A Unified Global Convergence Analysis of Multiplicative Update Rules for Nonnegative Matrix Factorization

Author(s): Norikazu Takahashi, Jiro Katayama, Masato Seki and Jun'ichi Takeuchi

Journal: Computational Optimization and Applications
Volume: 71
Number: 1
Pages: 221–250
Month: September
Year: 2018
DOI: 10.1007/s10589-018-9997-y

This is a post-peer-review, pre-copyedit version of an article published in Computational Optimization and Applications. The final authenticated version is available online at: http://dx.doi.org/10.1007/s10589-018-9997-y.

Noname manuscript No.
(will be inserted by the editor)

A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization

Norikazu Takahashi · Jiro Katayama · Masato Seki · Jun'ichi Takeuchi

Received: date / Accepted: date

Abstract Multiplicative update rules are a well-known computational method for nonnegative matrix factorization. Depending on the error measure between two matrices, various types of multiplicative update rules have been proposed so far. However, their convergence properties are not fully understood. This paper provides a sufficient condition for a general multiplicative update rule to have the global convergence property in the sense that any sequence of solutions has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the optimization problem. Using this condition, it is proved that many of the existing multiplicative update rules have the global convergence property if they are modified slightly so that all variables take positive values. This paper also proposes new multiplicative update rules based on Kullback-Leibler, Gamma, and Rényi divergences. It is shown that these three rules have the global convergence property if the same modification as above is made.

Part of this paper was presented in 2014 International Symposium on Nonlinear Theory and its Applications [34, 37].

N. Takahashi
Graduate School of Natural Science and Technology, Okayama University, 3–1–1 Tsushima-naka, Kita-ku, Okayama 700–8530, Japan
Tel.: +81–86–251–8179
Fax: +81–86–251–8256
E-mail: takahashi@cs.okayama-u.ac.jp

J. Katayama
Department of Informatics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819–0395, Japan

M. Seki
Graduate School of Natural Science and Technology, Okayama University, 3–1–1 Tsushima-naka, Kita-ku, Okayama 700–8530, Japan

J. Takeuchi
Department of Informatics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819–0395, Japan

Keywords Nonnegative matrix factorization · Multiplicative update rule · Global convergence

Mathematics Subject Classification (2010) 90C30 · 90C90 · 68W40 · 15A23

1 Introduction

Nonnegative matrix factorization (NMF), which was first introduced by Paatero and Tapper [32], is a mathematical operation that decomposes a given nonnegative matrix into two nonnegative low-rank factor matrices. Since the seminal papers by Lee and Seung [28,29], NMF has attracted a lot of of attention from researchers in different disciplines, and has been shown to be useful for many applications in pattern recognition [18], text mining [3], document clustering [35,43], signal processing [10], music analysis [20,33], graph analysis [40], cyber security [44], and so on.

NMF is formulated as an optimization problem in which an error between the given nonnegative matrix and the product of two factor matrices is minimized subject to the nonnegativity constraints on the factor matrices. Unlike nonnegative rank factorization [2, 4, 5] which has a longer history, the ranks of the factor matrices in NMF are allowed to be less than the maximum. However, because the objective function is nonconvex with respect to the two factor matrices, finding a global optimal solution is very difficult in general. In fact, it has been proved that the NMF optimization problem is NP-hard when the error is measured by Euclidean distance [39]. Therefore the goal in solving an NMF optimization problem is to find a local optimal solution.

During the last two decades, many algorithms for NMF have been developed [6–9, 11–13, 15, 17, 19, 21, 23–25, 27–29, 31, 45]. The most well-known and widely used algorithms are the two multiplicative update rules proposed by Lee and Seung [28,29]: one is for NMF with Euclidean distance and the other is for NMF with the generalized Kullback-Leibler divergence, also known as I-divergence. The main idea behind their algorithms is to repeat the minimization of an auxiliary function with respect to a subset of the variables until a certain stopping condition is satisfied. Because the minimization of the auxiliary function does not increase the error, it is guaranteed that the error decreases monotonically. This idea was later applied to various error measures such as Bregman divergence [11], a parametric divergence containing Euclidean distance and I-divergence as special cases [27], α-divergence [7], Itakura-Saito divergence [12], and β-divergence [13]. Furthermore, a unified method for deriving multiplicative update rules from a wide variety of error measures was proposed by Yang and Oja [45]. Using this method, they obtained eleven different multiplicative update rules from a collection of error measures including the ones mentioned above, Kullback-Leibler divergence, γ-divergence, Rényi divergence, and so on. Like the multiplicative update rules of Lee and Seung, all update rules obtained by this method have the property that the error decreases monotonically.

Although the original multiplicative update rules have some good properties as stated above, they have a serious drawback that the convergence to a stationary point is not guaranteed theoretically. The error decreases monotonically, but this does not mean the convergence of the sequence of solutions. In fact, it was shown by numerical experiments that the multiplicative update rule based on Euclidean distance does not converge to an local optimal solution for some problems [16]. Another drawback is that every multiplicative update is not well-defined in the feasible region of the NMF optimization problem. For example, if one of the two factor matrices is a zero matrix, the update rule is not defined because the denominator is zero. Here we should note that these two drawbacks are not independent but closely related to each other.

Many authors have studied the convergence property of the multiplicative update rules proposed by Lee and Seung and their variants. Finesso and Spreij [14] considered the I-divergence based multiplicative update rule and proved the convergence of the left factor matrix and the product of the two factor matrices under the assumption that the right factor matrix is normalized after each update. Lin [30] considered a modified version of the Euclidean distance based multiplicative update rule and proved that the sequence of solutions converges to a stationary point in some sense. Gillis and Glineur [15] considered another modified version of the Euclidean distance based multiplicative update rule, which returns a user-specified positive constant if the original update rule returns a value less than the constant, and proved that if the sequence of solutions converges then it is a stationary point. Takahashi and Hibi [36] proved that the multiplicative update rules based on Euclidean distance and I-divergence combined with the modification of Gillis and Glineur have the global convergence property in the sense that any sequence of solutions has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point. Badeau *et al.* [1] studied the local stability of a generalized multiplicative update rules by Lyapunov's stability theory and showed that the local optimal solution is asymptotically stable if one of the two factor matrices is fixed to a nonnegative constant matrix. Zhao and Tan recently performed a unified convergence analysis of the multiplicative update rules for NMF with ℓ_1 regularization [47].

In this paper, we consider a wide class of error measures including Euclidean distance, I-divergence and all other divergences mentioned above, and provide a general sufficient condition on the error measure and the auxiliary function for the obtained multiplicative update rule combined with the modification of Gillis and Glineur [15] to have the global convergence property in the same sense as Takahashi and Hibi [36]. We then apply our results to the eleven error measures considered by Yang and Oja [45], and show that all of them satisfy the sufficient condition for the global convergence except Kullback-Leibler divergence, γ-divergence and Rényi divergence. Finally, we propose three new error measures based on these divergences and show that the obtained update rules have the global convergence property.

It is often said that the multiplicative update rules are slow. In fact, many faster algorithms have been proposed for NMF with Euclidean distance [8,

9, 15, 17, 21, 23–25, 31]. Nevertheless, the multiplicative update rules are still important because of its simplicity, easiness to implement, and applicability to a wide range of error measures. We therefore focus our attention on the multiplicative update rules in this paper, and clarify the global convergence condition. For the convergence properties of other algorithms, see [6,19,25,26] and references therein.

The rest of this paper is organized as follows. In Section 2, NMF and the multiplicative update rule are briefly reviewed. In Section 3, the global convergence of the modified multiplicative update rule is proved under some mild assumptions on the error measure and the auxiliary function. In Section 4, the result in Section 3 is applied to the eleven multiplicative updates obtained by the unified method of Yang and Oja, and their global convergence is proved except Kullback-Leibler divergence, γ-divergence, and Rényi divergence. In Section 5, three new multiplicative update rules corresponding to these three error measures are derived, and their global convergence is proved. In Section 6 we conclude the paper.

2 NMF and Multiplicative Update Rules

2.1 NMF Optimization Problem

Given a nonnegative matrix $\boldsymbol{X} \in \mathbb{R}_+^{m\times n}$, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, we 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 m and n. The operation that decomposes a given nonnegative matrix $\boldsymbol{X}$ into two nonnegative factor matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ as shown in (1) is called nonnegative matrix factorization (NMF). Although it is important in NMF how to choose the value of r, we do not consider this issue 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 formulated as a constrained optimization problem:

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

where $D(\boldsymbol{W},\boldsymbol{H})$ is a function representing the error between $\boldsymbol{X}$ and $\boldsymbol{W}\boldsymbol{H}$ and $\boldsymbol{O}_{m\times r}$ ($\boldsymbol{O}_{r\times n}$, resp.) is the $m \times r$ ($r \times n$, resp.) zero matrix. The inequality signs between matrices are to be interpreted componentwise. In the remainder of this paper, we call $D(\boldsymbol{W},\boldsymbol{H})$ the error function. Also, we denote the feasible region of (2) by $\mathcal{F}_0$, that is, $\mathcal{F}_0 = \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{r\times n}$.

So far, various kinds of error functions such as the one based on Euclidean distance:

$$D(\boldsymbol{W},\boldsymbol{H}) = \sum_{ij} (X_{ij} - (\boldsymbol{W}\boldsymbol{H})_{ij})^2 \tag{3}$$

where $(\boldsymbol{W}\boldsymbol{H})_{ij}$ denotes the (i,j)-th entry of the matrix $\boldsymbol{W}\boldsymbol{H} \in \mathbb{R}^{m\times n}$, and the one based on I-divergence:

$$D(\boldsymbol{W},\boldsymbol{H}) = \sum_{ij}\left(X_{ij}\ln\frac{X_{ij}}{(\boldsymbol{W}\boldsymbol{H})_{ij}} - X_{ij} + (\boldsymbol{W}\boldsymbol{H})_{ij}\right)$$

have been used [45]. All of those error functions are in general continuous in the interior of $\mathcal{F}_0$, which is denoted by $\operatorname{int}\mathcal{F}_0$ in this paper. However, some of them are not well-defined on the boundary of $\mathcal{F}_0$. Hence, when using such an error function, we allow it take the value $+\infty$, that is, we regard it as an extended real-valued function defined on $\mathcal{F}_0$.

2.2 Multiplicative Update Rules

Because (2) is not a convex optimization problem, it is difficult to find a global optimal solution. As an approach to find a local optimal solution, a class of iterative methods called multiplicative update rules [14, 28, 29, 45] are widely used. Given an error function $D(\boldsymbol{W},\boldsymbol{H})$, a multiplicative update rule is obtained in the following manner. First, an auxiliary function of the error function is constructed in some way. The definition of the auxiliary function is given below. Second, for each variable, a problem of finding a unique minimum point of the auxiliary function under some constraints is solved. If the minimum point is explicitly expressed in terms of the current values of $\boldsymbol{W}$ and $\boldsymbol{H}$, this expression leads to the multiplicative update rule.

Let us now give a formal definition of the auxiliary function.

Definition 1 (Auxiliary Function) Given an error function $D(\boldsymbol{W},\boldsymbol{H})$, any $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) : \operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0 \to \mathbb{R}$ that satisfies

$$\forall(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \in \operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0,\ \ \bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \geq D(\boldsymbol{W},\boldsymbol{H}) \tag{4}$$

and

$$\forall(\boldsymbol{W},\boldsymbol{H}) \in \operatorname{int}\mathcal{F}_0,\ \ \bar{D}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{W},\boldsymbol{H}) = D(\boldsymbol{W},\boldsymbol{H}) \tag{5}$$

is called an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$.

