{aligned} M_3 &= \mathbb{E}[x^{*T}(\beta v^T + \beta' u^T)\varepsilon A^*_{R\bullet} x^*\varepsilon_R^T] = \mathbb{E}[A^*_{R\bullet}x^*x^{*T}(\beta v^T+\beta' u^T)\varepsilon\varepsilon_R^T] \\ &= A^*_{R\bullet}\mathbb{E}[x^*x^{*T}](\beta v^T+\beta' u^T)\mathbb{E}[\varepsilon\varepsilon_R^T] \\ &= A^*_{R\bullet}Q(\beta v^T+\beta' u^T)\mathbb{E}[\varepsilon\varepsilon_R^T], \end{aligned}$$

and

$$\begin{aligned} M_4 &= \mathbb{E}[x^{*T}(\beta v^T+\beta' u^T)\varepsilon\varepsilon_R x^{*T}A^{*T}_{R\bullet}] = \mathbb{E}[\varepsilon_R\varepsilon^T(v\beta^T+u\beta'^T)x^*x^{*T}A^{*T}_{R\bullet}] \\ &= \mathbb{E}[\varepsilon_R\varepsilon^T](v\beta^T+u\beta'^T)\mathbb{E}[x^*x^{*T}]A^{*T}_{R\bullet} \\ &= \mathbb{E}[\varepsilon_R\varepsilon^T](v\beta^T+u\beta'^T)QA^{*T}_{R\bullet}, \end{aligned}$$

and the fifth term $M_5 = \mathbb{E}[u^T\varepsilon\varepsilon^T v A^*_{R\bullet}x^*x^{*T}A^{*T}_{R\bullet}] = \sigma_\varepsilon^2 u^T v A^*_{R\bullet}\mathbb{E}[x^*x^{*T}]A^{*T}_{R\bullet} = \sigma_\varepsilon^2 u^T v A^*_{R\bullet}QA^{*T}_{R\bullet}$. We already have $\|\mathbb{E}[\varepsilon\varepsilon_R^T]\| = \sigma_\varepsilon^2$, $\|Q\| \leq O(k/m)$ and $|u^Tv| \leq \widetilde{O}(k)$ (proof of Claim 9), then the remaining work is to bound $\|\beta v^T + \beta' u^T\|$, then the bound of $v\beta^T + u\beta'^T$ directly follows. We have $\|\beta v^T\| = \|A^* uv^T\| \leq \|A^*\|\|u\|\|v\| \leq \widetilde{O}(k)$. Therefore, all three terms M_3, M_4 and M_5 are bounded in norm by $\widetilde{O}(\sigma_\varepsilon^2 k^2/m) \leq \widetilde{O}(k^3/mn)$.

The remaining term is

$$\begin{aligned} M_8 &= \mathbb{E}[u^T\varepsilon\varepsilon^T v\varepsilon_R\varepsilon_R^T] = \mathbb{E}[\big(\sum_{i,j} u_i v_j\varepsilon_i\varepsilon_j\big)\varepsilon_R\varepsilon_R^T] \\ &= \mathbb{E}[\big(\sum_{i\in R} u_i v_i\varepsilon_i^2\varepsilon_R\varepsilon_R^T\big)] + \mathbb{E}[\big(\sum_{i\neq j} u_i v_j\varepsilon_i\varepsilon_j\big)\varepsilon_R\varepsilon_R^T] \\ &= \sigma_\varepsilon^4 u_R v_R^T \end{aligned}$$

where $u_R = A^*_{R\bullet}\alpha + (\varepsilon_u)_R$ and $v_R = A^*_{R\bullet}\alpha' + (\varepsilon_v)_R$. We can see that $\|u_R\| \leq \|A^*_{R\bullet}\|\|\alpha\| + \|(\varepsilon_u)_R\| \leq \widetilde{O}(\sqrt{k})$. Therefore, $\|M8\| \leq \widetilde{O}(\sigma_\varepsilon^4 k) = \widetilde{O}(k^3/n^2)$. Since $m = O(n)$ and $k \leq O^*(\frac{\sqrt{n}}{\log n})$, then we can bound all the above terms by $O^*(k/m\log n)$ and finish the proof of Claim 7. ■

Combine the results of Claim 6 and 7, we complete the proof of Lemma 2.

Appendix C. Analysis of Main Algorithm

C.1 Simple Encoding

We can see that $(A^s x - y)\text{sgn}(x)^T$ is random over y and x that is obtained from the encoding step. We follow (Arora et al., 2015) to derive the closed form of $g^s = \mathbb{E}[(A^s x - y)\text{sgn}(x)^T]$ by proving that the encoding recovers the sign of x^* with high probability as long as A^s is close enough to A^*.

Lemma 8 *Assume that A^s is δ-close to A^* for $\delta = O(r/n\log n)$ and $\mu \le \frac{\sqrt{n}}{2k}$, and $k \ge \Omega(\log m)$ then with high probability over random samples $y = A^* x^* + \varepsilon$*

$$\text{sgn}(\text{threshold}_{C/2}\big((A^s)^T y\big) = \text{sgn}(x^*) \tag{7}$$

Proof [Proof of Lemma 8] We follow the same proof strategy from (Arora et al., 2015) (Lemmas 16 and 17) to prove a more general version in which the noise ε is taken into account. Write $S = \text{supp}(x^*)$ and skip the superscript s on A^s for the readability. What we need is to show $S = \{i \in [m] : \langle A_{\bullet i}, y\rangle \ge C/2\}$ and then $\text{sgn}(\langle A^s_{\bullet i}, y\rangle) = \text{sgn}(x^*_i)$ for each $i \in S$ with high probability. Following the same argument of (Arora et al., 2015), we prove in below a stronger statement that, even conditioned on the support S, $S = \{i \in [m] : |\langle A_{\bullet i}, y\rangle| \ge C/2\}$ with high probability.

Rewrite

$$\langle A_{\bullet i}, y\rangle = \langle A_{\bullet i}, A^* x^* + \varepsilon\rangle = \langle A_{\bullet i}, A^*_{\bullet i}\rangle x^*_i + \sum_{j\neq i}\langle A_{\bullet i}, A^*_{\bullet j}\rangle x^*_j + \langle A_{\bullet i}, \varepsilon\rangle,$$

and observe that, due to the closeness of $A_{\bullet i}$ and $A^*_{\bullet i}$, the first term is either close to x^*_i or equal to 0 depending on whether or not $i \in S$. Meanwhile, the rest are small due to the incoherence and the concentration in the weighted average of noise. We will show that both $Z_i = \sum_{S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\bullet j}\rangle x^*_j$ and $\langle A_{\bullet i}, \varepsilon\rangle$ are bounded by $C/8$ with high probability.

The cross-term $Z_i = \sum_{S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\bullet j}\rangle x^*_j$ is a sum of zero-mean independent sub-Gaussian random variables, which is another sub-Gaussian random variable with variance $\sigma^2_{Z_i} = \sum_{S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\bullet j}\rangle^2$. Note that

$$\langle A_{\bullet i}, A^*_{\bullet j}\rangle^2 \le 2\big(\langle A^*_{\bullet i}, A^*_{\bullet j}\rangle^2 + \langle A_{\bullet i} - A^*_{\bullet i}, A^*_{\bullet j}\rangle^2\big) \le 2\mu^2/n + 2\langle A_{\bullet i} - A^*_{\bullet i}, A^*_{\bullet j}\rangle^2,$$

where we use Cauchy-Schwarz inequality and the μ-incoherence of A^*. Therefore,

$$\sigma^2_{Z_i} \le 2\mu^2 k/n + 2\|A^{*T}_{\bullet S}(A_{\bullet i} - A^*_{\bullet i})\|^2_F \le 2\mu^2 k/n + 2\|A^*_{\bullet S}\|^2\|A_{\bullet i} - A^*_{\bullet i}\|^2 \le O(1/\log n),$$

under $\mu \le \frac{\sqrt{n}}{2k}$, to conclude $2\mu^2 k/n \le O(1/\log n)$ we need $1/k = O(1/\log n)$, i.e. $k = \Omega(\log n)$. Applying Bernstein's inequality, we get $|Z_i| \le C/8$ with high probability. What remains is to bound the noise term $\langle A_{\bullet i}, \varepsilon\rangle$. In fact, $\langle A_{\bullet i}, \varepsilon\rangle$ is sum of n Gaussian random variables, which is a sub-Gaussian with variance σ^2_ε. It is easy to see that $|\langle A_{\bullet i}, \varepsilon\rangle| \le \sigma_\varepsilon \log n$ with high probability. Notice that $\sigma_\varepsilon = O(1/\sqrt{n})$.

Finally, we combine these bounds to have $|Z_i + \langle A_{\bullet i}, \varepsilon\rangle| \le C/4$. Therefore, for $i \in S$, then $|\langle A_{\bullet i}, y\rangle| \ge C/2$ and negligible otherwise. Using union bound for every $i = 1, 2, \ldots, m$, we finish the proof of the Lemma. □

Lemma 8 enables us to derive the expected update direction $g^s = \mathbb{E}[(A^s x - y)\text{sgn}(x)^T]$ explicitly.

