A. Analysis of Algorithm 1

A.1. Analysis of coding step

Proof of Lemma 1. Denote $S = \text{supp}(x^*)$ and skip the superscript s on A^s for simplicity of notation. We will argue that w.h.p. $S = \{i \in [m] : \frac{1}{\rho}|\langle A_{\bullet i}, y\rangle| \geq C/2\}$ and $\text{sgn}(\langle A^s_{\bullet i}, y\rangle) = \text{sgn}(x^*_i)$ for every $i \in S$.

First, we write the element-wise estimate before thresholding in the encoding step as follows:

$$\frac{1}{\rho}\langle A_{\bullet i}, y\rangle = \frac{1}{\rho}\langle A_{\bullet i}, \mathcal{P}_\Gamma(A^* x^*)\rangle = \frac{1}{\rho}\langle A_{\bullet i}, A^*_{\Gamma\bullet} x^*\rangle = \frac{1}{\rho}\langle A_{\bullet i}, A^*_{\Gamma,i}\rangle x^*_i + \frac{1}{\rho}\sum_{j\neq i}\langle A_{\bullet i}, A^*_{\Gamma,j}\rangle x^*_j. \tag{5}$$

We expect that $Z_i = (1/\rho)\sum_{S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\Gamma,j}\rangle x^*_j$ is negligible based on the closeness of $A_{\bullet i}$ and $A^*_{\bullet i}$ and the democracy of A^*. More precisely, we want to upper bound it by $C/4$ with high probability. Here, C is the lower bound of the nonzero coefficients in x^*. In fact, since Γ and x^*_j are independent, Z_i is a sub-Gaussian random variable with variance

$$\sigma^2_{Z_i} = \frac{1}{\rho^2}\mathbb{E}[\sum_{j\in S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\Gamma,j}\rangle^2] = \sum_{j\in S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\bullet j}\rangle^2 + \frac{1-\rho}{\rho}\sum_{j\in S\setminus\{i\}, l\in[n]} A^2_{li}A^{*2}_{lj}. \tag{6}$$

The second term in (6) can be bounded by using the facts that $\|A^*\|_{\max} \leq O(1/\sqrt{n})$ and $\|A_{\bullet i}\| = 1$. Specifically,

$$\sum_{j\in S\setminus\{i\}, l\in[n]} A^2_{li}A^{*2}_{lj} \leq \sum_{j\in S\setminus\{i\}, l\in[n]} O(1/n)A^2_{li} \leq O(k/n)\|A_{\bullet i}\|^2 = O(k/n),$$

Moreover, since $k \leq \rho\sqrt{n}/\log n$, the second term in (6) is bounded by $O((1-\rho)/\sqrt{n}\log n) = o(C)$.

We bound the first term in (6) by using the incoherence and closeness. For each $j \in S\setminus\{i\}$, we have

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

since $|\langle A^*_{\bullet i}, A^*_{\bullet j}\rangle| \leq \mu/\sqrt{n}$ due to the μ-incoherence of A^*. Now, we combine the term across j and get a matrix form to leverage the spectral norm bound. In particular,

$$\sum_{S\setminus\{i\}}\langle A_{\bullet i}, A^*_{\bullet j}\rangle^2 \leq 2\mu^2 k/n + 2\|A^{*T}_{\bullet S}(A_{\bullet i} - A^*_{\bullet i})\|^2_F \leq 2\mu^2 k/n + 2\|A^*_{\bullet S}\|^2\|A_{\bullet i} - A^*_{\bullet i}\|^2 \leq O(1/\log n),$$

where we have used $m = O(n)$ and $\|A^*_{\bullet S}\| \leq O(1)$. Also, we made use of the condition that $\mu \leq \frac{\sqrt{n}}{2k}$ and $k = \Omega(\log n)$. Putting these together, we get $\sigma^2_{Z_i} \leq O(1/\log n)$. By an application of Bernstein's inequality, we get that $|Z_i| \leq C/4$ w.h.p.

We now argue that $(1/\rho)\langle A_{\bullet i}, y\rangle$ is small when $i \notin S$ and big otherwise. Clearly, when $i \notin S$, $(1/\rho)\langle A_{\bullet i}, y\rangle = Z_i$ is less than $C/4$ in magnitude w.h.p. On the contrary, when $i \in S$, then $|x^*_i| \geq C$, and using the Chernoff bound for $\langle A_{\bullet i}, A^*_{\Gamma,i}\rangle = \sum_{l=1}^n A_{li}A^*_{li}\mathbf{1}[l \in \Gamma]$, we see that

$$(1/\rho)\langle A_{\bullet i}, A^*_{\Gamma,i}\rangle \geq \langle A_{\bullet i}, A^*_{\bullet i}\rangle - o(1) \geq 1 - o(1)$$

w.h.p. because $\langle A_{\bullet i}, A^*_{\bullet i}\rangle \geq 1 - \delta^2/2$. Hence, $|(1/\rho)\langle A_{\bullet i}, y\rangle| \geq C/2$ holds with high probability.

Finally, we take the union bound over all $i = 1, 2, \ldots, m$ to finish the proof. □

A.2. Analysis of the update g^s (in expectation)

Lemma 1 is the key to analyzing the approximate gradient update term

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

This section presents a rigorous analysis of g^s, and is a key step towards achieving the descent property stated in Theorem 4. In essence, we make use of the distributions of x^*, together with its estimate, x to simplify the expectation in g^s. The result is the following:

Lemma 5. *The column-wise expected value g^s of the update rule is of the form*

$$g_i^s = \rho p_i q_i(\lambda_i^s A_{\bullet i}^s - A_{\bullet i}^*) + \rho p_i A_{\bullet -i}^s \text{diag}(q_{ij}) A_{\bullet -i}^{sT} A_{\bullet i}^* + (1-\rho) p_i q_i \text{diag}(A_{\bullet i}^* \circ A_{\bullet i}^s) A_{\bullet i}^s + (1-\rho) \sum_{j \neq i} p_i q_{ij} \text{diag}(A_{\bullet j}^* \circ A_{\bullet j}^s) A_{\bullet j}^s \pm \gamma,$$

where $p_i = \mathbb{E}[x_i \text{sgn}(x_i^*) | i \in S]$, $q_i = \mathbb{P}[i \in S]$ and $q_{ij} = \mathbb{P}[i, j \in S]$. Additionally, $\lambda_i^s = \langle A_{\bullet i}^s, A_{\bullet i}^* \rangle$ and $A_{\bullet -i}^s$ denotes A^s with its i^{th} column removed. In particular, if A^s is $(\delta, 2)$-near to A^* for $\delta = O^*(1/\log n)$, then all the additive terms in g_i^s, except the first term, have norm of order $o(\rho p_i q_i)$.

