On the Convergence of Multiplicative Update Algorithms for Non-negative Matrix Factorization

Chih-Jen Lin
Department of Computer Science
National Taiwan University, Taipei 106, Taiwan
cjlin@csie.ntu.edu.tw

Abstract

Non-negative matrix factorization (NMF) is useful to find basis information of non-negative data. Currently multiplicative updates are a simple and popular way to find the factorization. However, there is no proofs showing that such updates converge to a stationary point of the NMF optimization problem. Stationarity is important as it is a necessary condition of a local minimum. This paper discusses the difficulty of proving the convergence. We propose slight modifications of existing updates and prove their convergence. Techniques invented in this paper can be applied to prove the convergence for other bound-constrained optimization problems.

1 Introduction

Non-negative matrix factorization (NMF) is useful to find basis information of non-negative data (Paatero and Tapper, 1994; Lee and Seung, 1999). Given an $n \times m$ data matrix V with $V_{ij} \geq 0$ and a pre-determined positive integer $r < \min(n, m)$, NMF finds two non-negative matrices $W \in R^{n\times r}$ and $H \in R^{r\times m}$ so that

$$V \approx WH.$$

If each column of V is an object, this method approximates it by a linear combination of r "basis" columns in W. NMF has been applied to many application areas.

The usual way to find W and H is by the following optimization problem, which minimizes the Euclidean distance between V and WH:

$$\begin{aligned} \min_{W,H} \quad & f(W,H) \equiv \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\big(V_{ij} - (WH)_{ij}\big)^2 \\ \text{subject to} \quad & W_{ia} \geq 0, H_{bj} \geq 0, \forall i, a, b, j. \end{aligned} \tag{1}$$

Each non-negative constraint is a "bound constraint," as it relates to only a single variable. We also note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\big(V_{ij} - (WH)_{ij}\big)^2 = \|V - WH\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm.

One may also minimize the (generalized) Kullback-Leibler divergence between V and WH:

$$\begin{aligned} \min_{W,H} \quad & f(W,H) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right) \\ \text{subject to} \quad & W_{ia} \geq 0, H_{bj} \geq 0, \forall i, a, b, j. \end{aligned} \tag{2}$$

A commonly used approach to solve NMF optimization problems ((1)) and (2) is a multiplicative update algorithm by Lee and Seung (2001). Though some papers (Lin, 2005) pointed out its possible slow convergence, this method is popular due to the simplicity. Lee and Seung (2001) proved that the update causes the function value to be non-increasing, but there is no proof yet showing that any limit point is stationary. While optimization problems here may be non-convex and finding a global minimum is difficult, the stationarity is still important–it is a necessary condition of a local minimum. Therefore, existing multiplicative update algorithms lack sound optimization properties. Gonzales and Zhang (2005) presented numerical examples where Lee and Seung's algorithm fails to approach a stationary point. However, due to possible numerical inaccuracy, we think either a convergence proof or a non-convergence example is desired. This paper conducts a detailed study about the convergence properties of multiplicative update methods.

The main difficulty of proving the stationarity comes from the non-negativity constraints. Though multiplicative updates are close to standard fixed-point methods, existing proofs mainly deal with unconstrained situations. Section 2 reviews Lee and Seung's algorithm for (1) and discuss difficulties of proving the convergence. Section 3 proposes a modified algorithm, which has the same computational complexity per iteration. We prove that any limit point is stationary. For the optimization formula (2), Sections 4-5 give two modified algorithms with convergence proofs. Discussion and conclusions are in Section 6.

2 Multiplicative Update for (1) and Its Convergence Issues

Lee and Seung (2001) proposed the following algorithm to solve (1).

Algorithm 1 Multiplicative update for solving (1)

For $k = 1, 2, \ldots$

$$H_{bj}^{k+1} = H_{bj}^{k} \frac{((W^k)^T V)_{bj}}{((W^k)^T W^k H^k)_{bj}}, \quad \forall b, j. \tag{3}$$

$$W_{ia}^{k+1} = W_{ia}^{k} \frac{(V (H^{k+1})^T)_{ia}}{(W^k H^{k+1} (H^{k+1})^T)_{ia}}, \quad \forall i, a. \tag{4}$$

This procedure is not well-defined if denominators in (3) or (4) are zero. Lin (2005) discusses conditions avoiding such difficulties:

Theorem 1 (Theorem 1 in Lin (2005)) *If V has neither zero column nor row, and $W_{ia}^1 > 0$ and $H_{bj}^1 > 0, \forall i, a, b, j$, then*

$$W_{ia}^k > 0 \textit{ and } H_{bj}^k > 0, \forall i, a, b, j, \forall k \geq 1.$$

We hope that any limit point of $\{W^k, H^k\}$ is stationary as any local minimum must be a stationary point. By definition (W, H) is a stationary point of (1) if it satisfies the Karush-Kuhn-Tucker (KKT) optimality condition (e.g., (Bertsekas, 1999)):

$$\begin{aligned}
&W_{ia} \geq 0, H_{bj} \geq 0, \\
&\nabla_W f(W,H)_{ia} \geq 0, \nabla_H f(W,H)_{bj} \geq 0, \\
&W_{ia} \cdot \nabla_W f(W,H)_{ia} = 0, \text{ and } H_{bj} \cdot \nabla_H f(W,H)_{bj} = 0, \forall i, a, b, j,
\end{aligned} \tag{5}$$

where

$$\nabla_W f(W,H) = (WH - V)H^T \text{ and } \nabla_H f(W,H) = W^T(WH - V), \tag{6}$$

are respectively partial derivatives to elements in W and H.

Lee and Seung (2001) proved the following properties:

1. The function value is non-increasing after every update:

$$f(W^k, H^{k+1}) \le f(W^k, H^k) \qquad \text{and} \qquad f(W^{k+1}, H^{k+1}) \le f(W^k, H^{k+1}). \tag{7}$$

2. If $H_{bj}^k > 0$ and $\nabla_H f(W^k, H^k)_{bj} \neq 0, \forall b, j$, then the first inequality in (7) is strict. Similarly, the second inequality is strict under conditions on W.

Gonzales and Zhang (2005); Lin (2005) pointed out that such properties do not imply the convergence to a stationary point. Clearly Algorithm 1 intends to have a fixed-point update: If $H_{bj}^{k+1} = H_{bj}^k > 0$ and $((W^k)^T W^k H^k)_{bj} \neq 0$, then

$$((W^k)^T V)_{bj} = ((W^k)^T W^k H^k)_{bj} \qquad \text{implies} \qquad \nabla_H f(W^k, H^k)_{bj} = 0,$$

which is part of the KKT condition (5). A convergence proof of fix-point methods for minimizing an unconstrained function $f(\mathbf{x})$ usually involves the following steps:

1. $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ if $\nabla f(\mathbf{x}^k) \neq \mathbf{0}$.

2. From a limit point $\mathbf{x}^*$, if $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$, we can

$$\text{update } \mathbf{x}^* \text{ to } \mathbf{x}^{*+1} \text{ such that } f(\mathbf{x}^{*+1}) < f(\mathbf{x}^*). \tag{8}$$

3. If we have the continuity of $f(\mathbf{x})$ and $\lim_{k\to\infty} \mathbf{x}^{k+1} = \mathbf{x}^{*+1}$, then

$$f(\mathbf{x}^*) \le \lim_{k\to\infty} f(\mathbf{x}^{k+1}) = f(\mathbf{x}^{*+1}) < f(\mathbf{x}^*) \tag{9}$$

causes a contradiction.

Clearly this framework cannot be directly used here because of two difficulties:

1. Though Theorem 1 proves that $W_{ia}^k > 0$ and $H_{bj}^k > 0$, it is unclear if $W_{ia}^* > 0$ and $H_{bj}^* > 0$ or not. Hence in (8) an update from a limit point (W^*, H^*) to (W^*, H^{*+1}) may not be well-defined.

2. If $H^*_{bj} = 0$, we must prove $\nabla_H f(W^*, H^*)_{bj} \geq 0$. This KKT condition is due to non-negative constraints. The above framework does not reveal how to have this result.

Gonzales and Zhang (2005) numerically showed that Algorithm 1 may fail to converge to a stationary point. However, Lin (2005) stated that due to possible numerical inaccuracy, a mathematical example is desired before drawing conclusions. Thus the convergence issue remains open. In the next section we will modify Algorithm 1 so that the two difficulties are conquered. Then any limit point is stationary.

