# On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization

Chih-Jen Lin, *Senior Member, IEEE*

***Abstract*—Nonnegative matrix factorization (NMF) is useful to find basis information of nonnegative data. Currently, multiplicative updates are a simple and popular way to find the factorization. However, for the common NMF approach of minimizing the Euclidean distance between approximate and true values, no proof has shown that multiplicative 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 may be applied to prove the convergence for other bound-constrained optimization problems.**

***Index Terms*—Asymptotic convergence, multiplicative updates, nonnegative matrix factorization (NMF), stationarity.**

## I. Introduction

NONNEGATIVE matrix factorization (NMF) is useful to find basis information of nonnegative data [1], [2]. Given an $n \times m$ data matrix $V$ with $V_{ij} \geq 0$ and a predetermined positive integer $r < \min(n, m)$, NMF finds two nonnegative matrices $W \in R^{n \times r}$ and $H \in R^{r \times m}$ so that

$$V \approx WH.$$

If each column of $V$ represents an object, this method approximates it by a linear combination of $r$ "basis" columns in $W$. NMF has been applied to many application areas. A recent NMF survey is [3].

A common way to find $W$ and $H$ is by minimizing the Euclidean distance between $V$ and $WH$

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

Each nonnegative inequality is a "bound constraint," as it relates to only a single variable. We also note that

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

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

Manuscript received May 1, 2006; revised September 24, 2006 and December 1, 2006; accepted February 7, 2007. This work was supported in part by the National Science Council of Taiwan under Grant NSC 93-2213-E-002-030.

The author is with the Department of Computer Science, National Taiwan University, Taipei 106, Taiwan (e-mail: cjlin@csie.ntu.edu.tw).

Digital Object Identifier 10.1109/TNN.2007.895831

A popular approach to solve the NMF optimization problems(1) is a multiplicative update algorithm by Lee and Seung [4]. Though some papers such as [5] pointed out its possible slow convergence, this method is popular due to the simplicity. Lee and Seung [4] proved that the update causes the function value to be nonincreasing, but there is no proof yet showing that any limit point is stationary. While optimization problems here may be nonconvex 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 [6] presented numerical examples showing that Lee and Seung's algorithm in [4] fails to approach a stationary point. However, due to possible numerical inaccuracy, we think either a convergence proof or a nonconvergence example is desired. Other work which has touched the convergence issues includes [3] and [7]. This paper conducts a detailed study about the convergence properties of multiplicative update methods.

Besides multiplicative updates, other methods are available to minimize the NMF problem (1). Some examples are [3], [5], and [8]. These approaches may be more efficient, but are also more complicated.

The main difficulty of proving the convergence of multiplicative updates comes from the nonnegativity constraints. Though multiplicative updates are close to standard fixed-point methods, existing fixed-point proofs mainly deal with unconstrained situations. Section II reviews Lee and Seung's algorithm for (1) and discusses difficulties of proving the convergence. Section III proposes a modified algorithm, which has the same computational complexity per iteration. For this modified procedure, Section IV then proves that any limit point is stationary. We also show that iterations are in a closed and bounded set, so at least one limit point exists.

Another NMF formulation minimizes the (generalized) Kullback–Leibler divergence between $V$ and $WH$

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

Lee and Seung also proposed a multiplicative algorithm to minimize (2). By transforming (2) to another form, Finesso and Spreij [9] successfully analyzed the convergence property. Multiplicative updates for (2) are close to expectation–maximization (EM) algorithm in maximum likelihood, so their analysis is quite different from ours for (1). Section VI investigates the difference. The same section also discusses possible future issues, and gives conclusions of this paper.

## II. Multiplicative Update for (1) and Its Convergence Issues

Lee and Seung [4] proposed the following algorithm to solve (1).

**Algorithm 1:** Multiplicative update to solve (1)

For $k = 1, 2, \ldots$

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

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

This procedure is not well defined if denominators in (3) or (4) are zero. Moreover, the initial point is a concern; some use positive matrices, but some merely consider nonnegative ones. Lin [5] discusses conditions so the procedure is well defined.

