PROVABLE ONLINE DICTIONARY LEARNING AND SPARSE CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the dictionary learning problem, where the aim is to model the given data as a linear combination of a few columns of a matrix known as a dictionary, where the sparse weights forming the linear combination are known as coefficients. Since both the dictionary and coefficients parameterizing the linear model are unknown, the corresponding optimization is inherently non-convex. This was a major challenge until recently, when provable algorithms for dictionary learning were proposed. Yet, these provide guarantees only on the recovery of the dictionary, without explicit recovery guarantees on the coefficients. Moreover, any estimation error in the dictionary adversely impacts the ability to successfully localize and estimate the coefficients. This potentially limits the utility of existing provable dictionary learning methods in applications where coefficient recovery is of interest. To this end, we develop a simple online alternating optimization-based algorithm for dictionary learning, which recovers both the dictionary and coefficients exactly. Specifically, we show that – when initialized appropriately – the algorithm linearly converges to the true factors. Our algorithm is also scalable and amenable for large scale distributed implementations in neural architectures, by which we mean that it only involves simple linear and non-linear operations. Finally, we corroborate these theoretical results via empirical evaluation of convergence properties, and phase transition in sample complexity.

1 INTRODUCTION

Sparse models avoid overfitting by favoring simple yet highly expressive representations. Since signals of interest may not be inherently sparse, expressing them as a sparse linear combination of a few columns of a dictionary is used to exploit the sparsity properties. Of specific interest are overcomplete dictionaries, since they provide a flexible way of capturing the richness of a dataset, while yielding sparse representations that are robust to noise; see Mallat and Zhang (1993); Chen et al. (1998); Donoho et al. (2006). In practice however, these dictionaries may not be known, warranting a need to learn such representations – known as dictionary learning (DL) or sparse coding (Olshausen and Field, 1997). Formally, this entails learning an a priori unknown dictionary \( \mathbf{A} \in \mathbb{R}^{n \times m} \) and sparse coefficients \( \mathbf{x}^*_j \in \mathbb{R}^m \) from data samples \( y_j \in \mathbb{R}^n \) generated as

\[
y_j = \mathbf{A}^* \mathbf{x}^*_j, \quad \|\mathbf{x}^*_j\|_0 \leq k \quad \text{for all} \quad j = 1, 2, \ldots
\]

This particular model can also be viewed as an extension of the low-rank model (Pearson, 1901). Here, instead of sharing a low-dimensional structure, each data vector can now reside in a separate low-dimensional subspace. Therefore, together the data matrix admits a union-of-subspace model. As a result of this additional flexibility, DL finds applications in a wide range of signal processing and machine learning tasks, such as denoising (Elad and Aharon, 2006), image inpainting (Mairal et al., 2009), clustering and classification (Ramirez et al., 2010), and analysis of deep learning primitives (Ranzato et al., 2008; Gregor and LeCun, 2010); see also Elad (2010), and references therein.

Notwithstanding the non-convexity of the associated optimization problems (since both factors are unknown), alternating minimization-based dictionary learning techniques have enjoyed significant success in practice. Popular heuristics include regularized least squares-based (Olshausen and Field, 1997; Lee et al., 2007; Mairal et al., 2009; Lewicki and Sejnowski, 2000; Kreutz-Delgado et al., 2003), and greedy approaches such as the method of optimal directions (MOD) (Engan et al., 1999) and k-SVD (Aharon et al., 2006). However, these are difficult to analyze in theory.
To this end, motivated from a string of recent theoretical works (Gribonval and Schnass, 2010; Jenatton et al., 2012; Geng and Wright, 2014), provable algorithms for DL have been proposed recently to explain the success of aforementioned alternating minimization-based algorithms (Agarwal et al., 2014; Arora et al., 2014; 2015). However, these works exclusively focus on guarantees for dictionary recovery. On the other hand, for applications of DL in tasks such as classification and clustering – which rely on coefficient recovery – it is crucial to have guarantees on coefficients recovery as well.

Contrary to conventional prescription, a sparse approximation step after recovery of the dictionary does not help; since any error in the dictionary – which leads to an error-in-variables (EIV) (Fuller, 2009) model for the dictionary – degrades our ability to even recover the support of the coefficients (Wainwright, 2009). Further, when this error is non-negligible, the existing results guarantee recovery of the sparse coefficients only in $\ell_2$-norm sense (Donoho et al., 2006). As a result, there is a need for scalable dictionary learning techniques with guaranteed recovery of both factors.

1.1 Summary of Our Contributions

In this work, we present a simple online DL algorithm motivated from the following regularized least squares-based problem, where $S(\cdot)$ is a nonlinear function that promotes sparsity.

$$
\min_{\mathbf{A}, \{ \mathbf{x}_j \}} \sum_{j=1}^{p} \| \mathbf{y}_j - \mathbf{A} \mathbf{x}_j \|_2^2 + \sum_{j=1}^{p} S(\mathbf{x}_j).
$$

(P1)

Although our algorithm does not optimize this objective, it leverages the fact that the problem (P1) is convex w.r.t $\mathbf{A}$, given the sparse coefficients $\{ \mathbf{x}_j \}$. Following this, we recover the dictionary by choosing an appropriate gradient descent-based strategy (Arora et al., 2015; Engan et al., 1999). To recover the coefficients, we develop an iterative hard thresholding (IHT)-based update step (Haupt and Nowak, 2006; Blumensath and Davies, 2009), and show that – given an appropriate initial estimate of the dictionary and a mini-batch of $p$ data samples at each iteration of the online algorithm – alternating between this IHT-based update for coefficients, and a gradient descent-based update for the dictionary, recovers the true factors exactly.

In addition to achieving exact recovery of the two factors, our algorithm – Neurally plausible alternating Optimization-based Online Dictionary Learning (NOODL) – has linear convergence properties. Furthermore, it is scalable, and involves simple operations, making it an attractive choice for practical DL applications. Our major contributions are summarized as follows:

- **Provable coefficient recovery:** To the best of our knowledge, this is the first result on exact recovery of the sparse coefficients $\{ \mathbf{x}_j^* \}$, including their support recovery, for the DL problem. The proposed IHT-based strategy to update coefficient under the EIV model, is of independent interest for recovery of the sparse coefficients via IHT, which is challenging even when the dictionary is known; see also Yuan et al. (2016).
- **Unbiased estimation of factors and linear convergence:** The recovery guarantees on the coefficients also helps us to get rid of the bias incurred by the prior-art in dictionary estimation. Furthermore, our technique geometrically converges to the true factors.
- **Online nature and neural implementation:** The online nature of algorithm, makes it suitable for machine learning applications with streaming data. In addition, the separability of the coefficient update allows for distributed implementations in neural architectures (only involves simple linear and non-linear operations) to solve large-scale problems. To showcase this, we also present a prototype neural implementation of NOODL.

In addition, we also verify these theoretical properties of NOODL through experimental evaluations on synthetic data, and compare its performance with state-of-the-art provable DL techniques.

1.2 Related Works

With the success of the alternating minimization-based techniques in practice, a push to study the DL problem began when Gribonval and Schnass (2010) showed that for $m = n$, the solution pair $(\mathbf{A}^*, \mathbf{X}^*)$ lies at a local minima of the following non-convex optimization program, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p]$, with high probability over the randomness of the coefficients,

$$
\min_{\mathbf{A}, \mathbf{X}} \| \mathbf{X} \|_1 \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{A} \mathbf{X}, \quad \| \mathbf{A}_i \| = 1, \forall \ i \in [m],
$$

(2)
Following this, Geng and Wright (2014) and Jenatton et al. (2012) extended these results to the overcomplete case \((n < m)\), and the noisy case, respectively. Concurrently, Jung et al. (2014; 2016) studied the nature of the DL problem for \(S(\cdot) = \| \cdot \|_1\) (in (P1)), and derived a lower-bound on the minimax risk of the DL problem. However, these works do not provide any algorithms for DL.

Motivated from these theoretical advances, Spielman et al. (2012) proposed an algorithm for the under-complete case \(n \geq m\) that works up-to a sparsity of \(k = O(\sqrt{n})\). Later, Arora et al. (2014) and Arora et al. (2014) proposed clustering-based provable algorithms for the overcomplete setting, motivated from MOD (Engan et al., 1999) and k-SVD (Aharon et al., 2006), respectively. Here, in addition to requiring stringent conditions on dictionary initialization, Agarwal et al. (2014) alternates between solving a quadratic program for coefficients and an MOD-like (Engan et al., 1999) update for the dictionary, which is too expensive in practice. Recently, a DL algorithm that work for almost linear sparsity was proposed by Barak et al. (2015); however, as shown in Table 1, this algorithm may result in exponential running time. Finally, Arora et al. (2015) proposed a provable online DL algorithm, which provided improvements on initialization, sparsity, and sample complexity, and is closely related to our work. We summarize the provable techniques in Table 1.

The provable DL techniques discussed above implicitly assume that coefficients can be recovered – after dictionary recovery – via some sparse approximation technique. However, as alluded to earlier, the guarantees for coefficient recovery – when the dictionary is known approximately – may be limited to some \(\ell_2\) norm bounds (Donoho et al., 2006). This means that, the resulting coefficient estimates may not even be sparse. In addition, for practical applications, there is a need for efficient online algorithms that do not rely on computationally expensive post-processing steps for the coefficient recovery. These serve as the primary motivations for our work.

**Notation.** Given an integer \(n\), we denote \([n] = \{1, 2, \ldots, n\}\). The bold upper-case and lowercase letters are used to denote matrices and vectors \(\mathbf{M}\) and \(\mathbf{v}\), respectively. \(\mathbf{M}, \mathbf{M}_{(i,:)}, \mathbf{M}_{,:}, \mathbf{v}_i, \mathbf{v}(i)\) (and \(\mathbf{v}(i)\)) denote the \(i\)-th column, \(i\)-th row, \((i, j)\) element of a matrix, and \(i\)-th element of a vector, respectively. We denote \(\mathbf{D}(\mathbf{v})\) as a diagonal matrix with elements of a vector \(\mathbf{v}\) on the diagonal. The superscript \((\cdot)^{(n)}\) denotes the \(n\)-th iterate, while the subscript \((\cdot)_{(n)}\) is reserved for the \(n\)-th data sample. Given a matrix \(\mathbf{M}\), we use \(\mathbf{M}_{+} \) to denote a resulting matrix without \(i\)-th column, \(\|\mathbf{M}\|\) and \(\|\mathbf{M}\|_F\) as the spectral norm and Frobenius norm. Given a vector \(\mathbf{v}\), we use \(\|\mathbf{v}\|, \|\mathbf{v}\|_0\), and \(\|\mathbf{v}\|_1\) to denote the \(\ell_2\) norm, \(\ell_0\) (number of non-zero entries), and \(\ell_1\) norm, respectively. We also use standard notations \(O(\cdot), \Omega(\cdot)\) \((\bar{O}(\cdot), \bar{\Omega}(\cdot))\) to indicate the asymptotic behavior (ignoring logarithmic factors). Further, we use \(g(n) = O^\kappa(f(n))\) to indicate that \(g(n) \leq L f(n)\) for a small enough constant \(L\), which is independent of \(n\). We use \(c(\cdot)\) for constants parameterized by the quantities in \((\cdot)\). \(T_r(z) := z \cdot 1_{|z| \geq r}\) denotes the hard-thresholding operator, where \(\bar{1}\) is the indicator function. We use \(\text{supp}(\cdot)\) for the support (the set of non-zero elements) and \(\text{sign}(\cdot)\) for the element-wise sign.

## 2 ALGORITHM

We now detail the specifics of our algorithm – NOODL, outlined in Algorithm 1. NOODL recovers both the dictionary and the coefficients exactly given an appropriate initial estimate \(\mathbf{A}^{(0)}\) of the dictionary. Specifically, we require \(\mathbf{A}^{(0)}\) to be \((\epsilon_0, 2)\)-close to \(\mathbf{A}^*\) for \(\epsilon_0 = O^\kappa(1/\log(n))\), where \((\epsilon, \kappa)\)-closeness is defined as follows. This implies that, the initial dictionary estimate needs to be column-wise, and in spectral norm sense, close to \(\mathbf{A}^*\), which can be achieved via certain initialization algorithms, such as the initialization algorithm presented in Arora et al. (2015).
Input: Fresh data samples $y_{(j)} \in \mathbb{R}^p$ for $j \in [p]$ at each iteration $t$ generated as per (1), where $|x_i^*| \geq C$ for $i \in \text{supp}(x^*)$. Parameters $\eta_A$, $\eta_x^{(r)}$ and $\tau^{(r)}$ chosen as per A.5 and A.6. No. of iterations $T = \Omega(\log(1/\epsilon_T))$ and $R = \Omega(\log(1/\delta_R))$, for target tolerances $\epsilon_T$ and $\delta_R$.

Output: The dictionary $A^{(T)}$ and coefficient estimates $\hat{x}_{(j)}^{(t)}$ for $j \in [p]$ at each iterate $t$.

Initialize: Estimate $A^{(0)}$, which is $(\epsilon_0, 2)$-near to $A^*$ for $\epsilon_0 = \mathcal{O}^*(1/\log(n))$. For $t = 0$ to $T - 1$ do

**Predict:** (Estimate Coefficients)

for $j = 1$ to $p$ do

Initialize: $x_{(j)}^{(0)} = Tc/2(A^{(t)^\top}y_{(j)})$ (3)

for $r = 0$ to $R - 1$ do

Update: $x_{(j)}^{(r+1)} = T_{x(r)}(x_{(j)}^{(r)} - \eta_x^{(r)} A^{(t)^\top}(A^{(t)}x_{(j)}^{(r)} - y_{(j)}))$ (4)

end

end

$\hat{x}_{(j)}^{(t)} := x_{(j)}^{(R)}$ for $j \in [p]$

**Learn:** (Update Dictionary)

Form empirical gradient estimate: $\hat{g}^{(t)} = \frac{1}{p} \sum_{j=1}^p (A^{(t)^\top}x_{(j)}^{(t)} - y_{(j)}) \text{sign}(\hat{x}_{(j)}^{(t)^\top})$ (5)

Take a gradient descent step: $A_{(t+1)} = A_{(t)} - \eta_A \hat{g}^{(t)}$ (6)

Normalize: $A_{(t+1)} = A_{(t+1)}/\|A_{(t+1)}\| \forall i \in [m]$ end

**Definition 1** ($(\epsilon, \kappa)$-closeness). A dictionary $A$ is $(\epsilon, \kappa)$-close to $A^*$ if $\|A - A^*\| \leq \kappa\|A^*\|$, and if there is a permutation $\pi : [m] \rightarrow [m]$ and a collection of signs $\sigma : [m] \rightarrow \{\pm 1\}$ such that $\|\sigma(i)A_{\pi(i)} - A_{\pi(i)}\| \leq \epsilon, \forall i \in [m]$. Due to the streaming nature of the incoming data, NOODL takes a mini-batch of $p$ data samples at the $t$-th iteration of the algorithm, as shown in Algorithm 1. It then proceeds by alternating between two update stages: coefficient estimation (“Predict”) and dictionary update (“Learn”) as follows.

**Predict Stage:** For a general data sample $y = A^*x^*$, the algorithm begins by forming an initial coefficient estimate $x^{(0)}$ based on a hard thresholding (HT) step as shown in (3). Given this initial estimate $x^{(0)}$, the algorithm iterates over $R = \Omega(\log(1/\delta_R))$ IHT-based steps (4) to achieve a target tolerance of $\delta_R$, such that $(1 - \eta_x^{(r)})R \leq \delta_R$. Here, $\eta_x^{(r)}$ is the learning rate, and $\tau^{(r)}$ is the threshold at the $r$-th iterate of the IHT. In practice, these can be fixed to some constants for all iterations; see A.6 for details. Finally at the end of this stage, we have estimate $\hat{x}^{(t)} := x^{(R)}$ of $x^*$.

**Learn Stage:** Using this estimate of the coefficients, we update the dictionary at $t$-th iteration $A^{(t)}$ by an approximate gradient descent step (6), using the empirical gradient estimate (5) and the learning rate $\eta_A = \Theta(m/k)$; see also A.5. Finally, we normalize the columns of the dictionary and continue to the next batch. The running time of each step $t$ of NOODL is therefore $\mathcal{O}(m+n\log(1/\delta_R))$. For a target tolerance of $\epsilon_T$, such that $\|A_{(t)}A_{\pi(i)} - A_{\pi(i)}\| \leq \epsilon_T, \forall i \in [m]$, we choose $T = \Omega(\log(1/\epsilon_T))$.

NOODL uses an initial HT step and an approximate gradient descent-based strategy as in Arora et al. (2015). Following which, our IHT-based coefficient update step yields an estimate of the coefficients at each iteration of the online algorithm. Coupled with the guaranteed progress made on the dictionary, this also removes the bias in dictionary estimation. Further, the simultaneous recovery of both factors also avoids an often expensive post-processing step for recovery of the coefficients.

3 Main Result

We start by introducing a few important definitions. First, as discussed in the previous section we require that the initial estimate $A^{(0)}$ of the dictionary is $(\epsilon_0, 2)$-close to $A^*$. In fact, we require this closeness property to hold at each subsequent iteration $t$, which is a key ingredient in our analysis. This initialization achieves two goals. First, the $\|\sigma(i)A_{\pi(i)} - A_{\pi(i)}\| \leq \epsilon_0$ condition ensures that
the signed-support of the coefficients are recovered correctly (with high probability) by the hard thresholding-based coefficient initialization step, where signed-support is defined as follows.

**Definition 2.** The signed-support of a vector $\mathbf{x}$ is defined as $\text{sign}(\mathbf{x}) \cdot \mathbf{x}(\text{supp}(\mathbf{x}))$.

Next, the $\|A - A^*\| \leq \|A^*\|$ condition keeps the dictionary estimates close to $A^*$ and is used in our analysis to ensure that the gradient direction (5) makes progress. Further, in our analysis, we ensure $\epsilon_t$ (defined as $\|A_{(t)}^* - A^*_t\| \leq \epsilon_t$) contracts at every iteration, and assume $\epsilon_0, \epsilon_t = O^*(1/\log(n))$. Also, we assume that the dictionary $A$ is fixed (deterministic) and $\mu$-incoherent, defined as follows.

**Definition 3.** A matrix $A \in \mathbb{R}^{n \times m}$ with unit-norm columns is $\mu$-incoherent if for all $i \neq j$ the inner-product between the columns of the matrix follow $|\langle A_i^*, A_j^* \rangle| \leq \mu / \sqrt{n}$.

The incoherence parameter measures the degree of closeness of the dictionary elements. Smaller values (i.e., close to 0) of $\mu$ are preferred, since they indicate that the dictionary elements do not resemble each other. This helps us to effectively tell dictionary elements apart (Donoho and Huo, 2001; Candès and Romberg, 2007). We assume that $\mu = O(\log(n))$ (Donoho and Huo, 2001). Next, we assume that the coefficients are drawn from a distribution class $\mathcal{D}$ defined as follows.

**Definition 4** (Distribution class $\mathcal{D}$). The coefficient vector $\mathbf{x}^*$ belongs to an unknown distribution $\mathcal{D}$, where the support $S = \text{supp}(\mathbf{x}^*)$ is at most of size $k$, $\Pr[i \in S] = \Theta(k/m)$ and $\Pr[i, j \in S] = \Theta(k^2/m^2)$. Moreover, the distribution is normalized such that $E[|x_i^*|^2] = 0$ and $E[|x_i^*|^2] = 1$, and when $i \in S$, $|x_i^*| \geq C$ for some constant $C \leq 1$. In addition, the non-zero entries are pairwise independent conditioned on the support.

The randomness of the coefficient is necessary for our finite sample analysis of the convergence. Here, there are two sources of randomness. The first is the randomness of the support, where the non-zero elements are assumed to pair-wise independent. The second is the value an element in the support takes, which is assumed to be zero mean with variance one, and bounded in magnitude.

Similar conditions are also required for support recovery of sparse coefficients, even when the dictionary is known (Wainwright, 2009; Yuan et al., 2016). Note that, although we only consider the case $|x_i^*| \geq C$ for ease of discussion, analogous results may hold more generally for $x_i$'s drawn from a distribution with sufficiently (exponentially) small probability of taking values in $[-C, C]$.

Recall that, given the coefficients, we recover the dictionary by making progress on the least squares objective (P1) (ignoring the term penalizing $S(\cdot)$). Note that, our algorithm is based on finding an appropriate direction to ensure descent than the geometry of the objective. This results in an appropriate direction to ensure descent and guarantees the same geometry of the objective. To this end, we adopt a gradient descent-based strategy for dictionary update. However, since the coefficients are not exactly known, this results in an approximate gradient descent-based approach, where the empirical gradient estimate is formed as (5). In our analysis, we establish the conditions under which both the empirical gradient vector (corresponding to each dictionary element) and the gradient matrix concentrate around their means. To ensure progress at each iterate $t$, we show that the expected gradient vector is $\Omega(k/m), \Omega(m/k), 0$-correlated with the descent direction, and the gradient matrix concentrate around their means. To ensure progress at each iterate $t$, we show that the expected gradient vector is $\Omega(k/m), \Omega(m/k), 0$-correlated with the descent direction, defined as follows.

