2014 International Symposium on Nonlinear Theory and its Applications
NOLTA2014, Luzern, Switzerland, September 14-18, 2014

A Generalized Sufficient Condition for Global Convergence of Modified Multiplicative Updates for NMF

Norikazu Takahashi[†,♯], Jiro Katayama[‡] and Jun'ichi Takeuchi[♭,♯]

† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama 700–8530, Japan
‡ Graduate School of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819–0395 Japan
♭ Faculty of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819–0395 Japan
♯ Institute of Systems, Information Technologies and Nanotechnologies
2–1–22 Momochihama, Sawara-ku, Fukuoka 814–0001 Japan
Email: takahashi@cs.okayama-u.ac.jp

Abstract—Multiplicative updates are widely used computational methods for nonnegative matrix factorization (NMF). However, the global convergence of the original updates is not theoretically guaranteed. By the global convergence, we mean that the sequence of solutions contains at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the NMF optimization problem. In this paper, we consider a modified multiplicative update for a general error function and give a sufficient condition for the global convergence.

1. Introduction

Nonnegative matrix factorization (NMF) is a technique to decompose a given nonnegative matrix into the product of two nonnegative matrices. Since NMF is useful for finding a set of nonnegative basis vectors for the given nonnegative data, it has attracted great attention from researchers in machine learning, signal processing, pattern classification, statistics, and so on.

Multiplicative updates proposed by Lee and Seung [1,2] are widely used as an efficient computational approach to NMF. They derived two kinds of multiplicative updates based on Euclidean distance and I-divergence by taking the following two steps. The first step is to construct an auxiliary function for the original error function, and the second one is to find its unique minimum point. Yang and Oja [3] recently generalized this procedure and derived eleven multiplicative updates including the ones of Lee and Seung. However, all of the eleven multiplicative updates are not well-defined because each of them contains a rational function and its denominator can become zero. A simple way to avoid this problem is to apply the modification proposed by Gillis and Glineur [4]. Moreover, it has recently been proved that this modification makes the updates of Lee and Seung globally convergent [5,6].

In this paper, we generalize the global convergence analysis presented in Reference [6]. More specifically, we consider a general error function and show that under some conditions on the auxiliary function the modified multiplicative update is globally convergent. We also show that eight among eleven updates presented in Reference [3] satisfy those conditions.

2. NMF and Multiplicative Updates

Given a nonnegative matrix $\boldsymbol{X} \in \mathbb{R}_+^{m\times n}$, where $\mathbb{R}_+$ denotes the set of nonnegative numbers, let us consider the problem of finding two nonnegative matrices $\boldsymbol{W} \in \mathbb{R}_+^{m\times r}$ and $\boldsymbol{H} \in \mathbb{R}_+^{r\times n}$ such that

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \tag{1}$$

where r is a positive integer less than $\min\{m, n\}$. A variety of techniques to find $\boldsymbol{W}$ and $\boldsymbol{H}$ in (1) are called nonnegative matrix factorization (NMF). Although it is important in NMF how to set the value of r, we will not consider this problem in this paper: we simply assume that the value of r is given together with $\boldsymbol{X}$. Also, we assume throughout this paper that every row and column of $\boldsymbol{X}$ has at least one nonzero entry.

The problem of finding $\boldsymbol{W}$ and $\boldsymbol{H}$ in (1) is usually formulated as a mathematical programming problem of the following form:

$$\begin{array}{ll} \text{minimize} & D(\boldsymbol{W}, \boldsymbol{H}) \\ \text{subject to} & \boldsymbol{W} \geq \boldsymbol{O}_{m\times r},\ \boldsymbol{H} \geq \boldsymbol{O}_{r\times n} \end{array} \tag{2}$$

where $D(\boldsymbol{W}, \boldsymbol{H})$ is an error function and $\boldsymbol{O}_{m\times r}$ ($\boldsymbol{O}_{r\times n}$, resp.) is the $m \times r$ ($r \times n$, resp.) zero matrix. So far, various error functions such as Euclidean distance and I-divergence have been used for NMF.

This work was partially supported by KAKENHI 24560076, 24500018 and 23310104, and by Proactive Response Against Cyber-attacks Through International Collaborative Exchange (PRACTICE), Ministry of Internal Affairs and Communications, Japan.

