Provable Alternating Gradient Descent for Non-negative Matrix Factorization with Strong Correlations

Yuanzhi Li [1] **Yingyu Liang** [1]

Abstract

Non-negative matrix factorization is a basic tool for decomposing data into the feature and weight matrices under non-negativity constraints, and in practice is often solved in the alternating minimization framework. However, it is unclear whether such algorithms can recover the ground-truth feature matrix when the weights for different features are highly correlated, which is common in applications. This paper proposes a simple and natural alternating gradient descent based algorithm, and shows that with a mild initialization it provably recovers the ground-truth in the presence of strong correlations. In most interesting cases, the correlation can be in the same order as the highest possible. Our analysis also reveals its several favorable features including robustness to noise. We complement our theoretical results with empirical studies on semi-synthetic datasets, demonstrating its advantage over several popular methods in recovering the ground-truth.

1. Introduction

Non-negative matrix factorization (NMF) is an important tool in data analysis and is widely used in image processing, text mining, and hyperspectral imaging (e.g., (Lee & Seung, 1997; Blei et al., 2003; Yang & Leskovec, 2013)). Given a set of observations $\mathbf{Y} = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$, the goal of NMF is to find a feature matrix $\mathbf{A} = \{a_1, a_2, \dots, a_D\}$ and a non-negative weight matrix $\mathbf{X} = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ such that $y^{(i)} \approx \mathbf{A}x^{(i)}$ for any i, or $\mathbf{Y} \approx \mathbf{AX}$ for short. The intuition of NMF is to write each data point as a *non-negative* combination of the features. By doing so, one can avoid cancellation of different features and improve interpretability by thinking of each $x^{(i)}$ as a (unnormalized) probability distribution over the features. It is also observed empirically that the non-negativity constraint on the coefficients can lead to better features and improved downstream performance of the learned features.

Unlike the counterpart which factorizes $\mathbf{Y} \approx \mathbf{AX}$ without assuming non-negativity of $\mathbf{X}$, NMF is usually much harder to solve, and can even by NP-hard in the worse case (Arora et al., 2012b). This explains why, despite all the practical success, NMF largely remains a mystery in theory. Moreover, many of the theoretical results for NMF were based on very technical tools such has algebraic geometry (e.g., (Arora et al., 2012b)) or tensor decomposition (e.g. (Anandkumar et al., 2012)), which undermine their applicability in practice. Arguably, the most widely used algorithms for NMF use the *alternative minimization scheme*: In each iteration, the algorithm alternatively keeps $\mathbf{A}$ or $\mathbf{X}$ as fixed and tries to minimize some distance between $\mathbf{Y}$ and $\mathbf{AX}$. Algorithms in this framework, such as multiplicative update (Lee & Seung, 2001) and alternative non-negative least square (Kim & Park, 2008), usually perform well on real world data. However, alternative minimization algorithms are usually notoriously difficult to analyze. This problem is poorly understood, with only a few provable guarantees known (Awasthi & Risteski, 2015; Li et al., 2016). Most importantly, these results are only for the case when the coordinates of the weights are from essentially independent distributions, while in practice they are known to be correlated, for example, in correlated topic models (Blei & Lafferty, 2006). As far as we know, there exists no rigorous analysis of practical algorithms for the case with strong correlations.

In this paper, we provide a theoretical analysis of a natural algorithm AND (**A**lternative **N**on-negative gradient **D**escent) that belongs to the practical framework, and show that it probably recovers the ground-truth given a mild initialization. It works under general conditions on the feature matrix and the weights, in particular, allowing strong correlations. It also has multiple favorable features that are unique to its success. We further complement our theoretical analysis by experiments on semi-synthetic data, demon-

---

Authors listed in alphabetic order. [1]Princeton University, Princeton, NJ, USA. Correspondence to: Yuanzhi Li <yuanzhil@cs.princeton.edu>, Yingyu Liang <yingyul@cs.princeton.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

strating that the algorithm converges faster to the ground-truth than several existing practical algorithms, and providing positive support for some of the unique features of our algorithm. Our contributions are detailed below.

1.1. Contributions

In this paper, we assume a generative model of the data points, given the ground-truth feature matrix $\mathbf{A}^*$. In each round, we are given $y = \mathbf{A}^* x$,[1] where x is sampled i.i.d. from some *unknown* distribution μ and the goal is to recover the ground-truth feature matrix $\mathbf{A}^*$. We give an algorithm named AND that starts from a mild initialization matrix and provably converges to $\mathbf{A}^*$ in polynomial time. We also justify the convergence through a sequence of experiments. Our algorithm has the following favorable characteristics.

1.1.1. Simple Gradient Descent Algorithm

The algorithm AND runs in stages and keeps a working matrix $\mathbf{A}^{(t)}$ in each stage. At the t-th iteration in a stage, after getting one sample y, it performs the following:

$$\text{(Decode)} \quad z = \phi_\alpha\left((\mathbf{A}^{(0)})^\dagger y\right),$$
$$\text{(Update)} \quad \mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta\left(yz^\top - \mathbf{A}^{(t)} zz^\top\right),$$

where α is a threshold parameter,

$$\phi_\alpha(x) = \begin{cases} x & \text{if } x \geq \alpha, \\ 0 & \text{otherwise,} \end{cases}$$

$(\mathbf{A}^{(0)})^\dagger$ is the Moore-Penrose pesudo-inverse of $\mathbf{A}^{(0)}$, and η is the update step size. The decode step aims at recovering the corresponding weight for the data point, and the update step uses the decoded weight to update the feature matrix. The final working matrix at one stage will be used as the $\mathbf{A}^{(0)}$ in the next stage. See Algorithm 1 for the details.

At a high level, our update step to the feature matrix can be thought of as a gradient descent version of alternative non-negative least square (Kim & Park, 2008), which at each iteration alternatively minimizes $L(\mathbf{A}, \mathbf{Z}) = \|\mathbf{Y} - \mathbf{A}\mathbf{Z}\|_F^2$ by fixing $\mathbf{A}$ or $\mathbf{Z}$. Our algorithm, instead of performing an complete minimization, performs only a stochastic *gradient descent* step on the feature matrix. To see this, consider one data point y and consider minimizing $L(\mathbf{A}, z) = \|y - \mathbf{A}z\|_F^2$ with z fixed. Then the gradient of $\mathbf{A}$ is just $-\nabla L(\mathbf{A}) = (y - \mathbf{A}z)z^\top$, which is exactly the update of our feature matrix in each iteration.

As to the decode step, when $\alpha = 0$, our decoding can be regarded as a one-shot approach minimizing $\|\mathbf{Y} - \mathbf{A}\mathbf{Z}\|_F^2$ restricted to $\mathbf{Z} \geq 0$. Indeed, if for example projected gradient descent is used to minimize $\|\mathbf{Y} - \mathbf{A}\mathbf{Z}\|_F^2$, then the projection step is exactly applying ϕ_α to $\mathbf{Z}$ with $\alpha = 0$. A key ingredient of our algorithm is choosing α to be larger than zero and then decreasing it, which allows us to outperform the standard algorithms.

Perhaps worth noting, our decoding only uses $\mathbf{A}^{(0)}$. Ideally, we would like to use $(\mathbf{A}^{(t)})^\dagger$ as the decoding matrix in each iteration. However, such decoding method requires computing the pseudo-inverse of $\mathbf{A}^{(t)}$ at every step, which is extremely slow. Instead, we divide the algorithm into stages and in each stage, we only use the starting matrix in the decoding, thus the pseudo-inverse only needs to be computed once per stage and can be used across all iterations inside. We can show that our algorithm converges in polylogarithmic many stages, thus gives us to a much better running time. These are made clear when we formally present the algorithm in Section 4 and the theorems in Section 5 and 6.

1.1.2. Handling strong correlations

The most notable property of AND is that it can provably deal with *highly* correlated distribution μ on the weight x, meaning that the coordinates of x can have very strong correlations with each other. This is important since such correlated x naturally shows up in practice. For example, when a document contains the topic "machine learning", it is more likely to contain the topic "computer science" than "geography" (Blei & Lafferty, 2006).

Most of the previous theoretical approaches for analyzing alternating between decoding and encoding, such as (Awasthi & Risteski, 2015; Li et al., 2016; Arora et al., 2015), require the coordinates of x to be pairwise-independent, or almost pairwise-independent (meaning $\mathbb{E}_\mu[x_i x_j] \approx \mathbb{E}_\mu[x_i]\mathbb{E}_\mu[x_j]$). In this paper, we show that algorithm AND can recover $\mathbf{A}^*$ even when the coordinates are highly correlated. As one implication of our result, when the sparsity of x is $O(1)$ and each entry of x is in $\{0, 1\}$, AND can recover $\mathbf{A}^*$ even if each $\mathbb{E}_\mu[x_i x_j] = \Omega(\min\{\mathbb{E}_\mu[x_i], \mathbb{E}_\mu[x_j]\})$, matching (up to constant) the highest correlation possible. Moreover, we do not assume any prior knowledge about the distribution μ, and the result also extends to general sparsities as well.

1.1.3. Pseudo-inverse decoding

One of the feature of our algorithm is to use Moore-Penrose pesudo-inverse in decoding. Inverse decoding was also used in (Li et al., 2016; Arora et al., 2015; 2016). However, their algorithms require carefully finding an inverse such that certain norm is minimized, which is not as efficient as the vanilla Moore-Penrose pesudo-inverse. It was also observed in (Arora et al., 2016) that Moore-Penrose

[1] We also consider the noisy case; see 1.1.5.

pesudo-inverse works equally well in practice, but the experiment was done only when $\mathbf{A} = \mathbf{A}^*$. In this paper, we show that Moore-Penrose pesudo-inverse also works well when $\mathbf{A} \neq \mathbf{A}^*$, both theoretically and empirically.

1.1.4. THRESHOLDING AT DIFFERENT α

Thresholding at a value $\alpha > 0$ is a common trick used in many algorithms. However, many of them still only consider a fixed α throughout the entire algorithm. Our contribution is a new method of thresholding that first sets α to be high, and gradually decreases α as the algorithm goes. Our analysis naturally provides the explicit rate at which we decrease α, and shows that our algorithm, following this scheme, can provably converge to the ground-truth $\mathbf{A}^*$ in polynomial time. Moreover, we also provide experimental support for these choices.

1.1.5. ROBUSTNESS TO NOISE

We further show that the algorithm is robust to noise. In particular, we consider the model $y = \mathbf{A}^* x + \zeta$, where ζ is the noise. The algorithm can tolerate a general family of noise with bounded moments; we present in the main body the result for a simplified case with Gaussian noise and provide the general result in the appendix. The algorithm can recover the ground-truth matrix up to a small blow-up factor times the noise level in *each example*, when the ground-truth has a good condition number. This robustness is also supported by our experiments.

2. Related Work

Practical algorithms. Non-negative matrix factorization has a rich empirical history, starting with the practical algorithms of (Lee & Seung, 1997; 1999; 2001). It has been widely used in applications and there exist various methods for NMF, e.g., (Kim & Park, 2008; Lee & Seung, 2001; Cichocki et al., 2007; Ding et al., 2013; 2014). However, they do not have provable recovery guarantees.

Theoretical analysis. For theoretical analysis, (Arora et al., 2012b) provided a fixed-parameter tractable algorithm for NMF using algebraic equations. They also provided matching hardness results: namely they show there is no algorithm running in time $(mW)^{o(D)}$ unless there is a sub-exponential running time algorithm for 3-SAT. (Arora et al., 2012b) also studied NMF under separability assumptions about the features, and (Bhattacharyya et al., 2016) studied NMF under related assumptions. The most related work is (Li et al., 2016), which analyzed an alternating minimization type algorithm. However, the result only holds with strong assumptions about the distribution of the weight x, in particular, with the assumption that the coordinates of x are independent.

Topic modeling. Topic modeling is a popular generative model for text data (Blei et al., 2003; Blei, 2012). Usually, the model results in NMF type optimization problems with $\|x\|_1 = 1$, and a popular heuristic is *variational inference*, which can be regarded as alternating minimization in KL-divergence. Recently, there is a line of theoretical work analyzing tensor decomposition (Arora et al., 2012a; 2013; Anandkumar et al., 2013) or combinatorial methods (Awasthi & Risteski, 2015). These either need strong structural assumptions on the word-topic matrix $\mathbf{A}^*$, or need to know the distribution of the weight x, which is usually infeasible in applications.

3. Problem and Definitions

We use $\|\mathbf{M}\|_2$ to denote the 2-norm of a matrix $\mathbf{M}$. $\|x\|_1$ is the 1-norm of a vector x. We use $[\mathbf{M}]_i$ to denote the i-th row and $[\mathbf{M}]^i$ to denote the i-th column of a matrix $\mathbf{M}$. $\sigma_{\max}(\mathbf{M})(\sigma_{\min}(\mathbf{M}))$ stands for the maximum (minimal) singular value of $\mathbf{M}$, respectively. We consider a generative model for non-negative matrix factorization, where the data y is generated from[2]

$$y = \mathbf{A}^* x, \quad \mathbf{A}^* \in \mathbb{R}^{W \times D}$$

where $\mathbf{A}^*$ is the ground-truth feature matrix, and x is a non-negative random vector drawn from an unknown distribution μ. The goal is to recover the ground-truth $\mathbf{A}^*$ from i.i.d. samples of the observation y.

Since the general non-negative matrix factorization is NP-hard (Arora et al., 2012b), some assumptions on the distribution of x need to be made. In this paper, we would like to allow distributions as general as possible, especially those with strong correlations. Therefore, we introduce the following notion called (r, k, m, λ)-general correlation conditions (GCC) for the distribution of x.

Definition 1 (General Correlation Conditions, GCC)**.** *Let $\mathbf{\Delta} := \mathbb{E}[xx^\top]$ denote the second moment matrix.*

1. $\|x\|_1 \le r$ *and* $x_i \in [0,1], \forall i \in [D]$.
2. $\mathbf{\Delta}_{i,i} \le \frac{2k}{D}, \forall i \in [D]$.
3. $\mathbf{\Delta}_{i,j} \le \frac{m}{D^2}, \forall i \neq j \in [D]$.
4. $\mathbf{\Delta} \succeq \frac{k}{D}\lambda \mathbf{I}$.

The first condition regularizes the sparsity of x.[3] The second condition regularizes each coordinate of x_i so that there is no x_i being large too often. The third condition

[2] Section 6.2 considers the noisy case.

[3] Throughout this paper, the sparsity of x refers to the ℓ_1 norm, which is much weaker than the ℓ_0 norm (the support sparsity). For example, in LDA, the ℓ_1 norm of x is always 1.

regularizes the maximum pairwise correlation between x_i and x_j. The fourth condition always holds for $\lambda = 0$ since $\mathbb{E}[xx^\top]$ is a PSD matrix. Later we will assume this condition holds for some $\lambda > 0$ to avoid degenerate cases. Note that we put the weight k/D before λ such that λ defined in this way will be a positive constant in many interesting examples discussed below.

To get a sense of what are the ranges of k, m, and λ given sparsity r, we consider the following most commonly studied non-negative random variables.

Proposition 1 (Examples of GCC)**.**

1. *If x is chosen uniformly over s-sparse random vectors with $\{0,1\}$ entries, then $k = r = s$, $m = s^2$ and $\lambda = 1 - \frac{1}{s}$.*
2. *If x is uniformly chosen from Dirichlet distribution with parameter $\alpha_i = \frac{s}{D}$, then $r = k = 1$ and $m = \frac{1}{sD}$ with $\lambda = 1 - \frac{1}{s}$.*

For these examples, the result in this paper shows that we can recover $\mathbf{A}^*$ for aforementioned random variables x as long as $s = O(D^{1/6})$. In general, there is a wide range of parameters (r, k, m, λ) such that learning $\mathbf{A}^*$ is doable with polynomially many samples of y and in polynomial time.

However, just the GCC condition is not enough for recovering $\mathbf{A}^*$. We will also need a mild initialization.

Definition 2 (ℓ-initialization)**.** *The initial matrix $\mathbf{A}_0$ satisfies for some $\ell \in [0,1)$,*

1. $\mathbf{A}_0 = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E})$*, for some diagonal matrix $\mathbf{\Sigma}$ and off-diagonal matrix $\mathbf{E}$.*
2. $\|\mathbf{E}\|_2 \le \ell$, $\|\mathbf{\Sigma} - \mathbf{I}\|_2 \le \frac{1}{4}$.

The condition means that the initialization is not too far away from the ground-truth $\mathbf{A}^*$. For any $i \in [D]$, the i-th column $[\mathbf{A}_0]^i = \mathbf{\Sigma}_{i,i}[\mathbf{A}^*]^i + \sum_{j \neq i} \mathbf{E}_{j,i}[\mathbf{A}^*]^j$. So the condition means that each feature $[\mathbf{A}_0]^i$ has a large fraction of the ground-truth feature $[\mathbf{A}^*]^i$ and a small fraction of the other features. $\mathbf{\Sigma}$ can be regarded as the magnitude of the component from the ground-truth in the initialization, while $\mathbf{E}$ can be regarded as the magnitude of the error terms. In particular, when $\mathbf{\Sigma} = \mathbf{I}$ and $\mathbf{E} = 0$, we have $\mathbf{A}_0 = \mathbf{A}^*$. The initialization allows $\mathbf{\Sigma}$ to be a constant away from $\mathbf{I}$, and the error term $\mathbf{E}$ to be ℓ (in our theorems ℓ can be as large as a constant).

In practice, such an initialization is typically achieved by setting the columns of $\mathbf{A}_0$ to reasonable "pure" data points that contain one major feature and a small fraction of some others (e.g. (lda, 2016; Awasthi & Risteski, 2015)).