Computational complexity is another concern as we hope that our modifications are not more time consuming. Here we analyze the cost of Algorithm 1. Lin (2005) indicated that in (3) one should calculate $W(HH^T)$ but not $(WH)H^T$ as $r < \min(n, m)$. Hence the main cost is on calculating $(W^k)^T V$ and $V(H^{k+1})^T$ in (3) and (4), each of which takes $O(nmr)$ operations. Therefore, the complexity of Algorithm 1 is

$$\#\text{iterations} \times O(nmr).$$

3 A Modified Multiplicative Update and Its Convergence

Lee and Seung (2001) mentioned that the two update rules (3) and (4) are the same as

$$H^{k+1}_{bj} = H^k_{bj} - \frac{H^k_{bj}}{((W^k)^T W^k H^k)_{bj}} \nabla_H f(W^k, H^k)_{bj}, \quad \forall b, j, \tag{10}$$

$$W^{k+1}_{ia} = W^k_{ia} - \frac{W^k_{ia}}{(W^k H^{k+1} (H^{k+1})^T)_{ia}} \nabla_W f(W^k, H^{k+1})_{ia}, \quad \forall i, a. \tag{11}$$

The algorithm is thus a gradient descent method. For updating H^k_{bj},

$$\frac{H^k_{bj}}{((W^k)^T W^k H^k)_{bj}}$$

is referred to as the step size. The two difficulties raised in Section 2 can be reinterpreted as

1. The denominator of the step size may be zero.

2. If H_{bj}^k, numerator of the step size, is zero, and the gradient $\nabla_H f(W^k, H^k)_{bj} < 0$, H_{bj}^{k+1} is not changed. Hence one cannot use the strategy (8) for proving fixed-point convergence.

Therefore, we propose modifying the step size to:

$$\frac{\bar{H}_{bj}^k}{((W^k)^T W^k \bar{H}^k)_{bj} + \delta},$$

where

$$\bar{H}_{bj}^k \equiv \begin{cases} H_{bj}^k & \text{if } \nabla_H f(W^k, H^k)_{bj} \geq 0, \\ \max(H_{bj}^k, \sigma) & \text{if } \nabla_H f(W^k, H^k)_{bj} < 0. \end{cases} \tag{12}$$

Both σ and δ are pre-defined small positive numbers. Similarly we can define $\bar{W}_{ai}^k$. The modified algorithm is as the following:

Algorithm 2 A modified algorithm for minimizing (1)

1. Given $\sigma > 0$ and $\delta > 0$. Initialize $W_{ia}^1 \geq 0, H_{bj}^1 \geq 0, \forall i, a, b, j$.

2. For $k = 1, 2, \ldots$

   (a) If (W^k, H^k) is stationary, stop.

   (b) Else

$$H_{bj}^{k+1} = H_{bj}^k - \frac{\bar{H}_{bj}^k}{((W^k)^T W^k \bar{H}^k)_{bj} + \delta} \nabla_H f(W^k, H^k)_{bj}, \quad \forall b, j, \tag{13}$$

$$W_{ia}^{k+1} = W_{ia}^k - \frac{\bar{W}_{ia}^k}{(\bar{W}^k H^{k+1} (H^{k+1})^T)_{ia} + \delta} \nabla_W f(W^k, H^{k+1})_{ia}, \quad \forall i, a. \tag{14}$$

This modified algorithm requires extra operations: To find $\bar{H}^k$ (or $\bar{W}^k$) and add δ, each takes $O(mr^2)$ (or $O(nr^2)$). Then $(W^k)^T W^k \bar{H}^k$ costs $O(mr^2)$. All are less less than $O(nmr)$, so the complexity per iteration remains the same.

Due to δ, the new algorithm requires only $W_{ia}^1 \geq 0$ and $H_{bj}^1 \geq 0$ to make denominators in (13) and (14) positive. This result is proved in the following theorem:

Theorem 2 *If $W_{ia}^1 \geq 0$ and $H_{bj}^1 \geq 0, \forall i, a, b, j$, then*

$$W_{ia}^k \geq 0 \text{ and } H_{bj}^k \geq 0, \forall i, a, b, j, \forall k \geq 1. \tag{15}$$

Proof. When $k = 1$, (15) holds by the assumption of this theorem. Using induction, we assume results are correct at k. Then at $(k+1)$, we note that the step size for updating H is non-negative:

$$\frac{\bar{H}^k_{bj}}{((W^k)^T W^k \bar{H}^k)_{bj} + \delta} \geq 0. \tag{16}$$

We then consider two situations:
Case 1: If $\nabla_H f(W^k, H^k)_{bj} < 0$, then using (16),

$$H^{k+1}_{bj} = H^k_{bj} - \frac{\bar{H}^k_{bj}}{((W^k)^T W^k \bar{H}^k)_{bj} + \delta} \nabla_H f(W^k, H^k)_{bj} \geq H^k_{bj} \geq 0.$$

Case 2: If $\nabla_W f(W^k, H^k)_{bj} \geq 0$, then according to (12),

$$\bar{H}^k_{bj} = H^k_{bj}. \tag{17}$$

As $\bar{H}^k$'s components are not smaller than those of H^k, and by assumption (W^k, H^k) have non-negative elements, we have

$$\frac{\bar{H}^k_{bj}}{((W^k)^T W^k \bar{H}^k)_{bj} + \delta} \leq \frac{\bar{H}^k_{bj}}{((W^k)^T W^k H^k)_{bj} + \delta}. \tag{18}$$

Using (17) and (18),

$$\begin{aligned} H^{k+1}_{bj} &\geq H^k_{bj} - \frac{H^k_{bj}}{((W^k)^T W^k H^k)_{bj} + \delta} \nabla_H f(W^k, H^k)_{bj} \\ &= H^k_{bj} \frac{((W^k)^T V)_{bj} + \delta}{((W^k)^T W^k H^k)_{bj} + \delta} \geq 0. \end{aligned}$$

The proof of $W^{k+1}_{ia} \geq 0$ is similar. □

Next we show that from H^k to H^{k+1} all components not satisfying KKT conditions are changed and the function value is decreased. In contrast, elements satisfying KKT conditions remain the same. When W^k is fixed, the function $f(W^k, H)$ is the sum of m functions, each of which relates to only one column of H. It is hence sufficient to consider any column $\mathbf{h}$ and discuss the function

$$\bar{f}(\mathbf{h}) \equiv \frac{1}{2} \|\mathbf{v} - W\mathbf{h}\|^2, \tag{19}$$

where $\mathbf{v}$ and $W = W^k$ are considered as constants.

Lee and Seung (2001) then consider an auxiliary function

$$A(\mathbf{h}, \mathbf{h}^k) \equiv \bar{f}(\mathbf{h}^k) + (\mathbf{h} - \mathbf{h}^k)^T \nabla \bar{f}(\mathbf{h}^k) + \frac{1}{2}(\mathbf{h} - \mathbf{h}^k)^T D(\mathbf{h} - \mathbf{h}^k), \tag{20}$$

where D is a diagonal matrix with

$$D_{bb} \equiv \frac{(W^T W \mathbf{h}^k)_b}{h_b^k}, \forall b = 1, \ldots, r. \tag{21}$$

They proved that

$$\bar{f}(\mathbf{h}) \leq A(\mathbf{h}, \mathbf{h}^k) \leq A(\mathbf{h}^k, \mathbf{h}^k) = f(\mathbf{h}^k). \tag{22}$$

Minimizing $A(\mathbf{h}, \mathbf{h}^k)$ leads to the update rule (3). In addition, if $\bar{f}(\mathbf{h}^k)_b \neq 0$, then $h_b^{k+1} \neq h_b^k$. From (21), this auxiliary function is not well-defined if $h_b^k = 0$. Now we hope that if $h_b^k = 0$ and $\nabla \bar{f}(\mathbf{h}^k)_b < 0$ (i.e., another situation violating the KKT condition), then $A(\mathbf{h}, \mathbf{h}^k)$ is well-defined and h_b^k can be changed as well. Therefore, we define a new auxiliary function on non-KKT indices:

$$\bar{A}(\mathbf{h}, \mathbf{h}^k) \equiv \bar{f}(\mathbf{h}^k) + (\mathbf{h} - \mathbf{h}^k)_I^T \nabla \bar{f}(\mathbf{h}^k)_I + \frac{1}{2}(\mathbf{h} - \mathbf{h}^k)_I^T \bar{D}_{II}(\mathbf{h} - \mathbf{h}^k)_I, \tag{23}$$

where

$$\begin{aligned} I &\equiv \{b \mid h_b^k > 0, \nabla \bar{f}(\mathbf{h}^k)_b \neq 0 \text{ or } h_b^k = 0, \nabla \bar{f}(\mathbf{h}^k)_b < 0\} \\ &= \{b \mid \bar{h}_b^k > 0, \nabla \bar{f}(\mathbf{h}^k)_b \neq 0\}, \end{aligned} \tag{24}$$

and

$$\bar{D}_{bb} \equiv \begin{cases} \frac{((W^k)^T W^k \bar{\mathbf{h}})_b + \delta}{\bar{h}_b^k} & \text{if } b \in I, \\ 0 & \text{if } b \notin I. \end{cases} \tag{25}$$