Because (2) is a nonconvex optimization problem, it is difficult to find its global optimal solution. As an approach to find local optimal solutions, multiplicative updates [1–3, 7] are widely used. For example, the multiplicative update rule developed by Lee and Seung [1, 2] for Euclidean distance error function $D(\boldsymbol{W},\boldsymbol{H}) = \sum_{ij}(X_{ij} - (\boldsymbol{W}\boldsymbol{H})_{ij})^2$ is given by

$$W_{ik}^{(l+1)} = W_{ik}^{(l)} \frac{(\boldsymbol{X}(\boldsymbol{H}^{(l)})^T)_{ik}}{(\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)}(\boldsymbol{H}^{(l)})^T)_{ik}}, \quad (3)$$

$$H_{kj}^{(l+1)} = H_{kj}^{(l)} \frac{((\boldsymbol{W}^{(l+1)})^T\boldsymbol{X})_{kj}}{((\boldsymbol{W}^{(l+1)})^T\boldsymbol{W}^{(l+1)}\boldsymbol{H}^{(l)})_{kj}}, \quad (4)$$

where l represents the number of iterations and $W_{ik}^{(l)}$ ($H_{kj}^{(l)}$, resp.) is the value of W_{ik} (H_{kj}, resp.) after l updates. One can easily see from (3) and (4) that if the initial matrices $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$ are positive then $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$ are positive for all $l \geq 1$. Also, it is known that the sequence $\{D(\boldsymbol{W}^{(l)},\boldsymbol{H}^{(l)})\}_{l=1}^{\infty}$ is monotone decreasing [2], which means that the sequence converges to some constant because it is bounded from below. However, this does not imply that the sequence $\{(\boldsymbol{W}^{(l)},\boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ converges to a local optimal solution of (2).

The multiplicative update rule given by (3) and (4) is derived by constructing an auxiliary function for the error function $D(\boldsymbol{W},\boldsymbol{H}) = \sum_{ij}(X_{ij} - (\boldsymbol{W}\boldsymbol{H})_{ij})^2$ and minimizing it [2]. In this paper, a function $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) : \mathbb{R}_{++}^{m\times r} \times \mathbb{R}_{++}^{r\times n} \times \mathbb{R}_{++}^{m\times r} \times \mathbb{R}_{++}^{r\times n} \to \mathbb{R}$, where $\mathbb{R}_{++}$ denotes the set of positive numbers, is called an auxiliary function of $\boldsymbol{D}(\boldsymbol{W},\boldsymbol{H})$ if the following two conditions are satisfied:

$$\forall \boldsymbol{W} > \boldsymbol{O}_{m\times r}, \boldsymbol{H} > \boldsymbol{O}_{r\times n}, \widetilde{\boldsymbol{W}} > \boldsymbol{O}_{m\times r}, \widetilde{\boldsymbol{H}} > \boldsymbol{O}_{r\times n}, \quad \bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \geq D(\boldsymbol{W},\boldsymbol{H}),$$

$$\forall \boldsymbol{W} > \boldsymbol{O}_{m\times r}, \boldsymbol{H} > \boldsymbol{O}_{r\times n}, \quad \bar{D}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{W},\boldsymbol{H}) = D(\boldsymbol{W},\boldsymbol{H}).$$

Note that this approach is not restricted to Euclidean distance error function but can be applied to various error functions. In fact, Yang and Oja [3] proposed a unified approach to develop multiplicative update rules and applied it to eleven error functions (see Table 1).

3. Global Convergence of Modified Update Rules

The multiplicative update rule given by (3) and (4) and other update rules shown in Table 1 have a common serious problem that they are not well-defined. In order to solve this problem, Gillis and Glineur [4] proposed to modify (3) and (4) as

$$W_{ik}^{(l+1)} = \max\left(\epsilon, W_{ik}^{(l)} \frac{(\boldsymbol{X}(\boldsymbol{H}^{(l)})^T)_{ik}}{(\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)}(\boldsymbol{H}^{(l)})^T)_{ik}}\right), \quad (5)$$

$$H_{kj}^{(l+1)} = \max\left(\epsilon, H_{kj}^{(l)} \frac{((\boldsymbol{W}^{(l+1)})^T\boldsymbol{X})_{kj}}{((\boldsymbol{W}^{(l+1)})^T\boldsymbol{W}^{(l+1)}\boldsymbol{H}^{(l)})_{kj}}\right), \quad (6)$$