Algorithm 1 Alternating Non-negative gradient Descent (AND)

Input: Threshold values $\{\alpha_0, \alpha_1, \ldots, \alpha_s\}$, T, $\mathbf{A}_0$
1: $\mathbf{A}^{(0)} \leftarrow \mathbf{A}_0$
2: **for** $j = 0, 1, \ldots, s$ **do**
3:   **for** $t = 0, 1, \ldots, T$ **do**
4:     On getting sample $y^{(t)}$, do:
5:     $z^{(t)} \leftarrow \phi_{\alpha_j}\left((\mathbf{A}^{(0)})^\dagger y^{(t)}\right)$
6:     $\mathbf{A}^{(t+1)} \leftarrow \mathbf{A}^{(t)} + \eta\left(y^{(t)} - \mathbf{A}^{(t)} z^{(t)}\right)\left(z^{(t)}\right)^\top$
7:   **end for**
8:   $\mathbf{A}^{(0)} \leftarrow \mathbf{A}^{(T+1)}$
9: **end for**
Output: $\mathbf{A} \leftarrow \mathbf{A}^{(T+1)}$

4. Algorithm

The algorithm is formally describe in Algorithm 1. It runs in s stages, and in the j-th stage, uses the same threshold α_j and the same matrix $\mathbf{A}^{(0)}$ for decoding, where $\mathbf{A}^{(0)}$ is either the input initialization matrix or the working matrix obtained at the end of the last stage. Each stage consists of T iterations, and each iteration decodes one data point and uses the decoded result to update the working matrix. It can use a batch of data points instead of one data point, and our analysis still holds.

By running in stages, we save most of the cost of computing $(\mathbf{A}^{(0)})^\dagger$, as our results show that only polylogarithmic stages are needed. For the simple case where $x \in \{0,1\}^D$, the algorithm can use the same threshold value $\alpha = 1/4$ for all stages (see Theorem 1), while for the general case, it needs decreasing threshold values across the stages (see Theorem 4). Our analysis provides the hint for setting the threshold; see the discussion after Theorem 4, and Section 7 for how to set the threshold in practice.

5. Result for A Simplified Case

In this section, we consider the following simplified case:

$$y = \mathbf{A}^* x, \ \ x \in \{0,1\}^D. \tag{5.1}$$

That is, the weight coordinates x_i's are binary.

Theorem 1 (Main, binary)**.** *For the generative model (5.1), there exists $\ell = \Omega(1)$ such that for every (r, k, m, λ)-GCC x and every $\epsilon > 0$, Algorithm AND with $T = \mathsf{poly}(D, \frac{1}{\epsilon})$, $\eta = \frac{1}{\mathsf{poly}(D, \frac{1}{\epsilon})}$, $\{\alpha_i\}_{i=1}^s = \{\frac{1}{4}\}_{i=1}^s$ for $s = \mathsf{polylog}(D, \frac{1}{\epsilon})$ and an ℓ initialization matrix $\mathbf{A}_0$, outputs a matrix $\mathbf{A}$ such that there exists a diagonal matrix $\mathbf{\Sigma} \succeq \frac{1}{2}\mathbf{I}$ with $\|\mathbf{A} - \mathbf{A}^*\mathbf{\Sigma}\|_2 \le \epsilon$ using $\mathsf{poly}(D, \frac{1}{\epsilon})$ samples and iterations, as long as*

$$m = O\left(\frac{kD\lambda^4}{r^5}\right).$$

Therefore, our algorithm recovers the ground-truth $\mathbf{A}^*$ up to scaling. The scaling in unavoidable since there is no assumption on $\mathbf{A}^*$, so we cannot, for example, distinguish $\mathbf{A}^*$ from $2\mathbf{A}^*$. Indeed, if we in addition assume each column of $\mathbf{A}^*$ has norm 1 as typical in applications, then we can recover $\mathbf{A}^*$ directly. In particular, by normalizing each column of $\mathbf{A}$ to have norm 1, we can guarantee that $\|\mathbf{A} - \mathbf{A}^*\|_2 \leq O(\epsilon)$.

In many interesting applications (for example, those in Proposition 1), k, r, λ are constants. The theorem implies that the algorithm can recover $\mathbf{A}^*$ even when $m = O(D)$. In this case, $\mathbb{E}_\mu[x_i x_j]$ can be as large as $O(1/D)$, the same order as $\min\{\mathbb{E}_\mu[x_i], \mathbb{E}_\mu[x_j]\}$, which is the highest possible correlation.

5.1. Intuition

The intuition comes from assuming that we have the "correct decoding", that is, suppose magically for every $y^{(t)}$, our decoding $z^{(t)} = \phi_{\alpha_j}(\mathbf{A}^\dagger y^{(t)}) = x^{(t)}$. Here and in this subsection, $\mathbf{A}$ is a shorthand for $\mathbf{A}^{(0)}$. The gradient descent is then $\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta(y^{(t)} - \mathbf{A}^{(t)}x^{(t)})(x^{(t)})^\top$. Subtracting $\mathbf{A}^*$ on both side, we will get

$$(\mathbf{A}^{(t+1)} - \mathbf{A}^*) = (\mathbf{A}^{(t)} - \mathbf{A}^*)(\mathbf{I} - \eta x^{(t)}(x^{(t)})^\top)$$

Since $x^{(t)}(x^{(t)})^\top$ is positive semidefinite, as long as $\mathbb{E}[x^{(t)}(x^{(t)})^\top] \succ 0$ and η is sufficiently small, $\mathbf{A}^{(t)}$ will converge to $\mathbf{A}^*$ eventually.

However, this simple argument does not work when $\mathbf{A} \neq \mathbf{A}^*$ and thus we do not have the correct decoding. For example, if we just let the decoding be $\tilde{z}^{(t)} = \mathbf{A}^\dagger y^{(t)}$, we will have $y^{(t)} - \mathbf{A}\tilde{z}^{(t)} = y^{(t)} - \mathbf{A}^\dagger \mathbf{A} y^{(t)} = (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{A}^* x^{(t)}$. Thus, using this decoding, the algorithm can never make *any* progress once $\mathbf{A}$ and $\mathbf{A}^*$ are in the same subspace.

The most important piece of our proof is to show that after *thresholding*, $z^{(t)} = \phi_\alpha(\mathbf{A}^\dagger y^{(t)})$ is much closer to $x^{(t)}$ than $\tilde{z}^{(t)}$. Since $\mathbf{A}$ and $\mathbf{A}^*$ are in the same subspace, inspired by (Li et al., 2016) we can write $\mathbf{A}^*$ as $\mathbf{A}(\mathbf{\Sigma} + \mathbf{E})$ for a diagonal matrix $\mathbf{\Sigma}$ and an off-diagonal matrix $\mathbf{E}$, and thus the decoding becomes $z^{(t)} = \phi_\alpha(\mathbf{\Sigma}x^{(t)} + \mathbf{E}x^{(t)})$. Let us focus on one coordinate of $z^{(t)}$, that is, $z_i^{(t)} = \phi_\alpha(\mathbf{\Sigma}_{i,i}x_i^{(t)} + \mathbf{E}_i x^{(t)})$, where $\mathbf{E}_i$ is the i-th row of $\mathbf{E}_i$. The term $\mathbf{\Sigma}_{i,i}x_i^{(t)}$ is a nice term since it is just a rescaling of $x_i^{(t)}$, while $\mathbf{E}_i x^{(t)}$ mixes different coordinates of $x^{(t)}$. For simplicity, we just assume for now that $x_i^{(t)} \in \{0, 1\}$ and $\mathbf{\Sigma}_{i,i} = 1$. In our proof, we will show that the threshold will *remove* a large fraction of $\mathbf{E}_i x^{(t)}$ when $x_i^{(t)} = 0$, and keep a large fraction of $\mathbf{\Sigma}_{i,i}x_i^{(t)}$ when $x_i^{(t)} = 1$. Thus, our decoding is much more accurate than without thresholding. To show this, we maintain a crucial property that for our decoding matrix, we always have $\|\mathbf{E}_i\|_2 = O(1)$. Assuming this, we first consider two extreme cases of $\mathbf{E}_i$.

1. Ultra dense: all coordinates of $\mathbf{E}_i$ are in the order of $\frac{1}{\sqrt{d}}$. Since the sparsity of $x^{(t)}$ is r, as long as $r = o(\sqrt{d})\alpha$, $\mathbf{E}_i x^{(t)}$ will not pass α and thus $z_i^{(t)}$ will be decoded to zero when $x_i^{(t)} = 0$.
2. Ultra sparse: $\mathbf{E}_i$ only has few coordinate equal to $\Omega(1)$ and the rest are zero. Unless $x^{(t)}$ has those exact coordinates equal to 1 (which happens not so often), then $z_i^{(t)}$ will still be zero when $x_i^{(t)} = 0$.

Of course, the real $\mathbf{E}_i$ can be anywhere in between these two extremes, and thus we need more delicate decoding lemmas, as shown in the complete proof.

Furthermore, more complication arises when each $x_i^{(t)}$ is not just in $\{0, 1\}$ but can take fractional values. To handle this case, we will set our threshold α to be large at the beginning and then keep shrinking after each stage. The intuition here is that we first decode the coordinates that we are most confident in, so we do not decode $z_i^{(t)}$ to be non-zero when $x_i^{(t)} = 0$. Thus, we will still be able to remove a large fraction of error caused by $\mathbf{E}_i x^{(t)}$. However, by setting the threshold α so high, we may introduce more errors to the nice term $\mathbf{\Sigma}_{i,i}x_i^{(t)}$ as well, since $\mathbf{\Sigma}_{i,i}x_i^{(t)}$ might not be larger than α when $x_i^{(t)} \neq 0$. Our main contribution is to show that there is a nice trade-off between the errors in $\mathbf{E}_i$ terms and those in $\mathbf{\Sigma}_{i,i}$ terms such that as we gradually decreases α, the algorithm can converge to the ground-truth.

5.2. Proof Sketch

For simplicity, we only focus on one stage and the expected update. The expected update of $\mathbf{A}^{(t)}$ is given by

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta(\mathbb{E}[yz^\top] - \mathbf{A}^{(t)}\mathbb{E}[zz^\top]).$$

Let us write $\mathbf{A}^{(0)} = \mathbf{A}^*(\mathbf{\Sigma}_0 + \mathbf{E}_0)$ where $\mathbf{\Sigma}_0$ is diagonal and $\mathbf{E}_0$ is off-diagonal. Then the decoding is given by

$$z = \phi_\alpha((\mathbf{A}^{(0)})^\dagger y) = \phi_\alpha((\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}x).$$

Let $\mathbf{\Sigma}, \mathbf{E}$ be the diagonal part and the off-diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$.

The key lemma for decoding says that under suitable conditions, z will be close to $\mathbf{\Sigma}x$ in the following sense.

Lemma 2 (Decoding, informal)**.** *Suppose* $\mathbf{E}$ *is small and* $\mathbf{\Sigma} \approx \mathbf{I}$. *Then with a proper threshold value* α*, we have*

$$\mathbb{E}[\mathbf{\Sigma}xx^\top] \approx \mathbb{E}[zx^\top], \;\; \mathbb{E}[\mathbf{\Sigma}xz^\top] \approx \mathbb{E}[zz^\top].$$

Now, let us write $\mathbf{A}^{(t)} = \mathbf{A}^*(\mathbf{\Sigma}_t + \mathbf{E}_t)$. Then applying the above decoding lemma, the expected update of $\mathbf{\Sigma}_t + \mathbf{E}_t$ is

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(\mathbf{I} - \mathbf{\Sigma\Delta\Sigma}) + \mathbf{\Sigma}^{-1}(\mathbf{\Sigma\Delta\Sigma}) + \mathbf{R}_t$$

where $\mathbf{\Delta} = \mathbb{E}[xx^\top]$ and $\mathbf{R}_t$ is a small error term.

Our second key lemma is about this update.

Lemma 3 (Update, informal)**.** *Suppose the update rule is*

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(1 - \eta\mathbf{\Lambda}) + \eta\mathbf{Q}\mathbf{\Lambda} + \eta\mathbf{R}_t$$

for some PSD matrix $\mathbf{\Lambda}$ and $\|\mathbf{R}_t\|_2 \le C''$. Then

$$\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{Q}\|_2 \le \|\mathbf{\Sigma}_0 + \mathbf{E}_0 - \mathbf{Q}\|_2(1 - \eta\lambda_{\min}(\mathbf{\Lambda}))^t + \frac{C''}{\lambda_{\min}(\mathbf{\Lambda})}.$$

Applying this on our update rule with $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ and $\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}$, we know that when the error term is sufficiently small, we can make progress on $\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}^{-1}\|_2$. Furthermore, by using the fact that $\mathbf{\Sigma}_0 \approx \mathbf{I}$ and $\mathbf{E}_0$ is small, and the fact that $\mathbf{\Sigma}$ is the diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$, we can show that after sufficiently many iterations, $\|\mathbf{\Sigma}_t - \mathbf{I}\|_2$ blows up slightly, while $\|\mathbf{E}_t\|_2$ is reduced significantly. Repeating this for multiple stages completes the proof.

We note that most technical details are hidden, especially for the proofs of the decoding lemma, which need to show that the error term $\mathbf{R}_t$ is small. This crucially relies on the choice of α, and relies on bounding the effect of the correlation. These then give the setting of α and the bound on the parameter m in the final theorem.

6. More General Results

6.1. Result for General x

This subsection considers the general case where $x \in [0,1]^D$. Then the GCC condition is not enough for recovery, even for $k, r, m = O(1)$ and $\lambda = \Omega(1)$. For example, GCC does not rule out the case that x is drawn uniformly over $(r-1)$-sparse random vectors with $\{\frac{1}{D}, 1\}$ entries, when one cannot recover even a reasonable approximation of $\mathbf{A}^*$ since a common vector $\frac{1}{D}\sum_i [\mathbf{A}^*]^i$ shows up in all the samples. This example shows that the difficulty arises if each x_i constantly shows up with a small value. To avoid this, a general and natural way is to assume that each x_i, once being non-zero, has to take a large value with sufficient probability. This is formalized as follows.

Definition 3 (Decay condition)**.** *A distribution of x satisfies the order-q decay condition for some constant $q \ge 1$, if for all $i \in [D]$, x_i satisfies that for every $\alpha > 0$,*

$$\Pr[x_i \le \alpha \mid x_i \ne 0] \le \alpha^q.$$

When $q = 1$, each x_i, once being non-zero, is uniformly distributed in the interval $[0,1]$. When q gets larger, each x_i, once being non-zero, will be more likely to take larger values. We will show that our algorithm has a better guarantee for larger q. In the extreme case when $q = \infty$, x_i will only take $\{0,1\}$ values, which reduces to the binary case.

In this paper, we show that this simple decay condition, combined with the GCC conditions and an initialization with constant error, is sufficient for recovering $\mathbf{A}^*$.

Theorem 4 (Main)**.** *There exists $\ell = \Omega(1)$ such that for every (r,k,m,λ)-GCC x satisfying the order-q condition, every $\epsilon > 0$, there exists T, η and a sequence of $\{\alpha_i\}$ [4] such that Algorithm AND, with ℓ-initialization matrix $\mathbf{A}_0$, outputs a matrix $\mathbf{A}$ such that there exists a diagonal matrix $\mathbf{\Sigma} \succeq \frac{1}{2}\mathbf{I}$ with $\|\mathbf{A} - \mathbf{A}^*\mathbf{\Sigma}\|_2 \le \epsilon$ with poly$(D, \frac{1}{\epsilon})$ samples and iterations, as long as*

$$m = O\left(\frac{kD^{1-\frac{1}{q}}\lambda^{4+\frac{4}{q}}}{r^{5+\frac{6}{q+1}}}\right).$$

As mentioned, in many interesting applications, $k = r = \lambda = \Theta(1)$, where our algorithm can recover $\mathbf{A}^*$ as long as $m = O(D^{1-\frac{1}{q+1}})$. This means $\mathbb{E}_\mu[x_i x_j] = O(D^{-1-\frac{1}{q+1}})$, a factor of $D^{-\frac{1}{q+1}}$ away from the highest possible correlation $\min\{\mathbb{E}_\mu[x_i], \mathbb{E}_\mu[x_j]\} = O(1/D)$. Then, the larger q, the higher correlation it can tolerate. As q goes to infinity, we recover the result for the case $x \in \{0,1\}^D$, allowing the highest order correlation.

The analysis also shows that the decoding threshold should be $\alpha = \left(\frac{\lambda\|\mathbf{E}_0\|_2}{r}\right)^{\frac{2}{q+1}}$ where $\mathbf{E}_0$ is the error matrix at the beginning of the stage. Since the error decreases exponentially with stages, this suggests to decrease α exponentially with stages. This is crucial for AND to recover the ground-truth; see Section 7 for the experimental results.

6.2. Robustness to Noise

We now consider the case when the data is generated from $y = \mathbf{A}^*x + \zeta$, where ζ is the noise. For the sake of demonstration, we will just focus on the case when $x_i \in \{0,1\}$ and ζ is random Gaussian noise $\zeta \sim \gamma\mathcal{N}\left(0, \frac{1}{W}\mathbf{I}\right)$. [5] A more general theorem can be found in the appendix.

Definition 4 ((ℓ,ρ)-initialization)**.** *The initial matrix $\mathbf{A}_0$ satisfies for some $\ell, \rho \in [0,1)$,*

1. *$\mathbf{A}_0 = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}$, for some diagonal matrix $\mathbf{\Sigma}$ and off-diagonal matrix $\mathbf{E}$.*
2. *$\|\mathbf{E}\|_2 \le \ell$, $\|\mathbf{\Sigma} - \mathbf{I}\|_2 \le \frac{1}{4}$, $\|\mathbf{N}\|_2 \le \rho$.*

Theorem 5 (Noise, binary)**.** *Suppose each $x_i \in \{0,1\}$. There exists $\ell = \Omega(1)$ such that for every (r,k,m,λ)-GCC x, every $\epsilon > 0$, Algorithm AND with $T =$ poly$(D, \frac{1}{\epsilon}), \eta =$*

[4] In fact, we will make the choice explicit in the proof.

[5] we make this scaling so $\|\zeta\|_2 \approx \gamma$.

