A UNIFIED CONVERGENCE ANALYSIS OF THE MULTIPLICATIVE UPDATE ALGORITHM FOR NONNEGATIVE MATRIX FACTORIZATION

Renbo Zhao, Vincent Y. F. Tan

Department of Electrical and Computer Engineering & Department of Mathematics,
National University of Singapore

ABSTRACT

The multiplicative update (MU) algorithm has been used extensively to estimate the basis and coefficient matrices in nonnegative matrix factorization (NMF) problems under a wide range of divergences and regularizations. However, theoretical convergence guarantees have only been derived for a few special divergences. In this work, we provide a conceptually simple, self-contained, and unified proof for the convergence of the MU algorithm applied on NMF with a wide range of divergences and regularizations. Our result shows the sequence of iterates (i.e., pairs of basis and coefficient matrices) produced by the MU algorithm converges to the set of stationary points of the NMF (optimization) problem. Our proof strategy has the potential to open up new avenues for analyzing similar problems.

Index Terms— Nonnegative Matrix Factorization, Multiplicative Update Algorithm, Convergence Analysis, Nonconvex Optimization, Stationary Points

1. INTRODUCTION

Nonnegative Matrix Factorization (NMF) has been a popular dimensionality reduction technique, due to its non-subtractive and parts-based interpretation on the learned basis [1]. In the general formulation of NMF, given a nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, one seeks a nonnegative basis matrix $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and a nonnegative coefficient matrix $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ such that $\mathbf{V} \approx \mathbf{WH}$. One usually solves

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0} \left[\overline{\ell}(\mathbf{W},\mathbf{H}) \triangleq \overline{D}(\mathbf{V}\|\mathbf{WH})\right]. \tag{1}$$

In (1), $\overline{D}(\cdot\|\cdot)$ denotes the divergence, or distance, between two nonnegative matrices. In the NMF literature, many algorithms have been proposed to solve (1), including multiplicative updates (MU) [2,3], block principal pivoting (BPP) [4], projected gradient descent (PGD) [5], active set methods (ASM) [6] and the alternating direction method of multipliers (ADMM) [7]. However, some algorithms only solve (1) for certain divergences $\overline{D}(\cdot\|\cdot)$. For example, the BPP and ASM algorithms are only applicable to the squared-Frobenius loss. Among all algorithms, the MU algorithm is arguably the most widely applicable—it has been applied to NMF with the α-divergence [8], the β-divergence [3], the γ-divergence [9], the $\alpha\beta$-divergence [10], etc. However, despite its popularity and wide applicability, it is largely an heuristic algorithm in the sense that little of its convergence properties is known. In particular, most works [2,3,8] show that the sequence of objective values $\{\overline{\ell}(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$ in the MU algorithm is nonincreasing and hence converges. However, the convergence of objective values does not imply the convergence of the sequence of matrix pairs $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$, whose limit points (if they exist) serve as candidates for the output of the MU algorithm. Moreover, when the MU algorithm is used on real applications, such as music analysis [11], topic modeling [1] and source separation [8], the limit points of $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$ are meaningful and representative of the latent factors. Thus, the convergence properties of $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$, and in particular the properties of its limit points, are of fundamental importance.

1.1. Related Works

Due to the nonconvex nature of (1), algorithms that guarantee to converge to the global (or local) minima of (1) are in general out-of-reach. Indeed, [12] has shown that (1) is NP-hard. Thus existing works mainly study convergence to the stationary points (see Definition 3) of (1).[1] For the MU algorithm, some previous works on its convergence include [15–17]. For simplicity, all of the MU algorithms in these works only consider a special case of (1), namely $\overline{D}(\mathbf{V}\|\mathbf{WH}) = \frac{1}{2}\|\mathbf{V}-\mathbf{WH}\|_F^2$. In particular, a principled and rigorous analysis was performed in [15]. In [15], Lin modifies the MU algorithm proposed in [2], and shows the sequence of iterates $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$ generated by this algorithm converges to the set of stationary points[2] of (1). Later, the authors of [16] and [17] propose different modifications of the MU algorithm in [2] and then provide sound convergence analyses accordingly. In another interesting research direction, [18] studies the stability of local minima of (1) under the MU algorithm, when $\overline{D}(\cdot\|\cdot)$ belongs to the family of β-divergences. However, it cannot resolve whether (and when) the MU algorithm converges to any local minimum of $\overline{\ell}(\cdot,\cdot)$. For other algorithms that aim to solve (1), some rigorous convergence analyses have been done in [19–21]. However, all of the analyses are confined to some special cases of $\overline{D}(\cdot\|\cdot)$, including the Itakura-Saito (IS), (generalized) Kullback-Leibler (KL) or squared-Frobenius losses.