where ϵ is a small positive constant specified by the user. This simple modification can be applied to all multiplicative updates. Note that, with this modification, the problem (2) has to be modified as

$$\begin{array}{ll} \text{minimize} & D(\boldsymbol{W},\boldsymbol{H}) \\ \text{subject to} & \boldsymbol{W} \geq \epsilon\mathbf{1}_{m\times r},\ \boldsymbol{H} \geq \epsilon\mathbf{1}_{r\times n} \end{array} \quad (7)$$

where $\mathbf{1}_{m\times r}$ ($\mathbf{1}_{r\times n}$, resp.) is the $m \times r$ ($r \times n$, resp.) matrix consisting of all ones. Hibi and Takahashi [5] proved that the update rule given by (5) and (6) has the global convergence property in the sense of Zangwill [8]. Also, they have recently shown that this modification can also guarantee the global convergence of the multiplicative update for I-divergence [6].

In the following, we will consider a general error function $D(\boldsymbol{W},\boldsymbol{H})$ and give a sufficient condition on the auxiliary function $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ for the modified update rule to be globally convergent. Before proceeding further, we need to introduce some notations. The feasible region of the problem (7) is denoted by F_ϵ, that is,

$$F_\epsilon = \{(\boldsymbol{W},\boldsymbol{H}) \,|\, \boldsymbol{W} \geq \epsilon\mathbf{1}_{m\times r},\ \boldsymbol{H} \geq \epsilon\mathbf{1}_{r\times n}\}.$$

A point $(\boldsymbol{W},\boldsymbol{H})$ that satisfies Karush-Kuhn-Tucker (KKT) conditions for (7):

$$\begin{aligned} \boldsymbol{W} &\geq \epsilon\mathbf{1}_{m\times r}, \\ \boldsymbol{H} &\geq \epsilon\mathbf{1}_{r\times n}, \\ \frac{\partial D(\boldsymbol{W},\boldsymbol{H})}{\partial W_{ik}} &\geq 0, \quad \forall i,k, \\ \frac{\partial D(\boldsymbol{W},\boldsymbol{H})}{\partial H_{kj}} &\geq 0, \quad \forall k,j, \\ \frac{\partial D(\boldsymbol{W},\boldsymbol{H})}{\partial W_{ik}}(\epsilon - W_{ik}) &= 0, \quad \forall i,k, \\ \frac{\partial D(\boldsymbol{W},\boldsymbol{H})}{\partial H_{kj}}(\epsilon - H_{kj}) &= 0, \quad \forall k,j, \end{aligned}$$

is called a stationary point of (7). The set of stationary points is denoted by S_ϵ.

The following assumptions are also needed for later discussions.

Assumption 1 For any $\widetilde{\boldsymbol{W}} \in \mathbb{R}_{++}^{m\times r}$ and $\widetilde{\boldsymbol{H}} \in \mathbb{R}_{++}^{r\times n}$, $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is differentiable with respect to W_{ik}, and satisfies

$$\left.\frac{\partial \bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})}{\partial W_{ik}}\right|_{(\boldsymbol{W},\boldsymbol{H})=(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})} = \left.\frac{\partial D(\boldsymbol{W},\boldsymbol{H})}{\partial W_{ik}}\right|_{(\boldsymbol{W},\boldsymbol{H})=(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})}.$$

Assumption 2 $\bar{D}(\boldsymbol{W},\widetilde{\boldsymbol{H}},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ can be expressed as $\sum_{ik} u_{ik}(W_{ik},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ and $u_{ik}(W_{ik},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is strictly convex with respect to W_{ik} on $\mathbb{R}_{++}$. Also, for each $(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \in \mathbb{R}_{++}^{m\times r} \times \mathbb{R}_{++}^{r\times n}$, the problem

$$\begin{array}{ll} \text{minimize} & u_{ik}(W_{ik},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \\ \text{subject to} & W_{ik} > 0 \end{array}$$

Table 1: Error functions and multiplicative update rules [3]. $\boldsymbol{Z} = (Z_{ij})$ is defined by $Z_{ij} = X_{ij}/(\boldsymbol{W}\boldsymbol{H})_{ij}$.