*Theorem 1 [5, Th. 1]:* If $V$ has neither zero column nor row, and $W_{ia}^1 > 0$ and $H_{bj}^1 > 0, \forall\, i, a, b, j$, then

$$w_{ia}^k > 0 \text{ and } h_{bj}^k > 0 \qquad \forall\, i, a, b, j \qquad \forall\, k \geq 1.$$

We hope that any limit point of $\{W^k, H^k\}$ is stationary as a 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., [10]). For all $i$, $a$, $b$, and $j$

$$\begin{aligned} W_{ia} \geq 0 \quad & H_{bj} \geq 0 \\ \nabla_W f(W,H)_{ia} \geq 0 \quad & \nabla_h f(W,H)_{bj} \geq 0 \\ W_{ia} \cdot \nabla_W f(W,H)_{ia} = 0 \quad & H_{bj} \cdot \nabla_h f(W,H)_{bj} = 0 \end{aligned} \tag{5}$$

where

$$\begin{aligned} \nabla_W f(W,H) &= (WH - V)H^T \\ \nabla_H f(W,H) &= W^T(WH - V) \end{aligned} \tag{6}$$

are, respectively, partial derivatives to elements in $W$ and $H$.

Lee and Seung [4] proved the following properties.

1) The function value is nonincreasing after every update

$$\begin{aligned} f(W^k, H^{k+1}) &\leq f(W^k, H^k) \\ f(W^{k+1}, H^{k+1}) &\leq f(W^k, H^{k+1}). \end{aligned} \tag{7}$$

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 $W$.

Several papers such as [5] and [6] 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

$$\begin{aligned} &((W^k)^T V)_{bj} = ((W^k)^T W^k H^k)_{bj} \\ &\text{implies } \nabla_H F(W^k, H^k)_{bj} = 0 \end{aligned}$$

which is part of the KKT condition (5). A convergence proof of fixed-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 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

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

causes a contradiction.

Clearly, this framework cannot be directly used here because of the following 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 nonnegative constraints. The previous framework does not reveal how to have this result.

Gonzales and Zhang [6] numerically showed that Algorithm 1 may fail to converge to a stationary point. However, Lin [5] stated that due to possible numerical inaccuracy, a mathematical example is desired before drawing conclusions. Thus, the convergence issue remains open. In Section III, we will slightly 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 [5] 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 $O(\mathrm{nmr})$ operations. Therefore, the complexity of Algorithm 1 is

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

Some implementations of the multiplicative updates normalize $(W, H)$ at each iteration so that $W$'s column sums or $H$'s row sums are ones. This normalization does not change the function value as for any for any $r \times r$ positive diagonal matrix $S$, $f(WS, S^{-1}H) = f(W, H)$. We will consider this operation in our proposed algorithm.

## III. Modified Multiplicative Update

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

$$\begin{aligned} H_{bj}^{k+1} = H_{bj}^k - \frac{H_{bj}^k}{((W^k)^T W^k H^k)_{bj}} \\ \times \nabla_H f(W^k, H^k)_{bj} \qquad \forall\, b, j \end{aligned} \tag{10}$$

$$\begin{aligned} W_{ia}^{k+1} = W_{ia}^k - \frac{W_{ia}^k}{(W^k H^{k+1}(H^{k+1})^T)_{ia}} \\ \times \nabla_W f(W^k, H^{k+1})_{ia} \qquad \forall\, i, a. \end{aligned} \tag{11}$$

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 II can be reinterpreted as follows.

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

Therefore, we propose modifying the step size to

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

where

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

Both $\sigma$ and $\delta$ are predefined small positive numbers. Similarly, we can define $\bar{W}^k_{ia}$. The modified algorithm is as the following.

**Algorithm 2:** A modified algorithm for minimizing (1)

1) Given $\sigma > 0$ and $\delta > 0$. Initialize $W^1_{ia} \geq 0, H^1_{bj} \geq 0$, for all $i$, $a$, $b$, and $j$.