Proof. For notational simplicity, we skip the superscript s on A^s and g^s. Recall from Lemma 1 that the sign of x^* is recovered w.h.p. from the encoding step. Then under the event that $\text{supp}(x) = \text{supp}(x^*) \equiv S$, we can write $Ax = A_S x_S = \frac{1}{\rho} A_S A_S^T y$. Let us consider the i^{th} column of g, g_i, given by:

$$\begin{aligned} g_i &= \mathbb{E}[(\frac{1}{\rho} A_{\Gamma,S} A_S^T - I)\, y\, \text{sgn}(x_i)] \pm \gamma \\ &= \mathbb{E}[(\frac{1}{\rho} A_{\Gamma,S} A_S^T - I)\, y\, \text{sgn}(x_i^*)] \pm \gamma \\ &= \mathbb{E}[(\frac{1}{\rho} A_{\Gamma,S} A_S^T - I) A_{\Gamma \bullet}^* x^* \text{sgn}(x_i^*)] \pm \gamma \\ &= \mathbb{E}[\big(\frac{1}{\rho} \sum_{j \in S} A_{\Gamma,j} A_{\bullet j}^T - I\big) A_{\Gamma,i}^* x_i^* \text{sgn}(x_i^*)] \pm \gamma. \end{aligned}$$

Here, we make use of the fact that nonzero entries are conditionally independent given the support and have zero mean; therefore $\mathbb{E}[x_j^* \text{sgn}(x_i^*) | S] = 0$ for all $j \neq i$. In the expression, γ denotes any vector whose norm is sufficiently small because of the sign consistency and bounded $(\mathcal{P}_\Gamma(A^s x) - y)\text{sgn}(x)^T$ (see Claim 2 in Appendix C).

We continue simplifying the form of g_i by denoting $p_i = \mathbb{E}[x_i^* \text{sgn}(x_i^*) | i \in S]$, $q_i = \mathbb{P}[i \in S]$ and $q_{ij} = \mathbb{P}[i, j \in S]$. Then,

$$\begin{aligned} g_i &= \mathbb{E}_\Gamma[\frac{1}{\rho} \sum_{j=1}^m p_j q_{ij} A_{\Gamma,j} A_{\bullet j}^T A_{\Gamma,i}^* - p_i q_i A_{\Gamma,i}^*] \pm \gamma \\ &= \frac{1}{\rho} \sum_{j=1}^m p_i q_{ij} \mathbb{E}_\Gamma[A_{\Gamma,j} A_{\Gamma,i}^{*T}] A_{\bullet j} - \rho p_i q_i A_{\bullet i}^* \pm \gamma. \end{aligned}$$

In the final step, we calculate $\mathbb{E}_\Gamma[A_{\Gamma,j} A_{\Gamma,i}^{*T}]$ over the random Γ. One can easily show that

$$\mathbb{E}_\Gamma[A_{\Gamma,j} A_{\Gamma,i}^{*T}] = \rho^2 A_{\bullet j} A_{\bullet i}^{*T} + \rho(1-\rho)\text{diag}(A_{\bullet i}^* \circ A_{\bullet j}),$$

where we use $\text{diag}(v)$ to denote a diagonal matrix with entries in v and $\circ$ to denote the element-wise Hadamard product. As a result, g_i is expressed as follows:

$$\begin{aligned} g_i &= \rho \sum_{j=1}^m p_j q_{ij} A_{\bullet j} A_{\bullet i}^{*T} A_{\bullet j} + (1-\rho) \sum_{j=1}^m p_i q_{ij} \text{diag}(A_{\bullet i}^* \circ A_{\bullet j}) A_{\bullet j} - \rho p_i q_i A_{\bullet i}^* \pm \gamma \\ &= \rho p_i q_i (\lambda_i A_{\bullet i} - A_{\bullet i}^*) + \rho p_i A_{\bullet -i} \text{diag}(q_{ij}) A_{\bullet -i}^T A_{\bullet i}^* + (1-\rho) p_i q_i \text{diag}(A_{\bullet i}^* \circ A_{\bullet i}) A_{\bullet i} \\ &\quad + (1-\rho) \sum_{j \neq i} p_i q_{ij} \text{diag}(A_{\bullet i}^* \circ A_{\bullet j}) A_{\bullet j} \pm \gamma, \end{aligned} \tag{7}$$

where $\lambda_i = \langle A_{\bullet i}, A_{\bullet i}^* \rangle$. Furthermore, $A_{\bullet -i}^T$ denotes the matrix A whose i^{th} column is removed, and $\text{diag}(q_{ij})$ denotes the diagonal matrix of $(q_{i1}, q_{i2} \ldots, q_{im})^T$ without entry $q_{ii} = q_i$.

We will prove that $\rho p_i q_i (\lambda_i A_{\bullet i} - A_{\bullet i}^*)$ is the dominant term in (7). In the special case when $\rho = 1$, g_i is well studied in (Arora et al., 2015). Here we follow the same strategy and give upper bounds for the remaining terms. First, from the

nearness we have $\|A\| \le \|A - A^*\| + \|A^*\| \le O(\sqrt{m/n})$, and also $\|A^*_{\bullet i}\| = 1$; hence

$$\|\rho p_i A_{\bullet -i} \mathrm{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i}\| \le (\rho p_i \max_{j\neq i} q_{ij})\|A\|^2 \tag{8}$$

$$\le O(\rho p_i q_i \max_{j\neq i} q_{ij}/q_i) = o(\rho p_i q_i), \tag{9}$$

for $q_{ij} = \Theta(k^2/m^2)$ and $q_i = \Theta(k/m)$. The remaining terms can be bounded using the max norm constraint and the closeness of A and A^*. More precisely,

$$\begin{aligned}\|\mathrm{diag}(A^*_{\bullet i} \circ A_{\bullet i}) A_{\bullet i}\| &\le \|\mathrm{diag}(A^*_{\bullet i} \circ A_{\bullet i})\| \\ &\le \|\mathrm{diag}(A^*_{\bullet i} \circ A^*_{\bullet i})\| + \|\mathrm{diag}(A^*_{\bullet i} \circ (A_{\bullet i} - A^*_{\bullet i}))\| \\ &\le O(1/n) + O(\delta/\sqrt{n}) \\ &\le O(\delta/\sqrt{n}),\end{aligned} \tag{10}$$

since $\|A\|_{\max} \le O(1/\sqrt{n})$ and $\|A_{\bullet i} - A^*_{\bullet i}\| \le \delta$. Since $(1-\rho)/\rho \le k$ and $k \le O^*(\rho\sqrt{n}/\log n)$, then

$$\|(1-\rho) p_i q_i \mathrm{diag}(A^*_{\bullet i} \circ A_{\bullet i}) A_{\bullet i}\| \le O(\rho p_i q_i k\delta/\sqrt{n}) = o(\rho p_i q_i).$$