C.2 Approximate Gradient in Expectation

Proof [Proof of Lemma 5] Having the result from Lemma 8, we are now able to study the expected update direction $g^s = \mathbb{E}[(A^s x - y)\text{sgn}(x)^T]$. Recall that A^s is the update at the s-th iteration and $x \triangleq \text{threshold}_{C/2}((A^s)^T y)$. Based on the generative model, denote $p_i = \mathbb{E}[x_i^* \text{sgn}(x_i^*) | i \in S], q_i = \mathbb{P}[i \in S]$ and $q_{ij} = \mathbb{P}[i, j \in S]$. Throughout this section, we will use ζ to denote any vector whose norm is negligible although they can be different across their appearances. A_{-i} denotes the sub-matrix of A whose i-th column is removed. To avoid overwhelming appearance of the superscript s, we skip it from A^s for neatness. Denote $\mathcal{F}_{x^*}$ is the event under which the support of x is the same as that of x^*, and $\bar{\mathcal{F}}_{x^*}$ is its complement. In other words, $\mathbf{1}_{\mathcal{F}_{x^*}} = \mathbf{1}[\text{sgn}(x) = \text{sgn}(x^*)]$ and $\mathbf{1}_{\mathcal{F}_{x^*}} + \mathbf{1}_{\bar{\mathcal{F}}_{x^*}} = 1$.

$$g^s_{\bullet i} = \mathbb{E}[(Ax - y)\text{sgn}(x_i)] = \mathbb{E}[(Ax - y)\text{sgn}(x_i)\mathbf{1}_{\mathcal{F}_{x^*}}] \pm \zeta$$

Using the fact that $y = A^* x^* + \varepsilon$ and that under $\mathcal{F}_{x^*}$ we have $Ax = A_{\bullet S} x_S = A_{\bullet S} A^T_{\bullet S} y = A_{\bullet S} A^T_{\bullet S} A^* x^* + A_{\bullet S} A^T_{\bullet S} \varepsilon$. Using the independence of ε and x^* to get rid of the noise term, we get

$$\begin{aligned}
g^s_{\bullet i} &= \mathbb{E}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^* x^* \mathbf{1}_{\mathcal{F}_{x^*}}] + \mathbb{E}[(A_{\bullet S} A^T_{\bullet S} - I_n)\varepsilon \text{sgn}(x_i)\mathbf{1}_{\mathcal{F}_{x^*}}] \pm \zeta \\
&= \mathbb{E}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^* x^* \text{sgn}(x_i)\mathbf{1}_{\mathcal{F}_{x^*}}] \pm \zeta \quad \text{(Independence of } \varepsilon \text{ and } x\text{'s)} \\
&= \mathbb{E}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^* x^* \text{sgn}(x_i^*)(1 - \mathbf{1}_{\bar{\mathcal{F}}_{x^*}})] \pm \zeta \quad \text{(Under } \mathcal{F}_{x^*} \text{ event)} \\
&= \mathbb{E}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^* x^* \text{sgn}(x_i^*)] \pm \zeta
\end{aligned}$$

Recall from the generative model assumptions that $S = \text{supp}(x^*)$ is random and the entries of x^* are pairwise independent given the support, so

$$\begin{aligned}
g^s_{\bullet i} &= \mathbb{E}_S \mathbb{E}_{x^*|S}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^* x^* \text{sgn}(x_i^*)] \pm \zeta \\
&= p_i \mathbb{E}_{S, i \in S}[(A_{\bullet S} A^T_{\bullet S} - I_n) A^*_{\bullet i}] \pm \zeta \\
&= p_i \mathbb{E}_{S, i \in S}[(A_{\bullet i} A^T_{\bullet i} - I_n) A^*_{\bullet i}] + p_i \mathbb{E}_{S, i \in S}[\sum_{l \in S, l \neq i} A_{\bullet l} A^T_{\bullet l} A^*_{\bullet i}] \pm \zeta \\
&= p_i q_i (A_{\bullet i} A^T_{\bullet i} - I_n) A^*_{\bullet i} + p_i \sum_{l \in [m], l \neq i} q_{il} A_{\bullet l} A^T_{\bullet l} A^*_{\bullet i} \pm \zeta \\
&= p_i q_i (\lambda_i A_{\bullet i} - A^*_{\bullet i}) + p_i A_{\bullet -i} \text{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i} \pm \zeta
\end{aligned}$$

where $\lambda_i^s = \langle A^s_{\bullet i}, A^*_{\bullet i} \rangle$. Let $\xi_i^s = A_{R,-i} \text{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i} / q_i$ for $j = 1, \ldots, m$, we now have the full expression of the expected approximate gradient at iteration s:

$$g^s_{R,i} = p_i q_i (\lambda_i A^s_{R,i} - A^*_{R,i} + \xi_i^s) \pm \zeta_R. \tag{8}$$

What remains is to bound norms of ξ_s and ζ. We have $\|A^s_{R,-i}\| \leq \|A^s_{-i}\| \leq O(\sqrt{m/n})$ w.h.p. Then, along with the fact that $\|A^*_i\| = 1$, we can bound $\|\xi_i^s\|$

$$\|\xi_i^s\| \leq \|A^s_{R_i,-i}\| \max_{j \neq i} \frac{q_{ij}}{q_i} \|A^s_{-i}\| \leq O(k/n). \tag{9}$$

Next, we show that norm of ζ is negligible. In fact, $\mathcal{F}_{x^*}$ happens with very high probability, then it suffices to bound norm of $(Ax - y)\text{sgn}(x_i)$ which will be done using Lemma 12 and Lemma 11 in Section D. This concludes the proof for Lemma 5. ■

Appendix D. Sample Complexity

In previous sections, we rigorously analyzed both initialization and learning algorithms as if the expectations g^s, e and $M_{u,v}$ were given. Here we show that corresponding estimates based on empirical means are sufficient for the algorithms to succeed, and identify how may samples are required. Technically, this requires the study of their concentrations around their expectations. Having had these concentrations, we are ready to prove Theorems 4 and 5.

The entire section involves a variety of concentration bounds. Here we make heavy use of Bernstein's inequality for different types of random variables (including scalar, vector and matrix). The Bernstein's inequality is stated as follows.

Lemma 9 (Bernstein's Inequality) *Suppose that $Z^{(1)}, Z^{(2)}, \dots, Z^{(p)}$ are p i.i.d. samples from some distribution $\mathcal{D}$. If $\mathbb{E}[Z] = 0$, $\|Z^{(j)}\| \leq \mathcal{R}$ almost surely and $\|\mathbb{E}[Z^{(j)}(Z^{(j)})^T\| \leq \sigma^2$ for each j, then*

$$\frac{1}{p}\Big\|\sum_{j=1}^{p} Z^{(j)}\Big\| \leq \widetilde{O}\Big(\frac{\mathcal{R}}{p} + \sqrt{\frac{\sigma^2}{p}}\Big) \tag{10}$$

holds with probability $1 - n^{-\omega(1)}$.

Since all random variables (or their norms) are not bounded almost surely in our model setting, we make use of a technical lemma that is used in Arora et al. (2015) to handle the issue.

Lemma 10 (Arora et al. (2015)) *Suppose a random variable Z satisfies $\mathbb{P}[\|Z\| \geq \mathcal{R}(\log(1/\rho))^C] \leq \rho$ for some constant $C > 0$, then*

(a) If $p = n^{O(1)}$, it holds that $\|Z^{(j)}\| \leq \widetilde{O}(\mathcal{R})$ for each j with probability $1 - n^{-\omega(1)}$.

(b) $\|\mathbb{E}[Z\mathbf{1}_{\|Z\| \geq \widetilde{\Omega}(\mathcal{R})}]\| = n^{-\omega(1)}$.

This lemma suggests that if $\frac{1}{p}\sum_{i=1}^{p} Z^{(j)}(1 - \mathbf{1}_{\|Z^{(j)}\| \geq \widetilde{\Omega}(\mathcal{R})})$ concentrates around its mean with high probability, then so does $\frac{1}{p}\sum_{i=1}^{p} Z^{(j)}$ because the part outside the truncation level can be ignored. Since all random variables of our interest are sub-Gaussian or a product of sub-Gaussian that satisfy this lemma, we can apply Lemma 9 to the corresponding truncated random variables with carefully chosen truncation levels. Then the original random variables concentrate likewise.

In the next proofs, we define suitable random variables and identify good bounds of $\mathcal{R}$ and σ^2 for them. Note that in this section, the expectations are taken over y by conditioning on u and v. This aligns with the construction that the estimators of e and $M_{u,v}$ are empirical averages over i.i.d. samples of y, while u and v are kept fixed. Due to the dependency on u and v, these (conditional) expectations inherit randomness from u and v, and we will formulate probabilistic bounds for them.