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

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

Else

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

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

b) Normalize $W^{k,n}$ and $H^{k,n}$ to $H^{k+1}$ and $W^{k+1}$, respectively, so that $W^{k+1}$'s column sum is one. If the whole column is zero, then this as well as the corresponding row in $H$ are unchanged.

We denote $W^{k,n}$ and $H^{k,n}$ as intermediate matrices before normalization. Following [9], we impose the normalization operation in order to prove that $\{W^k, H^k\}$ is in a bounded set (see Theorem 8).

This modified algorithm requires the following extra operations:

1) calculate $\bar{H}^k$ (or $\bar{W}^k$);
2) calculate $(W^k)^T W^k \bar{H}^k$ (or $\bar{W}^k H^{k,n} (H^{k,n})^T$);
3) add $\delta$.

All are less than $O(\mathrm{nmr})$, so the complexity per iteration remains the same. The use of $\delta$ follows from some NMF papers (e.g., [11] and [12]), which includes this factor to avoid division by zero. This is also related to penalty terms added to the objective function. We discuss this aspect in more detail in Section VI.

The new algorithm is well defined without requiring any condition on $V$. One could even start from only nonnegative matrices: $W^1_{ia} \geq 0$ and $H^1_{bj} \geq 0$.

*Theorem 2:* If $W^1_{ia} > 0$ and $H^1_{bj} > 0, \forall\, i, a, b, j$, then

$$W^k_{ia} > 0 \text{ and } H^k_{bj} > 0 \qquad \forall\, i, a, b, j \qquad \forall\, k \geq 1. \quad (15)$$

Moreover, if $W^1_{ia} \geq 0$ and $H^1_{bj} \geq 0, \forall\, i, a, b, j$, then

$$W^k_{ia} \geq 0 \text{ and } H^k_{bj} \geq 0 \qquad \forall\, i, a, b, j \qquad \forall\, k \geq 1.$$

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

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

We then consider the following two situations:

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

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

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

$$\bar{H}^k_{bj} = H^k_{bj}. \quad (17)$$

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

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

Using (17) and(18)

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

After normalization, $H^{k+1}_{bj} > 0, \forall b, j$. The proof of $W^{k+1}_{ia} > 0$ is similar.

If $W^1_{ia} \geq 0$ and $H^1_{bj} \geq 0, \forall\, i, a, b, j$, then the proof is the same. ■

## IV. Convergence Analysis

We begin the convergence analysis by showing that from $H^k$ to $H^{k,n}$, all components not satisfying KKT conditions are changed and the function value is strictly 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 $H$. Hence, it is 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 $\mathbf{v}$ is the corresponding column in $V$ and $W = W^k$. Here, both $\mathbf{v}$ and $W$ are treated as constants. Lee and Seung 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

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

They proved that

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

Minimizing $A(\mathbf{h}, \mathbf{h}^k)$ with respect to $\mathbf{h}$ leads to the update rule (3). In addition, if $\bar{f}(\mathbf{h}^k)_b \neq 0$, then $h_b^{k,n} \neq h_b^k$. 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

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

and $\bar{D}_{II}$ is the submatrix of a diagonal matrix $\bar{D}$ with elements

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

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 nonincreasing property (22). This result is shown in Theorem 3.

*Theorem 3:* Let $\sigma$ and $\delta$ 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, $\mathbf{h}^{k,n}$ defined by (13) satisfies

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

and

$$\bar{f}(\mathbf{h}^{k,n}) \le \bar{A}(\mathbf{h}^{k,n}, \mathbf{h}^k) \le \bar{A}(\mathbf{h}^k, \mathbf{h}^k) = \bar{f}(\mathbf{h}^k). \tag{28}$$

Further, the following three properties are equivalent:

1) both inequalities in (28) are strict;
2) $\nabla \bar{f}(\mathbf{h}^k)_I \neq \mathbf{0}$;
3) $\mathbf{h}^{k,n} \neq \mathbf{h}^k$.