Similarly, we have

$$\begin{aligned}\Big\|\sum_{j\neq i} q_{ij} \mathrm{diag}(A^*_{\bullet i} \circ A_{\bullet j}) A_{\bullet j}\Big\|^2 &= \sum_{l=1}^n \big(\sum_{j\neq i} q_{ij} A^*_{li} A^2_{lj}\big)^2 \\ &\le \sum_{l=1}^n (\max_{j\neq i} q_{ij}\|A\|_{\max})^2 (\sum_{j\neq i} A^2_{lj})^2 \\ &\le (\max_{j\neq i} q_{ij}\|A\|_{\max})^2 \sum_{l=1}^n \|A_{l\bullet}\|^4\end{aligned} \tag{11}$$

Moreover $\|A_{l\bullet}\| \le \|A\| \le O(1)$, $\|A_{\max}\| \le O(1/\sqrt{n})$ and $k \le O^*(\rho\sqrt{n}/\log n)$, then

$$\Big\|(1-\rho)\sum_{j\neq i} q_{ij} \mathrm{diag}(A^*_{\bullet j} \circ A_{\bullet j}) A_{\bullet j}\Big\| \le O\Big(\rho p_i q_i \frac{1-\rho}{\rho} \max_{j\neq i} q_{ij}/q_i\Big) \tag{12}$$

$$= O\Big(\rho p_i q_i \frac{(1-\rho)k}{m\rho}\Big) = o(\rho p_i q_i) \tag{13}$$

□

From (7), (8), (10) and (12), we have the additive terms in (7) (excluding γ) bounded by $o(\rho p_i q_i)$, hence we can write g_i as $g_i = \rho p_i q_i(\lambda_i A_{\bullet i} - A^*_{\bullet i}) + o(\rho p_i q_i)$. Moreover, $A_{\bullet i}$ is 2δ-close to $A^*_{\bullet i}$, then $\lambda_i = \langle A_{\bullet i}, A^*_{\bullet i}\rangle \ge 1 - \delta \approx 1$. Therefore, the update rule g_i approximately aligns with the desired direction $A_{\bullet i} - A^*_{\bullet i}$, which leads to the descent property argued in the next section.

A.3. Descent property of g_i^s

We now prove:

Lemma 6. *The update g_i^s is correlated with the desired direction $A^s_{\bullet i} - A^*_{\bullet i}$; that is,*

$$\langle g_i^s, A^s_{\bullet i} - A^*_{\bullet i}\rangle \ge \rho p_i q_i (2 - \zeta^2)\|A^s_{\bullet i} - A^*_{\bullet i}\|^2 + \frac{1}{8\rho p_i q_i}\|g_i\|^2 - \frac{\epsilon^2}{4\rho p_i q_i},$$

for $\zeta = 1 + 2\frac{1-\rho}{\rho}\|A^*\|_{\max} = 1 + o(1)$ and $\epsilon = O(k^2/n^2)$.

Proof. We prove this lemma by mainly using the results in the above section. We first rewrite g_i in Equation (7) in terms of the desired update direction $A^s_{\bullet i} - A^*_{\bullet i}$ and everything else. For simplicity, we omit the superscript s and $2\alpha = \rho p_i q_i$ throughout the proof. We have:

$$\begin{aligned} g_i &= \rho p_i q_i(\lambda_i A_{\bullet i} - A^*_{\bullet i}) + \rho p_i A_{\bullet -i}\text{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i} + (1-\rho) p_i q_i \text{diag}(A^*_{\bullet i} \circ A_{\bullet i}) A_{\bullet i} \\ &\qquad + (1-\rho)\sum_{j\neq i} p_i q_{ij} \text{diag}(A^*_{\bullet i} \circ A_{\bullet j}) A_{\bullet j} \ \pm \gamma \\ &= 2\alpha(A_i - A^*_i) + v, \end{aligned} \tag{14}$$

in which v has the form:

$$\begin{aligned} v &= 2\alpha(\lambda_i - 1)A_{\bullet i} + 2\alpha \frac{1-\rho}{\rho}\text{diag}(A^*_{\bullet i} \circ A_{\bullet i}) A_{\bullet i} \\ &\quad + 2\rho p_i A_{\bullet -i}\text{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i} + 2(1-\rho) p_i \sum_{j\neq i} q_{ij} \text{diag}(A^*_{\bullet i} \circ A_{\bullet j}) A_{\bullet j} \ \pm \gamma. \end{aligned}$$

First, we bound $\|v\|$ in terms of $\|A_{\bullet i} - A^*_{\bullet i}\|$. Since $A_{\bullet i}$ is δ-close to $A^*_{\bullet i}$ and both have unit norm, then $\|2\alpha(\lambda_i - 1)A_{\bullet i}\| = \alpha\|A_{\bullet i} - A^*_{\bullet i}\|^2 \leq \alpha\|A_{\bullet i} - A^*_{\bullet i}\|$. Along with the bound of the second term obtained in (10), we have

$$\|v\| \leq \alpha\left(1 + 2\frac{1-\rho}{\rho} O(1/\sqrt{n})\right)\|A_{\bullet i} - A^*_{\bullet i}\| + \epsilon = \alpha\zeta\|A_{\bullet i} - A^*_{\bullet i}\| + \epsilon, \tag{15}$$

where $\epsilon = \|2\rho p_i A_{\bullet -i}\text{diag}(q_{ij}) A^T_{\bullet -i} A^*_{\bullet i} + 2(1-\rho) p_i \sum_{j\neq i} q_{ij}\text{diag}(A^*_{\bullet i} \circ A_{\bullet j}) A_{\bullet j} \pm \gamma\| = O(\rho k^2/m^2) + O((1-\rho)k^2/m^2) = O(k^2/m^2)$ due to (8) and (12). Here, ζ denotes the factor inside the parentheses.

Now, we look at the correlation of g_i and $A_{\bullet i} - A^*_{\bullet i}$ from (14):

$$\langle 2g_i, A_{\bullet i} - A^*_{\bullet i}\rangle = 4\alpha\|A_{\bullet i} - A^*_{\bullet i}\|^2 + \langle 2v, A_{\bullet i} - A^*_{\bullet i}\rangle. \tag{16}$$

Moreover, squaring both sides of (14) and re-arranging leads to

$$\begin{aligned} 2\langle v, A_{\bullet i} - A^*_{\bullet i}\rangle &= \frac{1}{2\alpha}\|g_i\|^2 - 2\alpha\|A_{\bullet i} - A^*_{\bullet i}\|^2 - \frac{1}{2\alpha}\|v\|^2 \\ &\geq \frac{1}{2\alpha}\|g_i\|^2 - 2\alpha\|A_{\bullet i} - A^*_{\bullet i}\|^2 - \alpha\zeta^2\|A_{\bullet i} - A^*_{\bullet i}\|^2 - \frac{\epsilon^2}{\alpha}, \end{aligned} \tag{17}$$

where in the last step we have used the Cauchy-Schwarz inequality:

$$\|v\|^2 \leq 2(\alpha^2\zeta^2\|A_{\bullet i} - A^*_{\bullet i}\|^2 + \epsilon^2),$$

applied to the right hand side of (15).