Our new auxiliary function looks like a straightforward extension of the original one, but this modification is not trivial. While we define $\bar{A}(\mathbf{h}, \mathbf{h}^k)$ so that indices satisfying $h_b^k = 0$ and $\nabla \bar{f}(\mathbf{h}^k)_b < 0$ are taken care of, simultaneously we also need that $\bar{A}(\mathbf{h}, \mathbf{h}^k)$ leads to the non-increasing property (22). This result is shown in the following theorem:

Theorem 3 *Let σ and δ be given in Algorithm 2 and $\mathbf{h}^k$ be any column of H^k. Let I and $\bar{D}$ be defined as in* (24) *and* (25)*, respectively. Let $I' \equiv \{1, \ldots, r\} \backslash I$. Then*

$$\arg\min_{\mathbf{h}_I} \bar{A}(\mathbf{h}, \mathbf{h}^k) = \mathbf{h}_I^k - \bar{D}_{II}^{-1} \nabla \bar{f}(\mathbf{h}^k)_I. \tag{26}$$

Moreover, $\mathbf{h}^{k+1}$ *defined by* (13) *satisfies*

$$\mathbf{h}_I^{k+1} = \arg\min_{\mathbf{h}_I} \bar{A}(\mathbf{h}, \mathbf{h}^k) \text{ and } \mathbf{h}_{I'}^{k+1} = \mathbf{h}_{I'}^k, \tag{27}$$

and

$$\bar{f}(\mathbf{h}^{k+1}) \leq \bar{A}(\mathbf{h}^{k+1}, \mathbf{h}^k) \leq \bar{A}(\mathbf{h}^k, \mathbf{h}^k) = f(\mathbf{h}^k). \tag{28}$$

We further have that the following three properties are equivalent:

1. *Inequalities in* (28) *are strict.*
2. $\nabla \bar{f}(\mathbf{h}^k)_I \neq \mathbf{0}$.
3. $\mathbf{h}^{k+1} \neq \mathbf{h}^k$.

Proof. As $\bar{D}_{II}$ is positive definite, $\bar{A}(\mathbf{h}, \mathbf{h}^k)$ is a strictly convex function of $\mathbf{h}_I$, and has a unique minimum satisfying

$$\bar{D}_{II}(\mathbf{h} - \mathbf{h}^k)_I + \nabla \bar{f}(\mathbf{h}^k)_I = \mathbf{0}. \tag{29}$$

Thus, (26) follows. This result and the update rule (13) then imply (27).

Now $\bar{f}(\mathbf{h})$ is a quadratic function, so

$$\bar{f}(\mathbf{h}) = \bar{f}(\mathbf{h}^k) + (\mathbf{h} - \mathbf{h}^k)^T \nabla \bar{f}(\mathbf{h}^k) + \frac{1}{2}(\mathbf{h} - \mathbf{h}^k)^T (W^T W)(\mathbf{h} - \mathbf{h}^k).$$

For any $\mathbf{h}$ with $\mathbf{h}_{I'} = \mathbf{h}_{I'}^k$,

$$\bar{A}(\mathbf{h}, \mathbf{h}^k) - \bar{f}(\mathbf{h}) = \frac{1}{2}(\mathbf{h} - \mathbf{h}^k)_I^T (\bar{D} - W^T W)_{II} (\mathbf{h} - \mathbf{h}^k)_I. \tag{30}$$

We use a technical Lemma 1 in Appendix A to show that $(\bar{D} - W^T W)_{II}$ is positive definite. Then (30) is non-negative. With (27), the result (28) follows.

Next we prove the three equivalent conditions on the strict decrease of the function value. Clearly (26) and (27) imply that $\nabla \bar{f}(\mathbf{h}^k)_I \neq \mathbf{0}$ if and only if $\mathbf{h}_I^{k+1} \neq \mathbf{h}_I^k$. Using (30) and

$$\bar{A}(\mathbf{h}^{k+1}, \mathbf{h}^k) - \bar{A}(\mathbf{h}^k, \mathbf{h}^k) = -\frac{1}{2}(\mathbf{h}^{k+1} - \mathbf{h}^k)_I^T \bar{D}_{II} (\mathbf{h}^{k+1} - \mathbf{h}^k)_I, \tag{31}$$

(28) is strict if and only if $\mathbf{h}^{k+1} \neq \mathbf{h}^k$. □

Theorem 3 immediately implies that the function value is non-increasing:

Theorem 4 *If Algorithm 2 generates an infinite sequence* $\{W^k, H^k\}$*, then*

$$f(W^{k+1}, H^{k+1}) \le f(W^k, H^{k+1}) \le f(W^k, H^k), \forall k. \tag{32}$$

Moreover, one of the two inequalities is strict.

At this stage one may think that we will use (9) to finish the convergence proof. Instead we first show that H^k and H^{k+1} converge to the same point. With this property, the convergence proof is easier than using (9).

Theorem 5 *Assume* $\{W^k, H^k\}, k \in \mathcal{K}$ *is a convergent sub-sequence and*

$$\lim_{k \in \mathcal{K}, k \to \infty} (W^k, H^k) = (W^*, H^*). \tag{33}$$

Then

$$\lim_{k \in \mathcal{K}, k \to \infty} H^{k+1} = H^*.$$

Proof.

Theorem 4 and (33) imply that $\{f(W^k, H^k)\}$ is a bounded decreasing sequence, which converges to $f(W^*, H^*)$. Thus

$$\lim_{k \in \mathcal{K}, k \to \infty} f(W^k, H^{k+1}) - f(W^k, H^k) = 0.$$

Since $f(W^k, H^{k+1}) - f(W^k, H^k)$ is the sum of the difference at each column, we have

$$\lim_{k \in \mathcal{K}, k \to \infty} \bar{f}(\mathbf{h}^{k+1}) - \bar{f}(\mathbf{h}^k) = 0, \tag{34}$$

where $\mathbf{h}^k$ is any column of H^k. If this theorem is wrong, there is a component b in a column $\mathbf{h}^k$, a value $\epsilon > 0$, and an infinite subset $\hat{\mathcal{K}}$ of $\mathcal{K}$ such that

$$|h_b^{k+1} - h_b^*| \ge \epsilon, \forall k \in \hat{\mathcal{K}}.$$

Using (33), there is an infinite subset $\bar{\mathcal{K}} \subset \hat{\mathcal{K}}$ such that

$$|h_b^k - h_b^*| \le \epsilon/2, \forall k \in \bar{\mathcal{K}}.$$

Combining the above two inequalities we have

$$|h_b^{k+1} - h_b^k| \ge \epsilon/2, \forall k \in \bar{\mathcal{K}}. \tag{35}$$

We claim that $\forall k \in \bar{\mathcal{K}}$, $\bar{h}_b^k > 0$. Otherwise, $\bar{h}_b^k = 0$ implies $h_b^{k+1} = h_b^k$ in (12), which violates (35). Using (28) and (31),

$$\bar{f}(\mathbf{h}^{k+1}) - \bar{f}(\mathbf{h}^k) \leq -\frac{(h_b^{k+1} - h_b^k)^2 \bar{D}_{bb}}{2} \leq -\frac{(h_b^{k+1} - h_b^k)^2 \delta}{2\bar{h}_b^k} \leq -\frac{(h_b^{k+1} - h_b^k)^2 \delta}{2\max(h_b^k, \delta)} \leq 0.$$

With (34), taking the limit of the above inequality we have

$$\lim_{k \in \bar{\mathcal{K}}, k \to \infty} h_b^{k+1} - h_b^k = 0,$$

a contradiction to (35). $\square$

Now we are ready to prove that at any limit point (W^*, H^*), the matrix H^* satisfies KKT optimality conditions:

Theorem 6 *Assume* $\{W^k, H^k\}, k \in \mathcal{K}$ *is a convergent sequence and*

$$\lim_{k \in \mathcal{K}, k \to \infty} (W^k, H^k) = (W^*, H^*).$$

We have that

$$\textit{if } H_{bj}^* > 0, \ \textit{then } \nabla_H f(W^*, H^*)_{bj} = 0, \tag{36}$$

and

$$\textit{if } H_{bj}^* = 0, \ \textit{then } \nabla_H f(W^*, H^*)_{bj} \geq 0. \tag{37}$$

Proof.