$\frac{1}{\textsf{poly}(D,\frac{1}{\epsilon})}$, $\{\alpha_i\}_{i=1}^{s} = \{\frac{1}{4}\}_{i=1}^{4}$ *and an* (ℓ,ρ)*-initialization* $\mathbf{A}_0$ *for* $\rho = O(\sigma_{\min}(\mathbf{A}^*))$*, outputs* $\mathbf{A}$ *such that there exists a diagonal matrix* $\mathbf{\Sigma} \succeq \frac{1}{2}\mathbf{I}$ *with*

$$\|\mathbf{A} - \mathbf{A}^*\mathbf{\Sigma}\|_2 \le O\left(\epsilon + r\frac{\sigma_{\max}(\mathbf{A}^*)}{\sigma_{\min}(\mathbf{A}^*)\lambda}\gamma\right)$$

using $\textsf{poly}(D,\frac{1}{\epsilon})$ *iterations, as long as* $m = O\left(\frac{kD\lambda^4}{r^5}\right)$.

The theorem implies that the algorithm can recover the ground-truth up to $r\frac{\sigma_{\max}(\mathbf{A}^*)}{\sigma_{\min}(\mathbf{A}^*)\lambda}$ times γ, the noise level in each sample. Although stated here for Gaussian noise for simplicity, the analysis applies to a much larger class of noises, including adversarial ones. In particular, we only need to the noise ζ have sufficiently bounded $\|\mathbb{E}[\zeta\zeta^\top]\|_2$; see the appendix for the details. For the special case of Gaussian noise, by exploiting its properties, it is possible to improve the error term with a more careful calculation, though not done here.

7. Experiments

To demonstrate the advantage of AND, we complement the theoretical analysis with empirical study on semi-synthetic datasets, where we have ground-truth feature matrices and can thus verify the convergence. We then provide support for the benefit of using decreasing thresholds, and test its robustness to noise. In the appendix, we further test its robust to initialization and sparsity of x, and provide qualitative results in some real world applications. [6]

Setup. Our work focuses on convergence of the solution to the ground-truth feature matrix. However, real-world datasets in general do not have ground-truth. So we construct semi-synthetic datasets in topic modeling: first take the word-topic matrix learned by some topic modeling method as the ground-truth $\mathbf{A}^*$, and then draw x from some specific distribution μ. For fair comparison, we use one not learned by any algorithm evaluated here. In particular, we used the matrix with 100 topics computed by the algorithm in (Arora et al., 2013) on the NIPS papers dataset (about 1500 documents, average length about 1000). Based on this we build two semi-synthetic datasets:

1. DIR. Construct a 100×5000 matrix $\mathbf{X}$, whose columns are from a Dirichlet distribution with parameters $(0.05, 0.05, \ldots, 0.05)$. Then the dataset is $\mathbf{Y} = \mathbf{A}^*\mathbf{X}$.
2. CTM. The matrix $\mathbf{X}$ is of the same size as above, while each column is drawn from the logistic normal prior in the correlated topic model (Blei & Lafferty, 2006). This leads to a dataset with strong correlations.

Note that the word-topic matrix is non-negative. While some competitor algorithms require a non-negative feature matrix, AND does not need such a condition. To demonstrate this, we generate the following synthetic data:

3. NEG. The entries of the matrix $\mathbf{A}^*$ are i.i.d. samples from the uniform distribution on $[-0.5, 0.5)$. The matrix $\mathbf{X}$ is the same as in CTM.

Finally, the following dataset is for testing the robustness of AND to the noise:

4. NOISE. $\mathbf{A}^*$ and $\mathbf{X}$ are the same as in CTM, but $\mathbf{Y} = \mathbf{A}^*\mathbf{X} + \mathbf{N}$ where $\mathbf{N}$ is the noise matrix with columns drawn from $\gamma\mathcal{N}\left(0, \frac{1}{W}\mathbf{I}\right)$ with the noise level γ.

Competitors. We compare the algorithm AND to the following popular methods: Alternating Non-negative Least Square (ANLS (Kim & Park, 2008)), multiplicative update (MU (Lee & Seung, 2001)), LDA (online version (Hoffman et al., 2010)),[7] and Hierarchical Alternating Least Square (HALS (Cichocki et al., 2007)).

Evaluation criterion. Given the output matrix $\mathbf{A}$ and the ground truth matrix $\mathbf{A}^*$, the *correlation error* of the i-th column is given by

$$\varepsilon_i(\mathbf{A}, \mathbf{A}^*) = \min_{j\in[D],\sigma\in\mathbb{R}} \{\|[\mathbf{A}^*]^i - \sigma[\mathbf{A}]^j\|_2\}.$$

Thus, the error measures how well the i-th column of $\mathbf{A}^*$ is covered by the best column of $\mathbf{A}$ up to scaling. We find the best column since in some competitor algorithms, the columns of the solution $\mathbf{A}$ may only correspond to a permutation of the columns of $\mathbf{A}^*$.[8]

We also define the *total correlation error* as

$$\varepsilon(\mathbf{A}, \mathbf{A}^*) = \sum_{i=1}^{D} \varepsilon_i(\mathbf{A}, \mathbf{A}^*).$$

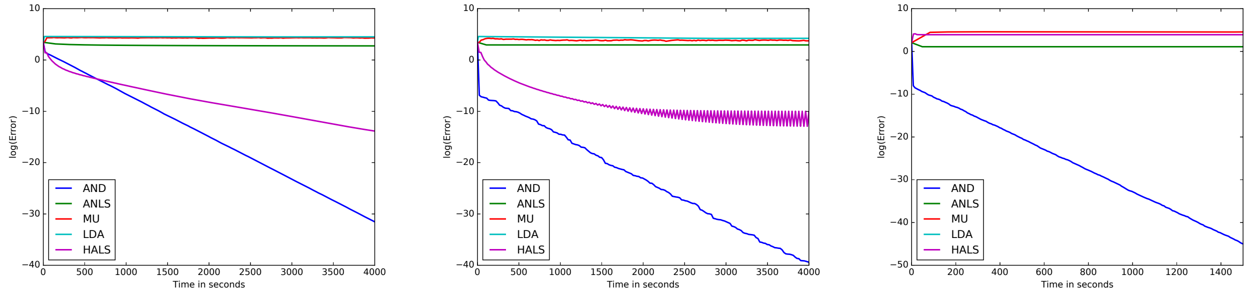
We report the total correlation error in all the experiments.

Initialization. In all the experiments, the initialization matrix $\mathbf{A}_0$ is set to $\mathbf{A}_0 = \mathbf{A}^*(\mathbf{I} + \mathbf{U})$ where $\mathbf{I}$ is the identity matrix and $\mathbf{U}$ is a matrix whose entries are i.i.d. samples from the uniform distribution on $[-0.05, 0.05)$. Note that this is a very weak initialization, since $[\mathbf{A}_0]^i = (1 + \mathbf{U}_{i,i})[\mathbf{A}^*]^i + \sum_{j\neq i}\mathbf{U}_{j,i}[\mathbf{A}^*]^j$ and the magnitude of the noise component $\sum_{j\neq i}\mathbf{U}_{j,i}[\mathbf{A}^*]^j$ can be larger than the signal part $(1 + \mathbf{U}_{i,i})[\mathbf{A}^*]^i$.

[6] The code is public on https://github.com/PrincetonML/AND4NMF.

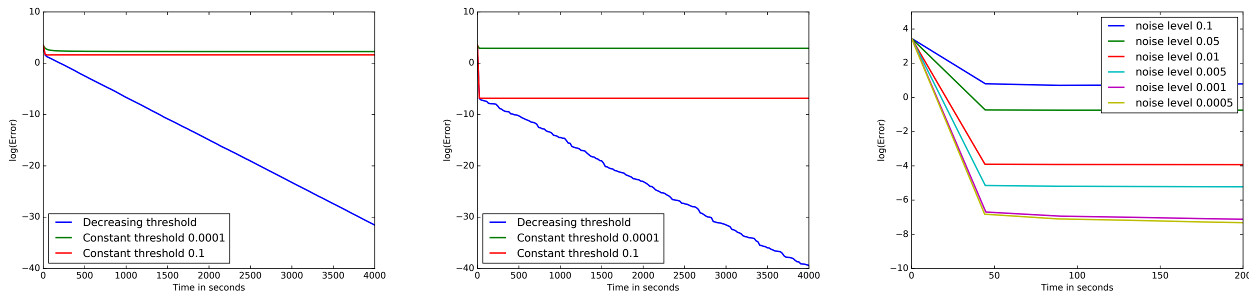
[7] We use the implementation in the sklearn package (http://scikit-learn.org/)

[8] In the Algorithm AND, the columns of $\mathbf{A}$ correspond to the columns of $\mathbf{A}^*$ without permutation.



(a) on DIR dataset (b) on CTM dataset (c) on NEG dataset

Figure 1. The performance of different algorithms on the three datasets. The x-axis is the running time (in seconds), the y-axis is the logarithm of the total correlation error.



(a) different thresholds on DIR (b) different thresholds on CTM (c) robustness to noise

Figure 2. The performance of the algorithm AND with different thresholding schemes, and its robustness to noise. The x-axis is the running time (in seconds), the y-axis is the logarithm of the total correlation error. (a)(b) Using different thresholding schemes on the DIR/CTM dataset. "Decreasing thresold" refers to the scheme used in the original AND, "Constant threshold c" refers to using the threshold value c throughout all iterations. (c) The performance in the presence of noises of various levels.

Hyperparameters and Implementations. For most experiments of AND, we used $T = 50$ iterations for each stage, and thresholds $\alpha_i = 0.1/(1.1)^{i-1}$. For experiments on the robustness to noise, we found $T = 100$ leads to better performance. Furthermore, for all the experiments, instead of using one data point at each step, we used the whole dataset for update.

7.1. Convergence to the Ground-Truth

Figure 1 shows the convergence rate of the algorithms on the three datasets. AND converges in linear rate on all three datasets (note that the y-axis is in log-scale). HALS converges on the DIR and CTM datasets, but the convergence is in slower rates. Also, on CTM, the error oscillates. Furthermore, it doesn't converge on NEG where the ground-truth matrix has negative entries. ANLS converges on DIR and CTM at a very slow speed due to the non-negative least square computation in each iteration. [9] All the other algorithms do not converge to the ground-truth, suggesting that they do not have recovery guarantees.

7.2. The Threshold Schemes

Figure 2(a) shows the results of using different thresholding schemes on DIR, while Figure 2(b) shows that those on CTM. When using a constant threshold for all iterations, the error only decreases for the first few steps and then stop decreasing. This aligns with our analysis and is in strong contrast to the case with decreasing thresholds.

7.3. Robustness to Noise

Figure 2(c) shows the performance of AND on the NOISE dataset with various noise levels γ. The error drops at the first few steps, but then stabilizes around a constant related to the noise level, as predicted by our analysis. This shows that it can recover the ground-truth to good accuracy, even when the data have a significant amount of noise.

[9] We also note that even the thresholding of HALS and ALNS designed for non-negative feature matrices is removed, they still do not converge on NEG.

Acknowledgements

This work was supported in part by NSF grants CCF-1527371, DMS-1317308, Simons Investigator Award, Simons Collaboration Grant, and ONR-N00014-16-1-2329. This work was done when Yingyu Liang was visiting the Simons Institute.

References

Lda-c software. https://github.com/blei-lab/lda-c/blob/master/readme.txt, 2016. Accessed: 2016-05-19.

Anandkumar, A., Kakade, S., Foster, D., Liu, Y., and Hsu, D. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. Technical report, 2012.

Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. Learning latent bayesian networks and topic models under expansion constraints. In *ICML*, 2013.

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond svd. In *FOCS*, 2012a.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. A practical algorithm for topic modeling with provable guarantees. In *ICML*, 2013.

Arora, S., Ge, R., Ma, T., and Moitra, A. Simple, efficient, and neural algorithms for sparse coding. In *COLT*, 2015.

Arora, Sanjeev, Ge, Rong, Kannan, Ravindran, and Moitra, Ankur. Computing a nonnegative matrix factorization–provably. In *STOC*, pp. 145–162. ACM, 2012b.

Arora, Sanjeev, Ge, Rong, Koehler, Frederic, Ma, Tengyu, and Moitra, Ankur. Provable algorithms for inference in topic models. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2859–2867, 2016.

Awasthi, Pranjal and Risteski, Andrej. On some provably correct cases of variational inference for topic models. In *NIPS*, pp. 2089–2097, 2015.

Bhattacharyya, Chiranjib, Goyal, Navin, Kannan, Ravindran, and Pani, Jagdeep. Non-negative matrix factorization under heavy noise. In *Proceedings of the 33nd International Conference on Machine Learning*, 2016.

Blei, David and Lafferty, John. Correlated topic models. *Advances in neural information processing systems*, 18: 147, 2006.

Blei, David M. Probabilistic topic models. *Communications of the ACM*, 2012.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Cichocki, Andrzej, Zdunek, Rafal, and Amari, Shun-ichi. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 169–176. Springer, 2007.

Ding, W., Rohban, M.H., Ishwar, P., and Saligrama, V. Topic discovery through data dependent and random projections. *arXiv preprint arXiv:1303.3664*, 2013.

Ding, W., Rohban, M.H., Ishwar, P., and Saligrama, V. Efficient distributed topic modeling with provable guarantees. In *AISTAT*, pp. 167–175, 2014.

Hoffman, Matthew, Bach, Francis R, and Blei, David M. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

Kim, Hyunsoo and Park, Haesun. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications*, 30(2):713–730, 2008.

Lee, Daniel D and Seung, H Sebastian. Unsupervised learning by convex and conic coding. *NIPS*, pp. 515–521, 1997.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Lee, Daniel D and Seung, H Sebastian. Algorithms for non-negative matrix factorization. In *NIPS*, pp. 556–562, 2001.

Li, Yuanzhi, Liang, Yingyu, and Risteski, Andrej. Recovery guarantee of non-negative matrix factorization via alternating updates. *Advances in neural information processing systems*, 2016.

Yang, Jaewon and Leskovec, Jure. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 587–596. ACM, 2013.

A. Complete Proofs

We now recall the proof sketch.

For simplicity, we only focus on one stage and the expected update. The expected update of $\mathbf{A}^{(t)}$ is given by

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta(\mathbb{E}[yz^\top] - \mathbf{A}^{(t)}\mathbb{E}[zz^\top]).$$

Let us write $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma}_0 + \mathbf{E}_0)$ where $\mathbf{\Sigma}_0$ is diagonal and $\mathbf{E}_0$ is off-diagonal. Then the decoding is given by

$$z = \phi_\alpha(\mathbf{A}^\dagger x) = \phi_\alpha((\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}x).$$

Let $\mathbf{\Sigma}, \mathbf{E}$ be the diagonal part and the off-diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$.

The first step of our analysis is a key lemma for decoding. It says that under suitable conditions, z will be close to $\mathbf{\Sigma}x$ in the following sense:

$$\mathbb{E}[\mathbf{\Sigma}xx^\top] \approx \mathbb{E}[zx^\top], \ \ \mathbb{E}[\mathbf{\Sigma}xz^\top] \approx \mathbb{E}[zz^\top].$$

This key decoding lemma is formally stated in Lemma 6 (for the simplified case where $x \in \{0,1\}^D$) and Lemma 8 (for the general case where $x \in [0,1]^D$).

Now, let us write $\mathbf{A}^{(t)} = \mathbf{A}^*(\mathbf{\Sigma}_t + \mathbf{E}_t)$. Then applying the above decoding lemma, the expected update of $\mathbf{\Sigma}_t + \mathbf{E}_t$ is

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{\Sigma}^{-1}(\mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{R}_t$$

where $\mathbf{R}_t$ is a small error term.

The second step is a key lemma for updating the feature matrix: for the update rule

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(1 - \eta\mathbf{\Lambda}) + \eta\mathbf{Q}\mathbf{\Lambda} + \eta\mathbf{R}_t$$

where $\mathbf{\Lambda}$ is a PSD matrix and $\|\mathbf{R}_t\|_2 \le C''$, we have

$$\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{Q}\|_2 \le \|\mathbf{\Sigma}_0 + \mathbf{E}_0 - \mathbf{Q}\|_2(1 - \eta\lambda_{\min}(\mathbf{\Lambda}))^t + \frac{C''}{\lambda_{\min}(\mathbf{\Lambda})}.$$

This key updating lemma is formally stated in Lemma 10.

Applying this on our update rule with $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ and $\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}$, we know that when the error term is sufficiently small, we can make progress on $\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}^{-1}\|_2$. Then, by using the fact that $\mathbf{\Sigma}_0 \approx \mathbf{I}$ and $\mathbf{E}_0$ is small, and the fact that $\mathbf{\Sigma}$ is the diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$, we can show that after sufficiently many iterations, $\|\mathbf{\Sigma}_t - \mathbf{I}\|_2$ blows up slightly, while $\|\mathbf{E}_t\|_2$ is reduced significantly (See Lemma 11 for the formal statement). Repeating this for multiple stages completes the proof.

Organization. Following the proof sketch, we first present the decoding lemmas in Section A.1, and then the update lemmas in Section A.2. Section A.3 then uses these lemmas to prove the main theorems (Theorem 1 and Theorem 4). Proving the decoding lemmas is highly non-trivial, and we collect the lemmas needed in Section A.4.

Finally, the analysis for the robustness to noise follows a similar proof sketch. It is presented in Section A.5.

A.1. Decoding

A.1.1. $x_i \in \{0,1\}$

Here we present the following decoding Lemma when $x_i \in \{0,1\}$.

Lemma 6 (Decoding)**.** *For every $\ell \in [0,1)$, every off-diagonal matrix $\mathbf{E}'$ such that $\|\mathbf{E}'\|_2 \le \ell$ and every diagonal matrix $\mathbf{\Sigma}'$ such that $\|\mathbf{\Sigma}' - \mathbf{I}\|_2 \le \frac{1}{2}$, let $z = \phi_\alpha((\mathbf{\Sigma}' + \mathbf{E}')x)$ for $\alpha \le \frac{1}{4}$. Then for every $\beta \in (0, 1/2]$,*

$$\|\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2, \|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2 = O(C_1)$$

where

$$C_1 = \frac{kr}{D} + \frac{m\ell^4r^2}{\alpha^3D^2} + \frac{\ell^2\sqrt{km}r^{1.5}}{D^{1.5}\beta} + \frac{\ell^4r^3m}{\beta^2D^2} + \frac{\ell^5r^{2.5}m}{D^2\alpha^2\beta}.$$

Proof of Lemma 6. We will prove the bound on $\|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2$, and a simlar argument holds for that on $\|\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2$.

