Alternating minimization for dictionary learning with random initialization

Niladri S. Chatterji
University of California, Berkeley
chatterji@berkeley.edu

Peter L. Bartlett
University of California, Berkeley
peter@berkeley.edu

Abstract

We present theoretical guarantees for an alternating minimization algorithm for the dictionary learning/sparse coding problem. The dictionary learning problem is to factorize vector samples $y^1, y^2, \ldots, y^n$ into an appropriate basis (dictionary) A^* and sparse vectors $x^{1*}, \ldots, x^{n*}$. Our algorithm is a simple alternating minimization procedure that switches between ℓ_1 minimization and gradient descent in alternate steps. Dictionary learning and specifically alternating minimization algorithms for dictionary learning are well studied both theoretically and empirically. However, in contrast to previous theoretical analyses for this problem, we replace a condition on the operator norm (that is, the largest magnitude singular value) of the true underlying dictionary A^* with a condition on the matrix infinity norm (that is, the largest magnitude term). Our guarantees are under a reasonable generative model that allows for dictionaries with growing operator norms, and can handle an arbitrary level of overcompleteness, while having sparsity that is information theoretically optimal. We also establish upper bounds on the sample complexity of our algorithm.

Erratum, August 7, 2019: An earlier version of this paper appeared in NIPS 2017 which had an erroneous claim about convergence guarantees with random initialization. The main result – Theorem 3 – has been corrected by adding an assumption about the initialization (Assumption B1).

1 Introduction

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

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

for $i \in \{1, 2, \ldots, n\}$, the goal is to recover a fixed dictionary $A^* \in \mathbb{R}^{d \times r}$ and s-sparse vectors $x^{i*} \in \mathbb{R}^r$. (An s-sparse vector has no more than s non-zero entries.) In many problems of interest, the dictionary is often overcomplete, that is, $r \geq d$. This is believed to add flexibility in modeling and robustness. This model was first proposed in neuroscience as an energy minimization heuristic that reproduces features of the V1 region of the visual cortex (Olshausen and Field, 1997; Lewicki and Sejnowski, 2000). It has also been an extremely successful approach to identifying low dimensional structure in high dimensional data; it is used extensively to find features in images, speech and video (see, for example, references in Elad and Aharon, 2006).

Most formulations of dictionary learning tend to yield non-convex optimization problems. For example, note that if either x^{i*} or A^* were known, given y^i, this would just be a (matrix/sparse) regression problem. However, estimating both x^{i*} and A^* simultaneously leads to both computational as well as statistical complications. The heuristic of alternating minimization works well empirically for dictionary learning. At each step, first an estimate of the dictionary is held fixed while the sparse

coefficients are estimated; next, using these sparse coefficients the dictionary is updated. Note that in *each step* the sub-problem has a convex formulation, and there is a range of efficient algorithms that can be used. This heuristic has been very successful empirically, and there has also been significant recent theoretical progress in understanding its performance, which we discuss next.

1.1 Related Work

A recent line of work theoretically analyzes local linear convergence rates for alternating minimization procedures applied to dictionary learning (Agarwal et al., 2014; Arora et al., 2015). Arora et al. (2015) present a neurally plausible algorithm that recovers the dictionary exactly for sparsity up to $s = \mathcal{O}(\sqrt{d}/(\mu \log(d)))$, where $\mu/\sqrt{d}$ is the level of incoherence in the dictionary (which is a measure of the similarity of the columns; see Assumption A1 below). Agarwal et al. (2014) analyze a least squares/ℓ_1 minimization scheme and show that it can tolerate sparsity up to $s = \mathcal{O}(d^{1/6})$. Both of these establish local linear convergence guarantees for the maximum column-wise distance. Exact recovery guarantees require a singular-value decomposition (SVD) or clustering based procedure to initialize their dictionary estimates (see also the previous work Arora et al., 2013; Agarwal et al., 2013).

For the undercomplete case (when $r \leq d$), Sun et al. (2017) provide a Riemannian trust region method that can tolerate sparsity $s = \mathcal{O}(d)$, while earlier work by Spielman et al. (2012) provides an algorithm that works in this setting for sparsity $\mathcal{O}(\sqrt{d})$.

Local and global optima of non-convex formulations for the problem have also been extensively studied in (Wu and Yu, 2015; Gribonval et al., 2015; Gribonval and Nielsen, 2003), among others. Apart from alternating minimization, other approaches (without theoretical convergence guarantees) for dictionary learning include K-SVD (Aharon et al., 2006) and MOD (Engan et al., 1999). There is also a nice formulation by Barak et al. (2015), based on the sum-of-squares hierarchy. Recently, Hazan and Ma (2016) provide guarantees for improper dictionary learning, where instead of learning a dictionary, they learn a comparable encoding via convex relaxations. Our work also adds to the recent literature on analyzing alternating minimization algorithms (Jain et al., 2013; Netrapalli et al., 2013, 2014; Hardt, 2014; Balakrishnan et al., 2017).

1.2 Contributions

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

Our results hold for a rather arbitrary level of overcompleteness, $r = \mathcal{O}(poly(d))$. We establish convergence results for sparsity level $s = \mathcal{O}(\sqrt{d})$, which is information theoretically optimal for incoherent dictionaries and improves the previously best known results in the overcomplete setting by a logarithmic factor. Our algorithm is simple, involving an ℓ_1-minimization step followed by a gradient update for the dictionary.

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

1.3 Organization

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

Notation

For a vector $v \in \mathbb{R}^d$, v_i denotes the ith component of the vector, $\|v\|_p$ denotes the ℓ_p norm, $supp(v)$ denotes the support of a vector v, that is, the set of non-zero entries of the vector, $sgn(v)$ denotes

---

Algorithm 1: Alternating Minimization for Dictionary Learning

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

 1 **for** $t = 1, 2, \ldots, T$ **do**

 2   **for** $k = 1, 2, \ldots, n$ **do**

 3 $w^{k,(t)} = MUS_{\gamma^{(t)},\lambda^{(t)},\nu^{(t)}}(y^k, A^{(t-1)}, R^{(t-1)})$

 4   **for** $l = 1, 2, 3 \ldots, r$ **do**

 5 $x_l^{k,(t)} = w_l^{k,(t)}\mathbb{I}\left(|w_l^{k,(t)}| > \tau^{(t)}\right),$ $(x^{k,(t)}$ is the sparse estimate$)$

 6   **end**

 7   **end**

 8   **for** $i = 1, 2, \ldots, d$ **do**

 9   **for** $j = 1, 2, \ldots, r$ **do**

10 $A_{ij}^{(t)} = A_{ij}^{(t-1)} - \frac{\eta}{n}\sum_{k=1}^n\left[\sum_{p=1}^r\left(A_{ip}^{(t-1)}x_p^{k,(t)}x_j^{k,(t)} - y_i^k x_j^{k,(t)}\right)\right]$

11   **end**

12   **end**

13 $R^{(t)} = \frac{7}{8}R^{(t-1)}.$

14 **end**

---

the sign of the vector v, that is, a vector with $sgn(v)_i = 1$ if $v_i > 0$, $sgn(v)_i = -1$ if $v_i < 0$ and $sgn(v)_i = 0$ if $v_i = 0$. For a matrix W, W_i denotes the ith column, W_{ij} is the element in the ith row and j^{th} column, $\|W\|_{op}$ denotes the operator norm, and $\|W\|_\infty$ denotes the maximum of the magnitudes of the elements of W. For a set J, we denote its cardinality by $|J|$. Throughout the paper, we use C multiple times to denote global constants that are independent of the problem parameters and dimension. We denote the indicator function by $\mathbb{I}(\cdot)$.

2   Algorithm

Given an initial estimate of the dictionary $A^{(0)}$ we alternate between an ℓ_1 minimization procedure (specifically the $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector—$MUS_{\gamma,\lambda,\nu}$ in the algorithm—followed by a thresholding step) and a gradient step, under sample ℓ_2 loss, to update the dictionary. We analyze this algorithm and demand linear convergence at a rate of 7/8; convergence analysis for other rates follows in the same vein with altered constants. Below we state the permitted range for the various parameters in the algorithm above.