By the definition (12),

$$\bar{H}_{bj}^k = \max(H_{bj}^k, \sigma) \qquad \text{or} \qquad H_{bj}^k,$$

so the sequence $\{\bar{H}_{bj}^k\}_{k \in \mathcal{K}}$ may have two convergent points H_{bj}^* or σ. Since the number of (b, j) is finite, there is an infinite set $\bar{\mathcal{K}} \subset \mathcal{K}$ such that

$$\tilde{H}^* \equiv \lim_{k \in \bar{\mathcal{K}}, k \to \infty} \bar{H}^k \text{ exists.} \tag{38}$$

From Theorem 5,

$$\lim_{k \in \bar{\mathcal{K}}, k \to \infty} H_{bj}^k - H_{bj}^{k+1} = \frac{\tilde{H}_{bj}^*}{((W^*)^T W^* \tilde{H}^*)_{bj} + \delta} \nabla_H f(W^*, H^*)_{bj} = 0. \tag{39}$$

Note that $\tilde{H}_{bj}^* \geq H_{bj}^*$. Hence if $H_{bj}^* > 0$, (39) immediately implies (36).

Next we prove (37). If it is wrong, there is (b, j) such that

$$H^*_{bj} = 0 \text{ and } \nabla_H f(W^*, H^*)_{bj} < 0.$$

For all $k \in \bar{\mathcal{K}}$ large enough, $\nabla_H f(W^k, H^k)_{bj} < 0$ and hence

$$\lim_{k \in \bar{\mathcal{K}}, k \to \infty} \bar{H}^k_{bj} = \tilde{H}^*_{bj} = \sigma.$$

Therefore,

$$\frac{\tilde{H}^*_{bj}}{((W^*)^T W^* \tilde{H}^*)_{bj} + \delta} \nabla_H f(W^*, H^*)_{bj} > 0,$$

an inequality contradicting (39). □

The main convergence statement is in the following theorem:

Theorem 7 *Any limit point of the sequence $\{W^k, H^k\}$ generated by Algorithm 2 is a stationary point of (1).*

Proof.

Theorem 6 implies the optimality condition on H^*. Using Theorem 5

$$\lim_{k \in \mathcal{K}, k \to \infty} (W^k, H^{k+1}) = (W^*, H^*).$$

We can then use the same proof in Theorem 6 to have the optimality condition on W^*. □

Earlier work such as (Hoyer, 2002; Piper et al., 2004) adds penalty terms to increase the sparsity of W and H:

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(V_{ij} - (WH)_{ij}\right)^2 + \delta \sum_{i,a} W_{ia} + \delta \sum_{b,j} H_{bj}. \tag{40}$$

Then the update rule (3) becomes

$$\begin{aligned} H^{k+1}_{bj} &= H^k_{bj} - \frac{H^k_{bj}}{((W^k)^T W^k H^k)_{bj} + \delta} (((W^k)^T W^k H^k)_{bj} + \delta - ((W^k)^T V)_{bj}) \\ &= H^k_{bj} \frac{((W^k)^T V)_{bj}}{((W^k)^T W^k H^k)_{bj} + \delta}. \end{aligned}$$

For this formulation, the penalty parameter δ can be directly used in Algorithm 2. The update rule is the same as (13), but $\nabla_H f(W^k, H^k)_{bj}$ involves an additional term δ.

4 Minimizing the Divergence: A Modified Algorithm

We switch to another NMF optimization problem (2), which minimizes the divergence. Lee and Seung (2001) proposed the following update rules:

Algorithm 3 Multiplicative update: minimizing the divergence

For $k = 1, 2, \ldots$

$$H_{bj}^{k+1} = H_{bj}^{k} \frac{\sum_{s=1}^{n} W_{sb}^{k} V_{sj} / (W^k H^k)_{sj}}{\mathbf{e}^T W_{:,b}^{k}} \quad \forall b, j. \tag{41}$$

$$W_{ia}^{k+1} = W_{ia}^{k} \frac{\sum_{s=1}^{m} V_{is} H_{as}^{k+1} / (W^k H^{k+1})_{is}}{H_{a,:}^{k+1} \mathbf{e}} \quad \forall i, a. \tag{42}$$

For matrices W and H, we denote $W_{:,b}$ as the bth column, and $H_{a,:}$ as the ath row, respectively. In (41) and (42), $\mathbf{e}$ is the vector of ones, so $\mathbf{e}^T W_{:,b}$ is the sum of W's bth column.

Similar to the update rules for (1), (41) and (42) can be rewritten as

$$H_{bj}^{k+1} = H_{bj}^{k} - \frac{H_{bj}^{k}}{\mathbf{e}^T W_{:,b}^{k}} \nabla_H f(W^k, H^k)_{bj}, \tag{43}$$

$$W_{ia}^{k+1} = W_{ia}^{k} - \frac{W_{ia}^{k}}{H_{a,:}^{k+1} \mathbf{e}} \nabla_W f(W^k, H^{k+1})_{ia},$$

where

$$\nabla_W f(W,H)_{ia} = H_{a,:}\mathbf{e} - \sum_{s=1}^{m} \frac{V_{is} H_{as}}{(WH)_{is}} \text{ and } \nabla_H f(W,H)_{bj} = \mathbf{e}^T W_{:,b} - \sum_{s=1}^{n} \frac{W_{sb} V_{sj}}{(WH)_{sj}}. \tag{44}$$

The update rule (41) involves two $O(nmr)$ operations:

$$W^k H^k \quad \text{and} \quad \sum_{s=1}^{n} \frac{W_{sb}^{k} V_{sj}}{(W^k H^k)_{sj}}, \quad \forall b, j.$$

Thus each iteration takes four $O(nmr)$ operations, twice of that in minimizing the Euclidean distance.

Before further discussion, we redefine the objective function in (2) as

$$f(W,H) = -\sum_{ij: V_{ij} > 0} V_{ij} \log(WH)_{ij} + \sum_{ij} (WH)_{ij}.$$

Thus $f(W,H)$ is well-defined if and only if $(WH)_{ij} > 0, \forall V_{ij} > 0$.

We may try the same strategy in Section 3 to modify the step size in (43) to

$$\frac{\bar{H}^k_{bj}}{\mathbf{e}^T W^k_{:,b} + \delta}, \tag{45}$$

where $\bar{H}^k_{bj}$ can be defined similarly to (12). However, two new difficulties occur:

1. In addition to $\mathbf{e}^T W_{:,b}$ in the denominator of the step size, $(WH)_{sj} = 0$ in the calculation of $\nabla_H f(W,H)$ may also cause a zero denominator.

2. For the convergence proof, similar to Section 3, we discuss a function of a column $\mathbf{h}$:

$$\bar{f}(\mathbf{h}) \equiv -\sum_{i:v_i>0} v_i \log(W\mathbf{h})_i + \sum_{i=1}^{n} (W\mathbf{h})_i, \tag{46}$$

   where $\mathbf{v}$ and W are constants. This function is different from (19) in Section 3. We need new strategies (e.g, different auxiliary functions) to have the non-increasing property (22).

To address the first issue, we design the algorithm so that it has the following property:

$$\text{If } V_{is} > 0, \text{ then } (W^k H^k)_{is} > 0 \;\; \forall k \text{ and } (W^* H^*)_{is} > 0, \tag{47}$$

where (W^*, H^*) is any limit point of the sequence $\{W^k, H^k\}$. Elements with $V_{ij} = 0$ are not a concern as they do not affect the function value calculation. Moreover, the gradient calculation should then be

$$\nabla_W f(W,H)_{ia} = H_{a,:}\mathbf{e} - \sum_{s:V_{is}>0} \frac{V_{is} H_{as}}{(WH)_{is}} \text{ and } \nabla_H f(W,H)_{bj} = \mathbf{e}^T W_{:,b} - \sum_{s:V_{sj}>0} \frac{W_{sb} V_{sj}}{(WH)_{sj}}. \tag{48}$$

The second issue is difficult. As $\bar{f}(\mathbf{h})$ is not a quadratic function any more, calculating the difference between $\bar{f}(\mathbf{h}^k)$ and $\bar{f}(\mathbf{h}^{k+1})$ is complicated. Earlier in (24) we simultaneously take care of the following two KKT-violating sets:

$$\begin{aligned} & \{b \mid h^k_b > 0, \nabla \bar{f}(\mathbf{h}^k)_b \neq 0\} \cup \{b \mid h^k_b = 0, \nabla \bar{f}(\mathbf{h}^k)_b < 0\} \\ = \;& \{b \mid h^k_b > 0, \nabla \bar{f}(\mathbf{h}^k)_b \neq 0\} \cup \{b \mid h^k_b \leq \sigma, \nabla \bar{f}(\mathbf{h}^k)_b < 0\}. \end{aligned} \tag{49}$$

For elements in the second set of (49), we used $\bar{h}_b^k$ to calculate the step size. The purpose is to avoid $\mathbf{h}_b^* = 0$ and $\nabla \bar{f}(\mathbf{h}^*)_b < 0$ at a limit point $\mathbf{h}^*$. Now we single this step out and have two stages:

1. For any element in the second set of (49), we modify it to $\bar{h}_b^k$ and ensure the strict decrease property
$$\bar{f}(\bar{\mathbf{h}}^k) < \bar{f}(\mathbf{h}^k). \tag{50}$$

2. We then update $\bar{\mathbf{h}}^k$ to $\mathbf{h}^{k+1}$ by the original formula. Of course we also need
$$\bar{f}(\mathbf{h}^{k+1}) < \bar{f}(\bar{\mathbf{h}}^k). \tag{51}$$

Combining (50) and (51) we have $\bar{f}(\mathbf{h}^{k+1}) < \bar{f}(\mathbf{h}^k)$.