1.2. Challenges and Main Contributions

Despite the rigorous analyses in previous works [15–17], some important questions still remain unresolved:

1. Is convergence analysis possible for the MU algorithm when $\overline{D}(\cdot\|\cdot)$ is not the squared-Frobenius loss?
2. In addition, is convergence analysis possible for the MU algorithm when the loss function $\overline{\ell}(\cdot,\cdot)$ also includes regularizers?

The authors (emails: {elezren,vtan}@nus.edu.sg) are supported by an NUS Young Investigator Award (R-263-000-B37-133).

[1] Under certain assumptions on the data matrix $\mathbf{V}$, e.g., the separability conditions proposed in [13], polynomial-time algorithms for exact NMF have been proposed, e.g., [14]. However, in most applications in signal processing and machine learning, $\mathbf{V}$ is contaminated by noise, thereby making the assumptions leveraged in these works invalid.

[2] See Definition 4 for the definition of convergence of a sequence to a set.

3. Furthermore, instead of a case-by-case study, is a unified convergence analysis possible?

These questions naturally arise due to the importance of utilizing h-divergences and regularizers in various applications. Indeed, in many practical applications, the objective function (1) is not the squared-Frobenius loss. For example, the IS divergence is used in music analysis [11] and the KL divergence is often used in topic learning [1]. The use of such divergences can be justified from both theoretical (i.e., maximum likelihood considerations) and practical viewpoints. For details, see [11, 22]. In addition, to enhance the interpretability of the learned dictionary and coefficient matrices, regularizers on $\mathbf{W}$ and/or $\mathbf{H}$ are typically employed. For example, the ℓ_1 regularization on columns of $\mathbf{H}$ promotes sparsity on the columns, hence each data sample (a column of $\mathbf{V}$) can be represented parsimoniously by a subset of feature vectors (columns of $\mathbf{W}$).

The above questions cannot be addressed by straightforward generalizations of the analysis techniques in [15–17]. Therefore, in this work, based on the *block majorization-minimization* framework [23,24], we propose a *unified* convergence analysis for the MU algorithm when $\overline{\ell}(\cdot,\cdot)$ includes both h-divergences and regularizers. We show that the sequence of iterates $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$ has at least one limit point and any limit point of this sequence is a stationary point of (1). We leverage the regularity properties of both the objective and surrogate functions.[3] In particular, the surrogate functions of interest to us here are termed *first-order surrogate functions*. Thus, as a *side contribution*, we also provide a principled and systematic way to construct first-order surrogate functions. Moreover, we also provide a theoretical justification of a popular heuristic, which involves adding a small positive constant to the denominator of the multiplicative factor. This heuristic not only preserves the numerical stability, but also ensures the joint coercivity of the loss function $\overline{\ell}(\cdot,\cdot)$. As a result, the existence of the limit point(s) of $\{(\mathbf{W}^t,\mathbf{H}^t)\}_{t=1}^{\infty}$ can be proved.

1.3. Notations

In this paper we use $\mathbb{R}_+$, $\mathbb{R}_{++}$ and $\mathbb{N}$ to denote the set of nonnegative real numbers, positive real numbers and natural numbers (excluding zero) respectively. For $n\in\mathbb{N}$, we define $[n]\triangleq\{1,2,\ldots,n\}$. We use boldface capital letters, boldface lowercase letters and plain lowercase letters to denote matrices, vectors and scalars respectively. For a vector $\mathbf{x}$, we denote its i-th entry, ℓ_1 and ℓ_2 norms as x_i, $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ respectively. For a matrix $\mathbf{X}$, we denote its (i,j)-th entry as x_{ij} and its $\ell_{1,1}$ norm as $\|\mathbf{X}\|_{1,1}\triangleq\sum_i\|\mathbf{x}_i\|_1$. In addition, we use $\mathbf{X}=0$ and $\mathbf{X}\geq 0$ to denote entrywise zero and nonnegativity. For matrices $\mathbf{X}$ and $\mathbf{Y}$, we use $\mathbf{X}\odot\mathbf{Y}$ and $\langle\mathbf{X},\mathbf{Y}\rangle$ to denote their Hadamard product and Frobenius inner product respectively. We use $\overset{\mathrm{c}}{=}$ to denote equality up to additive constants.