It is important to note that we use only a single auxiliary function in this paper, while two auxiliary functions have been used in the literature: one is for $\boldsymbol{W}$ and the other is for $\boldsymbol{H}$. The reason for doing this is to simplify the analysis. In fact, the update rule for the entries of $\boldsymbol{W}$ and the update rule for the entries of $\boldsymbol{H}$ are obtained from a common auxiliary function. When two auxiliary functions are given, one can immediately obtain a single auxiliary function by averaging them. If the obtained single auxiliary function satisfies Assumptions 1–3 given later, our method can be applied.

Let $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ be an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$. Let $\{(\boldsymbol{W}^{(l)},\boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ be any sequence such that i) $(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)}) \in \operatorname{int}\mathcal{F}_0$ and ii) for each $l \geq 0$,

$\boldsymbol{W}^{(l+1)}$ is obtained from $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$ as an optimal solution of the problem:

$$\begin{aligned} &\text{minimize} \quad \bar{D}(\boldsymbol{W}, \boldsymbol{H}^{(l)}, \boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \\ &\text{subject to } \boldsymbol{W} > \boldsymbol{O}_{m\times r}, \end{aligned} \tag{6}$$

and iii) for each $l \geq 0$, $\boldsymbol{H}^{(l+1)}$ is obtained from $\boldsymbol{W}^{(l+1)}$ and $\boldsymbol{H}^{(l)}$ as an optimal solution of the problem:

$$\begin{aligned} &\text{minimize} \quad \bar{D}(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}, \boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)}) \\ &\text{subject to } \boldsymbol{H} > \boldsymbol{O}_{r\times n}. \end{aligned} \tag{7}$$

Then the sequence $\{D(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ is monotone decreasing because

$$\begin{aligned} D(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)}) &\leq \bar{D}(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)}, \boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \\ &\leq \bar{D}(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)}, \boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \\ &\leq \bar{D}(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}, \boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \\ &= D(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \,. \end{aligned}$$

Here, the first inequality follows from (4), the second inequality follows from the fact that $\boldsymbol{H}^{(l+1)}$ is an optimal solution of (7), the third inequality follows from the fact that $\boldsymbol{W}^{(l+1)}$ is an optimal solution of (6), and the equality follows from (5).

For example, using Lemma 9 given in Appendix A, we can obtain an auxiliary function of (3) as follows:

$$\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \sum_{ij} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij} \sum_{k} \frac{W_{ik}^2 H_{kj}^2}{\widetilde{W}_{ik}\widetilde{H}_{kj}} - 2\sum_{ij} X_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij} + \sum_{ij} X_{ij}^2 \,. \tag{8}$$

If $(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \in \operatorname{int} \mathcal{F}_0$ then the problem (6) with (8) has the unique optimal solution, which is explicitly expressed by $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$. Similarly, if $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)}) \in \operatorname{int} \mathcal{F}_0$ then the problem (7) with (8) has the unique optimal solution, which is explicitly expressed by $\boldsymbol{W}^{(l+1)}$ and $\boldsymbol{H}^{(l)}$. From these expressions, a multiplicative update rule described by

$$W_{ik}^{(l+1)} = W_{ik}^{(l)} \frac{(\boldsymbol{X}(\boldsymbol{H}^{(l)})^{\mathrm{T}})_{ik}}{(\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)}(\boldsymbol{H}^{(l)})^{\mathrm{T}})_{ik}} \,, \tag{9}$$

$$H_{kj}^{(l+1)} = H_{kj}^{(l)} \frac{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}}\boldsymbol{X})_{kj}}{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}}\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)})_{kj}} \tag{10}$$

is obtained [28,29]. Note that the right-hand sides of (9) and (10) are positive because we have assumed that every row and column of $\boldsymbol{X}$ has at least one nonzero entry. Therefore, if $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \operatorname{int} \mathcal{F}_0$ then $(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) \in \operatorname{int} \mathcal{F}_0$ for all $l \geq 1$. Note also that the sequence $\{D(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=1}^{\infty}$ converges to some constant because it is monotone decreasing and 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). In addition, even if the sequence converges to some point, it is not guaranteed that the limit point belongs to $\operatorname{int} \mathcal{F}_0$.

The approach described above is not restricted to Euclidean distance but can be applied to various error functions. In fact, Yang and Oja [45] proposed a unified method to develop multiplicative update rules and applied it to eleven error functions. Details will be given in Section 4.

2.3 Modified Multiplicative Update Rules

The multiplicative update rule given by (9) and (10) and other multiplicative update rules [7, 11–13, 27, 45] have a common serious problem that they are not defined for all points in $\mathcal{F}_0$. For example, if all entries in the k-th row of $\boldsymbol{H}^{(l)}$ are zero then the denominator of the right-hand side of (9) becomes zero. One may think this is not a serious issue because if $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ belongs to $\operatorname{int} \mathcal{F}_0$ then so does $(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})$ for all $l \geq 0$, as mentioned above. However, it may occur that some of the entries of $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$ converges to zero as $l \to \infty$. In this case, some of the entries of $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$ may go to infinity with l. In order to avoid this situation, Gillis and Glineur [15] modified (9) and (10) as

$$W_{ik}^{(l+1)} = \max\left(\epsilon, W_{ik}^{(l)} \frac{(\boldsymbol{X}(\boldsymbol{H}^{(l)})^{\mathrm{T}})_{ik}}{(\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)}(\boldsymbol{H}^{(l)})^{\mathrm{T}})_{ik}}\right), \tag{11}$$

$$H_{kj}^{(l+1)} = \max\left(\epsilon, H_{kj}^{(l)} \frac{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}}\boldsymbol{X})_{kj}}{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}}\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l)})_{kj}}\right), \tag{12}$$

respectively, where ϵ is a user-specified positive constant. This simple modification is general and can be applied to all multiplicative update rules.

When a modified multiplicative update rule is used, it is natural to consider a modified optimization problem:

$$\begin{array}{l} \text{minimize} \quad D(\boldsymbol{W}, \boldsymbol{H}) \\ \text{subject to } \boldsymbol{W} \geq \epsilon \mathbf{1}_{m\times r},\ \boldsymbol{H} \geq \epsilon \mathbf{1}_{r\times n} \end{array} \tag{13}$$

instead of the original optimization problem (2), 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. Let the feasible region of the problem (13) be denoted by $\mathcal{F}_\epsilon$, that is,

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

If $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ belongs to $\mathcal{F}_\epsilon$ then the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by (11) and (12) is contained in $\mathcal{F}_\epsilon$. Moreover, if some entry of $\boldsymbol{W}^{(l)}$ and $\boldsymbol{H}^{(l)}$ converges as $l \to \infty$, its limit point belongs to $[\epsilon, \infty)$ because this is a closed set.

Although the original multiplicative update rule given by (9) and (10) often produces sparse matrices, the modified update rule given by (11) and (12) always produces dense matrices because ϵ is positive. However, if we replace all ϵ in the factor matrices produced by the modified update rule with zero, the resulting matrices are expected to be sparse because local optimal solutions of (2) or (13) are often located at the boundary of the feasible region.

3 Global Convergence of Modified Multiplicative Update Rules

3.1 Problem Setting and Main Result

In this section, we consider a general error function $D(\boldsymbol{W}, \boldsymbol{H})$, which is defined on $\mathcal{F}_0$ as an extended real-valued function and satisfies the following properties: i) it is continuously differentiable on $\operatorname{int}\mathcal{F}_0$ (see Assumption 1 given below), ii) $D(\boldsymbol{W}, \boldsymbol{H}) \geq 0$ for all $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}_0$, and iii) $D(\boldsymbol{W}, \boldsymbol{H}) = 0$ if $\boldsymbol{W}\boldsymbol{H} = \boldsymbol{X}$. We then give a sufficient condition on $D(\boldsymbol{W}, \boldsymbol{H})$ and the auxiliary function $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ for the modified update rule to have the global convergence property in the sense of Zangwill [46]. For the update rule we are considering, it is defined as follows.

Definition 2 (Global Convergence) Let ϵ be a user-specified positive constant. An update rule is said to have the global convergence property if, for any initial point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}_\epsilon$, the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by the update rule has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the optimization problem (13).

A stationary point of (13) is a point $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \mathcal{F}_\epsilon$ such that the following conditions are satisfied (see, *e.g.*, [36]):

$$\nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \geq \boldsymbol{O}_{m\times r}, \tag{14}$$

$$\nabla_{\boldsymbol{H}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \geq \boldsymbol{O}_{r\times n}, \tag{15}$$

$$\nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \odot (\epsilon \boldsymbol{1}_{m\times r} - \hat{\boldsymbol{W}}) = \boldsymbol{O}_{m\times r}, \tag{16}$$

$$\nabla_{\boldsymbol{H}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \odot (\epsilon \boldsymbol{1}_{r\times n} - \hat{\boldsymbol{H}}) = \boldsymbol{O}_{r\times n}, \tag{17}$$

where $\nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ ($\nabla_{\boldsymbol{H}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$, resp.) is the $m \times r$ ($r \times n$, resp.) matrix of which the (i,k)-th ((k,j)-th, resp.) entry is the value of $\partial D/\partial W_{ik}$ ($\partial D/\partial H_{kj}$, resp.) at $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $\odot$ denotes the componentwise multiplication of two matrices of the same dimension. In this paper, the set of stationary points of (13) is denoted by $\mathcal{S}_\epsilon$.

We now give three assumptions about the error function and the auxiliary function.

Assumption 1 $D(\boldsymbol{W}, \boldsymbol{H})$ and $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ are continuously differentiable on $\operatorname{int}\mathcal{F}_0$ and $\operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0$, respectively.

Assumption 2 For any $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int}\mathcal{F}_0$, the following equalities hold:

$$\nabla_{\boldsymbol{W}} \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}),$$
$$\nabla_{\boldsymbol{H}} \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{H}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$$

where $\nabla_{\boldsymbol{W}} \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ ($\nabla_{\boldsymbol{H}} \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$, resp.) is the $m \times r$ ($r \times n$, resp.) matrix of which (i,k)-th ((k,j)-th, resp.) entry is the value of $\partial \bar{D}/\partial W_{ik}$ ($\partial \bar{D}/\partial H_{kj}$, resp.) at $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$.

Assumption 3 $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ satisfies the following conditions.

1. $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is separable in the sense that it is expressed as
$$\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \sum_{ijk} \bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) + \bar{D}^2(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) + \text{constant}\,. \tag{18}$$
2. For each $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$,
$$u_{ik}(W_{ik}) \triangleq \sum_j \bar{D}^1_{ijk}(W_{ik}, \hat{H}_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$$
is strictly convex in $\mathbb{R}_{++}$ and minimized at $W^*_{ik} \in \mathbb{R}_{++}$ which is explicitly expressed as
$$W^*_{ik} = f_{ik}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\,,$$
where $\mathbb{R}_{++}$ is the set of positive numbers. Also, for each $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$,
$$v_{kj}(H_{kj}) \triangleq \sum_i \bar{D}^1_{ijk}(\hat{W}_{ik}, H_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$$
is strictly convex in $\mathbb{R}_{++}$ and minimized at $H^*_{kj} \in \mathbb{R}_{++}$ which is explicitly expressed as
$$H^*_{kj} = g_{kj}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\,.$$
3. $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ is continuous in $\operatorname{int} \mathcal{F}_0$ and, for each $\epsilon > 0$, there exist $c_{ik} > 0$ and $\phi_{ik} < 1$ such that
$$\forall (\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}_\epsilon, \quad f_{ik}(\boldsymbol{W}, \boldsymbol{H}) \le c_{ik} W_{ik}^{\phi_{ik}}\,. \tag{19}$$
Also, $g_{kj}(\boldsymbol{W}, \boldsymbol{H})$ is continuous in $\operatorname{int} \mathcal{F}_0$ and, for each $\epsilon > 0$, there exist $d_{kj} > 0$ and $\psi_{kj} < 1$ such that
$$\forall (\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}_\epsilon, \quad g_{kj}(\boldsymbol{W}, \boldsymbol{H}) \le d_{kj} H_{kj}^{\psi_{kj}}\,. \tag{20}$$

Under Assumptions 1–3, we can prove the global convergence of the modified update rule, as stated in the following theorem. The proof is given in the next subsection.