To have (50), for H^k we define

$$B_j \equiv \{b \mid H_{bj}^k \leq \sigma, \nabla_H f(W^k, H^k)_{bj} < 0\}, j = 1, \ldots, m. \tag{52}$$

The index k in B_j is omitted for simplification. We then define

$$\bar{H}_{bj}^k \equiv \begin{cases} H_{bj}^k - \frac{\nabla_H f(W^k, H^k)_{bj}}{M^k} & \text{if } b \in B_j, \\ H_{bj}^k & \text{otherwise,} \end{cases} \tag{53}$$

where M^k is a large value defined as

$$M^k \equiv 1 + \max_{j:B_j \neq \emptyset} \frac{\left(\sum_{b\in B_j} -\nabla_H f(W^k, H^k)_{bj} \mathbf{e}^T W_{:,b}^k\right)^2}{\sum\limits_{b\in B_j} \nabla_H f(W^k, H^k)_{bj}^2 \cdot \min\limits_{i:V_{ij}>0} (W^k H^k)_{ij}}. \tag{54}$$

If $(W^k H^k)_{ij} > 0, \forall V_{ij} > 0$, then M^k is well-defined. This is related to (47) discussed earlier. The constant 1 in (54) avoids zero denominator in (53). The update rule (53) takes the negative gradient direction, so a sufficiently small step guarantee the strict decrease of the function value. The real difficulty of defining M^k is that we must have that as $k \to \infty$, M^k does not approach ∞. This property is needed in the convergence proof.

For W^k, we can define similar sets A_i and a large value N^k to have $\bar{W}^k$. Our new algorithm is then as the following:

Algorithm 4 Minimizing the divergence: a modified algorithm

1. Given $\sigma > 0$ and $\delta > 0$. Initialize $W_{ia}^1 > 0, H_{bj}^1 > 0, \forall i, a, b, j$.
2. For $k = 1, 2, \ldots$
   (a) Update H^k to $\bar{H}^k$ by (53).
   (b)
   $$H_{bj}^{k+1} = \bar{H}_{bj}^k - \frac{\bar{H}_{bj}^k}{\mathbf{e}^T W_{:,b}^k + \delta} \nabla_H f(W^k, \bar{H}^k)_{bj}, \forall b, j. \tag{55}$$
   (c) Update W^k to $\bar{W}^k$.
   (d)
   $$W_{ia}^{k+1} = \bar{W}_{ia}^k - \frac{\bar{W}_{ia}^k}{H_{a,:}^{k+1}\mathbf{e} + \delta} \nabla_W f(\bar{W}^k, H^{k+1})_{ia}, \forall i, a. \tag{56}$$

This new algorithm doubles the cost per iteration as from H^k to H^{k+1} it calculates the gradient twice: $\nabla_H f(W^k, H^k)$ in (53) and $\nabla_H f(W^k, \bar{H}^k)$ in (55). However, rarely at a limit point

$$H_{bj}^* = 0 \text{ and } \nabla_H f(W^*, H^*)_{bj} = 0$$

both occur, a situation referred to as "degenerate" in optimization. Thus if σ is chosen to be small, in final iterations all B_j are empty sets. Thus $\bar{H}^k = H^k$ and the cost per iteration is the same as that of the original algorithm.

We then prove the convergence of Algorithm 4. The following theorem indicates that all iterations are strictly positive:

Theorem 8 *If $W_{ia}^1 > 0$ and $H_{bj}^1 > 0, \forall i, a, b, j$, then*

$$W_{ia}^k > 0 \text{ and } H_{bj}^k > 0, \forall i, a, b, j, \forall k \geq 1. \tag{57}$$

The proof is omitted due to the similarity to Theorem 2. The next two theorems prove the decreasing properties (50) and (51):

Theorem 9 *Let $\mathbf{h}$ be the jth column of H and $\mathbf{v}$ be the jth column of V, respectively. Assume $\bar{f}(\mathbf{h})$ is well-defined. Let*

$$B \text{ be any non-empty subset of } \{b \mid \nabla \bar{f}(\mathbf{h})_b < 0\}. \tag{58}$$

If $\mathbf{h}$ *is updated to* $\bar{\mathbf{h}}$ *by*

$$\bar{h}_b \equiv \begin{cases} h_b - \frac{\nabla \bar{f}(\mathbf{h})_b}{M} & \textit{if } b \in B, \\ h_b & \textit{otherwise}, \end{cases} \tag{59}$$

where

$$M > \frac{\max\limits_{i:v_i>0} \sum\limits_{b\in B} -W_{ib}\nabla \bar{f}(\mathbf{h})_b}{\min\limits_{i:v_i>0} (W\mathbf{h})_i} \cdot \frac{\sum\limits_{b\in B} -\nabla \bar{f}(\mathbf{h})_b \mathbf{e}^T W_{:,b}}{\sum\limits_{b\in B} \nabla \bar{f}(\mathbf{h})_b^2}, \tag{60}$$

then $\bar{f}(\bar{\mathbf{h}})$ *is well-defined and*

$$\bar{f}(\bar{\mathbf{h}}) < \bar{f}(\mathbf{h}).$$

Proof.

The assumption that $\bar{f}(\mathbf{h})$ is well-defined means that $(W\mathbf{h})_i > 0$ if $v_i > 0$. With $(W\bar{\mathbf{h}})_i \geq (W\mathbf{h})_i$, $\bar{f}(\bar{\mathbf{h}})$ is well-defined.

Using the inequality $\log x \leq x - 1, \forall x > 0$,

$$\begin{aligned} & \bar{f}(\bar{\mathbf{h}}) - \bar{f}(\mathbf{h}) \\ = & \sum_{i:v_i>0} v_i \log \frac{(W\mathbf{h})_i}{(W\bar{\mathbf{h}})_i} + \sum_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i \\ \leq & -\sum_{i:v_i>0} v_i \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\bar{\mathbf{h}})_i} + \sum_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i. \end{aligned} \tag{61}$$

We denote $\nabla \bar{f}(\mathbf{h})_b$ as the difference between two terms:

$$\nabla \bar{f}(\mathbf{h})_b = \Delta_b - \Gamma_b, \text{ where } \Delta_b \equiv \mathbf{e}^T W_{:,b} \text{ and } \Gamma_b \equiv \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h})_i}. \tag{62}$$

We next claim that

$$(W\bar{\mathbf{h}})_i > (W\mathbf{h})_i. \tag{63}$$

If this claim is wrong,

$$W_{ib} = 0, \forall b \in B, \forall i$$

implies $\nabla \bar{f}(\mathbf{h})_b = 0, b \in B$. As B is nonempty from (58), there is a contradiction.