*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 Section A of the Appendix to show that $(\bar{D} - W^T W)_{II}$ is positive definite. Then, (30) is nonnegative. 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 $\mathbf{h}_I^{k,n} \neq \mathbf{h}_I^k$. Using (30) and

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

both inequalities in (28) are strict if and only if $\mathbf{h}^{k,n} \neq \mathbf{h}^k$. ■

Theorem 3 immediately implies that the function value is strictly decreasing:

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

$$\begin{aligned} f(W^{k+1}, H^{k+1}) &= f(W^{k,n}, H^{k,n}) \\ &\le f(W^k, H^{k,n}) \le f(W^k, H^k) \qquad \forall\, k. \end{aligned} \tag{32}$$

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 show that $H^k$ and $H^{k,n}$ converge to the same point. With this property, the convergence proof is easier than using (9).

*Theorem 5:* Assume $\{H^k\}$, $k \in \mathcal{K}$, is a convergent subsequence and

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

Then

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

*Proof:* Theorem 4 and the property $f(W,H) \geq 0, \forall\, W, H$, imply that $\{f(W^k, H^k)\}$ is a bounded decreasing sequence, which globally converges. Using (32)

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

Since $f(W^k, H^{k,n}) - f(W^k, H^k)$ is the sum of the difference at each column and the difference is nonpositive due to (28), we have

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

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

$$\left|h_b^{k,n} - h_b^*\right| \geq \epsilon \qquad \forall k \in \hat{\mathcal{K}}.$$

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

$$\left|h_b^k - h_b^*\right| \leq \frac{\epsilon}{2} \qquad \forall\, k \in \bar{\mathcal{K}}.$$

Combining the previous two inequalities, we have

$$\left|h_b^{k,n} - h_b^k\right| \geq \frac{\epsilon}{2} \qquad \forall\, k \in \bar{\mathcal{K}}. \tag{35}$$

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

$$\begin{aligned}\bar{f}(\mathbf{h}^{k,n}) - \bar{f}(\mathbf{h}^k) &\leq -\frac{\left(h_b^{k,n} - h_b^k\right)^2 \bar{D}_{bb}}{2} \\ &\leq -\frac{\left(h_b^{k,n} - h_b^k\right)^2 \delta}{2\bar{h}_b^k} \\ &\leq -\frac{\left(h_b^{k,n} - h_b^k\right)^2 \delta}{2\max\left(h_b^k, \delta\right)} \leq 0.\end{aligned}$$

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

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

a contradiction to (35). ∎

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 subsequence and

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

We have that

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

and

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

*Proof:* By the definition (12)

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

so the sequence $\{\bar{H}_{bj}^k\}_{k\in\mathcal{K}}$ may have two convergent points $H_{bj}^*$ or $\sigma$. Since the number of indices $(b,j)$ 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

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

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

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

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 Theorem 7.

*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

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

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

The remaining task is to prove that at least one limit point exists.

*Theorem 8:* The sequence $\{W^k, H^k\}$ has at least one limit point.

*Proof:* It suffices to prove that $(W^k, H^k)$, $k = 1, \ldots, \infty$, are in a compact (i.e., closed and bounded) set. Since $W^k$ is normalized to have column sum one (or zero for exceptional situations), we only need to show that $\{H^k\}$ is bounded. If this result is wrong, there is a component $H_{bj}$ and an infinite index set $\mathcal{K}$ such that[1]

$$\lim_{k\in\mathcal{K}} H_{bj}^k \to \infty, \quad H_{bj}^k < H_{bj}^{k+1} \qquad \forall\, k \in \mathcal{K} \tag{40}$$

and

$$\lim_{k\in\mathcal{K}, k\to\infty} W_{ib}^k = W_{ib}^* \text{ exists} \qquad \forall\, i. \tag{41}$$

[1]More formally, we have a subsequence so that $\lim_{k\in\mathcal{K}} H_{bj}^k \to \infty$ first. From that, a sub-subsequence satisfies $H_{bj}^k < H_{bj}^{k+1}$. Then, a further subsequence satisfies (41).