Error function	Multiplicative updates for W_{ik}
Euclidean distance	$W_{ik}^{\text{new}} = W_{ik} \frac{(\boldsymbol{X}\boldsymbol{H}^T)_{ik}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T)_{ik}}$
I-divergence	$W_{ik}^{\text{new}} = W_{ik} \frac{(\boldsymbol{Z}\boldsymbol{H}^T)}{\sum_j H_{kj}}$
Dual I-divergence	$W_{ik}^{\text{new}} = W_{ik} \exp\left(\frac{\sum_j (\ln Z_{ij}) H_{kj}}{\sum_j H_{kj}}\right)$
Itakura-Saito divergence	$W_{ik}^{\text{new}} = W_{ik} \sqrt{\frac{\sum_j X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^{-2} H_{kj}}{\sum_j (\boldsymbol{W}\boldsymbol{H})_{ij}^{-1} H_{kj}}}$
α-divergence	$W_{ik}^{\text{new}} = \begin{cases} W_{ik}\left(\frac{\sum_j Z_{ij}^{\alpha} H_{kj}}{\sum_j H_{kj}}\right)^{\frac{1}{\alpha}}, & \alpha \neq 0 \\ W_{ik}\exp\left(\frac{\sum_j (\ln Z_{ij}) H_{kj}}{\sum_j H_{kj}}\right), & \alpha = 0 \end{cases}$
β-divergence	$W_{ik}^{\text{new}} = W_{ik}\left(\frac{\sum_j X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^{\beta-1} H_{kj}}{\sum_j (\boldsymbol{W}\boldsymbol{H})_{ij}^{\beta} H_{kj}}\right)^{\eta}, \quad \eta = \begin{cases} \frac{1}{\beta}, & \beta > 1 \\ 1, & 0 < \beta \leq 1 \\ \frac{1}{1-\beta}, & \beta < 0 \end{cases}$
Log-Quad cost	$W_{ik}^{\text{new}} = W_{ik} \sqrt{\frac{(\boldsymbol{Z}\boldsymbol{H}^T + 2\boldsymbol{X}\boldsymbol{H}^T)_{ik}}{\sum_j H_{kj} + 2(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T)_{ik}}}$
$\alpha\beta$-Bregman divergence	$W_{ik}^{\text{new}} = W_{ik}\left(\frac{\alpha(\alpha-1)\sum_j X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^{\alpha-2} H_{kj} + \beta(1-\beta)\sum_j X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^{\beta-2} H_{kj}}{\alpha(\alpha-1)\sum_j (\boldsymbol{W}\boldsymbol{H})_{ij}^{\alpha-1} H_{kj} + \beta(1-\beta)\sum_j (\boldsymbol{W}\boldsymbol{H})_{ij}^{\beta-1} H_{kj}}\right)^{\frac{1}{\alpha-\beta+1}} \quad (\alpha \geq 1, 0 < \beta < 1)$
Kullback-Leibler divergence	$W_{ik}^{\text{new}} = W_{ik} \frac{(\boldsymbol{Z}\boldsymbol{H}^T)_{ik}}{\sum_j H_{kj}} \sum_{ab} (\boldsymbol{W}\boldsymbol{H})_{ab}$
γ-divergence	$W_{ik}^{\text{new}} = W_{ik}\left(\frac{\sum_j X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^{\gamma-1} H_{kj}}{\sum_j (\boldsymbol{W}\boldsymbol{H})_{ij}^{\gamma} H_{kj}} \cdot \frac{\sum_{ab} (\boldsymbol{W}\boldsymbol{H})_{ab}^{1+\gamma}}{\sum_{ab} X_{ab}(\boldsymbol{W}\boldsymbol{H})_{ab}^{\gamma}}\right)^{\eta}, \quad \eta = \begin{cases} \frac{1}{1+\gamma}, & \gamma > 0 \\ \frac{1}{1-\gamma}, & \gamma < 0 \end{cases}$
Rényi divergence	$W_{ik}^{\text{new}} = W_{ik}\left(\frac{\sum_j Z_{ij}^{r} H_{kj}}{\sum_j H_{kj}} \cdot \frac{\sum_{ab} (\boldsymbol{W}\boldsymbol{H})_{ab}}{\sum_{ab} X_{ab}^{r}(\boldsymbol{W}\boldsymbol{H})_{ab}^{1-r}}\right)^{\eta}, \quad \eta = \begin{cases} \frac{1}{r}, & r > 1 \\ 1, & 0 < r < 1 \end{cases}$