The application of Bernstein's inequality requires a bound on $\|\mathbb{E}[ZZ^T(1 - \mathbf{1}_{\|Z\| \geq \widetilde{\Omega}(\mathcal{R}))}]\|$. We achieve that by the following technical lemma, where $\tilde{Z}$ is a standardized version of Z.

Lemma 11 *Suppose a random variable $\tilde{Z}\tilde{Z}^T = aT$ where $a \geq 0$ and T is positive semi-definite. They are both random. Suppose $\mathbb{P}[a \geq \mathcal{A}] = n^{-\omega(1)}$ and $\mathcal{B} > 0$ is a constant. Then,*

$$\|\mathbb{E}[\tilde{Z}\tilde{Z}^T(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\| \leq \mathcal{A}\|\mathbb{E}[T]\| + O(n^{-\omega(1)})$$

Proof To show this, we make use of the decomposition $\tilde{Z}\tilde{Z}^T = aT$ and a truncation for a. Specifically,

$$\begin{aligned}
\|\mathbb{E}[\tilde{Z}\tilde{Z}^T(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\| &= \mathbb{E}[aT(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})] \\
&\leq \|\mathbb{E}[a(1 - \mathbf{1}_{a \geq \mathcal{A}})T(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\| + \|\mathbb{E}[a\mathbf{1}_{a \geq \mathcal{A}}T(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\| \\
&\leq \|\mathbb{E}[a(1 - \mathbf{1}_{a \geq \mathcal{A}})T]\| + \mathbb{E}[a\mathbf{1}_{a \geq \mathcal{A}}\|T\|(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})] \\
&\leq \mathcal{A}\|\mathbb{E}[T]\| + \big(\mathbb{E}[\|aT\|^2(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\mathbb{E}[\mathbf{1}_{a \geq \mathcal{A}}]\big)^{1/2} \\
&\leq \mathcal{A}\|\mathbb{E}[T]\| + \big(\mathbb{E}[\|\tilde{Z}\|^4(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})]\mathbb{P}[a \geq \mathcal{A}]\big)^{1/2} \\
&\leq \mathcal{A}\|\mathbb{E}[T]\| + \mathcal{B}^2\big(\mathbb{P}[a \geq \mathcal{A}]\big)^{1/2} \\
&\leq \mathcal{A}\|\mathbb{E}[T]\| + O(n^{-\omega(1)}),
\end{aligned}$$

where at the third step we used $T(1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}})] \preceq T$ because of the fact that T is the positive semi-definite and $1 - \mathbf{1}_{\|\tilde{Z}\| \geq \mathcal{B}} \in \{0, 1\}$. Then, we finish the proof of the lemma. ■

D.1 Sample Complexity of Algorithm 1

In Algorithm 1, we empirically compute the "scores" $\widehat{e}$ and the reduced weighted covariance matrix $\widehat{M}_{u,v}$ to produce an estimate for each column of A^*. Since the construction of $\widehat{M}_{u,v}$ depends upon the support estimate $\widehat{R}$ given by ranking $\widehat{e}$, we denote it by $\widehat{M}^{\widehat{R}}_{u,v}$. We will show that we only need $p = \widetilde{O}(m)$ samples to be able to recover the support of one particular atom and up to some specified level of column-wise error with high probability.

Lemma 12 *Consider Algorithm 1 in which p is the given number of samples. For any pair u and v, then with high probability a) $\|\widehat{e} - e\| \leq O^*(k/m \log^2 n)$ when $p = \widetilde{\Omega}(m)$ and b) $\|\widehat{M}^{\widehat{R}}_{u,v} - M^R_{u,v}\| \leq O^*(k/m \log n)$ when $p = \widetilde{\Omega}(mr)$ where $\widehat{R}$ and R are respectively the estimated and correct support sets of one particular atom.*

D.1.1 Proof of Theorem 4

Using Lemma 12, we are ready to prove the Theorem 4. According to Lemma 1 when $U \cap V = \{i\}$, we can write $\widehat{e}$ as

$$\widehat{e} = q_i c_i \beta_i \beta_i' A^*_{R,i} \circ A^*_{R,i} + \text{ perturbation terms} + (\widehat{e} - e),$$

and consider $\widehat{e} - e$ as an additional perturbation with the same magnitude $O^*(k/m \log^2 n)$ in the sense of $\|\cdot\|_\infty$ w.h.p. The first part of Lemma 3 suggests that when u and v share exactly one atom i, then the set $\widehat{R}$ including r largest elements of $\widehat{e}$ is the same as $\text{supp}(A^*_i)$ with high probability.

Once we have $\widehat{R}$, we again write $\widehat{M}^{\widehat{R}}_{u,v}$ using Lemma 2 as

$$\widehat{M}^{\widehat{R}}_{u,v} = q_i c_i \beta_i \beta_i' A^*_{R,i} A^{*T}_{R,i} + \text{ perturbation terms} + (\widehat{M}^{\widehat{R}}_{u,v} - M^R_{u,v}),$$

and consider $\widehat{M}^{\widehat{R}}_{u,v} - M^R_{u,v}$ as an additional perturbation with the same magnitude $O^*(k/m \log n)$ in the sense of the spectral norm $\|\cdot\|$ w.h.p. Using the second part of Lemma 3, we have the top singular vectors of $\widehat{M}^{\widehat{R}}_{u,v}$ is $O^*(1/\log n)$ -close to $A^*_{R,i}$ with high probability.

Since every vector added to the list L in Algorithm 1 is close to one of the dictionary, then A^0 must be δ-close to A^*. In addition, the nearness ofA^0 to A^* is guaranteed via an appropriate projection onto the convex set $\mathcal{B} = \{A | A \text{ close to } A^0 \text{ and } \|A\| \leq 2\|A^*\|\}$. Finally, we finish the proof of Theorem 4. □

D.1.2 Proof of Lemma 12, Part a

For some fixed $l \in [n]$, consider p i.i.d. realizations $Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}$ of the random variable $Z \triangleq \langle y, u\rangle\langle y, v\rangle y_l^2$, then $\widehat{e}_l = \frac{1}{p}\sum_{i=1}^p Z^{(i)}$ and $e_l = \mathbb{E}[Z]$. To show that $\|\widehat{e} - e\|_\infty \leq O^*(k/m \log^2 n)$ holds with high probability, we first study the concentration for the l-th entry of $\widehat{e} - e$ and then take the union bound over all $l = 1, 2, \ldots, n$. We derive upper bounds for $|Z|$ and its variance $\mathbb{E}[Z^2]$ in order to apply Bernstein's inequality in (12) to the truncated version of Z.

Claim 8 *$|Z| \leq \widetilde{O}(k)$ and $\mathbb{E}[Z^2] \leq \widetilde{O}(k^2/m)$ with high probability.*

Again, the expectation is taken over y by conditioning on u and v, and therefore is still random due to the randomness of u and v. To show Claim 8, we begin with proving the following auxiliary claim.

Claim 9 *$\|y\| \leq \widetilde{O}(\sqrt{k})$ and $|\langle y, u\rangle| \leq \widetilde{O}(\sqrt{k})$ with high probability.*

Proof From the generative model, we have

$$\|y\| = \|A^*_{\bullet S} x^*_S + \varepsilon\| \leq \|A^*_{\bullet S} x^*_S\| + \|\varepsilon\| \leq \|A^*_{\bullet S}\|\|x^*_S\| + \|\varepsilon\|,$$

where $S = \text{supp}(x^*)$. From Claim 2, $\|x^*_S\| \leq \widetilde{O}(\sqrt{k})$ and $\|\varepsilon\| \leq \widetilde{O}(\sigma_\varepsilon\sqrt{n})$ w.h.p. In addition, A^* is overcomplete and has bounded spectral norm, then $\|A^*_{\bullet S}\| \leq \|A^*\| \leq O(1)$. Therefore, $\|y\| \leq \widetilde{O}(\sqrt{k})$ w.h.p., which is the first part of the proof. To bound the second term, we write it as

$$|\langle y, u\rangle| = |\langle A^*_{\bullet S} x^*_S + \varepsilon, u\rangle| \leq |\langle x^*_S, A^{*T}_{\bullet S} u\rangle| + |\langle \varepsilon, u\rangle|.$$

Similar to y, we have $\|u\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. and hence $\|A^{*T}_{\bullet S} u\| \leq \|A^{*T}_{\bullet S}\|\|u\| \leq O(\sqrt{k})$ with high probability. Since u and x^* are independent sub-Gaussian and $\langle x^*_S, A^{*T}_{\bullet S} u\rangle$ are sub-exponential with variance at most $O(\sqrt{k})$, $|\langle x^*_S, A^{*T}_{\bullet S} u\rangle| \leq \widetilde{O}(k)$ w.h.p. Similarly, $|\langle \varepsilon, u\rangle| \leq \widetilde{O}(\sqrt{k})$ w.h.p. Consequently, $|\langle y, u\rangle| \leq \widetilde{O}(\sqrt{k})$ w.h.p., and we conclude the proof of the claim. ■