Theorem 1 Suppose that an error function $D(\boldsymbol{W}, \boldsymbol{H})$ and an auxiliary function $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ of it satisfy Assumptions 1–3. Then, for any positive constant ϵ, the modified multiplicative update rule described by
$$W^{(l+1)}_{ik} = \max(\epsilon, f_{ik}(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}))\,, \tag{21}$$
$$H^{(l+1)}_{kj} = \max(\epsilon, g_{kj}(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l)})) \tag{22}$$
has the global convergence property, where f_{ik} and g_{kj} are defined in Assumption 3.

If a modified update rule has the global convergence property, it is easy to obtain an algorithm that terminates within a finite number of iterations after finding an approximate solution. Let us relax the conditions (14)–(17) as follows:

$$\nabla_{\boldsymbol{W}} D(\boldsymbol{W}, \boldsymbol{H}) \geq -\delta_1 \mathbf{1}_{m\times r}, \tag{23}$$

$$\nabla_{\boldsymbol{H}} D(\boldsymbol{W}, \boldsymbol{H}) \geq -\delta_1 \mathbf{1}_{r\times n}, \tag{24}$$

$$(\nabla_{\boldsymbol{W}} D(\boldsymbol{W}, \boldsymbol{H}))_{ik} > \delta_1 \Rightarrow W_{ik} - \epsilon \leq \delta_2\,, \tag{25}$$

$$(\nabla_{\boldsymbol{H}} D(\boldsymbol{W}, \boldsymbol{H}))_{kj} > \delta_1 \Rightarrow H_{kj} - \epsilon \leq \delta_2\,, \tag{26}$$

where δ_1 and δ_2 are positive constants specified by the user. Employing these conditions as a stopping criterion, we obtain the following algorithm.

Algorithm 1 Modified Multiplicative Update Algorithm

Input: $\boldsymbol{X} \in \mathbb{R}_+^{m\times n}$, $r \in \mathbb{N}$, $\epsilon > 0$, $\delta_1 > 0$, $\delta_2 > 0$
Output: A point $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}_\epsilon$ that satisfies (23)–(26)
1: Choose $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}_\epsilon$ and set $l = 0$.
2: Find $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)})$ by the update rule described by (21) and (22).
3: If $(\boldsymbol{W}, \boldsymbol{H}) = (\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)}) \in \mathcal{F}_\epsilon$ satisfies (23)–(26) then return $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)})$. Otherwise add 1 to l and go to Step 2.

The finite termination of this algorithm is guaranteed by the following theorem. We omit the proof because it is the same as that of Theorem 2 in Reference [36].

Theorem 2 Suppose that an error function $D(\boldsymbol{W}, \boldsymbol{H})$ and its auxiliary function $\bar{D}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ satisfy Assumptions 1–3. Then, for any positive constants ϵ, δ_1 and δ_2, Algorithm 1 stops within a finite number of iterations.

3.2 Relation to Existing Work

There are many results on the convergence properties of NMF algorithms in the literature [6, 15, 17, 21, 23, 25, 27, 29, 31]. However, most of them are insufficient to prove that an approximate stationary point is always obtained by the algorithm. As mentioned before, the monotone decrease of the error value does not imply the convergence of the sequence of solutions. It is claimed in some papers [6, 15, 21, 23, 25, 31] that every limit point of the sequence of solutions is a stationary point. However, this claim says nothing about the convergence of the sequence or even the existence of a convergent subsequence.

The convergence of the sequence of solutions generated by the multiplicative update rule was proved by some authors in different settings [14, 30, 47]. Finesso and Spreij [14] considered a variant of the I-divergence based multiplicative update rule, and proved the convergence of the right factor matrix and the product of the two factor matrices. Lin [30] proved the convergence

of a variant of the Euclidean distance based multiplicative update rule. A common key feature of these two algorithms is that one of the two factor matrices is normalized after each update. This guarantees the boundedness of the corresponding factor matrix. Zhao and Tan [47] recently performed a unified convergence analysis of the multiplicative update rules for NMF with ℓ_1 regularization. They proved the boundedness of the sequence of solutions by showing that the level set of the objective function is bounded. However, this result relies on the existence of ℓ_1 regularization terms, and cannot be directly applied to NMF without regularizations.

The unified global convergence analysis in the present paper is a generalization of the results given by Takahashi and Hibi [36]. They showed that the modification of Gillis and Glineur [15] is sufficient to guarantee the global convergence of the multiplicative update rules based on Euclidean distance and I-divergence. However, because these update rules were studied separately, it was not clear what kind of conditions are required in general for the error function and the auxiliary function in order for the obtained multiplicative update rule to have the global convergence property.

Finally, it is important to make a comment on the stopping condition. Some authors [17, 31] used a stopping condition similar to the one in Algorithm 1. Roughly speaking, their condition corresponds to (23)–(26) with $\delta_2 = 0$. In this case, the finite termination of the algorithm is not guaranteed.

3.3 Proof of Theorem 1

We prove Theorem 1 by Zangwill's global convergence theorem [46], which has been used extensively to prove the convergence of various algorithms (see [38, 42] for example).

Theorem 3 (Zangwill [46]) Let A be a point-to-set mapping defined on a space V that assigns to every point $x \in V$ a subset of V. Let $\{x^{(l)}\}_{l=0}^{\infty}$ be a sequence generated satisfying $x^{(0)} \in V$ and $x^{(l+1)} \in A(x^{(l)})$. Also let a solution set $\Omega \subset V$ be given. Suppose that the following three statements hold true.

1. All points $x^{(l)}$ are in a compact set $X \subset V$.
2. There is a continuous function $Z : V \to \mathbb{R}$ such that i) if x is not a solution then $Z(y) < Z(x)$ for any $y \in A(x)$, and ii) if x is a solution then $Z(y) \leq Z(x)$ for any $y \in A(x)$.
3. The mapping A is closed at points outside Ω.

Then the limit of any convergent subsequence of $\{x^{(l)}\}_{l=0}^{\infty}$ is a solution.

In the following discussion, we express for simplicity (21) and (22) 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. Furthermore, we express the update process from $(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})$ to $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)})$ as $(\boldsymbol{W}^{(l+1)}, \boldsymbol{H}^{(l+1)}) = A(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})$. Then the mapping A is expressed in terms of A_1 and A_2 as follows:

$$A(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}) = (A_1(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}), A_2(A_1(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)}), \boldsymbol{H}^{(l)})) .$$

Note that A is not a point-to-set mapping but a point-to-point mapping defined on $\mathcal{F}_\epsilon$. In this case, the closedness of A in Theorem 3 reduces to the continuity of A. Therefore, it suffices to show that the following statements hold true.

1. (Boundedness) For any initial point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}_\epsilon$, the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by the update rule A belongs to a compact subset of $\mathcal{F}_\epsilon$.
2. (Monotoneness) There exists a function $Z : \mathcal{F}_\epsilon \to \mathbb{R}$ such that

$$(\boldsymbol{W}, \boldsymbol{H}) \notin \mathcal{S}_\epsilon \Rightarrow Z(A(\boldsymbol{W}, \boldsymbol{H})) < Z(\boldsymbol{W}, \boldsymbol{H}),$$
$$(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{S}_\epsilon \Rightarrow Z(A(\boldsymbol{W}, \boldsymbol{H})) \leq Z(\boldsymbol{W}, \boldsymbol{H}),$$

   where $\mathcal{S}_\epsilon$ is the set of stationary points of (13).
3. (Continuity) A is continuous in $\mathcal{F}_\epsilon \setminus \mathcal{S}_\epsilon$.

We first consider the first statement. Our proof is based on the following lemma.

Lemma 1 (Katayama et al. [22]) Let ϵ be any positive constant. Let f be a mapping from $[\epsilon, \infty)$ to $\mathbb{R}$. If there exist a positive constant c and a constant ϕ less than 1 such that

$$\forall x \geq \epsilon, \quad f(x) \leq cx^{\phi}$$

then any sequence $\{x^{(l)}\}_{l=0}^{\infty}$ generated by the update rule:

$$x^{(l+1)} = \max(\epsilon, f(x^{(l)})), \quad l = 0, 1, 2, \ldots$$

with the initial value $x^{(0)} \geq \epsilon$ is contained in a closed and bounded set.

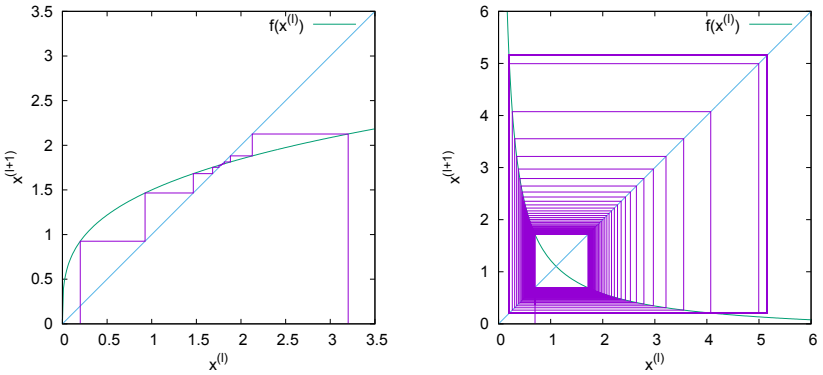
Some illustrative examples for Lemma 1 are shown in Fig. 1. In Fig. 1(a), two sequences generated by $x^{(l+1)} = \max(\epsilon, f(x^{(l)}))$ with $\epsilon = 0.1$ and $f(x) = 1.5x^{0.3}$ are plotted. The sequence starting from $x^{(0)} = 0.2$ increases monotonically and converges to the unique fixed point. The sequence starting from $x^{(0)} = 3.2$ decreases monotonically and converges to the fixed point. In Fig. 1(b), the sequence generated by $x^{(l+1)} = \max(\epsilon, f(x^{(l)}))$ with $x^{(0)} = 0.7$, $\epsilon = 0.2$ and $f(x) = 1.5x^{-0.8} - 0.28$ is plotted. We see from the figure that it converges to periodic sequence with period 2.

In view of Assumption 3 and Lemma 1, we immediately obtain the following lemma.

Lemma 2 Let ϵ be any positive constant. If Assumption 3 holds then, for any initial point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}_\epsilon$, the sequence $\{(\boldsymbol{W}^{(l)}, \boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ generated by (21) and (22) is contained in a closed and bounded set.

We next consider the third statement. The continuity of the mapping A is shown as follows.

Lemma 3 Let ϵ be any positive constant. If Assumption 3 holds then the mapping $A : \mathcal{F}_\epsilon \to \mathcal{F}_\epsilon$ is continuous.