TABLE I
IMAGE DATA: AVERAGE OBJECTIVE VALUE AND PROJECTED-GRADIENT NORMS OF USING THREE INITIAL POINTS UNDER SPECIFIED TIME LIMITS

| Problem | | CBCL | | ORL | | Natural | |
|---|---|---|---|---|---|---|---|
| Size ($n$ $r$ $m$) | | 361 49 | 2,429 | 10,304 25 | 400 | 288 72 | 10,000 |
| Time limit (in seconds) | | 25 | 50 | 25 | 50 | 25 | 50 |
| Objective Value | Original | 954.10 | 907.10 | 15.52 | 14.11 | 362865.54 | 351006.54 |
| | Modified | 971.47 | 917.86 | 16.20 | 14.31 | 377460.30 | 353900.09 |
| Iteration | Original | 256.33 | 412.00 | 138.00 | 274.00 | 79.00 | 156.00 |
| | Modified | 224.00 | 361.67 | 117.00 | 232.00 | 59.67 | 117.67 |

We claim that $W^*_{ib} = 0, \forall\, i$. Otherwise, there is an index $i$ such that

$$\lim_{k\in\mathcal{K},k\to\infty}(W^kH^k)_{ij} \geq \lim_{k\in\mathcal{K},k\to\infty} W^k_{ib}H^k_{bj} = \infty.$$

Then

$$\lim_{k\to\infty} f(W^k,H^k) \geq \lim_{k\in\mathcal{K},k\to\infty} \left|(W^kH^k)_{ij} - V_{ij}\right|^2 = \infty$$

contradicts Theorem 4, which shows that $f(W^k,H^k)$ is strictly decreasing. Since $W^k$'s column sum is either one or zero, $W^*_{ib} = 0, \forall\, i$ implies

$$W^k_{ib} = 0 \qquad \forall\, i \qquad \forall\, k \in \mathcal{K} \text{ large enough.} \tag{42}$$

Then

$$\begin{aligned}\nabla_H f(W^k,H^k)_{bj} &= 0 \qquad \forall\, k \in \mathcal{K}\\ \text{so } H^{k,n}_{bj} &= H^k_{bj} \qquad \forall\, k \in \mathcal{K}.\end{aligned} \tag{43}$$

Moreover, by (42) and a similar argument in Theorem 5

$$\lim_{k\in\mathcal{K},k\to\infty} W^{k,n}_{ib} = \lim_{k\in\mathcal{K},k\to\infty} W^k_{ib} = 0 \qquad \forall\, i.$$

Thus, in normalizing $(W^{k,n},H^{k,n})$, $H^{k,n}$'s $b$th row is either unchanged ($W^{k,n}_{ib} = 0, \forall\, i$) or decreased $\left(\sum_i W^{k,n}_{ib} < 1\right)$. With (43)

$$H^{k+1}_{bj} \leq H^k_{bj} \qquad \forall\, k \in \mathcal{K}$$

an inequality contradicting (40). Thus, $\{H^k\}$ is bounded.

Therefore, $\{W^k,H^k\}$ is in a compact set, so there is at least one convergent subsequence. ■

## V. EXPERIMENTS

Though the modified algorithm has the same computational complexity per iteration, it is important to check its practical performance. We implemented both original and modified multiplicative updates in MATLAB. We set $\sigma = \delta = 10^{-8}$ in Algorithm 2 and give the code in Section B of the Appendix. Clearly, our modifications can be easily implemented.

We consider the following three image problems used in [5] and [13]:

1) Center for Biological and Computational Learning (CBCL, MIT, Cambridge, MA) face image database;
2) ORL face image database[2];
3) natural image data set [11].

Details of these sets and settings are in [13]. For the original multiplicative updates, we also normalize $W$ and $H$ after each iteration. We compare objective values and number of iterations after running 25 and 50 s. Experiments were conducted on an Intel Xeon 2.8-GHz computer.