1. Step size: We need to set the step size in the range $3r/4s < \eta < r/s$.

2. Threshold: At each step set the threshold at $\tau^{(t)} = 16R^{(t-1)}M(R^{(t-1)}(s+1) + s/\sqrt{d})$.

3. Tuning parameters: We need to pick $\lambda^{(t)}$ and $\nu^{(t)}$ such that the assumption (D5) is satisfied. A choice that is suitable that satisfies this assumption is

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

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

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

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

2.1   Sparse Regression Estimator

Our proof of convergence for Algorithm 1 also goes through with a different choices of robust sparse regression estimators, however, we can establish the tightest guarantees when the $\{\ell_1, \ell_2, \ell_\infty\}$-MU

Selector is used in the sparse regression step. The $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector (Belloni et al., 2014) was established as a modification of the Dantzig selector to handle uncertainty in the dictionary. There is a beautiful line of work that precedes this that includes (Rosenbaum et al., 2010, 2013; Belloni et al., 2016). There are also modified non-convex LASSO programs that have been studied (Loh and Wainwright, 2011) and Orthogonal Matching Pursuit algorithms under in-variable errors (Chen and Caramanis, 2013). However these estimators require the error in the dictionary to be stochastic and zero mean which makes them less suitable in this setting. Also note that standard ℓ_1 minimization estimators like the LASSO and Dantzig selector are highly unstable under errors in the dictionary and would lead to much worse guarantees in terms of radius of convergence (as studied in Agarwal et al., 2014). We establish the error guarantees for a robust sparse estimator $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector under fixed corruption in the dictionary. We also establish that this estimator is minimax optimal when the error in the sparse estimate is measured in infinity norm $\|\hat{\theta} - \theta^*\|_\infty$ and the dictionary is corrupted.

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

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

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

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

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

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

To identify the signs of the sparse covariates correctly using this class of thresholded estimators, we would like in the first step to use an estimator $\hat{\theta}$ that is *optimal*, as this would lead to tighter control over the radius of convergence. This makes the choice of $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector natural, as we will show it is minimax optimal under certain settings.

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

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

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

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

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

4

2.2 Gradient Update for the dictionary

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

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

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

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

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

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

3 Main Results and Assumptions

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

3.1 Assumptions

Assumptions on A^*

(A1) **Incoherence:** *We assume the the true underlying dictionary is $\mu/\sqrt{d}$-incoherent*

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

This is a standard assumption in the sparse regression literature when support recovery is of interest. It was introduced in (Fuchs, 2004; Tropp, 2006) in signal processing and independently in (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006) in statistics. We can also establish guarantees under the strictly weaker ℓ_∞-sensitivity condition (cf. Gautier and Tsybakov, 2011) used in analyzing sparse estimators under in-variable uncertainty in (Belloni et al., 2016; Rosenbaum et al., 2013). The $\{\ell_1, \ell_2, \ell_\infty\}$-MU selector that we use for our sparse recovery step also works with this more general assumption, however for ease of exposition we assume A^* to be $\mu/\sqrt{d}$-incoherent.

(A2) **Normalized Columns**: *We assume that all the columns of A^* are normalized to* 1,

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

Note that the samples $\{y^i\}_{i=1}^n$ are invariant when we scale the columns of A^* or under permutations of its columns. Thus we restrict ourselves to dictionaries with normalized columns and label the entire equivalence class of dictionaries with permuted columns and varying signs as A^*.

(A3) **Bounded max-norm:** *We assume that A^* is bounded in matrix infinity norm*

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

where $C_b = 1/2000M^2$. This is in contrast with previous work that imposes conditions on the operator norm of A^* (Arora et al., 2015; Agarwal et al., 2014; Arora et al., 2013). Our assumptions help provide guarantees under alternate assumptions and it also allows the operator norm to grow with dimension, whereas earlier work requires A^* to be such that $\|A^*\|_{op} \leq C\left(\sqrt{r/d}\right)$. In general the infinity norm and operator norm balls are hard to

compare. However, one situation where a comparison is possible is if we assume the entries of the dictionary to be drawn iid from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then by standard concentration theorems, for the operator norm condition to be satisfied we would need the variance (σ^2) of the distribution to scale as $\mathcal{O}(1/d)$ while, for the infinity norm condition to be satisfied we need the variance to be $\tilde{\mathcal{O}}(1/s^2)$. This means that modulo constants the variance can be much larger for the infinity norm condition to be satisfied than for the operator norm condition.

(A4) **Separation:** *We assume that* $\forall\, i \in \{1, \ldots, r\}$

$$\|A_i^*\|_\infty > \frac{3C_b}{4s}, \text{ and, } \min_{z \in \{-1,1\}} \|A_i^* - zA_j^*\|_\infty \geq \frac{3C_b}{2s} \qquad \forall\, j \neq i \in \{1, \ldots, r\}.$$

This condition ensures that two dictionaries in the equivalence class with varying signs of columns or permutations are separated in infinity norm. The first condition ensures that for any column A_i^* and $-A_i^*$ are separated $\|A_i^* - (-A_i^*)\|_\infty \geq 3C_b/2s$.

Assumption on the initial estimate and initialization

(B1) *We require an initial estimate for the dictionary $A^{(0)}$ that is close in infinity norm,*

$$\|A^{(0)} - A^*\|_\infty \leq \frac{C_b}{2s}.$$

This initialization combined with the separation condtion above ensures that the initial estimate $A^{(0)}$ is close to *only one* dictionary in the equivalence class. The algorithm is going to be contractive, hence this will hold true throughout the run of the algorithm.

Assumptions on x^*

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

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

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

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

where C is an appropriate global constant such that A^* satisfies assumption (D3), see Remark 15. For incoherent dictionaries, this upper bound is tight up to constant factors for sparse recovery to be feasible (Donoho and Huo, 2001; Gribonval and Nielsen, 2003).

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

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

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

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

$$\mathbb{P}r(x_i^* \neq 0) = \frac{s}{r} \qquad \forall\, i \in \{1, \ldots, r\},$$

$$\mathbb{P}r(x_i^*, x_j^* \neq 0) = \frac{s(s-1)}{r(r-1)} \qquad \forall\, i \neq j \in \{1, \ldots, r\}.$$

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

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

We note that these assumptions (A1), (A2) and (C1) - (C5) are similar to those made in (Arora et al., 2015; Agarwal et al., 2014). Agarwal et al. (2014) require the matrices to satisfy the restricted isometry property, which is strictly weaker than $\mu/\sqrt{d}$-incoherence, however they can tolerate a much lower level of sparsity ($d^{1/6}$).

3.2 Main Result

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

$$\|A^{(T)} - A^*\|_\infty \leq \left(\frac{7}{8}\right)^T R^{(0)}, \qquad \text{with probability } 1 - T\delta.$$

4 Proof of Convergence

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

By Lemma 8, Algorithm 1 is guaranteed to estimate the sign pattern correctly at every round of the algorithm, $sgn(x) = sgn(x^*)$ (see proof in Appendix A.1). Also note that by assumption (B1), the initial dictionary $A^{(0)}$ is close to *only one* dictionary A^* in the equivalence class.

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

$$\hat{g}_{ij} = \frac{1}{n}\sum_{m=1}^n\left[\sum_{k=1}^r\left(a_{ik}x_k^m x_j^m\right) - y_i^m x_j^m\right] = \frac{1}{n}\sum_{m=1}^n\left[\sum_{k=1}^r\left(a_{ik}x_k^m - a_{ik}^* x_k^{m*}\right)x_j^m\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^r\left(a_{ik}x_k - a_{ik}^* x_k^*\right)x_j\right]$$

$$\quad + \left[\frac{1}{n}\sum_{m=1}^n\left[\sum_{k=1}^r\left(a_{ik}x_k^m - a_{ik}^* x_k^{m*}\right)x_j^m\right] - \mathbb{E}\left[\sum_{k=1}^r\left(a_{ik}x_k - a_{ik}^* x_k^*\right)x_j\right]\right]$$

$$= g_{ij} + \underbrace{\hat{g}_{ij} - g_{ij}}_{\triangleq \epsilon_n},$$

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

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

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

$$g_{ij} = \mathbb{E}\left(a_{ij}x_j^2 - a_{ij}^* x_j^* x_j\right) + \mathbb{E}\left[\sum_{k\neq j}a_{ik}x_k x_j - a_{ik}^* x_k^* x_j\right]$$

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

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

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

$$a_{ij}^{'} = a_{ij} - \eta \hat{g}_{ij} = a_{ij} - \eta(g_{ij} + \epsilon_n) = a_{ij} - \eta \left[g_{ij}^c + (\Xi_1 + \Xi_2) + \epsilon_n \right].$$

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

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

Taking absolute values, we get

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

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

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

Coupled with Lemma 6, Inequality (i) follows from $\eta \le \frac{r}{s}$ while (ii) follows from $\eta \ge \frac{3r}{4s}$. We also have by Theorem 10 that $\eta|\epsilon_n| \le R/8$ with probability $1 - \delta$. So if we unroll the bound for t steps we have,

$$|a_{ij}^{(t)} - a_{ij}^*| \le \frac{3}{4}R^{(t-1)} + \eta|\epsilon_n| \le \frac{3}{4}R^{(t-1)} + \frac{1}{8}R^{(t-1)} = \frac{7}{8}R^{(t-1)} \le \left(\frac{7}{8}\right)^t R^{(0)}.$$

We also have $\eta|\epsilon_n| \le R/8 \le R^{(0)}/8$ with probability at least $1 - \delta$ in each iteration, for all $t \in \{1, \dots, T\}$; thus by taking a union bound over the iterations we are guaranteed to remain in our initial ball of radius $R^{(0)}$ with high probability, completing the proof. ∎

5 Conclusion

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

References

Agarwal, A., A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon (2014). Learning sparsely used overcomplete dictionaries. In *COLT*, pp. 123–137.