Proof [Proof of Claim 8] We have $Z = \langle y, u\rangle\langle y, v\rangle y_l^2 = \langle y, u\rangle\langle y, v\rangle(\langle A^*_{l\bullet}, x^*\rangle + \varepsilon_l)^2$ with $\langle y, u\rangle\langle y, v\rangle \leq \widetilde{O}(k)$ w.h.p. according to Claim 9. What remains is to bound $y_l^2 = (\langle A^*_{l\bullet}, x^*\rangle + \varepsilon_l)^2$. Because $\langle A^*_{l\bullet}, x^*\rangle$ is sub-Gaussian with variance $\mathbb{E}_S(\sum_{i\in S} A^{*2}_{li}) \leq \|A^{*T}\|^2_{1,2} = O(1)$,

then $|\langle A^*_{l\bullet}, x^*\rangle| \leq O(\log n)$ w.h.p. Similarly for ε_l, $|\varepsilon_l| \leq O(\sigma_\varepsilon \log n)$ w.h.p. Ultimately, $|\langle A^*_{l\bullet}, x^*\rangle + \varepsilon_l| \leq O(\log n)$, and hence we obtain with high probability the bound $|Z| \leq \widetilde{O}(k)$.

To bound the variance term, we write $Z^2 = \langle y, v\rangle^2 y_l^2 \langle y, u\rangle^2 y_l^2$. Note that, from the first part, we get $\langle y, v\rangle^2 y_l^2 \leq \widetilde{O}(k)$ and $|Z| \leq \widetilde{O}(k)$ w.h.p.. We apply Lemma 11 with some appropriate scaling to both terms, then

$$\mathbb{E}[Z^2(1 - \mathbf{1}_{|Z|\geq\widetilde{\Omega}(k)})] \leq \widetilde{O}(k)\mathbb{E}[\langle y, u\rangle^2 y_l^2] + O(n^{-\omega(1)}),$$

where $\mathbb{E}[\langle y, u\rangle^2 y_l^2]$ is equal to e_l for pair u, v with $v = u$. From Lemma 1 and its proof in Appendix Section "Analysis of Initialization Algorithm",

$$\mathbb{E}[\langle y, u\rangle^2 y_l^2] = \sum_{i=1}^{m} q_i c_i \beta_i^2 A_{li}^{*2} + \text{ perturbation terms},$$

in which the perturbation terms are bounded by $O^*(k/m \log^2 n)$ w.h.p. (following Claims 4 and 5). The dominant term $\sum_i q_i c_i \beta_i^2 A_{li}^{*2} \leq (\max q_i c_i \beta_i^2)\|A^*_{l\bullet}\|^2 \leq \widetilde{O}(k/m)$ w.h.p. because $|\beta_i| \leq O(\log m)$ (Claim 3). Then we complete the proof of the second part. ■

Proof [Proof of Lemma 12, Part a] We are now ready to prove Part a of Lemma 12. We apply Bernstein's inequality in Lemma 9 for the truncated random variable $Z^{(i)}(1 - \mathbf{1}_{|Z^{(i)}|\geq\widetilde{\Omega}(\mathcal{R})})$ with $\mathcal{R} = \widetilde{O}(k)$ and variance $\sigma^2 = \widetilde{O}(k^2/m)$ from Claim 8, then

$$\left\| \frac{1}{p}\sum_{i=1}^{p} Z^{(i)}(1 - \mathbf{1}_{|Z^{(i)}|\geq\widetilde{\Omega}(\mathcal{R})}) - \mathbb{E}[Z(1 - \mathbf{1}_{|Z|\geq\widetilde{\Omega}(\mathcal{R})})] \right\| \leq \frac{\widetilde{O}(k)}{p} + \sqrt{\frac{\widetilde{O}(k^2/m)}{p}} \leq O^*(k/m \log n), \tag{11}$$

w.h.p. for $p = \widetilde{\Omega}(m)$. Then $\widehat{e}_l = \frac{1}{p}\sum_{i=1}^{p} Z^{(i)}$ also concentrates with high probability. Take the union bound over $l = 1, 2, \ldots, n$, we get $\|\widehat{e} - e\|_\infty \leq O^*(k/m \log n)$ with high probability and complete the proof of 12, Part a. □

D.1.3 Proof of Lemma 12, Part b

Next, we will prove that $\|\widehat{M}^{\widehat{R}}_{u,v} - M^R_{u,v}\| \leq O^*(k/m \log n)$ with high probability. We only need to prove the concentration inequalities for the case when conditioned on the event that $\widehat{R}$ is equivalent to R w.h.p. Again, what we need to derive are an upper norm bound $\mathcal{R}$ of the matrix random variable $Z \triangleq \langle y, u\rangle\langle y, v\rangle y_R y_R^T$ and its variance.

Claim 10 *$\|Z\| \leq \widetilde{O}(kr)$ and $\|\mathbb{E}[ZZ^T]\| \leq \widetilde{O}(k^2 r/m)$ hold with high probability.*

Proof We have $\|Z\| \leq |\langle y, u\rangle\langle y, v\rangle|\|y_R\|^2$ with $|\langle y, u\rangle\langle y, v\rangle| \leq \widetilde{O}(k)$ w.h.p. (according to Claim 9) whereas $\|y_R\|^2 = \sum_{i\in R} y_l^2 \leq O(r \log^2 n)$ w.h.p. because $y_l \leq O(\log n)$ w.h.p. (proof of Claim 8). This implies $\|Z\| \leq \widetilde{O}(kr)$ w.h.p. The second part is handled similarly as in the proof of Claim 8. We take advantage of the bounds of $\widehat{M}_{u,v}$ in Lemma 2. Specifically, using the first part $\|Z\| \leq \widetilde{O}(kr)$ and $\langle y, v\rangle^2\|y_R\|^2 \leq \widetilde{O}(kr)$, and applying Lemma 11, then

$$\|\mathbb{E}[ZZ^T(1 - \mathbf{1}_{\|Z\|\geq\widetilde{\Omega}(kr)})]\| \leq \widetilde{O}(kr)\|\mathbb{E}[\langle y, u\rangle^2 y_R y_R^T]\| + \widetilde{O}(kr)O(n^{-\omega(1)}) \leq \widetilde{O}(kr)\|M_{u,u}\|,$$

where $M_{u,u}$ arises from the application of Lemma 2. Recall that

$$M_{u,u} = \sum_i q_i c_i \beta_i^2 A^*_{R,i} A^{*T}_{R,i} + \text{ perturbation terms},$$

where the perturbation terms are all bounded by $O^*(k/m \log n)$ w.h.p. by Claims 6 and 7. In addition,

$$\|\sum_i q_i c_i \beta_i^2 A^*_{R,i} A^{*T}_{R,i}\| \leq (\max_i q_i c_i \beta_i^2) \|A^*_{R\bullet}\|^2 \leq \widetilde{O}(k/m)\|A^*\|^2 \leq \widetilde{O}(k/m)$$

w.h.p. Finally, the variance bound is $\widetilde{O}(k^2 r/m)$ w.h.p. ■

Then, applying Bernstein's inequality in Lemma 9 to the truncated version of Z with $\mathcal{R} = \widetilde{O}(kr)$ and variance $\sigma^2 = \widetilde{O}(k^2 r/m)$ and obtain the concentration for the full Z to get

$$\|\widehat{M}^R_{u,v} - M^R_{u,v}\| \leq \frac{\widetilde{O}(kr)}{p} + \sqrt{\frac{\widetilde{O}(k^2 r/m)}{p}} \leq O^*(k/m \log n)$$

w.h.p. when the number of samples is $p = \widetilde{\Omega}(mr)$ under Assumption **A4.1**.

