Recovery Guarantee of Non-negative Matrix Factorization via Alternating Updates

Yuanzhi Li, Yingyu Liang, Andrej Risteski
Computer Science Department at Princeton University
35 Olden St, Princeton, NJ 08540
{yuanzhil, yingyul, risteski}@cs.princeton.edu

Abstract

Non-negative matrix factorization is a popular tool for decomposing data into feature and weight matrices under non-negativity constraints. It enjoys practical success but is poorly understood theoretically. This paper proposes an algorithm that alternates between decoding the weights and updating the features, and shows that assuming a generative model of the data, it provably recovers the ground-truth under fairly mild conditions. In particular, its only essential requirement on features is linear independence. Furthermore, the algorithm uses ReLU to exploit the non-negativity for decoding the weights, and thus can tolerate adversarial noise that can potentially be as large as the signal, and can tolerate unbiased noise much larger than the signal. The analysis relies on a carefully designed coupling between two potential functions, which we believe is of independent interest.

1 Introduction

In this paper, we study the problem of non-negative matrix factorization (NMF), where given a matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$, the goal to find a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a *non-negative* matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$ such that $\mathbf{Y} \approx \mathbf{A}\mathbf{X}$.[1] $\mathbf{A}$ is often referred to as *feature matrix* and $\mathbf{X}$ referred as *weights*. NMF has been extensively used in extracting a parts representation of the data (e.g., [LS97, LS99, LS01]). It has been shown that the non-negativity constraint on the coefficients forcing features to combine, but not cancel out, can lead to much more interpretable features and improved downstream performance of the learned features.

Despite all the practical success, however, this problem is poorly understood theoretically, with only few provable guarantees known. Moreover, many of the theoretical algorithms are based on heavy tools from algebraic geometry (e.g., [AGKM12]) or tensors (e.g. [AKF$^+$12]), which are still not as widely used in practice primarily because of computational feasibility issues or sensitivity to assumptions on $\mathbf{A}$ and $\mathbf{X}$. Some others depend on specific structure of the feature matrix, such as separability [AGKM12] or similar properties [BGKP16].

A natural family of algorithms for NMF alternate between decoding the weights and updating the features. More precisely, in the decoding step, the algorithm represents the data as a non-negative combination of the current set of features; in the updating step, it updates the features using the decoded representations. This meta-algorithm is popular in practice due to ease of implementation, computational efficiency, and empirical quality of the recovered features. However, even less theoretical analysis exists for such algorithms.

[1] In the usual formulation of the problem, $\mathbf{A}$ is also assumed to be non-negative, which we will not require in this paper.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

This paper proposes an algorithm in the above framework with provable recovery guarantees. To be specific, the data is assumed to come from a generative model $y = \mathbf{A}^* x^* + \nu$. Here, $\mathbf{A}^*$ is the ground-truth feature matrix, x^* are the non-negative ground-truth weights generated from an unknown distribution, and ν is the noise. Our algorithm can provably recover $\mathbf{A}^*$ under mild conditions, even in the presence of large adversarial noise.

Overview of main results. The existing theoretical results on NMF can be roughly split into two categories. In the first category, they make heavy structural assumptions on the feature matrix $\mathbf{A}^*$ such as separability ([AGM12]) or allowing running time exponential in n ([AGKM12]). In the second one, they impose strict distributional assumptions on x^* ([AKF$^+$12]), where the methods are usually based on the method of moments and tensor decompositions and have poor tolerance to noise, which is very important in practice.

In this paper, we present a very simple and natural alternating update algorithm that achieves the best of both worlds. First, we have minimal assumptions on the feature matrix $\mathbf{A}^*$: the only essential condition is linear independence of the features. Second, it is robust to adversarial noise ν which in some parameter regimes be potentially be on the same order as the signal $\mathbf{A}^* x^*$, and is robust to unbiased noise potentially even higher than the signal by a factor of $O(\sqrt{n})$. The algorithm does not require knowing the distribution of x^*, and allows a fairly wide family of interesting distributions. We get this at a rather small cost of a mild "warm start". Namely, we initialize each of the features to be "correlated" with the ground-truth features. This type of initialization is often used in practice as well, for example in LDA-c, the most popular software for topic modeling ([lda16]).

A major feature of our algorithm is the significant robustness to noise. In the presence of adversarial noise on each entry of y up to level C_ν, the noise level $\|\nu\|_1$ can be in the same order as the signal $\mathbf{A}^* x^*$. Still, our algorithm is able to output a matrix $\mathbf{A}$ such that the *final* $\|\mathbf{A}^* - \mathbf{A}\|_1 \le O(\|\nu\|_1)$ in the order of the noise in *one* data point. If the noise is unbiased (i.e., $\mathbb{E}[\nu | x^*] = 0$), the noise level $\|\nu\|_1$ can be $\Omega(\sqrt{n})$ times larger than the signal $\mathbf{A}^* x^*$, while we can still guarantee $\|\mathbf{A}^* - \mathbf{A}\|_1 \le O\left(\|\nu\|_1 \sqrt{n}\right)$ – so our algorithm is not only tolerant to noise, but also has very strong denoising effect. Note that even for the unbiased case the noise can potentially be correlated with the ground-truth in very complicated manner, and also, all our results are obtained only requiring the columns of $\mathbf{A}^*$ are independent.