First consider the term $|\mathbf{\Sigma}'_{i,i}x_i - z_i|$ for a fixed $i \in [D]$. Due to the decoding, we have $z_i = \phi_\alpha([(\mathbf{\Sigma}' + \mathbf{E}')]_i x) = \phi_\alpha(\mathbf{\Sigma}'_{i,i}x_i + \langle e_i, x\rangle)$ where e_i is the i-th row of $\mathbf{E}'$.

Claim 7.

$$\mathbf{\Sigma}'_{i,i}x_i - z_i = a_{x,1}\phi_\alpha(-\langle e_i, x\rangle) + a_{x,2}\phi_\alpha(\langle e_i, x\rangle) - \langle e_i, x\rangle x_i \tag{A.1}$$

where $a_{x,1}, a_{x,2} \in [-1, 1]$ that depends on x.

Proof. To see this, we split into two cases:

1. When $x_i = 0$, then $|\mathbf{\Sigma}'_{i,i}x_i - z_i| = |z_i| \le \phi_\alpha(\langle e_i, x\rangle)$.
2. When $x_i = 1$, then $z_i = 0$ only when $-\langle e_i, x\rangle \ge \frac{1}{2} - \alpha \ge \alpha$, which implies that $|\mathbf{\Sigma}'_{i,i}x_i - z_i + \langle e_i, x\rangle| \le \alpha \le \phi_\alpha(-\langle e_i, x\rangle)$. When $\mathbf{\Sigma}'_{i,i}x_i - z_i + \langle e_i, x\rangle \ne 0$, then $\mathbf{\Sigma}'_{i,i}x_i - z_i = -\langle e_i, x\rangle$.

Putting everything together, we always have:

$$|\mathbf{\Sigma}'_{i,i}x_i - z_i + \langle e_i, x\rangle x_i| \le \phi_\alpha(|\langle e_i, x\rangle|)$$

which means that there exists $a_{x,1}, a_{x,2} \in [-1, 1]$ that depend on x such that

$$\mathbf{\Sigma}'_{i,i}x_i - z_i = a_{x,1}\phi_\alpha(-\langle e_i, x\rangle) + a_{x,2}\phi_\alpha(\langle e_i, x\rangle) - \langle e_i, x\rangle x_i.$$

□

Consider the term $\langle e_i, x\rangle x_i$, we know that for every $\beta \ge 0$,

$$|\langle e_i, x\rangle - \phi_\beta(\langle e_i, x\rangle) + \phi_\beta(-\langle e_i, x\rangle)| \le \beta.$$

Therefore, there exists $b_x \in [-\beta, \beta]$ that depends on x such that

$$\langle e_i, x\rangle x_i = \phi_\beta(\langle e_i, x\rangle) - \phi_\beta(-\langle e_i, x\rangle) - b_x x_i.$$

Putting into (A.1), we get:

$$\mathbf{\Sigma}'_{i,i}x_i - z_i = a_{x,1}\phi_\alpha(-\langle e_i, x\rangle) + a_{x,2}\phi_\alpha(\langle e_i, x\rangle) - \phi_\beta(\langle e_i, x\rangle)x_i + \phi_\beta(-\langle e_i, x\rangle)x_i + b_x x_i.$$

For notation simplicity, let us now write

$$z_i = (\mathbf{\Sigma}'_{i,i} - b_x)x_i + a_i + b_i$$

where

$$a_i = -a_{x,1}\phi_\alpha(-\langle e_i, x\rangle) - a_{x,2}\phi_\alpha(\langle e_i, x\rangle), \quad b_i = \phi_\beta(\langle e_i, x\rangle)x_i - \phi_\beta(-\langle e_i, x\rangle)x_i.$$

We then have

$$(\mathbf{\Sigma}'_{i,i}x_i - z_i)z_j = (b_x x_i - a_i - b_i)((\mathbf{\Sigma}'_{j,j} - b_x)x_j + a_j + b_j).$$

Let us now construct matrix $\mathbf{M}_1, \cdots \mathbf{M}_9$, whose entries are given by

1. $(\mathbf{M}_1)_{i,j} = b_x x_i(\mathbf{\Sigma}'_{j,j} - b_x)x_j$
2. $(\mathbf{M}_2)_{i,j} = b_x x_i a_j$

3. $(\mathbf{M}_3)_{i,j} = b_x x_i b_j$
4. $(\mathbf{M}_4)_{i,j} = -a_i(\mathbf{\Sigma}'_{j,j} - b_x)x_j$
5. $(\mathbf{M}_5)_{i,j} = -a_i a_j$
6. $(\mathbf{M}_6)_{i,j} = -a_i b_j$
7. $(\mathbf{M}_7)_{i,j} = -b_i(\mathbf{\Sigma}'_{j,j} - b_x)x_j$
8. $(\mathbf{M}_8)_{i,j} = -b_i a_j$
9. $(\mathbf{M}_9)_{i,j} = -b_i b_j$

Thus, we know that $\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top] = \sum_{i=1}^{9} \mathbb{E}[\mathbf{M}_i]$. It is sufficient to bound the spectral norm of each matrices separately, as we discuss below.

1. $\mathbf{M}_2, \mathbf{M}_4$: these matrices can be bounded by Lemma 14, term 1.
2. $\mathbf{M}_5$: this matrix can be bounded by Lemma 14, term 2.
3. $\mathbf{M}_6, \mathbf{M}_8$: these matrices can be bounded by Lemma 15, term 3.
4. $\mathbf{M}_3, \mathbf{M}_7$: these matrices can be bounded by Lemma 15, term 2.
5. $\mathbf{M}_9$: this matrix can be bounded by Lemma 15, term 1.
6. $\mathbb{E}[\mathbf{M}_1]$: this matrix is of the form $\mathbb{E}[b_x x(x \odot d_x)^\top]$, where d_x is a vector whose j-th entry is $(\mathbf{\Sigma}'_{j,j} - b_x)$.

   To bound the this term, we have that for any u, v such that $\|u\|_2 = \|v\|_2 = 1$,

   $$u^\top \mathbb{E}[b_x x(x \odot d_x)^\top]v \quad = \quad \mathbb{E}[b_x \langle u, x\rangle \langle v, x \odot d_x\rangle].$$

   When $\beta \le \frac{1}{2}$, since x is non-negative, we know that the maximum of $\mathbb{E}[b_x \langle u, x\rangle\langle v, x \odot d_x\rangle]$ is obtained when $b_x = \beta$, $d_x = (2, \cdots, 2)$ and u, v are all non-negative, which gives us

   $$\mathbb{E}[b_x \langle u, x\rangle \langle v, x \odot d_x\rangle] \le 2\beta \|\mathbb{E}[xx^\top]\|_2 \le \sqrt{\|\mathbb{E}[xx^\top]\|_1 \|\mathbb{E}[xx^\top]\|_\infty} = \|\mathbb{E}[xx^\top]\|_1.$$

   Now, for each row of $\|\mathbb{E}[xx^\top]\|_1$, we know that $[\mathbb{E}[xx^\top]]_i \le \mathbb{E}[x_i \sum_j x_j] \le \frac{2rk}{D}$, which gives us

   $$\|\mathbb{E}[\mathbf{M}_1]\|_2 \le \frac{4\beta rk}{D}.$$

Putting everything together gives the bound on $\|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2$. A similar proof holds for the bound on $\|\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2$. □

A.1.2. GENERAL x_i

We have the following decoding lemma for the general case when $x_i \in [0, 1]$ and the distribution of x satisfies the order-q decay condition.

Lemma 8 (Decoding II). *For every $\ell \in [0, 1)$, every off-diagonal matrix $\mathbf{E}'$ such that $\|\mathbf{E}'\|_2 \le \ell$ and every diagonal matrix $\mathbf{\Sigma}'$ such that $\|\mathbf{\Sigma}' - \mathbf{I}\|_2 \le \frac{1}{2}$, let $z = \phi_\alpha((\mathbf{\Sigma}' + \mathbf{E}')x)$ for $\alpha \le \frac{1}{4}$. Then for every $\beta \in (0, \alpha]$,*

$$\|\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2, \|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2 = O(C_2)$$

where

$$C_2 = \frac{\ell^4 r^3 m}{\alpha \beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2 \alpha^{2.5} \beta} + \frac{\ell^2 kr}{D\beta}\left(\frac{m}{Dk}\right)^{\frac{q}{2q+2}} + \frac{\ell^3 r^2 \sqrt{km}}{D^{1.5}\alpha^2} + \frac{\ell^6 r^4 m}{\alpha^4 D^2} + k\beta\left(\frac{r}{D}\right)^{\frac{2q+1}{2q+2}} + \frac{kr}{D}\alpha^{\frac{q+1}{2}}.$$

Proof of Lemma 8. We consider the bound on $\|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2$, and that on $\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2$ can be proved by a similar argument.

Again, we still have

$$z_i = \phi_\alpha([(\mathbf{\Sigma}' + \mathbf{E}')]_i x) = \phi_\alpha(\mathbf{\Sigma}'_{i,i} x_i + \langle e_i, x\rangle).$$

However, this time even when $x_i \neq 0$, x_i can be smaller than α. Therefore, we need the following inequality.

Claim 9.

$$|\mathbf{\Sigma}'_{i,i} x_i - z_i + \langle e_i, x\rangle \mathbf{1}_{x_i \geq 4\alpha}| \leq \phi_{\alpha/2}(|\langle e_i, x\rangle|) + 2x_i \mathbf{1}_{x_i \in (0,4\alpha)}.$$

Proof. To see this, we can consider the following four events:

1. $x_i = 0$, then $|\mathbf{\Sigma}'_{i,i} x_i - z_i + \langle e_i, x\rangle \mathbf{1}_{x_i \geq 2\alpha}| = |z_i| \leq \phi_\alpha(\langle e_i, x\rangle)$
2. $x_i \geq 4\alpha$. $|\mathbf{\Sigma}'_{i,i} x_i - z_i + \langle e_i, x\rangle \mathbf{1}_{x_i \geq 4\alpha}| = |\mathbf{\Sigma}'_{i,i} x_i + \langle e_i, x\rangle - \phi_\alpha(\mathbf{\Sigma}'_{i,i} x_i + \langle e_i, x\rangle)|$. Since $\mathbf{\Sigma}'_{i,i} x_i \geq 2\alpha$, we can get the same bound.
3. $x_i \in (\alpha/4, 4\alpha)$: then if $z_i \neq 0$, $|\mathbf{\Sigma}'_{i,i} x_i - z_i + \langle e_i, x\rangle| = 0$. Which implies that

$$|\mathbf{\Sigma}'_{i,i} x_i - z_i| \leq |\langle e_i, x\rangle| \leq \phi_{\alpha/2}(|\langle e_i, x\rangle|) + \frac{\alpha}{2} \leq \phi_{\alpha/2}(|\langle e_i, x\rangle|) + 2x_i \mathbf{1}_{x_i \in (0,4\alpha)}$$

   If $z_i = 0$, then

$$|\mathbf{\Sigma}'_{i,i} x_i - z_i| = \mathbf{\Sigma}'_{i,i} x_i \leq 2x_i \mathbf{1}_{x_i \in (0,2\alpha)}$$

4. $x_i \in (0, \alpha/4)$, then $\mathbf{\Sigma}'_{i,i} x_i \leq \frac{\alpha}{2}$, therefore, $z_i \neq 0$ only when $\langle e_i, x\rangle \geq \frac{\alpha}{2}$. We still have: $|\mathbf{\Sigma}'_{i,i} x_i - z_i| \leq \phi_{\alpha/2}(\langle e_i, x\rangle)$ If $z_0 = 0$, then $|\mathbf{\Sigma}'_{i,i} x_i - z_i| \leq 2x_i \mathbf{1}_{x_i \in (0,4\alpha)}$ as before.

Putting everything together, we have the claim. □

Following the exact same calculation as in Lemma 6, we can obtain

$$\begin{aligned}
\mathbf{\Sigma}'_{i,i} x_i - z_i &= a_{x,1}\phi_{\alpha/2}(-\langle e_i, x\rangle) + a_{x,2}\phi_{\alpha/2}(\langle e_i, x\rangle) \\
&\quad - \phi_\beta(\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha} + \phi_\beta(-\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha} \\
&\quad + b_x \mathbf{1}_{x_i \geq 4\alpha} + c_x 2x_i \mathbf{1}_{x_i \in (0,4\alpha)}
\end{aligned}$$

for $a_{x,1}, a_{x,1}, c_x \in [-1, 1]$ and $b_x \in [-\beta, \beta]$.

Therefore, consider a matrix $\mathbf{M}$ whose (i, j)-th entry is $(\mathbf{\Sigma}'_{i,i} x_i - z_i)z_j$. This entry can be written as the summation of the following terms.

1. Terms that can be bounded by Lemma 14. These include

$$a_{x,1}\phi_{\alpha/2}(-\langle e_i, x\rangle)x_j, \quad a_{x,2}\phi_{\alpha/2}(\langle e_i, x\rangle)x_j, \quad a_{x,u}a_{x,v}\phi_{\alpha/2}((-1)^u\langle e_i, x\rangle)\phi_{\alpha/2}((-1)^v\langle e_j, x\rangle)$$

   for $u, v \in \{1, 2\}$, and

$$a_{x,u}b_x\phi_{\alpha/2}((-1)^u\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha}, \quad 2a_{x,u}c_x\phi_{\alpha/2}((-1)^u\langle e_i, x\rangle)x_j\mathbf{1}_{x_j \in (0,4\alpha)}$$

   by using $0 \leq \mathbf{1}_{x_j \geq 4\alpha} \leq \frac{x_j}{4\alpha}$ and $0 \leq x_j \mathbf{1}_{x_j \in (0,4\alpha)} \leq x_j$.
2. Terms that can be bounded by Lemma 21. These include

$$-\phi_\beta(\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha}x_j, \quad \phi_\beta(-\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha}x_j,$$

$$(-1)^u a_{x,v}\phi_\beta((-1)^{1+u}\langle e_i, x\rangle)\mathbf{1}_{x_i \geq 4\alpha}\phi_{\alpha/2}((-1)^v\langle e_j, x\rangle),$$

$$(-1)^{u+v}\phi_\beta((-1)^{1+u}\langle e_i, x\rangle)\mathbf{1}_{x_i\geq 4\alpha}\mathbf{1}_{x_j\geq 4\alpha}\phi_\beta((-1)^{1+v}\langle e_j, x\rangle)$$

for $u, v \in \{1, 2\}$. Also include

$$2(-1)^u c_x \phi_\beta((-1)^{1+u}\langle e_i, x\rangle)\mathbf{1}_{x_i\geq 4\alpha} x_j \mathbf{1}_{x_j\in(0,4\alpha)}$$

by using$0 \leq x_j\mathbf{1}_{x_j\in(0,4\alpha)} \leq x_j$. Also include

$$2(-1)^u b_x \phi_\beta((-1)^{1+u}\langle e_i, x\rangle)\mathbf{1}_{x_i\geq 4\alpha}\mathbf{1}_{x_j\geq 4\alpha}$$

by using $0 \leq \mathbf{1}_{x_j\geq 4\alpha} \leq \frac{x_j}{4\alpha}$.

3. Terms that can be bounded by Lemma 18. These include

$$b_x\mathbf{1}_{x_i\geq 4\alpha}x_j, \quad b_x^2\mathbf{1}_{x_i\geq 4\alpha}\mathbf{1}_{x_j\geq 4\alpha}, \quad 2b_x c_x\mathbf{1}_{x_i\geq 4\alpha}x_i\mathbf{1}_{x_j\in(0,4\alpha)}x_j.$$

Where agin we use the fact that $0 \leq \mathbf{1}_{x_i\geq 4\alpha} \leq \frac{x_j}{4\alpha}$ and $0 \leq x_i, \mathbf{1}_{x_j\in(0,4\alpha)} \leq 1$

4. Terms that can be bounded by Lemma 17. These include

$$c_x 2x_i\mathbf{1}_{x_i\in(0,4\alpha)}x_j, \quad 4c_x^2 x_i\mathbf{1}_{x_i\in(0,4\alpha)}x_j\mathbf{1}_{x_j\in(0,4\alpha)}.$$

Where we use the fact that $0 \leq \mathbf{1}_{x_j\in(0,4\alpha)} \leq 1$.

Putting everything together, when $0 < \beta \leq \alpha$,

$$\|\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2 = O\left(C_2\right)$$

where

$$C_2 = \frac{\ell^4 r^3 m}{\alpha\beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2\alpha^{2.5}\beta} + \frac{\ell^2 kr}{D\beta}\left(\frac{m}{Dk}\right)^{\frac{q}{2q+2}} + \frac{\ell^3 r^2\sqrt{km}}{D^{1.5}\alpha^2} + \frac{\ell^6 r^4 m}{\alpha^4 D^2} + k\beta\left(\frac{r}{D}\right)^{\frac{2q+1}{2q+2}} + \frac{kr}{D}\alpha^{\frac{q+1}{2}}.$$

This gives the bound on $\mathbb{E}[(\mathbf{\Sigma}'x - z)z^\top]\|_2$. The bound on $\mathbb{E}[(\mathbf{\Sigma}'x - z)x^\top]\|_2$ can be proved by a similar argument.

□

A.2. Update

A.2.1. General Update Lemma

Lemma 10 (Update). *Suppose $\mathbf{\Sigma}_t$ is diagonal and $\mathbf{E}_t$ is off-diagonal for all t. Suppose we have an update rule that is given by*

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(1 - \eta\mathbf{\Delta}) + \eta\mathbf{\Sigma}\mathbf{\Delta} + \eta\mathbf{R}_t$$

for some positive semidefinite matrix $\mathbf{\Delta}$ and some $\mathbf{R}_t$ such that $\|\mathbf{R}_t\|_2 \leq C''$. Then for every $t \geq 0$,

$$\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}\|_2 \leq \|\mathbf{\Sigma}_0 + \mathbf{E}_0 - \mathbf{\Sigma}\|_2(1 - \eta\lambda_{\min}(\mathbf{\Delta}))^t + \frac{C''}{\lambda_{\min}(\mathbf{\Delta})}.$$

Proof of Lemma 10. We know that the update is given by

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} - \mathbf{\Sigma} = (\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma})(1 - \eta\mathbf{\Delta}) + \eta\mathbf{R}_t.$$

If we let

$$\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma} = (\mathbf{\Sigma}_0 + \mathbf{E}_0 - \mathbf{\Sigma})(1 - \eta\mathbf{\Delta})^t + \mathbf{C}_t.$$

Then we can see that the update rule of $\mathbf{C}_t$ is given by

$$\mathbf{C}_0 = 0,$$
$$\mathbf{C}_{t+1} = \mathbf{C}_t(1 - \eta\mathbf{\Delta}) + \eta\mathbf{R}_t$$

which implies that $\forall t \geq 0, \|\mathbf{C}_t\|_2 \leq \frac{C''}{\lambda_{\min}(\mathbf{\Delta})}$.