**Definition 5.** A vector $g^{(t)}$ is $(\rho_-, \rho_+, \zeta_t)$-correlated with a vector $z^*$ if $\langle g^{(t)}, z^{(t)} - z^* \rangle \geq \rho_- ||z^{(t)} - z^*||^2 + \rho_+ ||g^{(t)}||^2 - \zeta_t$.

This can be viewed as a local descent condition which leads to the true dictionary columns; see also (Candès et al., 2015), (Chen and Wainwright, 2015) and (Arora et al., 2015). In convex optimization literature, this condition is implied by the $2\rho$-strong convexity, and $1/2\rho$-smoothness of the objective. We show that for NOODL, $\zeta_t = 0$, which facilitates linear convergence to $A^*$ without incurring any bias. Overall our specific model assumptions for the analysis can be formalized as:

- **A1.** $A^*$ is $\mu$-incoherent (Def. 3), where $\mu = O(\log(n))$, $\|A^*\| = O(\sqrt{m/n})$ and $m = O(n)$;
- **A2.** The coefficients are drawn from the distribution class $\mathcal{D}$, as per Def. 4;
- **A3.** The sparsity $k$ satisfies $k = O(\sqrt{n}/\mu \log(n))$;
- **A4.** $A^{(0)}$ is $(\epsilon_0, 2)$-close to $A^*$ (Def. 1), and $\epsilon_0$ and each subsequent $\epsilon_t$ are $\epsilon_t, \epsilon_0 = O^*(1/\log(n))$;
- **A5.** The step-size for dictionary update satisfies $\eta_A = \Theta(m/k)$;
- **A6.** The step-size and threshold for coefficient estimation satisfies $\eta_k^{(r)} < c_1(\epsilon_t, \mu, n, k) < 1$ and $\tau^{(r)} = c_2(k)$ for small constants $c_1$ and $c_2$.

We are now ready to state our main result. A summary of the notation followed by a details of the analysis is provided in Appendix A and Appendix B, respectively.
**Theorem 1** (Main Result). Suppose that assumptions A.1-A.6 hold, and Algorithm 1 is provided with \( p = \Omega(mk^2) \) new samples generated according to model (1) at each iteration \( t \). Then for some \( 0 < \omega < 1/2 \), the estimate \( \hat{A}^{(t)} \) at \( t \)-th iteration satisfies
\[
\|A^{(t)} - A_*\|^2 \leq (1 - \omega)^t \|A^{(0)} - A_*\|^2, \text{ for all } t = 1, 2, \ldots.
\]
Furthermore, given \( R = \log(n) \), with probability at least \( (1 - \delta_{alg}) \) for some small constant \( \delta_{alg} \), the coefficient estimate \( \hat{x}^{(t)} \) at \( t \)-th iteration has the correct signed-support and satisfies
\[
(\hat{x}^{(t)} - x_*)^2 = O(k(1 - \omega)^t/2 \|A^{(0)} - A_*\|), \text{ for all } i \in \text{supp}(x_*).
\]
Our main result establishes that when the model satisfies A.1~A.3, the errors corresponding to the dictionary and coefficients geometrically decrease to the true model parameters, given appropriate dictionary initialization and learning parameters (step sizes and threshold); see A.4~A.6. In other words, to attain a target tolerance of \( \epsilon_T \) where \( \|A^{(T)} - A_*\| \leq \epsilon_T \), we require \( T = \Omega(1/\epsilon_T) \). Moreover, given \( R = \Omega(1/\delta_R) \) IHT iterations (for a target decay tolerance parameter \( \delta_R \) which removes the effect of the initial \( x^{(0)} \) and is defined as \( \delta_R \geq (1 - \eta_T)^R \)), the error in the coefficients only depends on the error in the dictionary and drops geometrically at every iteration \( t \). Also, note that NOODL can tolerate i.i.d. noise, as long as the noise variance is controlled to enable the concentration results to hold; we consider the noiseless case here for simplicity.

The introduction of IHT step adds complexity in the analysis of both the dictionary and coefficients. The primary condition that allows us to make progress on both factors is the signed-support recovery (Def. 2). To analyze the coefficients, in addition to deriving conditions on the parameters to preserve the correct signed-support, we analyze the recursive update step and decompose the noise term into a component that depends on the error in the dictionary, and the other that depends on the initial coefficient estimate. Further, for the dictionary update, we analyze the interactions between elements of the coefficient vector introduces by the IHT-based update step and show that the gradient vector for the dictionary update is \( \Omega(k/m), \Omega(m/k), 0 \)-correlated with the descent direction. In the end, this leads to exact recovery of the coefficients and removal of bias in the dictionary estimation. Note that our analysis pipeline is standard for the convergence analysis for iterative algorithms. However, the introduction of the IHT-based strategy for coefficient update makes the analysis highly involved as compared to existing results, e.g., the simple HT-based coefficient estimate in Arora et al. (2015).

NOODL has an overall running time of \( \mathcal{O}(mnp \log(1/\delta_R) \log(1/\epsilon_T)) \) to achieve \( \|A^{(T)} - A_*\| \leq \epsilon_T, \forall i \in [m] \), with a total sample complexity of \( p \cdot T = \Omega(mk^2) \). Thus, the IHT-based update in NOODL introduces a factor of \( \log(1/\delta_R) \) in the computational complexity as compared to Arora et al. (2015) (has a total sample complexity of \( p \cdot T = \Omega(mk) \)) in order to remove the bias, and does not have the exponential running time and sample complexity as Barak et al. (2015); see Table 1.

**4 NEURAL IMPLEMENTATION OF NOODL**

The neural plausibility of our algorithm implies that it can be implemented as a neural network. This is because, NOODL employs simple linear and non-linear operations (such as inner-product and hard-thresholding) and the coefficient updates are separable across data samples, as shown in (4) of Algorithm 1. To this end, we present a neural implementation of the our algorithm in Fig. 1, which showcases the applicability of NOODL in large-scale distributed learning tasks, motivated from the implementations described in (Olshausen and Field, 1997) and Arora et al., 2015).

The neural architecture shown in Fig. 1(a) has three layers – input layer, weighted residual evaluation layer, and the output layer. The input to the network is a data and step-size pair \( (y^{(j)}, \eta_r) \) to each input node. Given an input, the second layer evaluates the weighted residuals as \( \eta_r(y^{(j)}) - A^{(t)}x^{(r+1)}_j \).

Finally, the output layer neurons evaluate the IHT iterates \( x^{(r+1)}_j \) (4). We illustrate the operation of this architecture using the timing diagram in Fig. 1(b). The main stages of operation are as follows.

**Initial Hard Thresholding Phase**: The coefficients initialized to zero, and an input \( (y^{(j)}, 1) \) is provided to the input layer at a time instant \( \ell = 0 \), which communicates these to the second layer. Therefore, the residual at the output of the weighted residual evaluation layer evaluates to \( y^{(j)} \) at \( \ell = 1 \). Next, at \( \ell = 2 \), this residual is communicated to the output layer, which results in evaluation
Hebbian Learning in the timing sequence. In this phase, each output layer neuron communicates the iterates $x_{(j)}^{(r+1)}$ to the second layer for evaluation of subsequent iterates as shown in Fig. 1(b). The process then continues till the time instance $\ell = 2R + 1$, for $R = \Omega(\log(1/\delta_R))$ to generate the final coefficient estimate $\hat{x}_{(j)}^{(T)} := x_{(j)}^{(R)}$ for the current batch of data. At this time, the input layer is again injected with $(y_{(j)}, 1)$ to prepare the network for residual sharing and gradient evaluation for dictionary update.

**Dictionary Update Phase:** The procedure now enters the dictionary update phase, denoted as “Hebbian Learning” in the timing sequence. In this phase, each output layer neuron communicates the final coefficient estimate $\hat{x}_{(j)}^{(i)} = x_{(j)}^{(R)}$ to the second layer, which evaluates the residual for one last time (with $\eta_x = 1$), and shares it across all second layer neurons (“Hebbian learning”). This allows each second layer neuron to evaluate the empirical gradient estimate (5), which is used to update the current dictionary estimate (stored as weights) via a approximate gradient descent step.

This completes one outer iteration of Algorithm 1, and the process continues for $T = \Omega(\log(1/\epsilon_T))$ to achieve $\|A_i^{(T)} - A^\ast\| \leq \epsilon_T, \forall i \in [m]$, with each step receiving a new mini-batch of data.

5 **Experiments**

In this section, we analyze the convergence properties and sample complexity of NOODL via experimental evaluations. Additional results are shown in Appendix E.

5.1 **Convergence Analysis**

We compare the performance of our algorithm NOODL with the current state-of-the-art alternating optimization-based algorithms presented in Arora et al. (2015). First of these, Arora15 (‘’biased’’), is a simple neurally plausible method which incurs a bias and has a sample complexity of $\Omega(mn)$. The other one, referred to as Arora15 (‘’unbiased’’), incurs no bias as per Arora et al. (2015), but the sample complexity results were not established.

**Data Generation:** We generate a $1000 \times 1500$ matrix, whose entries are drawn from the standard normal distribution. We then normalize the columns of this matrix to form the ground-truth dictionary $A^\ast$; i.e., here $n = 1000$ and $m = 1500$. Next, we perturb this dictionary with random Gaussian noise, such that the unit-norm columns of the resulting matrix, $A^{(0)}$ are $2/\log(n)$ away from $A^\ast$, in $\ell_2$-norm sense, i.e., $\|A_i^{(0)} - A^\ast\| \leq 2/\log(n)$; this satisfies the initialization assumptions in A.4. At each iteration, we generate $p = 5000$ samples by multiplying the dictionary $A^\ast$ with a coefficient matrix $X^\ast \in \mathbb{R}^{m \times k}$ with at most $k = 10, 20, 50$, and 100, entries per column, drawn from the Radamacher distribution. We report the results in terms of relative Frobenius error for all the ex-
Figure 2: Comparative analysis of convergence properties. Panels (a-i), (b-i), (c-i), and (d-i) show the convergence of NOODL, Arora15('biased'), and Arora15('unbiased') for different sparsity levels for $n = 1000$, $m = 1500$ and $p = 5000$. Since NOODL also recovers the coefficients, we show the corresponding recovery of the dictionary, coefficients, and overall fit in panels (a-ii), (b-ii), (c-ii), and (d-ii), respectively. Further, panels (e-i) and (e-ii) show the phase transition in samples $p$ (per iteration) with the size of the dictionary $m$ averaged across 10 Monte Carlo simulations for the two factors. Here, $n = 100$, $k = 3$, $\eta_k = 0.2$, $\tau = 0.1$, $\epsilon_0 = 2/\log(n)$, $\eta_A$ is chosen as per A.5. A trial is considered successful if the relative Frobenius error incurred by $\hat{A}$ and $\hat{X}$ is below $5 \times 10^{-7}$ after 50 iterations.

Experiments, i.e., for a recovered matrix $\hat{M}$, we report $\|\hat{M} - M^*\|_F/\|M^*\|_F$. Note that, although our phase transition analysis for NOODL shows that a batch size of $p = m$ suffices, we use $p = 5000$ in our convergence analysis for a fair comparison with related techniques.

Discussion: Fig. 2 panels (a-i), (b-i), (c-i), and (d-i) show the performance of the aforementioned methods for $k = 10, 20, 50$, and 100, respectively. Here, for all experiments we set $\eta_k = 0.2$ and $\tau = 0.1$. We terminate NOODL when the error in dictionary is less than $10^{-10}$, and otherwise when they converge. Also, for coefficient update, we terminate when change in the iterates is below $10^{-12}$. For $k = 10, 20$ and $k = 50$, we note that Arora15('biased') and Arora15('unbiased') incur significant bias, while NOODL converges to $A^*$ linearly. NOODL also converges for significantly higher choices of sparsity $k$, i.e., for $k = 100$ as shown in panel (d), beyond $k = \mathcal{O}(\sqrt{n})$, indicating a potential for improving this bound. Also, in panels (a-ii), (b-ii), (c-ii) and (d-ii) we show the corresponding performance of NOODL in terms of the error in the overall fit ($\|\hat{Y} - \hat{A}\hat{X}\|_F/\|Y\|_F$), and the error in the coefficients and the dictionary, in terms of relative Frobenius error metric discussed above. We observe that the error in dictionary and coefficients drops linearly as indicated by our main result.

5.2 Phase transitions

Fig. 2 panels (e-i) and (e-ii), shows the phase transition in number of samples with respect to the size of the dictionary $m$. We observe a sharp phase transition at $p/m = 1$ for the dictionary, and at $p/m = 0.75$ for the coefficients. This phenomenon is similar one observed by Agarwal et al. (2014) (however, theoretically they required $p = \mathcal{O}(m^2)$). Here, we confirm number of samples required by NOODL are linearly dependent on the dictionary elements $m$.

6 Conclusions

We present NOODL, to the best of our knowledge, the first provable online algorithm for exact recovery of both factors of the dictionary learning (DL) model. Here, we pose the DL problem as a factorization task, and develop an algorithm that alternates between: (a) an iterative hard thresholding (IHT)-based step for coefficient recovery, and (b) a gradient descent-based update for the dictionary, resulting in a simple and scalable algorithm, suitable for large-scale distributed implementations. Further, we show that once initialized appropriately, the sequence of estimates produced by NOODL converge linearly to the true dictionary and coefficients without incurring any bias in the estimation. Complementary to our theoretical and numerical results, we also design an implementation of NOODL in a neural architecture for use in practical applications. Finally, the analysis of this inherently non-convex problem impacts matrix and tensor factorization tasks arising in signal processing, collaborative filtering, and machine learning.
REFERENCES


A SUMMARY OF NOTATION

The following table summarizes the definitions of some frequently used symbols in our analysis.

Table 2: Frequently used symbols

<table>
<thead>
<tr>
<th>Dictionary Related</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Lambda^{(t)}_{ij} )</td>
<td>( \text{i-th column of the dictionary estimate at the } t\text{-th iterate.} )</td>
<td></td>
</tr>
<tr>
<td>( \epsilon_t )</td>
<td>( | \Lambda^{(t)}<em>{\beta} - \Lambda^*</em>{\beta} | \leq \epsilon_t = \mathcal{O}(\frac{1}{\sqrt{\log(n)}}) )</td>
<td>Upper-bound on column-wise error at the ( t)-th iterate.</td>
</tr>
<tr>
<td>( \mu_t )</td>
<td>( \frac{\mu_t}{\sqrt{n}} = \frac{1}{\sqrt{n}} + 2\epsilon_t )</td>
<td>Incoherence between the columns of ( \Lambda^{(t)} ); See Claim 1.</td>
</tr>
<tr>
<td>( \lambda^{(t)}_{ij} )</td>
<td>( \lambda^{(t)}_{ij} :=</td>
<td>(\Lambda^{(t)}<em>{ij} - \Lambda^*</em>{ij}, \Lambda^*_{ij})</td>
</tr>
<tr>
<td>( \Lambda^{(t)}_{S}(i,j) )</td>
<td>( \Lambda^{(t)}<em>{S}(i,j) = \begin{cases} \lambda^{(t)}</em>{ij}, &amp; \text{for } j = i, i \in S \ 0, &amp; \text{otherwise.} \end{cases} )</td>
<td>A diagonal matrix of size (</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient Related</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x^{(t)} )</td>
<td>( \text{i-th element the coefficient estimate at the } t\text{-th IHT iterate.} )</td>
<td></td>
</tr>
<tr>
<td>( C )</td>
<td>(</td>
<td>x^{(t)}_r</td>
</tr>
<tr>
<td>( S )</td>
<td>( S := \text{supp}(x^*) ) where (</td>
<td>S</td>
</tr>
<tr>
<td>( \delta_R )</td>
<td>( \delta_R := (1 - \eta_x + \frac{\eta_x \tau}{\sqrt{m}})^2 \geq (1 - \eta_x)^2 )</td>
<td>Decay parameter for coefficients.</td>
</tr>
<tr>
<td>( \delta^{(t)}_{C_i} )</td>
<td>( \delta^{(t)}_{C_i} :=</td>
<td>x^*_i - x^{(t)}_i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probabilities</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_t )</td>
<td>( q_t = \text{Pr}[i \in S] = \Theta(\frac{1}{\sqrt{m}}) )</td>
<td></td>
</tr>
<tr>
<td>( p_t )</td>
<td>( p_t = \mathbb{E}[x^{(t)}_i \text{sign}(x^*_i)</td>
<td>x^*_i \neq 0] )</td>
</tr>
<tr>
<td>( \delta^{(t)}_R )</td>
<td>( \delta^{(t)}_R = 2k \exp(-1/\mathcal{O}(\epsilon_t)) )</td>
<td></td>
</tr>
<tr>
<td>( \delta^{(t)}_{\delta^{(t)}_R} )</td>
<td>( \delta^{(t)}_{\delta^{(t)}_R} = \exp(-\Omega(k)) )</td>
<td></td>
</tr>
<tr>
<td>( \delta^{(t)}_{\delta^{(t)}_R} )</td>
<td>( \delta^{(t)}_{\delta^{(t)}_R} = (n + m) \exp(-\Omega(m \sqrt{\log(n)}) )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other terms</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \xi^{(t)}_j )</td>
<td>( \xi^{(t)}<em>j := \sum</em>{i \neq j}</td>
<td>\langle \Lambda^{(t)}_j - \Lambda^<em>_j, \Lambda^</em>_i \rangle + \langle \Lambda^*_j, \Lambda^{(t)}_i \rangle</td>
</tr>
<tr>
<td>( \beta^{(t)}_j )</td>
<td>( \beta^{(t)}<em>j := \sum</em>{i \neq j}</td>
<td>\langle \Lambda^<em>_j, \Lambda^</em>_i - \Lambda^{(t)}_j \rangle + \langle \Lambda^*_j - \Lambda^{(t)}_j, \Lambda^{(t)}_i \rangle</td>
</tr>
<tr>
<td>( \tau_j )</td>
<td>( \tau_j = \mathcal{O}(\sqrt{k} \epsilon_t) ) is an upper-bound on ( \beta^{(t)}_j ) with probability at least ( (1 - \delta^{(t)}_R) )</td>
<td></td>
</tr>
<tr>
<td>( \gamma^{(t)}_j )</td>
<td>( \gamma^{(t)}_j := \mathbb{E}[(\Lambda^{(t)}_j - \Lambda^<em>_j) \text{sign}(x^</em>_j)], ) See ( \dagger ) below.</td>
<td></td>
</tr>
<tr>
<td>( \hat{x}_j )</td>
<td>( \hat{x}_j := x^{(t)}_j - \gamma^{(t)}_j )</td>
<td></td>
</tr>
</tbody>
</table>

\( \dagger \) \( \mathbb{1}_{x^*} \) is the indicator function corresponding to the event that \( \text{sign}(x^*) = \text{sign}(\hat{x}) \), denoted by \( \mathcal{F}_{\hat{x}} \), and similarly for the complement \( \overline{\mathcal{F}_{\hat{x}}} \).
B PROOF OF THEOREM 1

We now prove our main result. The detailed proofs of intermediate lemmas and claims are organized in Appendix C and Appendix D, respectively. Furthermore, the standard concentration results are stated in Appendix F for completeness. Also, see Table 3 for a map of dependence between the results.

OVERVIEW

Given a \((\epsilon_0, 2)\)-close estimate of the dictionary, the main property that allows us to make progress on the dictionary is the recovery of the correct sign and support of the coefficients. Therefore, we first show that the initial coefficient estimate (3) recovers the correct signed-support in Step I.A. Now, the IHT-based coefficient update step also needs to preserve the correct signed-support. This is to ensure that the approximate gradient descent-based update for the dictionary makes progress. Therefore, in Step I.B, we derive the conditions under which the signed-support recovery condition is preserved by the IHT update.

To get a handle on the coefficients, in Step II.A, we derive an upper-bound on the error incurred by each non-zero element of the estimated coefficient vector, i.e., \(|\tilde{x}_i - x^*_i|\) for \(i \in S\) for a general coefficient vector \(x^*\), and show that this error only depends on \(\epsilon_t\) (the column-wise error in the dictionary) given enough IHT iterations \(R\) as per the chosen decay parameter \(\delta_R\). In addition, for analysis of the dictionary update, we develop an expression for the estimated coefficient vector in Step II.B.

We then use the coefficient estimate to show that the gradient vector satisfies the local descent condition (Def. 5). This ensures that the gradient makes progress after taking the gradient descent-based step (6). To begin, we first develop an expression for the expected gradient vector (corresponding to each dictionary element) in Step III.A. Here, we use the closeness property Def 1 of the dictionary. As a result, our main result establishes the conditions under which any coefficient estimate \(x^{(t)}\) at t-th iteration has the correct signed-support and satisfies \((\bar{x}_i^{(t)} - x_i^*)^2 \leq O(k(1 - \omega)^t/2\|A_i^{(0)} - A_i^*\|^2),\) for all \(i = 1, 2, \ldots\).

Next in Step IV.A and Step IV.B, we show that the updated dictionary estimate maintains the closeness property Def 1. This sets the stage for the next dictionary update iteration. As a result, our main result establishes the conditions under which any t-th iteration succeeds.

Our main result is as follows.