Technical contribution. The success of our algorithm crucially relies on exploiting the non-negativity of x^* by a ReLU thresholding step during the decoding procedure. Similar techniques have been considered in prior works on matrix factorization, however to the best of our knowledge, the analysis (e.g., [AGMM15]) requires that the decodings are correct in all the intermediate iterations, in the sense that the supports of x^* are recovered with no error. Indeed, we cannot hope for a similar guarantee in our setting, since we consider adversarial noise that could potentially be the same order as the signal. Our major technical contribution is a way to deal with the erroneous decoding through out all the intermediate iterations. We achieve this by a coupling between two potential functions that capture different aspects of the working matrix $\mathbf{A}$. While analyzing iterative algorithms like alternating minimization or gradient descent in non-convex settings is a popular topic in recent years, the proof usually proceeds by showing that the updates are approximately performing gradient descent on an objective with some local or hidden convex structure. Our technique diverges from the common proof strategy, and we believe is interesting in its own right.

Organization. After reviewing related work, we define the problem in Section 3 and describe our main algorithm in Section 4. To emphasize the key ideas, we first present the results and the proof sketch for a simplified yet still interesting case in Section 5, and then present the results under much more general assumptions in Section 6. The complete proof is provided in the appendix.

2 Related work

Non-negative matrix factorization relates to several different topics in machine learning.

Non-negative matrix factorization. The area of non-negative matrix factorization (NMF) has a rich empirical history, starting with the practical algorithm of [LS97].On the theoretical side, [AGKM12] provides a fixed-parameter tractable algorithm for NMF, which solves algebraic equations and thus has poor noise tolerance. [AGKM12] also studies NMF under separability assumptions about the features.

[BGKP16] studies NMF under heavy noise, but also needs assumptions related to separability, such as the existence of dominant features. Also, their noise model is different from ours.

Topic modeling. A closely related problem to NMF is topic modeling, a common generative model for textual data [BNJ03, Ble12]. Usually, $\|x^*\|_1 = 1$ while there also exist work that assume $x_i^* \in [0,1]$ and are independent [ZX12]. A popular heuristic in practice for learning $\mathbf{A}^*$ is *variational inference*, which can be interpreted as alternating minimization in KL divergence norm. On the theory front, there is a sequence of works by based on either spectral or combinatorial approaches, which need certain "non-overlapping" assumptions on the topics. For example, [AGH$^+$13] assume the topic-word matrix contains "anchor words": words which appear in a single topic. Most related is the work of [AR15] who analyze a version of the variational inference updates when documents are long. However, they require strong assumptions on both the warm start, and the amount of "non-overlapping" of the topics in the topic-word matrix.

ICA. Our generative model for x^* will assume the coordinates are independent, therefore our problem can be viewed as a non-negative variant of ICA with high levels of noise. Results here typically are not robust to noise, with the exception of [AGMS12] that tolerates Gaussian noise. However, to best of our knowledge, no result in this setting is provably robust to adversarial noise.

Non-convex optimization. The framework of having a "decoding" for the samples, along with performing an update for the model parameters has proven successful for dictionary learning as well. The original empirical work proposing such an algorithm (in fact, it suggested that the V1 layer processes visual signals in the same manner) was due to [OF97]. Even more, similar families of algorithms based on "decoding" and gradient-descent are believed to be neurally plausible as mechanisms for a variety of tasks like clustering, dimension-reduction, NMF, etc ([PC15, PC14]). A theoretical analysis came latter for dictionary learning due to [AGMM15] under the assumption that the columns of $\mathbf{A}^*$ are incoherent. The technique is not directly applicable to our case, as we don't wish to have any assumptions on the matrix $\mathbf{A}^*$. For instance, if $\mathbf{A}^*$ is non-negative and columns with l_1 norm 1, incoherence effectively means the the columns of $\mathbf{A}^*$ have very small overlap.

3 Problem definition and assumptions

Given a matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$, the goal of non-negative matrix factorization (NMF) is to find a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a non-negative matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$, so that $\mathbf{Y} \approx \mathbf{A}\mathbf{X}$. The columns of $\mathbf{Y}$ are called data points, those of $\mathbf{A}$ are features, and those of $\mathbf{X}$ are weights. We note that in the original NMF, $\mathbf{A}$ is also assumed to be non-negative, which is not required here. We also note that typically $m \gg n$, i.e., the features are a few representative components in the data space. This is different from dictionary learning where overcompleteness is often assumed.

The problem in the worst case is NP-hard [AGKM12], so some assumptions are needed to design provable efficient algorithms. In this paper, we consider a generative model for the data point

$$y = \mathbf{A}^* x^* + \nu \tag{1}$$

where $\mathbf{A}^*$ is the ground-truth feature matrix, x^* is the ground-truth non-negative weight from some unknown distribution, and ν is the noise. Our focus is to recover $\mathbf{A}^*$ given access to the data distribution, assuming some properties of $\mathbf{A}^*$, x^*, and ν. To describe our assumptions, we let $[\mathbf{M}]^i$ denote the i-th row of a matrix $\mathbf{M}$, $[\mathbf{M}]_j$ its i-th column, $\mathbf{M}_{i,j}$ its (i,j)-th entry. Denote its column norm, row norm, and symmetrized norm as $\|\mathbf{M}\|_1 = \max_j \sum_i |\mathbf{M}_{i,j}|$, $\|\mathbf{M}\|_\infty = \max_i \sum_j |\mathbf{M}_{i,j}|$, and $\|\mathbf{M}\|_{\mathrm{s}} = \max\left\{\|\mathbf{M}\|_1, \|\mathbf{M}\|_\infty\right\}$, respectively.