(a) (b)

Fig. 1 Illustrative examples for Lemma 1. (a) Two sequences generated by $x^{(l+1)} = \max(\epsilon, f(x^{(l)}))$ with $\epsilon = 0.1$ and $f(x) = 1.5x^{0.3}$. (b) A sequence generated by $x^{(l+1)} = \max(\epsilon, f(x^{(l)}))$ with $\epsilon = 0.2$ and $f(x) = 1.5x^{-0.8} - 0.28$.

Proof By Assumption 3, f_{ik} and g_{kj} are continuous. Also, for any positive constant ϵ, the mapping $\max(\epsilon, \cdot)$ is continuous. So (21), which is the composition of f_{ik} and $\max(\epsilon, \cdot)$, and (22), which is the composition of g_{kj} and $\max(\epsilon, \cdot)$, are both continuous. Furthermore, because the mapping A is the composition of (21) and (22), it is continuous. □

We finally consider the second statement. As the first step, we discuss the relationship between the stationary points of (13) and the mapping A in the following two lemmas.

Lemma 4 Let ϵ be any positive constant. Let $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ be any point in $\mathcal{F}_\epsilon$. If Assumption 3 holds then the optimization problem:

$$\begin{aligned} &\text{minimize} \quad \bar{D}(\boldsymbol{W}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &\text{subject to } \boldsymbol{W} \geq \epsilon \mathbf{1}_{m \times r} \end{aligned} \tag{27}$$

has a unique optimal solution which is given by $A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. Similarly, if Assumption 3 hold then the optimization problem:

$$\begin{aligned} &\text{minimize} \quad \bar{D}(\hat{\boldsymbol{W}}, \boldsymbol{H}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &\text{subject to } \boldsymbol{H} \geq \epsilon \mathbf{1}_{r \times n} \end{aligned} \tag{28}$$

has a unique optimal solution which is given by $A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$.

Proof We prove only the first part because the second one can be proved in the same way. Because $\bar{D}(\boldsymbol{W}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ is separable by Assumption 3, the problem (27) can be divided into mr independent problems of the form:

$$\begin{aligned} &\text{minimize} \quad \textstyle\sum_j \bar{D}^1_{ijk}(W_{ik}, \hat{H}_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &\text{subject to } W_{ik} \geq \epsilon\,. \end{aligned} \tag{29}$$

It follows from Assumption 3 that $u_{ik}(W_{ik}) \triangleq \sum_j \bar{D}^1_{ijk}(W_{ik}, \hat{H}_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ has a unique minimum point given by $f_{ik}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \mathbb{R}_{++}$. If $f_{ik}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \geq \epsilon$ then it is apparently the unique optimal solution of (29). Otherwise, ϵ is the unique optimal solution of (29) because, by Assumption 3, $u_{ik}(W_{ik})$ is strictly convex in $\mathbb{R}_{++}$. Therefore, the unique optimal solution of (29) is given by $\max(\epsilon, f_{ik}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) = (A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}))_{ik}$. □

Lemma 5 Under Assumptions 1–3, the necessary and sufficient condition for $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \mathcal{F}_\epsilon$ to be a stationary point of (13) is that $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ are the unique optimal solutions of (27) and (28), respectively.

Proof As shown in the proof of Lemma 4, the necessary and sufficient condition for $\hat{\boldsymbol{W}} \geq \epsilon \mathbf{1}_{m\times r}$ to be the unique optimal solution of (27) is that

$$\forall i,k, \quad u'_{ik}(\hat{W}_{ik}) \begin{cases} = 0, \text{ if } \hat{W}_{ik} > \epsilon\,, \\ \geq 0, \text{ if } \hat{W}_{ik} = \epsilon\,. \end{cases} \tag{30}$$

By Assumptions 2 and 3, we have

$$\begin{aligned} u'_{ik}(\hat{W}_{ik}) &= \left.\frac{\partial \bar{D}}{\partial W_{ik}}\right|_{(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})} \\ &= \left(\nabla_{\boldsymbol{W}} \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\right)_{ik} \\ &= \left(\nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\right)_{ik}\,. \end{aligned}$$

Therefore, (30) is rewritten as

$$\forall i,k, \quad \left(\nabla_{\boldsymbol{W}} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\right)_{ik} \begin{cases} = 0, \text{ if } \hat{W}_{ik} > \epsilon\,, \\ \geq 0, \text{ if } \hat{W}_{ik} = \epsilon\,, \end{cases}$$

which is equivalent to the conjunction of (14) and (16). We can show in the same way as above that the necessary and sufficient condition for $\hat{\boldsymbol{H}} \geq \epsilon \mathbf{1}_{r\times n}$ to be the unique optimal solution of (28) is equivalent to the conjunction of (15) and (17). □

Using Lemmas 4 and 5, we can prove the monotoneness of the error function as follows.

Lemma 6 Let ϵ be any positive constant. Under Assumptions 1–3, the following propositions hold true:

$$(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \notin \mathcal{S}_\epsilon \Rightarrow D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) < D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\,, \tag{31}$$

$$(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \mathcal{S}_\epsilon \Rightarrow D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) = D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})\,. \tag{32}$$

Proof We first prove (32). Let $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ be any point in $\mathcal{S}_\epsilon$. Then, it follows from Lemma 5 that $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ are the unique optimal solutions of (27) and (28), respectively. Furthermore, it follows from Lemma 4 that $\hat{\boldsymbol{W}} = A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $\hat{\boldsymbol{H}} = A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. Therefore, we have

$$\begin{aligned} A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) &= (A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}})) \\ &= (A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) \\ &= (\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \end{aligned}$$

which implies that $D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) = D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. We next prove (31). Note that if $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \notin \mathcal{S}_\epsilon$ then, by Lemma 5, at least one of the following two statements holds: i) $\hat{\boldsymbol{W}}$ is not the unique optimal solution of (27), that is, $\hat{\boldsymbol{W}} \neq A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and ii) $\hat{\boldsymbol{H}}$ is not the unique optimal solution of (28), that is, $\hat{\boldsymbol{H}} \neq A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. If the first statement holds, we have

$$\begin{aligned} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) &= \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &> \bar{D}(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &= D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) - D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) \\ &\quad + \bar{D}(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &\geq D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) \\ &= \bar{D}(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}, A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) \\ &\geq \bar{D}(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}), A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) \\ &= D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}})) \\ &\quad - D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}})) \\ &\quad + \bar{D}(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}), A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}}) \\ &\geq D(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), A_2(A_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{H}})) \\ &= D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}))\,. \end{aligned}$$

If the first statement does not hold but the second one does, we have

$$\begin{aligned} D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) &= \bar{D}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &> \bar{D}(\hat{\boldsymbol{W}}, A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &= D(\hat{\boldsymbol{W}}, A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) - D(\hat{\boldsymbol{W}}, A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) \\ &\quad + \bar{D}(\hat{\boldsymbol{W}}, A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}), \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \\ &\geq D(\hat{\boldsymbol{W}}, A_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) \\ &= D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}))\,. \end{aligned}$$

Therefore, the inequality $D(A(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})) < D(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ holds in both cases. □

Table 1 Error functions considered by Yang and Oja [45].

Name	$D(\boldsymbol{W},\boldsymbol{H})$
Euclidean distance	$\sum_{ij}(X_{ij}-(\boldsymbol{WH})_{ij})^2$
I-divergence	$\sum_{ij}\left(X_{ij}\ln\frac{X_{ij}}{(\boldsymbol{WH})_{ij}}-X_{ij}+(\boldsymbol{WH})_{ij}\right)$
Dual I-divergence	$\sum_{ij}\left((\boldsymbol{WH})_{ij}\ln\frac{(\boldsymbol{WH})_{ij}}{X_{ij}}-(\boldsymbol{WH})_{ij}+X_{ij}\right)$
Itakura-Saito divergence	$\sum_{ij}\left(-\ln\left(\frac{X_{ij}}{(\boldsymbol{WH})_{ij}}\right)+\frac{X_{ij}}{(\boldsymbol{WH})_{ij}}-1\right)$
α-divergence	$\frac{1}{\alpha(1-\alpha)}\sum_{ij}\left(\alpha X_{ij}+(1-\alpha)(\boldsymbol{WH})_{ij}-X_{ij}^{\alpha}(\boldsymbol{WH})_{ij}^{1-\alpha}\right)\ (\alpha\neq 0,1)$
β-divergence	$\sum_{ij}\left(X_{ij}\frac{X_{ij}^{\beta}-(\boldsymbol{WH})_{ij}^{\beta}}{\beta}-\frac{X_{ij}^{\beta+1}-(\boldsymbol{WH})_{ij}^{\beta+1}}{\beta+1}\right)\ (\beta\neq 0,-1)$
Log-Quad cost	$\sum_{ij}\left((X_{ij}-(\boldsymbol{WH})_{ij})^2+X_{ij}\ln\frac{X_{ij}}{(\boldsymbol{WH})_{ij}}-X_{ij}+(\boldsymbol{WH})_{ij}\right)$
$\alpha\beta$-Bregman divergence	$\sum_{ij}\Big(X_{ij}^{\alpha}-X_{ij}^{\beta}-(\boldsymbol{WH})_{ij}^{\alpha}+(\boldsymbol{WH})_{ij}^{\beta}$ $-(\alpha(\boldsymbol{WH})_{ij}^{\alpha-1}-\beta(\boldsymbol{WH})_{ij}^{\beta-1})(X_{ij}-(\boldsymbol{WH})_{ij})\Big)\ (\alpha\geq 1,\ 0<\beta<1)$
Kullback-Leibler divergence	$\sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}\ln\left(\frac{X_{ij}/\sum_{pq}X_{pq}}{(\boldsymbol{WH})_{ij}/\sum_{pq}(\boldsymbol{WH})_{pq}}\right)$
γ-divergence	$\frac{1}{\gamma(1+\gamma)}\Big(\ln\left(\sum_{ij}X_{ij}^{1+\gamma}\right)+\gamma\ln\left(\sum_{ij}(\boldsymbol{WH})_{ij}^{1+\gamma}\right)$ $-(1+\gamma)\ln\left(\sum_{ij}X_{ij}(\boldsymbol{WH})_{ij}^{\gamma}\right)\Big)\ (\gamma\neq 0,-1)$
Rényi divergence	$\frac{1}{\rho-1}\ln\left(\sum_{ij}\left(\frac{X_{ij}}{\sum_{pq}X_{pq}}\right)^{\rho}\left(\frac{(\boldsymbol{WH})_{ij}}{\sum_{pq}(\boldsymbol{WH})_{pq}}\right)^{1-\rho}\right)\ (\rho>0,\ \rho\neq 1)$

4 Application of Theorem 1 to Multiplicative Update Rules Derived by the Unified Method of Yang and Oja

Theorem 1 can be applied to various multiplicative update rules to prove their global convergence. In this section, we apply it to the eleven multiplicative update rules derived from the error functions shown in Table 1[1] by the unified method proposed by Yang and Oja [45]. To make the analysis simpler, we assume for the moment that $\boldsymbol{X}$ is a positive matrix. Then all error functions in Table 1 are well-defined on $\mathcal{F}_0$ as extended real-valued functions. For each of 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), we show later that the auxiliary function derived by the method of Yang and Oja satisfies all conditions in Assumptions 1–3. This means that the modified update rules obtained from these error functions have the global convergence property. In contrast, for the last three error functions, the global convergence of the modified update rule cannot be proved by Theorem 1 because the inequalities (19) and (20) are not satisfied. This issue is discussed in the next section.