Expressions (16) and (17) imply that

$$\langle 2g_i, A_{\bullet i} - A^*_{\bullet i}\rangle \geq \alpha(2 - \zeta^2)\|A_{\bullet i} - A^*_{\bullet i}\|^2 + \frac{1}{2\alpha}\|g_i\|^2 - \frac{\epsilon^2}{\alpha}.$$

Since $(1-\rho)/\rho \leq k \leq O(\rho\sqrt{n}/\log n)$ and $m = O(n)$, then $1 < \zeta^2 < 2$. Besides, we have $p_i = \Theta(k/m)$ and $q_i = \Theta(1)$, then $\alpha = (1/2)\rho p_i q_i = \Theta(\rho k/m)$, and $\epsilon^2/\alpha = O(k^3/\rho m^3)$ we have lower bound on the gradient. This is equivalent to saying that g^s_i is $(\Omega(k/m), \Omega(m/k), O(k^3/\rho m^3))$-correlated with the true solution $A^*_{\bullet i}$ (see (Arora et al., 2015).) □

Proof of Theorem 4. Having argued the correlation of g^s_i and $A_{\bullet i} - A^*_{\bullet i}$, we apply Theorem 6 in (Arora et al., 2015) to obtain the descent stated in Theorem 4. Next, we will establish the nearness for the update at step s. □

A.4. Nearness

The final step in analyzing Algorithm 1 is to show that the nearness of A^{s+1} to the ground truth A^* is maintained after each update. Clearly, A^{s+1} is columnwise close to A^*, which follows from Theorem 4. What remains is to argue that $\|A^{s+1} - A^*\| \leq 2\|A^*\|$ holds true, which is stated in Lemma 7. To this end, we require the sampling probability ρ to be constant. However, we can remove this condition by projecting each update of A on convex set $\mathcal{B} = \{A | A \text{ is } \delta\text{-close to } A^* \text{ and } \|A\| \leq 2\|A^*\|\}$ to guarantee the nearness. The details can be found in (Arora et al., 2015).

Lemma 7. *Provided that A^s is $(\delta, 2)$-near to A^* and that the probability ρ is a constant of n, then $\|A^{s+1} - A^*\| \leq 2\|A^*\|$.*

Proof. Notice from the update that $A^{s+1} - A^* = A^s - A^* - \eta g^s$. Using the column-wise g_i^s in (7), we have the matrix form for g^s as

$$-\eta g^s = -\eta g^s_{\big|_{\rho=1}} - \eta(1-\rho)(A^* \circ A^s \circ A^s)\text{diag}(p_i q_i) - \eta(1-\rho)Q \pm \eta\gamma, \tag{18}$$

where $Q \in \mathbb{R}^{n\times m}$ whose column Q_i equals to $\sum_{j\neq i} p_i q_{ij} \text{diag}(A^*_{\bullet j} \circ A_{\bullet j})A_{\bullet j}$. Since $\|A^s - A^*\| \leq 2\|A^*\|$, then to prove the lemma we need $\|\eta g^s\| \leq o(\|A^*\|)$. Arora et al. (2015) have shown the same nearness property for $\rho = 1$, i.e. $\|\eta g^s_{|_{\rho=1}}\| \leq o(\|A^*\|)$. We will show that the last two terms involving $1-\rho$ are negligible of $\|A^*\|$. From (11), we have bound on each column Q_i such that $\|Q_i\| \leq O(\max_{j\neq i} q_{ij})$. Then,

$$\|Q\| \leq \|Q\|_F \leq \sqrt{m}\max_i \|Q_i\| \leq O(\max_{j\neq i} q_{ij}\sqrt{m}) = O(k^2/m\sqrt{m}).$$

Moreover, $\eta = \Theta(m/\rho k)$ and $k \leq O^*(\rho\sqrt{n}/\log n)$, therefore

$$\eta(1-\rho)\|Q\| \leq O\Big(\frac{(1-\rho)k}{\rho\sqrt{m}}\Big) = o(1)$$

We now bound the term $\eta(1-\rho)(A^* \circ A^s \circ A^s)\text{diag}(p_i q_i)$ using the column-wise upper bound in (10). More specifically,

$$\|\eta(1-\rho)(A^* \circ A^s \circ A^s)\text{diag}(p_i q_i)\| \leq \sqrt{m}\|\eta(1-\rho)p_i q_i \text{diag}(A^*_{\bullet i} \circ A_{\bullet i})A_{\bullet i}\| \leq O(\frac{m}{\rho k}(1-\rho)p_i q_i \delta\sqrt{m/n}) \leq o(1)$$

for a constant ρ independent of n, $p_i q_i = \Theta(k/m)$ and $m = O(n)$. Put together, we complete the proof of Lemma 7. □

B. Analysis of Algorithm 2

Proof of Lemma 3. Recall the distributional properties of x^* that x_i^*'s are conditionally independent given $S = \text{supp}(x^*)$ and the summary statistics are $\mathbb{E}[x_i^{*4} | i \in S] = c_i \in (0,1)$, $\mathbb{E}[x_i^{*2} | i \in S] = 1$, $q_i = \mathbb{P}[i \in S]$ and $q_{ij} = \mathbb{P}[i, j \in S]$.

$$\begin{aligned}
M_{u,v} &= \frac{1}{\rho^4}\mathbb{E}[\langle y,u\rangle\langle y,v\rangle yy^T] = \frac{1}{\rho^2}\mathbb{E}\big[\langle x^*,\beta\rangle\langle x^*,\beta'\rangle A^*_{\Gamma\bullet}x^*x^{*T}A^{*T}_{\Gamma\bullet}\big] \\
&= \frac{1}{\rho^2}\mathbb{E}_\Gamma\mathbb{E}_{x^*}\Big[\sum_{i\in S}\beta_i x_i^* \sum_{i\in S}\beta_i' x_i^* \sum_{i,j\in S} x_i^* x_j^* A^*_{\Gamma,i}A^{*T}_{\Gamma,i}\Big] \\
&= \frac{1}{\rho^2}\sum_{i\in[m]} q_i c_i \beta_i\beta_i' \mathbb{E}_\Gamma[A^*_{\Gamma,i}A^{*T}_{\Gamma,i}] + \frac{1}{\rho^2}\sum_{i,j\in[m],j\neq i} q_{ij}\beta_i\beta_i'\mathbb{E}_\Gamma[A^*_{\Gamma,j}A^{*T}_{\Gamma,j}] + 2q_{ij}\beta_i\beta_j'\mathbb{E}_\Gamma[A^*_{\Gamma,i}A^{*T}_{\Gamma,j}],
\end{aligned}$$