**Theorem 1** (Main Result) Suppose that assumptions A.1-A.6 hold, and Algorithm 1 is provided with \(p = \Omega(mk^2)\) new samples generated according to model (1) at each iteration \(t\). Then for some \(0 < \omega < 1/2\), the estimate \(A^{(t)}\) at (t)-th iteration satisfies
\[\|A^{(t)} - A^*_i\|^2 \leq (1 - \omega)^t\|A^{(0)} - A^*_i\|^2,\] for all \(t = 1, 2, \ldots\). Furthermore, given \(R = \log(n)\), with probability at least \((1 - \delta_{alg})\) for some small constant \(\delta_{alg}\), the coefficient estimate \(\tilde{x}_i^{(t)}\) at t-th iteration has the correct signed-support and satisfies
\[\tilde{x}_i^{(t)} = O(k(1 - \omega)^{t/2}\|A^{(0)} - A^*_i\|),\] for all \(i = \supp(x^*)\). Here, \(\delta_{alg} = \frac{\delta^{(t)}}{2} + \delta_{HW} + \delta_{g_i} + \delta^{(t)},\) \(\delta^{(t)} = 2m \exp(-C^2/O^*(\epsilon_t^2)),\) \(\delta^{(t)} = 2k \exp(-1/O^*(\epsilon_t)),\) \(\delta_{HW} = \exp(-1/O^*(\epsilon_t)),\) \(\delta_{g_i} = \exp(-\Omega(k))\), and \(\delta^{(t)} = (n + m) \exp(-\Omega(m\sqrt{\log(n)}))\).

**STEP I: COEFFICIENTS HAVE THE CORRECT SIGNED-SUPPORT**

As a first step, we ensure that our coefficient estimate has the correct signed-support (Def. 2). To this end, we first show that the initialization has the correct signed-support, and then show that the iterative hard-thresholding (IHT)-based update step preserves the correct signed-support for a suitable choice of parameters.

- **Step I.A: Showing that the initial coefficient estimate has the correct signed-support**—Given a \((\epsilon_0, 2)\)-close estimate \(A^{(0)}\) of \(A^*\), we first show that for a general sample \(y\) the
initialization step (3) recovers the correct signed-support with probability at least \((1 - \delta_T^{(t)})\), where \(\delta_T^{(t)} = 2m \exp(-\frac{c^2}{\sigma^2 T^2})\). This is encapsulated by the following lemma.

**Lemma 1 (Signed-support recovery by coefficient initialization step).** Suppose \(A^{(t)}\) is \(\epsilon_t\)-close to \(A^*\). Then, if \(\mu = \mathcal{O}(\log(n))\), \(k = \sqrt{n}/\mu \log(n)\), and \(\epsilon_t = \mathcal{O}^*(1/\sqrt{\log(m)})\), with probability at least \((1 - \delta_T^{(t)})\) for each random sample \(y = A^* x^*\):

\[
\text{sign}(T_{C/2}/((A^{(t)})^T y)) = \text{sign}(x^*),
\]

where \(\delta_T^{(t)} = 2m \exp(-\frac{c^2}{\sigma^2 T^2})\).

Note that this result only requires the dictionary to be column-wise close to the true dictionary, and works for less stringent conditions on the initial dictionary estimate, i.e., requires \(\epsilon_t = \mathcal{O}^*(1/\sqrt{\log(m)})\) instead of \(\epsilon_t = \mathcal{O}^*(1/\log(m))\); see also (Arora et al., 2015).

- **Step II.B: The iterative IHT-type updates preserve the correct signed support**— Next, we show that the IHT-type coefficient update step (4) preserves the correct signed-support for an appropriate choice of step-size parameter \(\eta_x^{(r)}\) and threshold \(\tau^{(r)}\). The choice of these parameters arises from the analysis of the IHT-based update step. Specifically, we show that at each iterate \(r\), the step-size \(\eta_x^{(r)}\) should be chosen to ensure that the component corresponding to the true coefficient value is greater than the “interference” introduced by other non-zero coefficient elements. Then, if the threshold is chosen to reject this “noise”, each iteration of the IHT-based update step preserves the correct signed-support.

**Lemma 2 (IHT update step preserves the correct signed-support).** Suppose \(A^{(t)}\) is \(\epsilon_t\)-close to \(A^*\), \(\mu = \mathcal{O}(\log(n))\), \(k = \sqrt{n}/\mu \log(n)\), and \(\epsilon_t = \mathcal{O}^*(1/\log(m))\). Then, with probability at least \((1 - \delta_T^{(t)} - \delta_T^{(n)})\), each iterate of the IHT-based coefficient update step shown in (4) has the correct signed-support, if for a constant \(c_1^{(r)}(\epsilon_t, \mu, k, n)\), the step size is chosen as \(\eta_x^{(r)} \leq c_1^{(r)}\), and the threshold \(\tau^{(r)}\) is chosen as

\[
\tau^{(r)} = \eta_x^{(r)}(t_\beta + \frac{\mu}{\sqrt{n}}\|x^{(r-1)} - x^*\|_1) := c_2^{(r)}(k),
\]

for some constants \(c_1\) and \(c_2\). Here, \(t_\beta = \mathcal{O}(\sqrt{k\epsilon_t})\), \(\delta_T^{(t)} = 2m \exp(-\frac{c^2}{\sigma^2 T^2})\), and \(\delta_T^{(n)} = 2k \exp(-\frac{1}{\sigma^2})\).

Note that, although we have a dependence on the iterate \(r\) in choice of \(\eta_x^{(r)}\) and \(\tau^{(r)}\), these can be set to some constants independent of \(r\). In practice, this dependence allows for greater flexibility in the choice of these parameters.

**Step II: Analyzing the Coefficient Estimate**

We now derive an upper-bound on the error incurred by each non-zero coefficient element. Further, we derive an expression for the coefficient estimate at the \(t\)-th round of the online algorithm \(\hat{x}^{(t)} := x^{(t)}\); we use \(\hat{x}\) instead of \(\hat{x}^{(t)}\) for simplicity.

- **Step II.A: Derive a bound on the error incurred by the coefficient estimate**— Since Lemma 2 ensures that \(\hat{x}\) has the correct signed-support, we now focus on the error incurred by each coefficient element on the support by analyzing \(\hat{x}\). To this end, we carefully analyze the effect of the recursive update (4), to decompose the error incurred by each element on the support into two components – one that depends on the initial coefficient estimate \(x^{(0)}\) and other that depends on the error in the dictionary.

We show that the effect of the component that depends on the initial coefficient estimate diminishes by a factor of \((1 - \eta_x + \eta_x \frac{r_{\text{end}}}{t_{\text{end}}})\) at each iteration \(r\). Therefore, for a decay parameter \(\delta_T\), we can choose the number of IHT iterations \(R\), to make this component arbitrarily small. Therefore, the error in the coefficients only depends on the per column error in the dictionary, formalized by the following result.
Lemma 3 (Upper-bound on the error in coefficient estimation). With probability at least \((1 - \delta(t) - \delta_T(t))\) the error incurred by each element \(i_1 \in \text{supp}(x^*)\) of the coefficient estimate is upper-bounded as

\[
|\hat{x}_{i_1} - x_{i_1}^*| \leq O(t_\beta) + \left((R + 1)k\eta_i \frac{\rho}{\sqrt{n}} \max_i |x_i^{(0)} - x_i^*| + |x_i^{(0)} - x_i^*|\right) \delta_R = O(t_\beta)
\]

where \(t_\beta = O(\sqrt{k}\epsilon_t)\), \(\delta_R := (1 - \eta_\epsilon + \eta_\epsilon \frac{\rho}{\sqrt{n}})R\), \(\delta_T(t) = 2m \exp(-\frac{C^2}{C_T})\), \(\delta_\beta = 2k \exp(-\frac{1}{C_T})\), and \(\mu_i\) is the incoherence between the columns of \(A(t)\); see Claim 1.

This result allows us to show that if the column-wise error in the dictionary decreases at each iteration \(t\), then the corresponding estimates of the coefficients also improve.

- Step II.B: Developing an expression for the coefficient estimate– Next, we derive the expression for the coefficient estimate in the following lemma. This expression is used to analyze the dictionary update.

Lemma 4 (Expression for the coefficient estimate at the end of \(R\)-th IHT iteration). With probability at least \((1 - \delta_T(t) - \delta_\beta)\) the \(i_1\)-th element of the coefficient estimate, for each \(i_1 \in \text{supp}(x^*)\), is given by

\[
\hat{x}_{i_1} := x_{i_1}^{(R)} = x_{1_1}^* (1 - \lambda_{i_1}^{(t)}) + \vartheta_{i_1}^{(R)}.
\]

Here, \(\vartheta_{i_1}^{(R)} = |\vartheta_{i_1}^{(R)}| = O(t_\beta)\), where \(t_\beta = O(\sqrt{k}\epsilon_t)\). Further, \(\lambda_{i_1}^{(t)} = \|A_{i_1}^{(t)} - A_{i_1}^*, A_{i_1}^*_1\| \leq \frac{\epsilon_t\sqrt{T}}{2}, \delta_T(t) = 2m \exp(-\frac{C^2}{C_T})\) and \(\delta_\beta = 2k \exp(-\frac{1}{C_T})\).

We again observe that the error in the coefficient estimate depends on the error in the dictionary via \(\lambda_{i_1}^{(t)}\) and \(\vartheta_{i_1}^{(R)}\).

**STEP III: ANALYZING THE GRADIENT FOR DICTIONARY UPDATE**

Given the coefficient estimate we now show that the choice of the gradient as shown in (5) makes progress at each step. To this end, we analyze the gradient vector corresponding to each dictionary element to see if it satisfies the local descent condition of Def. 5. Our analysis of the gradient is motivated from Arora et al. (2015). However, as opposed to the simple HT-based coefficient update step used by Arora et al. (2015), our IHT-based coefficient estimate adds significant overhead in terms of analysis. Notwithstanding the complexity of the analysis, we show that this allows us to remove the bias in the gradient estimate.

To this end, we first develop an expression for each expected gradient vector, show that the empirical gradient estimate converges to its mean, and finally show that the empirical gradient vector is \((\Omega(k/m), \Omega(m/k), 0)\)-correlated with the descent direction, i.e. has no bias.

- **Step III.A: Develop an expression for the expected gradient vector corresponding to each dictionary element**—The expression for the expected gradient vector \(g_j^{(t)}\) corresponding to \(j\)-th dictionary element is given by the following lemma.

Lemma 5 (Expression for the expected gradient vector). Suppose that \(A^{(t)}\) is \((\epsilon_t, 2)\)-near to \(A^*\). Then, the dictionary update step in Algorithm 1 amounts to the following for the \(j\)-th dictionary element

\[
E[A_j^{(t+1)}] = A_j^{(t)} + \eta_A g_j^{(t)},
\]

where \(g_j^{(t)}\) is given by

\[
g_j^{(t)} = q_j p_j ((1 - \lambda_j^{(t)}) A_j^{(t)} - A^* + \frac{1}{q_j p_j} \Delta_j^{(t)} \pm \gamma),
\]

\[
\lambda_j^{(t)} = \|A_j^{(t)} - A^*, A^*_j\|, \text{ and } \Delta_j^{(t)} := E[A_j^{(t)} \vartheta_j^{(R)} \text{ sign}(x_j^*)], \text{ where } \|\Delta_j^{(t)}\| = O(\sqrt{m_{q_j} p_j q_j \epsilon_t \|A(t)\|}).
\]

- **Step III.B: Show that the empirical gradient vector concentrates around its expectation**—Since we only have access to the empirical gradient vectors, we show that these concentrate around their expected value via the following lemma.
Lemma 6 (Concentration of the empirical gradient vector). Given $p = \tilde{\Omega}(mk^2)$ samples, the empirical gradient vector estimate corresponding to the $i$-th dictionary element, $\hat{g}_i^{(t)}$, concentrates around its expectation, i.e.,

$$\|\hat{g}_i^{(t)} - g_i^{(t)}\| \leq o\left(\frac{k}{m}\epsilon_t\right).$$

with probability at least $(1 - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)})$, where $\delta_i^{(t)} = \exp(-\Omega(k))$.

- Step III.C: Show that the empirical gradient vector is correlated with the descent direction. Next, in the following lemma we show that the empirical gradient vector $\hat{g}_j^{(t)}$ is correlated with the descent direction. This is the main result which enables the progress in the dictionary (and coefficients) at each iteration $t$.

Lemma 7 (Empirical gradient vector is correlated with the descent direction). Suppose $A^{(t)}$ is $(\epsilon_t, 2)$-near to $A^*$, $k = O(\sqrt{m})$ and $\eta_A = O(m/k)$. Then, with probability at least $(1 - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)})$ the empirical gradient vector $\hat{g}_j^{(t)}$ is $(\Omega(k/m), \Omega(m/k), 0)$-correlated with $(A_j^{(t)} - A_j^*)$, and for any $t \in [T]$,

$$\|A_j^{(t)} - A_j^*\|^2 \leq (1 - \rho, \eta_A)\|A_j^{(t)} - A_j^*\|^2.$$

This result ensures for at any $t \in [T]$, the gradient descent-based updates made via (5) gets the columns of the dictionary estimate closer to the true dictionary, i.e., $\epsilon_{t+1} \leq \epsilon_t$. Moreover, this step requires closeness between the dictionary estimate $A^{(t)}$ and $A^*$, in the spectral norm-sense, as per Def 1.

**Step IV: Show that the dictionary maintains the closeness property**

As discussed above, the closeness property (Def 1) is crucial to show that the gradient vector is correlated with the descent direction. Therefore, we now ensure that the updated dictionary $A^{(t+1)}$ maintains this closeness property. Lemma 7 already ensures that $\epsilon_{t+1} \leq \epsilon_t$. As a result, we show that $A^{(t+1)}$ maintains closeness in the spectral norm-sense as required by our algorithm, i.e., that it is still $(\epsilon_{t+1}, 2)$-close to the true dictionary. Also, since we use the gradient matrix in this analysis, we show that the empirical gradient matrix concentrates around its mean.

- Step IV.A: The empirical gradient matrix concentrates around its expectation. We first show that the empirical gradient matrix concentrates as formalized by the following lemma.

Lemma 8 (Concentration of the empirical gradient matrix). With probability at least $(1 - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)})$, $\|g^{(t)}(t) - g^{(t)}\|$, is upper-bounded by $O^*(\frac{k}{m}\|A^*\|)$, where $\delta_i^{(t)} = (n + m)\exp(-\Omega(m\sqrt{\log(m)})$.

- Step IV.B: The “closeness” property is maintained after the updates made using the empirical gradient estimate: Next, the following lemma shows that the updated dictionary $A^{(t+1)}$ maintains the closeness property.

Lemma 9 ($A^{(t+1)}$ maintains closeness). Suppose $A^{(t)}$ is $(\epsilon_t, 2)$ near to $A^*$ with $\epsilon_t = O^*(1/\log(n))$, and number of samples used in step $t$ is $p = \tilde{\Omega}(mk^2)$, then with probability at least $(1 - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)})$, $A^{(t+1)}$ satisfies $\|A^{(t+1)} - A^*\| \leq 2\|A^*\|$.

**Step V: Combine results to show the main result**

Proof of Theorem 1. From Lemma 7 we have that with probability at least $(1 - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)} - \delta_i^{(t)})$, $g_j^{(t)}$ is $(\Omega(k/m), \Omega(m/k), 0)$-correlated with $A_j^*$. Further, Lemma 9 ensures that each iterate maintains the closeness property. Now, applying Lemma 15 we have that, $\eta_A \leq \Theta(m/k)$, with probability at least $1 - \delta_i^{(t)}$ any $t \in [T]$ satisfies

$$\|A_j^{(t)} - A_j^*\|^2 \leq (1 - \omega)\|A_j^{(0)} - A_j^*\|^2 \leq (1 - \omega)\epsilon_0^2,$$

where for $0 < \omega < 1/2$ with $\omega = \Omega(k/m)\eta_A$. That is, the updates converge geometrically to $A^*$. Further, from Lemma 3, we have that the result on the error incurred by the coefficients. Here,
\[ \delta_{\text{alg}}^{(t)} = \delta_T^{(t)} + \delta_R^{(t)} + \delta_{HW}^{(t)} + \delta_R^{(t)} + \delta_k^{(t)}. \]

That is, the updates converge geometrically to \( A^* \). Further, from Lemma 3, we have that the error in the coefficients only depends on the error in the dictionary, which leads us to our result on the error incurred by the coefficients. This completes the proof of our main result.

**Table 3:** Proof map: dependence of results.

<table>
<thead>
<tr>
<th>Lemmas</th>
<th>Result</th>
<th>Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma 1</td>
<td>Signed-support recovery by coefficient initialization step</td>
<td>–</td>
</tr>
<tr>
<td>Lemma 2</td>
<td>IHT update step preserves the correct signed-support</td>
<td>Claim 1, Lemma 1, and Claim 2</td>
</tr>
<tr>
<td>Lemma 3</td>
<td>Upper-bound on the error in coefficient estimation</td>
<td>Claim 1, Claim 2, Claim 3, and Claim 4</td>
</tr>
<tr>
<td>Lemma 4</td>
<td>Expression for the coefficient estimate at the end of ( t )-th IHT iteration</td>
<td>Claim 5</td>
</tr>
<tr>
<td>Lemma 5</td>
<td>Expression for the expected gradient vector</td>
<td>Lemma 4 and Claim 7</td>
</tr>
<tr>
<td>Lemma 6</td>
<td>Concentration of the empirical gradient vector</td>
<td>Claim 8 and Claim 9</td>
</tr>
<tr>
<td>Lemma 7</td>
<td>Empirical gradient vector is correlated with the descent direction</td>
<td>Claim 10</td>
</tr>
<tr>
<td>Lemma 8</td>
<td>Concentration of the empirical gradient matrix</td>
<td>Claim 8 and Claim 10</td>
</tr>
<tr>
<td>Lemma 9</td>
<td>( A^{(t+1)} ) maintains closeness</td>
<td>Lemma 5, Claim 7 and Lemma 8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Claims</th>
<th>Result</th>
<th>Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claim 1</td>
<td>Incoherence of ( A^{(t)} )</td>
<td>–</td>
</tr>
<tr>
<td>Claim 2</td>
<td>Bound on ( \beta_{\text{alg}}^{(t)} ): the noise component in coefficient estimate that depends on ( \epsilon_t )</td>
<td>–</td>
</tr>
<tr>
<td>Claim 3</td>
<td>Error in coefficient estimation for a general iterate ( (t + 1) )</td>
<td>–</td>
</tr>
<tr>
<td>Claim 4</td>
<td>An intermediate result for bounding the error in coefficient calculations</td>
<td>Claim 2</td>
</tr>
<tr>
<td>Claim 5</td>
<td>Bound on the noise term in the estimation of a coefficient element in the support</td>
<td>Claim 6</td>
</tr>
<tr>
<td>Claim 6</td>
<td>An intermediate result for ( \delta_{\text{alg}}^{(t)} ) calculations</td>
<td>Claim 3</td>
</tr>
<tr>
<td>Claim 7</td>
<td>Bound on the noise term in expected gradient vector estimate</td>
<td>Claim 6 and Claim 2</td>
</tr>
<tr>
<td>Claim 8</td>
<td>An intermediate result for concentration results</td>
<td>Lemma 1, Lemma 4 and Claim 5</td>
</tr>
<tr>
<td>Claim 9</td>
<td>Bound on variance parameter for concentration of gradient vector</td>
<td>Claim 5</td>
</tr>
<tr>
<td>Claim 10</td>
<td>Bound on variance parameter for concentration of gradient matrix</td>
<td>Lemma 1, Lemma 4 and Claim 5</td>
</tr>
</tbody>
</table>

C Appendix: Proof of Lemmas

We present the proofs of the Lemmas used to establish our main result. Also, see Table 3 for a map of dependence between the results, and Appendix D for proofs of intermediate results.

**Proof of Lemma 1.** Let \( y \in \mathbb{R}^n \) be general sample generated as \( y = A^*x^* \), where \( x^* \in \mathbb{R}^m \) is a sparse random vector with support \( S = \text{supp}(x^*) \) distributed according to \( D.4 \).