Agarwal, A., A. Anandkumar, and P. Netrapalli (2013). A clustering approach to learn sparsely-used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*.

Aharon, M., M. Elad, and A. Bruckstein (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing 54*(11), 4311–4322.

Arora, S., R. Ge, T. Ma, and A. Moitra (2015). Simple, efficient, and neural algorithms for sparse coding. In *COLT*, pp. 113–149.

Arora, S., R. Ge, and A. Moitra (2013). New algorithms for learning incoherent and overcomplete dictionaries.

Balakrishnan, S., M. J. Wainwright, B. Yu, et al. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics 45*(1), 77–120.

Barak, B., J. A. Kelner, and D. Steurer (2015). Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 143–151. ACM.

Belloni, A., M. Rosenbaum, and A. B. Tsybakov (2014). An $\{\ell_1, \ell_2, \ell_\infty\}$-Regularization Approach to High-Dimensional Errors-in-variables Models. *arXiv preprint arXiv:1412.7216*.

Belloni, A., M. Rosenbaum, and A. B. Tsybakov (2016). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Chen, Y. and C. Caramanis (2013). Noisy and Missing Data Regression: Distribution-Oblivious Support Recovery.

Donoho, D. L. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory 47*(7), 2845–2862.

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

Engan, K., S. O. Aase, and J. H. Husoy (1999). Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, Volume 5, pp. 2443–2446. IEEE.

Fuchs, J.-J. (2004). Recovery of exact sparse representations in the presence of noise. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, Volume 2, pp. ii–533. IEEE.

Gautier, E. and A. Tsybakov (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

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

Gribonval, R. and M. Nielsen (2003). Sparse representations in unions of bases. *IEEE transactions on Information theory 49*(12), 3320–3325.

Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660. IEEE.

Hazan, E. and T. Ma (2016). A Non-generative Framework and Convex Relaxations for Unsupervised Learning. In *Advances in Neural Information Processing Systems*, pp. 3306–3314.

Jain, P., P. Netrapalli, and S. Sanghavi (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM.

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

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

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics 141*(1), 148–188.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The annals of statistics*, 1436–1462.

Netrapalli, P., P. Jain, and S. Sanghavi (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pp. 2796–2804.

Netrapalli, P., U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain (2014). Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pp. 1107–1115.

Olshausen, B. A. and D. J. Field (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research 37*(23), 3311–3325.

Rigollet, P. and A. Tsybakov (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 731–771.

Rosenbaum, M., A. B. Tsybakov, et al. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics 38*(5), 2620–2651.

Rosenbaum, M., A. B. Tsybakov, et al. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 276–290. Institute of Mathematical Statistics.

Spielman, D. A., H. Wang, and J. Wright (2012). Exact Recovery of Sparsely-Used Dictionaries. In *COLT*, pp. 37–1.

Sun, J., Q. Qu, and J. Wright (2017). Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory 63*(2), 853–884.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory 52*(3), 1030–1051.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats.

Wu, S. and B. Yu (2015). Local identifiability of ℓ_1-minimization dictionary learning: a sufficient and almost necessary condition. *arXiv preprint arXiv:1505.04363*.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine learning research 7*(Nov), 2541–2563.

A Additional details for the proof of convergence

For Appendix A.1 and A.2, we borrow the notation from Section 4. In Appendix A.1 we prove Lemma 4 that controls an error term which will be useful in establishing Lemma 5 that bounds the error terms in the gradient, Ξ_1 and Ξ_2. Corollary 9 establishes the error bound for the sparse estimate while Lemma 8 establishes that the sparse estimate after the thresholding step has the correct sign. In Appendix A.2, we establish finite sample guarantees.

A.1 Proof of Auxillary Lemmas

Before we prove Lemma 5, which controls the terms in the gradient, we prove Lemma 4, which will be vital in controlling the cross-term in the gradient.

Lemma 4. *Let the assumptions stated in Section 3.1 hold. Then at each iteration step we have the guarantee that*

$$\left| \max_{k:k\neq j} \left\{ \mathbb{E}\left[a_{ik}x_kx_j - a_{ik}^*x_k^*x_j | x_k^* \neq 0, x_j^* \neq 0 \right] \right\} \right| \leq \frac{R}{6(s-1)}.$$

Proof Let us define

$$\Gamma \triangleq \max_{k:k\neq j} \left\{ \mathbb{E}\left[a_{ik}x_kx_j - a_{ik}^*x_k^*x_j | x_k^* \neq 0, x_j^* \neq 0 \right] \right\},$$

and let us define the event $\mathcal{E}_{jk} \triangleq \{x_j^* \neq 0, x_k^* \neq 0\}$. Expanding Γ,

$$\Gamma = \max_{k:k\neq j} \left\{ \mathbb{E}\left[a_{ik}(x_k - x_k^* + x_k^*)(x_j - x_j^* + x_j^*) - a_{ik}^*x_k^*(x_j - x_j^* + x_j^*) | \mathcal{E}_{jk} \right] \right\}$$

$$= \max_{k:k\neq j} \left\{ \underbrace{(a_{ik} - a_{ik}^*)\mathbb{E}\left[x_k^*(x_j - x_j^*) | \mathcal{E}_{jk} \right]}_{\triangleq n_1} + \underbrace{a_{ik}\mathbb{E}\left[(x_k - x_k^*)x_j^* | \mathcal{E}_{jk} \right]}_{\triangleq n_2} \right.$$

$$\left. + \underbrace{a_{ik}\mathbb{E}\left[(x_k - x_k^*)(x_j - x_j^*) | \mathcal{E}_{jk} \right]}_{\triangleq n_3} + \underbrace{(a_{ik} - a_{ik}^*)\mathbb{E}\left[x_k^*x_j^* | \mathcal{E}_{jk} \right]}_{\triangleq n_4} \right\}.$$

Given that the non-zero entries of x^* are independent and mean zero we have $n_4 = 0$. Next we see n_1, n_2 and n_3 are bounded above as

$$n_1 \leq |a_{ik} - a_{ik}^*|M\|x - x^*\|_\infty \leq RM\|x - x^*\|_\infty$$
$$n_2 \leq |a_{ik}|M\|x - x^*\|_\infty \leq (|a_{ik}^*| + R)M\|x - x^*\|_\infty$$
$$n_3 \leq |a_{ik}|\|x - x^*\|_\infty^2 \leq (R + |a_{ik}^*|)\|x - x^*\|_\infty^2,$$

these follow as $|x_j^*| \leq M$, $|x_j - x_j^*| \leq \|x - x^*\|_\infty$ and $|a_{ik} - a_{ik}^*| \leq R$. The goal now is to show that $n_1 \leq R/30(s-1)$, $n_2 \leq R/15(s-1)$ and $n_3 \leq R/15(s-1)$. Let us unwrap the first term of n_1

$$n_1 \leq RM\|x - x^*\|_\infty \overset{(i)}{\leq} \frac{R}{30(s-1)}\left[30(s-1)M \cdot 16RM\left(R(s+1) + \frac{s}{\sqrt{d}} \right) \right]$$

$$\overset{(ii)}{\leq} \frac{R}{30(s-1)}\left[30(s-1)M \cdot \frac{8C_bM}{s}\left(\frac{C_b(s+1)}{2s} + 2 \right) \right]$$

$$= \frac{R}{30(s-1)}\left[240M^2 \underbrace{\left(\frac{s-1}{s} \right)}_{\leq 1} C_b \left(C_b \underbrace{\left(\frac{(s+1)}{2s} \right)}_{\leq 3/4} + 2 \right) \right]$$

$$\leq \frac{R}{30(s-1)}\underbrace{\left[240M^2C_b\left(\frac{3}{4}C_b + 2 \right) \right]}_{\triangleq \xi_1}, \tag{5}$$

where (i) follows by invoking Corollary 9 and (ii) follows as $s \leq 2\sqrt{d}$ and $R \leq C_b/2s$. Our choice $C_b = 1/2000M^2$ ensures that $\xi_1 \leq 1$. The second term in the upper bound on n_2 can be bounded by the same technique as we used to bound n_1, giving $RM\|x - x^*\|_\infty \leq R/30(s-1)$. For the first term in n_2, we have

$$
\begin{aligned}
|a_{ik}^*|M\|x - x^*\|_\infty &\leq \frac{R}{30(s-1)} \left[480\frac{(s-1)}{s} M^2 C_b \left(R(s+1) + \frac{s}{\sqrt{d}} \right) \right] \\
&\leq \frac{R}{30(s-1)} \underbrace{\left[480 M^2 C_b \left(C_b \frac{(s+1)}{2s} + 2 \right) \right]}_{\triangleq \xi_2},
\end{aligned}
\tag{6}
$$

where these inequalities follow by invoking Corollary 9 and by the upper bounds on $|a_{ik}^*|$ and R. Again our choice $C_b = 1/2000M^2$ ensures that $\xi_2 \leq 1$ which leaves us with the upper bound on $n_2 \leq \frac{R}{15(s-1)}$. Finally to bound n_3 we observe that the first term is bounded as follows,

$$
\begin{aligned}
R\|x - x^*\|_\infty^2 &\leq \frac{R}{30(s-1)} \left[30(s-1) \cdot 16^2 R^2 M^2 \left(R(s+1) + \frac{s}{\sqrt{d}} \right)^2 \right] \\
&\leq \frac{R}{30(s-1)} \left[\sqrt{30(s-1)} \cdot 16 \frac{C_b}{2s} M \left(C_b \frac{(s+1)}{2s} + 2 \right) \right]^2 \\
&\leq \frac{R}{30(s-1)},
\end{aligned}
$$

where the last inequality is due to the fact that $\xi_1 \leq 1$. We have $|a_{ik}^*| \leq C_b/s$ and similar arguments as above can be used to show that the second term in n_3 is also bounded above by $\frac{R}{30(s-1)}$. Having controlled n_1, n_2 and n_3 at the appropriate levels completes the proof and yields the desired bound on Γ. ∎