We continue calculating the expectations over Γ. All of those terms are of the same form:

$$\mathbb{E}_\Gamma[A^*_{\Gamma,i}A^{*T}_{\Gamma,j}] = \rho(1-\rho)\text{diag}(A^*_{\bullet i} \circ A^*_{\bullet j}) + \rho^2 A^*_{\bullet i}A^{*T}_{\bullet j}.$$

Plug in this expression into $M_{u,v}$ to have,

$$\begin{aligned}
M_{u,v} &= \sum_{i\in U\cap V} q_i c_i \beta_i\beta_i' A^*_{\bullet i}A^{*T}_{\bullet i} + \sum_{i\notin U\cap V} q_i c_i\beta_i\beta_i' A^*_{\bullet i}A^{*T}_{\bullet i} + \sum_{j\neq i} q_{ij}\beta_i\beta_i' A^*_{\bullet j}A^{*T}_{\bullet j} + 2\sum_{j\neq i} q_{ij}\beta_i\beta_j' A^*_{\bullet i}A^{*T}_{\bullet j} \\
&\quad + \frac{1-\rho}{\rho}\sum_{i\in[m]} q_i\beta_i\beta_i'\text{diag}(A^*_{\bullet i} \circ A^*_{\bullet i}) + \frac{1-\rho}{\rho}\sum_{j\neq i} q_{ij}\beta_i\beta_i'\text{diag}(A^*_{\bullet j} \circ A^*_{\bullet j}) + 2q_{ij}\beta_i\beta_j'\text{diag}(A^*_{\bullet i} \circ A^*_{\bullet j}) \\
&= \sum_{i\in U\cap V} q_i c_i\beta_i\beta_i' A^*_{\bullet i}A^{*T}_{\bullet i} + \text{ perturbation terms},
\end{aligned}$$

where all the terms except $\sum_{i\in U\cap V} q_i c_i \beta_i \beta_i' A^*_{\bullet i} A^{*T}_{\bullet i}$ are expected to be small enough. When $\rho = 1$, then $M_{u,v}$ simply includes the first four terms, which is exactly the weighted matrix studied in (Arora et al., 2015) for regular sparse coding. We will adapt bounds for these terms that now depend on ρ. First of all, for $i \notin U \cap V$ assume $\alpha_i = 0$, using Claim 2 and $|\alpha_i'| \le O(\log n)$ we have $|\beta_i \beta_i'| \le |(\beta_i - \alpha_i)(\beta_i' - \alpha_i')| + |\beta_i \alpha_i'| \le O^*(1/\log n)$, then

$$\Big\| \sum_{i\notin U\cap V} q_i c_i \beta_i \beta_i' A^*_{\bullet i} A^{*T}_{\bullet i} \Big\| \le O^*(k/m \log n), \tag{19}$$

for $q_i = \Theta(k/m)$. For next two perturbation terms, recall from Claim 2 β and β' has norms bounded by $O(\sqrt{k}\log n/\rho)$ and $q_{ij} = \Theta(k^2/m^2)$. We again use the results from (Arora et al., 2015) to get

$$\Big\| \sum_{j\neq i} q_{ij} (\beta_i \beta_i' A^*_{\bullet j} A^{*T}_{\bullet j} + 2\beta_i \beta_j' A^*_{\bullet i} A^{*T}_{\bullet j}) \Big\| \le O\Big(\frac{k^3 \log^2 n}{\rho^2 m^2}\Big). \tag{20}$$

Now, we will handle the terms involving the diagonal matrices as follows,

$$\begin{aligned}
\Big\| \sum_{i\in[m]} q_i \beta_i \beta_i' \mathrm{diag}(A^*_{\bullet i} \circ A^*_{\bullet i}) \Big\| &= \max_{j\in[n]} \Big| \sum_{i\in[m]} q_i \beta_i \beta_i' A^{*2}_{ji} \Big| \le \max_{i,j} (q_i A^{*2}_{ji}) \Big| \sum_{i\in m} \beta_i \beta_i' \Big| \\
&\le \max_i q_i \|A^*\|^2_{\max} \|\beta\| \|\beta'\| = O\Big(\frac{k^2 \log^2 n}{\rho^2 mn}\Big)
\end{aligned} \tag{21}$$

because of the fact that $\|A^*\|_{\max} \le O(1/\sqrt{n})$. Similarly, we also have the same bound for the below term

$$\begin{aligned}
\Big\| \sum_{j\neq i} q_{ij} \beta_i \beta_i' \mathrm{diag}(A^*_{\bullet j} \circ A^*_{\bullet j}) \Big\| &= \max_{l\in[n]} \Big| \sum_{j\neq i} q_{ij} \beta_i \beta_i' A^{*2}_{lj} \Big| = \max_{l\in[n]} \Big| \sum_i \beta_i \beta_i' \sum_{j\neq i} q_{ij} A^{*2}_{lj} \Big| \\
&\le \max_{i,l} \Big(\sum_{j\neq i} q_{ij} A^{*2}_{lj} \Big) \Big| \sum_{i\in m} \beta_i \beta_i' \Big| \le \max_{i,l} \Big(\sum_{j\neq i} q_{ij} A^{*2}_{lj} \Big) \|\beta\| \|\beta'\| \\
&= O\Big(\frac{k^2 \log^2 n}{\rho^2 mn}\Big),
\end{aligned} \tag{22}$$

where we used $\sum_{j\neq i} q_{ij} A^{*2}_{lj} \le \max_{i\neq j} q_{ij} \|A^*_{l\bullet}\|^2 \le O(k^2/mn)$ since $\|A^*_{l\bullet}\| \le \|A^*\| \le O(\sqrt{m/n})$.

We bound the last term using a result from (Nguyen et al., 2018) (proof of Claim 4) that $\sum_{j\neq i} q_{ij} \beta_i \beta_j' A^*_{li} A^*_{lj} = A^{*T}_{l\bullet} Q_\beta A^*_{l\bullet}$ where $(Q_\beta)_{ij} = q_{ij}\beta_i\beta_j'$ for $i \neq j$ and $(Q_\beta)_{ij} = 0$ for $i = j$, so

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

Moreover, $\|Q_\beta\|_F^2 = \sum_{i\neq j} q_{ij}^2 \beta_i^2 (\beta_j')^2 \le (\max_{i\neq j} q_{ij}^2) \sum_i \beta_i^2 \sum_j (\beta_j')^2 \le (\max_{i\neq j} q_{ij}^2) \|\beta\|^2 \|\beta'\|^2$, then

$$\begin{aligned}
\Big\| \sum_{j\neq i} q_{ij} \beta_i \beta_j' \mathrm{diag}(A^*_{\bullet i} \circ A^*_{\bullet j} \Big\| &= \max_{l\in[n]} \Big| \sum_{j\neq i} q_{ij} \beta_i \beta_j' A^*_{li} A^*_{lj} \Big| = \max_{l\in[n]} |A^{*T}_{l\bullet} Q_\beta A^*_{l\bullet}| \\
&\le (\max_{i\neq j} q_{ij}^2) \|\beta\|^2 \|\beta'\|^2 \le O\Big(\frac{k^2 \log^2 n}{\rho^2 m^2}\Big).
\end{aligned} \tag{23}$$