Putting everything together completes the proof. $\square$

Lemma 11 (Stage)**.** *In the same setting as Lemma 10, suppose initially for $\ell_1, \ell_2 \in [0, \frac{1}{8})$ we have*

$$\|\mathbf{\Sigma}_0 - \mathbf{I}\|_2 \leq \ell_1, \|\mathbf{E}_0\| \leq \ell_2, \mathbf{\Sigma} = (\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}])^{-1}.$$

Moreover, suppose in each iteration, the error $\mathbf{R}_t$ satisfies that $\|\mathbf{R}_t\|_2 \leq \frac{\lambda_{\min}(\mathbf{\Delta})}{160}\ell_2$.

Then after $t = \frac{\log \frac{400}{\ell_2}}{\eta\lambda_{\min}(\mathbf{\Delta})}$ iterations, we have

1. $\|\mathbf{\Sigma}_t - \mathbf{I}\|_2 \leq \ell_1 + 4\ell_2$,
2. $\|\mathbf{E}_t\|_2 \leq \frac{1}{40}\ell_2$.

Proof of Lemma 11. Using Taylor expansion, we know that

$$\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}] = \mathbf{\Sigma}_0^{-1} + \sum_{i=1}^{\infty} \mathbf{\Sigma}_0^{-1/2}\mathsf{Diag}[(-\mathbf{\Sigma}_0^{-1/2}\mathbf{E}_0\mathbf{\Sigma}_0^{-1/2})^i]\mathbf{\Sigma}_0^{-1/2}.$$

Since $\|\mathsf{Diag}(\mathbf{M})\|_2 \leq \|\mathbf{M}\|_2$ for any matrix $\mathbf{M}$,

$$\begin{aligned}
\|\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}] - \mathbf{\Sigma}_0^{-1}\|_2 &= \|\sum_{i=1}^{\infty} \mathbf{\Sigma}_0^{-1/2}\mathsf{Diag}[(-\mathbf{\Sigma}_0^{-1/2}\mathbf{E}_0\mathbf{\Sigma}_0^{-1/2})^i]\mathbf{\Sigma}_0^{-1/2}\|_2 \\
&\leq \|\sum_{i=1}^{\infty} \mathbf{\Sigma}_0^{-1/2}(-\mathbf{\Sigma}_0^{-1/2}\mathbf{E}_0\mathbf{\Sigma}_0^{-1/2})^i\mathbf{\Sigma}_0^{-1/2}\|_2 \\
&= \|[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}](-\mathbf{E}_0\mathbf{\Sigma}_0^{-1})\|_2 \\
&\leq \frac{\ell_2}{(1-\ell_1)(1-\ell_1-\ell_2)} \leq \frac{32}{21}\ell_2.
\end{aligned}$$

Therefore,

$$\|\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}]\mathbf{\Sigma}_0 - \mathbf{I}\|_2 \leq \frac{\ell_2(1+\ell_1)}{(1-\ell_1)(1-\ell_1-\ell_2)} \leq \ell := \frac{12}{7}\ell_2.$$

which gives us

$$\|\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}]^{-1}\mathbf{\Sigma}_0^{-1} - \mathbf{I}\|_2 \leq \frac{\ell}{1-\ell} \leq \frac{24}{11}\ell_2.$$

This then leads to

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_2 \leq \|\mathsf{Diag}[(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}]^{-1} - \mathbf{\Sigma}_0\|_2 \leq \frac{(1+\ell_1)\ell}{1-\ell} \leq 3\ell_2.$$

Now since $\|\mathbf{\Sigma}_0 + \mathbf{E}_0 - \mathbf{\Sigma}\|_2 \leq 4\ell_2 \leq 1$, after $t = \frac{\log \frac{400}{\ell_2}}{\eta\lambda_{\min}(\mathbf{\Delta})}$ iterations, we have

$$\|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}\|_2 \leq \frac{1}{80}\ell_2.$$

Then since $\mathbf{\Sigma}_t - \mathbf{\Sigma} = \mathsf{Diag}[\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}]$, we have

$$\|\mathbf{\Sigma}_t - \mathbf{\Sigma}\|_2 \le \|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}\|_2 \le \frac{1}{80}\ell_2.$$

This implies that

$$\|\mathbf{E}_t\|_2 \le \|\mathbf{\Sigma}_t - \mathbf{\Sigma}\|_2 + \|\mathbf{\Sigma}_t + \mathbf{E}_t - \mathbf{\Sigma}\|_2 \le \frac{1}{40}\ell_2$$

and

$$\|\mathbf{\Sigma}_t - \mathbf{I}\|_2 \le \|\mathbf{\Sigma}_t - \mathbf{\Sigma}\|_2 + \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_2 + \|\mathbf{\Sigma}_0 - \mathbf{I}\|_2 \le \frac{1}{80}\ell_2 + 3\ell_2 + \ell_1 \le \ell_1 + 4\ell_2.$$

□

Corollary 12 (Corollary of Lemma 11)**.** *Under the same setting as Lemma 11, suppose initially $\ell_1 \le \frac{1}{17}$, then*

1. $\ell_1 \le \frac{1}{8}$ *holds true through all stages,*
2. $\ell_2 \le \left(\frac{1}{40}\right)^t$ *after t stages.*

Proof of Corollary 12. The second claim is trivial. For the first claim, we have

$$(\ell_1)_{stage\ s+1} \le (\ell_1)_{stage\ s} + 4(\ell_2)_{stage\ s} \le \cdots \le \frac{1}{17} + \frac{1}{8}\sum_i (1/40)^i \le \frac{1}{8}.$$

□

A.3. Proof of the Main Theorems

With the update lemmas, we are ready to prove the main theorems.

Proof of Theorem 1. For simplicity, we only focus on the expected update. The on-line version can be proved directly from this by noting that the variance of the update is polynomial bounded and setting accordingly a polynomially small η. The expected update of $\mathbf{A}^{(t)}$ is given by

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta(\mathbb{E}[yz^\top] - \mathbf{A}^{(t)}\mathbb{E}[zz^\top])$$

Let us pick $\alpha = \frac{1}{4}$, focus on one stage and write $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma}_0 + \mathbf{E}_0)$. Then the decoding is given by

$$z = \phi_\alpha(\mathbf{A}^\dagger x) = \phi_\alpha((\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}x).$$

Let $\mathbf{\Sigma}, \mathbf{E}$ be the diagonal part and the off diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$. By Lemma 6,

$$\|\mathbb{E}[(\mathbf{\Sigma}x - z)x^\top \mathbf{\Sigma}]\|_2, \|\mathbb{E}[(\mathbf{\Sigma}x - z)z^\top\|_2 = O(C_1).$$

Now, if we write $\mathbf{A}^{(t)} = \mathbf{A}^*(\mathbf{\Sigma}_t + \mathbf{E}_t)$, then the expected update of $\mathbf{\Sigma}_t + \mathbf{E}_t$ is given by

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{\Sigma}^{-1}(\mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{R}_t$$

where $\|\mathbf{R}_t\|_2 = O(C_1)$.

By Lemma 11, as long as $C_1 = O(\sigma_{\min}(\mathbf{\Delta})\|\mathbf{E}_0\|_2) = O\left(\frac{k\lambda}{D}\|\mathbf{E}_0\|_2\right)$, we can make progress. Putting in the expression of C_1 with $\ell \ge \|\mathbf{E}_0\|_2$, we can see that as long as

$$\frac{\beta k r}{D} + \frac{m\ell^4 r^2}{\alpha^3 D^2} + \frac{\ell^2\sqrt{kmr}^{1.5}}{D^{1.5}\beta} + \frac{\ell^4 r^3 m}{\beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2\alpha^2\beta} = O\left(\frac{k\lambda}{D}\ell\right),$$

we can make progress. By setting $\beta = O\left(\frac{\lambda\ell}{r}\right)$, with Corollary 12 we completes the proof. □

Proof of Theorem 4. For simplicity, we only focus on the expected update. The on-line version can be proved directly from this by setting a polynomially small η. The expected update of $\mathbf{A}^{(t)}$ is given by

$$\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)} + \eta(\mathbb{E}[yz^\top] - \mathbf{A}^{(t)}\mathbb{E}[zz^\top]).$$

Let us focus on one stage and write $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma}_0 + \mathbf{E}_0)$. Then the decoding is given by

$$z = \phi_\alpha(\mathbf{A}^\dagger x) = \phi_\alpha((\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}x).$$

Let $\mathbf{\Sigma}, \mathbf{E}$ be the diagonal part and the off diagonal part of $(\mathbf{\Sigma}_0 + \mathbf{E}_0)^{-1}$. By Lemma 8,

$$\|\mathbb{E}[(\mathbf{\Sigma}x - z)x^\top \mathbf{\Sigma}]\|_2, \|\mathbb{E}[(\mathbf{\Sigma}x - z)z^\top\|_2 = O(C_2).$$

Now, if we write $\mathbf{A}^{(t)} = \mathbf{A}^*(\mathbf{\Sigma}_t + \mathbf{E}_t)$, then the expected update of $\mathbf{\Sigma}_t + \mathbf{E}_t$ is given by

$$\mathbf{\Sigma}_{t+1} + \mathbf{E}_{t+1} = (\mathbf{\Sigma}_t + \mathbf{E}_t)(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{\Sigma}^{-1}(\mathbf{\Sigma}\mathbf{\Delta}\mathbf{\Sigma}) + \mathbf{R}_t$$

where $\|\mathbf{R}_t\|_2 = O(C_2)$.

By Lemma 11, as long as $C_2 = O(\sigma_{\min}(\mathbf{\Delta})\|\mathbf{E}_0\|_2) = O\left(\frac{k\lambda}{D}\|\mathbf{E}_0\|_2\right)$, we can make progress. Putting in the expression of C_2 with $\ell \geq \|\mathbf{E}_0\|_2$, we can see that as long as

$$C_2 = \frac{\ell^4 r^3 m}{\alpha\beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2\alpha^{2.5}\beta} + \frac{\ell^2 kr}{D\beta}\left(\frac{m}{Dk}\right)^{\frac{q}{2q+2}} + \frac{\ell^3 r^2\sqrt{km}}{D^{1.5}\alpha^2} + \frac{\ell^6 r^4 m}{\alpha^4 D^2} + k\beta\left(\frac{r}{D}\right)^{\frac{2q+1}{2q+2}} + \frac{kr}{D}\alpha^{\frac{q+1}{2}} = O\left(\frac{k\lambda}{D}\ell\right),$$

we can make progress. Now set

$$\beta = \frac{\lambda\ell}{D\left(\frac{r}{D}\right)^{\frac{2q+1}{2q+2}}}, \alpha = \left(\frac{\lambda\ell}{r}\right)^{\frac{2}{q+1}}$$

and thus in C_2,

1. First term
$$\frac{\ell^{2-\frac{2}{q+1}} k^0 r^{\frac{5q+6}{q+1}} m}{\lambda^{2+\frac{2}{q+1}} D^{2-\frac{1}{q+1}}}$$

2. Second term
$$\frac{\ell^{4-\frac{5}{q+1}} k^0 r^{\frac{7q+16}{2q+2}} m}{\lambda^{1+\frac{5}{q+1}} D^{2-\frac{1}{2q+2}}}$$

3. Third term
$$\frac{\ell^1 k^{\frac{q+2}{2q+2}} r^{\frac{4q+3}{2q+2}} m^{\frac{q}{2q+2}}}{\lambda D^{\frac{3q+1}{2q+2}}}$$

4. Fourth term
$$\frac{\ell^{3-\frac{4}{q+1}} k^{\frac{1}{2}} r^{\frac{4}{q+1}+2} m^{\frac{1}{2}}}{\lambda^{\frac{4}{q+1}} D^{-\frac{3}{2}}}$$

5. Fifth term
$$\frac{\ell^{6-\frac{8}{q+1}} k^0 r^{4+\frac{8}{q+1}} m}{\lambda^{\frac{8}{q+1}} D^2}$$

We need each term to be smaller than $\frac{\lambda k\ell}{D}$, which implies that (we can ignore the constant ℓ)

1. First term:
$$m \le \frac{kD^{1-\frac{1}{q+1}}\lambda^{3+\frac{2}{q+1}}}{r^{5+\frac{1}{q+1}}}$$

2. Second term:
$$m \le \frac{kD^{1-\frac{1}{2q+2}}\lambda^{2+\frac{5}{q+1}}}{r^{\frac{7}{2}+\frac{9}{2q+2}}}$$

3. Third term:
$$m \le \frac{kD^{\frac{q-1}{q}}\lambda^{4+\frac{4}{q}}}{r^{4+\frac{2}{q}}}$$

4. Fourth term:
$$m \le \frac{kD\lambda^{2+\frac{8}{q+1}}}{r^{4+\frac{8}{q+1}}}$$

5. Fifth term:
$$m \le \frac{kD\lambda^{1+\frac{8}{q+1}}}{r^{4+\frac{8}{q+1}}}$$

This is satisfied by our choice of m in the theorem statement.

Then with Corollary 12 we completes the proof. □

A.4. Expectation Lemmas

In this subsection, we assume that x follows (r, k, m, λ)-GCC. Then we show the following lemmas.

A.4.1. Lemmas with only GCC

Lemma 13 (Expectation)**.** *For every $\ell \in [0, 1)$, every vector e such that $\|e\|_2 \le \ell$, for every α such that $\alpha > 2\ell$, we have*

$$\mathbb{E}[\phi_\alpha(\langle e, x\rangle)] \le \frac{16m\ell^4 r^2}{\alpha^2(\alpha - 2\ell)D^2}.$$

Proof of Lemma 13. Without lose of generality, we can assume that all the entries of e are non-negative. Let us denote a new vector g such that

$$g_i = \begin{cases} e_i & \text{if } e_i \ge \frac{\alpha}{2r}; \\ 0 & \text{otherwise.} \end{cases}$$

Due to the fact that $\|x\|_1 \le r$, we can conclude $\langle e - g, x\rangle \le \frac{\alpha}{2r} \times r = \frac{\alpha}{2}$, which implies

$$\phi_{\frac{\alpha}{2}}(\langle g, x\rangle) \ge \frac{1}{2}\phi_\alpha(\langle e, x\rangle).$$

Now we can only focus on g. Since $\|g\|_2 \le \ell$, we know that g has at most $\frac{4\ell^2 r^2}{\alpha^2}$ non-zero entries. Let us then denote the set of non-zero entries of g as $\mathcal{E}$, so we have $|\mathcal{E}| \le \frac{4\ell^2 r^2}{\alpha^2}$.

Suppose the all the x such that $\langle g, x\rangle \ge \frac{\alpha}{2}$ forms a set $\mathcal{S}$ of size S, each $x^{(s)} \in \mathcal{S}$ has probability p_t. Then we have:

$$\mathbb{E}[\phi_\alpha(\langle e, x\rangle)] \le 2\sum_{s\in[S]} p_s\langle g, x^{(s)}\rangle = 2\sum_{s\in[S], i\in\mathcal{E}} p_s g_i x_i^{(s)}.$$

On the other hand, we have:

1. $\forall s \in [S]: \quad \sum_{i\in\mathcal{E}} g_i x_i^{(s)} \ge \frac{\alpha}{2}$.

2. $\forall i \neq j \in [D]: \sum_{s \in [S]} p_s x_i^{(s)} x_j^{(s)} \leq \frac{m}{D^2}$. This is by assumption 5 of the distribution of x.

Using (2) and multiply both side by $g_i g_j$, we get

$$\sum_{s \in [S]} p_s (g_i x_i^{(s)})(g_j x_j^{(s)}) \leq \frac{m g_i g_j}{D^2}$$

Sum over all $j \in \mathcal{E}, j \neq i$, we have:

$$\sum_{s \in [S]} \sum_{j \in \mathcal{E}, j \neq i} p_s (g_i x_i^{(s)})(g_j x_j^{(s)}) \leq \frac{m g_i}{D^2} \left(\sum_{j \in \mathcal{E}, j \neq i} g_j \right) \leq \frac{m g_i}{D^2} \sum_{j \in \mathcal{E}} g_j$$

By (1), and since $\sum_{j \in \mathcal{E}} g_j x_j^{(s)} \geq \frac{\alpha}{2}$ and $g_i \leq \ell, x_i^{(s)} \leq 1$, we can obtain $\sum_{j \in \mathcal{E}, j \neq i} g_j x_j^{(s)} \geq \frac{\alpha}{2} - \ell$. This implies

$$\sum_{s \in [S]} p_s (g_i x_i^{(s)}) \leq \frac{1}{\sum_{j \in \mathcal{E}, j \neq i} g_j x_j^{(s)}} \left(\frac{m g_i}{D^2} \sum_{j \in \mathcal{E}} g_j \right) \leq \frac{2m}{(\alpha - 2\ell) D^2} g_i \sum_{j \in \mathcal{E}} g_j.$$

Summing over i,

$$\sum_{s \in [S], i \in \mathcal{E}} p_s g_i x_i^{(s)} \leq \frac{2m}{(\alpha - 2\ell) D^2} \left(\sum_{j \in \mathcal{E}} g_j \right)^2 \leq \frac{2m}{(\alpha - 2\ell) D^2} |\mathcal{E}| \|g\|_2^2 \leq \frac{8 m \ell^4 r^2}{\alpha^2 (\alpha - 2\ell) D^2}.$$

Putting everything together we complete the proof. ☐

Lemma 14 (Expectation, Matrix)**.** *For every $\ell, \ell' \in [0,1)$, every matrices $\mathbf{E}, \mathbf{E}' \in \mathbb{R}^{D \times D}$ such that $\|\mathbf{E}\|_2, \|\mathbf{E}'\|_2 \leq \ell$, $\alpha \geq 4\ell$ and every $b_x \in [-1,1]$ that depends on x, the following hold.*

1. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \phi_\alpha(\langle [\mathbf{E}]_i, x \rangle) x_j$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{8 \ell^3 r^2 \sqrt{km}}{D^{1.5} \alpha^2}.$$

2. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \phi_\alpha(\langle [\mathbf{E}]_i, x \rangle) \phi_\alpha(\langle [\mathbf{E}']_j, x \rangle)$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{32 \ell^6 r^4 m}{\alpha^4 D^2}.$$

Proof of Lemma 14. Since all the $\phi_\alpha(\langle [\mathbf{E}]_i, x \rangle)$ and x_i are non-negative, without lose of generality we can assume that $b_x = 1$.