We assume the following hold for parameters $C_1, c_2, C_2, \ell, C_\nu$ to be determined in our theorems.

- **(A1)** The columns of $\mathbf{A}^*$ are linearly independent.
- **(A2)** For all $i \in [n]$, $x_i^* \in [0,1]$, $\mathbb{E}[x_i^*] \le \frac{C_1}{n}$ and $\frac{c_2}{n} \le \mathbb{E}[(x_i^*)^2] \le \frac{C_2}{n}$, and x_i^*'s are independent.
- **(A3)** The initialization $\mathbf{A}^{(0)} = \mathbf{A}^*(\mathbf{\Sigma}^{(0)} + \mathbf{E}^{(0)}) + \mathbf{N}^{(0)}$, where $\mathbf{\Sigma}^{(0)}$ is diagonal, $\mathbf{E}^{(0)}$ is off-diagonal, and
$$\mathbf{\Sigma}^{(0)} \succeq (1-\ell)\mathbf{I}, \quad \left\|\mathbf{E}^{(0)}\right\|_{\mathrm{s}} \le \ell.$$

We consider two noise models.

(**N1**) Adversarial noise: only assume that $\max_i |\nu_i| \le C_\nu$ almost surely.

(**N2**) Unbiased noise: $\max_i |\nu_i| \le C_\nu$ almost surely, and $\mathbb{E}[\nu|x^*] = 0$.

Remarks. We make several remarks about each of the assumptions.
(**A1**) is the assumption about $\mathbf{A}^*$. It only requires the columns of $\mathbf{A}^*$ to be linear independent, which is very mild and needed to ensure identifiability. Otherwise, for instance, if $(\mathbf{A}^*)_3 = \lambda_1(\mathbf{A}^*)_1 + \lambda_2(\mathbf{A}^*)_2$, it is impossible to distinguish between the case when $x_3^* = 1$ and the case when $x_2^* = \lambda_1$ and $x_1^* = \lambda_2$. In particular, we do not restrict the feature matrix to be non-negative, which is more general than the traditional NMF and is potentially useful for many applications. We also do not make incoherence or anchor word assumptions that are typical in related work.

(**A2**) is the assumption on x^*. First, the coordinates are non-negative and bounded by 1; this is simply a matter of scaling. Second, the assumption on the moments requires that, roughly speaking, each feature should appear with reasonable probability. This is expected: if the occurrences of the features are extremely unbalanced, then it will be difficult to recover the rare ones. The third requirement on independence is motivated by that the features should be different so that their occurrences are not correlated. Here we do not stick to a specific distribution, since the moment conditions are more general, and highlight the essential properties our algorithm needs. Example distributions satisfying our assumptions will be discussed later.

The warm start required by (**A3**) means that each feature $A_i^{(0)}$ has a large fraction of the ground-truth feature $\mathbf{A}_i^*$ and a small fraction of the other features, plus some noise outside the span of the ground-truth features. We emphasize that $\mathbf{N}^{(0)}$ is the component of $\mathbf{A}^{(0)}$ outside the column space of $\mathbf{A}^*$, and is not the difference between $\mathbf{A}^{(0)}$ and $\mathbf{A}^*$. This requirement is typically achieved in practice by setting the columns of $\mathbf{A}^{(0)}$ to reasonable "pure" data points that contains one major feature and a small fraction of some other features (e.g. [lda16, AR15]); in this initialization, it is generally believed that $\mathbf{N}^{(0)} = 0$. But we state our theorems to allow some noise $\mathbf{N}^{(0)}$ for robustness in the initialization.

The adversarial noise model (**N1**) is very general, only imposing an upper bound on the entry-wise noise level. Thus, ν can be correlated with x^* in some complicated unknown way. (**N2**) additionally requires it to be zero mean, which is commonly assumed and will be exploited by our algorithm to tolerate larger noise.

4 Main algorithm

```
Algorithm 1 Purification
Input: initialization $\mathbf{A}^{(0)}$, threshold $\alpha$, step size $\eta$, scaling factor $r$, sample size $N$, iterations $T$
for $t = 0, 1, 2, ..., T-1$ do
    Draw examples $y_1, \dots, y_N$.
    (Decode) Compute $\mathbf{A}^\dagger$, the pseudo-inverse of $\mathbf{A}^{(t)}$ with minimum $\|(\mathbf{A})^\dagger\|_\infty$.
        Set $x = \phi_\alpha(\mathbf{A}^\dagger y)$ for each example $y$.    // $\phi_\alpha$ is ReLU activation; see (2) for the definition
    (Update) Update the feature matrix
        $\mathbf{A}^{(t+1)} = (1-\eta)\,\mathbf{A}^{(t)} + r\eta\hat{\mathbb{E}}\left[(y-y')(x-x')^\top\right]$
    where $\hat{\mathbb{E}}$ is over independent uniform $y, y'$ from $\{y_1, \dots, y_N\}$, and $x, x'$ are their decodings.
Output: $\mathbf{A} = \mathbf{A}^{(T)}$
```

Our main algorithm is presented in Algorithm 1. It keeps a working feature matrix and operates in iterations. In each iteration, it first compute the weights for a batch of N examples (*decoding*), and then uses the computed weights to update the feature matrix (*updating*).