has a unique optimal solution that can be explicitly expressed as $W_{ik}^* = f_{ik}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$. Furthermore, $f_{ik}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is continuous and, for each $\epsilon > 0$, there exist $c_{ik} > 0$ and $\nu_{ik} < 1$ such that

$$\forall \widetilde{\boldsymbol{W}} \geq \epsilon \mathbf{1}_{m\times r}, \widetilde{\boldsymbol{H}} \geq \epsilon \mathbf{1}_{r\times n}, \quad f_{ik}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \leq c_{ik} \widetilde{W}_{ik}^{\nu_{ik}}. \quad (8)$$

Assumption 3 For any $\widetilde{\boldsymbol{W}} \in \mathbb{R}_{++}^{m\times r}$ and $\widetilde{\boldsymbol{H}} \in \mathbb{R}_{++}^{r\times n}$, $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is differentiable with respect to H_{kj}, and satisfies

$$\left.\frac{\partial \bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})}{\partial H_{kj}}\right|_{(\boldsymbol{W},\boldsymbol{H})=(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})} = \left.\frac{\partial D(\boldsymbol{W}, \boldsymbol{H})}{\partial H_{kj}}\right|_{(\boldsymbol{W},\boldsymbol{H})=(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})}.$$

Assumption 4 $\bar{D}(\widetilde{\boldsymbol{W}}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ can be expressed as $\sum_{kj} v_{kj}(H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ and $v_{kj}(H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is strictly convex with respect to H_{kj} on $\mathbb{R}_{++}$. Also, for each $(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \in \mathbb{R}_{++}^{m\times r} \times \mathbb{R}_{++}^{r\times n}$, the problem

$$\begin{array}{ll} \text{minimize} & v_{ik}(H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \\ \text{subject to} & H_{kj} > 0 \end{array}$$

has a unique optimal solution that can be explicitly expressed as $H_{kj}^* = g_{kj}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$. Furthermore, $g_{kj}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is continuous and, for each $\epsilon > 0$, there exist $d_{kj} > 0$ and $\mu_{kj} < 1$ such that

$$\forall \widetilde{\boldsymbol{W}} \geq \epsilon \mathbf{1}_{m\times r}, \widetilde{\boldsymbol{H}} \geq \epsilon \mathbf{1}_{r\times n}, \quad g_{kj}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \leq d_{kj} \widetilde{H}_{kj}^{\mu_{kj}}. \quad (9)$$

The next theorem is the main result of this paper.

Theorem 1 Let ϵ be any positive number. If an auxiliary function $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ of $D(\boldsymbol{W}, \boldsymbol{H})$ satisfies Assumptions 1–4 then, for any initial matrices $\boldsymbol{W}^{(0)} \geq \epsilon \mathbf{1}_{m\times r}$ and $\boldsymbol{H}^{(0)} \geq \epsilon \mathbf{1}_{r\times n}$, the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by the update rule given by

$$\begin{aligned} W_{ik}^{(l+1)} &= \max(\epsilon, f_{ik}(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})) \\ H_{kj}^{(l+1)} &= \max(\epsilon, g_{kj}(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)})) \end{aligned}$$

has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the problem (7).

Proof Proof sketch will be given in the next section. □

Let us now consider eleven update rules shown in Table 1. By analyzing the unified method [3] for developing multiplicative update rules, we can prove that the auxiliary functions for the first eight error functions (Euclidean distance, I-divergence, Dual I-divergence, Itakura-Saito divergence, α-divergence, β-divergence, Log-Quad cost and $\alpha\beta$-Bregman divergence) satisfy Assumptions 1–4. Therefore, we can conclude that the modified update rules corresponding to the first eight update rules in Table 1 have a global convergence property. On the other hand, the global convergence of the last three update rules cannot be proved by Theorem 1 because the inequalities (8) and (9) are not satisfied [9]. This issue will be discussed in [10].

4. Proof Sketch of Theorem 1

Let (5) and (6) be expressed as $W^{(l+1)} = A_1(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})$ and $H^{(l+1)} = A_2(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)})$, respectively. Moreover, let the update from $(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})$ to $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)})$ be expressed as $A(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) = (\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)})$. Then we have

$$\begin{aligned}
&(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)}) \\
&\quad = A(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \\
&\quad = (A_1(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}), A_2(A_1(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}), \boldsymbol{H}^{(l)}))\,.
\end{aligned}$$

We prove Theorem 1 by using Zangwill's global convergence theorem [8]. To do so, we need to show that the following statements hold true.