Table I reports the average results of using the following two random initial points:

$$W_{ia} = |N(0,1)| \quad H_{bj} = |N(0,1)|$$

where $N(0,1)$ is the normal distribution.

Clearly, the modified algorithm is slower as the objective value is higher and the number of iterations is smaller. However, the difference between two objective values is rather small. Hence, in practice, if one would like to safely use multiplicative update algorithms, our modification is a possible choice.

## VI. DISCUSSION AND CONCLUSIONS

Earlier work such as [11] and [12] adds penalty terms to increase the sparsity of $W$ and $H$

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{ij} - (WH)_{ij}\right)^2 + \delta\sum_{i,a} W_{ia} + \delta\sum_{b,j} H_{bj}. \tag{44}$$

Then, the update rule (3) becomes

$$\begin{aligned}H^{k,n}_{bj} &= H^k_{bj} - \frac{H^k_{bj}}{((W^k)^TW^kH^k)_{bj}+\delta}\\ &\quad\times\left(((W^k)^TW^kH^k)_{bj} + \delta - ((W^k)^TV)_{bj}\right)\\ &= H^k_{bj}\frac{((W^k)^TV)_{bj}}{((W^k)^TW^kH^k)_{bj}+\delta}.\end{aligned}$$

For this formulation, the penalty parameter $\delta$ 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 $\delta$.

Section I mentioned that [9] addressed the convergence of minimizing the KL divergence formula (2). Here, we discuss the difference between their and our work. In [9], $W$ and $H$ are reparameterized to two other matrix variables. Multiplicative update rules are also reformulated accordingly. In contrast, we keep working on $W$ and $H$, but slightly modify the update rule. Then, [9] nicely proves the global convergence of the two new matrix variables. Since we try to specifically control the step size, our modified update rules help the numerical stability. The analysis in [9] does not have such a property. In fact, for the analysis, they define $0/0 = 0$, which may cause problems in practical implementations.

[2]http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Gonzales and Zhang [6] proposed a method to accelerate Lee and Seung's multiplicative updates. Instead of using

$$h_b^{k,n} = h_b^k - \frac{\bar{h}_b^k}{((W^k)^T W^k \bar{\mathbf{h}}^k)_b + \delta} \nabla \bar{f}(\mathbf{h}^k)_b, \qquad b = 1, \ldots, r$$

they consider

$$h_b^{k,n} = h_b^k - \alpha \frac{\bar{h}_b^k}{((W^k)^T W^k \bar{\mathbf{h}}^k)_b + \delta} \nabla \bar{f}(\mathbf{h}^k)_b, \qquad b = 1, \ldots, r$$

where $\alpha > 1$. Thus, they use a larger step size along the negative gradient direction. This modification causes problems in proving Lemma A. Hence, we cannot directly extend Theorem 3 to prove the strict decrease of function values. How to analyze the convergence of Gonzales and Zhang's method is an interesting future research issue.

In summary, this paper has the following two main contributions.

1) Under minor modifications, any limit point of [4]'s multiplicative update algorithms is a stationary point.
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 may apply the same approach to prove the convergence.

## APPENDIX

### A. Technical Lemma

*Lemma 1:* Given $\delta > 0$, an $r \times r$ symmetric positive–semidefinite 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

$$x_b > 0 \text{ if } b \in I \tag{45}$$

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

$$\bar{D}_{bb} \equiv \begin{cases} \frac{(A\mathbf{x})_b + \delta}{x_b}, & \text{if } b \in I \\ 0, & \text{if } b \notin I \end{cases}. \tag{46}$$