Lemma 5. *Let the assumptions stated in Section 3.1 hold. Then at each iteration step we can bound the error terms in the gradient as*

$$
|\Xi_1| = \left| a_{ij}^* \frac{s}{r} \mathbb{E}\left[(x_j - x_j^*)x_j | x_j^* \neq 0 \right] \right| \leq \frac{s}{6r} R
$$

$$
|\Xi_2| = \left| \mathbb{E}\left[\sum_{k \neq j} a_{ik} x_k x_j - a_{ik}^* x_k^* x_j \right] \right| \leq \frac{s}{6r} R.
$$

Proof *Part 1*-We first prove the bound on Ξ_1. We start by unpacking Ξ_1

$$
\begin{aligned}
|\Xi_1| &= \left| \frac{s}{r} a_{ij}^* \mathbb{E}\left[(x_j - x_j^*)x_j | x_j^* \neq 0 \right] \right| \\
&\leq \frac{s}{r}|a_{ij}^*| \cdot \left| \mathbb{E}\left[(x_j - x_j^*)(x_j^* + x_j - x_j^*) | x_j^* \neq 0 \right] \right| \\
&\overset{(i)}{\leq} \frac{s}{r}|a_{ij}^*|M \cdot \mathbb{E}\left[\|x - x^*\|_\infty | x_j^* \neq 0 \right] + \frac{s}{r}|a_{ij}^*|\mathbb{E}\left[\|x - x^*\|_\infty^2 | x_j^* \neq 0 \right] \\
&\overset{(ii)}{\leq} \frac{s}{r}|a_{ij}^*|M \cdot \left(16RM\left((s+1)R + \frac{s}{\sqrt{d}} \right) \right) + \frac{s}{r}|a_{ij}^*|\left(16RM\left((s+1)R + \frac{s}{\sqrt{d}} \right) \right)^2 \\
&= \frac{s}{6r}R \left\{ 96|a_{ij}^*|M^2\left(R(s+1) + \frac{s}{\sqrt{d}} \right) + 6|a_{ij}^*|R\left(16M\left(R(s+1) + \frac{s}{\sqrt{d}} \right) \right)^2 \right\} \\
&\leq \frac{s}{6r}R,
\end{aligned}
\tag{7}
$$

where (i) follows by triangle inequality and $|x_j^*| \leq M$ and, (ii) follows by Corollary 9. It can be shown that in (7) the term in the curly braces is ≤ 1 by arguments similar to those used in Lemma 4 (because $R \leq 1/4000M^2s$, $s \leq 2\sqrt{d}$ and $|a_{ij}^*| \leq 1/2000M^2s$), thus establishing the desired bound on $|\Xi_1|$.

Part 2- Expanding Ξ_2 we find

$$
\begin{aligned}
|\Xi_2| &= \left| \mathbb{E}\left[\sum_{k \neq j} a_{ik}x_k x_j - a_{ik}^* x_k^* x_j^* \right] \right| \\
&\stackrel{(i)}{=} \frac{s(s-1)}{r(r-1)} \left| \mathbb{E}\left[\sum_{k \neq j} a_{ik}x_k x_j - a_{ik}^* x_k^* x_j | x_k^* \neq 0, x_j^* \neq 0 \right] \right| \\
&\leq \frac{s(s-1)}{r(r-1)} \cdot (r-1) \left| \max_{k \neq j} \left\{ \mathbb{E}\left[a_{ik}x_k x_j - a_{ik}^* x_k^* x_j | x_k^* \neq 0, x_j^* \neq 0 \right] \right\} \right| \\
&= \frac{s}{6r} R \left(\frac{6(s-1)}{R} \left| \max_{k \neq j} \left\{ \mathbb{E}\left[a_{ik}x_k x_j - a_{ik}^* x_k^* x_j | x_k^* \neq 0, x_j^* \neq 0 \right] \right\} \right| \right) \\
&\stackrel{(ii)}{\leq} \frac{s}{6r} R,
\end{aligned}
$$

where (i) follows from assumption (C4) and (ii) follows by invoking Lemma 4. ∎

Lemma 6. *Let the assumptions stated in Section 3.1 hold. Then at each iteration step we can bound the variance of the estimate,*

$$
\frac{2}{3} \leq \mathbb{E}\left[x_j^2 | x_j^* \neq 0 \right] \leq \frac{4}{3}.
$$

Proof Consider the expectation of the random variable $x_j^2 - x_j^{*2} | x_j^* \neq 0$. We have

$$
\begin{aligned}
x_j^2 - x_j^{*2} &\leq |x_j + x_j^*| \|x - x^*\|_\infty \\
&= |2x_j^* + x_j - x_j^*| \|x - x^*\|_\infty \leq 2|x_j^*| \|x - x^*\|_\infty + \|x - x^*\|_\infty^2 \\
&\leq \underbrace{2M\|x - x^*\|_\infty + \|x - x^*\|_\infty^2}_{\triangleq \xi_3}.
\end{aligned}
$$

Note that $\xi_3 \leq \frac{1}{3}$, if $\|x - x^*\|_\infty \leq \frac{1}{3}\left(\sqrt{3}\sqrt{3M^2+1} - 3M \right)$. We also have an upper bound on $\|x - x^*\|_\infty$ by Corollary 9

$$
\begin{aligned}
\|x - x^*\|_\infty &\leq 16RM\left(R(s+1) + \frac{s}{\sqrt{d}} \right) \leq \underbrace{\frac{8}{s}}_{\leq 4} C_b M \left(C_b \underbrace{\frac{s+1}{2s}}_{\leq 3/4} + \underbrace{\frac{s}{\sqrt{d}}}_{\leq 2} \right) \\
&\leq 4C_b M \left(\frac{3}{4} C_b + 2 \right).
\end{aligned}
$$

Our choice $C_b = 1/2000M^2$ with $M > 1$ guarantees that

$$
4C_b M \left(\frac{3}{4} C_b + 2 \right) \leq \frac{1}{3}\left(\sqrt{3}\sqrt{3M^2+1} - 3M \right), \tag{8}
$$

this yields the claimed bound. ∎

The next corollary establishes an infinity norm bound on the error in the sparse estimate under the assumptions made in Section 3.1 and choice of parameters in Section 2.

Corollary 7. *Under the assumptions specified in Section 3.1 and choice of parameters for Algorithm 1 in Section 2 we have the bound for all $t \in \{1, \dots, T\}$ and $k \in \{1, \dots, n\}$,*

$$
\|w^{k,(t)} - x^{k*}\|_\infty \leq 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}} \right),
$$

where $w^{k,(t)}$ is as defined in Algorithm 1.

Proof We have $\|x^{k*}\|_2 \leq \sqrt{s}M$, $\|x^{k*}\|_\infty \leq M$ thus plugging this into Theorem 16 gives us the desired result. ∎

The next theorem guarantees that at each round of the algorithm, under the assumptions stated in Section 3.1, we correctly predict the sign pattern.