The initial decoding step at the \( t \)-th iteration (shown in Algorithm 1) involves evaluating the inner-product between the estimate of the dictionary \( A^{(t)} \), and \( y \). The \( i \)-th element of the resulting vector can be written as

\[ \langle A_i^{(t)}, y \rangle = \langle A_i^{(t)}, A_i^* \rangle x_i^* + w_i, \]

where \( w_i = \langle A_i^{(t)}, A_{-i}^* x_{-i}^* \rangle \). Now, since \( \|A_i^* - A_i^{(t)}\|_2 \leq \epsilon_t \) and

\[ \|A_i^* - A_i^{(t)}\|_2^2 = \|A_i^*\|^2 + \|A_i^{(t)}\|^2 - 2\langle A_i^{(t)}, A_i^* \rangle = 2 - 2\|\langle A_i^{(t)}, A_i^* \rangle\|, \]

we have

\[ \|\langle A_i^{(t)}, A_i^* \rangle\| \geq 1 - \epsilon_t^2/2. \]
Therefore, the term
\[ |\langle A_i^{(t)}, A_i^* \rangle x_i^*| \left\{ \begin{array}{ll} \geq (1 - \frac{\gamma^2}{2})C & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{array} \right. \]

Now, we focus on the \( w_i \) and show that it is small. By the definition of \( w_i \), we have
\[ w_i = \langle A_i^{(t)}, A_i^*, x_i^* - i \rangle = \sum_{t \neq i} \langle A_i^{(t)}, A_i^* \rangle x_i^* = \sum_{\ell \in S \setminus \{i\}} \langle A_i^{(t)}, A_i^* \rangle x_i^*. \]

Here, since \( \text{var}(x_i^*) = 1 \), \( w_i \) is a zero-mean random variable with variance
\[ \text{var}(w_i) = \sum_{\ell \in S \setminus \{i\}} (\langle A_i^{(t)}, A_i^* \rangle)^2. \]

Now, each term in this sum can be bounded as,
\[ \langle A_i^{(t)}, A_i^* \rangle^2 = (\langle A_i^{(t)} - A_i^*, A_i^* \rangle + \langle A_i^*, A_i^* \rangle)^2 \leq 2(\langle A_i^{(t)} - A_i^*, A_i^* \rangle^2 + \langle A_i^*, A_i^* \rangle^2) \leq 2(\langle A_i^{(t)} - A_i^*, A_i^* \rangle^2 + \mu_i^2). \]

Next, \( \sum_{\ell \neq i} \langle A_i^{(t)} - A_i^*, A_i^* \rangle^2 \) can be upper-bounded as
\[ \sum_{\ell \in S \setminus \{i\}} (\langle A_i^{(t)} - A_i^*, A_i^* \rangle)^2 \leq \|A_i^*\|_2^2. \]

Therefore, we have the following as per our assumptions on \( \mu \) and \( k \),
\[ \|A_i^*\|_2^2 \leq (1 + k \frac{h}{\sqrt{n}}) \leq 2, \]
using Gershgorin Circle Theorem (Gershgorin, 1931). Therefore, we have
\[ \sum_{\ell \in S \setminus \{i\}} (\langle A_i^{(t)} - A_i^*, A_i^* \rangle)^2 \leq 2\epsilon_i^2. \]

Finally, we have that
\[ \sum_{\ell \in S \setminus \{i\}} (\langle A_i^{(t)}, A_i^* \rangle)^2 \leq 2(2\epsilon_i^2 + k \frac{\mu^2}{n}) = O(\epsilon_i^2). \]

Now, we apply the Chernoff bound for sub-Gaussian random variables \( w_i \) (shown in Lemma 12) to conclude that
\[ \Pr[|w_i| \geq C/4] \leq 2 \exp\left(-\frac{C^2}{8(\epsilon_i^2)}\right). \]

Further, \( w_i \) corresponding to each \( m \) should follow this bound, applying union bound we conclude that
\[ \Pr[\max_i |w_i| \geq C/4] \leq 2m \exp\left(-\frac{C^2}{8(\epsilon_i^2)}\right) \cdot \delta_i(t). \]

\( \square \)

**Proof of Lemma 2.** Consider the \( (r + 1) \)-th iterate \( x^{(r+1)} \) for the \( t \)-th dictionary iterate, where
\[ \|A_i^{(t)} - A_i^*\| \leq \epsilon_i \text{ for all } i \in [1, m] \]
evaluated as the following by the update step described in Algorithm 1,
\[ x^{(r+1)} = x^{(r)} - \eta_x^{(r+1)} A^{(t)\top} (A^{(t)} x^{(r)} - y) = (I - \eta_x^{(r+1)} A^{(t)\top}) x^{(r)} - \eta_x^{(r+1)} A^{(t)\top} A^* x^*, \]
(7)
where \( \eta_x^{(1)} < 1 \) is the learning rate or the step-size parameter. Now, using Lemma 1 we know that \( x^{(0)} \) (3) has the correct signed-support with probability at least \( 1 - \delta_i(t) \). Further, since \( A^{(t)\top} A^* \) can be written as
\[ A^{(t)\top} A^* = (A^{(t)} - A^*)^\top A^* + A^* A^*, \]
we can write the \((r + 1)\)-th iterate of the coefficient update step using (7) as
\[
x^{(r+1)} = (I - \eta_x^{(r+1)} A^{(t)\top} A^{(t)}) x^{(r)} - \eta_x^{(r+1)} (A^{(t)} - A^{*\top}) A^{*} x^{*} + \eta_x^{(r+1)} A^{*\top} A^{*} x^{*}.
\]

Further, the \(j\)-th entry of this vector is given by
\[
x_j^{(r+1)} = (I - \eta_x^{(r+1)} A^{(t)\top} A^{(t)}) x_j^{(r)} - \eta_x^{(r+1)} ((A^{(t)} - A^{*\top}) A^{*}) x_j^{*} + \eta_x^{(r+1)} (A^{*\top} A^{*}) x_j^{*}.
\]

We now develop an expression for the \(j\)-th element of each of the term in (8) as follows. First, we can write the first term as
\[
(I - \eta_x^{(r+1)} A^{(t)\top} A^{(t)}) x_j^{(r)} = (1 - \eta_x^{(r+1)}) x_j^{(r)} - \eta_x^{(r+1)} \sum_{i \neq j} \langle A^{(t)}_j, A^{(t)}_i \rangle x_i^{(r)}.
\]

Next, the second term in (8) can be expressed as
\[
\eta_x^{(r+1)} ((A^{(t)} - A^{*})^\top A^{*}) x_j^{*} = \eta_x^{(r+1)} \sum_i \langle A^{(t)}_j - A^{*}_j, A^{*}_i \rangle x_i^{*} = \eta_x^{(r+1)} (A^{(t)} - A^{*}_j, A^{*}_j) x_j^{*} + \eta_x^{(r+1)} \sum_{i \neq j} (A^{(t)}_j - A^{*}_j, A^{*}_i) x_i^{*}.
\]

Finally, we have the following expression for the third term,
\[
\eta_x^{(r+1)} (A^{*\top} A^{*}) x_j^{*} = \eta_x^{(r+1)} x_j^{*} + \eta_x^{(r+1)} \sum_{i \neq j} \langle A^{*}_j, A^{*}_i \rangle x_i^{*}.
\]

Now using our definition of \( \lambda_j^{(t)} = |\langle A^{(t)}_j - A^{*}_j, A^{*}_j \rangle| \leq \frac{\sqrt{e}}{2} \), combining all the results for (8), and using the fact that since \( A^{(t)} \) is close to \( A^{*} \), vectors \( A^{(t)}_j - A^{*}_j \) and \( A^{*}_j \) enclose an obtuse angle, we have the following for the \( j \)-th entry of the \((r + 1)\)-th iterate, \( x^{(r+1)} \) is given by
\[
x_j^{(r+1)} = (1 - \eta_x^{(r+1)}) x_j^{(r)} + \eta_x^{(r+1)} (1 - \lambda_j^{(t)}) x_j^{*} + \eta_x^{(r+1)} \xi_j^{(r+1)}.
\]

Here \( \xi_j^{(r+1)} \) is defined as
\[
\xi_j^{(r+1)} := \sum_{i \neq j} \langle A^{(t)}_j - A^{*}_j, A^{*}_i \rangle x_i^{*} - \sum_{i \neq j} \langle A^{(t)}_j, A^{*}_i \rangle x_i^{(r)}.
\]

Since, \( \langle A^{*}_j, A^{*}_i - A^{*}_j \rangle = \langle A^{*}_j, A^{*}_i \rangle + \langle A^{*}_j - A^{*}_j, A^{*}_i \rangle \), we can write \( \xi^{(r+1)} \), as
\[
\xi_j^{(r+1)} = \beta_j^{(t)} + \sum_{i \neq j} \langle A^{(t)}_j, A^{(t)}_i \rangle (x_i^{*} - x_i^{(r)}),
\]

where \( \beta_j^{(t)} \) is defined as
\[
\beta_j^{(t)} := \sum_{i \neq j} \langle A^{*}_j, A^{*}_i - A^{*}_j \rangle + \langle A^{*}_j - A^{*}_j, A^{*}_i \rangle + \langle A^{(t)}_j - A^{*}_j, A^{*}_i \rangle x_i^{*}.
\]

Note that \( \beta_j^{(t)} \) does not change for each iteration \( r \) of the coefficient update step. Further, by Claim 2 we show that \( |\beta_j^{(t)}| \leq t_{\beta} \) with probability at least \((1 - \delta_{\beta}^{(t)})\). Next, we define \( \xi_j^{(r+1)} \) as
\[
\xi_j^{(r+1)} := \beta_j^{(t)} + \sum_{i \neq j} |\langle A^{(t)}_j, A^{(t)}_i \rangle| |x_i^{*} - x_i^{(r)}|.
\]

Therefore, for the \((r + 1)\)-th iteration, we choose the threshold to be
\[
\tau^{(r+1)} := \eta_x^{(r+1)} \xi_{\max}^{(r+1)},
\]
and the step-size by setting the “noise” component of \( \eta_{(r+1)} \) to be smaller than the “signal” part, specifically, half the signal component, i.e.,

\[
\eta_{x}^{(r+1)}\xi_{\max}^{(r+1)} \leq \frac{1-n_x^{(r+1)}}{2} x_{\min}^{(r)} + \frac{n_x^{(r+1)}}{2} (1 - \frac{\epsilon^2}{2}) C,
\]

Also, since we choose the threshold as \( \tau^{(r)} := \eta_{x}^{(r)} \xi_{\max}^{(r)} \), \( x_{\min}^{(r)} = \eta_{x}^{(r)} \xi_{\max}^{(r)} \), where \( x_{\min}^{(0)} = C/2 \), we have the following for the \( (r+1) \)-th iteration,

\[
\eta_{x}^{(r+1)}\xi_{\max}^{(r+1)} \leq \frac{1-n_x^{(r+1)}}{2} \eta_{x}^{(r)} \xi_{\max}^{(r)} + \frac{n_x^{(r+1)}}{2} (1 - \frac{\epsilon^2}{2}) C.
\]

Therefore, for this step we choose \( \eta_{x}^{(r+1)} \) as

\[
\eta_{x}^{(r+1)} \leq \frac{\eta_{x}^{(r)} \xi_{\max}^{(r)}}{\xi_{\max}^{(r+1)}} + \frac{n_x^{(r+1)}}{2} (1 - \frac{\epsilon^2}{2}) C.
\]

Therefore, since \( \xi_{\max}^{(r)} \) are proportional to \( \|x^{(r+1)} - x^{*}\|_1 = O(k), \eta_{x}^{(r+1)} \) can be chosen as

\[
\eta_{x}^{(r+1)} \leq c^{(r+1)}(\epsilon, \mu, k, n),
\]

for a small constant \( c^{(r+1)}(\epsilon, \mu, k, n) \), \( \eta_{x}^{(r+1)} \). Further, since we initialize with the hard-thresholding step, the entries in \( x^{(0)} \) \( \geq C/2 \). Here, we define \( \xi_{\max}^{(0)} = C \) and \( \eta_{x}^{(0)} = 1/2 \), and set the threshold for initial step as \( \eta_{x}^{(0)} = 0 \).

\[
\Box
\]

**Proof of Lemma 3.** Using the definition of \( \xi_{i_1}^{(t)} \) as in (12), we have

\[
\xi_{i_1}^{(t)} = \beta_{i_1}^{(t)} + \sum_{i_2 \neq i_1} \|A_{i_1}^{(t)}, A_{i_2}^{(t)}\|_2 x_{i_2}^{*} - x_{i_2}^{(t-1)}\|_1.
\]

From Claim 2 we have that \( |\beta_{i_1}^{(t)}| \leq t_3 \) with probability at least \( (1 - \delta_{\beta}^{(t)} \). Further, using Claim 1, and letting \( C_{i_2}^{(t)} := \|x_{i_2}^{*} - x_{i_2}^{(t)}\|_1 = \|x_{i_2}^{*} - x_{i_2}^{(t-1)}\|_1 \), \( \xi_{i_1}^{(t)} \) can be upper-bounded as

\[
\xi_{i_1}^{(t)} \leq \beta_{i_1}^{(t)} + \frac{\mu_\ell}{2} \sum_{i_2 \neq i_1} C_{i_2}^{(t-1)}.
\]

Rearranging the expression for \( (r+1) \)-th update (9), and using (15) we have the following upper-bound

\[
C_{i_1}^{(r+1)} \leq (1 - \eta_{x}^{(r+1)}) C_{i_1}^{(r)} + \eta_{x}^{(r+1)} \lambda_{i_1}^{(t)} |x_{i_1}^{*} | + \eta_{x}^{(r+1)} \xi_{i_1}^{(r+1)}.
\]

Next, recursively substituting in for \( C_{i_1}^{(t)} \), where we define \( \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}) = 1, \)

\[
C_{i_1}^{(r+1)} \leq C_{i_1}^{(0)} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}) + \lambda_{i_1}^{(t)} |x_{i_1}^{*} | \sum_{q=1}^{r+1} \eta_{x}^{(q+1)} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}) + \sum_{q=0}^{t} \eta_{x}^{(q+1)} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}).
\]

Substituting for the upper-bound of \( \xi_{i_1}^{(t)} \) from (15),

\[
C_{i_1}^{(r+1)} \leq \lambda_{i_1}^{(t)} |x_{i_1}^{*} | \sum_{q=1}^{r+1} \eta_{x}^{(q+1)} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}) + \mu_\ell \sum_{q=0}^{t} \eta_{x}^{(q+1)} \sum_{q=1}^{r+1} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}) - \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}).
\]

Here, \( \alpha_{i_1}^{(r+1)} \) is defined as

\[
\alpha_{i_1}^{(r+1)} = C_{i_1}^{(0)} \prod_{q=0}^{r} (1 - \eta_{x}^{(q+1)}) + \lambda_{i_1}^{(t)} |x_{i_1}^{*} | \prod_{q=0}^{r} \eta_{x}^{(q+1)} \prod_{q=0}^{t} (1 - \eta_{x}^{(q+1)}).
\]

Our aim now will be to express \( C_{i_1}^{(t)} \) for \( \ell > 0 \) in terms of \( C_{i_2}^{(0)} \). Let each \( \alpha_{i_j}^{(t)} \leq \alpha_{i_1}^{(t)} \) where \( j = i_1, i_2, \ldots, i_k \). Similarly, let \( \alpha_{i_j}^{(t)} \leq C_{i_2}^{(0)} \) for \( j = i_1, i_2, \ldots, i_k \), and all \( \eta_{x}^{(t)} = \eta_{x} \). Then, using Claim 3 we have the following expression for \( C_{i_1}^{(R+1)} \),

\[
C_{i_1}^{(R+1)} \leq \alpha_{i_1}^{(R+1)} + \eta_{x} \mu_\ell \sum_{q=0}^{R} \alpha_{i_1}^{(t)} \prod_{q=0}^{R} (1 - \eta_{x} + \eta_{x} \mu_\ell)^{R-\ell}
\]
\[ + (k-1)\eta_x \frac{\mu}{\sqrt{n}} C_{i_{\max}}^0 (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^R. \]

Here, \((1 - \eta_x)^R \leq (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^R \leq \delta_R\). Next from Claim 4 we have that with probability at least \((1 - \delta_{\beta}^t)\),
\[
\sum_{\ell=1}^R \alpha_{i_{\max}}^t (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-\ell} \leq C_{i_{\max}}^0 R \delta_R + \frac{1}{\eta_x (1 - \frac{\mu}{\sqrt{n}})} \left( \frac{1}{2} |x_{i_{\max}}^*| + t_{\beta} \right).
\]

Therefore, for \(c_x = \frac{\mu}{\sqrt{n}} / (1 - \frac{\mu}{\sqrt{n}})\)
\[
C_{i_{i_{t+1}}}^{(R+1)} \leq \alpha_{i_{i_{t+1}}}^{(R+1)} + (k-1)c_x \left( \frac{1}{2} |x_{i_{i_{t+1}}}^*| + t_{\beta} \right) + (R+1)(k-1)\eta_x \frac{\mu}{\sqrt{n}} C_{i_{\max}}^0 \delta_R.
\]

Now, using the definition of \(\alpha_{i_{i_{t+1}}}^{(R+1)}\), and using the result on sum of geometric series, we have
\[
\alpha_{i_{i_{t+1}}}^{(R+1)} = C_{i_{i_{t}}}^0 (1 - \eta_x)^{R+1} + \frac{\lambda_{i_{i_{t}}}^t |x_{i_{i_{t}}}^*| + \beta_{i_{i_{t}}}^t}{R+1} \sum_{s=1}^{R+1} \eta_x (1 - \eta_x)^{R-s+1},
\]
\[
= C_{i_{i_{t}}}^0 \delta_R + \lambda_{i_{i_{t}}}^t |x_{i_{i_{t}}}^*| + \beta_{i_{i_{t}}}^t \leq C_{i_{i_{t}}}^0 \delta_R + \frac{1}{2} |x_{i_{i_{t}}}^*| + t_{\beta}.
\]

Therefore, \(C_{i_{i_{t}}}^{(R)}\) is upper-bounded as
\[
C_{i_{i_{t}}}^{(R)} \leq (c_x k + 1) \left( \frac{1}{2} |x_{i_{i_{t}}}^*| + t_{\beta} \right) + (R+1)k\eta_x \frac{\mu}{\sqrt{n}} C_{i_{\max}}^0 \delta_R + C_{i_{i_{t}}}^{(R)} \delta_R.
\]

Further, since \(k = \mathcal{O}(\sqrt{n}/\mu \log(n))\), \(kc_x < 1\), therefore, we have
\[
C_{i_{i_{t}}}^{(R)} \leq \mathcal{O}(t_{\beta}) + (R+1)k\eta_x \frac{\mu}{\sqrt{n}} C_{i_{\max}}^0 \delta_R + C_{i_{i_{t}}}^{(R)} \delta_R,
\]
with probability at least \((1 - \delta_{\beta}^t)\). Here, \((R+1)k\eta_x \frac{\mu}{\sqrt{n}} C_{i_{\max}}^0 \delta_R + C_{i_{i_{t}}}^{(R)} \delta_R \approx 0\) for an appropriately large \(R\). Therefore, the error in each non-zero coefficient is
\[
C_{i_{i_{t}}}^{(R)} = \mathcal{O}(t_{\beta}).
\]
\[\text{with probability at least } (1 - \delta_{\beta}^t). \qedhere \]

**Proof of Lemma 4.** Using the expression for \(x_{i_{i_{t}}}^{(R)}\) as defined in (9), and recursively substituting for \(x_{i_{i_{t}}}^{(r)}\) we have
\[
x_{i_{i_{t}}}^{(R)} = (1 - \eta_x)^R x_{i_{i_{t}}}^{(0)} + \sum_{r=1}^R \eta_x (1 - \lambda_{i_{i_{t}}}^t) (1 - \eta_x)^{R-r} + \sum_{r=1}^R \eta_x \xi_{i_{i_{t}}}^{(r)} (1 - \eta_x)^{R-r},
\]
where we set all \(\eta_x^r\) to be \(\eta_x\). Further, on defining
\[
\hat{\psi}_{i_{i_{t}}}^{(R)} := \sum_{r=1}^R \eta_x \xi_{i_{i_{t}}}^{(r)} (1 - \eta_x)^{R-r} + \gamma_{i_{i_{t}}}^{(R)},
\]
where \(\gamma_{i_{i_{t}}}^{(R)} := (1 - \eta_x)^R (x_{i_{i_{t}}}^{(0)} - x_{i_{i_{t}}}^* (1 - \lambda_{i_{i_{t}}}^t) )\), we have
\[
x_{i_{i_{t}}}^{(R)} = (1 - \eta_x)^R x_{i_{i_{t}}}^{(0)} + \sum_{r=1}^R \eta_x \xi_{i_{i_{t}}}^{(r)} (1 - \eta_x)^{R-r},
\]
\[
= x_{i_{i_{t}}}^* (1 - \lambda_{i_{i_{t}}}^t) + \hat{\psi}_{i_{i_{t}}}^{(R)} . \quad (19)
\]

Note that \(\gamma_{i_{i_{t}}}^{(R)}\) can be made appropriately small by choice of \(R\). Further, by Claim 5 we have
\[
|\hat{\psi}_{i_{i_{t}}}^{(R)}| \leq \mathcal{O}(t_{\beta}),
\]
with probability at least \((1 - \delta_{\beta}^t)\), where \(t_{\beta} = \mathcal{O}(\sqrt{k\epsilon t})\). \qedhere
Proof of Lemma 5. From Lemma 4 we have that for each $j \in S$,
\[
\hat{x}_S := x_S^{(R)} = (I - \Lambda_S^{(t)})x_S^* + \theta_S^{(R)},
\]
with probability at least $(1 - \delta_1^{(t)} - \delta_2^{(t)})$. Further, let $F_{\hat{x}^*}$ be the event that $\text{sign}(x^*) = \text{sign}(\hat{x})$, and let $1_{F_{\hat{x}^*}}$ denote the indicator function corresponding to this event. As we show in Lemma 2, this event occurs with probability at least $(1 - \delta_1^{(t)} - \delta_2^{(t)})$. Using this, we can write the expected gradient vector corresponding to the $j$-th sample as $1_{F_{\hat{x}^*}}$.