The decoding is simply multiplying the example by the pseudo-inverse of the current feature matrix and then passing it through the rectified linear unit (ReLU) ϕ_α with offset α. The pseudo-inverse with minimum infinity norm is used so as to maximize the robustness to noise (see the theorems). The ReLU function ϕ_α operates element-wisely on the input vector v, and for an element v_i, it is defined as

$$\phi_\alpha(v_i) = \max\{v_i - \alpha, 0\}. \tag{2}$$

To get an intuition why the decoding makes sense, suppose the current feature matrix is the ground-truth. Then $\mathbf{A}^\dagger y = \mathbf{A}^\dagger \mathbf{A}^* x^* + \mathbf{A}^\dagger \nu = x^* + \mathbf{A}^\dagger \nu$. So we would like to use a small $\mathbf{A}^\dagger$ and use threshold to remove the noise term.

In the encoding step, the algorithm move the feature matrix along the direction $\mathbb{E}\left[(y - y')(x - x')^\top\right]$. To see intuitively why this is a good direction, note that when the decoding is perfect and there is no noise, $\mathbb{E}\left[(y - y')(x - x')^\top\right] = \mathbf{A}^*$, and thus it is moving towards the ground-truth. Without those ideal conditions, we need to choose a proper step size, which is tuned by the parameters η and r.

5 Results for a simplified case

Our intuitions can be demonstrated in a simplified setting with (**A1**), (**A2'**), (**A3**), and (**N1**), where

(**A2'**) x_i^*'s are independent, and $x_i^* = 1$ with probability s/n and 0 otherwise for a constant $s > 0$.

Furthermore, let $\mathbf{N}^{(0)} = 0$. This is a special case of our general assumptions, with $C_1 = c_2 = C_2 = s$ where s is the parameter in (**A2'**). It is still an interesting setting; as far as we know, there is no existing guarantee of alternating type algorithms for it.

To present our results, we let $(\mathbf{A}^*)^\dagger$ denote the matrix satisfying $(\mathbf{A}^*)^\dagger \mathbf{A}^* = \mathbf{I}$; if there are multiple such matrices we let it denote the one with minimum $\|(\mathbf{A}^*)^\dagger\|_\infty$.

Theorem 1 (Simplified case, adversarial noise)**.** *There exists a* absolute constant $\mathcal{G}$ *such that when Assumption (A1)(A2')(A3) and (N1) are satisfied with* $l = 1/10$, $C_\nu \le \frac{\mathcal{G}c}{\max\left\{m, n\left\|(\mathbf{A}^*)^\dagger\right\|_\infty\right\}}$ *for some* $0 \le c \le 1$, *and* $\mathbf{N}^{(0)} = 0$, *then there exist* α, η, r *such that for every* $0 < \epsilon, \delta < 1$ *and* $N = \text{poly}(n, m, 1/\epsilon, 1/\delta)$ *the following holds with probability at least* $1 - \delta$.

After $T = O\left(\ln \frac{1}{\epsilon}\right)$ *iterations, Algorithm 1 outputs a solution* $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}$ *where* $\mathbf{\Sigma} \succeq (1 - \ell)\mathbf{I}$ *is diagonal,* $\|\mathbf{E}\|_1 \le \epsilon + c$ *is off-diagonal, and* $\|\mathbf{N}\|_1 \le c$.

Remarks. Consequently, when $\|\mathbf{A}^*\|_1 = 1$, we can do normalization $\hat{\mathbf{A}}_i = \mathbf{A}_i / \|\mathbf{A}_i\|_1$, and the normalized output $\hat{\mathbf{A}}$ satisfies

$$\|\hat{\mathbf{A}} - \mathbf{A}^*\|_1 \le \epsilon + 2c.$$

So under mild conditions and with proper parameters, our algorithm recovers the ground-truth in a geometric rate. It can achieve arbitrary small recovery error in the noiseless setting, and achieve error up to the noise limit even with adversarial noise whose level is comparable to the signal.

The condition on ℓ means that a constant warm start is sufficient for our algorithm to converge, which is much better than previous work such as [AR15]. Indeed, in that work, the ℓ needs to even depend on the dynamic range of the entries of $\mathbf{A}^*$ which is problematic in practice.

It is shown that with large adversarial noise, the algorithm can still recover the features up to the noise limit. When $m \ge n\left\|(\mathbf{A}^*)^\dagger\right\|_\infty$, each data point has adversarial noise with ℓ_1 norm as large as $\|\nu\|_1 = C_\nu m = \Omega(c)$, which is in the same order as the signal $\|\mathbf{A}^* x^*\|_1 = O(1)$. Our algorithm still works in this regime. Furthermore, the *final* error $\|\mathbf{A} - \mathbf{A}^*\|_1$ is $O(c)$, in the same order as the *adversarial* noise in *one* data point.