Lemma 8. *Under the assumptions (A1)-(A6),(B1) and (C1)-(C5) stated in Section 3.1 with*

$$32R^{(0)}M\left(R^{(0)}(s+1) + \frac{s}{\sqrt{d}}\right) < m, \tag{9}$$

and under the choice of the parameters $\eta, R^{(t)}, \tau^{(t)}, \gamma^{(t)}, \lambda^{(t)}$ and $\nu^{(t)}$ specified in Section 2 for all $t \in \{1, 2, \ldots, T\}$ we have the guarantee that Algorithm 1 returns a sparse estimate $\{x^{k,(t)}\}_{k=1}^{n}$ such that,

$$sgn(x^{k,(t)}) = sgn(x^{k*}), \qquad\qquad \forall k \in \{1, 2, \ldots, n\}.$$

Proof Under the assumptions stated we can invoke Corollary 7 to get,

$$\|w^{k,(t)} - x^{k*}\|_\infty \leq 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right) \quad \forall k \in \{1, \ldots, n\}, t \in \{1, \ldots, T\}, \tag{10}$$

where $w^{k,(t)}$ is defined as in Algorithm 1. Note that the thresholds are defined by the schedule,

$$\tau^{(t)} = 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right).$$

By definition $x^{k,(t)}$ is the coordinate-wise thresholded estimate,

$$x_l^{k,(t)} = w_l^{k,(t)}\mathbb{I}\left(|w_l^{k,(t)}| > \tau^{(t)}\right) \qquad\qquad \forall l \in \{1, 2, \ldots, r\}.$$

We know that for all $t > 1$ we have $R^{(t)} < R^{(0)}$. So by the infinity norm bound in the above display (10) and, by the assumptions on the distribution of x^*, we have that

$$sgn\left(x^{k,(t)}\right) = sgn\left(x^{k*}\right) \qquad\qquad \forall k \in \{1, 2, \ldots, n\}.$$

This follows as the thresholding step only zeros out the non-zero elements in $x^{k,(t)}$ that are not in $supp(x^{k*})$. ∎

Corollary 9. *Under the assumptions specified in Section 3.1 and choice of parameters for Algorithm 1 in Section 2 we have the bound for all $t \in \{1, \ldots, T\}$ and $k \in \{1, \ldots, n\}$,*

$$\|x^{k,(t)} - x^{k*}\|_\infty \leq 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right),$$

where $x^{k,(t)}$ is as defined in Algorithm 1.

Proof Note that by Lemma 8 we have that $sgn(x^{k,(t)}) = sgn(x^{k*})$. Thus for any $l \in \{1, \ldots, r\}$ if $l \notin supp(x^{k*})$ then the choice of threshold of $\tau^{(t)} = 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right)$ implies that,

$$|x_l^{k,(t)} - x_l^{k*}| = |x_l^{k,(t)}| \leq 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right).$$

While for $l \in supp(x^{k*})$ Corollary 7 implies

$$|x_l^{k,(t)} - x_l^{k*}| \leq 16R^{(t-1)}M\left(R^{(t-1)}(s+1) + \frac{s}{\sqrt{d}}\right).$$

This completes the proof. ∎

A.2 Finite Sample Guarantees

In this section, we establish finite sample guarantees and state convergence results used in the proof of convergence of our algorithm.

Theorem 10. *Let $\epsilon_n \leq \frac{R}{8\eta}$, where $\frac{3r}{4s} \leq \eta \leq \frac{r}{s}$ is the step-size used at each gradient step. If we are given n i.i.d. samples at each round where $n = \Omega(\frac{r}{sR^2}\log(dr/\delta))$, then we have the guarantee that*

$$\max_{i \in \{1,\ldots,r\}, j \in \{1,\ldots,d\}} \{|\hat{g}_{ij} - g_{ij}|\} \leq \epsilon_n,$$

with probability $1 - \delta$.

Proof We define the set $W = \{m : j \in supp(x^{m*})\}$ and then we have that

$$\hat{g}_{ij} = \frac{|W|}{n} \cdot \underbrace{\frac{1}{|W|} \sum_{m \in W} \left(\sum_k a_{ik}x_k^m - a_{ik}^* x_k^{m*} \right) x_j^m}_{\triangleq \hat{g}_{ij}^W}.$$

Let x^{l*} be a sample such that $l \in W$. We will bound the term $\Lambda = |\left(\sum_k a_{ik}x_k^l - a_{ik}^* x_k^{l*} \right) x_j^l|$ and later invoke McDiarmid's inequality. To ease notation we drop the superscript l. Expanding Λ we get

$$\Lambda = \left| \sum_{k=1}^r (a_{ik} - a_{ik}^*)(x_k - x_k^*)(x_j - x_j^*) + (a_{ik} - a_{ik}^*)(x_k - x_k^*)x_j^* \right.$$

$$\left. + (a_{ik} - a_{ik}^*)x_k^*(x_j - x_j^*) + (a_{ik} - a_{ik}^*)x_k^* x_j^* - a_{ik}^* x_k^*(x_j - x_j^*) - a_{ik}^* x_k^* x_j^* \right|$$

Recall that by Lemma 8 we have that $sgn(x^l) = sgn(x^{l*})$, and x^{l*} is s-sparse thus only s terms in the above sum are non-zero. We repeatedly use the bounds,

1. $|a_{ik}^*| \leq \frac{C_b}{s}$.
2. $|a_{ik} - a_{ik}^*| \leq R \leq R^{(0)} \leq \frac{C_b}{2s}$.
3. $\|x - x^*\|_\infty \leq 16RM \left(R(s+1) + \frac{s}{\sqrt{d}} \right)$.
4. $2 \leq s \leq 2\sqrt{d}$.

Using these we can upper bound Λ by

$$\Lambda \quad \leq \quad \frac{3C_b M^2}{4} \quad + \quad \frac{10C_b^2 M^2}{s} \left(\frac{C_b(s+1)}{2s} + 2 \right) \quad + \quad \frac{C_b}{2} \left(\frac{8C_b M}{s} \left(\frac{C_b(s+1)}{2s} + 2 \right) \right)^2.$$

By our choice of $C_b = 1/2000M^2$, where $M > 1$ we have that

$$\Lambda \leq B,$$

for an appropriate global constant B (independent of s and M).

By simple concentration arguments we can get that $|W|/n$ is close to s/r. Conditioned on a value of $|W|$ by invoking McDiarmid's inequality (Theorem 11), we have that $|\hat{g}_{ij}^W - \mathbb{E}\left[\hat{g}_{ij}|j \in supp(x^*)\right]| \leq \epsilon_{W,n}$ with probability $1 - 2e^{-2|W|\epsilon_{W,n}^2/B^2}$. We demand

$$\epsilon_{W,n} = \frac{C \cdot r \cdot R}{8s\eta}, \tag{11}$$

with probability $1 - c\delta/dr$ for every (i,j), where c and C are appropriate constants such that $|\hat{g}_{ij} - g_{ij}| \leq R/8\eta$ with probability at least $1 - \delta/dr$. Thus we need $|W| = \Omega((\frac{s\eta}{rR})^2 \log(dr/\delta))$. As η is proportional to r/s, this implies that for (11) to hold, we need that $|W| = \Omega(1/R^2 \log(dr/\delta))$.

As stated above we have that $|W|/n$ is close to s/r so if $|W| = \Omega(1/R^2 \log(dr/\delta))$ it suffices to have $n = \Omega(\frac{r}{sR^2}\log(dr/\delta))$. We finish the proof by a union bound over all entries of the matrix. ∎

A.3 Concentration Theorems

We recall McDiarmid's inequality (McDiarmid, 1989).

Theorem 11. *Let $X_1, \ldots, X_m$ be independent random variables all taking values in the set $\mathcal{X}$. Further, let $f : \mathcal{X}^m \mapsto \mathbb{R}$ be a function of $X_1, X_2, \ldots, X_m$ that satisfies $\forall i, \forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$,*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i.$$

Then for all $\epsilon > 0$,

$$\mathbb{P}(f - \mathbb{E}[f] \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

Next we present a concentration theorem for a sum of the squares of d independent Gaussian random variables each with variance σ^2 (χ^2-concentration theorem).

Theorem 12 (Gaussian concentration inequality, see Theorem 5.6 in (Boucheron et al., 2013)). *Let $X = (X_1, \ldots, X_n)$ be a vector of n independent standard normal random variables. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ denote an L-Lipschitz function with respect to Euclidean distance. Then, for all $t > 0$,*

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq e^{-t^2/(2L^2)}.$$

Lemma 13. *If $\{Z_k\}_{k=1}^d \sim \mathcal{N}(0,1)$ are i.i.d. standard normal variables, then $Y \triangleq \sigma^2 \sum_{k=1}^d Z_k^2$ is a scaled chi-squared variate with d degrees of freedom. Define $V \triangleq \sqrt{Y}$, then for all $\delta > 0$ we have*

$$\mathbb{P}\left[V \geq \sigma\sqrt{d} + \delta\right] \leq \exp\left(-\frac{\delta^2}{2\sigma^2}\right).$$

Proof Note that by definition $V(Z_1, \ldots, Z_d)$ is a σ-Lipschitz function of d standard normal variables. By Jensen's inequality we have,

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}[V^2]} = \sigma\sqrt{d}.$$

Thus by applying Theorem 12 to V we have the claimed bound. ∎

B   Analysis of Robust Sparse Estimator

Analysis of the $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector (2) is presented in (Belloni et al., 2014), which we adapt here to present guarantees for deterministic (worst case) perturbations to the dictionary. The analysis in (Belloni et al., 2014) is in a setting where the error in the A is random with zero mean. Here, we consider the error to be deterministic (worst case). Let us start by introducing some notation and important definitions.