\[
g_j^{(t)} = E[(A^{(t)}\hat{x} - y)\text{sign}(x_j^*)1_{F_{\hat{x}^*}}] + E[(A^{(t)}\hat{x} - y)\text{sign}(x_j^*)1_{\neg F_{\hat{x}^*}}],
\]

Here, $\gamma := E[(A^{(t)}\hat{x} - y)\text{sign}(x_j^*)1_{F_{\hat{x}^*}}]$ is small and depends on $\delta_1^{(t)}$ and $\delta_2^{(t)}$, which in turn drops with $\epsilon_t$. Therefore, $\gamma$ diminishes with $\epsilon_t$. Further, since $1_{F_{\hat{x}^*}} + 1_{\neg F_{\hat{x}^*}} = 1$, and $Pr[F_{\hat{x}^*}] = (1 - \delta_1^{(t)} - \delta_2^{(t)})$, is very large,
\[
g_j^{(t)} = E[(A^{(t)}\hat{x} - y)\text{sign}(x_j^*)1_{\neg F_{\hat{x}^*}}] \pm \gamma,
\]
Therefore, we can write $g_j^{(t)}$ as
\[
g_j^{(t)} = E[(A^{(t)}\hat{x} - y)\text{sign}(x_j^*)] \pm \gamma,
= E[(1 - \eta_j)^R A_S^{(t)}\hat{x}_S^{(0)} + A_S^{(t)}(1 - \Lambda_S^{(t)})x_S^* + A_S^{(t)}\theta_S^{(R)} - A_S^*x_S^*)\text{sign}(x_j^*)] \pm \gamma.
\]
Since $E[(1 - \eta_j)^R A_S^{(t)}\hat{x}_S^{(0)}]$ can be made very small by choice of $R$, we absorb this term in $\gamma$. Therefore,
\[
g_j^{(t)} = E[A_S^{(t)}(1 - \Lambda_S^{(t)})x_S^* + A_S^{(t)}\theta_S^{(R)} - A_S^*x_S^*)\text{sign}(x_j^*)] \pm \gamma.
\]
Writing the expectation by sub-conditioning on the support,
\[
g_j^{(t)} = E_S[A_S^{(t)}(1 - \Lambda_S^{(t)})x_S^*\text{sign}(x_j^*) - A_S^*x_S^*)\text{sign}(x_j^*)] \pm \gamma,
= E_S[A_S^{(t)}(1 - \Lambda_S^{(t)})E_{x_S^*}[x_S^*\text{sign}(x_j^*)] - A_S^*E_{x_S^*}[x_S^*\text{sign}(x_j^*)]] \pm \gamma,
= E_S[p_j(1 - \lambda_j^{(t)})A_j^{(t)} - p_jA_j^*] + \Delta_j^{(t)} \pm \gamma,
\]
where we have used the fact that $E_{x_S^*}[\text{sign}(x_j^*)] = 0$ and introduced
\[
\Delta_j^{(t)} = E[A_S^{(t)}\theta_S^{(R)}\text{sign}(x_j^*)].
\]
Next, since $p_j = E_{x_S^*}[x_S^*\text{sign}(x_j^*)][j \in S]$, therefore,
\[
g_j^{(t)} = E_S[p_j(1 - \lambda_j^{(t)})A_j^{(t)} - p_jA_j^*] + \Delta_j^{(t)} \pm \gamma.
\]
Further, since $q_j = Pr[j \in S] = O(k/m)$,
\[
g_j^{(t)} = q_j p_j (1 - \lambda_j^{(t)}) A_j^{(t)} - A_j^* + \frac{1}{q_j p_j} \Delta_j^{(t)} \pm \gamma.
\]
Further, by Claim 7 we have that
\[
||\Delta_j^{(t)}|| = O(\sqrt{m}q_i \epsilon_t ||A^{(t)}||).
\]
This completes the proof. \hfill \Box

Proof of Lemma 6. Let $W = \{j : i \in \text{supp}(x_{(j)})\}$ and then we have that
\[
\hat{g}_i^{(t)} = \frac{|W|}{p} \frac{1}{|W|} \sum_{j \in W} (y_{(j)} - A^{(t)}\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)}(i)),
\]

where \( \hat{x}_{(j)}(i) \) denotes the \( i \)-th element of the coefficient estimate corresponding to the \( (j) \)-th sample. Here, for \( \ell = |W| \) the summation
\[
\sum_{j=1}^{\ell} \frac{1}{\ell} (y_{(j)} - A^{(t)} \hat{x}_{(j)}(i)) \text{sign}(\hat{x}_{(j)}(i)),
\]
has the same distribution as \( \Sigma_{j=1}^{\ell} z_j \), where each \( z_j \) belongs to a distribution as
\[
z := \frac{1}{\ell} (y - A^{(t)} \hat{x}) \text{sign}(\hat{x}) |i \in S|.
\]
Also, \( E[y - A^{(t)} \hat{x}] \text{sign}(\hat{x})] = q_i E[z] \), where \( q_i = \Pr[x^*_i \neq 0] = \Theta(\frac{k}{m}) \). Therefore, since \( p = \tilde{\Omega}(mk^2) \), we have \( \ell = pq_i = \tilde{\Omega}(k^3) \) non-zero vectors,
\[
\|\tilde{g}_i^{(t)} - g_i^{(t)}\| \leq O(\frac{k}{m}) \|\Sigma_{j=1}^{\ell} (z_j - E[z])\|.
\]
Let \( w_j = z_j - E[z] \), we will now apply the vector Bernstein result shown in Lemma 11. For this, we require bounds on two parameters for these \( L := \|w_j\| \) and \( \sigma^2 := \|\Sigma_j E[|w_j|]^2\| \). Note that, since the quantity of interest is a function of \( x^*_i \), which are sub-Gaussian, they are only bounded almost surely. To this end, we will employ Lemma 14 (Lemma 45 in (Arora et al., 2015)) to get a handle on the concentration.

**Bound on the norm \( \|w\| \):** This bound is evaluated in Claim 8, which states that with probability at least \( (1 - \delta^{(t)}_\beta - \delta^{(t)}_\ell - \delta^{(t)}_{\text{HW}}) \),
\[
L := \|w\| = \|z - E[z]\| = \frac{2}{\ell} \|y - A^{(t)} \hat{x} \text{sign}(\hat{x})|i \in S|\| \leq \frac{2}{\ell} \|y - A^{(t)} \hat{x}\| = \tilde{O}(\frac{kt}{\ell^2}).
\]

**Bound on variance parameter \( E[\|w\|^2] \):** Using Claim 9, we have \( E[\|w\|^2] = O(k\epsilon^2_t) + O(kt^2) \). Therefore, the bound on the variance parameter \( \sigma^2 \) is given by
\[
\sigma^2 := \|\Sigma_j E[|w_j|^2]\| \leq \|\Sigma_j E[|z_j|^2]\| \leq \tilde{O}((\frac{kt^2}{\ell^2}) + O(\frac{kt^2}{\ell^2})).
\]

From Claim 2 we have that with probability at least \( (1 - \delta^{(t)}_\ell) \), \( t_\beta = O(\sqrt{k\epsilon_t}) \). Applying vector Bernstein inequality shown in Lemma 11 and using Lemma 14 (Lemma 45 in (Arora et al., 2015)), choosing \( \ell = \tilde{\Omega}(k^3) \), we conclude
\[
\|\sum_{j=1}^{\ell} z_j - E[z]\| = O(L) + O(\sigma) = o(\epsilon_t),
\]
with probability at least \( (1 - \delta^{(t)}_\ell) \), where \( \tilde{g}_i^{(t)} = \exp(-\Omega(k)) \). Finally, substituting in (20) we have
\[
\|\tilde{g}_i^{(t)} - g_i^{(t)}\| = O(\frac{k}{m}) o(\epsilon_t),
\]
with probability at least \( (1 - \delta^{(t)}_\ell - \delta^{(t)}_\beta - \delta^{(t)}_\ell - \delta^{(t)}_{\text{HW}}) \). \( \square \)

**Proof of Lemma 7.** Since we only have access to the empirical estimate of the gradient \( \tilde{g}_i^{(t)} \), we will show that this estimate is correlated with \( (A^{(t)}_j - A^*_j) \). To this end, first from Lemma 6 we have that the empirical gradient vector concentrates around its mean, specifically,
\[
\|\tilde{g}_i^{(t)} - g_i^{(t)}\| \leq O(\frac{k}{m}) \epsilon_t,
\]
with probability at least \( (1 - \delta^{(t)}_\ell - \delta^{(t)}_\beta - \delta^{(t)}_\ell - \delta^{(t)}_{\text{HW}}) \). From Lemma 5, we have the following expression for the expected gradient vector
\[
g_i^{(t)} = p_j q_j (A^{(t)}_j - A^*_j) + p_j q_j (-\lambda^{(t)}_j A^{(t)}_j + \frac{1}{p_j q_j} \Delta^{(t)}_j \pm \gamma).
\]
Let \( \tilde{g}_i^{(t)} = 4\rho_\gamma (A^{(t)}_j - A^*_j) + v \), where \( 4\rho_\gamma = p_j q_j \) and \( v \) is defined as
\[
v = p_j q_j (-\lambda^{(t)}_j A^{(t)}_j + \frac{1}{p_j q_j} \Delta^{(t)}_j \pm \gamma).
\]
Then, \( \tilde{g}_i^{(t)} \) can be written as
\[
\tilde{g}_i^{(t)} = g_i^{(t)} + \tilde{g}_i^{(t)}.
\]
Here, we use the fact that R.H.S. can be minimized only if
\[ \tilde{v} = v + (\tilde{g}_i^{(t)} - g_i^{(t)}) \]. Let \( \|\tilde{v}\| \leq \rho_1\|A_i^{(t)} - A_i^*\| \). Using the definition of \( v \) as shown in (21) we have
\[ \|\tilde{v}\| \leq q_j p_j \lambda_j^{(t)} \|A_j^{(t)}\| + \|\Delta_j^{(t)}\| + o(k \|\epsilon_t\|) \pm \gamma. \]

Now for the first term, since \( \|A_j^{(t)}\| = 1 \), we have
\[ \lambda_j^{(t)} = |\langle A_j^{(t)} - A_j^* , A_j^* \rangle| = \frac{1}{2} \|A_j^{(t)} - A_j^*\|^2. \]
Therefore
\[ q_j p_j \lambda_j^{(t)} \|A_j^{(t)}\| = q_j p_j \frac{1}{2} \|A_j^{(t)} - A_j^*\|^2, \]
Further, using Claim 7
\[ \|\Delta_j^{(t)}\| = O(\sqrt{m} p_{i,j} \epsilon_t \|A^{(t)}\|). \]

Now, since \( \|A^{(t)} - A^*\| \leq 2 \|A^*\| \), and further since \( \|A^*\| = O(\sqrt{m/n}) \), we have that
\[ \|A^{(t)}\| \leq \|A^{(t)} - A^*\| + \|A^*\| = O(\sqrt{m/n}). \]
Therefore, we need
\[ \|\Delta_j^{(t)}\| + o(k \|\epsilon_t\|) \pm \gamma = O(\sqrt{m} p_{i,j} \epsilon_t \|A^{(t)}\|) \leq \frac{\eta_4 p_j}{\sqrt{m}} \|A_j^{(t)} - A_j^*\|. \]

Here, we use the fact that \( \gamma \) drops with decreasing \( \epsilon_t \) as argued in Lemma 5. Therefore, if \( k = O(\sqrt{m}) \) we have
\[ \|\tilde{v}\| \leq q_j p_j \|A_j^{(t)} - A_j^*\|. \]

Now, using (22), we have
\[ \|\tilde{g}_i^{(t)}\| \leq 4 \rho_1 \|A_i^{(t)} - A_i^*\| + \|\tilde{v}\|. \]

Substituting for \( \|\tilde{v}\| \), this implies that \( \|\tilde{g}_i^{(t)}\|^2 \leq 25 \rho_1^2 \|A_i^{(t)} - A_i^*\|^2 \). Further, we also have the following lower-bound
\[ \langle \tilde{g}_i^{(t)} , A_i^{(t)} - A_i^* \rangle \geq 4 \rho_1 \|A_i^{(t)} - A_i^*\|^2 - \|\tilde{v}\| \|A_i^{(t)} - A_i^*\| - \rho_4 \|A_i^{(t)} - A_i^*\|^2 - \frac{1}{100 \rho_4} \|\tilde{g}_i^{(t)}\|^2, \]

Here, we use the fact that R.H.S. can be minimized only if \( \tilde{v} \) is directed opposite to the direction of \( A_i^{(t)} - A_i^* \). Now, we show that this gradient is \( (\rho_1 , 1/100 \rho_4, 0) \) correlated,
\[ \langle \tilde{g}_i^{(t)} , A_i^{(t)} - A_i^* \rangle - \rho_4 \|A_i^{(t)} - A_i^*\|^2 - \frac{1}{100 \rho_4} \|\tilde{g}_i^{(t)}\|^2, \]
\[ \geq 4 \rho_1 \|A_i^{(t)} - A_i^*\|^2 - \|\tilde{v}\| \|A_i^{(t)} - A_i^*\| - \rho_4 \|A_i^{(t)} - A_i^*\|^2 - \frac{1}{100 \rho_4} \|\tilde{g}_i^{(t)}\|^2, \]
\[ \geq 4 \rho_1 \|A_i^{(t)} - A_i^*\|^2 - 2 \rho_4 \|A_i^{(t)} - A_i^*\|^2 - \frac{25 \rho_1^2 \|A_i^{(t)} - A_i^*\|^2}{100 \rho_4}, \]
\[ \geq \rho_4 \|A_i^{(t)} - A_i^*\|^2 \geq 0. \]

Therefore, for this choice of \( k \), there is no bias. Further, since we choose \( 4 \rho_1 = p_j q_j \), we have that \( \rho_* = \Theta(k/m) \), therefore \( \rho_+ = 1/100 \rho_* = \Omega(m/k) \). Applying Lemma 15 we have
\[ \|A_i^{(t+1)} - A_i^*\|^2 \leq (1 - \rho_4 \eta_4) \|A_i^{(t)} - A_i^*\|^2, \]
for \( \eta_4 = O(m/k) \) with probability at least \( (1 - \delta_T^{(i)} - \delta_{\tilde{g}_i^{(t)}}^{(i)}). \)

**Proof of Lemma 8.** Here, we will prove that \( \tilde{g}_i^{(t)} \) defined as
\[ \tilde{g}_i^{(t)} = \sum_j \langle x_i^{(t)} - A_j^{(t)} \rangle \tilde{x}_j^{(t)} \text{sign}(\tilde{x}_j^{(t)})^\top, \]

is directed opposite to the direction of \( A_i^{(t)} - A_i^* \).
concentrates around its mean. Notice that each summand \( (y_{(j)} - A^{(t)}\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top \) is a random matrix of the form \( (y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top \). Also, we have \( g^{(t)} \) defined as
\[
g^{(t)} = E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top].
\]
To bound \( \|g^{(t)} - g^{(t)}\| \), we are interested in \( \|\sum_{j=1}^{p} W_j\| \), where each matrix \( W_j \) is given by
\[
W_j = \frac{1}{p}(y_{(j)} - A^{(t)}\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top - \frac{1}{p}E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top].
\]
Noting that \( E[W_j] = 0 \), we will employ the matrix Bernstein result (Lemma 10) to bound \( \|\hat{g}^{(t)} - g^{(t)}\| \). To this end, we will bound \( \|W_j\| \) and the variance proxy
\[
v(W_j) = \max\{\|\sum_{j=1}^{p} E[W_j W_j^\top]\|, \|\sum_{j=1}^{p} E[W_j W_j^\top]\| \}.
\]
**Bound on \( \|W_j\|\)– First, we can bound both terms in the expression for \( W_j \) by triangle inequality as
\[
\|W_j\| \leq \frac{1}{p}\|\|y_{(j)} - A^{(t)}\hat{x}_{(j)}\|\|\|\text{sign}(\hat{x}_{(j)})^\top\|\| + \frac{1}{p}\|\|E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top]\|\|
\]
\[
\leq \frac{2}{p}\|\|y_{(j)} - A^{(t)}\hat{x}\|\|\|\text{sign}(\hat{x})^\top\|\|.
\]
Here, we use Jensen’s inequality for the second term, followed by upper-bounding the expected value of the argument by \( \|\|y_{(j)} - A^{(t)}\hat{x}\|\|\|\text{sign}(\hat{x})^\top\|\| \).

Next, using Claim 8 we have that with probability at least \( (1 - \delta_\beta - \delta_T - \delta_{\text{HW}}) \), \( \|y - A^{(t)}\hat{x}\| \) is \( \tilde{O}(kt_\beta) \), and the fact that \( \|\text{sign}(x)^\top\| = \sqrt{k} \),
\[
\|W_j\| \leq \frac{2}{p}\|\sqrt{k}\|\|y - A^{(t)}\hat{x}\|\| = \tilde{O}(k\sqrt{kt_\beta}).
\]

**Bound on the variance statistic \( v(W_j)\)– For the variance statistic, we first look at \( \|\sum E[W_j W_j^\top]\| \),
\[
E[W_j W_j^\top] = \frac{1}{p}E[(y_{(j)} - A^{(t)}\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top - E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top]\]
\[
\times [\text{sign}(\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top - (E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top])] + \frac{1}{p}\|E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top]\|.
\]
Since \( E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top]\) is positive semi-definite,
\[
E[W_j W_j^\top] \leq \frac{1}{p}E[(y_{(j)} - A^{(t)}\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top\text{sign}(\hat{x}_{(j)}) (y_{(j)} - A^{(t)}\hat{x}_{(j)})^\top].
\]
Now, since each \( \hat{x}_{(j)} \) has \( k \) non-zeros, \( \text{sign}(\hat{x}_{(j)})^\top\text{sign}(\hat{x}_{(j)}) = k \), and using Claim 10, with probability at least \( (1 - \delta_\beta - \delta_T) \)
\[
\|\sum E[W_j W_j^\top]\| \leq \frac{1}{p}\|E[(y_{(j)} - A^{(t)}\hat{x}_{(j)})(y_{(j)} - A^{(t)}\hat{x}_{(j)})]\|,
\]
\[
= \tilde{O}(\frac{k^2 t_\beta^2}{pm})\|\mathbf{A}\|^2.
\]
Similarly, expanding \( E[W_j^\top W_j] \), and using the fact that \( E[(y - A^{(t)}\hat{x})\text{sign}(\hat{x})^\top]\) is positive semi-definite. Now, using Claim 8 and the fact that entries of \( E[\text{sign}(\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top] \) are \( q_{ij} \) on the diagonal and zero elsewhere, where \( q_{ij} = O(k/m) \),
\[
\|\sum E[W_j^\top W_j]\| \leq \frac{1}{p}\|E[\text{sign}(\hat{x}_{(j)})\text{sign}(\hat{x}_{(j)})^\top\text{sign}(\hat{x}_{(j)})^\top]\|\|y_{(j)} - A^{(t)}\hat{x}_{(j)}\|^2,
\]
\[
\leq O(\frac{k^2 t_\beta^2}{mp}) \tilde{O}(k^2 t_\beta^2) = \tilde{O}(\frac{k^2 t_\beta^2}{mp}).
\]
Now, we are ready to apply the matrix Bernstein result. Since, \( m = O(n) \) the variance statistic comes out to be \( \tilde{O}(\frac{k^2 t_\beta^2}{pm})\|\mathbf{A}\|^2 \), then as long as we choose \( p = \tilde{O}(mk^2) \), with probability at least
\[
(1 - \delta_\beta - \delta_T - \delta_{\text{HW}} - \delta_g)
\]
\[
\|g^{(t)} - g^{(t)}\| \leq \tilde{O}(\frac{k\sqrt{kt_\beta}}{p}) + \|\mathbf{A}\|\|\sqrt{\tilde{O}(\frac{k^3 t_\beta^2}{pm})}
\]
\[
= \tilde{O}(\|\mathbf{A}\|\|\)).
\]
where \( \delta_g = (n + m) \exp(-\Omega(m\sqrt{\log(n)})) \).
Proof of Lemma 9. The update step for the $i$-th dictionary element at the $s + 1$ iteration can be written as
\[
A_i^{(t+1)} - A_i^* = A_i^{(t)} - A_i^* - \eta A \hat{g}_i^{(t)},
\]
\[
= A_i^{(t)} - A_i^* - \eta A \hat{g}_i^{(t)} - \eta A (\hat{g}_i^{(t)} - g_i^{(t)}).
\]
Here, $g_i^{(t)}$ is given by the following as per Lemma 5 with probability at least $(1 - \delta_\beta^{(t)} - \delta_\delta^{(t)})$
\[
g_i^{(t)} = q_i p_i (A_i^{(t)} - A_i^*) + q_i p_i (\lambda_i^{(t)} A_i^{(t)} + \frac{1}{q_i p_i} \Delta_i^{(t)} \mp \gamma).
\]
Substituting the expression for $g_i^{(t)}$ in the dictionary update step,
\[
A_i^{(t+1)} - A_i^* = (1 - \eta A p_i q_i) (A_i^{(t)} - A_i^*) - \eta A p_i q_i \lambda_i^{(t)} A_i^{(t)} - \eta A (\hat{g}_i^{(t)} - g_i^{(t)}) \mp \gamma,
\]
where $\Delta_j^{(t)} = E[A_i^{(t)} q_j^{(R)} \text{sign}(x_j^*)]$. Therefore, the update step for the dictionary (matrix) can be written as
\[
A^{(t+1)} - A^* = (A^{(t)} - A^*) \text{diag}((1 - \eta A p_i q_i)) + \eta A U - \eta A V \pm \gamma - \eta A (\hat{g}^{(t)} - g^{(t)}),
\] (24)
where, $U = A^{(t)} \text{diag}(p_i q_i \lambda_i^*)$ and $V = A^{(t)} Q$, with the matrix $Q$ given by,
\[
Q_{i,j} = q_i E_{x} [q_{j}^{(R)} \text{sign}(x_j^*)] S,
\]
and using the following intermediate result shown in Claim 7,
\[
E_{x} [q_{j}^{(R)} \text{sign}(x_j^*)] S \leq \gamma_{i}^{(R)},
\]
for $i = j$, and
\[
= O(p_j \epsilon_i),
\]
for $i \neq j$, we have $\|Q\| = O(\sqrt{mq_{i,j}p_i \epsilon_i})$. Hence, we have
\[
\|Q\|_F \leq O(mq_{i,j}p_i \epsilon_i).
\]
Therefore,
\[
\|V\| \leq \|A^{(t)} \|_F \leq \|A^{(t)}\| \|Q\|_F = O(mq_{i,j}p_i \epsilon_i) = O(\frac{k^2}{m \log(n)}) \|A^*\|.
\]
We will now proceed to bound each term in (24). Starting with $(A^{(t)} - A^*) \text{diag}(1 - \eta A p_i q_i)$, and using the fact that $p_i = O(1)$, $q_i = O(k/m)$, and $\|A^{(t)} - A^*\| \leq 2 \|A^*\|$, we have
\[
\|(A^{(t)} - A^*) \text{diag}(1 - \eta A p_i q_i)\| \leq (1 - \min_i \eta A p_i q_i) \|(A^{(t)} - A^*)\| \leq 2(1 - \Omega(\eta A k/m)) \|A^*\|.
\]
Next, since $\|A_j^*\| = 1$, we have $\lambda_j^{(t)} = \| (A_j^{(t)} - A^*, A_j^*) \| = \frac{1}{2} \| A_j^{(t)} - A_j^* \|^2$, and $\lambda_i^{(t)} \leq \frac{\epsilon_i^2}{2k}$.
\[
\|U\| = \|A^{(t)} \text{diag}(p_i q_i \lambda_i^*)\| \leq \max_i p_i q_i \frac{\epsilon_i^2}{2k} \|A^{(t)} - A^* + A^*\| \leq O(\sqrt{k/m}) \|A^*\|.
\]
Using the results derived above, and the the result derived in Lemma 8 which states that with probability at least $(1 - \delta_\beta^{(t)} - \delta_\delta^{(t)} - \delta_{HW}^{(t)} - \delta_{\hat{g}}^{(t)})$, $\|\hat{g}^{(t)} - g^{(t)}\| = O(\sqrt{k/m} \|A^*\|)$ we have
\[
\|A^{(t+1)} - A^*\| = \| (A^{(t)} - A^*) D (1 - \eta A p_i q_i) \| + \eta A \| U \| + \eta A \| V \| + \eta A \| (\hat{g}^{(t)} - g^{(t)}) \| \| \pm \gamma,
\]
\[
\leq 2(1 - \Omega(\eta A \frac{k}{m})) \|A^*\| + O(\eta A \frac{k}{m}) \|A^*\| + O(\eta A \frac{k^2}{m \log(n)}) \|A^*\| + O(\eta A \frac{k^2}{m \log(n)}) \|A^*\| + O(\eta A \frac{k}{m}) \|A^*\| \| \pm \gamma,
\]
\[
\leq 2 \|A^*\|.
\]
D. Appendix: Proofs of Intermediate Results