Note the appearance of $\left\|(\mathbf{A}^*)^\dagger\right\|_\infty$ is not surprising. The case when the columns are the canonical unit vectors for instance, which corresponds to $\left\|(\mathbf{A}^*)^\dagger\right\|_\infty = 1$, is expected to be easier than the case when the columns are nearly the same, which corresponds to large $\left\|(\mathbf{A}^*)^\dagger\right\|_\infty$.

A similar theorem holds for the unbiased noise model.

Theorem 2 (Simplified case, unbiased noise)**.** *If Assumption (A1)(A2')(A3) and (N2) are satisfied with* $C_\nu = \frac{\mathcal{G}c\sqrt{n}}{\max\left\{m, n\left\|(\mathbf{A}^*)^\dagger\right\|_\infty\right\}}$ *and the other parameters set as in Theorem 1, then the same guarantee in holds.*

Remarks. With unbiased noise which is commonly assumed in many applications, the algorithm can tolerate noise level $\sqrt{n}$ larger than the adversarial case. When $m \geq n\| \left(\mathbf{A}^*\right)^\dagger \|_\infty$, each data point has adversarial noise with ℓ_1 norm as large as $\|\nu\|_1 = C_\nu m = \Omega(c\sqrt{n})$, which can be $\Omega(\sqrt{n})$ times larger than the signal $\|\mathbf{A}^* x^*\|_1 = O(1)$. The algorithm can recover the ground-truth in this heavy noise regime. Furthermore, the *final* error $\left\|\mathbf{A} - \mathbf{A}^*\right\|_1$ is $O\left(\|\nu\|_1/\sqrt{n}\right)$, which is only $O(1/\sqrt{n})$ fraction of the noise in *one* data point. This is very strong denoising effect and a bit counter-intuitive. It is possible since we exploit the average of the noise for cancellation, and also use thresholding to remove noise spread out in the coordinates.

5.1 Analysis: intuition

A natural approach typically employed to analyze algorithms for non-convex problems is to define a function on the intermediate solution $\mathbf{A}$ and the ground-truth $\mathbf{A}^*$ measuring their distance and then show that the function decreases at each step. However, a single potential function will not be enough in our case, as we argue below, so we introduce a novel framework of maintaining two potential functions which capture different aspects of the intermediate solutions.

Let us denote the intermediate solution and the update as (omitting the superscript (t))

$$\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}, \quad \hat{\mathbb{E}}[(y - y')(x - x')^\top] = \mathbf{A}^*(\widetilde{\mathbf{\Sigma}} + \widetilde{\mathbf{E}}) + \widetilde{\mathbf{N}},$$

where $\mathbf{\Sigma}$ and $\widetilde{\mathbf{\Sigma}}$ are diagonal, $\mathbf{E}$ and $\widetilde{\mathbf{E}}$ are off-diagonal, and $\mathbf{N}$ and $\widetilde{\mathbf{N}}$ are the terms outside the span of $\mathbf{A}^*$ which is caused by the noise. To cleanly illustrate the intuition behind ReLU and the coupled potential functions, we focus on the noiseless case and assume that we have infinite samples.

Since $\mathbf{A}_i = \mathbf{\Sigma}_{i,i}\mathbf{A}^*_i + \sum_{j \neq i} \mathbf{E}_{j,i}\mathbf{A}^*_j$, if the ratio between $\|\mathbf{E}_i\|_1 = \sum_{j\neq i} |\mathbf{E}_{j,i}|$ and $\mathbf{\Sigma}_{i,i}$ gets smaller, then the algorithm is making progress; if the ratio is large at the end, a normalization of $\mathbf{A}_i$ gives a good approximation of $\mathbf{A}^*_i$. So it suffices to show that $\mathbf{\Sigma}_{i,i}$ is always about a constant while $\|\mathbf{E}_i\|_1$ decreases at each iteration. We will focus on $\mathbf{E}$ and consider the update rule in more detail to argue this. After some calculation, we have

$$\mathbf{E} \leftarrow (1 - \eta)\mathbf{E} + r\eta\widetilde{\mathbf{E}}, \qquad \widetilde{\mathbf{E}} = \mathbb{E}[(x^* - (x')^*)\,(x - x')^\top], \tag{3}$$

where x, x' are the decoding for $x^*, (x')^*$ respectively:

$$x = \phi_\alpha\left((\mathbf{\Sigma} + \mathbf{E})^{-1}x^*\right), \qquad x' = \phi_\alpha\left((\mathbf{\Sigma} + \mathbf{E})^{-1}(x')^*\right). \tag{4}$$

To see why the ReLU function matters, consider the case when we do not use it.

$$\begin{aligned}\widetilde{\mathbf{E}} &= \mathbb{E}(x^* - (x')^*)\left[\mathbf{A}^\dagger\mathbf{A}^*(x^* - (x')^*)\right]^\top = \mathbb{E}\left[(x^* - (x')^*)(x^* - (x')^*)^\top\right]\left[(\mathbf{\Sigma} + \mathbf{E})^{-1}\right]^\top \\ &\propto \left[(\mathbf{\Sigma} + \mathbf{E})^{-1}\right]^\top \approx \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{E}\mathbf{\Sigma}^{-1}.\end{aligned}$$

where we used Taylor expansion and the fact that $\mathbb{E}\left[(x^* - (x')^*)(x^* - (x')^*)^\top\right]$ is a scaling of identity. Hence, if we think of $\mathbf{\Sigma}$ as approximately $\mathbf{I}$ and take an appropriate r, the update to the matrix $\mathbf{E}$ is approximately $\mathbf{E} \leftarrow \mathbf{E} - \eta\mathbf{E}^\top$. Since we do not have control over the signs of $\mathbf{E}$ throughout the iterations, the problematic case is when the entries of $\mathbf{E}^\top$ and $\mathbf{E}$ roughly match in signs, which would lead to the entries of $\mathbf{E}$ increasing.