2. PROBLEM FORMULATION

2.1. Definition of h-Divergences

Before introducing the notion of *h-divergences*, we first define an important function

$$h(\nu,t)\triangleq\begin{cases}(\nu^t-1)/t, & t\in\mathbb{R}\setminus\{0\}\\ \log\nu, & t=0\end{cases},\ \nu\in\mathbb{R}_{++}. \tag{2}$$

[3] Informally, a surrogate function is a function that upper bounds the original function and is tight at some point(s) in the domain. See Definition 2 for a precise definition.

Definition 1 (h-divergences; [25])**.** Given any $\mathbf{V}\in\mathbb{R}_+^{F\times N}$, a divergence $D(\mathbf{V}\|\cdot):\mathbb{R}_+^{F\times N}\to\mathbb{R}_+$ is called a h-divergence if for any $\widehat{\mathbf{V}}\in\mathbb{R}_+^{F\times N}$, there exist a constant $P\in\mathbb{N}\setminus\{1\}$, such that

$$D(\mathbf{V}\|\widehat{\mathbf{V}})\overset{\mathrm{c}}{=}\sum_{p=1}^{P}\mu_p h\left(\sum_{i=1}^{F}\sum_{j=1}^{N}\nu_{pij}h(\widehat{v}_{ij},\zeta_p),\,\xi_p\right), \tag{3}$$

where '$\overset{\mathrm{c}}{=}$' omits constants that are independent of $\widehat{\mathbf{V}}$ and μ_p,ν_{pij},ζ_p and ξ_p are all real constants independent of $\widehat{\mathbf{V}}$. Moreover, $\{\zeta_p\}_{p=1}^{P}$ are distinct.

Remark 1. First, note that the h-divergences include many important classes of divergences, including the families of $\alpha\,(\alpha\neq 0)$, β, γ, α-β and Rényi divergences.[4] All of these divergences have been applied in the NMF literature [3,8–10,27]. Second, when $\mu_p=\xi_p=1$, for any $p\in[P]$, $D(\mathbf{V}\|\cdot)$ is separable across the entries of $\widehat{\mathbf{V}}$, i.e., $D(\mathbf{V}\|\widehat{\mathbf{V}})\overset{\mathrm{c}}{=}\sum_{i=1}^{F}\sum_{j=1}^{N}\sum_{p=1}^{P}\nu_{pij}h(\widehat{v}_{ij},\zeta_p)$. In the sequel, we term such a divergence as *separable h-divergence*. In particular, any member in the families of $\alpha\,(\alpha\neq 0)$ or β-divergences is separable. For example, taking $P=2$, $\nu_{1ij}=-v_{ij}$, $\zeta_1=0$, $\nu_{2ij}=1$ and $\zeta_2=1$, we obtain the KL divergence, which belongs to both the α- and β-divergence families.

2.2. Optimization Problem

We focus on the following optimization problem

$$\min_{\mathbf{W}\in\mathbb{R}_+^{F\times K},\mathbf{H}\in\mathbb{R}_+^{K\times N}}\ell(\mathbf{W},\mathbf{H}), \tag{4}$$

where $K<\min(F,N)$ and

$$\ell(\mathbf{W},\mathbf{H})\triangleq D(\mathbf{V}\|\mathbf{W}\mathbf{H})+\sum_{i=1}^{2}\lambda_i\phi_i(\mathbf{W})+\sum_{j=1}^{2}\widetilde{\lambda}_j\phi_j(\mathbf{H}). \tag{5}$$

In (5), $\mathbf{V}\in\mathbb{R}_{++}^{F\times N}$, $\{\lambda_1,\widetilde{\lambda}_1\}\subseteq\mathbb{R}_{++}$, $\{\lambda_2,\widetilde{\lambda}_2\}\subseteq\mathbb{R}_+$ and for any nonnegative matrix $\mathbf{X}$, $\phi_1(\mathbf{X})\triangleq\|\mathbf{X}\|_{1,1}$ and $\phi_2(\mathbf{X})\triangleq\|\mathbf{X}\|_F^2$.