[1] The error function based on Kullback-Leibler divergence in Table 1 is slightly different from the one in [45]. Instead of assuming that $\sum_{ij}X_{ij}=1$, X_{ij} has been replaced with $X_{ij}/\sum_{pq}X_{pq}$ so that the result can be applied to a general nonnegative matrix $\boldsymbol{X}$.

Let $D(\boldsymbol{W}, \boldsymbol{H})$ be a given error function. The procedure of the unified method of Yang and Oja to derive a multiplicative update rule from $D(\boldsymbol{W}, \boldsymbol{H})$ is summarized as follows.

1. If $D(\boldsymbol{W}, \boldsymbol{H})$ contains one or more logarithmic functions then we replace each of them with a generalized polynomial by using the following relationship:
$$\ln x = \lim_{\mu \to 0^+} \frac{x^\mu - 1}{\mu}. \tag{33}$$
2. Applying Lemmas 8–10 given in Appendix A and then taking the limit $\mu \to 0^+$ if necessary, we can derive an auxiliary function of $D(\boldsymbol{W}, \boldsymbol{H})$. Note that we may need to apply L'Hôpital's rule when taking the limit. The derived auxiliary function is expressed in the form of (18) and, for any $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$, both $u_{ik}(W_{ik}) \triangleq \sum_j \bar{D}^1_{ijk}(W_{ik}, \hat{H}_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $v_{kj}(H_{kj}) \triangleq \sum_i \bar{D}^1_{ijk}(\hat{W}_{ik}, H_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ are strictly convex in $\mathbb{R}_{++}$.
3. Let $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$ be the current solution. Solving the equations $u'_{ik}(W_{ik}) = 0$ and $v'_{kj}(H_{kj}) = 0$, we obtain multiplicative update formulae for W_{ik} and H_{kj}, respectively.

Carrying out this procedure for each of the eleven error functions shown in Table 1, we obtain eleven auxiliary functions in the form of (18). Table 2 shows the formula for the first term $\bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ of the auxiliary function for each error function. It is easily seen that every auxiliary function shown in Table 2 satisfies Assumptions 1, Assumption 2, and the first and second conditions of Assumption 3. It is also seen that every $\bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ is expressed as $a_{1ijk}(W_{ik}H_{kj})^{b_1} + a_{2ijk}(W_{ik}H_{kj})^{b_2}$ where $b_1 \neq b_2$ and a_{1ijk} and a_{2ijk} are independent of W_{ik} and H_{kj}. Therefore, letting $u_{ik}(W_{ik}) \triangleq \sum_j \bar{D}^1_{ijk}(W_{ik}, \hat{H}_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ where $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$ and solving the equation $u'_{ik}(W_{ik}) = 0$, we obtain $W_{ik} = f_{ik}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. The formulae for $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ are shown in Reference [45][2]. Similarly, letting $v_{kj}(H_{kj}) \triangleq \sum_i \bar{D}^1_{ijk}(\hat{W}_{ik}, H_{kj}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and solving the equation $v'_{kj}(H_{kj}) = 0$, we obtain $H_{kj} = g_{kj}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$. Finally, by simple algebraic manipulations, we can prove that $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ and $g_{kj}(\boldsymbol{W}, \boldsymbol{H})$ are upper bounded for the first eight error functions. Upper bounds for $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ are explicitly given in Reference [22].

Let us next consider the case where $\boldsymbol{X}$ has a zero entry. For Dual I-divergence and α-divergence with a negative α, there exists a pair (i, k) such that both $\sum_j \bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ and $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ are not defined. Therefore, these error functions need a stronger assumption that $\boldsymbol{X}$ is positive. For Itakura-Saito divergence, β-divergence with β less than -1, γ-divergence with γ less than -1, $D(\boldsymbol{W}, \boldsymbol{H})$ contains a constant term which takes $+\infty$. However, if we construct a new error function from the remaining terms, it is well-defined on $\mathcal{F}_0$ as an extended real-valued function. Also, both $\sum_j \bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$

[2] In the case of Kullback-Leibler divergence, X_{ij} in $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ must be replaced with $X_{ij} / \sum_{pq} X_{pq}$.

and $f_{ik}(\boldsymbol{W},\boldsymbol{H})$ are well-defined on $\mathbb{R}_{++}\times\mathbb{R}_{++}\times\operatorname{int}\mathcal{F}_0$ and $\operatorname{int}\mathcal{F}_0$, respectively, for all (i,k). As for other functions, $D(\boldsymbol{W},\boldsymbol{H})$ is well-defined on $\mathcal{F}_0$ even though some entries of $\boldsymbol{X}$ are zero. In addition, $\sum_j \bar{D}^1_{ijk}(W_{ik},H_{kj},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ and $f_{ik}(\boldsymbol{W},\boldsymbol{H})$ are well-defined on $\mathbb{R}_{++}\times\mathbb{R}_{++}\times\operatorname{int}\mathcal{F}_0$ and $\operatorname{int}\mathcal{F}_0$, respectively, for all (i,k).

From these observations, we obtain the following lemma.

Lemma 7 For Euclidean distance, I-divergence, Itakura-Saito divergence, α-divergence with a positive α, β-divergence, Log-Quad cost and $\alpha\beta$-Bregman divergence, the auxiliary function shown in Table 2 satisfies all conditions in Assumptions 1–3. For Dual I-divergence and α-divergence with a negative α, the auxiliary function satisfies all conditions in Assumptions 1–3 if $\boldsymbol{X}$ is positive.

By Lemma 7 and Theorem 1, we immediately obtain the following theorem.

Theorem 4 For Euclidean distance, I-divergence, Itakura-Saito divergence, α-divergence with a positive α, β-divergence, Log-Quad cost and $\alpha\beta$-Bregman divergence, the modified multiplicative update rule obtained by the unified method of Yang and Oja has the global convergence property for any positive constant ϵ. For Dual I-divergence and α-divergence with a negative α, the modified multiplicative update rule has the global convergence property for any positive constant ϵ if $\boldsymbol{X}$ is positive.

5 New Multiplicative Update Rules for Kullback-Leibler, Gamma and Rényi Divergences

In the previous section, we proved the global convergence of the modified multiplicative update rule described by (21) and (22) for the first eight error functions shown in Table 1. As for the last three (Kullback-Leibler divergence, γ-divergence and Rényi divergence), the multiplicative update rule can be derived, but the boundedness of the sequence $\{(\boldsymbol{W}^{(l)},\boldsymbol{H}^{(l)})\}_{l=0}^{\infty}$ cannot be proved. A possible reason is that the value of the error function does not change even though $\boldsymbol{W}$ and $\boldsymbol{H}$ are multiplied by any positive scalar. In other words, we can increase the values of nonzero entries of $\boldsymbol{W}$ and $\boldsymbol{H}$ as much as we want, while keeping the value of the error function fixed. As a simple way to avoid this situation, we add a regularization term

$$\frac{C}{2}\left(\sum_{ij}X_{ij}-\sum_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}\right)^2 \tag{34}$$

to each error function, where C is any positive constant. Applying Lemmas 8–10, we obtain new auxiliary functions for Kullback-Leibler divergence, γ-divergence and Rényi divergence. How to derive the auxiliary function for Kullback-Leibler divergence is explained in Appendix B. Although other auxiliary function is omitted due to space constraints, they are obtained in a

Table 2 Auxiliary functions obtained from the error functions in Table 1 by the method of Yang and Oja.

Error function	$\bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$
Euclidean distance	$-2X_{ij}W_{ik}H_{kj} + (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1}(W_{ik}H_{kj})^2$
I-divergence	$-X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1}\widetilde{W}_{ik}\widetilde{H}_{kj}\ln(W_{ik}H_{kj}) + W_{ik}H_{kj}$
Dual I-divergence	$\left(\ln(W_{ik}H_{kj}) + \ln\left(X_{ij}^{-1}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1}\right) - 1\right) W_{ik}H_{kj}$
Itakura-Saito divergence	$(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1}W_{ik}H_{kj} + X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-2}(\widetilde{W}_{ik}\widetilde{H}_{kj})^2(W_{ik}H_{kj})^{-1}$
α-divergence	$\frac{1}{\alpha}W_{ik}H_{kj} - \frac{1}{\alpha(1-\alpha)}X_{ij}^{\alpha}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-\alpha}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{\alpha}(W_{ik}H_{kj})^{1-\alpha} \quad (\alpha \neq 0, 1)$
β-divergence	
1) $\beta > 1$	$-X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta-1}W_{ik}H_{kj} + \frac{1}{\beta+1}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-\beta}(W_{ik}H_{kj})^{\beta+1}$
2) $0 < \beta \leq 1$	$-\frac{1}{\beta}X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta-1}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-\beta}(W_{ik}H_{kj})^{\beta}$ $+\frac{1}{\beta+1}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-\beta}(W_{ik}H_{kj})^{\beta+1}$
3) $\beta < 0$, $\beta \neq -1$	$-\frac{1}{\beta}X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta-1}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-\beta}(W_{ik}H_{kj})^{\beta} + (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta}W_{ik}H_{kj}$
Log-Quad cost	$\left((\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij} + \frac{1}{2}\right)(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1}(W_{ik}H_{kj})^2$ $-X_{ij}\left((\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1} + 2\right)\widetilde{W}_{ik}\widetilde{H}_{kj}\ln(W_{ik}H_{kj})$
$\alpha\beta$-Bregman divergence	$\left((\alpha-1)(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\alpha-1} + \frac{\beta(1-\beta)}{\alpha}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta-1}\right)(\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-\alpha}(W_{ik}H_{kj})^{\alpha}$ $+\left(\frac{\alpha(\alpha-1)}{1-\beta}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\alpha-2} + \beta(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\beta-2}\right)X_{ij}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{2-\beta}(W_{ik}H_{kj})^{\beta-1}$ $(\alpha \geq 1,\ 0 < \beta < 1)$
Kullback-Leibler divergence	$(\sum_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq})^{-1}W_{ik}H_{kj} - (\sum_{pq}X_{pq})^{-1}X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1}\widetilde{W}_{ik}\widetilde{H}_{kj}\ln(W_{ik}H_{kj})$
γ-divergence	
1) $\gamma > 0$	$\frac{1}{1+\gamma}\left(\sum_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^{1+\gamma}\right)^{-1}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\gamma}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-\gamma}(W_{ik}H_{kj})^{1+\gamma}$ $-\left(\sum_{pq}X_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^{\gamma}\right)^{-1}X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\gamma-1}\widetilde{W}_{ik}\widetilde{H}_{kj}\ln(W_{ik}H_{kj})$
2) $\gamma < 0$, $\gamma \neq -1$	$\left(\sum_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^{1+\gamma}\right)^{-1}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\gamma}W_{ik}H_{kj}$ $-\frac{1}{\gamma}\left(\sum_{pq}X_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^{\gamma}\right)^{-1}X_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\gamma-1}\left(\widetilde{W}_{ik}\widetilde{H}_{kj}\right)^{1-\gamma}(W_{ik}H_{kj})^{\gamma}$
Rényi divergence	
1) $\rho > 1$	$-\frac{1}{1-\rho}\left(\sum_{pq}X_{pq}^{\rho}(\widetilde{W}\widetilde{H})_{pq}^{1-\rho}\right)^{-1}X_{ij}^{\rho}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-\rho}\left(\widetilde{W}_{ik}\widetilde{H}_{kj}\right)^{\rho}(W_{ik}H_{kj})^{1-\rho}$ $+\left(\sum_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}\right)^{-1}W_{ik}H_{kj}$
2) $0 < \rho < 1$	$-\left(\sum_{pq}X_{pq}^{\rho}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^{1-\rho}\right)^{-1}X_{ij}^{\rho}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-\rho}\widetilde{W}_{ik}\widetilde{H}_{kj}\ln(W_{ik}H_{kj})$ $+\left(\sum_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}\right)^{-1}W_{ik}H_{kj}$

similar way. All of the three auxiliary functions satisfy Assumption 1, Assumption 2, and the first and second conditions of Assumption 3. Also, each $\bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ can be expressed as $a_{1ijk}(W_{ik}H_{kj})^{b_1} + a_{2ijk}(W_{ik}H_{kj})^{b_2}$ where $b_1 \neq b_2$ and a_{1ijk} and a_{2ijk} are independent of W_{ik} and H_{kj}. Therefore, the multiplicative update rules can be obtained from the auxiliary functions. The formulae for $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ for the three error functions are shown in Table 3. Furthermore, for each of the three error functions, it is easy to prove that $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ and $g_{jk}(\boldsymbol{W}, \boldsymbol{H})$ satisfy the third condition of Assumption 3. Upper bounds for $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ on $\mathcal{F}_\epsilon$ are shown in Table 4, and how to derive the upper bound for Kullback-Leibler divergence is described in Appendix C.