1. (Boundedness) For each initial value $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in F_\epsilon$, the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by the update rule A belongs to a compact subset of F_ϵ.
2. (Monotoneness) There exists a function $h : F_\epsilon \to \mathbb{R}$ such that
$$\begin{aligned}
(\boldsymbol{W}, \boldsymbol{H}) \notin S_\epsilon &\Rightarrow h(A(\boldsymbol{W}, \boldsymbol{H})) < h(\boldsymbol{W}, \boldsymbol{H})\,, \\
(\boldsymbol{W}, \boldsymbol{H}) \in S_\epsilon &\Rightarrow h(A(\boldsymbol{W}, \boldsymbol{H})) \leq h(\boldsymbol{W}, \boldsymbol{H})\,.
\end{aligned}$$
3. (Continuity) A is continuous in $F_\epsilon \setminus S_\epsilon$.

Among these three statements, the boundedness of A can be directly proved by Assumptions 2 and 4 and [9, Lemma 1]. The continuity of A is also apparent from Assumptions 2 and 4. So it suffices for us to show the monotoneness of A. This can be done by using Assumptions 1–4 and the following lemmas. However, we omit the details due to lack of space.

Lemma 1 Let $(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ be any point in F_ϵ. Then

$$\begin{array}{ll} \text{minimize} & u(\boldsymbol{W}) = \bar{D}(\boldsymbol{W}, \widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \\ \text{subject to} & \boldsymbol{W} \geq \epsilon \mathbf{1}_{m \times r} \end{array} \tag{10}$$

has a unique optimal solution $\boldsymbol{W}^* = A_1(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$. Also,

$$\begin{array}{ll} \text{minimize} & v(\boldsymbol{H}) = \bar{D}(\widetilde{\boldsymbol{W}}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \\ \text{subject to} & \boldsymbol{H} \geq \epsilon \mathbf{1}_{r \times n} \end{array} \tag{11}$$

has a unique optimal solution $\boldsymbol{H}^* = A_2(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$.

Lemma 2 The necessary and sufficient condition for $(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ to be a stationary point of (7) is that $\widetilde{\boldsymbol{W}}$ is the unique optimal solution of (10) and $\widetilde{\boldsymbol{H}}$ is the unique optimal solution of (11).

5. Conclusion

We have given a sufficient condition for the modified multiplicative update for NMF to be globally convergent. Because the sufficient condition is very general, the result of this paper can be applied to many multiplicative updates. In fact, the global convergence of the modified versions of eight among the eleven multiplicative updates presented in [3] is guaranteed by Theorem 1.

References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–792, 1999.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[3] Z. Yang and E. Oja, "Unified development of multiplicative algorithm for linear and quadratic nonnegative matrix factorization," *IEEE Trans. Neural Networks*, vol. 22, pp. 1878–1891, December 2011.

[4] N. Gillis and F. Glineur, "Nonnegative factorization and the maximum edge biclique problem," *arXiv e-prints*, October 2008.

[5] R. Hibi and N. Takahashi, "A modified multiplicative update algorithm for Euclidean distance-based nonnegative matrix factorization and its global convergence," in *Proc. ICONIP2011, Part-II*, pp. 655–662, November 2011.

[6] N. Takahashi and R. Hibi, "Global convergence of modified multiplicative updates for nonnegative matrix factorization," *Computational Optimization and Applications*, vol. 57, pp. 417–440, March 2014.

[7] L. Finesso and P. Spreij, "Nonnegative matrix factorization and I-divergence alternating minimization," *Linear Algebra and its Applications*, vol. 416, pp. 270–287, 2006.

[8] W. I. Zangwill, *Nonlinear programming: A unified approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.

[9] J. Katayama, N. Takahashi, and J. Takeuchi, "Boundedness of modified multiplicative updates for nonnegative matrix factorization," in *Proc. CAMSAP2013*, pp. 252–255, December 2013.

[10] M. Seki and N. Takahashi, "New updates based on Kullback-Leibler, gamma, and Rényi divergences for nonnegative matrix factorization," to appear in *Proc. NOLTA2014*.