Now we consider the decoding to see why ReLU is important. Ignoring the higher order terms and regarding $\mathbf{\Sigma} = \mathbf{I}$, we have

$$x = \phi_\alpha\left((\mathbf{\Sigma} + \mathbf{E})^{-1}x^*\right) \approx \phi_\alpha\left(\mathbf{\Sigma}^{-1}x^* - \mathbf{\Sigma}^{-1}\mathbf{E}\mathbf{\Sigma}^{-1}x^*\right) \approx \phi_\alpha\left(x^* - \mathbf{E}x^*\right). \tag{5}$$

The problematic term is $\mathbf{E}x^*$. These errors when summed up will be comparable or even larger than the signals, and the algorithm will fail. However, since the signals are non-negative and most coordinates with errors only have small values, thresholding with ReLU properly can remove those errors while keeping a large fraction of the signals. This leads to large $\widetilde{\mathbf{\Sigma}}_{i,i}$ and small $\widetilde{\mathbf{E}}_{j,i}$'s, and then we can choose an r such that $\mathbf{E}_{j,i}$'s keep decreasing while $\mathbf{\Sigma}_{i,i}$'s stay in a certain range.

To get a quantitative bound, we divide $\mathbf{E}$ into its positive part $\mathbf{E}_+$ and its negative part $\mathbf{E}_-$:

$$[\mathbf{E}_+]_{i,j} = \max\left\{\mathbf{E}_{i,j}, 0\right\}, \qquad [\mathbf{E}_-]_{i,j} = \max\left\{-\mathbf{E}_{i,j}, 0\right\}. \tag{6}$$

The reason to do so is the following: when $\mathbf{E}_{i,j}$ is negative, by the Taylor expansion approximation, $\left[(\mathbf{\Sigma}+\mathbf{E})^{-1}x^*\right]_i$ will tend to be more positive and will not be thresholded most of the time. Therefore, $\mathbf{E}_{j,i}$ will turn more positive at next iteration. On the other hand, when $\mathbf{E}_{i,j}$ is positive, $\left[(\mathbf{\Sigma}+\mathbf{E})^{-1}x^*\right]_i$ will tend to be more negative and zeroed out by the threshold function. Therefore, $\mathbf{E}_{j,i}$ will *not* be more negative at next iteration. We will show for positive and negative parts of $\mathbf{E}$:

postive$^{(t+1)} \leftarrow (1-\eta)$positive$^{(t)}+(\eta)$negative$^{(t)}$, negative$^{(t+1)} \leftarrow (1-\eta)$negative$^{(t)}+(\varepsilon\eta)$positive$^{(t)}$

for a small $\varepsilon \ll 1$. Due to ϵ, we can couple the two parts so that a weighted average of them will decrease, which implies that $\|\mathbf{E}\|_s$ is small at the end. This leads to our coupled potential function.[2]

5.2 Analysis: proof sketch

Here we describe a proof sketch for the simplified case while the complete proof for the general case is presented in the appendix. The lemmas here are direct corollaries of those in the appendix.

One iteration. We focus on one update and omit the superscript (t). Recall the definitions of $\mathbf{E}$, $\mathbf{\Sigma}$ and $\mathbf{N}$ in (5.1), and $\widetilde{\mathbf{E}}$, $\widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{N}}$ in (5.1). Our goal is to derive lower and upper bounds for $\widetilde{\mathbf{E}}$, $\widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{N}}$, assuming that $\mathbf{\Sigma}_{i,i}$ falls into some range around 1, while $\mathbf{E}$ and $\mathbf{N}$ are small. This will allow doing induction on them.

First, begin with the decoding. Some calculation shows that, the decoding for $y = \mathbf{A}^*x^* + \nu$ is

$$x = \phi_\alpha\left(\mathbf{Z}x^* + \xi\right), \text{ where } \mathbf{Z} = (\mathbf{\Sigma}+\mathbf{E})^{-1}, \ \xi = -\mathbf{A}^\dagger\mathbf{N}\mathbf{Z}x^* + \mathbf{A}^\dagger\nu. \tag{7}$$

Now, we can present our key lemmas bounding $\widetilde{\mathbf{E}}$, $\widetilde{\mathbf{\Sigma}}$, and $\widetilde{\mathbf{N}}$.