From these results, we obtain the following theorem.

Table 3 Multiplicative update rules obtained from the last three error functions in Table 1 with regularization term.

Error function	$f_{ik}(\boldsymbol{W},\boldsymbol{H})$
Kullback-Leibler divergence	$W_{ik}\left(\frac{(\sum_{pq}X_{pq})^{-1}\sum_j X_{ij}(\boldsymbol{WH})_{ij}^{-1}H_{kj}+C\sum_{pq}X_{pq}\sum_j H_{kj}}{(\sum_{pq}(\boldsymbol{WH})_{pq})^{-1}\sum_j H_{kj}+C\sum_{pq}(\boldsymbol{WH})_{pq}\sum_j H_{kj}}\right)^{\frac{1}{2}}$
γ-divergence	$W_{ik}\left(\frac{(\sum_{pq}X_{pq}(\boldsymbol{WH})_{pq}^{\gamma})^{-1}\sum_j X_{ij}(\boldsymbol{WH})_{ij}^{\gamma-1}H_{kj}+C\sum_{pq}X_{pq}\sum_j H_{kj}}{(\sum_{pq}(\boldsymbol{WH})_{pq}^{1+\gamma})^{-1}\sum_j(\boldsymbol{WH})_{ij}^{\gamma}H_{kj}+C\sum_{pq}(\boldsymbol{WH})_{pq}\sum_j H_{kj}}\right)^{\eta}$ where $\eta=\begin{cases}1/(1+\gamma), \text{ if } \gamma>0\\ 1/(1-\gamma), \text{ if } \gamma<0,\ \gamma\neq-1\end{cases}$
Rényi divergence	$W_{ik}\left(\frac{(\sum_{pq}X_{pq}^{\rho}(\boldsymbol{WH})_{pq}^{1-\rho})^{-1}\sum_j X_{ij}^{\rho}(\boldsymbol{WH})_{ij}^{-\rho}H_{kj}+C\sum_{pq}X_{pq}\sum_j H_{kj}}{(\sum_{pq}(\boldsymbol{WH})_{pq})^{-1}\sum_j H_{kj}+C\sum_{pq}(\boldsymbol{WH})_{pq}\sum_j H_{kj}}\right)^{\eta}$ where $\eta=\begin{cases}1/\rho, \text{ if } \rho>1\\ 1, \quad \text{if } 0<\rho<1\end{cases}$

Table 4 Upper bounds for $f_{ik}(\boldsymbol{W},\boldsymbol{H})$ in Table 3 on $\mathcal{F}_\epsilon$.

Error function	Upper bound
Kullback-Leibler divergence	$\left(\frac{1}{\epsilon^3 nrC}+\frac{1}{\epsilon n}\sum_{pq}X_{pq}\right)^{\frac{1}{2}}W_{ik}^{\frac{1}{2}}$
γ-divergence	$\left(\frac{1}{\epsilon^3 mrC}+\frac{1}{\epsilon n}\sum_{pq}X_{pq}\right)^{\eta}W_{ik}^{1-\eta}$ where $\eta=\begin{cases}1/(1+\gamma), \text{ if } \gamma>0\\ 1/(1-\gamma), \text{ if } \gamma<0,\ \gamma\neq-1\end{cases}$
Rényi divergence	$\left(\frac{1}{\epsilon^3 mrC}+\frac{1}{\epsilon n}\sum_{pq}X_{pq}\right)^{\eta}W_{ik}^{1-\eta}$ where $\eta=\begin{cases}1/\rho, \text{ if } \rho>1\\ 1, \quad \text{if } 0<\rho<1\end{cases}$

Theorem 5 For Kullback-Leibler divergence, γ-divergence and Rényi divergence with the regularization term (34), the modified multiplicative update rule obtained by the unified method of Yang and Oja has the global convergence property for any positive constant ϵ.

6 Conclusions

A unified global convergence analysis of the multiplicative update rule for NMF has been presented. We have given a sufficient condition on the error function and the auxiliary function for a slightly modified version of the multiplicative update rule to have the global convergence property in the sense that any sequence of solutions contains at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the corresponding optimization problem. This result can be applied to a wide variety of multiplicative update rules to examine their global convergence. In fact, we have proved the global convergence of eleven different multiplicative update rules.

Recently, many variants of NMF have been proposed (see the review paper by Wang and Zhang [41] and references therein). Extending the results in this paper to such variants is an interesting and important direction for future

research. Another interesting direction is to develop a method for determining the global convergence of the modified multiplicative update rule only from the error function. The motivation of this problem comes from the observation that the multiplicative update rule is uniquely determined by the error function if we assume that the auxiliary function is always derived by the method of Yang and Oja.

A How to Derive Auxiliary Functions

In the unified method of Yang and Oja [45], an auxiliary function is systematically derived from a given generalized polynomial by using three rules. They are described by the following lemmas. Because the mathematical expressions differ from those in [45] due to the introduction of the framework of a single auxiliary function, we provide proofs for the sake of readers' convenience.

Lemma 8 Suppose that the error function is expressed as $D(\boldsymbol{W},\boldsymbol{H}) = a\left(\sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c\right)^d$ where a and c are nonzero constants, b_{ij} are positive constants, and d is a constant other than 0 or 1. If $\xi(x) \triangleq ax^d$ is convex in $\mathbb{R}_{++}$, let

$$\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = a\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^{d-1} \sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c \left(\frac{(\boldsymbol{W}\boldsymbol{H})_{ij}}{(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}}\right)^{cd}.$$

If $\xi(x)$ is concave in $\mathbb{R}_{++}$, let

$$\begin{aligned}\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = a\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^{d} + ad\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^{d-1} \\ \times \left(\sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c - \sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right).\end{aligned}$$

Then $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$, and satisfies the conditions in Assumptions 1 and 2.

Proof There are two cases to consider: One is that $\xi(x) \triangleq ax^d$ is convex, and the other is that $\xi(x)$ is concave. In either case, it is easy to see that the following statements hold true:

1. $\bar{D}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = D(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})$ for all $(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) \in \mathcal{F}_0$,
2. $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is differentiable at any point in $\operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0$, and
3. $\nabla_{\boldsymbol{W}}\bar{D}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{W}}D(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})$ and $\nabla_{\boldsymbol{H}}\bar{D}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{H}}D(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})$ for all $(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) \in \operatorname{int}\mathcal{F}_0$.

Therefore, it suffices for us to show that

$$\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \geq D(\boldsymbol{W},\boldsymbol{H})$$

for all $(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) \in \operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0$. Suppose first that $\xi(x) = ax^d$ is convex in $\mathbb{R}_{++}$. Then, for any numbers $x_{11}, x_{12}, \ldots, x_{mn}$ and any positive numbers $\lambda_{11}, \lambda_{12}, \ldots, \lambda_{mn}$ such that $\sum_{ij}\lambda_{ij} = 1$, it follows from Jensen's inequality that

$$\xi\left(\sum_{ij} x_{ij}\right) = \xi\left(\sum_{ij}\lambda_{ij}\cdot\frac{x_{ij}}{\lambda_{ij}}\right) \leq \sum_{ij}\lambda_{ij}\xi\left(\frac{x_{ij}}{\lambda_{ij}}\right).$$

Substituting $x_{ij} = b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c$ and $\lambda_{ij} = b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c / \sum_{pq} b_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^c$ into this equation, we have

$$\begin{aligned}
D(\boldsymbol{W},\boldsymbol{H}) &= a\left(\sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c\right)^d \\
&\le a\sum_{ij}\frac{b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c}{\sum_{pq} b_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^c}\left(\frac{b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c}{b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c / \sum_{pq} b_{pq}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}^c}\right)^d \\
&= a\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^{d-1}\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\left(\frac{(\boldsymbol{W}\boldsymbol{H})_{ij}}{(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}}\right)^{cd} \\
&= \bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})\,.
\end{aligned}$$

Suppose next that $\xi(x) = ax^d$ is concave in $\mathbb{R}_{++}$. Then, for any positive numbers x and $\tilde{x}$, the following inequality holds:

$$\xi(x) \le \xi(\tilde{x}) + \xi'(\tilde{x})(x - \tilde{x})\,.$$

Substituting $x = \sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c$ and $\tilde{x} = \sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c$ into this inequality, we have

$$\begin{aligned}
D(\boldsymbol{W},\boldsymbol{H}) &= a\left(\sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c\right)^d \\
&\le a\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^d \\
&\quad + ad\left(\sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right)^{d-1}\left(\sum_{ij} b_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^c - \sum_{ij} b_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^c\right) \\
&= \bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})
\end{aligned}$$

which completes the proof. □

Lemma 9 Suppose that the error function is expressed as $D(\boldsymbol{W},\boldsymbol{H}) = \sum_{ij} a_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^b$ where a_{ij} are nonzero constants and b is a constant other than 0 or 1. If $\xi_{ij}(x) \triangleq a_{ij}x^b$ is convex in $\mathbb{R}_{++}$, let

$$\bar{D}_{ij}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = a_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{b-1}\sum_k (\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-b}(W_{ik}H_{kj})^b\,.$$

If $\xi_{ij}(x)$ is concave in $\mathbb{R}_{++}$, let

$$\bar{D}_{ij}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = a_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^b + a_{ij}b(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{b-1}\left((\boldsymbol{W}\boldsymbol{H})_{ij} - (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)\,.$$

Then $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \sum_{ij}\bar{D}_{ij}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$, and satisfies the conditions in Assumptions 1 and 2.

Proof Let $D_{ij}(\boldsymbol{W},\boldsymbol{H}) = a_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}^b$. There are two cases to consider: One is that $\xi_{ij}(x) \triangleq a_{ij}x^b$ is convex and the other is that $\xi_{ij}(x)$ is concave. In either case, we easily see that the following statements hold true:

1. $\bar{D}_{ij}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = D_{ij}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})$ for all $(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) \in \operatorname{int}\mathcal{F}_0$,
2. $\bar{D}_{ij}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is differentiable at any point in $\operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0$, and

3. $\nabla_{\boldsymbol{W}} \bar{D}_{ij}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{W}} D_{ij}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $\nabla_{\boldsymbol{H}} \bar{D}_{ij}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{H}} D_{ij}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ for all $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0$.