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

*Proof:* For any vector $\mathbf{v}$ with $\mathbf{v}_I \neq \mathbf{0}$

$$\mathbf{v}_I^T (\bar{D} - A)_{II} \mathbf{v}_I$$

$$= \sum_{a \in I} v_a^2 \frac{\delta}{x_a} + \sum_{a \in I} v_a^2 \frac{(A\mathbf{x})_a}{x_a} - \sum_{a,b \in I} v_a v_b A_{ab} \tag{47}$$

$$> \sum_{a \in I} v_a^2 \frac{\sum_{b \in I} A_{ab} x_b}{x_a} - \sum_{a,b \in I} v_a v_b A_{ab} \tag{48}$$

$$= \frac{1}{2} \sum_{a,b \in I} v_a^2 \frac{A_{ab} x_b}{x_a} + \frac{1}{2} \sum_{a,b \in I} v_b^2 \frac{A_{ba} x_a}{x_b} - \sum_{a,b \in I} v_a v_b A_{ab} \tag{49}$$

$$= \frac{1}{2} \sum_{a,b} A_{ab} \left( \sqrt{\frac{x_b}{x_a}} v_a - \sqrt{\frac{x_a}{x_b}} v_b \right)^2 \geq 0. \tag{50}$$

The condition (45) ensures (47) to be well defined. From (47) and (48), we use the properties $A_{ab} \geq 0$ and $x_b \geq 0, \forall\, b = 1, \ldots, r$. From (49) and (50), the symmetry of $A$ is used. ■

### B. MATLAB Code of Algorithm 2

```
function [W,H,iter] = multconv(V,W0,H0,timelimit,maxiter)

sigma = 1.0e-8; delta = sigma;

W = W0; H = H0;

[m,n] = size(V);

initt = cputime; t = initt;

for inter = 1:maxiter,
    if (iter == maxiter | cputime-initt > timelimit),
     fprintf ('iter %d n', iter);
     break
    end

    WtW = W' * W;
    gradH = WtW * H - W' * V;
    Hb = max(H, (gradH < 0) * sigma);
    H = H - Hb./(WtW * Hb + delta). * gradH;

    HHt = H * H';
    gradW = W * (HHt) - V * H';
    Wb = max(H, (gradH < 0) * sigma);
    W = W - Wb./(Wb * HHt + delta). * gradW;

    S = sum(W,1);
    W = W./repmat(S,m,1); H = H. * repmat(S',1,n);
end
```

## REFERENCES


[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] M. Berry, M. Browne, A. Langville, P. Pauca, and R. Plemmon, "Algorithms and applications for approximate nonnegative matrix factorization," *Comput. Statist. Data Anal.*, 2007, to be published.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 556–562.

[5] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Comput.*, 2007, to be published.

[6] E. F. Gonzales and Y. Zhang, "Accelerating the Lee-Seung algorithm for non-negative matrix factorization," Dept. Comput. Appl. Math., Rice Univ., Houston, TX, Tech. Rep., 2005.

[7] M. Catral, L. Han, M. Neumann, and R. Plemmons, "On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices," *Linear Algebra Appl.*, vol. 393, pp. 107–126, 2004.
[8] M. Chu, F. Diele, R. Plemmons, and S. Ragni, Theory, Numerical Methods Appl. Nonnegative Matrix Factorization Tech. Rep., 2004 [Online]. Available: http://www.wfu.edu/~plemmons/papers/chu_ple.pdf
[9] L. Finesso and P. Spreij, "Nonnegative matrix factorization and I-divergence alternating minimization," *Linear Algebra Appl.*, vol. 416, no. 2-3, pp. 270–287, 2006.
[10] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999, pp. 02178–9998.
[11] P. O. Hoyer, "Non-negative sparse coding," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, 2002, pp. 557–565.
[12] J. Piper, P. Pauca, R. Plemmons, and M. Giffin, "Object characterization from spectral data using nonnegative factorization and information theory," in *Proc. AMOS Tech. Conf.*, 2004.
[13] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

**Chih-Jen Lin** (S'91–M'98–SM'06) received the B.S. degree in mathematics from the National Taiwan University, Taipei, Taiwan, in 1993 and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1996 and 1998, respectively.

Currently, he is a Professor at the Department of Computer Science, National Taiwan University. His research interests include machine learning, data mining, and related applications.