From (63), $\sum_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i > 0$, so we can compare the two terms in (61) by the following formula:

$$
\begin{aligned}
& \sum_{i:v_i>0} v_i \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\bar{\mathbf{h}})_i} \Bigg/ \sum_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i \\
= & \sum_{i:v_i>0} v_i \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i \frac{(W\bar{\mathbf{h}})_i}{(W\mathbf{h})_i}} \Bigg/ \sum_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i \qquad (64) \\
\geq & \frac{\sum\limits_{i:v_i>0} v_i \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i}}{\sum\limits_{i=1}^{n} (W\bar{\mathbf{h}} - W\mathbf{h})_i} \cdot \frac{1}{\max\limits_{i:v_i>0} \frac{(W\bar{\mathbf{h}})_i}{(W\mathbf{h})_i}} \\
= & \frac{\sum\limits_{i:v_i>0} v_i \sum\limits_{b\in B} \frac{W_{ib}(\bar{h}_b - h_b)}{(W\mathbf{h})_i}}{\sum\limits_{i=1}^{n} \sum\limits_{b\in B} W_{ib}(\bar{h}_b - h_b)} \cdot \frac{1}{1 + \max\limits_{i:v_i>0} \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i}} \\
= & \frac{\sum\limits_{b\in B} (\bar{h}_b - h_b)\Gamma_b}{\sum\limits_{b\in B} (\bar{h}_b - h_b)\Delta_b} \cdot \frac{1}{1 + \max\limits_{i:v_i>0} \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i}}. \qquad (65)
\end{aligned}
$$

To have that (61) is less than zero, it suffices to prove that (65) is greater than 1. This is equivalent to

$$
\max_{i:v_i>0} \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i} < \frac{\sum_{b\in B} (\bar{h}_b - h_b)\Gamma_b}{\sum_{b\in B} (\bar{h}_b - h_b)\Delta_b} - 1. \qquad (66)
$$

We further simplify the right-hand-side of (66) to

$$
\frac{\sum_{b\in B} (\bar{h}_b - h_b)(\Gamma_b - \Delta_b)}{\sum_{b\in B} (\bar{h}_b - h_b)\Delta_b} = \frac{\sum_{b\in B} \nabla \bar{f}(\mathbf{h})_b^2}{\sum_{b\in B} -\nabla \bar{f}(\mathbf{h})_b \Delta_b}. \qquad (67)
$$

Using the definition of M in (60),

$$
\begin{aligned}
& \max_{i:v_i>0} \frac{(W\bar{\mathbf{h}} - W\mathbf{h})_i}{(W\mathbf{h})_i} \\
\leq & \frac{\max\limits_{i:v_i>0} (W\bar{\mathbf{h}} - W\mathbf{h})_i}{\min\limits_{i:v_i>0} (W\mathbf{h})_i} \\
= & \frac{1}{M} \cdot \frac{\max\limits_{i:v_i>0} \sum\limits_{b\in B} -W_{ib} \nabla \bar{f}(\mathbf{h})_b}{\min\limits_{i:v_i>0} (W\mathbf{h})_i} < (67).
\end{aligned}
$$

Thus (65) is greater than one and the proof is complete. □

Clearly our setting in (53) and (54) satisfies conditions of Theorem 9. From (60), we use

$$\sum_{b\in B} -\nabla \bar{f}(\mathbf{h})_b \mathbf{e}^T W_{:,b} \geq \max_{i:v_i>0} \sum_{b\in B} -W_{ib}^k \nabla \bar{f}(\mathbf{h})_b$$

so that M^k in (54) has a simpler form. The next theorem discusses the change of function values from $\bar{\mathbf{h}}^k$ to $\mathbf{h}^{k+1}$:

Theorem 10 *Let* $\mathbf{h}$ *be the* j*th column of* H *and* $\mathbf{v}$ *be the* j*th column of* V*, respectively. Assume* $\bar{f}(\mathbf{h})$ *is well-defined. If* $\mathbf{h}$ *is updated to* $\mathbf{h}^n$ *by*

$$h_b^n = h_b - \frac{h_b}{\mathbf{e}^T W_{:,b} + \delta} \nabla \bar{f}(\mathbf{h})_b,$$

then $\bar{f}(\mathbf{h}^n)$ *is well-defined and*

$$\bar{f}(\mathbf{h}^n) \leq \bar{f}(\mathbf{h}). \tag{68}$$

Moreover, if $\nabla \bar{f}(\mathbf{h}) \neq \mathbf{0}$*, then the above inequality is strict.*

Proof.

The proof that $\bar{f}(\mathbf{h}^n)$ is well-defined is straightforward, so we directly prove (68). Using (61),

$$\begin{aligned}
&\bar{f}(\mathbf{h}^n) - \bar{f}(\mathbf{h}) \\
\leq\ & \sum_{i:v_i>0} v_i \frac{(W\mathbf{h} - W\mathbf{h}^n)_i}{(W\mathbf{h}^n)_i} + \sum_{i=1}^{n} (W\mathbf{h}^n - W\mathbf{h})_i \\
=\ & \sum_{b=1}^{r} (h_b^n - h_b) \left(\mathbf{e}^T W_{:,b} - \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h}^n)_i} \right) \\
=\ & \sum_{b=1}^{r} (h_b^n - h_b)\Delta_b - \sum_i v_i + \sum_{b=1}^{r} h_b \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h}^n)_i} \\
=\ & -\sum_{b=1}^{r} \frac{h_b}{\Delta_b + \delta} (\Delta_b - \Gamma_b)\Delta_b - \sum_i v_i + \sum_{b=1}^{r} h_b \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h}^n)_i} \\
=\ & -\sum_{b=1}^{r} \frac{h_b \Delta_b^2}{\Delta_b + \delta} + \sum_{b=1}^{r} h_b \Gamma_b - \sum_{b=1}^{r} \frac{\delta h_b \Gamma_b}{\Delta_b + \delta} - \sum_i v_i + \sum_{b=1}^{r} h_b \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h}^n)_i}
\end{aligned} \tag{69}$$

We have

$$\sum_{b=1}^{r} h_b \Gamma_b = \sum_{b=1}^{r} h_b \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h})_i} = \sum_i v_i,$$

and

$$
\begin{aligned}
&\sum_{b=1}^{r} h_b \sum_{i:v_i>0} \frac{W_{ib}v_i}{\sum_{a=1}^{r} W_{ia}h_a \frac{\Gamma_a+\delta}{\Delta_a+\delta}} \\
= &\sum_{i:v_i>0} \frac{v_i}{\sum_{a=1}^{r} \frac{W_{ia}h_a}{(W\mathbf{h})_i} \frac{\Gamma_a+\delta}{\Delta_a+\delta}} \\
\leq &\sum_{i:v_i>0} v_i \sum_{a=1}^{r} \frac{W_{ia}h_a}{(W\mathbf{h})_i} \frac{\Delta_a+\delta}{\Gamma_a+\delta} \qquad (70) \\
= &\sum_{a=1}^{r} h_a \frac{\Delta_a+\delta}{\Gamma_a+\delta} \Gamma_a,
\end{aligned}
$$

where (70) is from Jensen's inequality. Therefore,

$$
\begin{aligned}
(69) \leq &\sum_{b=1}^{r} h_b \left(\frac{-\Delta_b^2 - \delta\Gamma_b}{\Delta_b+\delta} + \frac{\Delta_b+\delta}{\Gamma_b+\delta}\Gamma_b \right) \\
= &-\sum_{b=1}^{r} h_b \frac{\delta(\Delta_b - \Gamma_b)^2}{(\Delta_b+\delta)(\Gamma_b+\delta)} \leq 0. \qquad (71)
\end{aligned}
$$

Hence (68) holds. If $\nabla \bar{f}(\mathbf{h})_b = \Delta_b - \Gamma_b \neq 0$ for some b, then clearly (71) becomes a strict inequality. □

What Lee and Seung proved is the case when $\delta = 0$ and $\mathbf{e}^T W_{:,b} > 0$. Their proof does not extend to the case of $\delta > 0$, so we have a very different derivation here.

From the above two theorems the non-increasing property of function values follows:

Theorem 11 *If Algorithm 4 generates an infinite sequence* $\{W^k, H^k\}$, *then*

$$
f(W^{k+1}, H^{k+1}) \leq f(\bar{W}^k, H^{k+1}) \leq f(W^k, H^{k+1}) \leq f(W^k, \bar{H}^k) \leq f(W^k, H^k), \forall k. \qquad (72)
$$

Moreover, one of the above inequalities is strict.

We omit the proof as it is similar to Theorem 4. In Section 3 we then have Theorems 5 and 6 to finish the convergence proof. However, here we follow the standard fixed-point proof of using (9). In Section 3 the auxiliary function gives useful information about $\mathbf{h}^{k+1} - \mathbf{h}^k$, so we prove Theorem 5 first. Here such information is however not that obvious.

Theorem 12 *Assume*

$$\lim_{k\in\mathcal{K},k\to\infty}(W^k, H^k) = (W^*, H^*). \tag{73}$$

Then

1.

$$(W^*H^*)_{ij} > 0 \text{ if } V_{ij} > 0. \tag{74}$$

2. *We have*

$$\text{if } H^*_{bj} = 0, \text{ then } \nabla_H f(W^*, H^*)_{bj} \geq 0, \tag{75}$$

and

$$\text{if } H^*_{bj} > 0, \text{ then } \nabla_H f(W^*, H^*)_{bj} = 0. \tag{76}$$

3.

$$\lim_{k\in\mathcal{K},k\to\infty} H^{k+1} = H^*. \tag{77}$$

Proof.