Therefore, it suffices for us to show that

$$\bar{D}_{ij}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \geq D_{ij}(\boldsymbol{W}, \boldsymbol{H})$$

for all $(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \in \operatorname{int} \mathcal{F}_0 \times \operatorname{int} \mathcal{F}_0$. Suppose first that $\xi_{ij}(x) = a_{ij} x^b$ is convex in $\mathbb{R}_{++}$. Then, for any numbers $x_1, x_2, \ldots, x_r$ and any positive numbers $\lambda_1, \lambda_2, \ldots, \lambda_r$ such that $\sum_k \lambda_k = 1$, it follows from Jensen's inequality that

$$\xi_{ij}\left(\sum_k x_k\right) = \xi_{ij}\left(\sum_k \lambda_k \cdot \frac{x_k}{\lambda_k}\right) \leq \sum_k \lambda_k \xi_{ij}\left(\frac{x_k}{\lambda_k}\right) .$$

Substituting $x_k = W_{ik} H_{kj}$ and $\lambda_k = \widetilde{W}_{ik} \widetilde{H}_{kj} / (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}$ into this equation, we have

$$\begin{aligned} D_{ij}(\boldsymbol{W}, \boldsymbol{H}) &= a_{ij} (\boldsymbol{W}\boldsymbol{H})_{ij}^b \\ &\leq \sum_k \frac{\widetilde{W}_{ik} \widetilde{H}_{kj}}{(\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}} a_{ij} \left(\frac{W_{ik} H_{kj}}{\widetilde{W}_{ik} \widetilde{H}_{kj} / (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}} \right)^b \\ &= a_{ij} (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}^{b-1} \sum_k (\widetilde{W}_{ik} \widetilde{H}_{kj})^{1-b} (W_{ik} H_{kj})^b \\ &= \bar{D}_{ij}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) . \end{aligned}$$

Suppose next that $\xi_{ij}(x) = a_{ij} x^b$ is concave in $\mathbb{R}_{++}$. Then, for any positive numbers x and $\tilde{x}$, the following inequality holds:

$$\xi_{ij}(x) \leq \xi_{ij}(\tilde{x}) + \xi'_{ij}(\tilde{x})(x - \tilde{x}) .$$

Substituting $x = (\boldsymbol{W}\boldsymbol{H})_{ij}$ and $\tilde{x} = (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}$ into this inequality, we have

$$\begin{aligned} D_{ij}(\boldsymbol{W}, \boldsymbol{H}) &= a_{ij} (\boldsymbol{W}\boldsymbol{H})_{ij}^b \\ &\leq a_{ij} (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}^b + a_{ij} b (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij}^{b-1} \left((\boldsymbol{W}\boldsymbol{H})_{ij} - (\widetilde{\boldsymbol{W}} \widetilde{\boldsymbol{H}})_{ij} \right) \\ &= \bar{D}_{ij}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \end{aligned}$$

which completes the proof. □

Lemma 10 Suppose that the error function is expressed as $D(\boldsymbol{W}, \boldsymbol{H}) = \sum_{tijk} a_{tijk} (W_{ik} H_{kj})^{b_t}$ where a_{tijk} are nonzero constants and b_t are constants, $a_{tijk} x^{b_t}$ is convex in $\mathbb{R}_{++}$, and $\{b_t\}$ contains at least two distinct nonzero numbers. Let $b_{\max} = \max\{b_t \,|\, b_t \neq 0\}$ and $b_{\min} = \min\{b_t \,|\, b_t \neq 0\}$. Let us define $\bar{D}_{tijk}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ on $\operatorname{int} \mathcal{F}_0 \times \operatorname{int} \mathcal{F}_0$ as follows:

1. If $b_t \in \{b_{\min}, b_{\max}, 0\}$, let

$$\bar{D}_{tijk}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = a_{tijk} (W_{ik} H_{kj})^{b_t} .$$

2. If $b_t \notin \{b_{\min}, b_{\max}, 0\}$ and
   (a) if $(b_t > 1) \vee ((b_t = 1) \wedge (a_{tijk} > 0))$, let

$$\begin{aligned} \bar{D}_{tijk}(\boldsymbol{W}, \boldsymbol{H}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = {} & \frac{a_{tijk} b_t}{b_{\max}} (\widetilde{W}_{ik} \widetilde{H}_{kj})^{b_t - b_{\max}} \left(W_{ik} H_{kj} \right)^{b_{\max}} \\ & + a_{tijk} (\widetilde{W}_{ik} \widetilde{H}_{kj})^{b_t} \left(1 - \frac{b_t}{b_{\max}} \right) , \end{aligned}$$

(b) if $(b_t < 1) \vee ((b_t = 1) \wedge (a_{tijk} < 0))$, let

$$\bar{D}_{tijk}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \frac{a_{tijk}b_t}{b_{\min}}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{b_t-b_{\min}}\left(W_{ik}H_{kj}\right)^{b_{\min}} + a_{tijk}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{b_t}\left(1-\frac{b_t}{b_{\min}}\right).$$

Then $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \sum_{tijk}\bar{D}_{tijk}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$, and strictly convex in $\operatorname{int}\mathcal{F}_0$. Furthermore, $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ satisfies the conditions in Assumptions 1 and 2.

Proof Let $D_{tijk}(\boldsymbol{W},\boldsymbol{H}) = a_{tijk}(W_{ik}H_{kj})^{b_t}$. There are three cases to consider depending on the values of a_{tijk} and b_t. In either case, we easily see that $\bar{D}_{tijk}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is differentiable at any point in $\operatorname{int}\mathcal{F}_0 \times \operatorname{int}\mathcal{F}_0$ and that

$$\nabla_{\boldsymbol{W}}\bar{D}_{tijk}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{W}}D_{tijk}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}),$$
$$\nabla_{\boldsymbol{H}}\bar{D}_{tijk}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}},\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) = \nabla_{\boldsymbol{H}}D_{tijk}(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}})$$

hold for all $(\hat{\boldsymbol{W}},\hat{\boldsymbol{H}}) \in \mathcal{F}_0$. Also, it has already been shown by Yang and Oja [45, Lemma 2] that $\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})$ is an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$. □

B Derivation of Auxiliary Function for Kullback-Leibler Divergence with Regularization Term

We derive an auxiliary function for Kullback-Leibler divergence with the regularization term by using the unified method of Yang and Oja. First of all, we rewrite the error function by using (33) as follows:

$$D(\boldsymbol{W},\boldsymbol{H}) = \lim_{\mu\to 0^+}\frac{1}{\mu}\Big(D_1(\boldsymbol{W},\boldsymbol{H}) + D_2(\boldsymbol{W},\boldsymbol{H}) + D_3(\boldsymbol{W},\boldsymbol{H}) + D_4(\boldsymbol{W},\boldsymbol{H})\Big) + \sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}\ln\left(\frac{X_{ij}}{\sum_{pq}X_{pq}}\right) + \frac{C}{2}\left(\sum_{ij}X_{ij}\right)^2$$

where

$$D_1(\boldsymbol{W},\boldsymbol{H}) = \left(\sum_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}\right)^{\mu},$$
$$D_2(\boldsymbol{W},\boldsymbol{H}) = -\sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}(\boldsymbol{W}\boldsymbol{H})_{ij}^{\mu},$$
$$D_3(\boldsymbol{W},\boldsymbol{H}) = -\mu C\sum_{ij}X_{ij}\cdot\sum_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij},$$
$$D_4(\boldsymbol{W},\boldsymbol{H}) = \frac{\mu C}{2}\left(\sum_{ij}(\boldsymbol{W}\boldsymbol{H})_{ij}\right)^2.$$

Let us assume that μ is a sufficiently small positive constant. Applying Lemmas 8 and 9 to these functions, we have the following auxiliary functions:

$$\overline{D_1}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \mu\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)^{\mu-1}\sum_{ijk}W_{ik}H_{kj} + (1-\mu)\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)^{\mu},$$

$$\overline{D_2}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = -\sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{\mu-1}\sum_{k}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-\mu}(W_{ik}H_{kj})^{\mu},$$
$$\overline{D_3}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = -\mu C\sum_{ij}X_{ij}\cdot\sum_{ijk}W_{ik}H_{kj},$$
$$\overline{D_4}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \frac{\mu C}{2}\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\cdot\sum_{ijk}\left(\widetilde{W}_{ik}\widetilde{H}_{kj}\right)^{-1}(W_{ik}H_{kj})^{2}.$$

The exponents of $W_{ik}H_{kj}$ in these auxiliary functions are 1, μ, 1 and 2. The minimum is μ and the maximum is 2. So we apply Lemma 10 to some of these auxiliary functions to obtain an auxiliary function of $D(\boldsymbol{W},\boldsymbol{H})$ such that the exponents of $W_{ik}H_{kj}$ are restricted to μ and 2. Applying Lemma 10 to $\overline{D_1}$, we obtain another auxiliary function of D_1 as follows:

$$\overline{\overline{D_1}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \frac{\mu}{2}\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)^{\mu-1}\sum_{ijk}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1}(W_{ik}H_{kj})^{2} + \left(1-\frac{\mu}{2}\right)\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)^{\mu}.$$

Applying Lemma 10 to $\overline{D_3}$, we obtain another auxiliary function of D_3 as follows:

$$\overline{\overline{D_3}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = -C\sum_{ij}X_{ij}\cdot\sum_{ijk}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{1-\mu}(W_{ik}H_{kj})^{\mu} + (1-\mu)C\sum_{ij}X_{ij}\cdot\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}.$$

As a result, we have the following auxiliary function:

$$\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \lim_{\mu\to 0+}\frac{1}{\mu}\left(\overline{\overline{D_1}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{D_2}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{\overline{D_3}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{D_4}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})\right) + \sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}\ln\left(\frac{X_{ij}}{\sum_{pq}X_{pq}}\right) + \frac{C}{2}\left(\sum_{ij}X_{ij}\right)^{2}.$$

Because

$$\lim_{\mu\to 0+}\left(\overline{\overline{D_1}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{D_2}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{\overline{D_3}}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) + \overline{D_4}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}})\right) = 0,$$

we apply L'Hôpital's rule. Then we have

$$\bar{D}(\boldsymbol{W},\boldsymbol{H},\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{H}}) = \frac{1}{2}\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)^{-1}\sum_{ijk}(\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1}(W_{ik}H_{kj})^{2} - \frac{1}{2} + \ln\left(\sum_{ij}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right) - \sum_{ij}\frac{X_{ij}}{\sum_{pq}X_{pq}}(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1}\ln\left((\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}\right)\sum_{k}\widetilde{W}_{ik}\widetilde{H}_{kj}$$

$$
\begin{aligned}
&- \sum_{ij} \frac{X_{ij}}{\sum_{pq} X_{pq}} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1} \sum_{k} \widetilde{W}_{ik}\widetilde{H}_{kj} \ln\left(\frac{W_{ik}H_{kj}}{\widetilde{W}_{ik}\widetilde{H}_{kj}}\right) \\
&- C \sum_{ij} X_{ij} \cdot \sum_{ijk} \widetilde{W}_{ik}\widetilde{H}_{kj} \ln\left(\frac{W_{ik}H_{kj}}{\widetilde{W}_{ik}\widetilde{H}_{kj}}\right) \\
&- C \sum_{ij} X_{ij} \cdot \sum_{ij} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij} \\
&+ \frac{C}{2} \sum_{ij} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij} \cdot \sum_{ijk} (\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1} (W_{ik}H_{kj})^2 \\
&+ \sum_{ij} \frac{X_{ij}}{\sum_{pq} X_{pq}} \ln\left(\frac{X_{ij}}{\sum_{pq} X_{pq}}\right) + \frac{C}{2}\left(\sum_{ij} X_{ij}\right)^2
\end{aligned}
$$

which can be rewritten in the form of (18) with

$$
\begin{aligned}
\bar{D}^1_{ijk}(W_{ik}, H_{kj}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \frac{1}{2}&\left(\sum_{pq} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq}\right)^{-1} (\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1} (W_{ik}H_{kj})^2 \\
&- \frac{X_{ij}}{\sum_{pq} X_{pq}} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{ij}^{-1} \widetilde{W}_{ik}\widetilde{H}_{kj} \ln(W_{ik}H_{ij}) \\
&- C \sum_{pq} X_{pq} \cdot \widetilde{W}_{ik}\widetilde{H}_{kj} \ln(W_{ik}H_{kj}) \\
&+ \frac{C}{2} \sum_{pq} (\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{H}})_{pq} \cdot (\widetilde{W}_{ik}\widetilde{H}_{kj})^{-1} (W_{ik}H_{kj})^2 .
\end{aligned}
$$