Since $(1-\rho)/\rho \le k$ and $m = O(n)$, then (20), (21), (22) and (23) are all bounded by $O\Big(\frac{k^3 \log^2 n}{\rho^2 mn}\Big)$. Besides, we know that $k \le O^*(\frac{\rho\sqrt{n}}{\log n})$, then all the perturbation terms are bounded by $O^*(k/m \log n)$. We have finished the proof of Lemma 3. □

C. Sample Complexity

In this section, we give concentration bounds for the finite-sample estimates $\widehat{g}^s$ and $\widehat{M}_{u,v}$ and prove Theorem 3 and Theorem 5 . We employ the same technique used in (Arora et al., 2015), which basically apply Bernstein inequalities for proper vector and matrix random variables. The inequality is generally stated in the following lemma.

Lemma 8. *Suppose that $Z^{(1)}, Z^{(2)}, \ldots, Z^{(p)}$ are p i.i.d. samples drawn from some distribution $\mathcal{D}$ such that $\mathbb{E}[Z^{(j)}] = 0$, $\|Z^{(j)}\| \leq R$ almost surely and $\|\mathbb{E}[Z^{(j)}(Z^{(j)})^T\| \leq \sigma^2$ for each j, then*

$$\frac{1}{p}\Big\|\sum_{j=1}^{p} Z^{(j)}\Big\| \leq \widetilde{O}\Big(\frac{R}{p} + \sqrt{\frac{\sigma^2}{p}}\Big) \tag{24}$$

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

In order to apply the above inequality, we need bounds on the random variable Z and its covariance. However, these quantities are not bounded almost surely, and hence we use the common trick of analyzing a truncated version of Z to overcome this issue. Lemma 9 provides sufficient conditions for the truncation trick to work

Lemma 9 (Arora et al. (2015))**.** *Suppose a random variable Z satisfies $\mathbb{P}[\|Z\| \geq R(\log(1/\rho))^C] \leq \rho$ for some constant $C > 0$, then*

1. *If $p = n^{O(1)}$, it holds that $\|Z^{(j)}\| \leq \widetilde{O}(R)$ for each j with probability $1 - n^{-\omega(1)}$.*
2. *$\|\mathbb{E}[Z\mathbf{1}_{\|Z\| \geq \widetilde{\Omega}(R)}]\| = n^{-\omega(1)}$.*

Note that there is a slight abuse of notation here: the constant C and ρ are only used in the context of the above lemma and are not related to those used in our generative model. Since the random components in $\widehat{g}$ and $\widehat{M}_{u,v}$ are products of sub-Gaussian random variables, we can apply Lemma 8 and Lemma 9 to show the concentration of $\frac{1}{p}\sum_{i=1}^{p} Z^{(j)}(1 - \mathbf{1}_{\|Z^{(j)}\| \geq \widetilde{\Omega}(R)})$, then conclude about the concentration of $\frac{1}{p}\sum_{i=1}^{p} Z^{(j)}$ likewise.

In bounding $\|\mathbb{E}[ZZ^T(1 - \mathbf{1}_{\|Z\| \geq \widetilde{\Omega}(R))}]\|$, we sometimes need to take bounds of some random terms out of the expectation. In such case, the following lemma is often useful.

Lemma 10 (Nguyen et al. (2018))**.** *Suppose a random variable $\tilde{Z}\tilde{Z}^T = aT$ where $a \geq 0$ and T is positive semi-definite. Suppose $\mathbb{P}[a \geq \mathcal{A}] = n^{-\omega(1)}$ and $\mathcal{B} > 0$ is a constant. Then,*

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

Other details of these auxiliary lemmas can be found in (Arora et al., 2015; Nguyen et al., 2018).

C.1. Sample Complexity of Algorithm 1

C.1.1. Proof of Theorem 3

We start by using two key auxiliary lemmas for the concentration of $\widehat{g}$, both column-wise as well as for the whole matrix.

Lemma 11. *At iteration s of Algorithm 1, suppose that A^s is $(\delta_s, 2)$-near to A^*. Then $\|\widehat{g}_i^s - g_i^s\| \leq O(k/m)\cdot(o(\delta_s) + O(\epsilon_s))$ with high probability for $\delta_s = O^*(1/\log n)$ and $\epsilon_s = O(\sqrt{k/n})$ when $p = \widetilde{\Omega}(m)$.*

Lemma 12. *If A^s is $(\delta_s, 2)$-near to A^* and number of samples used in step s is $p = \widetilde{\Omega}(mk)$, then with high probability $\|A^{s+1} - A^*\| \leq 2\|A^*\|$.*

While the proof of Lemma 11 is provided below, Lemma 12 directly follows from Lemma 42 in Arora et al. (2015) and the number of samples being $\widetilde{\Omega}(mk)$.

Proof of Theorem 3. We can write $\widehat{g}_i^s$ as

$$\widehat{g}_i^s = g_i^s + (\widehat{g}_i^s - g_i^s) = g_i^s + O(k/m)\cdot(o(\delta_s) + O(\epsilon_s))$$

with high probability; then argue that $\widehat{g}_i^s$ is correlated with $A_{\bullet i} - A^*_{\bullet i}$ with high probability from Lemma 6. The descent property follows directly as Theorem 4 except that we have the expected $\langle \widehat{g}_i^s, A_{\bullet i} - A^*_{\bullet i}\rangle$ on the right hand side. The overall sample complexity is $\widetilde{O}(mk)$, which combines the complexities of having descent and maintaining nearness. □

C.1.2. PROOF OF LEMMA 11

Notice that $\widehat{g}_i^s$ is a sum of p random vectors of the form $(\mathcal{P}_\Gamma(Ax)-y)\text{sgn}(x_i)$. We will show the concentration of $\widehat{g}_i^s$ by applying the Bernstein inequality on $Z \triangleq (\mathcal{P}_\Gamma(Ax)-y)\text{sgn}(x_i)$. Nevertheless, the inequality does not give a sharp bound for such sparse Z, so we instead consider $Z \triangleq (\mathcal{P}_\Gamma(Ax)-y)\text{sgn}(x_i)|i\in S$, with $S=\text{supp}(x^*)$ and $x=\text{threshold}_{C/2}(A^Ty)$.