Remark 2. We explain why we focus on the so-called *elastic-net regularizer* [28] on $(\mathbf{W},\mathbf{H})$. This regularizer includes the $\ell_{1,1}$ and Tikhonov regularizers as special cases, both of which are widely used in NMF. Specifically, the $\ell_{1,1}$ regularizer promotes element-wise sparsity on the basis matrix $\mathbf{W}$ and coefficient matrix $\mathbf{H}$ [29]. The Tikhonov regularizer promotes smoothness on $(\mathbf{W},\mathbf{H})$ and also prevents overfitting [30]. Second, the positivity of λ_1 and $\widetilde{\lambda}_1$ originates from a commonly used heuristic in the MU algorithm that ensures numerical stability in the updates. See Remark 4 for details.

3. ALGORITHMS

3.1. First-Order Surrogate Functions and General Framework

Definition 2. Given a finite-dimensional real Banach space $\mathcal{X}=\Pi_{i=1}^{n}\mathcal{X}_i$ and let $x\triangleq(x_1,\ldots,x_n)\in\mathcal{X}$, where for any $i\in[n]$, $x_i\in\mathcal{X}_i$ is the i-th block of x. Consider a differentiable function

[4] In particular, some important instances in the h-divergences include the Hellinger, IS, KL and squared-Frobenius divergences. When $\alpha=0$, the corresponding divergence is called the dual (generalized) KL divergence. With slight modifications of our methodology, all the propositions and theorems in this paper will also hold for this case. See Section 5.3 in the extended version [26] for details.

$f : \mathcal{X} \to \mathbb{R}$. For any $i \in [n]$, a first-order surrogate function of f for the i-th block x_i, $F_i(\cdot \,|\, \cdot) : \mathcal{X}_i \times \mathcal{X} \to \mathbb{R}$ satisfies

(P1) $F_i(\widetilde{x}_i \,|\, \widetilde{x}) = f(\widetilde{x})$, for any $\widetilde{x} \in \mathcal{X}$,

(P2) $F_i(x_i \,|\, \widetilde{x}) \geq f(\widetilde{x}_1, \ldots, x_i, \ldots, \widetilde{x}_n)$, for any $(x_i, \widetilde{x}) \in \mathcal{X}_i \times \mathcal{X}$.

(P3) $F_i(\cdot \,|\, \cdot)$ is differentiable on $\mathcal{X}_i \times \mathcal{X}$ and for any $\widetilde{x} \in \mathcal{X}$, there exists a function $g(\cdot \,|\, \widetilde{x}) : \mathcal{X}_i \to \mathbb{R}$ such that $\nabla_{x_i} F_i(x_i \,|\, \widetilde{x}) = g(x_i / \widetilde{x}_i \,|\, \widetilde{x})$, for any $x_i \in \mathcal{X}_i$,

(P4) $\nabla_{x_i} F_i(\widetilde{x}_i \,|\, \widetilde{x}) = \nabla f(\widetilde{x})$, for any $\widetilde{x} \in \mathcal{X}$,

(P5) $F_i(\cdot \,|\, \widetilde{x})$ is strictly convex on $\mathcal{X}_i$, for any $\widetilde{x} \in \mathcal{X}$.

If $F_i(\cdot \,|\, \cdot)$ only satisfies **(P1)** to **(P3)**, it is called a *surrogate function* of f for x_i.

Remark 3. We now explain the implications of the five properties in Definition 2. First define

$$x_i^* \triangleq \operatorname*{arg\,min}_{x_i \in \mathcal{X}_i} F_i(x_i | \widetilde{x}), \tag{6}$$

where the uniqueness of the minimizer in (6) is guaranteed by **(P5)**. Moreover, define $x^* \triangleq (\widetilde{x}_1, \ldots, x_i^*, \ldots, \widetilde{x}_n)$, then **(P1)** and **(P2)** together ensure $f(x^*) \leq f(\widetilde{x})$. **(P3)** ensures the minimization in (6) yields the multiplicative update. **(P4)** justifies the term "first-order", and its implication will be seen in the proof of Theorem 1.

The framework of multiplicative updates for the h-divergences is shown in Algorithm 1, where $G_1(\cdot|\cdot)$ and $G_2(\cdot|\cdot)$ denote the first-order surrogate functions of ℓ for $\mathbf{W}$ and $\mathbf{H}$ respectively.