C Derivation of Upper Bound for Multiplicative Update Rule Obtained from Kullback-Leibler Divergence With Regularization Term

For the first multiplicative update rule shown in Table 3, which is obtained from Kullback-Leibler divergence with the regularization term, we derive an upper bound for $f_{ik}(\boldsymbol{W}, \boldsymbol{H})$ on $\mathcal{F}_\epsilon$. By simple mathematical manipulations, we have the following inequalities:

$$
\begin{aligned}
f_{ik}(\boldsymbol{W}, \boldsymbol{H}) &< W_{ik} \left(\frac{\sum_j \frac{X_{ij}}{\sum_{pq} X_{pq}} (\boldsymbol{W}\boldsymbol{H})_{ij}^{-1} H_{kj} + C \sum_{pq} X_{pq} \sum_j H_{kj}}{C \sum_{pq} (\boldsymbol{W}\boldsymbol{H})_{pq} \sum_j H_{kj}}\right)^{\frac{1}{2}} \\
&\le W_{ik} \left(\frac{\sum_j \frac{X_{ij}}{\sum_{pq} X_{pq}} (\boldsymbol{W}\boldsymbol{H})_{ij}^{-1} H_{kj}}{C \sum_{pq} (\boldsymbol{W}\boldsymbol{H})_{pq} \sum_j H_{kj}} + \frac{\sum_{pq} X_{pq}}{\sum_{pq} (\boldsymbol{W}\boldsymbol{H})_{pq}}\right)^{\frac{1}{2}} .
\end{aligned}
$$

Here note that

$$
\begin{aligned}
\frac{\sum_j \frac{X_{ij}}{\sum_{pq} X_{pq}} (\boldsymbol{W}\boldsymbol{H})_{ij}^{-1} H_{kj}}{\sum_j H_{kj}} &= \sum_j \frac{X_{ij}}{\sum_{pq} X_{pq}} (\boldsymbol{W}\boldsymbol{H})_{ij}^{-1} \frac{H_{kj}}{\sum_q H_{kq}} \\
&< \frac{1}{\epsilon^2 r} \frac{\sum_j X_{ij}}{\sum_{pq} X_{pq}} \\
&< \frac{1}{\epsilon^2 r}
\end{aligned}
$$

and

$$\frac{1}{\sum_{pq}(\boldsymbol{WH})_{pq}} < \frac{1}{\sum_q W_{ik}H_{kq}} = \frac{1}{W_{ik}\sum_q H_{kq}} \leq \frac{1}{\epsilon n W_{ik}}.$$

Therefore we have

$$f_{ik}(\boldsymbol{W},\boldsymbol{H}) \leq W_{ik}^{\frac{1}{2}}\left(\frac{1}{\epsilon^3 nrC} + \frac{1}{\epsilon n}\sum_{pq} X_{pq}\right)^{\frac{1}{2}}.$$

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number JP15K00035.

References

1. Badeau, R., Bertin, N., Vincent, E.: Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. IEEE Transactions on Neural Networks **21**(12), 1869–1881 (2010)
2. Berman, A., Plemmons, R.: Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York (1979)
3. Berry, M.W., Browne, M.: Email surveillance using non-negative matrix factorization. Computational and Mathematical Organization Theory **11**, 249–264 (2005)
4. Campbell, S.L., Poole, G.D.: Computing nonnegative rank factorizations. Linear Algebra and its Applications **35**, 175–182 (1981)
5. Chen, J.C.: The nonnegative rank factorizations of nonnegative matrices. Linear algebra and its applications **62**, 207–217 (1984)
6. Chi, E.C., Kolda, T.G.: On tensors, sparsity, and nonnegative factorizations. SIAM Journal on Matrix Analysis and Applications **33**(4), 1272–1299 (2012)
7. Cichocki, A., Lee, H., Kim, Y.D., Choi, S.: Non-negative matrix factorization with α-divergence. Pattern Recognition Letters **29**(9), 1433–1440 (2008)
8. Cichocki, A., Phan, A.H.: Fast local algorithms for large scale nonnegative matrix and tensor factorization. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **E92-A**(3), 708–721 (2009)
9. Cichocki, A., Zdunek, R., Amari, S.I.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: Lecture Notes in Computer Science, vol. 4666, pp. 169–176. Springer (2007)
10. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. John Wiley & Sons, West Sussex, U.K. (2009)
11. Dhillon, I.S., Sra, S.: Generalized nonnegative matrix approximations with Bregman divergences. In: Advances in Neural Information Processing Systems, pp. 283–290 (2005)
12. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Computation **21**(3), 793–830 (2009)
13. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β-divergence. Neural Computation **23**(9), 2421–2456 (2011)
14. Finesso, L., Spreij, P.: Nonnegative matrix factorization and I-divergence alternating minimization. Linear Algebra and its Applications **416**, 270–287 (2006)
15. Gillis, N., Glineur, F.: Nonnegative factorization and the maximum edge biclique problem. arXiv e-prints (2008)
16. Gonzalez, E.F., Zhang, Y.: Accelerating the Lee-Seung algorithm for non-negative matrix factorization. Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02 (2005)
17. Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: an optimal gradient method for nonnegative matrix factorization. IEEE Transactions on Signal Processing **60**(6), 2882–2898 (2012)
18. Guillamet, D., Vitria, J.: Non-negative matrix factorization for face recognition. In: Lecture Notes in Artificial Intelligence, pp. 336–344. Springer (2002)

19. Hansen, S., Plantenga, T., Kolda, T.G.: Newton-based optimization for Kullback-Leibler nonnegative tensor factorizations. Optimization Methods and Software **30**(5), 1002–1029 (2015)
20. Holzapfel, A., Stylianou, Y.: Musical genre classification using nonnegative matrix factorization-based features. IEEE Transactions on Audio, Speech, and Language Processing **16**(2), 424–434 (2008)
21. Hsieh, C.J., Dhillon, I.S.: Fast coordinate descent methods with variable selection for non-negative matrix factorization. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1064–1072. ACM (2011)
22. Katayama, J., Takahashi, N., Takeuchi, J.: Boundedness of modified multiplicative updates for nonnegative matrix factorization. In: Proceedings of the Fifth International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, pp. 252–255 (2013)
23. Kim, D., Sra, S., Dhillon, I.S.: Fast newton-type methods for the least squares nonnegative matrix approximation problem. In: Proceedings of the Sixth SIAM International Conference on Data Mining, pp. 343–354. SIAM (2007)
24. Kim, H., Park, H.: Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. SIAM Journal on Matrix Analysis and Applications **30**(2), 713–730 (2008)
25. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorization: a unified view based on block coordinate descent framework. Journal of Global Optimization **58**(2), 285–319 (2014)
26. Kimura, T., Takahashi, N.: Global convergence of a modified HALS algorithm for nonnegative matrix factorization. In: Proceedings of 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, pp. 21–24 (2015)
27. Kompass, R.: A generalized divergence measure for nonnegative matrix factorization. Neural Computation **19**(3), 780–791 (2007)
28. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–792 (1999)
29. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: T.K. Leen, T.G. Dietterich, V. Tresp (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 556–562 (2001)
30. Lin, C.J.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Transactions on Neural Networks **18**(6), 1589–1596 (2007)
31. Lin, C.J.: Projected gradient methods for non-negative matrix factorization. Neural Computation **19**(10), 2756–2779 (2007)
32. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)
33. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Transactions on Audio, Speech, and Language Processing **18**(3), 576–588 (2010)
34. Seki, M., Takahashi, N.: New updates based on Kullback-Leibler, gamma, and Rényi divergences for nonnegative matrix factorization. In: Proceedings of 2014 International Symposium on Nonlinear Theory and its Applications, pp. 48–51 (2014)
35. Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing and Management **42**, 373–386 (2006)
36. Takahashi, N., Hibi, R.: Global convergence of modified multiplicative updates for nonnegative matrix factorization. Computational Optimization and Applications **57**, 417–440 (2014)
37. Takahashi, N., Katayama, J., Takeuchi, J.: A generalized sufficient condition for global convergence of modified multiplicative updates for NMF. In: Proceedings of 2014 International Symposium on Nonlinear Theory and its Applications, pp. 44–47 (2014)
38. Takahashi, N., Nishi, T.: Global convergence of decomposition learning methods for support vector machines. IEEE Transactions on Neural Networks **17**(6), 1362–1369 (2006)

39. Vavasis, S.A.: On the complexity of nonnegative matrix factorization. SIAM Journal on Optimization **20**(3), 1364–1377 (2009)
40. Wang, R.S., Zhang, S., Wang, Y., Zhang, X.S., Chen, L.: Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. Neurocomputing **72**, 134–141 (2008)
41. Wang, Y.X., Zhang, Y.J.: Nonnegative matrix factorization: A comprehensive review. IEEE Transactions on Knowledge and Data Engineering **25**(6), 1336–1353 (2013)
42. Wu, C.F.J.: On the convergence properties of the EM algorithm. The Ananls of Statistics **11**(1), 95–103 (1983)
43. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273. ACM (2003)
44. Yamauchi, S., Kawakita, M., Takeuchi, J.: Botnet detection based on non-negative matrix factorization and the MDL principle. In: Proceedings of 19th International Conference on Neural Information Processing, pp. 400–409. Springer (2012)
45. Yang, Z., Oja, E.: Unified development of multiplicative algorithm for linear and quadratic nonnegative matrix factorization. IEEE Transactions on Neural Networks **22**(12), 1878–1891 (2011)
46. Zangwill, W.: Nonlinear Programming: A Unified Approach. Prentice-Hall, Englewood Cliffs, NJ (1969)
47. Zhao, R., Tan, V.Y.: A unified convergence analysis of the multiplicative update algorithm for nonnegative matrix factorization. arXiv preprint arXiv:1609.00951 (2016)

Erratum Sheet

Norikazu Takahashi, Jiro Katayama, Masato Seki and Jun'ichi Takeuchi, "A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization", Computational Optimization and Applications, vol.71, no.1, pp.221–250, 2018.

1. Eq.(10) on Page 226 is not correct. It should be replaced with

$$H_{kj}^{(l+1)} = H_{kj}^{(l)} \frac{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}} \boldsymbol{X})_{kj}}{((\boldsymbol{W}^{(l+1)})^{\mathrm{T}} \boldsymbol{W}^{(l+1)} \boldsymbol{H}^{(l)})_{kj}} \tag{10}$$

(Last updated: April 5, 2019)