1. We have

$$\sum_{j \in [D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\phi_\alpha(\langle [\mathbf{E}]_i, x \rangle) \sum_{j \in [D]} x_j \right] \leq r \mathbb{E}[\phi_\alpha(\langle [\mathbf{E}]_i, x \rangle)] \leq \frac{16 \ell^4 r^3 m}{\alpha^2 (\alpha - 2\ell) D^2}.$$

On the other hand,

$$\sum_{i \in [D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\left(\sum_{i \in [D]} \phi_\alpha(\langle [\mathbf{E}]_i, x \rangle) \right) x_j \right] \leq \mathbb{E}[(u_x \mathbf{E} x) x_j]$$

where u_x is a vector with each entry either 0 or 1 depend on $\langle [\mathbf{E}]_i, x\rangle \geq \alpha$ or not. Note that $\sum_i \langle [\mathbf{E}]_i, x\rangle^2 \leq \ell^2 r$, so u_x can only have at most $\frac{\ell^2 r}{\alpha^2}$ entries equal to 1, so $\|u_x\|_2 \leq \frac{\ell\sqrt{r}}{\alpha}$. This implies that

$$(u_x \mathbf{E} x) \leq \frac{\ell^2 r}{\alpha}.$$

Therefore, $\sum_{i\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] \leq \frac{2\ell^2 rk}{\alpha D}$, which implies that

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \frac{4\sqrt{2}\ell^3 r^2\sqrt{km}}{D^{1.5}\alpha^{1.5}\sqrt{(\alpha-2\ell)}}.$$

2. We have

$$\|\mathbb{E}[\mathbf{M}]\|_1 \leq \max_i \sum_{j\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] = \max_i \mathbb{E}\left[\phi_\alpha(\langle [\mathbf{E}]_i, x\rangle) \sum_{j\in[D]} \phi_\alpha(\langle [\mathbf{E}']_j, x\rangle)\right] \leq \frac{\ell^2 r}{\alpha}\frac{16\ell^4 r^3 m}{D^2\alpha^2(\alpha-2\ell)}.$$

In the same way we can bound $\|\mathbb{E}[\mathbf{M}]\|_\infty$ and get the desired result.

□

A.4.2. Lemmas with $x_i \in \{0,1\}$

Here we present some expectation lemmas when $x_i \in \{0,1\}$.

Lemma 15 (Expectation, Matrix). *For every $\ell, \ell' \in [0,1)$, every matrices $\mathbf{E}, \mathbf{E}' \in \mathbb{R}^{D\times D}$ such that $\|\mathbf{E}\|_2, \|\mathbf{E}'\|_2 \leq \ell$, and $\forall i \in [D], |\mathbf{E}_{i,i}||\mathbf{E}'_{i,i}| \leq \ell'$, then for every $\beta > 4\ell'$ and $\alpha \geq 4\ell$ and every $b_x \in [-1,1]$ that depends on x, the following hold.*

1. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j}\sqrt{\|\mathbb{E}[\mathbf{M}]\|_1\|\mathbb{E}[\mathbf{M}]\|_\infty} = b_x\phi_\beta(\langle [\mathbf{E}]_i, x\rangle)x_i x_j$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1\|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{8\ell^{1.5}r^{1.25}m}{D^{1.5}\beta^{0.5}}.$$

2. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x\phi_\beta(\langle [\mathbf{E}]_i, x\rangle)x_i x_j \phi_\beta(\langle [\mathbf{E}']_j, x\rangle)$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1\|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{8\ell^4 r^3 m}{\beta^2 D^2}.$$

3. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x\phi_\beta(\langle [\mathbf{E}]_i, x\rangle)x_i \phi_\alpha(\langle [\mathbf{E}']_j, x\rangle)$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1\|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{16\ell^5 r^{2.5} m}{D^2\alpha^2\beta}.$$

Proof. This Lemma is a special case of Lemma 21 by setting $\gamma = 1$. □

A.4.3. Lemmas with general x_i

Here we present some expectation lemmas for the general case where $x_i \in [0,1]$ and the distribution of x satisfies the order-q decay condition.

Lemma 16 (General expectation). *Suppose the distribution of x satisfies the order-q decay condition.*

$$\forall i \in [D], \quad \mathbb{E}[x_i] \leq \Pr[x_i \neq 0] \leq \frac{(q+2)2k}{qD}.$$

Proof. Denote $s = \Pr[x_i \neq 0]$. By assumption, $\Pr[x_i \leq \sqrt{\alpha} \mid x_i \neq 0] \leq \alpha^{q/2}$, which implies that $\Pr[x_i > \sqrt{\alpha}] > s(1-\alpha^{q/2})$. Now, since

$$\mathbb{E}[x_i^2] = \int_0^1 \Pr[x_i^2 \geq \alpha] = \int_0^1 \Pr[x_i \geq \sqrt{\alpha}] \leq \frac{2k}{D},$$

We obtain

$$s \leq \frac{2k}{D} \frac{1}{\int_0^1 (1-\alpha^{q/2}) d\alpha} \leq \frac{q+2}{q} \frac{2k}{D}.$$

□

Lemma 17 (Truncated covariance)**.** *For every $\alpha > 0$, every $b_x \in [-1,1]$ that depends on x, the following holds. Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \mathbf{1}_{x_i \leq \alpha} x_i x_j$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{6kr}{D} \alpha^{\frac{q+1}{2}}.$$

Proof of Lemma 17. Again, without lose of generality we can assume that b_x are just 1.

On one hand,

$$\sum_{j \in [D]} \mathbb{E}[\mathbf{1}_{x_i \leq \alpha} x_i x_j] \leq r \mathbb{E}[\mathbf{1}_{x_i \leq \alpha} x_i] \leq r\alpha \mathbb{E}[\mathbf{1}_{0 < x_i \leq \alpha}] = r\alpha \Pr[x_i \in (0, \alpha]].$$

By Lemma 16,

$$\Pr[x_i \in (0,\alpha]] = \Pr[x_i \neq 0] \Pr[x_i \leq \alpha \mid x_i \neq 0] \leq \frac{(q+2)2k}{qD} \alpha^q,$$

and thus

$$\sum_{j \in [D]} \mathbb{E}[\mathbf{1}_{x_i \leq \alpha} x_i x_j] \leq \frac{2(q+2)kr}{qD} \alpha^{q+1} \leq \frac{6kr}{D} \alpha^{q+1}.$$

On the other hand,

$$\sum_{i \in [D]} \mathbb{E}[\mathbf{1}_{x_i \leq \alpha} x_i x_j] \leq \mathbb{E}\left[\left(\sum_{i \in [D]} x_i\right) x_j\right] \leq \frac{6kr}{D}.$$

Putting everything together we completes the proof. □

Lemma 18 (Truncated half covariance)**.** *For every $\alpha > 0$, every $b_x \in [-1,1]$ that depends on x, the following holds. Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \mathbf{1}_{x_i \geq \alpha} x_j$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq 12k \left(\frac{r}{D}\right)^{\frac{2q+1}{2q+2}}$$

Proof of Lemma 18. Without lose of generality, we can assume $b_x = 1$. We know that

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{x_i \geq \alpha} x_j] &\leq \Pr[x_i \neq 0] \mathbb{E}[x_j \mid x_i \neq 0] \\
&\leq \frac{1}{s} \Pr[x_i \neq 0] \mathbb{E}[\mathbf{1}_{x_i \geq s} x_i x_j \mid x_i \neq 0] + \Pr[x_i \neq 0] \mathbb{E}[\mathbf{1}_{x_i < s} \mid x_i \neq 0] \\
&\leq \frac{1}{s} \mathbb{E}[x_i x_j] + s^q \Pr[x_i \neq 0].
\end{aligned}$$

From Lemma 16 we know that $\Pr[x_i \neq 0] \leq \frac{6k}{D}$, which implies that

$$\sum_{i \in [D]} \mathbb{E}[x_i x_j] = \mathbb{E}[\sum_{i \in [D]} x_i x_j] \leq r \mathbb{E}[x_j] \leq r \Pr[x_j \neq 0] \leq \frac{6kr}{D}.$$

Therefore,

$$\sum_{i\in[D]} \mathbb{E}[\mathbf{1}_{x_i\geq\alpha} x_j] \leq 6ks^q + \frac{1}{s}\frac{6kr}{D}$$

Choosing the optimal s, we are able to obtain

$$\sum_{i\in[D]} \mathbb{E}[\mathbf{1}_{x_i\geq\alpha} x_j] \leq \left(\frac{r}{qD}\right)^{q/(q+1)} 6k(q+1) \leq 24k\left(\frac{r}{D}\right)^{q/(q+1)}.$$

On the other hand,

$$\sum_{j\in[D]} \mathbb{E}[\mathbf{1}_{x_i\geq\alpha} x_j] \leq r\mathbb{E}[\mathbf{1}_{x_i>0}] \leq \frac{6k}{D}r.$$

Putting everything together we get the desired bound. $\square$

Lemma 19 (Expectation)**.** *For every $\ell\in[0,1)$, every vector e such that $\|e\|_2\leq\ell$, for every $i\in[D], \alpha>2|e_i|, \gamma>0$, the following hold.*

1.

$$\forall i\in[D]: \ \mathbb{E}[\phi_\alpha(\langle e,x\rangle)\mathbf{1}_{x_i\geq\gamma}] \leq \frac{4\ell^2 rm}{\gamma D^2(\alpha-2|e_i|)}.$$

2. If $e_i=0$, then

$$\forall i\in[D]: \ \mathbb{E}[\phi_\alpha(\langle e,x\rangle)\mathbf{1}_{x_i\geq\gamma}] \leq \frac{24kr\ell^2}{D\alpha}\left(\frac{m}{Dk}\right)^{q/(q+1)}.$$

Proof of Lemma 19. We define g as in Lemma 13. We still have $\|g\|_1 \leq \frac{\|g\|_2^2}{\frac{\alpha}{2r}} \leq \frac{2r\ell^2}{\alpha}$.

1. The value $\phi_\alpha(\langle e,x\rangle)1_{x_i\geq\gamma}$ is non-zero only when $x_i\geq\gamma$. Therefore, we shall only focus on this case.

   Let us again suppose x such that $\langle g,x\rangle\geq\frac{\alpha}{2}$ and $x_i\geq\gamma$ forms a set $\mathcal{S}$ of size S, each $x^{(s)}\in\mathcal{S}$ has probability p_s.

   Claim 20. *(1) $\forall s\in[S]:\quad \sum_{j\in\mathcal{E}} g_j x_j^{(s)} \geq \frac{\alpha}{2}$.*
   (2) $\forall j\neq i\in[D]: \ \sum_{s\in[S]} p_s x_j^{(s)} x_i^{(s)} \leq \frac{m}{D^2}$.
   (3) By Lemma 18,

   $$\forall j\neq i\in[D]: \ \sum_{a\in[S]} p_a x_j^{(a)} \leq \frac{1}{s}\frac{m}{D^2} + s^q\frac{6k}{D} = \frac{6k(q+1)}{D}\left(\frac{m}{6Dkq}\right)^{q/(q+1)}$$

   by choosing optimal s. Moreover, we can directly calculate that

   $$\frac{6k(q+1)}{D}\left(\frac{m}{6Dkq}\right)^{q/(q+1)} \leq \frac{6k}{D}\left(\frac{m}{Dk}\right)^{q/(q+1)}.$$

   With Claim 20(2), multiply both side by g_j and taking the summation,

   $$\sum_{s\in[S], j\in\mathcal{E}, j\neq i} p_s g_j x_j^{(s)} x_i^{(s)} \leq \frac{m}{D^2}\sum_{j\in\mathcal{E}, j\neq i} g_j.$$

   Using the fact that $x_i^{(s)}\geq\gamma$ for every $s\in[S]$, we obtain

   $$\sum_{s\in[S], j\in\mathcal{E}, j\neq i} p_s g_j x_j^{(s)} \leq \frac{m}{\gamma D^2}\sum_{j\in\mathcal{E}, j\neq i} g_j.$$

On the other hand, by Claim 20(1) and the fact that $|e_i| \geq g_i \geq 0$, we know that

$$\sum_{j \in \mathcal{E}, j \neq i} g_j x_j^{(s)} \geq \frac{\alpha}{2} - |e_i|.$$

Using the fact that $x_i^{(s)} \geq \gamma$ for every $s \in [S]$, we obtain

$$\sum_{s \in [S]} p_s \leq \frac{2m}{\gamma(\alpha - 2|e_i|)D^2} \sum_{j \in \mathcal{E}, j \neq i} g_j.$$

Therefore, since $g_i \leq |e_i|$,

$$\begin{aligned} \sum_{s \in [S], j \in \mathcal{E}} p_s g_j x_j^{(s)} &\leq \sum_{s \in [S], j \in \mathcal{E}, j \neq i} p_s g_j x_j^{(s)} + \sum_{s \in [S]} p_s g_i x_i^{(s)} \\ &\leq \frac{m}{\gamma D^2} \left(1 + \frac{2g_i}{\alpha - 2|e_i|}\right) \sum_{j \in \mathcal{E}, j \neq i} g_j \leq \frac{m}{\gamma D^2} \frac{\alpha}{\alpha - 2|e_i|} \|g\|_1 \leq \frac{2\ell^2 rm}{\gamma D^2(\alpha - 2|e_i|)}. \end{aligned}$$

2. When $e_i = 0$, in the same manner, but using Claim 20(3), we obtain

$$\sum_{s \in [S]} p_s g_j x_j^{(s)} \leq \frac{6k}{D} \left(\frac{m}{Dk}\right)^{q/(q+1)} g_j.$$

Summing over $j \in \mathcal{E}, j \neq i$ we have:

$$\sum_{s \in [S], j \in \mathcal{E}, j \neq i} p_s g_j x_j^{(s)} \leq \frac{6k}{D} \left(\frac{m}{Dk}\right)^{q/(q+1)} \frac{2r\ell^2}{\alpha}.$$

$\square$

Lemma 21 (Expectation, Matrix)**.** *For every $\ell, \ell' \in [0, 1)$, every matrices $\mathbf{E}, \mathbf{E}' \in \mathbb{R}^{D \times D}$ such that $\|\mathbf{E}\|_2, \|\mathbf{E}'\|_2 \leq \ell$, and $\forall i \in [D], |\mathbf{E}_{i,i}|, |\mathbf{E}'_{i,i}| \leq \ell'$, every $\beta > 4\ell'$ and $\alpha \geq 4\ell$, every $\gamma > 0$ and every $b_x \in [-1, 1]$ that depends on x, the following hold.*

1. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \phi_\beta(\langle [\mathbf{E}]_i, x\rangle) \mathbf{1}_{x_i \geq \gamma} x_j$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \min\left\{\frac{8\ell^2\sqrt{k} r^{1.5}\sqrt{m}}{\sqrt{\gamma} D^{1.5}\beta}, \frac{12k\ell^2 r}{D\beta}\left(\frac{m}{Dk}\right)^{q/(2q+2)}\right\}.$$

2. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \phi_\beta(\langle [\mathbf{E}]_i, x\rangle) \mathbf{1}_{x_i \geq \gamma} \mathbf{1}_{x_j \geq \gamma} \phi_\beta(\langle [\mathbf{E}']_j, x\rangle)$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{8\ell^4 r^2 m}{\gamma \beta^2 D^2}.$$

3. *Let $\mathbf{M}$ be a matrix such that $[\mathbf{M}]_{i,j} = b_x \phi_\beta(\langle [\mathbf{E}]_i, x\rangle) \mathbf{1}_{x_i \geq \gamma} \phi_\alpha(\langle [\mathbf{E}']_j, x\rangle)$, then*

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1 \|\mathbb{E}[\mathbf{M}]\|_\infty} \leq \frac{16\ell^5 r^{2.5} m}{\sqrt{\gamma} D^2 \alpha^2 \beta}.$$

Proof of Lemma 21. Without loss of generality, assume $b_x = 1$.

1. Since every entry of $\mathbf{M}$ is non-negative, by Lemma 19,

$$\sum_{j\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}\sum_{j\in[D]} x_j\right] \leq r\mathbb{E}[\phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}] \leq \frac{4\ell^2 r^2 m}{\gamma D^2(\beta-2\ell')}$$

and

$$\sum_{j\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] \leq \frac{24kr\ell^2}{D\beta}\left(\frac{m}{Dk}\right)^{q/(q+1)}.$$

On the other hand, as in Lemma 14, we know that

$$\sum_{j\in[D]} \phi_\beta(\langle[\mathbf{E}']_j, x\rangle) \leq \frac{\ell^2 r}{\beta}.$$

Therefore,

$$\sum_{i\in[D]} \mathbb{E}[\phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}x_j] \leq \frac{\ell^2 r}{\beta}\mathbb{E}[x_j] \leq \frac{6k\ell^2 r}{\beta D}.$$

Now, since each entry of $\mathbf{M}$ is non-negative, using $\|\mathbb{E}[\mathbf{M}]\|_2 \leq \sqrt{\|\mathbb{E}[\mathbf{M}]\|_1\|\mathbb{E}[\mathbf{M}]\|_\infty}$, we obtain the desired bound.