We have proved that $\|\widehat{M}^R_{u,v} - M^R_{u,v}\| \leq O^*(k/m \log n)$ as conditioned on the support consistency event holds w.h.p. $\|\widehat{M}^{\widehat{R}}_{u,v} - M^R_{u,v}\| \leq O^*(k/m \log n)$ is easily followed by the law of total probability through the tail bounds on the conditional and marginal probabilities (i.e. $\mathbb{P}[\|\widehat{M}^R_{u,v} - M^R_{u,v}\| \leq O^*(k/m \log n) | \widehat{R} = R]$) and $\mathbb{P}[\widehat{R} \neq R]$. We finish the proof of Lemma 12, Part b for both cases of the spectral bounds. ■

D.2 Proof of Theorem 5 and Sample Complexity of Algorithm 2

In this section, we prove Theorem 5 and identify sample complexity per iteration of Algorithm 2. We divide the proof into two steps: 1) show that when A^s is $(\delta_s, 2)$-near to A^* for $\delta_s = O^*(1/\log n)$, the approximate gradient estimate $\widehat{g}^s$ is $(\alpha, \beta, \gamma_s)$-correlated-whp with A^* with $\gamma_s \leq O(k^2/mn) + \alpha o(\delta_s^2)$, and 2) show that the nearness is preserved at each iteration. These correspond to showing the following lemmas:

Lemma 13 *At iteration s of Algorithm 2, suppose that A^s has each column correctly supported and is $(\delta_s, 2)$-near to A^* and that $\eta = O(m/k)$. Denote $R = \text{supp}(A^s_{\bullet i})$, then the update $\widehat{g}^s_{R,i}$ is $(\alpha, \beta, \gamma_s)$-correlated-whp with $A^*_{R,i}$ where $\alpha = \Omega(k/m)$, $\beta = \Omega(m/k)$ and $\gamma_s \leq O(k^2/mn) + \alpha o(\delta_s^2)$ for $\delta_s = O^*(1/\log n)$.*

Note that this is a finite-sample version of Lemma 6.

Lemma 14 *If A^s is $(\delta_s, 2)$-near to A^* and number of samples used in step s is $p = \widetilde{\Omega}(m)$, then with high probability $\|A^{s+1} - A^*\| \leq 2\|A^*\|$.*

Proof [Proof of Theorem 5] The correlation of $\widehat{g}_i$ with A^*_i, described in Lemma 13, implies the descent of column-wise error according to Theorem 1. Along with Lemma 14, the theorem follows directly.

D.2.1 Proof of Lemma 13

We prove Lemma 13 by obtaining a tail bound on the difference between $\widehat{g}^s_{R,i}$ and $g^s_{R,i}$ using the Bernstein's inequality in Lemma 9.

Lemma 15 *At iteration s of Algorithm 2, suppose that A^s has each column correctly supported and is $(\delta_s, 2)$-near to A^*. For $R = \mathrm{supp}(A^s_i) = \mathrm{supp}(A^*_i)$, then $\|\widehat{g}^s_{R,i} - g^s_{R,i}\| \leq O(k/m) \cdot (o(\delta_s) + O(\epsilon_s))$ with high probability for $\delta_s = O^*(1/\log n)$ and $\epsilon_s = O(\sqrt{k/n})$ when $p = \widetilde{\Omega}(m + \sigma_\varepsilon^2 \frac{mnr}{k})$.*

To prove this lemma, we study the concentration of $\widehat{g}^s_{R,i}$, which is a sum of random vector of the form $(y - Ax)_R \mathrm{sgn}(x_i)$. We consider random variable $Z \triangleq (y - Ax)_R \mathrm{sgn}(x_i) | i \in S$, with $S = \mathrm{supp}(x^*)$ and $x = \mathrm{threshold}_{C/2}(A^T y)$. Then, using the following technical lemma to bridge the gap in concentration of the two variables. We adopt this strategy from Arora et al. (2015) for our purpose.

Claim 11 *Suppose that $Z^{(1)}, Z^{(2)}, \ldots, Z^{(N)}$ are i.i.d. samples of the random variable $Z = (y - Ax)_R \mathrm{sgn}(x_i) | i \in S$. Then,*

$$\Big\| \frac{1}{N} \sum_{j=1}^{N} Z^{(j)} - \mathbb{E}[Z] \Big\| \leq o(\delta_s) + O(\epsilon_s) \tag{12}$$

holds with probability when $N = \widetilde{\Omega}(k + \sigma_\varepsilon^2 nr)$, $\delta_s = O^*(1/\log n)$ and $\epsilon_s = O(\sqrt{k/n})$.

Proof [Proof of Lemma 15] Once we have done the proof of Claim 11, we can easily prove Lemma 15. We recycle the proof of Lemma 43 in Arora et al. (2015).

Write $W = \{j : i \in \mathrm{supp}(x^{*(j)})\}$ and $N = |W|$, then express $\widehat{g}_{R,i}$ as

$$\widehat{g}_{R,i} = \frac{N}{p} \frac{1}{N} \sum_j (y^{(j)} - Ax^{(j)})_R \mathrm{sgn}(x_i^{(j)}),$$

where $\frac{1}{|W|} \sum_j (y^{(j)} - Ax^{(j)})_R \mathrm{sgn}(x_i^{(j)})$ is distributed as $\frac{1}{N} \sum_{j=1}^{N} Z^{(j)}$ with $N = |W|$. Note that $\mathbb{E}[(y - Ax)_R \mathrm{sgn}(x_i)] = \mathbb{E}[(y - Ax)_R \mathrm{sgn}(x_i) \mathbf{1}_{i \in S}] = \mathbb{E}[Z] \mathbb{P}[i \in S] = q_i \mathbb{E}[Z]$ with $q_i = \Theta(k/m)$. Following Claim 11, we have

$$\|\widehat{g}^s_{R,i} - g^s_{R,i}\| \leq O(k/m) \Big\| \frac{1}{N} \sum_{j=1}^{N} Z^{(j)} - \mathbb{E}[Z] \Big\| \leq O(k/m) \cdot (o(\delta_s) + O(\epsilon_s)),$$

holds with high probability as $p = \Omega(mN/k)$. Substituting N in Claim 11, we obtain the results in Lemma 15. □

Proof [Proof of Claim 11] We are now ready to prove the claim. What we need are good bounds for $\|Z\|$ and its variance, then we can apply Bernstein's inequality in Lemma 9 for the truncated version of Z, then Z is also concentrates likewise.

Claim 12 *$\|Z\| \leq \mathcal{R}$ holds with high probability for $\mathcal{R} = \widetilde{O}(\delta_s \sqrt{k} + \mu k/\sqrt{n} + \sigma_\varepsilon \sqrt{r})$ with $\delta_s = O^*(1/\log n)$.*

Proof From the generative model and the support consistency of the encoding step, we have $y = A^*x^* + \varepsilon = A^*_{\bullet S}x^*_S + \varepsilon$ and $x_S = A^T_{\bullet S}y = A^T_{\bullet S}A^*_{\bullet S}x^*_S + A^T_{\bullet S}\varepsilon$. Then,

$$\begin{aligned}(y - Ax)_R &= (A^*_{R,S}x^*_S + \varepsilon_R) - A_{R,S}A^T_{\bullet S}A^*_{\bullet S}x^*_S - A_{R,S}A^T_{\bullet S}\varepsilon \\ &= (A^*_{R,S} - A_{R,S})x^*_S + A_{R,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})x^*_S + (I_n - A_{\bullet S}A^T_{\bullet S})_{R\bullet}\varepsilon.\end{aligned}$$

Using the fact that x^*_S and ε are sub-Gaussian and that $\|Mw\| \leq \widetilde{O}(\sigma_w\|M\|_F)$ holds with high probability for a fixed M and a sub-Gaussian w of variance σ_w^2, we have

$$\|(y-Ax)_R\mathrm{sgn}(x_i)\| \leq \widetilde{O}(\|A^*_{R,S}-A_{R,S}\|_F+\|A_{R,S}(I_k-A^T_{\bullet S}A^*_{\bullet S})\|_F+\sigma_\varepsilon\|(I_n-A_{\bullet S}A^T_{\bullet S})_{R\bullet}\|_F).$$

Now, we need to bound those Frobenius norms. The first quantity is easily bounded as

$$\|A^*_{R,S} - A_{R,S}\|_F \leq \|A^*_{\bullet S} - A_{\bullet S}\|_F \leq \delta_s\sqrt{k}, \tag{13}$$

since A is δ_s-close to A^*. To handle the other two, we use the fact that $\|UV\|_F \leq \|U\|\|V\|_F$. Using this fact for the second term, we have

$$\|A_{R,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F \leq \|A_{R,S}\|\|(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F,$$

where $\|A_{R,S}\| \leq \|A_{R\bullet}\| \leq O(1)$ due to the nearness. The second part is rearranged to take advantage of the closeness and incoherence properties:

$$\begin{aligned}\|I_k - A^T_{\bullet S}A^*_{\bullet S}\|_F &\leq \|I_k - A^{*T}_{\bullet S}A^*_{\bullet S} - (A_{\bullet S} - A^*_{\bullet S})^TA^*_{\bullet S}\|_F \\ &\leq \|I_k - A^{*T}_{\bullet S}A^*_{\bullet S}\|_F + \|(A_{\bullet S} - A^*_{\bullet S})^TA^*_{\bullet S}\|_F \\ &\leq \|I_k - A^{*T}_{\bullet S}A^*_{\bullet S}\|_F + \|A^*_{\bullet S}\|\|A_{\bullet S} - A^*_{\bullet S}\|_F \\ &\leq \mu k/\sqrt{n} + O(\delta_s\sqrt{k}),\end{aligned}$$

where we have used $\|I_k - A^{*T}_{\bullet S}A^*_{\bullet S}\|_F \leq \mu k/\sqrt{n}$ because of the μ-incoherence of A^*, $\|A_{\bullet S} - A^*_{\bullet S}\|_F \leq \delta_s\sqrt{k}$ in (13) and $\|A^*_{\bullet S}\| \leq \|A^*\| \leq O(1)$. Accordingly, the second Frobenius norm is bounded by

$$\|A_{R,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F \leq O\big(\mu k/\sqrt{n} + \delta_s\sqrt{k}\big). \tag{14}$$