Lemma 3 (Simplified bound on $\widetilde{\mathbf{E}}$, informal)**.** *(1) if* $\mathbf{Z}_{i,j} < 0$*, then* $\left|\widetilde{\mathbf{E}}_{j,i}\right| \le O\left(\frac{1}{n^2}\left(|\mathbf{Z}_{i,j}| + c\right)\right)$, *(2) if* $\mathbf{Z}_{i,j} \ge 0$*, then* $-O\left(\frac{c}{n^2} + \frac{c}{n}|\mathbf{Z}_{i,j}| + \frac{1}{n^2}|\mathbf{Z}_{i,j}|\right) \le \left|\widetilde{\mathbf{E}}_{j,i}\right| \le O\left(\frac{1}{n}\|\mathbf{Z}_{i,j}\|\right)$.

Note that $\mathbf{Z} \approx \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{E}\mathbf{\Sigma}^{-1}$, so $\mathbf{Z}_{i,j} < 0$ corresponds roughly to $\mathbf{E}_{i,j} > 0$. In this case, the upper bound on $|\widetilde{\mathbf{E}}_{j,i}|$ is very small and thus $|\mathbf{E}_{j,i}|$ decreases, as described in the intuition. What is most interesting is the case when $\mathbf{Z}_{i,j} \ge 0$ (roughly $\mathbf{E}_{i,j} < 0$). The upper bound is much larger, corresponding to the intuition that negative $\mathbf{E}_{i,j}$ can contribute a large positive value to $\mathbf{E}_{j,i}$. Fortunately, the lower bounds are of much smaller absolute value, which allows us to show that a potential function that couples Case (1) and Case (2) in Lemma 3 actually decreases; see the induction below.

Lemma 4 (Simplified bound on $\widetilde{\mathbf{\Sigma}}$, informal)**.** $\widetilde{\mathbf{\Sigma}}_{i,i} \ge \Omega(\mathbf{\Sigma}_{i,i}^{-1} - \alpha)/n$.

Lemma 5 (Simplified bound on $\widetilde{\mathbf{N}}$, adversarial noise, informal)**.** $\left|\widetilde{\mathbf{N}}_{i,j}\right| \le O(C_\nu/n)$.

Induction by iterations. We now show how to use the three lemmas to prove the theorem for the adversarial noise, and that for the unbiased noise is similar.

Let $a_t := \left\|\mathbf{E}_+^{(t)}\right\|_{\mathrm{s}}$ and $b_t := \left\|\mathbf{E}_-^{(t)}\right\|_{\mathrm{s}}$, and choose $\eta = \ell/6$. We begin with proving the following three claims by induction on t: at the beginning of iteration t,

(1) $(1-\ell)\mathbf{I} \preceq \mathbf{\Sigma}^{(t)}$

(2) $\left\|\mathbf{E}^{(t)}\right\|_{\mathrm{s}} \le 1/8$, and if $t > 0$, then $a_t + \beta b_t \le \left(1 - \frac{1}{25}\eta\right)(a_{t-1} + \beta b_{t-1}) + \eta h$, for some $\beta \in (1, 8)$, and some small value h,

(3) $\left\|\mathbf{N}^{(t)}\right\|_{\mathrm{s}} \le c/10$.

The most interesting part is the second claim. At a high level, by Lemma 3, we can show that

$$a_{t+1} \le \left(1 - \frac{3}{25}\eta\right)a_t + 7\eta b_t + \eta h, \qquad b_{t+1} \le \left(1 - \frac{24}{25}\eta\right)b_t + \frac{1}{100}\eta a_t + \eta h.$$

[2] Note that since intuitively, $\mathbf{E}_{i,j}$ gets affected by $\mathbf{E}_{j,i}$ after an update, if we have a row which contains negative entries, it is possible that $\|\mathbf{A}_i - \mathbf{A}_i^*\|_1$ increases. So we cannot simply use $\max_i \|\mathbf{A}_i - \mathbf{A}_i^*\|_1$ as a potential function.

Notice that the contribution of b_t to a_{t+1} is quite large (due to the larger upper bound in Case (2) in Lemma 3), but the other terms are much nicer, such as the small contribution of a_t to b_{t+1}. This allows to choose a $\beta \in (1, 8)$ so that $a_{t+1} + \beta b_{t+1}$ leads to the desired recurrence in the second claim. In other words, $a_{t+1} + \beta b_{t+1}$ is our potential function which decreases at each iteration up to the level h. The other claims can also be proved by the corresponding lemmas. Then the theorem follows from the induction claims.

6 More general results

More general weight distributions. Our argument holds under more general assumptions on x^*.

Theorem 6 (Adversarial noise)**.** *There exists an absolute constant $\mathcal{G}$ such that when Assumption (**A0**)-(**A3**) and (**N1**) are satisfied with $l = 1/10$, $C_2 \le 2c_2$, $C_1^3 \le \mathcal{G} c_2^2 n$, $C_\nu \le \left\{ \frac{c_2^2 \mathcal{G} c}{C_1^2 m}, \frac{c_2^4 \mathcal{G} c}{C_1^5 n \left\| (\mathbf{A}^*)^\dagger \right\|_\infty} \right\}$ for $0 \le c \le 1$, and $\left\| \mathbf{N}^{(0)} \right\|_\infty \le \frac{c_2^2 \mathcal{G} c}{C_1^3 \| (\mathbf{A}^*)^\dagger \|_\infty}$, then there exist α, η, r such that for every $0 < \epsilon, \delta < 1$ and $N = \mathrm{poly}(n, m, 1/\epsilon, 1/\delta)$, with probability at least $1 - \delta$ the following holds.*

After $T = O\left(\ln \frac{1}{\epsilon}\right)$ iterations, Algorithm 1 outputs a solution $\mathbf{A} = \mathbf{A}^*(\mathbf{\Sigma} + \mathbf{E}) + \mathbf{N}$ where $\mathbf{\Sigma} \succeq (1 - \ell)\mathbf{I}$ is diagonal, $\|\mathbf{E}\|_1 \le \epsilon + c/2$ is off-diagonal, and $\|\mathbf{N}\|_1 \le c/2$.