Claim 1 (Incoherence of $A^{(t)}$). If $A^* \in \mathbb{R}^{n \times m}$ is $\mu$-incoherent and $\|A^*_i - A^{(t)}_i\| \leq \epsilon_t$ holds for each $i \in [1 \ldots m]$, then $A^{(t)} \in \mathbb{R}^{n \times m}$ is $\mu_t$-incoherent, where $\mu_t = \mu + 2\sqrt{n}t$.

Proof of Claim 1. We start by looking at the incoherence between the columns of $A^*$, for $j \neq i$,

$$
\langle A^*_i, A^*_j \rangle = \langle A^*_i - A^{(t)}_i, A^*_j \rangle + \langle A^{(t)}_i, A^*_j \rangle,
$$

$$
= \langle A^*_i - A^{(t)}_i, A^*_j \rangle + \langle A^{(t)}_i, A^*_j - A^{(t)}_j \rangle + \langle A^{(t)}_i, A^{(t)}_j \rangle.
$$

Since $\langle A^*_i, A^*_j \rangle \leq \frac{\mu}{\sqrt{n}}$,

$$
|\langle A^{(t)}_i, A^{(t)}_j \rangle| \leq |\langle A^*_i, A^*_j \rangle - \langle A^*_i - A^{(t)}_i, A^*_j \rangle - \langle A^*_i, A^*_j - A^{(t)}_j \rangle|,
$$

$$
\leq \frac{\mu}{\sqrt{n}} + 2\epsilon_t.
$$

QED

Claim 2 (Bound on $\beta^{(t)}_j$: the noise component in coefficient estimate that depends on $\epsilon_t$). With probability $(1 - \delta^{(t)}_\beta)$, $|\beta^{(t)}_j|$ is upper-bounded by $t_\beta = \mathcal{O}(\sqrt{\epsilon_t})$, where $\delta^{(t)}_\beta = 2k \exp(-\frac{1}{18k\epsilon_t})$.

Proof of Claim 2. We have the following definition for $\beta^{(t)}_j$ from (11),

$$
\beta^{(t)}_j = \sum_{i \neq j} (\langle A^*_j, A^*_i - A^{(t)}_i \rangle + \langle A^*_j - A^{(t)}_j, A^{(t)}_i \rangle + \langle A^{(t)}_j - A^*_j, A^*_i \rangle) x^*_j.
$$

Here, since $x^*_j$ are independent sub-Gaussian random variables, $\beta^{(t)}_j$ is a sub-Gaussian random variable with the variance parameter evaluated as shown below

$$
\text{var} [\beta^{(t)}_j] = \sum_{i \neq j} (\|A^*_j\|^2 + \|A^*_j - A^{(t)}_j\|^2 + \|A^{(t)}_j - A^*_j\|^2) \leq 9k\epsilon_t^2.
$$

Therefore, by Lemma 12

$$
\Pr [ |\beta^{(t)}_j| > t_\beta ] \leq 2 \exp(-\frac{t_\beta^2}{18k\epsilon_t}).
$$

Now, we need this for each $\beta^{(t)}_j$ for $j \in \text{supp}(x^*)$, union bounding over $k$ coefficients

$$
\Pr [ \max |\beta^{(t)}_j| > t_\beta ] \leq \delta^{(t)}_\beta,
$$

where $\delta^{(t)}_\beta = 2k \exp(-\frac{t_\beta^2}{18k\epsilon_t})$. Choosing $t_\beta = \mathcal{O}(\sqrt{\epsilon_t})$, we have that $\delta^{(t)}_\beta = 2k \exp(-\frac{1}{18k\epsilon_t})$. QED

Claim 3 (Error in coefficient estimation for a general iterate $(r + 1)$). The error in a general iterate $r$ of the coefficient estimation is upper-bounded as

$$
C_{i_1}^{(r+1)} \leq \alpha_1^{(r+1)} + (k - 1) \eta_x \frac{\mu}{\sqrt{n}} \sum_{\ell=1}^{r} \sigma^{(t)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{r-\ell}
$$

$$
+ (k - 1) \eta_x \frac{\mu}{\sqrt{n}} C_{\max}^{(0)} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{r}.
$$

Proof of Claim 3. From (16) we have the following expression for $C_{i_1}^{(r+1)}$

$$
C_{i_1}^{(r+1)} \leq \alpha_1^{(r+1)} + \frac{\mu}{\sqrt{n}} \sum_{\ell=1}^{r} \eta_x^{(t)} \sum_{i_2 \neq i_1} C_{i_2}^{(\ell-1)} \prod_{q=\ell}^{r+1} (1 - \eta_x^{(q+1)}).
$$

Our aim will be to recursively substitute for $C_{i_1}^{(\ell-1)}$ to develop an expression for $C_{i_1}^{(r+1)}$ as a function of $C_{\max}^{(0)}$. To this end, we start by analyzing the iterates $C_{i_1}^{(1)}, C_{i_1}^{(2)}, \text{and so on to develop an expression for } C_{i_1}^{(r+1)} \text{ as follows.}
Expression for $C_{i_1}^{(1)}$ - Consider $C_{i_1}^{(1)}$

$$C_{i_1}^{(1)} = \alpha_{i_1}^{(1)} + \frac{\mu_i}{\sqrt{n}} \sum_{\ell=1}^{1} \eta_{i} \sum_{i_2 \neq i_1} C_{i_2}^{(\ell-1)} \prod_{q=\ell}^{1} (1 - \eta_x),$$

$$= \alpha_{i_1}^{(1)} + \eta_{x} \left( \frac{\mu_i}{\sqrt{n}} \sum_{i_2 \neq i_1} C_{i_2}^{(0)} \right). \quad (25)$$

Expression for $C_{i_1}^{(2)}$ - Next, $C_{i_1}^{(2)}$ is given by

$$C_{i_1}^{(2)} = \alpha_{i_1}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{\ell=1}^{2} \sum_{i_2 \neq i_1} C_{i_2}^{(\ell-1)} \prod_{q=\ell}^{2} (1 - \eta_x),$$

$$\leq \alpha_{i_1}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( \sum_{i_2 \neq i_1} C_{i_2}^{(1)} + \sum_{i_2 \neq i_1} C_{i_2}^{(0)} (1 - \eta_x) \right).$$

Further, we know from (25) we have

$$C_{i_2}^{(1)} = \alpha_{i_2}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} C_{i_3}^{(0)}.$$ 

Therefore, since $\sum_{i_2 \neq i_1} \sum_{i_3 \neq i_1} = \sum_{i_1 \neq i_1}$,

$$C_{i_1}^{(2)} \leq \alpha_{i_1}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( \sum_{i_2 \neq i_1} \left( \alpha_{i_2}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} C_{i_3}^{(0)} \right) + \sum_{i_2 \neq i_1} C_{i_2}^{(0)} (1 - \eta_x) \right),$$

$$= \alpha_{i_1}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_2 \neq i_1} \left( \alpha_{i_2}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} C_{i_3}^{(0)} \right) + \sum_{i_2 \neq i_1} C_{i_2}^{(0)} (1 - \eta_x) + C_{i_2}^{(2)}). \quad (26)$$

Expression for $C_{i_1}^{(3)}$ - Next, we writing $C_{i_1}^{(3)}$,

$$C_{i_1}^{(3)} \leq \alpha_{i_1}^{(3)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{\ell=1}^{3} \sum_{i_2 \neq i_1} C_{i_2}^{(\ell-1)} (1 - \eta_x)^{3-\ell},$$

$$= \alpha_{i_1}^{(3)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_2 \neq i_1} \left( \alpha_{i_2}^{(0)} (1 - \eta_x)^2 + C_{i_2}^{(1)} (1 - \eta_x) + C_{i_2}^{(2)} \right),$$

$$\leq \alpha_{i_1}^{(3)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_2 \neq i_1} \left( C_{i_2}^{(0)} (1 - \eta_x)^2 + \alpha_{i_2}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} C_{i_3}^{(0)} (1 - \eta_x) + C_{i_2}^{(2)} \right).$$

Here, using (26) we have the following expression for $C_{i_2}^{(2)}$

$$C_{i_2}^{(2)} \leq \alpha_{i_2}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} \alpha_{i_3}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_4 \neq i_3} C_{i_4}^{(0)} + \sum_{i_3 \neq i_2} C_{i_3}^{(0)} (1 - \eta_x) \right).$$

Substituting for $C_{i_2}^{(2)}$ in the expression for $C_{i_1}^{(3)}$, and rearranging the terms in the expression for $C_{i_1}^{(3)}$, we have

$$C_{i_1}^{(3)} \leq \alpha_{i_1}^{(3)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_2 \neq i_1} \alpha_{i_2}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( (1 - \eta_x) \sum_{i_2 \neq i_1} \alpha_{i_2}^{(1)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} \alpha_{i_3}^{(1)} \right)$$

$$+ \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( (1 - \eta_x)^2 \sum_{i_2 \neq i_1} C_{i_2}^{(0)} + 2(1 - \eta_x) \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2} C_{i_3}^{(0)} + (1 - \eta_x)^2 \sum_{i_4 \neq i_2} C_{i_4}^{(0)} \right). \quad (27)$$

Expression for $C_{i_1}^{(4)}$ - Now, consider $C_{i_1}^{(4)}$

$$C_{i_1}^{(4)} \leq \alpha_{i_1}^{(4)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{\ell=1}^{4} \sum_{i_2 \neq i_1} C_{i_2}^{(\ell-1)} (1 - \eta_x)^{4-\ell},$$

$$\leq \alpha_{i_1}^{(4)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( \sum_{i_2 \neq i_1} C_{i_2}^{(0)} (1 - \eta_x)^3 + \sum_{i_2 \neq i_1} C_{i_2}^{(1)} (1 - \eta_x)^2 + \sum_{i_2 \neq i_1} C_{i_2}^{(2)} (1 - \eta_x)^3 \right)$$

$$+ \sum_{i_2 \neq i_1} C_{i_2}^{(3)} (1 - \eta_x)^4.$$

Substituting for $C_{i_2}^{(3)}$ from (27), $C_{i_2}^{(2)}$ from (26), $C_{i_2}^{(1)}$ using (25), and rearranging,

$$C_{i_1}^{(4)} \leq \alpha_{i_1}^{(4)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \left( \sum_{i_2 \neq i_1} \alpha_{i_2}^{(3)} + (1 - \eta_x)^1 \sum_{i_2 \neq i_1} \alpha_{i_2}^{(2)} + \eta_{x} \frac{\mu_i}{\sqrt{n}} \sum_{i_3 \neq i_2, i_1} \alpha_{i_3}^{(2)} \right).$$
where

\[ C_i^{(4)} \leq \alpha_i^{(4)} + \eta x \mu x \sqrt{n} \left[ (k-1) \alpha_i^{(3)} + \alpha_i^{(2)} \left( (1 - \eta x)^1 (k - 1) + \eta x \mu x \right) \right] \]

\[ + \alpha_i^{(1)} \left( (1 - \eta x)^2 + 2(1 - \eta x)(1 - \eta x) + (1 - \eta x)^2 \right) \]

\[ + \eta x \mu x C_i^{(0)} \left[ (k-1)(1 - \eta x)^3 + 3(1 - \eta x)^2 \right] + 3(k-3)(1 - \eta x)^3 (1 - \eta x) + (k-4)(1 - \eta x) \sum_{i=1}^{R} \alpha_i^{(0)} \]

Notice that the terms have a binomial series like form. To reveal this structure, let each \( \alpha_i^{(j)} \leq \alpha_{\text{max}}^{(j)} \) where \( j = i_1, i_2, \ldots, i_k \). Similarly, let \( C_j^{(0)} \leq C_{\text{max}}^{(0)} \) for \( j = i_1, i_2, \ldots, i_k \). Therefore, we have

\[ C_i^{(4)} \leq \alpha_i^{(4)} + \eta x \mu x \sqrt{n} \left[ (k-1) \alpha_i^{(3)} + \alpha_i^{(2)} \left( (1 - \eta x)^1 (k - 1) + \eta x \mu x \right) \right] \]

\[ + \alpha_i^{(1)} \left( (1 - \eta x)^2 + 2(1 - \eta x)(1 - \eta x) + (1 - \eta x)^2 \right) \]

\[ + \eta x \mu x C_i^{(0)} \left[ (k-1)(1 - \eta x)^3 + 3(1 - \eta x)^2 \right] + 3(k-3)(1 - \eta x)^3 (1 - \eta x) + (k-4)(1 - \eta x) \sum_{i=1}^{R} \alpha_i^{(0)} \]

Further upper-bounding the expression, we have

\[ C_i^{(4)} \leq \alpha_i^{(4)} + \eta x \mu x \sqrt{n} \left[ \alpha_i^{(3)} + \alpha_i^{(2)} \left( (1 - \eta x) + \eta x \mu x \right) \right] \]

\[ + \alpha_i^{(1)} \left( (1 - \eta x)^2 + 2(1 - \eta x)(1 - \eta x) + (1 - \eta x)^2 \right) \]

\[ + \eta x \mu x C_i^{(0)} \left[ (1 - \eta x)^3 + 3(1 - \eta x)^2 \right] + 3(k-3)(1 - \eta x)^3 (1 - \eta x) + (k-4)(1 - \eta x) \sum_{i=1}^{R} \alpha_i^{(0)} \]

Therefore,

\[ C_i^{(4)} \leq \alpha_i^{(4)} + (k-1) \eta x \mu x \sqrt{n} \left[ \alpha_i^{(3)} + \alpha_i^{(2)} \left( (1 - \eta x) + \eta x \mu x \right) \right] \]

\[ + \alpha_i^{(1)} \left( (1 - \eta x)^2 + 2(1 - \eta x)(1 - \eta x) + (1 - \eta x)^2 \right) \]

\[ + (k-1)(1 - \eta x) \sum_{i=1}^{R} \alpha_i^{(0)} \]

Expression for \( C_i^{(r+1)} \) - With this, we are ready to write the general term,

\[ C_i^{(r+1)} \leq \alpha_i^{(r+1)} + (k-1) \eta x \mu x \sqrt{n} \sum_{\ell=1}^{R} \alpha_i^{(\ell)} (1 - \eta x + \eta x \mu x) \]

\[ + (k-1) \eta x \mu x \sum_{i=1}^{R} \alpha_i^{(0)} (1 - \eta x + \eta x \mu x)^{r+1} \]

Claim 4 (An intermediate result for bounding the error in coefficient calculations). With probability \((1 - \delta_i^{(0)} - \delta_i^{(4)})\),

\[ \sum_{\ell=1}^{R} \alpha_i^{(\ell)} (1 - \eta x + \eta x \mu x)^{R-\ell} \leq C_i^{(0)} R \delta_R + \frac{1}{\eta x (1 - \eta x) \mu x} \left( \frac{\|x^*\|^2}{2} \right) \sum_{i=1}^{R} \alpha_i^{(0)} (1 - \eta x + \eta x \mu x)^{r+1} \]

Proof of Claim 4 Using (17), the quantity \( \alpha_i^{(\ell)} \) is defined as

\[ \alpha_i^{(\ell)} = C_i^{(0)} (1 - \eta x)^{\ell} + (\lambda_i^{(\ell)} |x^*_i| + \beta_i^{(\ell)}) \sum_{s=1}^{\ell} \eta x (1 - \eta x)^{\ell-s+1} \]

28
Therefore, we are interested in
\[
\sum_{\ell=1}^{R} C_i^{(0)} (1 - \eta_x)^{\ell} \left( 1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}} \right)^{R - \ell} 
+ \left( \lambda_i^{(t)} \right| x_i^* | + \beta_i^{(t)} \right) \sum_{\ell=1}^{R} (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})^{R - \ell} \sum_{s=1}^{\ell} \eta_x (1 - \eta_x)^{\ell - s + 1}.
\]

Consider the first term which depends on $C_i^{(0)}$. Since $(1 - \eta_x) \leq (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})$, we have
\[
C_i^{(0)} \sum_{\ell=1}^{R} (1 - \eta_x)^{\ell} (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})^{R - \ell} \leq C_i^{(0)} R \left( 1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}} \right)^R \leq C_i^{(0)} R \delta_R,
\]
where $\delta_R$ is a small constant, and a parameter which determines the number of iterations $R$ required for the coefficient update step. Now, coming back to the quantity of interest
\[
\sum_{\ell=1}^{R} \alpha_i^{(t)} (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})^{R - \ell} \leq C_i^{(0)} R \delta_R
+ \left( \lambda_i^{(t)} \right| x_i^* | + \beta_i^{(t)} \right) \sum_{\ell=1}^{R} (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})^{R - \ell} \sum_{s=1}^{\ell} \eta_x (1 - \eta_x)^{\ell - s + 1}.
\]

Now, using sum of geometric series result, we have that $\sum_{s=1}^{\ell} \eta_x (1 - \eta_x)^{\ell - s + 1}$, and
\[
\sum_{\ell=1}^{R} \left( 1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}} \right)^{R - \ell} = \frac{1 - (1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}})^R}{\eta_x - \eta_x \frac{\mu_r}{\sqrt{n}}} \leq \frac{1}{\eta_x (1 - \eta_x)^R}.
\]