3.2. Construction of First-Order Surrogate Functions and Derivation of Multiplicative Updates

Proposition 1. *Let $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ and $D(\mathbf{V}\|\cdot)$ be a separable h-divergence, then there exist $\zeta_{\min}, \zeta_{\max} \in \mathbb{R}$, $\zeta_{\min} < \zeta_{\max}$, such that*

$$G(\mathbf{W}|\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}}) \triangleq \sum_{i=1}^{F} \sum_{k=1}^{K} \left[(s_{ik}^+ + \lambda_1) \widetilde{w}_{ik} h\left(\frac{w_{ik}}{\widetilde{w}_{ik}}, \zeta_{\max}\right) + 2\lambda_2 \widetilde{w}_{ik}^2 h\left(\frac{w_{ik}}{\widetilde{w}_{ik}}, \zeta_{\max}\right) - s_{ik}^- \widetilde{w}_{ik} h\left(\frac{w_{ik}}{\widetilde{w}_{ik}}, \zeta_{\min}\right) \right] \tag{7}$$

is a first-order surrogate function of ℓ for $\mathbf{W}$ at $(\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}})$ up to some additive constant (independent of $\mathbf{W}$). Here $\mathbf{S}^+$ and $\mathbf{S}^-$ (both in $\mathbb{R}_+^{F \times K}$) are defined as the sums of positive and unsigned negative terms (cf. [2]) in $\nabla_{\mathbf{W}} D(\mathbf{V}\|\mathbf{W}\widetilde{\mathbf{H}})\big|_{\mathbf{W}=\widetilde{\mathbf{W}}}$ respectively.

Proof Sketch. First we show $G(\mathbf{W}|\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}})$ is a surrogate function, i.e., it satisfies **(P1)** to **(P3)**, by constructing it using the up-merging and down-merging techniques introduced in [25]. Indeed,[5]

$$\zeta_{\max} = \max\{\zeta'_p\}_{p=1}^P \cup \{\operatorname{sgn}(\lambda_1), 2\operatorname{sgn}(\lambda_2)\}, \tag{8}$$

$$\zeta_{\min} = \min\{\zeta'_p\}_{p=1}^P, \tag{9}$$

where for all $p \in [P]$, $\zeta'_p \triangleq 1$ if $\zeta_p \in (0, 1)$ and $\zeta'_p \triangleq \zeta_p$ otherwise. Proving **(P4)** involves verification of $\nabla_{\mathbf{W}} G(\mathbf{W}|\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}})\big|_{\mathbf{W}=\widetilde{\mathbf{W}}} = \nabla_{\mathbf{W}} \ell(\mathbf{W}, \widetilde{\mathbf{H}})\big|_{\mathbf{W}=\widetilde{\mathbf{W}}}$. To show **(P5)**, it suffices to show for any $(i,k) \in [F] \times [K]$ and $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, $\frac{\partial^2}{\partial w_{ik}^2} G(\mathbf{W}|\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}}) > 0$. See Section 5.1 in [26] for the detailed steps in the proof. □

[5] For a nonnegative scalar x, $\operatorname{sgn}(x) \triangleq 1$ if $x > 0$ and $\operatorname{sgn}(x) \triangleq 0$ otherwise.

Algorithm 1 General Framework for Multiplicative Updates

Input: Data matrix $\mathbf{V}$, latent dimension K, regularization weights $\{\lambda_1, \widetilde{\lambda}_1\} \subseteq \mathbb{R}_{++}$ and $\{\lambda_2, \widetilde{\lambda}_2\} \subseteq \mathbb{R}_+$

Initialize basis matrix $\mathbf{W}^0$, coefficient matrix $\mathbf{H}^0$ and iteration index $t := 0$

Repeat

$$\mathbf{W}^{t+1} := \operatorname*{arg\,min}_{\mathbf{W} \in \mathbb{R}_+^{F \times K}} G_1(\mathbf{W}|\mathbf{W}^t, \mathbf{H}^t) \tag{11}$$

$$\mathbf{H}^{t+1} := \operatorname*{arg\,min}_{\mathbf{H} \in \mathbb{R}_+^{K \times N}} G_2(\mathbf{H}|\mathbf{W}^{t+1}, \mathbf{H}^t) \tag{12}$$

$$t := t + 1 \tag{13}$$

Until some convergence criterion is met

Output: Learned basis matrix $\overline{\mathbf{W}}$ and coefficient matrix $\overline{\mathbf{H}}$

By setting $\nabla_{\mathbf{W}} G(\mathbf{W}|\widetilde{\mathbf{W}}, \widetilde{\mathbf{H}})$ to zero, we obtain the corresponding multiplicative updates.