Theorem 7 (Unbiased noise)**.** *If Assumption (**A0**)-(**A3**) and (**N2**) are satisfied with $C_\nu = \frac{c_2 \mathcal{G} \sqrt{cn}}{C_1 \max\left\{ m, n \left\| (\mathbf{A}^*)^\dagger \right\|_\infty \right\}}$ and the other parameters set as in Theorem 6, then the same guarantee holds.*

The conditions on C_1, c_2, C_2 intuitively mean that each feature needs to appear with reasonable probability. $C_2 \le 2c_2$ means that their proportions are reasonably balanced. This may be a mild restriction for some applications, and additionally we propose a pre-processing step that can relax this in the next subsection. The conditions allow a rather general family of distributions, so we point out an important special case to provide a more concrete sense of the parameters. For example, for the uniform independent distribution considered in the simplified case, we can actually allow s to be much larger than a constant; our algorithm just requires $s \le \mathcal{G} n$ for a fixed constant $\mathcal{G}$. So it works for uniform sparse distributions even when the sparsity is linear, which is an order of magnitude larger than in the dictionary learning regime. Furthermore, the distributions of x_i^* can be very different, since we only require $C_1^3 = O(c_2^2 n)$. Moreover, all these can be handled without specific structural assumptions on $\mathbf{A}^*$.

More general proportions. A mild restriction in Theorem 6 and 7 is that $C_2 \le 2c_2$, that is, $\max_{i \in [n]} \mathbb{E}[(x_i^*)^2] \le 2 \min_{i \in [n]} \mathbb{E}[(x_i^*)^2]$. To satisfy this, we propose a preprocessing algorithm for balancing $\mathbb{E}[(x_i^*)^2]$. The idea is quite simple: instead of solving $\mathbf{Y} \approx \mathbf{A}^*\mathbf{X}$, we could also solve $\mathbf{Y} \approx [\mathbf{A}^*\mathbf{D}][(\mathbf{D})^{-1}\mathbf{X}]$ for a positive diagonal matrix $\mathbf{D}$, where $\mathbb{E}[(x_i^*)^2]/\mathbf{D}_{i,i}^2$ is with in a factor of 2 from each other. We show in the appendix that this can be done under assumptions as the above theorems, and additionally $\mathbf{\Sigma} \preceq (1 + \ell)\mathbf{I}$ and $\mathbf{E}^{(0)} \ge$ entry-wise. After balancing, one can use Algorithm 1 on the new ground-truth matrix $[\mathbf{A}^*\mathbf{D}]$ to get the final result.

7 Conclusion

A simple and natural algorithm that alternates between decoding and updating is proposed for non-negative matrix factorization and theoretical guarantees are provided. The algorithm provably recovers a feature matrix close to the ground-truth and is robust to noise. Our analysis provides insights on the effect of the ReLU units in the presence of the non-negativity constraints, and the resulting interesting dynamics of the convergence.

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.

References

[AGH$^+$13] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML*, 2013.

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *STOC*, pages 145–162. ACM, 2012.

[AGM12] S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond svd. In *FOCS*, 2012.

[AGMM15] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *COLT*, 2015.

[AGMS12] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *NIPS*, pages 2375–2383, 2012.

[AKF$^+$12] A. Anandkumar, S. Kakade, D. Foster, Y. Liu, and D. Hsu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. Technical report, 2012.

[AR15] Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *NIPS*, pages 2089–2097, 2015.

[BGKP16] Chiranjib Bhattacharyya, Navin Goyal, Ravindran Kannan, and Jagdeep Pani. Nonnegative matrix factorization under heavy noise. In *Proceedings of the 33nd International Conference on Machine Learning*, 2016.

[Ble12] David M Blei. Probabilistic topic models. *Communications of the ACM*, 2012.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[lda16] Lda-c software. `https://github.com/blei-lab/lda-c/blob/master/readme.txt`, 2016. Accessed: 2016-05-19.

[LS97] Daniel D Lee and H Sebastian Seung. Unsupervised learning by convex and conic coding. *NIPS*, pages 515–521, 1997.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LS01] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.

[OF97] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[PC14] Cengiz Pehlevan and Dmitri B Chklovskii. A hebbian/anti-hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features. In *Asilomar Conference on Signals, Systems and Computers*, pages 769–775. IEEE, 2014.

[PC15] Cengiz Pehlevan and Dmitri Chklovskii. A normative theory of adaptive dimensionality reduction in neural networks. In *NIPS*, pages 2260–2268, 2015.

[ZX12] Jun Zhu and Eric P Xing. Sparse topical coding. *arXiv preprint arXiv:1202.3778*, 2012.