Therefore, with probability at least $(1 - \delta_{\beta}^{(t)})$,
\[
\sum_{\ell=1}^{R} \alpha_i^{(t)} \left( 1 - \eta_x + \eta_x \frac{\mu_r}{\sqrt{n}} \right)^{R - \ell} \leq C_i^{(0)} R \delta_R + \frac{1}{\eta_x (1 - \eta_x)^R} \left( \frac{\sum |x_i^*|^2}{\eta_x} + t_{\beta} \right),
\]
where $\lambda_i^{(t)} \leq \frac{\sum |x_i^*|^2}{\eta_x}$ and $|\beta_i^{(t)}| = t_{\beta}$ with probability at least $(1 - \delta_{\beta}^{(t)})$ using Claim 2. \hfill \Box

**Claim 5 (Bound on the noise term in the estimation of a coefficient element in the support).**

With probability $(1 - \delta_{\beta}^{(t)})$, each entry $\bar{\hat{\theta}}_{i_1}^{(R)}$ of $\hat{\theta}^{(R)}$ is upper-bounded as
\[
|\bar{\hat{\theta}}_{i_1}^{(R)}| \leq O(t_{\beta}).
\]

**Proof of Claim 5.** From (18) $\bar{\hat{\theta}}_{i_1}^{(R)}$ is defined as
\[
\bar{\hat{\theta}}_{i_1}^{(R)} := \sum_{r=1}^{R} \eta_x \xi_{i_1}^{(r)} (1 - \eta_x)^{R - r} + \gamma_{i_1}^{(R)},
\]
where $\gamma_{i_1}^{(R)} := (1 - \eta_x)^R (x_{i_1}^{(0)} - x_i^* (1 - \lambda_{i_1}^{(t)}))$. Further, $\xi_{i_1}^{(r)}$ is as defined in (10),
\[
\xi_{i_1}^{(r)} = \beta_{i_1}^{(t)} + \sum_{i_2 \neq i_1} \|A_{i_1}^{(t)}[A_{i_2}^{(t)}] \text{sign}(A_{i_1}^{(t)}[A_{i_2}^{(t)}])C_{i_1}^{(r-1)} \text{sign}(x_{i_1}^* - x_{i_2}^{(r)})\).
\]

Therefore, we have the following expression for $\bar{\hat{\theta}}_{i_1}^{(R)}$
\[
\bar{\hat{\theta}}_{i_1}^{(R)} = \beta_{i_1}^{(t)} \sum_{r=1}^{R} \eta_x (1 - \eta_x)^{R - r}
+ \sum_{r=1}^{R} \eta_x \sum_{i_2 \neq i_1} (A_{i_1}^{(t)}[A_{i_2}^{(t)}] \text{sign}(A_{i_1}^{(t)}[A_{i_2}^{(t)}])C_{i_1}^{(r-1)} \text{sign}(x_{i_1}^* - x_{i_2}^{(r)}) (1 - \eta_x)^{R - r} + \gamma_{i_1}^{(R)}).
\] (29)
Now \( \delta^{(R)}_{i_1} \) can be upper-bounded as
\[
\delta^{(R)}_{i_1} \leq \beta^{(t)}_{i_1} \sum_{r=1}^{R} \eta_x (1 - \eta_x)^{R-r} + \eta_x \frac{\mu}{\sqrt{n}} \sum_{r=1}^{R} \sum_{i_2 \neq i_1} C^{(R-1)}_{i_2} (1 - \eta_x)^{R-r} + \gamma_{i_1}^{(R)},
\]
\[
\leq \beta^{(t)}_{i_1} + (k-1) \eta_x \frac{\mu}{\sqrt{n}} \sum_{r=1}^{R} C^{(R-1)}_{i_2} (1 - \eta_x)^{R-r} + \gamma_{i_1}^{(R)}.
\]

Since from Claim 6 we have
\[
C^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} \leq (\lambda_{\max}^{(t)} |x_{\max}^*| + \beta^{(t)}_{\max}) \left[ \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{R-s} + k_c x (1 - \eta_x)^{R-r} \right] + k \eta_x \frac{\mu}{\sqrt{n}} C^{(0)}_{\max} \delta R - 2.
\]

Further, since \( 1 - (1 - \eta_x)^{r-1} \leq 1 \), we have that
\[
\sum_{r=1}^{R} \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{R-s} = \sum_{r=1}^{R} \eta_x (1 - \eta_x)^{R-r+1} \frac{1 - (1 - \eta_x)^{r-1}}{\eta_x} \leq \sum_{r=1}^{R} (1 - \eta_x)^{R-r+1} \leq \frac{1}{\eta_x}.
\]

Therefore,
\[
|\delta^{(R)}_{i_1}| \leq |\beta^{(t)}_{i_1}| + (k-1) \eta_x \frac{\mu}{\sqrt{n}} (\lambda_{\max}^{(t)} |x_{\max}^*| + \beta^{(t)}_{\max}) (1 + k_c x) + \left( k \eta_x \frac{\mu}{\sqrt{n}} \right)^2 R C^{(0)}_{\max} \delta R - 2 + \gamma_{i_1}^{(R)}.
\]

Now, since each \( |\beta^{(t)}_{i_1}| = t_\beta \) with probability at least \( (1 - \delta^{(t)}_{\beta}) \) for the \( t \)-th iterate, and \( k = O^*(\frac{\sqrt{R}}{\mu \log(n)}) \), therefore \( k c_x < 1 \), we have that
\[
|\delta^{(R)}_{i_1}| \leq O(t_\beta).
\]

with probability at least \( (1 - \delta^{(t)}_{\beta}) \).

\( \square \)

**Claim 6 (An intermediate result for \( \delta^{(R)}_{i_1} \) calculations).** For \( c_x = \frac{\mu}{\sqrt{n}} / (1 - \frac{\mu}{\sqrt{n}}) \), we have
\[
C^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} \leq (\lambda_{\max}^{(t)} |x_{\max}^*| + \beta^{(t)}_{\max}) \left[ \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{R-s} + k_c x (1 - \eta_x)^{R-r} \right] + k \eta_x \frac{\mu}{\sqrt{n}} C^{(0)}_{\max} \delta R - 2.
\]

**Proof of Claim 6.** Here, from Claim 3 we have that for any \( i_1 \),
\[
C^{(r+1)}_{i_1} \leq \alpha^{(r+1)}_{i_1} + k \eta_x \frac{\mu}{\sqrt{n}} \sum_{r=1}^{R} \alpha^{(t)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r} + k \eta_x \frac{\mu}{\sqrt{n}} C^{(0)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r}.
\]

Therefore, \( C^{(r-1)}_{i_2} \) is given by
\[
C^{(r-1)}_{i_2} \leq \alpha^{(r-1)}_{i_2} + k \eta_x \frac{\mu}{\sqrt{n}} \sum_{r=2}^{R} \alpha^{(t)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r+1} + k \eta_x \frac{\mu}{\sqrt{n}} C^{(0)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r+1}.
\]

Further, the term of interest \( C^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} \) can be upper-bounded by
\[
C^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} \leq \alpha^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} + (1 - \eta_x)^{R-r} k \eta_x \frac{\mu}{\sqrt{n}} \sum_{r=2}^{R} \alpha^{(t)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r+1} + k \eta_x \frac{\mu}{\sqrt{n}} C^{(0)}_{\max} (1 - \eta_x + \eta_x \frac{\mu}{\sqrt{n}})^{R-r+1}.
\]

From the definition of \( \alpha^{(t)}_{i_2} \) from (17), \( \alpha^{(r-1)}_{i_2} \) can be written as
\[
\alpha^{(r-1)}_{i_2} = C^{(0)}_{\max} (1 - \eta_x)^{r-1} + (\lambda_{\max}^{(t)} |x_{\max}^*| + \beta^{(t)}_{\max}) \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{r-s}.
\]

Therefore, we have
\[
\alpha^{(r-1)}_{i_2} (1 - \eta_x)^{R-r} = C^{(0)}_{\max} (1 - \eta_x)^{R-1} + (\lambda_{\max}^{(t)} |x_{\max}^*| + \beta^{(t)}_{\max}) \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{R-s}.
\]
Next, to get a handle on $\alpha_{\max}^{(r)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell-2}$, consider the following using the definition of $\alpha_{\max}^{(r)}$ from (17), where $\eta_x = \eta_x$ for all $i$,

$$
\sum_{\ell=1}^{r} \alpha_{\max}^{(r)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell} = \sum_{\ell=1}^{r} \alpha_{\max}^{(0)}(1 - \eta_x)^{r-\ell} + (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)}) \sum_{\ell=1}^{r} \alpha_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell} \sum_{s=1}^{\ell} \eta_x (1 - \eta_x)^{\ell-s+1},
$$

and

$$
\leq \sum_{\ell=1}^{r} \alpha_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r} + (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)}) \sum_{\ell=1}^{r} (1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell}.
$$

Therefore,

$$
(1 - \eta_x)^{R-r} \sum_{\ell=1}^{r} \alpha_{\max}^{(r)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell-2} \leq \sum_{\ell=1}^{r} \alpha_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-2} (1 - \eta_x)^{R-r}
$$

$$
+ (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)}) (1 - \eta_x)^{R-r} \sum_{\ell=1}^{r} \alpha_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell-2},
$$

$$
\leq (R - 2) C_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{R-2} + (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)}(1 - \eta_x)^{R-r} \frac{1}{\eta_x(1 - \eta_x \sqrt{\frac{\mu_x}{n}})}.
$$

Therefore,

$$
(1 - \eta_x)^{R-r} \sum_{\ell=1}^{r} \alpha_{\max}^{(r)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{r-\ell-2} \leq (r - 2) C_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{R-2} + (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)}(1 - \eta_x)^{R-r} \frac{1}{\eta_x(1 - \eta_x \sqrt{\frac{\mu_x}{n}})}.
$$

Furthermore, since $\eta_x \leq (1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}}),

$$
k_{\eta_x} \mu_x \sqrt{\frac{\mu_x}{n}} C_{\max}^{(0)}(1 - \eta_x + \eta_x \sqrt{\frac{\mu_x}{n}})^{R-2} (1 - \eta_x)^{R-r} \leq k_{\eta_x} \mu_x \sqrt{\frac{\mu_x}{n}} C_{\max}^{(0)} \frac{1}{\eta_x(1 - \eta_x \sqrt{\frac{\mu_x}{n}})}.
$$

Therefore, combining all the results we have that, for a constant $c_{x} = \frac{\mu_x}{\sqrt{n}}/(1 - \frac{\mu_x}{\sqrt{n}}),

$$
C_{x} (1 - \eta_x)^{R-r}
$$

$$
\leq (\lambda_{\max}^{(r)}|x_{\max}^{*} + \beta_{\max}^{(r)} \sum_{s=1}^{r-1} \eta_x (1 - \eta_x)^{R-s} + k_{c_{x}} (1 - \eta_x)^{R-r}) + k_{\eta_x} \mu_x \sqrt{\frac{\mu_x}{n}} C_{\max}^{(0)} \frac{1}{\eta_x(1 - \eta_x \sqrt{\frac{\mu_x}{n}})}.
$$

\begin{proof}

Claim 7 (Bound on the noise term in expected gradient vector estimate). $\|\Delta^{(t)}\|$ where $\Delta^{(t)} := E[A^{(t)}\vartheta^{(R)}]\text{sign}(x_{j}^{*})]$ is upper-bounded as,

$$
\|\Delta^{(t)}\| = O(\sqrt{m} q_{i,j} \epsilon_{j} \|A^{(t)}\|).
$$

Proof of Claim 7.

$$
\Delta^{(t)} = E[A^{(t)}\vartheta^{(R)}]\text{sign}(x_{j}^{*})] = E_{S}[A_{S}^{(t)}E_{x_{S}}^{(t)}[\vartheta^{(R)}_{S}\text{sign}(x_{j}^{*})]|S]]
$$

From (29) we have the following definition for $\vartheta^{(R)}_{j}$

$$
\vartheta^{(R)}_{j} = \beta_{j}^{(t)} + \sum_{i=1}^{R} \eta_{x} \sum_{j \neq j} \langle \langle A^{(t)}_{j}, A^{(t)}_{i} \rangle, \text{sign}(\langle A^{(t)}_{j}, A^{(t)}_{i} \rangle)C_{i}^{(t)}(1 - \eta_{x})^{R-r} + \gamma_{j}^{(R)}
$$

where $\beta_{j}^{(t)}$ is defined as the following (11)

$$
\beta_{j}^{(t)} = \sum_{i \neq j} \langle A_{j}^{(t)}, A_{i}^{(t)} \rangle + \langle A_{j}^{(t)} - A_{j}^{(t)}, A_{i}^{(t)} \rangle x_{j}^{*} + \sum_{j \neq \tilde{j}} \langle A_{j}^{(t)} - A_{j}^{(t)}, A_{i}^{(t)} \rangle x_{j}^{*}.$$

\end{proof}
Consider $E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S]$, where $\vartheta_i^{(R)}$ is a vector with each element as defined in (29). Therefore, the elements of the vector $E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S]$ are given by

$$E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S] = \begin{cases} E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S], & \text{for } i \neq j, \\
E_{\mathbf{x}^*_j} [\vartheta_j^{(R)} \text{sign}(\mathbf{x}^*_j)] | S], & \text{for } i = j. \end{cases}$$

Consider the general term of interest

$$E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S]$$

$$\leq \sum_{r = 1}^R \eta_x (1 - \eta_x)^{R-r} E_{\mathbf{x}^*_j} [\beta_i \text{sign}(\mathbf{x}^*_j)] | S]$$

$$+ \frac{\mu_x}{\sqrt{n}} \sum_{r = 1}^R \eta_x (1 - \eta_x)^{R-r} \sum_{s \neq i} E_{\mathbf{x}^*_j} [C^{(r-1)} \text{sign}(\mathbf{x}^*_j - \mathbf{x}_s^{(r)}) \text{sign}(\mathbf{x}^*_j)] | S] + \gamma_i^{(R)}.$$

Further, since

$$E_{\mathbf{x}^*_j} [\mathbf{x}^*_j \text{sign}(\mathbf{x}^*_j)] | S] = \begin{cases} 0, & \text{for } i \neq j, \\
\rho_j, & \text{for } i = j, \end{cases}$$

we have that

$$\mathbf{a} := E_{\mathbf{x}^*_j} [\beta_i^{(t)} \text{sign}(\mathbf{x}^*_j)] | S] \leq \begin{cases} 3\rho_j \epsilon_t, & \text{for } i \neq j, \\
0, & \text{for } i = j. \end{cases}$$

(30)

Further, for $\mathbf{a}_s := E_{\mathbf{x}^*_j} [C^{(r-1)} \text{sign}(\mathbf{x}^*_j - \mathbf{x}_s^{(r)}) \text{sign}(\mathbf{x}^*_j)] | S]$ we have that

$$\mathbf{a}_s = \begin{cases} E_{\mathbf{x}^*_j} [C^{(r-1)} \text{sign}(\mathbf{x}^*_j - \mathbf{x}_s^{(r)}) \text{sign}(\mathbf{x}^*_j)] | S] \leq C_j^{(r-1)}, & \text{for } s = j, \\
0, & \text{for } s \neq j. \end{cases}$$

In addition, for $\sum_{s \neq i} \mathbf{a}_s$ we have that

$$\sum_{s \neq i} \mathbf{a}_s = \begin{cases} C_j^{(r-1)}, & \text{for } i \neq j, \\
0, & \text{for } i = j. \end{cases}$$

(31)

Therefore, using the results for $\mathbf{a}$ and $\sum_{s \neq i} \mathbf{a}_s$, we have that $E_{\mathbf{x}^*_j} [\vartheta_j^{(R)} \text{sign}(\mathbf{x}^*_j)] | S] = \gamma_i^{(R)}$ for $i = j$, and for $i \neq j$ we have

$$E_{\mathbf{x}^*_j} [\vartheta_i^{(R)} \text{sign}(\mathbf{x}^*_j)] | S]$$

$$\leq 3\rho_j \epsilon_t + \frac{\mu_x}{\sqrt{n}} \sum_{r = 1}^R E_{\mathbf{x}^*_j} [C_j^{(r-1)} \text{sign}(\mathbf{x}^*_j - \mathbf{x}_s^{(r)}) \text{sign}(\mathbf{x}^*_j)] | S] \eta_x (1 - \eta_x)^{R-r} + \gamma_i^{(R)},$$

$$\leq 3\rho_j \epsilon_t + \frac{\mu_x}{\sqrt{n}} \sum_{r = 1}^R C_j^{(r-1)} \eta_x (1 - \eta_x)^{R-r} + \gamma_i^{(R)}.$$  

(32)

Here, from Claim 6, for $\epsilon_x = \frac{\mu_x}{\sqrt{n}} / (1 - \frac{\mu_x}{\sqrt{n}})$ we have

$$C_j^{(r-1)} (1 - \eta_x)^{R-r}$$

$$\leq \left( \lambda_{\text{max}}^{(t)} \mathbf{x}^*_j + \beta^{(t)} \right) \sum_{s = 1}^r \eta_x (1 - \eta_x)^{R-s} + k \epsilon_x (1 - \eta_x)^{R-r} + k \eta_x \frac{\mu_x}{\sqrt{n}} \mathbf{C}^{(0)} \delta_{R-2}.$$

Further, due to our assumptions on sparsity, $k \epsilon_x \leq 1$; in addition from Claim 2, and with probability at least $(1 - \delta^{(t)})$ we have $|\beta^{(t)}| \leq t^{\beta}$, substituting,

$$\sum_{r = 1}^R C_j^{(r-1)} \eta_x (1 - \eta_x)^{R-r}$$

32
result, we can write

$$\|\hat{x}_t\| = \|x_t^* + \hat{\theta}_t^R\|,$$

where

$$\hat{x}_t := x_t^* (1 - \lambda_t) + \hat{\theta}_t^R.$$

Therefore, the vector \(\hat{x}_S\), for \(S \in \text{supp}(x^*)\) can be written as

$$\hat{x}_S := x_S^* = (I - \Lambda_S^*(t))x_S^* + \hat{\theta}_S^R,$$

where \(\hat{x}\) has the correct signed-support with probability at least \((1 - \delta_T)\) using Lemma 2. Using this result, we can write \(\|y - A(t)\hat{x}\|\) as

$$\|y - A(t)\hat{x}\| = \|A_S^*x_S^* - (I - \Lambda_S^*(t))x_S^* - \Lambda_S^*(t)\hat{\theta}_S^R\|.$$ 

Now, since \(\Lambda_S^*(t) \leq \frac{c^2}{2}\) we have

$$\|y - A(t)\hat{x}\| \leq \|A_S^*x_S^* - (1 - \frac{c^2}{2})A_S^*(t)x_S^* - \Lambda_S^*(t)\hat{\theta}_S^R\|,$$

$$\leq \left\| (1 - \frac{c^2}{2})(A_S^* - \Lambda_S^*(t))x_S^* \right\| + \left\| \frac{c^2}{2}A_S^* - \Lambda_S^*(t)\hat{\theta}_S^R \right\|.$$ 

WITH \(x_S^*\) being independent and sub-Gaussian, using Lemma 13, which is a result based on the Hanson-Wright result (Hanson and Wright, 1971) for sub-Gaussian random variables, and since

$$\|A_S^*(t) - A_S^*\| \leq \|A_S^* - A_S^*\|_F \leq \sqrt{k}\epsilon_t,$$

we have that with probability at least \((1 - \delta_{HW}^*)\)

$$\|\Delta_S^*\|^2 \leq \|((1 - \frac{c^2}{2})(A_S^* - \Lambda_S^*(t)) + \frac{c^2}{2}A_S^*)x_S^*\| \leq \tilde{O}((1 - \frac{c^2}{2})(A_S^* - \Lambda_S^*(t)) + \frac{c^2}{2}A_S^*)_F,$$

where \(\delta_{HW}^* = \exp(-\frac{1}{\tilde{O}(\epsilon_t)}).\)

Now, consider the \(|\Delta|\|_F\), since \(\|A_S^*(t) - A_S^*\|_F \leq \sqrt{k}\epsilon_t\)

$$\|\Delta\|_F := \|((1 - \frac{c^2}{2})(A_S^* - \Lambda_S^*(t)) + \frac{c^2}{2}A_S^*)_F \leq \|((1 - \frac{c^2}{2})(A_S^* - \Lambda_S^*(t)) + \frac{c^2}{2}A_S^*)_F, \leq \sqrt{k}(1 - \frac{c^2}{2})\epsilon_t + \frac{c^2}{2}\|A_S^*\|_F.$$

Consider the \(|\Delta|\| term. Using Claim 5, each \(\hat{\theta}_j^R\) is bounded by \(O(t_\beta)\), with probability at least \((1 - \delta_\beta^*)\) Therefore,

$$\|\Delta\| = \|A_S^*(t)\hat{\theta}_S^R\| \leq \|A_S^*\|\|\hat{\theta}_S^R\| = \|A_S^*\|\sqrt{k}O(t_\beta).$$
Again, since $\|A_S^{(t)} - A_S\| \leq \|A_S^{(t)} - A_S^*\|_F \leq \sqrt{k}\epsilon_t,$

\[
\|A_S^{(t)}\| \leq \|A_S^{(t)} - A_S^* + A_S^*\| \leq \|A_S^{(t)} - A_S^*\| + \|A_S^*\| \leq \sqrt{k}\epsilon_t + 2.
\]

Finally, combining all the results and using the fact that $\|A_S^*\|_F \leq \sqrt{k}\|A_S^*\| \leq 2\sqrt{k},$

\[
\|y - A^{(t)}\bar{x}\| = \tilde{O}(\sqrt{k}(1 - \frac{1}{2}\epsilon_t + \epsilon_t^2)\sqrt{k}) + \|A_S^{(t)}\|\sqrt{k}O(t_\beta),
\]

\[
= \tilde{O}(kt_\beta).
\]

\[\square\]

**Claim 9 (Bound on variance parameter for concentration of gradient vector).** For $z := (y - A^{(t)}\bar{x})\text{sign}(\bar{x})|i \in S$ the variance parameter $E[\|z\|^2]$ is bounded as $E[\|z\|^2] = O(k\epsilon_t^2) + O(kt_\beta^2)$ with probability at least $1 - \delta^{(t)} - \delta^{(R)}$.