B.1   Notation and Definitions

Let $J \subset \{1, \ldots, r\}$ be a set of integers. For a vector $\theta = (\theta_1, \ldots, \theta_r) \in \mathbb{R}^r$ we denote by θ_J the vector in $\mathbb{R}^r$ whose j^{th} component satisfies $(\theta_J)_j = \theta_j$ if $j \in J$, and $(\theta_J)_j = 0$ otherwise. Let $diag(\cdot)$ be the matrix formed by just the diagonal entries and zeroing out the off diagonal terms. Also let $\Delta \triangleq \hat{\theta} - \theta^*$ and $W \triangleq A - A^*$, where θ^* is the true parameter and A^* is the true dictionary without error. Define the cone,

$$\mathcal{C}_J(u) \triangleq \{\Delta \in \mathbb{R}^r : \|\Delta_{J^c}\|_1 \leq u\|\Delta_J\|_1\},$$

where J is a subset of $\{1, \ldots, r\}$. For $q \in [1, \infty]$ and an integer $s \in [1, r]$, the ℓ_q-sensitivity (see for example Gautier and Tsybakov (2011); Rosenbaum et al. (2013); Belloni et al. (2014, 2016)) is defined as

$$\kappa_q(s, u) \triangleq \min_{J:|J| \leq s} \left(\min_{\Delta \in \mathcal{C}_J(u):\|\Delta\|_q=1} \frac{1}{d}\|A^{*\top} A^* \Delta\|_\infty\right).$$

The ℓ_q-sensitivity is routinely used to study convergence of estimators under sparsity constraints. If we have $\kappa_q(s, u) \geq cs^{-1/q}$ for some constant $c > 0$, this leads to optimal bounds for the errors. It has also been shown to be a strict generalization of the restricted eigenvalue property and of the mutual incoherence condition. Relations between these conditions are provided by Lemma 6 of Belloni et al. (2016). We restate that lemma here.

Lemma 14 (Restated from Belloni et al. (2016)). *Let $u > 0$. For any $\alpha \in (0, 1)$ there exists a $c > 0$ such that for $1 \le s \le r$ and $1 \le d \le r$ with $\mu/\sqrt{d} \le 1/(cs)$ then*

$$\kappa_\infty(s, u) \ge \alpha.$$

Furthermore, for any $1 \le q \le \infty$,

$$\kappa_q(s, u) \ge \left(\frac{1}{2s}\right)^{1/q} \kappa_\infty(s, u).$$

Next we highlight the assumptions under which we can establish guarantees for this estimator.

B.2 Assumptions

We make the following assumptions in the analysis of $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector.

(D1) We assume that the true dictionary A^* is deterministic. We also assume that A is deterministic.

(D2) We assume that the columns of A^* are normalized, that is, $\|A_i^*\|_2 = 1 \ \forall i \in \{1, 2, \ldots, r\}$.

(D3) For the matrix A^* we assume the ℓ_∞-sensitivity is bounded below

$$\kappa_\infty(s, 1 + \lambda + \nu) \ge 1/4.$$

(D4) We demand that $\|W\|_\infty \le R$.

(D5) Finally, the tuning parameters λ and ν are chosen such that

$$8s \underbrace{\left(\frac{\left(\sqrt{s}R^2 + \sqrt{\frac{s}{d}}R\right)\left(1 + \nu + \frac{2\lambda}{\sqrt{s}}\right)}{\lambda} + \frac{R^2(1 + \lambda)}{\nu}\right)}_{\triangleq \zeta} \le \frac{1}{2}.$$

Remark 15. *If the dictionary A^* is $\mu/\sqrt{d}$-incoherent and if the sparsity level $s \le C\sqrt{d}/\mu$ for an appropriate global constant C then by Lemma 14 Assumption (D3) holds for A^*.*

Theorem 16 (Adapted from Belloni et al. (2014)). *Let assumptions (D1) - (D5) hold. Assume that the true parameter θ^* is $s-$sparse and belongs to Θ. Let $0 < \lambda, \nu < \infty$, $\gamma = \sqrt{s}R^2 + \sqrt{\frac{s}{d}}R$, and let $\hat{\theta}$ be the $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector. Then*

$$\|\hat{\theta} - \theta^*\|_\infty \le 16(\gamma\|\theta^*\|_2 + R^2\|\theta^*\|_\infty).$$

Proof Throughout the proof, $J = \{j : \theta_j^* \ne 0\}$. We proceed in three steps. Step 1 establishes initial relations and the fact that $\Delta = \hat{\theta} - \theta^*$ belongs to $\mathcal{C}_J(1 + \lambda + \nu)$. Step 2 provides a bound on $\frac{1}{d}\|A^\top A\Delta\|_\infty$. Finally, Step 3 establishes the rate of convergence stated in the theorem. We also often use the inequality $\|\theta\|_\infty \le \|\theta\|_2 \le \|\theta\|_1, \forall \theta \in \mathbb{R}^r$.

Step 1: We first note that,

$$
\begin{aligned}
\frac{1}{d}\left\|A^\top(y - A\theta^*)\right\|_\infty &= \frac{1}{d}\|A^\top W\theta^*\|_\infty \\
&\overset{(i)}{\le} \frac{1}{d}\left\|A^{*\top}W\theta^*\right\|_\infty + \frac{1}{d}\left\|W^\top W\theta^*\right\|_\infty \\
&\overset{(ii)}{\le} \underbrace{\frac{1}{d}\left\|A^{*\top}W\theta^*\right\|_\infty}_{\triangleq n_1} + \underbrace{\frac{1}{d}\left\|(W^\top W - diag(W^\top W))\theta^*\right\|_\infty}_{\triangleq n_2} \\
&\qquad + \underbrace{\frac{1}{d}\left\|diag(W^\top W)\theta^*\right\|_\infty}_{\triangleq n_3},
\end{aligned}
$$

where both (i), (ii) follow by applications of the triangle inequality. Next we bound n_1

$$n_1 = \frac{1}{d}\|A^{*\top}W\theta^*\|_\infty.$$

17

Note that the columns of A^* are normalized, $\|A_i^*\|_2 = 1$ and we have $\|W\|_\infty \le R$, thus we have all elements of $A^{*\top}W$ are bounded by $\sqrt{d}R$. We also know that θ^* is s-sparse, combining these we get,

$$
\begin{aligned}
n_1 &= \frac{1}{d}\|A^{*\top}W\theta^*\|_\infty \\
&\le \frac{1}{d}(\|\theta^*\|_2)(\sqrt{s}\|A^{*\top}W\|_\infty) \\
&\le \frac{1}{d}(\|\theta^*\|_2)(\sqrt{s}dR) \\
&\le \|\theta^*\|_2\left(\sqrt{\frac{s}{d}}R\right),
\end{aligned}
$$

where the last step is by Cauchy-Schwartz. Next for n_2

$$
n_2 = \frac{1}{d}\left\|(W^\top W - diag(W^\top W))\theta^*\right\|_\infty.
$$

We know that $\|W\|_\infty \le R$, thus we have $\|W^\top W - diag(W^\top W)\|_\infty \le dR^2$. Again using the fact that θ^* is s-sparse we have,

$$
\begin{aligned}
n_2 &= \frac{1}{d}\left\|(W^\top W - diag(W^\top W))\theta^*\right\|_\infty \\
&\le \frac{1}{d}(\|\theta^*\|_2)(\sqrt{s}\|W^\top W - diag(W^\top W)\|_\infty) \\
&\le \frac{1}{d}(\|\theta^*\|_2)(\sqrt{s}dR^2) \\
&= \|\theta^*\|_2\sqrt{s}R^2,
\end{aligned}
$$

where the first inequality follows by an application of Cauchy-Schwartz. Finally for n_3, we again have $\|W^\top W\|_\infty \le dR^2$, thus by Cauchy-Schwartz inequality

$$
n_3 = \frac{1}{d}\left\|diag(W^\top W)\theta^*\right\|_\infty \le \|\theta^*\|_\infty R^2.
$$

Combining these together we get,

$$
\frac{1}{d}\|A^\top(y - A\theta^*)\|_\infty \le \left(\sqrt{s}R^2 + \sqrt{\frac{s}{d}}R\right)\|\theta^*\|_2 + R^2\|\theta^*\|_\infty. \tag{12}
$$

As $\gamma = \sqrt{s}R^2 + \sqrt{\frac{s}{d}}R$, this implies that $(\theta, t, u) = (\theta^*, \|\theta^*\|_2, \|\theta^*\|_\infty)$ is feasible. Let $(\hat\theta, \hat t, \hat u)$ be the optimal solution, then we have

$$
\|\hat\theta\|_1 + \lambda\|\hat\theta\|_2 + \nu\|\hat\theta\|_\infty \le \|\hat\theta\|_1 + \lambda\hat t + \nu\hat u \le \|\theta^*\|_1 + \lambda\|\theta^*\|_2 + \nu\|\theta^*\|_\infty.
$$

By rearranging terms and by triangle inequality we get the relation

$$
\|\hat\theta_{J^C}\|_1 \le (1 + \lambda + \nu)\|\hat\theta_J - \theta^*\|_1 = (1 + \lambda + \nu)\|\Delta_J\|_1.
$$