Proposition 2. *Let $\mathbf{V}$, $D(\mathbf{V}\|\cdot)$, $\zeta_{\max}$, $\zeta_{\min}$, $\mathbf{S}^+$ and $\mathbf{S}^-$ be given as in Proposition 1. For any $(i,k) \in [F] \times [K]$, the multiplicative update corresponding to* (11) *in Algorithm 1 admits the form*[6]

$$w_{ik} := \widetilde{w}_{ik} \left(\frac{s_{ik}^-}{s_{ik}^+ + 2\lambda_2 \widetilde{w}_{ik} + \lambda_1} \right)^{1/(\zeta_{\max} - \zeta_{\min})}. \tag{10}$$

Remark 4. In (10), the presence of a small $\lambda_1 > 0$ ensures numerical stability, i.e., it prevents division by extremely small numbers (which may lead to numerical overflow). As a popular heuristic [10], a small positive number is usually added to the denominator of the multiplicative factor artificially. Here we establish the connection between this small number and ℓ_1 regularization *for separable h-divergences*, thereby theoretically justifying this heuristic.[7]

Next, we consider nonseparable h-divergences. By the convexity (or concavity) of $h(\cdot, t)$, (3) is a difference-of-convex (DC) function [31]. Therefore, by using either a first-order Taylor expansion or Jensen's inequality, the nonseparable case can be easily converted to the separable case. Such standard techniques are well-studied in the literature. For details, see [25, 32].

To better illustrate our general multiplicative updates in (10), we employ the family of α-divergences as a concrete example.[8] The details are deferred to Sections 5.2 and 5.3 in [26].

4. CONVERGENCE ANALYSIS

4.1. Preliminaries

Definition 3 (Stationary points of constrained optimization problems)**.** Given a finite-dimensional real Banach space $\mathcal{X}$, a differentiable function $g : \mathcal{X} \to \mathbb{R}$ and a set $\mathcal{K} \subseteq \mathcal{X}$, $x_0 \in \mathcal{K}$ is a stationary point of the constrained optimization problem $\min_{x \in \mathcal{K}} g(x)$ if $\langle \nabla g(x_0), x - x_0 \rangle \geq 0$, for all $x \in \mathcal{K}$.

Define $\mathbf{X} \triangleq \begin{bmatrix} \mathbf{W}^T & \mathbf{H} \end{bmatrix} \in \mathbb{R}_+^{K \times (F+N)}$ and with a slight abuse of notation, we write $\ell(\mathbf{X}) \triangleq \ell(\mathbf{W}, \mathbf{H})$. Thus by Definition 3,

[6] Here $\widetilde{\mathbf{W}}$ (resp. $\widetilde{\mathbf{H}}$) denotes the value of basis (resp. coefficient) matrix at the *current* iteration (iteration t), and $\mathbf{W}$ (resp. $\mathbf{H}$) denotes the value of basis (resp. coefficient) matrix at the *next* iteration (iteration $t+1$).

[7] This connection has been observed for some special h-divergences [3, 29], but here we provide a more general and unified discussion.

[8] Both cases $\alpha \neq 0$ and $\alpha = 0$ will be discussed.

we have that $(\overline{\mathbf{W}}, \overline{\mathbf{H}})$ is a stationary point of (4) if and only if $\left\langle \nabla_{\mathbf{X}} \ell(\overline{\mathbf{X}}), \mathbf{X} - \overline{\mathbf{X}} \right\rangle \geq 0$, for any $\mathbf{X} \in \mathbb{R}_+^{K \times (F+N)}$, where $\overline{\mathbf{X}} \triangleq [\,\overline{\mathbf{W}}^T \ \overline{\mathbf{H}}\,]$. In particular, this is true if

$$\left\langle \nabla_{\mathbf{W}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}), \mathbf{W} - \overline{\mathbf{W}} \right\rangle \geq 0, \ \forall\, \mathbf{W} \in \mathbb{R}_+^{F \times K}, \tag{14}$$

$$\left\langle \nabla_{\mathbf{H}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}), \mathbf{H} - \overline{\mathbf{H}} \right\rangle \geq 0, \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \tag{15}$$

Remark 5. In some previous works (e.g., [15]), stationary points are defined in terms of KKT conditions, i.e.,[9]

$$\overline{\mathbf{W}} \geq 0, \quad \overline{\mathbf{H}} \geq 0 \tag{16}$$

$$\nabla_{\mathbf{W}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}) \geq 0, \quad \nabla_{\mathbf{H}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}) \geq 0 \tag{17}$$

$$\overline{\mathbf{W}} \odot \nabla_{\mathbf{W}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}) = 0, \quad \overline{\mathbf{H}} \odot \nabla_{\mathbf{H}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}}) = 0. \tag{18}$$

Since both $\overline{\mathbf{W}}$ and $\overline{\mathbf{H}}$ are nonnegative, it is easy to show these three conditions are equivalent to (14) and (15). In our analysis, we will use (14) and (15) for convenience.