The noise term is handled using the eigen-decomposition $U\Lambda U^T$ of $A_{\bullet S}A^T_{\bullet S}$, then with high probability

$$\|(I_n-A_{\bullet S}A^T_{\bullet S})_{R\bullet}\|_F = \|(UU^T-U\Lambda U^T)_{R\bullet}\|_F = \|U_{R\bullet}(I_n-\Lambda)\|_F \leq \|I_n-\Lambda\|\|U_{R\bullet}\|_F \leq O(\sqrt{r}), \tag{15}$$

where the last inequality $\|I_n - \Lambda\| \leq O(1)$ follows by $\|A_{\bullet S}\| \leq \|A\| \leq \|A - A^*\| + \|A^*\| \leq 3\|A^*\| \leq O(1)$ due to the nearness. Putting (13), (14) and (15) together, we obtain the bounds in Claim 12. ■

Next, we determine a bound for the variance of Z.

Claim 13 $\mathbb{E}[\|Z\|^2] = \mathbb{E}[\|(y - Ax)_R\mathrm{sgn}(x_i)\|^2 | i \in S] \leq \sigma^2$ *holds with high probability for* $\sigma^2 = O(\delta_s^2k + k^2/n + \sigma_\varepsilon^2 r)$ *with* $\delta_s = O^*(1/\log n)$.

Proof We explicitly calculate the variance using the fact that x_S^* is conditionally independent given S, and so is ε. x_S^* and ε are also independent and have zero mean. Then we can decompose the norm into three terms in which the dot product is zero in expectation and the others can be shortened using the fact that $E[x_S^* x_S^{*T}] = I_k$, $E[\varepsilon\varepsilon^T] = \sigma_\varepsilon I_n$.

$$\begin{aligned}\mathbb{E}[\|(y - Ax)_R \mathrm{sgn}(x_i)\|^2 | i \in S] &= \mathbb{E}[\|(A_{R,S}^* - A_{R,S}A_{\bullet S}^T A_{\bullet S}^*)x_S^* + (I_n - A_{\bullet S}A_{\bullet S}^T)_{R\cdot}\varepsilon\|^2 | i \in S]] \\ &= \mathbb{E}[\|A_{R,S}^* - A_{R,S}A_{\bullet S}^T A_{\bullet S}^*\|_F^2 | i \in S] + \sigma_\varepsilon^2 \mathbb{E}[\|I_n - A_{\bullet S}A_{\bullet S}^T)_{R\bullet}\|_F^2 | i \in S].\end{aligned}$$

Then, by re-writing $A_{R,S}^* - A_{R,S}A_{\bullet S}^T A_{\bullet S}^*$ as before, we get the form $(A_{R,S}^* - A_{R,S}) + A_{R,S}(I_k - A_{\bullet S}^T A_{\bullet S}^*)$ in which the first term has norm bounded by $\delta_s\sqrt{k}$. The second is further decomposed as

$$\mathbb{E}[\|A_{R,S}(I_k - A_{\bullet S}^T A_{\bullet S}^*)\|_F^2 | i \in S] \leq \sup_S \|A_{R,S}\|^2 \mathbb{E}[\|I_k - A_{\bullet S}^T A_{\bullet S}^*\|_F^2 | i \in S], \tag{16}$$

where $\sup_S \|A_{R,S}\| \leq \|A_{R\bullet}\| \leq O(1)$. We will bound $\mathbb{E}[\|I_k - A_{\bullet S}^T A_{\bullet S}^*\|_F^2 | i \in S] \leq O(k\delta_s^2) + O(k^2/n)$ using the proof from Arora et al. (2015):

$$\begin{aligned}&\mathbb{E}[\|I_k - A_{\bullet S}^T A_{\bullet S}^*\|_F^2 | i \in S] = \mathbb{E}[\sum_{j\in S}(1 - A_{\bullet j}^T A_{\bullet j}^*)^2 + \sum_{j \in S}\|A_{\bullet j}^T A_{\bullet,-j}^*\|^2 | i \in S] \\ &= \mathbb{E}[\sum_{j\in S}\frac{1}{4}\|A_{\bullet j} - A_{\bullet j}^*\|^2] + q_{ij}\sum_{j\neq i}\|A_{\bullet j}^T A_{\bullet,-j}^*\|^2 + q_i\|A_{\bullet i}^T A_{\bullet,-i}^*\|^2 + q_i\|A_{\bullet,-i}^T A_{\bullet i}^*\|^2,\end{aligned}$$

where $A_{\bullet,-i}$ is the matrix A with the i-th column removed, $q_{ij} \leq O(k^2/m^2)$ and $q_i \leq O(k/m)$. For any $j = 1, 2, \ldots, m$,

$$\begin{aligned}\|A_{\bullet j}^T A_{\bullet,-j}^*\|^2 &= \|A_{\bullet j}^{*T} A_{\bullet,-j}^* + (A_{\bullet j} - A_{\bullet j}^*)^T A_{\bullet,-j}^*\|^2 \\ &\leq \sum_{l\neq j}\langle A_{\bullet j}^*, A_{\bullet l}^*\rangle^2 + \|(A_{\bullet j} - A_{\bullet j}^*)^T A_{\bullet,-j}^*\|^2 \\ &\leq \sum_{l\neq j}\langle A_{\bullet j}^*, A_{\bullet l}^*\rangle^2 + \|A_{\bullet j} - A_{\bullet j}^*\|^2\|A_{\bullet,-j}^*\|^2 \leq \mu^2 + \delta_s^2.\end{aligned}$$

The last inequality invokes the μ-incoherence, δ-closeness and the spectral norm of A^*. Similarly, we come up with the same bound for $\|A_{\bullet i}^T A_{\bullet,-i}^*\|^2$ and $\|A_{\bullet,-i}^T A_{\bullet i}^*\|^2$. Consequently,

$$\mathbb{E}[\|I_k - A_{\bullet S}^T A_{\bullet S}^*\|_F^2 | i \in S] \leq O(k\delta_s^2) + O(k^2/n). \tag{17}$$

For the last term, we invoke the inequality (15) (Claim 12) to get

$$\mathbb{E}[\|(I_n - A_{\bullet S}A_{\bullet S}^T)_{R\bullet}\|_F^2 | i \in S] \leq r \tag{18}$$

Putting (16), (17) and (18) together and using $\|A_{R\bullet}\| \leq 1$, we obtain the variance bound of Z: $\sigma^2 = O(\delta_s^2 k + k^2/n + \sigma_\varepsilon^2 r)$ with $\delta_s = O^*(1/\log n)$. Finally, we complete the proof. □

We now apply truncated Bernstein's inequality to the random variable $Z^{(j)}(1 - 1_{\|Z^{(j)}\| \geq \Omega(\mathcal{R})})$ with $\mathcal{R}$ and σ^2 in Claims 12 and 13, which are $\mathcal{R} = \widetilde{O}(\delta_s\sqrt{k} + \mu k/\sqrt{n} + \sigma_\varepsilon\sqrt{r})$ and $\sigma^2 = O(\delta_s^2 k + k^2/n + \sigma_\varepsilon^2 r)$. Then, $(1/N)\sum_{j=}^N Z^{(j)}$ also concentrates:

$$\Big\|\frac{1}{N}\sum_{i=1}^N Z^{(j)} - E[Z]\Big\| \leq \widetilde{O}\Big(\frac{\mathcal{R}}{N}\Big) + \widetilde{O}\Bigg(\sqrt{\frac{\sigma^2}{N}}\Bigg) = o(\delta_s) + O(\sqrt{k/n})$$

holds with high probability when $N = \widetilde{\Omega}(k + \sigma_\varepsilon^2 nr)$. Then, we finally finish the proof of Claim 11. ■

Proof [Proof of Lemma 13] With Claim 11, we study the concentration of $\widehat{g}^s_{R,i}$ around its mean $g^s_{R,i}$. Now, we consider this difference as an error term of the expectation $g^s_{R,i}$ and using Lemma 6 to show the correlation of $\widehat{g}^s_{R,i}$. Using the expression in Lemma 5 with high probability, we can write

$$\widehat{g}^s_{R,i} = g^s_{R,i} + (g^s_{R,i} - \widehat{g}^s_{R,i}) = 2\alpha(A_{R,i} - A^*_{R,i}) + v,$$

where $\|v\| \leq \alpha\|A_{R,i} - A^*_{R,i}\| + O(k/m) \cdot (o(\delta_s) + O(\epsilon_s))$. By Lemma 6, we have $\widehat{g}^s_{R,i}$ is $(\alpha, \beta, \gamma_s)$-correlated-whp with $A^*_{R,i}$ where $\alpha = \Omega(k/m)$, $\beta = \Omega(m/k)$ and $\gamma_s \leq O(k/m) \cdot (o(\delta_s) + O(\sqrt{k/n}))$, then we have done the proof Lemma 13. □