To prove (74) we assume that it is wrong. Then

$$f(W^k, H^k) = -\sum_{\substack{ij:\ V_{ij}>0\\(W^*H^*)_{ij}=0}} V_{ij}\log(W^kH^k)_{ij} - \sum_{\substack{ij:\ V_{ij}>0\\(W^*H^*)_{ij}>0}} V_{ij}\log(W^kH^k)_{ij} + \sum_{ij}(W^kH^k)_{ij}$$

goes to ∞ as $k \in \mathcal{K}, k \to \infty$. This result contradicts Theorem 11.

Next we prove (75). As the number of possible $B_j, j = 1, \ldots, m$ is finite, there are $B^*_j, j = 1, \ldots, m$ used at infinitely many iterations of $\mathcal{K}$. We use $\bar{\mathcal{K}} \subseteq \mathcal{K}$ to denote these iterations. Therefore,

$$\nabla_H f(W^k, H^k)_{bj} < 0, \forall b \in B^*_j, \forall k \in \bar{\mathcal{K}} \text{ implies } \nabla_H f(W^*, H^*)_{bj} \leq 0, \forall b \in B^*_j. \tag{78}$$

We further define a subset of B^*_j:

$$\bar{B}^*_j \equiv \{b \mid b \in B^*_j, \nabla_H f(W^*, H^*)_{bj} < 0\} \subseteq \{b \mid \nabla_H f(W^*, H^*)_{bj} < 0\}. \tag{79}$$

If $\cup_j \bar{B}_j^* \neq \emptyset$, then using (74) the following limit exists and is denoted as M^*:

$$\begin{aligned}
& M^* \\
\equiv \quad & \lim_{k\in\bar{\mathcal{K}},k\to\infty} M^k \\
= \quad & 1 + \max_{j:B_j^*\neq\emptyset} \frac{\left(\sum_{b\in B_j^*} -\nabla_H f(W^*,H^*)_{bj}\mathbf{e}^T W^*_{:,b}\right)^2}{\sum\limits_{b\in B_j^*} \nabla_H f(W^*,H^*)_{bj}^2 \cdot \min\limits_{i:V_{ij}>0} (W^*H^*)_{ij}} \\
= \quad & 1 + \max_{j:B_j^*\neq\emptyset} \frac{\left(\sum_{b\in \bar{B}_j^*} -\nabla_H f(W^*,H^*)_{bj}\mathbf{e}^T W^*_{:,b}\right)^2}{\sum\limits_{b\in \bar{B}_j^*} \nabla_H f(W^*,H^*)_{bj}^2 \cdot \min\limits_{i:V_{ij}>0} (W^*H^*)_{ij}}.
\end{aligned} \tag{80}$$

Using $\bar{B}_j^*$ and (80) respectively as B and M in Theorem 9, (79) implies (58) and we can define $\bar{H}^*$ according to (59). If the result (75) is wrong, then $\cup_j \bar{B}_j^* \neq \emptyset$, so

$$\bar{H}^* \neq H^* \text{ and } f(W^*, \bar{H}^*) < f(W^*, H^*). \tag{81}$$

We then claim that

$$\lim_{k\in\bar{\mathcal{K}},k\to\infty} \bar{H}^k = \bar{H}^*. \tag{82}$$

This property is proved by considering the following three situations:

1. $b \in \bar{B}_j^*$.

   (78) and (53) imply

   $$\bar{H}_{bj}^k = H_{bj}^k - \frac{\nabla_H f(W^k, H^k)_{bj}}{M^k}, \forall k \in \bar{\mathcal{K}}, \tag{83}$$

   Moreover, (79) and (53) imply

   $$\bar{H}_{bj}^* = H_{bj}^* - \frac{\nabla_H f(W^*, H^*)_{bj}}{M^*}.$$

   Using (80), (73) and taking the limit of (83) we have (82).

2. $b \in B_j^* \backslash \bar{B}_j^*$.

   We have

   $$\bar{H}_{bj}^k = H_{bj}^k - \frac{\nabla_H f(W^k, H^k)_{bj}}{M^k}, \forall k \in \bar{\mathcal{K}} \quad \text{and} \quad \bar{H}_{bj}^* = H_{bj}^*.$$

   Moreover, $\nabla_H f(W^*, H^*)_{bj} = 0$. Taking the limit and using (73) leads to (82).

3. $b \notin B_j^*$.

   Then

$$\bar{H}_{bj}^k = H_{bj}^k \ \ \forall k \in \bar{\mathcal{K}} \qquad \text{and} \qquad \bar{H}_{bj}^* = H_{bj}^*.$$

   Thus (82) also follows.

Using (82),

$$\lim_{k \in \bar{\mathcal{K}}, k \to \infty} f(W^k, \bar{H}^k) = f(W^*, \bar{H}^*).$$

This and (81) then contradicts Theorem 11. Thus (75) holds.

The proof for (76) is easier and similar. We omit it here.

Using (75), (76), and (82), we have

$$\lim_{k \in \bar{\mathcal{K}}, k \to \infty} \bar{H}_{bj}^k = \bar{H}_{bj}^* = H_{bj}^*.$$

Taking the limit of (55), and using (75)-(76), then (77) follows.

Note that δ in the denominator of the step size plays a role. Without it

$$\lim_{k \in \mathcal{K}, k \to \infty} \mathbf{e}^T W_{:,b}^k = \mathbf{e}^T W_{:,b}^*$$

may cause a zero denominator in taking the limit of (55). □

Now are have all required properties. The main convergence theorem follows from a similar proof for Theorem 7.

Theorem 13 *Any limit point of the sequence* $\{W^k, H^k\}$ *generated by Algorithm 4 is a stationary point of* (2).

5 Minimizing the Divergence: Update One Row at A Time

Due to the difficulty of handling the following two modifications together:

$$H \to \bar{H} \qquad \text{and} \qquad \mathbf{e}^T W_{:,b} \to \mathbf{e}^T W_{:,b} + \delta,$$

Section 4 considers a two-stage algorithm and proof. This section proposes a different approach: two modifications are implemented together, but at each iteration only one row of H is updated. Thus for any column $\mathbf{h}$ and the corresponding function $\bar{f}(\mathbf{h})$, only one element is updated per iteration. We will show that proving the decreasing property is easier.

Given $\sigma > 0$, we define $\bar{H}$ as the following:

$$\bar{H}_{bj} \equiv \begin{cases} \min\left(\min_{i:W_{ib}>0,V_{ib}>0} \frac{(W\mathbf{h})_i}{W_{ib}}, \sigma\right) & \text{if } H_{bj} \leq \sigma \text{ and } \nabla_H f(W,H)_{bj} < 0, \\ H_{bj} & \text{otherwise.} \end{cases} \tag{84}$$

A new algorithm is as follows:

Algorithm 5 Minimizing the divergence: updating one row (column) at a time

---

1. Given $\sigma > 0$ and $\delta > 0$. Initialize $W^1_{ia} > 0, H^1_{bj} > 0, \forall i, a, b, j,$
2. For $k = 1, 2, \ldots$
   (a) $H^{k,0} = H^k$
   (b) For $t = 1, \ldots, r$
   $H^{k,t}$ is the same as $H^{k,t-1}$ except the tth row is updated by
   $$H^{k,t}_{tj} = H^{k,t-1}_{tj} - \frac{\bar{H}^{k,t-1}_{tj}}{\mathbf{e}^T W^k_{:,b} + \delta} \nabla_H f(W^k, H^{k,t-1})_{tj}, \forall j. \tag{85}$$
   (c) $H^{k+1} = H^{k,r}$ and $W^{k,0} = W^k$.
   (d) For $t = 1, \ldots, r$
   $W^{k,t}$ is the same as $W^{k,t-1}$ except the tth column is updated by
   $$W^{k,t}_{it} = W^{k,t-1}_{it} - \frac{\bar{W}^{k,t-1}_{it}}{H^{k+1}_{a,:}\mathbf{e} + \delta} \nabla_W f(W^{k,t-1}, H^{k+1})_{it}, \forall i. \tag{86}$$
   (e) $W^{k+1} = W^{k,r}$.