2. Since now $\mathbf{M}$ is a "symmetric" matrix, we only need to look at $\sum_{j\in[D]}\mathbb{E}[\mathbf{M}_{i,j}]$, and a similar bound holds for $\sum_{i\in[D]}\mathbb{E}[\mathbf{M}_{i,j}]$.

$$\sum_{j\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}\sum_{j\in[D]} \phi_\beta(\langle[\mathbf{E}']_j, x\rangle)1_{x_j\geq\gamma}\right] \leq \frac{\ell^2 r}{\beta}\frac{4\ell^2 rm}{\gamma D^2(\beta-2\ell')}.$$

The conclusion then follows.

3. On one hand,

$$\sum_{j\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}\sum_{j\in[D]} \phi_\alpha(\langle[\mathbf{E}']_j, x\rangle)\right] \leq \frac{\ell^2 r}{\alpha}\frac{4\ell^2 rm}{\gamma D^2(\beta-2\ell')}.$$

On the other hand,

$$\sum_{i\in[D]} \mathbb{E}[\mathbf{M}_{i,j}] = \mathbb{E}\left[\left(\sum_{i\in[D]} \phi_\beta(\langle[\mathbf{E}]_i, x\rangle)1_{x_i\geq\gamma}\right)\phi_\alpha(\langle[\mathbf{E}']_j, x\rangle)\right] \leq \frac{\ell^2 r}{\beta}\frac{16\ell^4 r^2 m}{D^2\alpha^2(\alpha-2\ell)}.$$

Therefore,

$$\|\mathbb{E}[\mathbf{M}]\|_2 \leq \frac{8\ell^5 r^{2.5} m}{\sqrt{\gamma}D^2\alpha^{1.5}\beta^{0.5}\sqrt{\beta-2\ell'}\sqrt{\alpha-2\ell}} \leq \frac{16\ell^5 r^{2.5} m}{\sqrt{\gamma}D^2\alpha^2\beta}.$$

□

A.5. Robustness

In this subsection, we show that our algorithm is also robust to noise. To demonstrate the idea, we will present a proof for the case when $x_i \in \{0,1\}$. The general case when $x_i \in [0,1]$ follows from the same argument, just with more calculations.

Lemma 22 (Expectation)**.** *For every $\ell, \nu \in [0,1)$, every vector e such that $\|e\|_2 \leq \ell$, every α such that $\alpha > 2\ell + 2\nu$, the following hold.*

1. $\mathbb{E}[\phi_\alpha(\langle e, x\rangle + \nu)] \le \frac{16m\ell^4r^2}{\alpha^2(\alpha-2\ell-2\nu)D^2}$.

2. *If* $e_{i,i} = 0$*, then* $\mathbb{E}[|\langle e_i, x\rangle| x_i] \le \sqrt{\frac{2mkr}{D^3}}$.

Proof of Lemma 22. The proof of this lemma is almost the same as the proof of Lemma 13 with a few modifications.

1. Without lose of generality, we can assume that all the entries of e are non-negative. Let us denote a new vector g such that

$$g_i = \begin{cases} e_i & \text{if } e_i \ge \frac{\alpha}{2r}, \\ 0 & \text{otherwise.} \end{cases}$$

Due to the fact that $\|x\|_1 \le r$, we can conclude $\langle e - g, x\rangle \le \frac{\alpha}{2r} \times r = \frac{\alpha}{2}$, which implies

$$\phi_{\frac{\alpha}{2}}(\langle g, x\rangle + \nu) \ge \frac{1}{2}\phi_\alpha(\langle e, x\rangle + \nu).$$

Now we can only focus on g. Since $\|g\|_2 \le \ell$, we know that g has at most $\frac{4\ell^2r^2}{\alpha^2}$ non-zero entries. Let us then denote the set of non-zero entries of g as $\mathcal{E}$. Then we have $|\mathcal{E}| \le \frac{4\ell^2r^2}{\alpha^2}$.

Suppose the all the x such that $\langle g, x\rangle \ge \frac{\alpha}{2} - \nu$ forms a set $\mathcal{S}$ of size S, each $x^{(s)} \in \mathcal{S}$ has probability p_t. Then

$$\mathbb{E}[\phi_\alpha(\langle e, x\rangle)] \le 2\sum_{s\in[S]} p_s\langle g, x^{(s)}\rangle = 2\sum_{s\in[S], i\in\mathcal{E}} p_s g_i x_i^{(s)}.$$

On the other hand, we have the following claim.

Claim 23. *1.* $\forall s \in [S]: \quad \sum_{i\in\mathcal{E}} g_i x_i^{(s)} \ge \frac{\alpha}{2} - \nu$.

2. $\forall i \ne j \in [D]: \sum_{s\in[S]} p_s x_i^{(s)} x_j^{(s)} \le \frac{m}{D^2}$*. This is by the GCC conditions of the distribution of* x*.*

Using (2) and multiply both side by $g_i g_j$, we get

$$\sum_{s\in[S]} p_s(g_i x_i^{(s)})(g_j x_j^{(s)}) \le \frac{m g_i g_j}{D^2}.$$

Sum over all $j \in \mathcal{E}, j \ne i$,

$$\sum_{s\in[S]}\sum_{j\in\mathcal{E}, j\ne i} p_s(g_i x_i^{(s)})(g_j x_j^{(s)}) \le \frac{m g_i}{D^2}\left(\sum_{j\in\mathcal{E}, j\ne i} g_j\right) \le \frac{m g_i}{D^2}\sum_{j\in\mathcal{E}} g_j.$$

Using (1), and that $\sum_{j\in\mathcal{E}} g_j x_j^{(s)} \ge \frac{\alpha}{2} - \nu$ and $g_i \le \ell, x_i^{(s)} \le 1$, we can obtain

$$\sum_{j\in\mathcal{E}, j\ne i} g_j x_j^{(s)} \ge \frac{\alpha}{2} - \nu - \ell.$$

This implies

$$\sum_{s\in[S]} p_s(g_i x_i^{(s)}) \le \frac{1}{\sum_{j\in\mathcal{E}, j\ne i} g_j x_j^{(s)}}\left(\frac{m g_i}{D^2}\sum_{j\in\mathcal{E}} g_j\right) \le \frac{2m}{(\alpha - 2\nu - 2\ell)D^2} g_i \sum_{j\in\mathcal{E}} g_j.$$

Summing over i,

$$\sum_{s\in[S], i\in\mathcal{E}} p_s g_i x_i^{(s)} \le \frac{2m}{(\alpha - 4\ell)D^2}\left(\sum_{j\in\mathcal{E}} g_j\right)^2 \le \frac{2m}{(\alpha - 2\nu - 2\ell)D^2}|\mathcal{E}|\|g\|_2^2 \le \frac{8m\ell^4r^2}{\alpha^2(\alpha - 2\ell - 2\nu)D^2}.$$

2. We can directly bound this term as follows.

$$\mathbb{E}[|\langle e, x\rangle| x_i] \leq \sum_{j\neq i} |e_j| \mathbb{E}[x_i x_j] \leq \ell \sqrt{\sum_{j\neq i} \mathbb{E}[x_i x_j]^2} \leq \ell \sqrt{\frac{m}{D^2} \sum_{j \neq i} \mathbb{E}[x_i x_j]^2} \leq \ell \sqrt{\frac{2mkr}{D^3}}.$$

□

We show the following lemma saying that even with noise, $\mathbf{A}^\dagger \mathbf{A}^*$ is roughly $(\mathbf{\Sigma} + \mathbf{E})^{-1}$.

Lemma 24 (Noisy inverse)**.** *Let $\mathbf{A} \in \mathbb{R}^{W\times D}$ be a matrix such that $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}$, for diagonal matrix $\mathbf{\Sigma} \succeq \frac{1}{2}\mathbf{I}$, off diagonal matrix $\mathbf{E}$ with $\|\mathbf{E}\|_2 \leq \ell \leq \frac{1}{8}$ and $\|\mathbf{N}\|_2 \leq \frac{1}{4}\sigma_{\min}(\mathbf{A}^*)$. Then*

$$\|\mathbf{A}^\dagger \mathbf{A}^* - (\mathbf{\Sigma} + \mathbf{E})^{-1}\|_2 \leq \frac{2\|\mathbf{N}\|_2}{\left(\frac{1}{2} - \frac{3}{2}\ell\right)\sigma_{\min}(\mathbf{A}^*) - \|\mathbf{N}\|_2} \leq \frac{32\|\mathbf{N}\|_2}{\sigma_{\min}(\mathbf{A}^*)}.$$

Proof of Lemma 24.

$$\begin{aligned}
\|\mathbf{A}^\dagger \mathbf{A}^* - (\mathbf{\Sigma} + \mathbf{E})^{-1}\|_2 &\leq \|\mathbf{A}^\dagger(\mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N})(\mathbf{\Sigma} + \mathbf{E})^{-1} - (\mathbf{\Sigma} + \mathbf{E})^{-1}\|_2 + \|\mathbf{A}^\dagger \mathbf{N}\|_2 \|(\mathbf{\Sigma} + \mathbf{E})^{-1}\|_2 \\
&\leq \|\mathbf{A}^\dagger \mathbf{N}\|_2 \|(\mathbf{\Sigma} + \mathbf{E})^{-1}\| \\
&\leq \frac{2}{(1-\ell)\sigma_{\min}(\mathbf{A})}\|\mathbf{N}\|_2.
\end{aligned}$$

Since $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}$,

$$\sigma_{\min}(\mathbf{A}) \geq \sigma_{\min}(\mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E})) - \|\mathbf{N}\|_2 \geq \left(\frac{1}{2} - \ell\right)\sigma_{\min}(\mathbf{A}^*) - \|\mathbf{N}\|_2.$$

Putting everything together, we are able to obtain

$$\|\mathbf{A}^\dagger \mathbf{A}^* - (\mathbf{\Sigma} + \mathbf{E})^{-1}\|_2 \leq \frac{2}{(1-\ell)\left(\frac{1}{2} - \ell\right)\sigma_{\min}(\mathbf{A}^*) - \|\mathbf{N}\|_2}\|\mathbf{N}\|_2 \leq \frac{2\|\mathbf{N}\|_2}{\left(\frac{1}{2} - \frac{3}{2}\ell\right)\sigma_{\min}(\mathbf{A}^*) - \|\mathbf{N}\|_2}.$$

□

Lemma 25 (Noisy decoding)**.** *Suppose we have $z = \phi_\alpha((\mathbf{\Sigma}' + \mathbf{E}')x + \xi^x)$ for diagonal matrix $\|\mathbf{\Sigma}' - \mathbf{I}\|_2 \leq \frac{1}{2}$ and off diagonal matrix $\mathbf{E}'$ such that $\|\mathbf{E}'\|_2 \leq \ell \leq \frac{1}{8}$ and random variable ξ^x depend on x such that $\|\xi^x\|_\infty \leq \nu$. Then if $\frac{1}{4} > \alpha > 4\ell + 4\nu, m \leq \frac{D}{r^2}$, we have*

$$\|\mathbb{E}[(\mathbf{\Sigma}x - z)x^\top]\|_2, \|\mathbb{E}[(\mathbf{\Sigma}x - z)z^\top]\|_2 = O\left(C_3\right)$$

where

$$C_3 = (\nu + \beta)\frac{kr}{D} + \frac{m\ell^4 r^2}{\alpha^3 D^2} + \frac{\ell^2\sqrt{kmr}r^{1.5}}{D^{1.5}\beta} + \frac{\ell^4 r^3 m}{\beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2 \alpha^2 \beta}.$$

Proof of Lemma 25. Since we have now

$$z_i = \phi_\alpha(\mathbf{\Sigma}'_{i,i}x_i + \langle e_i, x\rangle + \xi^x_i).$$

Like in Lemma 6, we can still show that

$$|\mathbf{\Sigma}'_{i,i}x_i + \langle e_i, x\rangle x_i + \xi^x_i x_i - z_i| \leq \phi_\alpha(\langle e_i, x\rangle + \xi^x_i) \leq \phi_\alpha(\langle e_i, x\rangle + \nu)$$

which implies that there exists $a_{x,\xi} \in [-1, 1]$ that depends on x, ξ such that

$$z_i - \mathbf{\Sigma}'_{i,i}x_i = \langle e_i, x\rangle x_i + \xi^x_i x_i + a_{x,\xi}\phi_\alpha(\langle e_i, x\rangle + \nu).$$

Therefore,

$$
\begin{aligned}
\mathbb{E}[z_i^2] &\leq 3(\mathbf{\Sigma}'_{i,i}+\nu)^2\mathbb{E}[x_i^2]+3\mathbb{E}[\langle e_i,x\rangle^2 x_i^2]+3\mathbb{E}[\phi_\alpha(\langle e_i,x\rangle+\nu)^2] \\
&\leq \frac{6(2+\nu)^2 k}{D}+3\ell^2 r\sqrt{\frac{2mk}{D^3}}+\frac{48(\ell\sqrt{r}+\nu)m\ell^4 r^2}{\alpha^2(\alpha-2\nu-2\ell)D^2} \\
&\leq O\left(\frac{k}{D}\right).
\end{aligned}
$$

Again, from $z_i - \mathbf{\Sigma}'_{i,i}x_i = \langle e_i, x\rangle x_i + \xi_i^x x_i + a_{x,\xi}\phi_\alpha(\langle e_i,x\rangle+\nu)$, following the exact same calculation as in Lemma 6, but using Lemma 22 instead of Lemma 13, we obtain the result. □

Definition 5 ((γ_1,γ_2)-rounded)**.** *A random variable ζ is (γ_1,γ_2) rounded if*

$$\|\mathbb{E}[\zeta\zeta^\top]\|_2 \leq \gamma_1, \quad \|\zeta\|_2 \leq \gamma_2.$$

Theorem 26 (Noise)**.** *Suppose $\mathbf{A}_0$ is (ℓ,ρ)-initialization for $\ell = O(1), \rho = O(\sigma_{\min}(\mathbf{A}^*))$. Suppose that the data is generated from $y^{(t)} = \mathbf{A}^* x^{(t)} + \zeta^{(t)}$, where $\zeta^{(t)}$ is (γ_1,γ_2)-rounded, and $\gamma_2 = O(\sigma_{\min}(\mathbf{A}^*))$.*

Then after poly$(D,\frac{1}{\epsilon})$ iterations, Algorithm 1 outputs a matrix $\mathbf{A}$ such that there exists diagonal matrix $\tilde{\mathbf{\Sigma}} \succeq \frac{1}{2}\mathbf{I}$ with

$$\|\mathbf{A}-\mathbf{A}^*\tilde{\mathbf{\Sigma}}\|_2 = O\left(r\frac{\gamma_2}{\lambda}\frac{\sigma_{\max}(\mathbf{A}^*)}{\sigma_{\min}(\mathbf{A}^*)}+\frac{\sqrt{\gamma_1}}{\lambda}\sqrt{\frac{D}{k}}+\varepsilon\right).$$

Proof of Theorem 26. For notation simplicity, we only consider one stages, and we drop the round number here and let $\widetilde{\mathbf{A}} = \mathbf{A}^{(t+1)}$ and $\mathbf{A} = \mathbf{A}^{(t)}$, and we denote the new decomposition as $\widetilde{\mathbf{A}} = \mathbf{A}^*(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{E}})+\widetilde{\mathbf{N}}$.

Thus, the decoding of z is given by

$$z = \phi_\alpha(\mathbf{A}_0^\dagger(\mathbf{A}^*x+\zeta)).$$

By Lemma 24, there exists a matrix $\mathbf{R}$ such that $\|\mathbf{R}\|_2 \leq \frac{32\|\mathbf{N}_0\|_2}{\sigma_{\min}(\mathbf{A}^*)}$ with $\mathbf{A}_0^\dagger\mathbf{A}^* = (\mathbf{\Sigma}_0+\mathbf{E}_0)^{-1}+\mathbf{R}$. Now let $\mathbf{\Sigma}'+\mathbf{E}' = (\mathbf{\Sigma}_0+\mathbf{E}_0)^{-1}+\mathbf{R}$, where $\mathbf{\Sigma}'$ is diagonal and $\mathbf{E}'$ is off-diagonal. Then

$$z = \phi_\alpha((\mathbf{\Sigma}'+\mathbf{E}')x+\mathbf{A}_0^\dagger\zeta)$$

where

$$\nu := \|\mathbf{A}_0^\dagger\zeta\|_\infty \leq \frac{16\|\zeta\|_2}{\sigma_{\min}(\mathbf{A}^*)} \leq \frac{16\gamma_2}{\sigma_{\min}(\mathbf{A}^*)}.$$

For simplicity, we only focus on the expected update. The on-line version can be proved directly from this by setting a polynomially small η. The expected update is given by

$$\widetilde{\mathbf{A}} = \mathbf{A}+\eta\mathbb{E}[(\mathbf{A}^*x+\zeta)z^\top-\mathbf{A}zz^\top].$$

Therefore,

$$
\begin{aligned}
\mathbf{A}^*(\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{E}})+\widetilde{\mathbf{N}} &= \mathbf{A}+\eta\mathbb{E}[(\mathbf{A}^*x+\zeta)z^\top-\mathbf{A}zz^\top] \\
&= \mathbf{A}^*[(\mathbf{\Sigma}+\mathbf{E})(\mathbf{I}-\eta\mathbb{E}[zz^\top])+\eta\mathbb{E}[xz^\top]]+\mathbf{N}(\mathbf{I}-\eta\mathbb{E}[zz^\top])+\mathbb{E}[\zeta z^\top].
\end{aligned}
$$

So we still have

$$\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{E}} = (\mathbf{\Sigma}+\mathbf{E})(\mathbf{I}-\eta\mathbb{E}[zz^\top])+\eta\mathbb{E}[xz^\top], \quad \widetilde{\mathbf{N}} = \mathbf{N}(\mathbf{I}-\eta\mathbb{E}[zz^\top])+\mathbb{E}[\zeta z^\top].$$

By Lemma 25,

$$\widetilde{\mathbf{\Sigma}}+\widetilde{\mathbf{E}} = (\mathbf{\Sigma}+\mathbf{E})(\mathbf{I}-\eta\mathbf{\Sigma}'\mathbf{\Delta}\mathbf{\Sigma}')+\eta\mathbf{\Delta}\mathbf{\Sigma}'+\mathbf{C}_1$$

$$\widetilde{\mathbf{N}} = \mathbf{N}(\mathbf{I} - \eta \mathbf{\Sigma}' \mathbf{\Delta} \mathbf{\Sigma}') + \mathbb{E}[\zeta z^\top] + \mathbf{N}\mathbf{C}_2.$$

where $\|\mathbf{C}_1\|_2, \|\mathbf{C}_2\|_2 \le C_3$, and

$$C_3 = (\nu + \beta)\frac{kr}{D} + \frac{m\ell^4 r^2}{\alpha^3 D^2} + \frac{\ell^2\sqrt{km}r^{1.5}}{D^{1.5}\beta} + \frac{\ell^4 r^3 m}{\beta^2 D^2} + \frac{\ell^5 r^{2.5} m}{D^2 \alpha^2 \beta}.$$

First, consider the update on $\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{E}}$. By a similar argument as in Lemma 11, we know that as long as $C_3 = O\left(\frac{k}{D}\lambda \|\mathbf{E}_0\|_2\right)$ and $\nu = O(\ell)$, we can reduce the norm of $\mathbf{E}$ by a constant factor in polynomially many iterations. To satisfy the requirement on C_3, we will choose $\alpha = 1/4, \beta = \frac{\lambda}{r}$. Then to make the terms in C_3 small, m is set as follows.