D.2.2 Proof of Lemma 14

We have shown the correlation of $\widehat{g}^s$ with A^* w.h.p. and established the descent property of Algorithm 2. The next step is to show that the nearness is preserved at each iteration. To prove $\|A^{s+1} - A^*\| \leq 2\|A^*\|$ holds with high probability, we recall the update rule

$$A^{s+1} = A^s - \eta \mathcal{P}_H(\widehat{g}^s),$$

where $\mathcal{P}_H(\widehat{g}^s) = H \circ \widehat{g}^s$. Here $H = (h_{ij})$ where $h_{ij} = 1$ if $i \in \mathrm{supp}(A_{\bullet j})$ and $h_{ij} = 0$ otherwise. Also, note that A^s is $(\delta_s, 2)$-near to A^* for $\delta_s = O^*(1/\log n)$. We already proved that this holds for the exact expectation g^s in Lemma 7. To prove for $\widehat{g}^s$, we again apply matrix Bernstein's inequality to bound $\|\mathcal{P}_H(g^s) - \mathcal{P}_H(\widehat{g}^s)\|$ by $O(k/m)$ because $\eta = \Theta(m/k)$ and $\|A^*\| = O(1)$.

Consider a matrix random variable $Z \triangleq \mathcal{P}_H((y - Ax)\mathrm{sgn}(x)^T)$. Our goal is to bound the spectral norm $\|Z\|$ and, both $\|\mathbb{E}[ZZ^T]\|$ and $\|\mathbb{E}[Z^TZ]\|$ since Z is asymmetric. To simplify our notations, we denote by x_R the vector x by zeroing out the elements not in R. Also, denote $R_i = \mathrm{supp}(h_i)$ and $S = \mathrm{supp}(x)$. Then Z can be written explicitly as

$$Z = [(y - Ax)_{R_1}\mathrm{sgn}(x_1), \ldots, (y - Ax)_{R_m}\mathrm{sgn}(x_m)],$$

where many columns are zero since x is k-sparse. The following claims follow from the proof of Claim 42 in Arora et al. (2015). Here we state and detail some important steps.

Claim 14 *$\|Z\| \leq \widetilde{O}(k)$ holds with high probability.*

Proof With high probability

$$\|Z\| \leq \sqrt{\sum_{i \in S} \|(y - Ax)_{R_i}\mathrm{sgn}(x_i)\|^2} \leq \sqrt{k}\|(y - Ax)_{R_i}\|$$

where we use Claim 12 with $\|(y - Ax)_R\| \leq \widetilde{O}(\delta_s\sqrt{k})$ w.h.p., then $\|Z\| \leq \widetilde{O}(k)$ holds w.h.p. ■

Claim 15 *$\|\mathbb{E}[ZZ^T]\| \leq O(k^2/n)$ and $\|\mathbb{E}[Z^TZ]\| \leq \widetilde{O}(k^2/n)$ with high probability.*

Proof The first term is easily handled. Specifically, with high probability

$$\|\mathbb{E}[ZZ^T]\| \leq \|\mathbb{E}[\sum_{i\in S}(y-Ax)_{R_i}\mathrm{sgn}(x_i)^2(y-Ax)_{R_i}^T]\| = \|\mathbb{E}[\sum_{i\in S}(y-Ax)_{R_i}(y-Ax)_{R_i}^T]\| \leq O(k^2/n),$$

where the last inequality follows from the proof of Claim 42 in Arora et al. (2015), which is tedious to be repeated.

To bound $\|\mathbb{E}[Z^TZ]\|$, we use bound of the full matrix $(y-Ax)\mathrm{sgn}(x)^T$. Note that $\|y-Ax\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. is similar to what derived in Claim 12. Then with high probability,

$$\|\mathbb{E}[Z^TZ]\| \leq \|\mathbb{E}[\mathrm{sgn}(x)(y-Ax)^T(y-Ax)\mathrm{sgn}(x)^T]\| \leq \widetilde{O}(k)\|\mathbb{E}[\mathrm{sgn}(x)\mathrm{sgn}(x)^T]\| \leq \widetilde{O}(k^2/m).$$

where $\mathbb{E}[\mathrm{sgn}(x)\mathrm{sgn}(x)^T] = \mathrm{diag}(q_1, q_2, \ldots, q_m)$ has norm bounded by $O(k/m)$. We now can apply Bernstein's inequality for the truncated version of Z with $\mathcal{R} = \widetilde{O}(k)$ and $\sigma^2 = \widetilde{O}(k^2/m)$, then with $p = \widetilde{O}(m)$,

$$\|\mathcal{P}_H(g^s) - \mathcal{P}_H(\widehat{g}^s)\| \leq \frac{\widetilde{O}(k)}{p} + \sqrt{\frac{\widetilde{O}(k^2/m)}{p}} \leq O^*(k/m)$$

holds with high probability. Finally, we invoke the bound $\eta = O(m/k)$ and complete the proof. ■

Appendix E. A Special Case: Orthonormal A^*

We extend our results for the special case where the dictionary is orthonormal. As such, the dictionary is perfectly incoherent and bounded (i.e., $\mu = 0$ and $\|A^*\| = 1$).

Theorem 7 *Suppose that A^* is orthonormal. When $p_1 = \widetilde{\Omega}(n)$ and $p_2 = \widetilde{\Omega}(nr)$, then with high probability Algorithm 1 returns an initial estimate A^0 whose columns share the same support as A^* and with $(\delta, 2)$-nearness to A^* with $\delta = O^*(1/\log n)$. The sparsity of A^* can be achieved up to $r = O^*\big(\min(\frac{\sqrt{n}}{\log^2 n}, \frac{n}{k^2\log^2 n})\big)$.*

We use the same initialization procedure for this special case and achieve a better order of r. The proof of Theorem 7 follows the analysis for the general case with following two results:

Claim 16 (Special case of Claim 3) *Suppose that $u = A^*\alpha + \varepsilon_u$ is a random sample and $U = \mathrm{supp}(\alpha)$. Let $\beta = A^{*T}u$, then w.h.p., we have (a) $|\beta_i - \alpha_i| \leq \sigma_\varepsilon \log n$ for each i and (b) $\|\beta\| \leq O(\sqrt{k}\log n + \sigma_\varepsilon\sqrt{n}\log n)$.*

Proof We have $\beta = A^{*T}u = \alpha + A^{*T}\epsilon_u$, then $\beta_i - \alpha_i = \langle A^*_{\bullet i}, \epsilon_u\rangle$ and $\|\beta - \alpha\| = \|\epsilon_u\|$. Using probability bounds of $\langle A^*_{\bullet i}, \epsilon_u\rangle$, $\|\epsilon_u\|$ and $\|\alpha\|$ in Claim 2, we have the claim proved. □

We draw from the claim that for any $i \notin U \cap V$, $|\beta_i\beta_i'| \leq O(\sigma_\varepsilon \log^2 n)$ and have the following result:

Lemma 16 *Fix samples u and v and suppose that $y = A^*x^* + \varepsilon$ is a random sample independent of u, v. The expected value of the score for the l^{th} component of y is given by:*

$$e_l \triangleq \mathbb{E}[\langle y, u\rangle\langle y, v\rangle y_l^2] = \sum_{i\in U\cap V} q_i c_i \beta_i \beta_i' A_{li}^{*2} + \text{ perturbation terms}$$

where $q_i = \mathbb{P}[i \in S]$, $q_{ij} = \mathbb{P}[i, j \in S]$ and $c_i = \mathbb{E}[x_i^4 | i \in S]$. Moreover, the perturbation terms have absolute value at most $O^*\big(k/n \log^2 n \max(1/\sqrt{n}, k^2/n)\big)$.

Proof Lemma follows Lemma 1 via Claim 3 except that the second term of E_1 is bounded by $O(k \log^2 n/n^{3/2})$.

Appendix F. Extensions of Arora et al. (2015)

F.1 Sample complexity in noisy case

In this section, we study the sample complexity of the algorithms in Arora et al. (2015) in the presence of noise. While noise with order $\sigma_\varepsilon = O(1/\sqrt{n})$ does not change the sample complexity of the initialization algorithm, it affects that of the descent stage. The analysis involves producing a sharp bound for $\|\widehat{g}_{\bullet,i}^s - g_{\bullet i}^s\|$.