Claim 1. *Suppose that $Z^{(1)}, Z^{(2)}, \dots, Z^{(N)}$ are i.i.d. samples of the random variable $Z=\mathcal{P}_\Gamma(y-Ax)\text{sgn}(x_i)|i\in S$. Then,*

$$\Big\|\frac{1}{N}\sum_{j=1}^{N} Z^{(j)} - \mathbb{E}[Z]\Big\| \le o(\delta_s) + O(\epsilon_s) \tag{25}$$

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

Proof of Lemma 11. The lemma is easily proved by applying Claim 1. For the reader, we recycle the proof of Lemma 43 in (Arora et al., 2015).

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

$$\widehat{g}_i = \frac{N}{p}\frac{1}{N}\sum_j (\mathcal{P}_\Gamma(Ax^{(j)})-y^{(j)})\text{sgn}(x_i^{(j)}),$$

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

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

holds with high probability as $p=\Omega(mN/k)$. Substituting N in Claim 1, we obtain the results in Lemma 11. □

Proof of Claim 1. To prove it, we need to bound $\|Z\|$ and its variance (Lemma 2 and Lemma 3), then we can apply the Bernstein inequality in Lemma 8.

Claim 2. *$\|Z\| \le \widetilde{O}(\delta_s\sqrt{k}+\mu k/\sqrt{n})$ holds with high probability over the randomness of y.*

Proof. From the generative model and the support consistency of the encoding step, we have

$$y=\mathcal{P}_\Gamma(A^*x^*) = A^*_{\Gamma,S}x^*_S \text{ and } x_S = A^T_{\bullet S}y = A^T_{\bullet S}A^*_{\Gamma,S}x^*_S$$

and plug the following quantities into the

$$\begin{aligned} y-\mathcal{P}_\Gamma(Ax) &= A^*_{\Gamma,S}x^*_S - A_{\Gamma,S}A^T_{\bullet S}A^*_{\Gamma,S}x^*_S \\ &= (A^*_{\Gamma,S}-A_{\Gamma,S})x^*_S + A_{\Gamma,S}(I_n - A^T_{\bullet S}A^*_{\Gamma,S})x^*_S. \end{aligned}$$

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

$$\|(\mathcal{P}_\Gamma(Ax)-y)\text{sgn}(x_i)|i\in S\| \le \widetilde{O}(\|A^*_{\Gamma,S}-A_{\Gamma,S}\|_F + \|A_{\Gamma,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F).$$

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

$$\|A^*_{\Gamma,S}-A_{\Gamma,S}\|_F^2 = \sum_{i\in S}\|A_{\Gamma,i}-A^*_{\Gamma,i}\|^2 \le \delta_s^2 k \tag{26}$$

due to the δ-closeness of A and A^*. This leads to $\|A^*_{\Gamma,S} - A_{\Gamma,S}\|_F \le \delta_s\sqrt{k}$ w.h.p. To handle the other two, we use the fact that $\|UV\|_F \le \|U\|\|V\|_F$. For the second term, we have

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

where $\|A_{\Gamma,S}\| \le \|A_{\Gamma\bullet}\| \le O(1)$ due to the nearness.

The second part is rearranged to take advantage of the closeness and incoherence properties:

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

where we have used $\|I_k - A^{*T}_{\bullet S}A^*_{\bullet S}\|_F \le \mu k/\sqrt{n}$ because of the μ-incoherence of A^*, $\|A_{\bullet S} - A^*_{\bullet S}\|_F \le \delta_s\sqrt{k}$ in (26) and $\|A^*_{\bullet S}\| \le \|A^*\| \le O(1)$. Accordingly, the second Frobenius norm is bounded by

$$\|A_{\Gamma,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F \le O\big(\mu k/\sqrt{n} + \delta_s\sqrt{k}\big). \tag{27}$$

Claim 3. $\mathbb{E}[\|Z\|^2] \le O(\delta_s^2 k + k^2/n)$ *holds with* $\delta_s = O^*(1/\log n)$.

Proof. In the following proofs, we use x^*_S to mean a vector of size k obtained by selecting entries in S. Using the fact that $E[x^*_S x^{*T}_S] = I_k$, we can expand the expectation $\mathbb{E}[\|Z\|^2]$ as follows,

$$\begin{aligned}\mathbb{E}[\|\mathcal{P}_\Gamma(y - Ax)\mathrm{sgn}(x_i)\|^2 | i \in S] &= \mathbb{E}[\|(A^*_{\Gamma,S} - A_{\Gamma,S}A^T_{\bullet S}A^*_{\bullet S})x^*_S\|^2]\\ &= \mathbb{E}[\|A^*_{\Gamma,S} - A_{\Gamma,S}A^T_{\bullet S}A^*_{\bullet S}\|_F^2 | i \in S]\\ &\le \mathbb{E}[\|(A^*_{\Gamma,S} - A_{\Gamma,S})\|^2 | i \in S] + \mathbb{E}[\|A_{\Gamma,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|^2 | i \in S]\\ &\le \delta_s^2 k + \mathbb{E}[\|A_{\Gamma,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|^2 | i \in S].\end{aligned}$$

Here we have used the bound $\|(A^*_{\Gamma,S} - A_{\Gamma,S})\|^2 \le \delta_s^2 k$ for the first term shown in the previous claim. For the second term, we notice that

$$\mathbb{E}[\|A_{\Gamma,S}(I_k - A^T_{\bullet S}A^*_{\bullet S})\|_F^2 | i \in S] \le \sup_S \|A_{\Gamma,S}\|^2 \mathbb{E}[\|I_k - A^T_{\bullet S}A^*_{\bullet S}\|_F^2 | i \in S], \tag{28}$$

in which $\sup_S\|A_{\Gamma,S}\| \le \|A_{\Gamma\bullet}\| \le O(1)$. We will show that $\mathbb{E}[\|I_k - A^T_{\bullet S}A^*_{\bullet S}\|_F^2 | i \in S] \le O(k\delta_s^2) + O(k^2/n)$ by recycling the proof from (Arora et al., 2015):

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

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

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

The μ-incoherence, δ-closeness and the spectral norm of A^* have been used in the last step. Similarly, we can bound $\|A^T_{\bullet i}A^*_{\bullet,-i}\|^2$ and $\|A^T_{\bullet,-i}A^*_{\bullet i}\|^2$. As a result,

$$\mathbb{E}[\|I_k - A^T_{\bullet S}A^*_{\bullet S}\|_F^2 | i \in S] \le O(k\delta_s^2) + O(k^2/n). \tag{29}$$

Combining (28) and (29), we have shown that the covariance is bounded by: $\sigma^2 = O(\delta_s^2 k + k^2/n)$. □

Having had $R = \widetilde{O}(\delta_s\sqrt{k} + \mu k/\sqrt{n})$ and $\sigma^2 = O(\delta_s^2 k + k^2/n)$ in Claims 2 and 3, we are now ready to apply truncated Bernstein inequality to the random variable $Z^{(j)}(1 - 1_{\|Z^{(j)}\| \geq \Omega(R)})$, leading to the concentration of $\frac{1}{N}\sum_{j=1}^{N} Z^{(j)}$. More precisely,

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

holds with high probability when $N = \widetilde{\Omega}(k)$. As such, we finished the proof of Claim 1.