**Proof of Claim 9.** For the variance $E[\|z\|^2]$, we focus on the following,

\[
E[\|z\|^2] = E[\|(y - A^{(t)}\bar{x})\text{sign}(\bar{x})\|^2| i \in S].
\]

Here, $\bar{x}_S$ is given by

\[
\bar{x}_S = (I - \Lambda_S^{(t)})x^*_S + \vartheta^{(R)}_S.
\]

Therefore, $E[\|z\|^2]$ can be written as

\[
E[\|(y - A^{(t)}\bar{x})\text{sign}(\bar{x})\|^2| i \in S] = E[\|A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})\|^2| i \in S].
\]

We will now consider each term in (34) separately. We start with $\vartriangle$. Since $x^*_S$ are conditionally independent of $S$, $E[x_S^*x_S^{*\top}] = I$. Therefore, we can simplify this expression as

\[
\vartriangle := E[\|A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})\|^2| i \in S] = E[\|A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})\|^2| i \in S].
\]

Rearranging the terms we have the following for $\vartriangle$,

\[
\vartriangle = E[\|A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})\|^2| i \in S] = E[\|A_S^*A_S^{(t)} - (A_S^* - A_S^{(t)})(I - \Lambda_S^{(t)})\|^2| i \in S].
\]

Therefore, $\vartriangle$ can be upper-bounded as

\[
\vartriangle \leq E[\|A_S^*A_S^{(t)}\|^2| i \in S] + E[\|(A_S^* - A_S^{(t)})(I - \Lambda_S^{(t)})\|^2| i \in S]
\]

\[
\leq E[\|A_S^*A_S^{(t)}\|^2| i \in S] + 2E[\|A_S^* - A_S^{(t)}\|_F^2| \|A_S^* - A_S^{(t)}\|_F| i \in S].
\]

For $\vartriangle_1$, since $\|A_S^{(t)}\| \leq \sqrt{k}\epsilon_t + 2$, we have

\[
\vartriangle_1 := E[\|A_S^*A_S^{(t)}\|^2| i \in S] \leq E[\|A_S^*\|\|A_S^{(t)}\|^2| i \in S] \leq \|A_S^*\| \sum_{j \in S}(\lambda_j^{(t)})^2 \leq k(\sqrt{k}\epsilon_t + 2)^2\epsilon_t^2.
\]

Next, since $(1 - \lambda_j^{(t)}) \leq 1$, we have the following bound for $\vartriangle_2$

\[
\vartriangle_2 := E[\|(A_S^* - A_S^{(t)})(I - \Lambda_S^{(t)})\|^2| i \in S] \leq E[\|A_S^* - A_S^{(t)}\|^2| i \in S] \leq \|A_S^* - A_S^{(t)}\|^2 \leq k\epsilon_t^2.
\]

Further, $\vartriangle_3$ can be upper-bounded by using bounds for $\vartriangle_1$ and $\vartriangle_2$. Combining the results of upper-bounding $\vartriangle_1$, $\vartriangle_2$, and $\vartriangle_3$ we have the following for (35)

\[
\vartriangle \leq E[\|A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})\|^2| i \in S] = O(k\epsilon_t^2).
\]
Next, by Claim 5, \( \vartheta_j^{(R)} \) is upper-bounded as \(|\vartheta_j^{(R)}| \leq O(t_\beta) \). Therefore, the term \( \diamond \), the second term of (34), can be bounded as

\[
\diamond \leq \|A_S^{(t)} \vartheta_S^{(R)} \operatorname{sign}(\tilde{x}_i)\|^2 \leq (\sqrt{k} \epsilon_\ell + 2)^2 kO(t_\beta)^2 = O(k^2 t_\beta^2).
\]

Finally, combining all the results, the term of interest in (34) has the following form

\[
E[\| (y - A^{(t)} \tilde{x}) \operatorname{sign}(\tilde{x}_i) \|^2 | i \in S] = O(k \epsilon_\ell^2) + O(k t_\beta^2).
\]

Claim 10 (Bound on variance parameter for concentration of gradient matrix). With probability \((1 - \delta_T^{(t)} - \delta_\beta^{(t)})\), the variance parameter \(\|E[(y - A^{(t)} \tilde{x})(y - A \tilde{x})^\top]\|\) is upper-bounded by \(O(\frac{k^2 t_\beta^2}{m})\|A^*\|^2\).

Proof of Claim 10. Let \( \mathcal{F}_{x^*} \) be the event that \(\operatorname{sign}(x^*) = \operatorname{sign}(\tilde{x})\), and let \(1_{\mathcal{F}_{x^*}} \) denote the indicator function corresponding to this event. As we show in Lemma 2, this event occurs with probability at least \((1 - \delta_T^{(t)} - \delta_\beta^{(t)})\), therefore,

\[
E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top] = E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top 1_{\mathcal{F}_{x^*}}] + E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top 1_{\mathcal{F}_{x^*}}^c],
\]

Here, \( \gamma \) is small. Under the event \( \mathcal{F}_{x^*} \), \( \tilde{x} \) has the correct signed-support. Again, since \(1_{\mathcal{F}_{x^*}} = 1 - 1_{\mathcal{F}_{x^*}}^c\),

\[
E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top] = E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top (1 - 1_{\mathcal{F}_{x^*}}^c)] \pm \gamma,
\]

Now, using Lemma 4 with probability at least \((1 - \delta_T^{(t)} - \delta_\beta^{(t)})\), \( \tilde{x}_S \) admits the following expression

\[
\tilde{x}_S := x_S^{(R)} = (I - \Lambda_S^{(t)})x_S^* + \vartheta_S^{(R)}.
\]

Therefore we have

\[
y - A^{(t)} \tilde{x} = (A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)}))x_S^* - A_S^{(t)} \vartheta_S^{(R)}.
\]

Using the expression above \(E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top]\) can be written as

\[
E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top] = E[((A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)}))x_S^* - A_S^{(t)} \vartheta_S^{(R)})(A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)}))x_S^* - A_S^{(t)} \vartheta_S^{(R)})^\top].
\]

Sub-conditioning, we have

\[
E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top] = E_S[A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)})]E_{x_S^*} [x_S^*x_S^*^\top | S][A_S^* - (I - \Lambda_S^{(t)})A_S^{(t)}^\top] - E_S[A_S^{(t)}E_{x_S^*} [\vartheta_S^{(R)}x_S^*^\top | S][A_S^* - (I - \Lambda_S^{(t)})A_S^{(t)}^\top] - E_S[(A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)}))E_{x_S^*} [x_S^* \vartheta_S^{(R)}] | S][A_S^{(t)}] + E_S[A_S^{(t)}E_{x_S^*} [\vartheta_S^{(R)} \vartheta_S^{(R)}] | S][A_S^{(t)}].
\]

Now, since \( E_{x_S^*} [x_S^*x_S^*^\top | S] = I \),

\[
\|E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^\top]\| \leq \|E_S[(A_S^* - A_S^{(t)}(I - \Lambda_S^{(t)}))A_S^* - (I - \Lambda_S^{(t)})A_S^{(t)}^\top]\|.
\]
Therefore, since we have the following, in (37), which can be simplified as

\[
1 \times (37) \text{ which can also be bounded similarly as }
\]

\[
\]

Similarly,

\[
\]

Now consider each term of equation (37). First, since \( \|A_{s}^{(t)}(I - \Lambda_{s}^{(t)})A_{s}^{(t)\top}\| \) in (37), which can be simplified as

\[
\]

Let’s start with the first term (\( \mathcal{A}_{1} \)) of (36), which can be written as

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

Further, since \( A^{(t)} \) is \((\epsilon_{t}, 2)\)-close to \( A^{*} \), we have that \( \|A^{(t)}\| \leq \|A^{(t)} - A^{*}\| + \|A^{*}\| \leq 3\|A^{*}\| \), therefore

\[
\]

Similarly, (37) is also \( O(\frac{k^{2}}{m})\|A^{*}\|^{2} \). Next, we consider \( \mathcal{A}_{3} := \|E_{S}[A_{s}^{(t)}(I - \Lambda_{s}^{(t)})^{2}A_{s}^{(t)\top}]\| \) in (37) which can also be bounded similarly as

\[
\]

Therefore, we have the following for \( \mathcal{A} \) in (36)

\[
\]

Consider \( \mathcal{A} \) in (36). Letting \( M = E_{x_{S}}[\vartheta_{S}^{(R)}x_{S}^{\top} | S] \), and using the analysis similar to that shown in 7, we have that elements of \( M \in \mathbb{R}^{k \times k} \) are given by

\[
M_{i,j} = E_{x_{S}}[\vartheta_{i}^{(R)}x_{j}^{\top} | S] \begin{cases} \leq O(\gamma_{i}^{(R)}), & \text{for } i = j, \\ = O(\epsilon_{i}), & \text{for } i \neq j. \end{cases}
\]

We have the following,

\[
\]

Therefore, since \( E_{S}[I_{i,j \in S}] = O(\frac{k^{2}}{m}) \), and \( \|1^{m \times m}\| = m \),

\[
\]

36
\[ \| \frac{m}{i,j=1} M_{i,j} A_i^{(t)} (A_j^{(t)})^T - (1 - \lambda_j^{(t)}) A_j^{(t)} \|_{E_S[I_{i,j} \in S]} \| \\
= O(\epsilon_l) \| \frac{m}{i,j=1} A_i^{(t)} (A_j^{(t)})^T - (1 - \lambda_j^{(t)}) A_j^{(t)} \|_{E_S[I_{i,j} \in S]} \| \\
= O(\epsilon_l) O(\frac{k^2}{m}) (\| A^{(t)} 1_{m \times m} A^* \| + \| A^{(t)} 1_{m \times m} A^{(t)} \|) \\
= O(\epsilon_l) O(\frac{k^2}{m}) \| A^* \|^2. \]

Therefore,
\[ \mathbf{\blacklozenge} := \| E_S[A_S^{(t)} M(A_S^{(t)})^T - (I - A_S^{(t)}) A_S^{(t)}] \| = O(\frac{k^2}{m}) \epsilon_l \| A^* \|^2. \]

Similarly, \[ \heartsuit \] in (36) is also bounded as \[ \mathbf{\blacklozenge}. \] Next, we consider \[ \heartsuit \] in (36). In this case, letting \[ E_S[\hat{\theta}_j^{(R)} (\hat{\theta}_j^{(R)})^T] \| S \| = N, \] where \( N \in \mathbb{R}^{k \times k} \) is a matrix whose each entry \( N_{i,j} \leq \| \hat{\theta}_j^{(R)} \| \| \hat{\theta}_j^{(R)} \| . \)

Further, by Claim 5, each element \( \hat{\theta}_j^{(R)} \) is upper-bounded as
\[ | \hat{\theta}_j^{(R)} | \leq \mathcal{O}(t_{\beta}). \]

with probability at least \( (1 - \delta_j^{(t)}) \).

Therefore,
\[ \heartsuit := \| \frac{m}{i,j=1} N_{i,j} A_i^{(t)} A_j^{(t)}^T E_S[I_{i,j} \in S] \| = \max_{i,j} | \hat{\theta}_i^{(R)} | | \hat{\theta}_j^{(R)} | O(\frac{k^2}{m^2}) \sum_{i,j=1}^m A_i^{(t)} A_j^{(t)}^T. \]

Again, using the result on \( | \hat{\theta}_i^{(R)} | \), we have
\[ \diamond := \| E_S[A_S^{(t)} N A_S^{(t)}] \| = m \max_{i,j} | \hat{\theta}_i^{(R)} | | \hat{\theta}_j^{(R)} | O(\frac{k^2}{m^2}) \| A^{(t)} \| | A^{(t)} \| = O(\frac{k^2}{m^2}) \| A^* \|^2. \]

Combining all the results for \( \mathbf{\blacklozenge} \), \( \heartsuit \), \( \spadesuit \) and \( \diamond \), we have,
\[ \| E[(y - A^{(t)} \tilde{x})(y - A^{(t)} \tilde{x})^T] \| \\
= O(\frac{k^2}{m}) \| A^* \|^2 + O(\frac{k^2}{m}) \epsilon_l \| A^* \|^2 + O(\frac{k^2}{m}) \epsilon_l \| A^* \|^2 + O(\frac{k^2}{m}) \| A^* \|^2, \]
\[ = O(\frac{k^2}{m}) \| A^* \|^2. \]

\[ \square \]

### E Additional Experimental Results

We now present some additional results to highlight the features of NOODL. Here, we compare the performance of NOODL (for both dictionary and coefficient recovery) with the state-of-the-art provable techniques for DL (when the coefficients are recovered via a sparse approximation step after DL).

Table 4 summarizes the results of the convergence analysis shown in Fig. 2. Here, we compare the dictionary and coefficient recovery performance of NOODL with other techniques. For Arora15 (‘unbiased’) and Arora15 (‘unbiased’), we report the error in recovered coefficients after the HT step (\( X_{HT} \)) and the best error via sparse approximation using Lasso (\( X_{Lasso} \)) by scanning over 50 values of regularization parameters.

We observe that NOODL exhibits significantly superior performance across the board. Also, we observe that using sparse approximation after dictionary recovery, when the dictionary suffers from a bias, leads to poor coefficient recovery. This highlights the applicability of our approach in real-world machine learning tasks where coefficient recovery is of interest. In fact, it is a testament to the fact that, even in cases where dictionary recovery is the primary goal, making progress on the coefficients is also important for dictionary recovery.

In addition, the coefficient estimation is also online in case of NOODL, while for the state-of-the-art provable techniques (which only recover the dictionary and incur bias in estimation) need additional sparse approximation step for coefficient recovery. Moreover, these sparse approximation techniques (such as Lasso) are expensive to use in practice, and need significant tuning.
Table 4: Final error in recovery of the dictionary and the coefficients by various techniques corresponding to Fig. 2. We report the coefficient estimate after the HT step as per Arora’15 as XHT. Also, we scan across 50 values of the regularization parameter for coefficient estimation using Lasso after learning the dictionary (A), and report the optimal estimation error for the coefficients (XLasso). For k = 100, Arora’15 algorithms do not converge (shown as N/A).

<table>
<thead>
<tr>
<th>Technique</th>
<th>Recovered Factor</th>
<th>k = 10</th>
<th>k = 20</th>
<th>k = 50</th>
<th>k = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOODL</td>
<td>A</td>
<td>9.44 × 10^{-11}</td>
<td>8.82 × 10^{-11}</td>
<td>9.70 × 10^{-11}</td>
<td>7.53 × 10^{-11}</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>1.14 × 10^{-11}</td>
<td>1.76 × 10^{-11}</td>
<td>3.58 × 10^{-11}</td>
<td>4.74 × 10^{-11}</td>
</tr>
<tr>
<td>Arora’15 (unbiased)</td>
<td>A</td>
<td>0.011</td>
<td>0.027</td>
<td>0.148</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>XHT</td>
<td>0.078</td>
<td>0.122</td>
<td>0.371</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>XLasso</td>
<td>0.005</td>
<td>0.015</td>
<td>0.0921</td>
<td>N/A</td>
</tr>
<tr>
<td>Arora’15 (biased)</td>
<td>A</td>
<td>0.013</td>
<td>0.031</td>
<td>0.137</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>XHT</td>
<td>0.077</td>
<td>0.120</td>
<td>0.308</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>XLasso</td>
<td>0.006</td>
<td>0.018</td>
<td>0.097</td>
<td>N/A</td>
</tr>
</tbody>
</table>

F Appendix: Standard Results

Definition 6 (sub-Gaussian Random variable). Let \( x \sim \text{subGaussian}(\sigma^2) \). Then, for any \( t > 0 \), it holds that

\[
\Pr[|x| > t] \leq 2 \exp \left( \frac{-t^2}{2\sigma^2} \right).
\]

F.1 Concentration results

Lemma 10 (Matrix Bernstein (Tropp, 2015)). Consider a finite sequence \( W_k \in \mathbb{R}^{n \times m} \) of independent, random, centered matrices with dimension \( n \). Assume that each random matrix satisfies \( \mathbb{E}[W_k] = 0 \) and \( \|W_k\| \leq R \) almost surely. Then, for all \( t \geq 0 \),

\[
\Pr\{\|\sum_k W_k\| \geq t\} \leq (n + m) \exp\left(\frac{-t^2/2}{\sigma^2 + RT/3}\right),
\]

where \( \sigma^2 := \max\{\|\sum_k \mathbb{E}[W_k W_k^T]\|, \|\sum_k \mathbb{E}[W_k W_k]\|\} \).

Furthermore,

\[
\mathbb{E}[\|\sum_k W_k\|] \leq \sqrt{2\sigma^2 \log(n + m) + \frac{1}{2}R \log(n + m)}.
\]

Lemma 11 (Vector Bernstein (Tropp, 2015)). Consider a finite sequence \( w_k \in \mathbb{R}^n \) of independent, random, zero mean vectors with dimension \( n \). Assume that each random vector satisfies \( \mathbb{E}[w_k] = 0 \) and \( \|w_k\| \leq R \) almost surely. Then, for all \( t \geq 0 \),

\[
\Pr\{\|\sum_k w_k\| \geq t\} \leq 2n \exp\left(\frac{-t^2/2}{\sigma^2 + RT/3}\right),
\]

where \( \sigma^2 := \|\sum_k \mathbb{E}[\|w_k\|^2]\| \). Furthermore,

\[
\mathbb{E}[\|\sum_k w_k\|] \leq \sqrt{2\sigma^2 \log(2n) + \frac{1}{2}R \log(2n)}.
\]

Lemma 12. Chernoff Bound for sub-Gaussian Random Variables Let \( w \) be an independent sub-Gaussian random variables with variance parameter \( \sigma^2 \), then for any \( t > 0 \) it holds that

\[
\Pr[|w| > t] \leq 2 \exp(-\frac{t^2}{2\sigma^2}).
\]

Lemma 13 (Sub-Gaussian concentration (Rudelson and Vershynin, 2013)). Let \( M \in \mathbb{R}^{n \times m} \) be a fixed matrix. Let \( w \) be a vector of independent, sub-Gaussian random variables with mean zero and variance one. Then, for an absolute constant \( c \),

\[
\Pr[\|Mw\|_2 - \|M\|_F > t] \leq \exp(-\frac{t^2}{\|M\|_F^2}).
\]
F.2 RESULTS FROM (ARORA ET AL., 2015)

Lemma 14 ((Arora et al., 2015) Lemma 45). Suppose that the distribution of $Z$ satisfies $\Pr[\|Z\| \geq L(\log(1/\rho))^C] \leq \rho$ for some constant $C > 0$, then

1. If $p = n^{O(1)}$ then $\|Z^{(j)}\| \leq \tilde{O}(L)$ holds for each $j$ with probability at least $(1 - \rho)$ and,
2. $\|E[Z_1|_{\|Z\| \geq \tilde{O}(L)}]\| = n^{-\omega(1)}$.

In particular, if $\frac{1}{p} \sum_{j=1}^p Z^{(j)}(1 - 1_{\|Z\| \geq \tilde{O}(L)})$ is concentrated with probability $(1 - \rho)$, then so is $\frac{1}{p} \sum_{j=1}^p Z^{(j)}$.

Lemma 15 (Theorem 40 (Arora et al., 2015)). Suppose random vector $g_t^{(i)}$ is a $(\rho_*, \rho_+, \zeta_t)$-correlated with high probability with $z^*$ for $t \in [T]$ where $T \leq \text{poly}(n)$, and $\eta_A$ satisfies $0 < \eta_A \leq 2\rho_+$, then for any $t \in [T]$, $E[\|z^{(t+1)} - z^*\|^2] \leq (1 - 2\rho_+\eta_A)\|z^{(t)} - z^*\|^2 + 2\eta_A\zeta_t$.

In particular, if $\|z^{(0)} - z^*\| \leq \epsilon_0$ and $\zeta_t \leq (\rho_+)o((1 - 2\rho_+\eta)^t)\epsilon_0^2 + \zeta$, where $\zeta = \max_{t\in[T]}\zeta_t$, then the updates converge to $z^*$ geometrically with systematic error $\zeta/\rho_*$ in the sense that $E[\|z^{(t+1)} - z^*\|^2] \leq (1 - 2\rho_+\eta_A)^t\epsilon_0^2 + \zeta/\rho_*$. 

39