Lemma 17 *For a regular dictionary A^*, suppose A^s is $(\delta_s, 2)$-near to A^* with $\delta_s = O^*(1/\log n)$, then with high probability $\|\widehat{g}_{\bullet,i}^s - g_{\bullet i}^s\| \le O(k/m)\cdot(o(\delta) + O(\sqrt{k/n}))$ when $p = \widetilde{\Omega}(m + \sigma_\varepsilon^2 \frac{mn^2}{k})$.*

Proof This follows directly from Lemma 15 where $r = n$. □

We tighten the original analysis to obtain the complexity $\widetilde{\Omega}(m)$ instead of $\widetilde{\Omega}(mk)$ for the noiseless case. Putting together with $p = \widetilde{\Omega}(mk)$ required by the initialization, we then have the overall sample complexity $\widetilde{O}(mk + \sigma_\varepsilon^2 \frac{mn^2}{k})$ for the algorithms in Arora et al. (2015) in the noise regime.

F.2 Extension of Arora et al. (2015)'s initialization algorithm for sparse case

We study a simple and straightforward extension of the initialization algorithm of Arora et al. (2015) for the sparse case. This extension is produced by adding an extra projection, and is described in Figure 3. The recovery of the support of A^* is guaranteed by the following Lemma:

Lemma 18 *Suppose that $z^* \in \mathbb{R}^n$ is r-sparse whose nonzero entries are at least τ in magnitude. Provided z is δ-close to z^* and $z^0 = \mathcal{H}_r(z)$ with $\delta = O^*(1/\log n)$ and $r = O^*(\log^2 n)$, then z^0 and z^* has the same support.*

Proof Since z^0 is δ-close to z^*, then $\|z^0 - z^*\| \le \delta$ and $|z_i - z_i^*| \le \delta$ for every i. For $i \in \mathrm{supp}(z^*)$,

$$|z_i| \ge |z_i^*| - |z_i - z_i^*| \ge \tau - \delta$$

and for $i \notin \mathrm{supp}(z^*)$, $|z_i| \le \delta$. Since $\tau > O(1/\sqrt{r}) \gg \delta$, then the r-largest entries of z are in the support z^*, and hence z^0 and z^* has the same support. □

Algorithm 3 Pairwise Reweighting with Hard-Thresholding

Initialize $L = \emptyset$

Randomly divide p samples into two disjoint sets $\mathcal{P}_1$ and $\mathcal{P}_2$ of sizes p_1 and p_2 respectively

While $|L| < m$. Pick u and v from $\mathcal{P}_1$ at random

Construct the re-weighted covariance matrix $\widehat{M}_{u,v}$:

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

Compute the top singular values δ_1, δ_2 and top singular vector z of $\widehat{M}_{u,v}$

If $\delta_1 \geq \Omega(k/m)$ and $\delta_2 < O^*(k/m \log n)$

$z = \mathcal{H}_r(z)$, where $\mathcal{H}_r$ keeps r largest entries of z

If z is not within distance $1/\log n$ of any vector in L even with sign flip

$L = L \cup \{z\}$

Return $A^0 = (L_1, \ldots, L_m)$

Theorem 8 *Suppose that Assumptions* **B1-B4** *hold and Assumptions* **A1-A3** *satify with* $\mu = O^*\big(\frac{\sqrt{n}}{k\log^3 n}\big)$ *and* $r = O^*(\log^2 n)$. *When* $p_1 = \widetilde{\Omega}(m)$ *and* $p_2 = \widetilde{\Omega}(mk)$, *then with high probability Algorithm 3 returns an initial estimate* A^0 *whose columns share the same support as* A^* *and with* $(\delta, 2)$*-nearness to* A^* *with* $\delta = O^*(1/\log n)$.

This algorithm requires $r = O^*(\log^2 n)$, which is somewhat better than ours. However, the sample complexity and running time is inferior as compared with our novel algorithm.

Appendix G. Neural Implementation of Our Approach

We now briefly describe why our algorithm is "neurally plausible". Basically, similar to the argument in Arora et al. (2015), we describe at a very high level how our algorithm can be implemented via a neural network architecture. One should note that although both our initialization and descent stages are non-trivial modifications of those in Arora et al. (2015), both still inherit the nice neural plausiblity property.

G.1 Neural implementation of Stage 1: Initialization

Recall that the initialization stage includes two main steps: (i) estimate the support of each column of the synthesis matrix, and (ii) compute the top principal component(s) of a certain truncated weighted covariance matrix. Both steps involve simple vector and matrix-vector manipulations that can be implemented plausibly using basic neuronal manipulations.

For the support estimation step, we compute the product $\langle y, u\rangle\langle y, u\rangle y \circ y$, followed by a thresholding. The inner products, $\langle y, u\rangle$ and $\langle y, v\rangle$ can be computed using neurons via an online manner where the samples arrive in sequence; the thresholding can be implemented via a ReLU-type non-linearity.

For the second step, it is well known that the top principal components of a matrix can be computed in a neural (Hebbian) fashion using Oja's Rule Oja (1992).

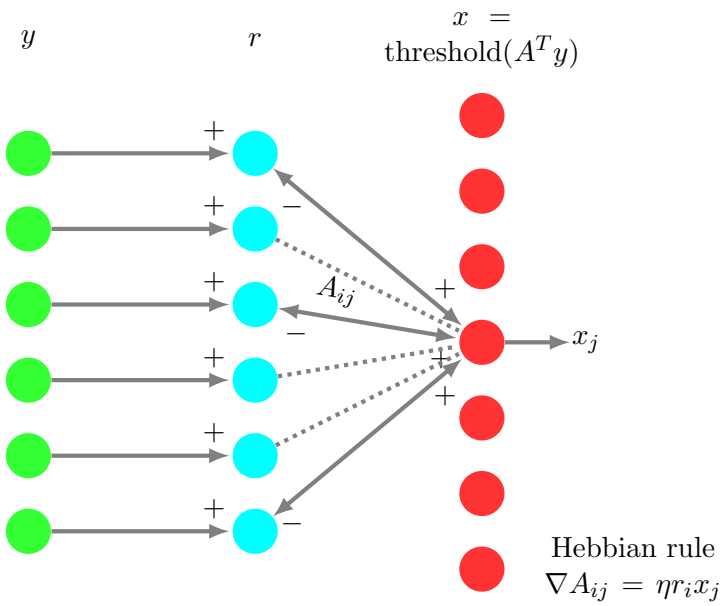

Figure 4: Neural network implementation of Algorithm 2. The network takes the image y as input and produces the sparse representation x as output. The hidden layer represents the residual between the image and its reconstruction Ax. The weights A_{ij}'s are stored on synapses, but most of them are zero and shown by the dotted lines.

G.2 Neural implementation of Stage 2: Descent

Our neural implementation of the descent stage (Algorithm 2), shown in Figure 4, mimics the architecture of Arora et al. (2015), which describes a simple two-layer network architecture for computing a single gradient update of A. The only difference in our case is that most of the value in A are set to zero, or in other words, our network is sparse. The network takes values y from the input layer and produce x as the output; there is an intermediate layer in between connecting the middle layer with the output via synapses. The synaptic weights are stored on A. The weights are updated by Hebbian learning. In our case, since A is sparse (with support given by R, as estimated in the first stage), we enforce the condition the corresponding synapses are inactive. In the output layer, as in the initialization stage, the neurons can use a ReLU-type non-linear activation function to enforce the sparsity of x.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014.

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

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014a.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014b.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pages 113–149, 2015.

Jarosław Błasiok and Jelani Nelson. An improved analysis of the er-spud dictionary learning algorithm. *arXiv preprint arXiv:1602.05719*, 2016.

Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE, 2010.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Niladri Chatterji and Peter Bartlett. Alternating minimization for dictionary learning with random initialization. 2017. arXiv:1711.03634v1.

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

Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 2443–2446. IEEE, 1999.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11): 6298–6319, 2015a.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015b.

Hamid Krim, Dewey Tucker, Stephane Mallat, and David Donoho. On denoising and best signal representation. *IEEE Transactions on Information Theory*, 45(7):2225–2238, 1999.

Rados law Adamczak. A note on the sample complexity of the er-spud algorithm by spielman, wang and wright for exact recovery of sparsely used dictionaries. *Journal of Machine Learning Research*, 17:1–18, 2016.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 689–696, 2009.

Arya Mazumdar and Ankit Singh Rawat. Associative memory using dictionary learning and expander decoding. In *Proc. Conf. American Assoc. Artificial Intelligence (AAAI)*, pages 267–273, 2017.

Erkki Oja. Principal components, minor components, and linear neural networks. *Neural networks*, 5(6):927–935, 1992.

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.

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

Ron Rubinstein, Michael Zibulevsky, and Michael Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Signal Processing*, 58 (3):1553–1564, 2010b.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1, 2012.

Jeremias Sulam, Boaz Ophir, Michael Zibulevsky, and Michael Elad. Trainlets: Dictionary learning in high dimensions. *IEEE Transactions on Signal Processing*, 64(12):3180–3193, 2016.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery using nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2351–2360, 2015.

Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1):3537–3580, 2016.