1. Second term:
$$m \le \frac{Dk\lambda}{r^2}.$$
2. Third term:
$$m \le \frac{D\lambda^4 k}{r^5}.$$
3. Fourth term:
$$m \le \frac{D\lambda^3 k}{r^5}.$$
4. Fifth term:
$$m \le \frac{D\lambda^2 k}{r^{3.5}}.$$

This implies that after $\mathsf{poly}(\frac{1}{\varepsilon})$ stages, the final $\mathbf{E}$ will have

$$\|\mathbf{E}\|_2 = O\left(\frac{r\gamma_2}{\lambda \sigma_{\min}(\mathbf{A}^*)} + \varepsilon\right).$$

Next, consider the update on $\widetilde{\mathbf{N}}$. Since the chosen value satisfies $C_3 \le \frac{1}{2}\sigma_{\min}(\mathbf{\Sigma}'\mathbf{\Delta}\mathbf{\Sigma}')$, we have

$$\|\widetilde{\mathbf{N}}\|_2 \le \max\left\{\|\mathbf{N}_0\|_2, \frac{2\|\mathbb{E}[\zeta z^\top]\|_2}{\sigma_{\min}(\mathbf{\Sigma}'\mathbf{\Delta}\mathbf{\Sigma}')}\right\}.$$

For the term $\mathbb{E}[\zeta z^\top]$, we know that for every vectors u, v with norm 1,

$$u^\top \mathbb{E}[\zeta z^\top] v \le \mathbb{E}[|\langle u, \zeta\rangle||\langle z, v\rangle|].$$

Since z is non-negative, we might without loss of generality assume that v is all non-negative, and obtain

$$\mathbb{E}[|\langle u, \zeta\rangle||\langle z, v\rangle|] \le \sqrt{\mathbb{E}[\langle u, \zeta\rangle^2]\mathbb{E}[\langle z, v\rangle^2]} \le \sqrt{\mathbb{E}[\langle u, \zeta\rangle^2]}\sqrt{\max_{i\in[D]}\mathbb{E}[z_i^2]} = O\left(\sqrt{\frac{k\gamma_1}{D}}\right).$$

Putting everything together and applying Corollary 12 across stages complete the proof. □

Proof of Theorem 5. The theorem follows from Theorem 26 and noting that $\frac{\sqrt{\gamma_1}}{\lambda}\sqrt{\frac{D}{k}}$ is smaller than $r\frac{\gamma_2}{\lambda}\frac{\sigma_{\max}(\mathbf{A}^*)}{\sigma_{\min}(\mathbf{A}^*)}$ in order. □

B. Additional Experiments

Here we provide additional experimental results. The first set of experiments in Section B.1 evaluates the performance of our algorithm in the presence of weak initialization, since for our theoretical analysis a warm start is crucial for the convergence. It turns out that our algorithm is not very sensitive to the warm start; even if there is a lot of noise in the initialization, it still produces reasonable results. This allows it to be used in a wide arrange of applications where a strong warm start is hard to achieve.

The second set of experiments in Section B.2 evaluates the performance of the algorithm when the weight x has large sparsity. Note that our current bounds have a slightly strong dependency on the ℓ_1 norm of x. We believe that this is only because we want to make our statement as general as possible, making only assumptions on the first two moments of x. If in addition, for example, x is assumed to have nice third moments, then our bound can be greatly improved. Here we show that empirically, our algorithm indeed works for typical distributions with large sparsity.

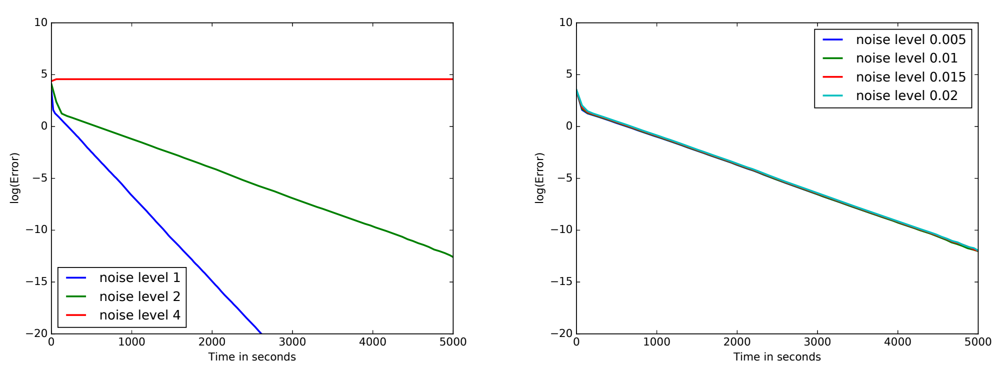
The final set of experiments in Section B.3 applies our algorithm on typical real world applications of NMF. In particular, we consider topic modeling on text data and component analysis for image data, and compare our method to popular existing methods.

B.1. Robustness to Initializations

In all the experiments in the main text, the initialization matrix $\mathbf{A}_0$ is set to $\mathbf{A}_0 = \mathbf{A}^*(\mathbf{I} + \mathbf{U})$ where $\mathbf{I}$ is the identity matrix and $\mathbf{U}$ is a matrix whose entries are i.i.d. samples from the uniform distribution on $[-0.05, 0.05]$. Note that this is a very weak initialization, since $[\mathbf{A}_0]^i = (1 + \mathbf{U}_{i,i})[\mathbf{A}^*]^i + \sum_{j\neq i} \mathbf{U}_{j,i}[\mathbf{A}^*]^j$ and the magnitude of the noise component $\sum_{j\neq i} \mathbf{U}_{j,i}[\mathbf{A}^*]^j$ can be larger than the signal part $(1 + \mathbf{U}_{i,i})[\mathbf{A}^*]^i$.

Here, we further explore even worse initializations: $\mathbf{A}_0 = \mathbf{A}^*(\mathbf{I} + \mathbf{U}) + \mathbf{N}$ where $\mathbf{I}$ is the identity matrix, $\mathbf{U}$ is a matrix whose entries are i.i.d. samples from the uniform distribution on $[-0.05, 0.05] \times r_l$ for a scalar r_l, $\mathbf{N}$ is an additive error matrix whose entries are i.i.d. samples from the uniform distribution on $[-0.05, 0.05] \times r_n$ for a scalar r_n. Here, we call $\mathbf{U}$ the in-span noise and $\mathbf{N}$ the out-of-span noise, since they introduce noise in or out of the span of $\mathbf{A}^*$.

We varied the values of r_l or r_n, and found that even when $\mathbf{U}$ violates our assumptions strongly, or the column norm of $\mathbf{N}$ becomes as large as the column norm of the signal $\mathbf{A}^*$, the algorithm can still recover the ground-truth up to small relative error. Figure 3(a) shows the results for different values of r_l. Note that when $r_l = 1$, the in-span noise already violates our assumptions, but as shown in the figure, even when $r_l = 2$, the ground-truth can still be recovered, though at a slower yet exponential rate. Figure 3(b) shows the results for different values of r_n. For these noise values, the column norm of the noise matrix $\mathbf{N}$ is comparable or even larger than the column norm of the signal $\mathbf{A}^*$, but as shown in the figure, such noise merely affects on the convergence.



(a) Initialization with in-span noise

(b) Initialization with out-of-span noise

Figure 3. The performance of the algorithm AND with weak initialization. The x-axis is the running time (in seconds), the y-axis is the logarithm of the total correlation error. (a) Using different values for the noise level r_l that controls the in-span noise in the initialization. (b) Using different values for the noise level r_n that controls the out-of-span noise in the initialization.

B.2. Robustness to Sparsity

We performed experiments on the DIR data with different sparsity. In particular, construct a 100×5000 matrix $\mathbf{X}$, where each column is drawn from a Dirichlet prior $D(\alpha)$ on $d = 100$ dimension, where $\alpha = (\alpha/d, \alpha/d, \ldots, \alpha/d)$ for a scalar α. Then the dataset is $\mathbf{Y} = \mathbf{A}^*\mathbf{X}$. We varied the α parameter of the prior to control the expected support sparsity, and ran the algorithm on the data generated.

Figure 4 shows the results. For α as large as 20, the algorithm still converges to the ground-truth in exponential rate. When $\alpha = 80$ meaning that the weight vectors (columns in $\mathbf{X}$) have almost full support, the algorithm still produces good results, stabilizing to a small relative error at the end. This demonstates that the algorithm is not sensitive to the support sparsity of the data.

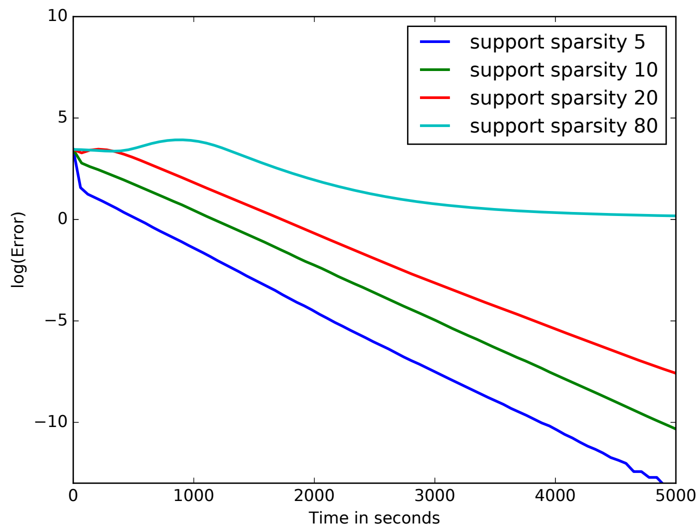


Figure 4. The performance of the algorithm AND on data generated from Dirichlet prior on x with different sparsities. The x-axis is the running time (in seconds), the y-axis is the logarithm of the total correlation error.

B.3. Qualitative Results on Some Real World Applications

We applied our algorithm to two popular applications with real world data to demonstrate the applicability of the method to real world scenarios. Note that the evaluations here are qualitative, due to that the guarantees for our algorithm is the convergence to the ground-truth, while there are no predefined ground-truth for these datasets in practice. Quantitative studies using other criteria computable in practice are left for future work.

B.3.1. Topic Modeling

Here our method is used to compute 10 topics on the 20newsgroups dataset, which is a standard dataset for the topic modeling setting. Our algorithm is initialized with 10 random documents from the dataset, and the hyperparameters like learning rate are from the experiments in the main text. Note that better initialization is possible, while here we keep things simple to demonstrate the power of the method.

Table 1 shows the results of the NMF method and the LDA method in the sklearn package,[10] and the result of our AND method. It shows that our method indeed leads to reasonable topics, with quality comparable to well implemented popular methods tuned to this task.

[10] http://scikit-learn.org/

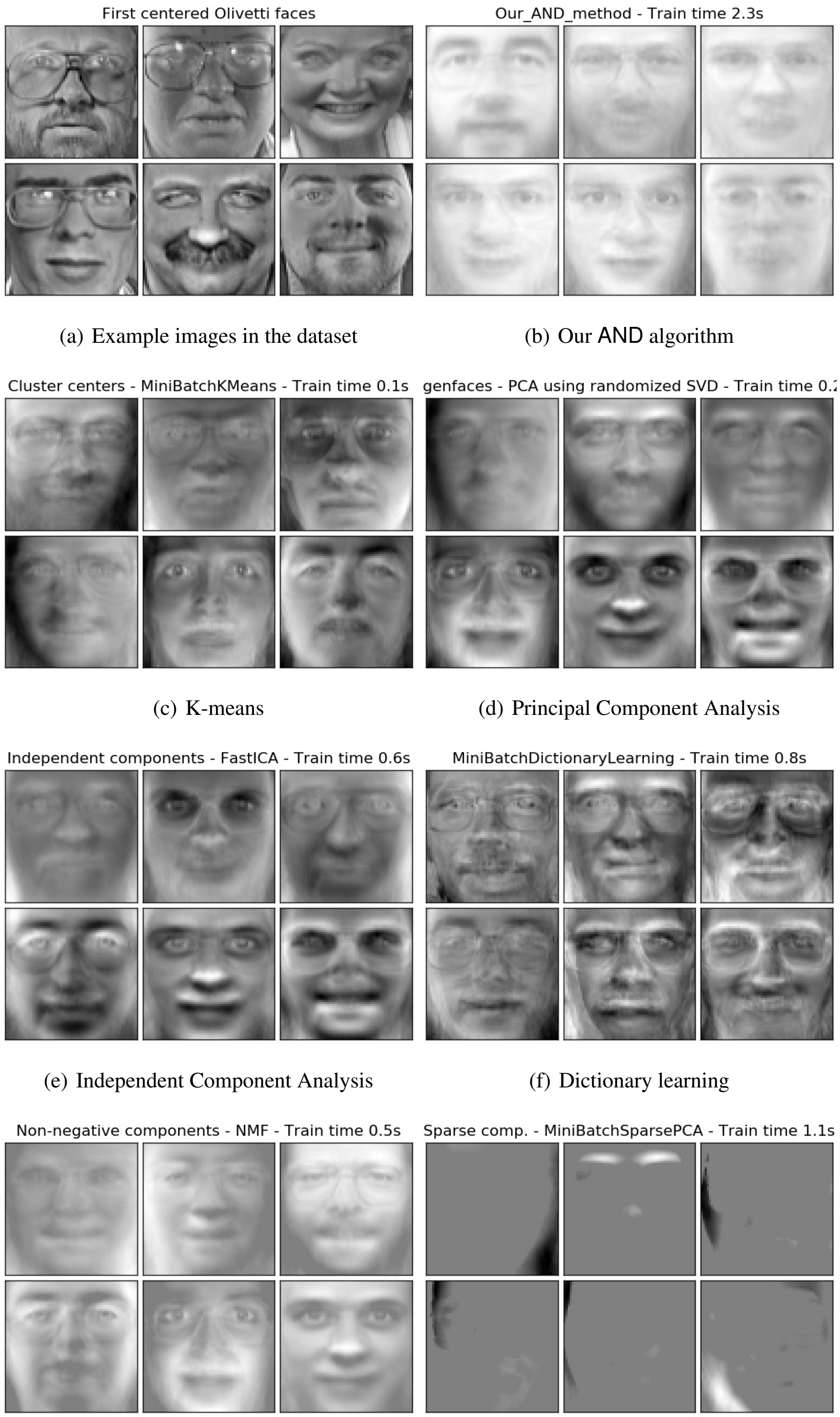
Method	Topic
NMF (sklearn)	just people don think like know time good make way
	windows use dos using window program os drivers application help
	god jesus bible faith christian christ christians does heaven sin
	thanks know does mail advance hi info interested email anybody
	car cars tires miles 00 new engine insurance price condition
	edu soon com send university internet mit ftp mail cc
	file problem files format win sound ftp pub read save
	game team games year win play season players nhl runs
	drive drives hard disk floppy software card mac computer power
	key chip clipper keys encryption government public use secure enforcement
LDA (sklearn)	edu com mail send graphics ftp pub available contact university
	don like just know think ve way use right good
	christian think atheism faith pittsburgh new bible radio games
	drive disk windows thanks use card drives hard version pc
	hiv health aids disease april medical care research 1993 light
	god people does just good don jesus say israel way
	55 10 11 18 15 team game 19 period play
	car year just cars new engine like bike good oil
	people said did just didn know time like went think
	key space law government public use encryption earth section security
AND (ours)	game team year games win play season players 10 nhl
	god jesus does bible faith christian christ new christians 00
	car new bike just 00 like cars power price engine
	key government chip clipper encryption keys use law public people
	young encrypted exactly evidence events especially error eric equipment entire
	thanks know does advance mail hi like info interested anybody
	windows file just don think use problem like files know
	drive drives hard card disk software floppy think mac power
	edu com soon send think mail ftp university internet information
	think don just people like know win game sure edu

Table 1. Results of different methods computing 10 topics on the 20newsgroups dataset. Each topic is visualized by using its top frequent words, and each line presents one topic.

B.3.2. Image Decomposition

Here our method is used to compute 6 components on the Olivetti faces dataset, which is a standard dataset for image decomposition. Our algorithm is initialized with 6 random images from the dataset, and the hyperparameters like learning rate are from the experiments in the main text. Again, note that better initialization is possible, while here we keep things simple to demonstrate the power of the method.

Figure 5 shows some examples from the dataset, the result of our AND method, and 6 other methods using the implementation in the sklearn package. It can be observed that our method can produce meaningful component images, and the non-negative matrix factorization implementation from sklearn produces component images of similar quality. The results of these two methods are generally better than those by the other methods.



(a) Example images in the dataset

(b) Our AND algorithm

(c) K-means

(d) Principal Component Analysis

(e) Independent Component Analysis

(f) Dictionary learning

(g) Non-negative matrix factorization (sklearn)

(h) Sparse Principal Component Analysis

Figure 5. The results of different methods computing 6 components on the Olivetti faces dataset. For all the competitors, we used the implementations in the sklearn package.