The time complexity per iteration is the same as that of Algorithm 3. From $H^{k,t-1}$ to $H^{k,t}$, the operations $\sum_{s=1}^n W^k_{st} V_{sj}, \forall j$ in calculating $\nabla_H f(W^k, H^{k,t-1})_{tj}$ takes $O(nm)$ operations. Thus from 1 to r the total is $O(nmr)$, the same as that in (41). Maintaining $W^k H^{k,t}$ can be time consuming, but we can take the following trick:

$$W^k H^{k,t} = W^k H^{k,t-1} + W^k_{:,t}(H^{k,t} - H^{k,t-1})_{t,:}. \tag{87}$$

The second term is a rank-one update involving the tth column of W^k and the change of the tth row of H. Then (87) takes $O(nm)$ operations. From $t = 1, \ldots, r$, the total is $O(nmr)$, again the same as that in Algorithm 3. Practically as operations are not matrix-based, this algorithm may be less efficient than Algorithms 3 and 4 when finely tuned numerical linear algebra subroutines are used.

We then prove the convergence of Algorithm 5. All details are similar to those in Section 4 except the decreasing property. We prove only this key result in the following theorem:

Theorem 14 *Let* $\mathbf{h}$ *be the* j*th column of* H *and* $\mathbf{v}$ *be the* j*th column of* V*. Assume* $\bar{f}(\mathbf{h})$ *is well-defined. From* $\mathbf{h}$ *we update its* b*th component by*

$$h_b^n = h_b - \frac{\bar{h}_b}{\mathbf{e}^T W_{:,b} + \delta} \nabla \bar{f}(\mathbf{h})_b \tag{88}$$

and have all other elements remain the same. Then $\bar{f}(\mathbf{h}^n)$ *is well-defined and*

$$\bar{f}(\mathbf{h}^n) \leq \bar{f}(\mathbf{h}). \tag{89}$$

Moreover, if

$$h_b = 0, \nabla \bar{f}(\mathbf{h})_b < 0 \text{ or } h_b > 0, \nabla \bar{f}(\mathbf{h})_b \neq 0,$$

then

$$\bar{f}(\mathbf{h}^n) < \bar{f}(\mathbf{h}).$$

Proof.

From (84), we have

$$(W\mathbf{h}^n)_i = (W\mathbf{h})_i - \frac{W_{ib}\bar{h}_b}{\mathbf{e}^T W_{:,b} + \delta} \nabla \bar{f}(\mathbf{h})_b \geq (W\mathbf{h})_i,$$

so $\bar{f}(\mathbf{h}^n)$ is well-defined.

Using (61),

$$\begin{aligned}
& \bar{f}(\mathbf{h}^n) - \bar{f}(\mathbf{h}) \\
\leq & \sum_{i:v_i>0} v_i \frac{(W\mathbf{h} - W\mathbf{h}^n)_i}{(W\mathbf{h}^n)_i} + \sum_{i=1}^{n} (W\mathbf{h}^n - W\mathbf{h})_i \\
= & (h_b^n - h_b) \left(\mathbf{e}^T W_{:,b} - \sum_{i:v_i>0} \frac{W_{ib} v_i}{(W\mathbf{h}^n)_i} \right).
\end{aligned}$$

If $h_b^n - h_b = 0$, then of course (89) holds.

If $h_b^n - h_b > 0$, then from (88), $-\nabla \bar{f}(\mathbf{h})_b > 0$. We consider two cases:

Case 1: $\mathbf{e}^T W_{:,b} = 0$. Then the non-negativity of $\sum_{i:v_i>0} \frac{W_{ib}v_i}{(W\mathbf{h}^n)_i}$ implies (89).

Case 2: $\mathbf{e}^T W_{:,b} > 0$. Using the definition of $\bar{h}_b$ in (84),

$$
\begin{aligned}
&\frac{\sum_{i:v_i>0} \frac{W_{ib}v_i}{(W\mathbf{h}^n)_i}}{\mathbf{e}^T W_{:,b}} \\
= &\frac{\sum_{i:v_i>0} \frac{W_{ib}v_i}{(W\mathbf{h})_i + \frac{-W_{ib}\bar{h}_b}{\mathbf{e}^T W_{:,b}+\delta}\nabla \bar{f}(\mathbf{h})_b}}{\mathbf{e}^T W_{:.b}} \\
> &\frac{\sum_{i:v_i>0} \frac{W_{ib}v_i}{(W\mathbf{h})_i + (W\mathbf{h})_i \frac{-\nabla \bar{f}(\mathbf{h})_b}{\mathbf{e}^T W_{:,b}}}}{\mathbf{e}^T W_{:,b}} \\
= &\frac{\sum_{i:v_i>0} \frac{W_{ib}v_i}{(W\mathbf{h})_i}}{\mathbf{e}^T W_{:,b}} \cdot \frac{1}{1 - \frac{\nabla \bar{f}(\mathbf{h})_b}{\mathbf{e}^T W_{:,b}}} = 1.
\end{aligned} \tag{90}
$$

For the case of $h_b^n - h_b < 0$, we then have $-\nabla \bar{f}(\mathbf{h})_b < 0$. It is impossible that $\mathbf{e}^T W_{:,b} = 0$ as otherwise $\nabla \bar{f}(\mathbf{h})_b \leq 0$ causes a contradiction. Then the remaining proof is similar to deriving (90). □

6 Discussion and Conclusions

Though we have proved that any limit point is stationary, it is unclear yet if the sequence $\{W^k, H^k\}$ has at least one limit point. Showing the existence of limit points is an interesting future issue.

Though this paper may have only theoretical values, it has two main contributions:

1. Under minor modifications, Lee and Seung's multiplicative update algorithms converge to stationary points.

2. Though bound constraints introduce difficulties in proving the convergence, we invent a technique to control the step size. For multiplicative update algorithms to solve other bound-constrained problems, we can apply the same approach to prove the convergence.

A A Technical Lemma

Lemma 1 *Given $\delta > 0$, an $r \times r$ symmetric positive semi-definite matrix A with $A_{ab} \geq 0, \forall a, b$, and a vector $\mathbf{x}$ with $x_b \geq 0, \forall b$. Let I be any index set such that*

$$x_b > 0 \text{ if } b \in I, \tag{91}$$

and define a diagonal matrix matrix $\bar{D}$ with

$$\bar{D}_{bb} \equiv \begin{cases} \frac{(A\mathbf{x})_b + \delta}{x_b} & \text{if } b \in I, \\ 0 & \text{if } b \notin I. \end{cases} \tag{92}$$

Then $(\bar{D} - A)_{II}$ is symmetric positive definite.

Proof.

For any vector $\mathbf{v}$ with $\mathbf{v}_I \neq \mathbf{0}$,

$$\begin{aligned}
& \mathbf{v}_I^T (\bar{D} - A)_{II} \mathbf{v}_I \\
= & \sum_{a \in I} v_a^2 \frac{\delta}{x_a} + \sum_{a \in I} v_a^2 \frac{(A\mathbf{x})_a}{x_a} - \sum_{a,b \in I} v_a v_b A_{ab} && (93) \\
> & \sum_{a \in I} v_a^2 \frac{\sum_{b \in I} A_{ab} x_b}{x_a} - \sum_{a,b \in I} v_a v_b A_{ab} && (94) \\
= & \frac{1}{2} \sum_{a,b \in I} v_a^2 \frac{A_{ab} x_b}{x_a} + \frac{1}{2} \sum_{a,b \in I} v_b^2 \frac{A_{ba} x_a}{x_b} - \sum_{a,b \in I} v_a v_b A_{ab} && (95) \\
= & \frac{1}{2} \sum_{a,b} A_{ab} \left(\sqrt{\frac{x_b}{x_a}} v_a - \sqrt{\frac{x_a}{x_b}} v_b \right)^2 \geq 0. && (96)
\end{aligned}$$

The condition (91) ensures (93) to be well-defined. From (93) to (94) we use the property $A_{ab} \geq 0$ and $x_b \geq 0$, $\forall b = 1, \ldots, r$. From (95) to (96) the symmetry of A is used. □

References

Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA 02178-9998, second edition, 1999.

Edward F. Gonzales and Yin Zhang. Accelerating the Lee-Seung algorithm for non-negative matrix factorization. Technical report, Department of Computational and Applied Mathematics, Rice University, 2005.

Patrik O. Hoyer. Non-negative sparse coding. In *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

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

Chih-Jen Lin. Projected gradient methods for non-negative matrix factorization. Technical report, Department of Computer Science, National Taiwan University, 2005. URL http://www.csie.ntu.edu.tw/~cjlin/papers/pgradnmf.pdf.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error. *Environmetrics*, 5:111–126, 1994.

Jon Piper, Paul Pauca, Robert Plemmons, and Maile Giffin. Object characterization from spectral data using nonnegative factorization and information theory. In *Proceedings of AMOS Technical Conference*, 2004.