C.2. Sample Complexity of Algorithm 2

In the next proofs, we argue the concentration inequality for $\widehat{M}_{u,v}$ computed in Algorithm 2, which is the empirical average over i.i.d. samples of y, then prove Theorem 5. We note that while u and v are fixed for one iteration, they are random. The (conditional) expectations contain randomness from u and v, hence in some high probability statement, we refer it to the randomness of u, v.

Lemma 13. *Consider Algorithm 2 in which p is the given number of incomplete samples. For any pair of full samples u and v, with high probability $\|\widehat{M}_{u,v} - M_{u,v}\| \leq O^*(k/m \log n)$ when $p = \widetilde{\Omega}(mk/\rho^4)$.*

C.2.1. Proof of Lemma 13

Consider a random matrix variable $Z \triangleq \langle y, u\rangle\langle y, v\rangle yy^T$. We have $\widehat{M}_{u,v} = \frac{1}{p}\sum_{i=1}^{p} Z^{(i)}/\rho^4$. To give a tail bound for $\|\widehat{M}_{u,v} - M_{u,v}\|$, all we need is derive are an upper norm bound R of the matrix random variable Z and its variance, then apply Bernstein inequality. These following claims provide bounds for $\|Z\|$ and $\|\mathbb{E}[ZZ^T]\|$.

Claim 4. *$\|y\| \leq \widetilde{O}(\sqrt{k})$ and $|\langle y, u\rangle| \leq \widetilde{O}(\sqrt{k})$ hold with high probability (over random samples u and v.)*

Proof. Under the generative model where $S = \text{supp}(x^*)$, we have

$$\|y\| = \|A^*_{\Gamma,S}x^*_S\| \leq \|A^*_{\Gamma,S}x^*_S\| \leq \|A^*_{\Gamma,S}\|\|x^*_S\|.$$

From Claim 2, $\|x^*_S\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. In addition, $\|A^*_{\Gamma,S}\| \leq \|A^*\| \leq O(1)$. Therefore, $\|y\| \leq \widetilde{O}(\sqrt{k})$ w.h.p., which is the first part of the proof. To bound the second term, we write it as

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

Even though u is fully observed sample, we can prove similarly that $\|u\| \leq \widetilde{O}(\sqrt{k})$ w.h.p. which results in $\|A^{*T}_{\bullet S}u\| \leq \|A^{*T}_{\bullet S}\|\|u\| \leq \widetilde{O}(\sqrt{k})$ with high probability. Consequently, $|\langle y, u\rangle| \leq \widetilde{O}(\sqrt{k})$ w.h.p., and we finish the proof of the claim. □

Claim 5. *$\|Z\| \leq \widetilde{O}(k^2)$ and $\|\mathbb{E}[ZZ^T]\| \leq \widetilde{O}(\rho^4 k^3/m)$ hold with high probability.*

Proof. First, it is obvious that

$$\|Z\| \leq |\langle y, u\rangle\langle y, v\rangle|\|y\|^2,$$

in which $|\langle y, u\rangle\langle y, v\rangle| \leq \widetilde{O}(k)$ and $\|y\|^2 \leq \widetilde{O}(k)$ w.h.p. (according to Claim 4). Clearly, $\|Z\| \leq \widetilde{O}(k^2)$ w.h.p.

For the second part, we use the auxiliary lemma 10 to take out the bound of $\|Z\|$. Specifically, we have just shown that $\|Z\| \leq \widetilde{O}(k^2)$ and $\langle y, v\rangle^2\|y\|^2 \leq \widetilde{O}(k^2)$, applying Lemma 10:

$$\|\mathbb{E}[ZZ^T(1 - \mathbf{1}_{\|Z\| \geq \widetilde{\Omega}(k^2)})]\| \leq \widetilde{O}(k^2)\|\mathbb{E}[\langle y, u\rangle^2 yy^T]\| + \widetilde{O}(k^2)O(n^{-\omega(1)}) \leq \widetilde{O}(k^2)\|\rho^4 M_{u,u}\|,$$

where $M_{u,u}$ is the expected weighted covariance matrix defined in Lemma 3 for u and $v = u$. From Lemma 3 we have

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

and the perturbation terms are all bounded by $O^*(k/m \log n)$ whereas

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

w.h.p. since $|\beta_i| \leq \log n$ w.h.p. Finally, the variance bound is $\widetilde{O}(\rho^4 k^3/m)$ w.h.p. □

Then, applying Bernstein inequality in Lemma 8 to the truncated version of Z with $R = \widetilde{O}(k^2)$ and variance $\sigma^2 = \widetilde{O}(\rho^4 k^3/m)$ and obtain the concentration for the full Z to get

$$\|\widehat{M}_{u,v} - M_{u,v}\| \leq \frac{\widetilde{O}(k^2)}{\rho^4 p} + \frac{1}{\rho^4}\sqrt{\frac{\widetilde{O}(\rho^4 k^3/m)}{p}} \leq O^*(k/m \log n)$$

w.h.p. when the number of samples is $p = \widetilde{\Omega}(mk/\rho^4)$. We finish the proof of Lemma 13. □

C.2.2. Proof of Theorem 5

We can write the empirical estimate $\widehat{M}_{u,v}$ in term of its expectation $M_{u,v}$ as

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

and the new term $\widehat{M}_{u,v} - M_{u,v}$ can be considered an additional perturbation with the same magnitude $O^*(k/m \log n)$ in spectral norm. As a consequence, as u and v share a unique element in their code supports, the top singular vectors of $\widehat{M}_{u,v}$ is $O^*(1/\log n)$ -close to $A^*_{\bullet i}$ with high probability using $p = \widetilde{O}(mk/\rho^4)$ partial samples.

Each vector added to the list L in Algorithm 2 is close to one of the dictionary, then it must be the case that A^0 is δ-close to A^*. In addition, the nearness of A^0 to A^* is guaranteed via an appropriate projection onto the convex set $\mathcal{B} = \{A | A \text{ close to } A^0 \text{ and } \|A\| \leq 2\|A^*\|\}$.

Finally, using the result in (Arora et al., 2015), the number of full samples in $\mathcal{P}_1$ is $\widetilde{O}(m)$ such that we can draw u, v share uniquely and estimate all the m dictionary atoms. Overall, the sample complexities of Algorithm 2 are $\widetilde{O}(m)$ full samples and $p = \widetilde{O}(mk/\rho^4)$ partial samples. We finish the proof of Theorem 5.