This proves that $\Delta \in \mathcal{C}_J(1 + \lambda + \nu)$. Also by similar arguments we get

$$
\hat t - \|\theta^*\|_2 \le \frac{\|\Delta\|_1 + \nu\|\Delta\|_\infty}{\lambda} \le \frac{(1 + \nu)\|\Delta\|_1}{\lambda} \tag{13}
$$

$$
\text{and,} \quad \hat u - \|\theta^*\|_\infty \le \frac{\|\Delta\|_1 + \lambda\|\Delta\|_2}{\nu} \le \frac{(1 + \lambda)}{\nu}\|\Delta\|_1. \tag{14}
$$

Step 2: By applications of the triangle inequality we have

$$
\begin{aligned}
\frac{1}{d}\|A^{*\top}A^*\Delta\|_\infty &\le \frac{1}{d}\left[\|A^\top A^*\Delta\|_\infty + \|W^\top A^*\Delta\|_\infty\right] \\
&\le \frac{1}{d}\left[\|A^\top A\Delta\|_\infty + \|A^\top W\Delta\|_\infty + \|W^\top A^*\Delta\|_\infty\right] \\
&\le \frac{1}{d}\Big[\underbrace{\|A^\top(y - A\theta^*)\|_\infty}_{\triangleq m_1} + \underbrace{\|A^\top(y - A\hat\theta)\|_\infty}_{\triangleq m_2} + \underbrace{\|A^\top W\Delta\|_\infty}_{\triangleq m_3} + \underbrace{\|W^\top A^*\Delta\|_\infty}_{\triangleq m_4}\Big].
\end{aligned}
$$

18

Now we bound each of these terms

$$
m_1 \overset{(i)}{\leq} d(\gamma \|\theta^*\|_2 + R^2 \|\theta\|_\infty)
$$

$$
m_2 \overset{(ii)}{\leq} d(\gamma \hat{t} + R^2 \hat{u}) \leq d \left(\gamma \|\theta^*\|_2 + R^2 \|\theta^*\|_\infty + \left\{ \frac{\gamma(1+\nu)}{\lambda} + \frac{R^2(1+\lambda)}{\nu} \right\} \|\Delta\|_1 \right)
$$

$$
m_3 \overset{(iii)}{\leq} \left(dR^2 + \sqrt{d}R \right) \|\Delta\|_1 + dR^2 \|\Delta\|_\infty
$$

$$
m_4 \overset{(iv)}{\leq} \sqrt{d}R \|\Delta\|_1,
$$

where (i) follows as $(\theta^*, \|\theta^*\|_2, \|\theta^*\|_\infty)$ is a feasible point, (ii) is because $(\hat{\theta}, \hat{t}, \hat{u})$ is a (optimal) feasible point along with (13), (14). Bound (iii) follows by similar arguments made to arrive at (12) and finally (iv) is due to Hölder's inequality. Combing these we have the following bound

$$
\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty + \left\{ \frac{\gamma(1+\nu)}{\lambda} + \frac{R^2(1+\lambda)}{\nu} \right\} \|\Delta\|_1
$$
$$
+ \frac{R}{\sqrt{d}} \|\Delta\|_1 + R^2 \|\Delta\|_\infty + \left(R^2 + \frac{R}{\sqrt{d}} \right) \|\Delta\|_1.
$$

Simplifying using $\|\Delta\|_\infty \leq \|\Delta\|_1$ and $\gamma = \sqrt{s} \left(R^2 + \frac{R}{\sqrt{d}} \right)$ we get

$$
\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty + \left(\gamma \frac{\left(1 + \nu + \frac{2\lambda}{\sqrt{s}} \right)}{\lambda} + \frac{R^2(1+\lambda)}{\nu} \right) \|\Delta\|_1. \quad (15)
$$

Step 3: Define

$$
\zeta \triangleq \left(\gamma \frac{\left(1 + \nu + \frac{2\lambda}{\sqrt{s}} \right)}{\lambda} + \frac{R^2(1+\lambda)}{\nu} \right).
$$

Rewriting (15) using the definition of ζ we have

$$
\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty + \zeta \|\Delta\|_1.
$$

By the assumption on ℓ_∞-sensitivity and Lemma 14 we have $\kappa_1(s, u) \geq \frac{1}{2s} \kappa_\infty(s, u) \geq \frac{1}{8s}$. Thus by definition of ℓ_1-sensitivity we have

$$
\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \geq \kappa_1(s, 1 + \lambda + \nu) \|\Delta\|_1.
$$

Combining this with the previous display gives us

$$
\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty + \frac{\zeta}{\kappa_1(s, 1 + \lambda + \nu)} \left(\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \right)
$$
$$
\leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty + 8s\zeta \left(\frac{1}{d} \|A^{*\top} A^* \Delta\|_\infty \right).
$$

By assumption (D5) $- 8s\zeta \leq 1/2$, therefore we have the claimed error bound

$$
\frac{1}{2d} \|A^{*\top} A^* \Delta\|_\infty \leq 2\gamma \|\theta^*\|_2 + 2R^2 \|\theta^*\|_\infty
$$
$$
\kappa_\infty(s, 1 + \lambda + \nu) \|\Delta\|_\infty \leq 4\gamma \|\theta^*\|_2 + 4R^2 \|\theta^*\|_\infty
$$
$$
\|\hat{\theta} - \theta^*\|_\infty \leq 16(\gamma \|\theta^*\|_2 + R^2 \|\theta^*\|_\infty).
$$

∎

C   Lower Bounds: Proof of Theorem 1

In this section we will show that when the uncertainty in the dictionary measured in matrix infinity norm scales as $R = \mathcal{O}(1/\sqrt{s})$, the $\{\ell_1, \ell_2, \ell_\infty\}$-MU Selector is information theoretically optimal up to logarithmic factors and the infinity norm of the error (in the worst case) is lower bounded by $CR\|\theta^*\|_2$. We will prove this by Fano's method (see for example review in Yu (1997); Tsybakov (2009)). The proof technique to show this estimator is minimax optimal is adapted from Belloni et al. (2016). We define the sets

$$B_0(s) = \{\theta : \|\theta\|_0 \leq s\} \qquad \text{and} \qquad S_2(L) = \{\theta : \|\theta\|_2 = L\},$$

where $L > 0$. We define the parameter set to be $\Theta = B_0(s) \cap S_2(L)$, which is the set of $s-$sparse vectors with $\|\cdot\|_2$ norm equal to L. To prove this theorem we will choose a particular probability distribution over the set of underlying *true dictionaries* $\mathbb{P}_{A^*}$ and also a distribution over the deviations from the true dictionary $\mathbb{P}_W$. We will assume that the entries of A^* are drawn i.i.d. from a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma_D^2)$ and the entries of W are chosen i.i.d. from a zero mean Gaussian distribution $\mathcal{N}(0, \sigma_E^2)$ independent of the distribution generating A^*. We set $\sigma_D = \mathcal{O}(1/\sqrt{d})$ and $\sigma_E = \mathcal{O}(R/\sqrt{log(dr)})$. We now restate a formal version of Theorem 1.

Theorem 17. *Let $r \geq 2$, $2 \leq s \leq r$, and $L > 0$. Let $y = A^*\theta^*$ where $A^* \in \mathbb{R}^{d \times r}$ and θ^* is a s-sparse vector with norm $\|\theta^*\|_2 = L$. Further let the entries of A^* be drawn from $\mathcal{N}(0, \sigma_D^2)$ and independently let the entries of the perturbation W be drawn from the distribution $\mathcal{N}(0, \sigma_E^2)$. Let $A = A^* + W$, $\sigma_D^2 = \mathcal{O}(1/d)$ and $\sigma_E^2 = R/\log(dr)$. Then there exists constants C and $C' > 0$ such that*

$$\inf_{\hat{T}} \sup_{\theta \in B_0(s) \cap S_2(L)} \mathbb{P}_{A^*,W} \left[\|\hat{T} - \theta\|_\infty \geq CRL\sqrt{1 - \frac{\log(s)}{\log(r)}} \right] > C',$$

where $\inf_{\hat{T}}$ denotes the infimum over all measurable estimators $\hat{T}$ with input (y, A, R).

Proof We define a finite set of "hypotheses" (packing set) included in $B_0(s) \cap S_2(L)$. To this end, we first introduce

$$\mathcal{M} = \{x \in \{0,1\}^{r-1} : \rho_H(\mathbf{0}, x) = s - 1\},$$

where ρ_H denotes the Hamming distance between elements of $\{0,1\}^{r-1}$, and $\mathbf{0}$ is the zero vector. Then there exists a subset $\mathcal{M}'$ of $\mathcal{M}$ such that for any x, x' in $\mathcal{M}'$ with $x \neq x'$, we have $\rho_H(x, x') > s/16$ and moreover the cardinality of $\mathcal{M}'$ is bounded below

$$\log|\mathcal{M}'| \geq Cs \log\left(\frac{r}{s}\right),$$

for some constant C. This follows from Varshamov-Gilbert bound (see Lemma 2.9 in Tsybakov (2009)) if $s - 1 > (r-1)/2$ and from Lemma A.3 in Rigollet and Tsybakov (2011) if $s - 1 \leq (r-1)/2$. We denote ω_j' to be the elements of the finite set $\mathcal{M}'$. For $j = 1, \ldots, |\mathcal{M}|$, we define the vectors $\omega_j \in \{0,1\}^r$ with components $\omega_{j1} = 0$ and $\omega_{jk} = \omega_{j(k-1)}'$ for $k > 2$, where ω_{jk} is the k-th component of ω_j. We also define ω_0 as the vector in $\{0,1\}^r$ with all components equal to 0 except the first one equal to 1. We now define the set of "hypotheses" (packing set of Θ) $(\bar{\omega}_j, j = 0, \ldots, |\mathcal{M}'| + 1)$, where $\bar{\omega}_0 = R\omega_0$ and

$$\bar{\omega}_j = \frac{L}{\sqrt{1 + \psi^2(s-1)}}(\omega_0 + \psi\omega_j), \qquad j = 1, \ldots, |\mathcal{M}'| + 1.$$

Here ψ is a positive parameter that will be chosen appropriately. Note that these vectors are s-sparse and have $\|\bar{\omega}_j\|_2 = L$. By Lemma 18 we have the KL divergence is bounded,

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_{\bar{\omega}_j}, \mathbb{P}_{\bar{\omega}_0}) &= \frac{d\sigma_D^2}{2\sigma_E^2 \|\bar{\omega}_0\|_2^2} \|\bar{\omega}_j - \bar{\omega}_0\|^2 \\
&\leq \frac{d\sigma_D^2}{2\sigma_E^2 L^2} \left(\frac{\psi^2 L^2 s}{1 + \psi^2(s-1)} \right) \\
&\leq \psi^2 \left(\frac{s d\sigma_D^2}{2\sigma_E^2 (1 + \psi^2(s-1))} \right).
\end{aligned}$$

20

If we choose $\psi = C\sqrt{\frac{\sigma_E^2 \log(r/s)}{d\sigma_D^2}}$ with C being an appropriately chosen constant independent of dimensions (s, d, r) and L we get that for all j,

$$\mathcal{K}(\mathbb{P}_{\bar{\omega}_j}, \mathbb{P}_{\bar{\omega}_0}) \leq \frac{1}{16} \log(|\mathcal{M}'|).$$

Thus for j and j' both different from 0,

$$\|\bar{\omega}_j - \bar{\omega}_{j'}\|_\infty = \frac{L\psi}{\sqrt{1 + \psi^2(s-1)}} \geq C \frac{L\sigma_E \sqrt{\log(r/s)}}{\sqrt{d}\sigma_D},$$

and for $j \neq 0$ we have

$$\|\bar{\omega}_j - \bar{\omega}_0\|_\infty \geq \frac{L\psi\|\omega_j\|_\infty}{\sqrt{1 + \psi^2(s-1)}} \geq C \frac{L\sigma_E \sqrt{\log(r/s)}}{\sqrt{d}\sigma_D}.$$

We want the columns of $\|A^*\|_2 \leq 1$ (upper bound used in the proof of Theorem 16), hence we want $\sigma_D = \mathcal{O}(1/\sqrt{d})$ (this follows by an application of Lemma 13 followed by a union bound over the r columns using the fact that $r = \mathcal{O}(poly(d))$). We also demand that our deviation from the true dictionary be bounded by R with high probability over all entries so we choose $\sigma_E \leq \mathcal{O}(R/\sqrt{log(dr)})$. Hence given our choices of σ_E and σ_D we have for any j, j'

$$\|\bar{\omega}_j - \bar{\omega}_{j'}\|_\infty \geq CLR\left(\sqrt{1 - \frac{\log(s)}{\log(r)}}\right).$$

We can now apply Theorem 2.7 in Tsybakov (2009) to complete the proof. ∎

Lemma 18. *Let $\theta_1 \in \mathbb{R}^r$ and $\theta_2 \in \mathbb{R}^r$ be such that $\|\theta_1\|_2 = \|\theta_2\|_2$. Under the assumptions stated in the Appendix C we have*

$$\mathcal{K}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{d\sigma_D^2}{2\sigma_E^2\|\theta_2\|_2^2}\|\theta_1 - \theta_2\|^2.$$

Proof By the properties of Kullback Leibler divergence between product measures, it suffices to prove the lemma for $d = 1$. Let $\theta \in \mathbb{R}^r$. Consider the random vector (U, V) where

$$V = (D_1 + E_1, \ldots, D_r + E_r),$$

with $D = (D_1, D_2, \ldots, D_r)^\top$ a zero-mean Gaussian vector with covariance $\sigma_D^2 I_{r\times r}$ and $E = (E_1, E_2, \ldots, E_r)^\top$ a zero mean Gaussian vector with covariance $\sigma_E^2 I_{r\times r}$ independent of A and

$$U = \sum_{j=1}^r \theta_j(V_j - E_j).$$

We introduce some variables

$$\tilde{\Sigma} = \frac{\sigma_E^2}{\sigma_D^2 + \sigma_E^2} I_{r\times r}, \quad \Pi = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_E^2} I_{r\times r}, \quad c_\theta = \theta^\top \Pi \theta = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_E^2}\|\theta\|_2^2.$$

We find the conditional distribution $\mathcal{L}_\theta(U|V)$ of U given V. Also note that the vector $(V_1, \ldots, V_r, E_1, \ldots, E_r)^\top$ is a zero-mean Gaussian random vector with covariance matrix

$$\begin{pmatrix} (\sigma_D^2 + \sigma_E^2)I_{r\times r} & \sigma_E^2 I_{r\times r} \\ \sigma_E^2 I_{r\times r} & \sigma_E^2 I_{r\times r} \end{pmatrix}.$$

So that $\mathcal{L}_\theta(E|V)$ is a Gaussian with mean $\tilde{\Sigma}V$ and covariance $\sigma_E^2(I_{r\times r} - \tilde{\Sigma})$. This implies that $\mathcal{L}_\theta(U|V)$ is Gaussian with mean $\theta^\top \Pi \theta$ and variance $c_\theta \sigma_E^2$. Then the logarithm of density of $\mathcal{L}_\theta(U|V)$, denoted by $\ell_\theta(U|V)$ satisfies

$$\ell_\theta(U|V) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(c_\theta\sigma_E^2) - \frac{1}{2c_\theta\sigma_E^2}(U - \theta^\top \Pi V)^2.$$

21

Now let $\theta_1 \in \mathbb{R}^r$ and $\theta_2 \in \mathbb{R}^r$ with $\|\theta_1\|_2 = \|\theta_2\|_2$. Then,

$$\ell_{\theta_1}(U|V) - \ell_{\theta_2}(U|V) = \frac{1}{2} \underbrace{\left(\log \left(\frac{c_{\theta_2}}{c_{\theta_1}} \right) \right)}_{=0} + \frac{1}{2c_{\theta_2}\sigma_E^2} \left((U - \theta_2^\top \Pi V)^2 - (U - \theta_1^\top \Pi V)^2 \right)$$

$$+ \underbrace{\left(\frac{1}{2c_{\theta_2}\sigma_E^2} - \frac{1}{2c_{\theta_1}\sigma_E^2} \right)}_{=0} (U - \theta_1^\top \Pi V)^2$$

$$= \frac{1}{2c_{\theta_2}\sigma_E^2} \left((U - \theta_2^\top \Pi V)^2 - (U - \theta_1^\top \Pi V)^2 \right).$$

Since the distribution of V does not depend on θ_1 we obtain that in the case $d = 1$,

$$\mathcal{K}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{1}{2c_{\theta_2}\sigma_E^2} \mathbb{E}_{\theta_1} \left[(U - \theta_2^\top \Pi V)^2 - (U - \theta_1^\top \Pi V)^2 \right]$$

$$= \frac{\sigma_D^2 + \sigma_E^2}{2\sigma_E^2 \sigma_D^2 \|\theta\|_2^2} \left[\sigma_D^2 (\theta_1^\top - \theta_2^\top \Pi) I_{r \times r} (\theta_1 - \Pi \theta_2) \right.$$

$$\left. - \sigma_D^2 (\theta_1^\top - \theta_1^\top \Pi) I_{r \times r} (\theta_1 - \Pi \theta_1) + (\theta_2^\top \Pi^2 \theta_2 - \theta_1^\top \Pi^2 \theta_1) \right].$$

Where in the final step the cross terms are zero by the independence of D and E. Developing this expression leaves us with

$$\mathcal{K}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{\sigma_D^2 + \sigma_E^2}{2\sigma_E^2 \sigma_D^2 \|\theta\|_2^2} \left[(\theta_1 - \theta_2)^\top \Pi \sigma_D^2 I_{r \times r} (\theta_1 - \theta_2) \right]$$

$$= \frac{\sigma_D^2}{2\sigma_E^2 \|\theta_2\|_2^2} \|\theta_1 - \theta_2\|^2.$$

$\blacksquare$