Definition 4 (Convergence of a sequence to a set)**.** Given a finite-dimensional real Banach space $\mathcal{X}$, a sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathcal{X}$ is said to converge to a set $\mathcal{A} \subseteq \mathcal{X}$, denoted as $x_n \to \mathcal{A}$, if $\lim_{n\to\infty} \inf_{a \in \mathcal{A}} \|x_n - a\| = 0$.

Lemma 1 ([33])**.** *Let $\mathcal{X}$, $\{x_n\}_{n=1}^{\infty}$ and $\mathcal{A}$ be given in Definition 4. $x_n \to \mathcal{A}$ if and only if every limit point of $\{x_n\}_{n=1}^{\infty}$ lies in $\mathcal{A}$.*

4.2. Main Result

Theorem 1. *For any $\mathbf{V} \in \mathbb{R}_+^{F \times N}$, $K \in \mathbb{N}$, $\{\lambda_1, \widetilde{\lambda}_1\} \subseteq \mathbb{R}_{++}$ and $\{\lambda_2, \widetilde{\lambda}_2\} \subseteq \mathbb{R}_+$, the sequence of iterates $\{(\mathbf{W}^t, \mathbf{H}^t)\}_{t=1}^{\infty}$ generated by Algorithm 1 converges to the set of stationary points of* (4)*.*

Proof. First, by Lemma 1, it suffices to show every limit point of $\{(\mathbf{W}^t, \mathbf{H}^t)\}_{t=1}^{\infty}$ is a stationary point of (4). Since $\{\lambda_1, \widetilde{\lambda}_1\} \subseteq \mathbb{R}_{++}$, $(\mathbf{W}, \mathbf{H}) \mapsto \ell(\mathbf{W}, \mathbf{H})$ is jointly coercive [33] in $(\mathbf{W}, \mathbf{H})$. In addition, the continuous differentiability of $h(\cdot, t)$ implies the joint continuous differentiability of $(\mathbf{W}, \mathbf{H}) \mapsto \ell(\mathbf{W}, \mathbf{H})$ in $(\mathbf{W}, \mathbf{H})$. Hence

$$\mathcal{S}_0 \triangleq \left\{ (\mathbf{W}, \mathbf{H}) \in \mathbb{R}_+^{F \times K} \times \mathbb{R}_+^{K \times N} \mid \ell(\mathbf{W}, \mathbf{H}) \leq \ell(\mathbf{W}^0, \mathbf{H}^0) \right\}$$

is compact. Since the sequence $\{\ell(\mathbf{W}^t, \mathbf{H}^t)\}_{t=1}^{\infty}$ is nonincreasing, $\{(\mathbf{W}^t, \mathbf{H}^t)\}_{t=1}^{\infty} \subseteq \mathcal{S}_0$. By the compactness of $\mathcal{S}_0$, $\{(\mathbf{W}^t, \mathbf{H}^t)\}_{t=1}^{\infty}$ has at least one limit point. Pick any such limit point and denote it as $(\mathring{\mathbf{W}}, \mathring{\mathbf{H}})$. For convenience, define

$$\mathbf{Z}^t \triangleq \begin{cases} \left(\mathbf{W}^{t/2}, \mathbf{H}^{t/2}\right), & t \text{ even} \\ \left(\mathbf{W}^{\lfloor t/2 \rfloor + 1}, \mathbf{H}^{\lfloor t/2 \rfloor}\right), & t \text{ odd} \end{cases} \quad \text{and} \quad \mathring{\mathbf{Z}} \triangleq \left(\mathring{\mathbf{W}}, \mathring{\mathbf{H}}\right).$$

Then there exists a subsequence $\left\{\mathbf{Z}^{t_j}\right\}_{j=1}^{\infty}$ that converges to $\mathring{\mathbf{Z}} \in \mathcal{S}_0$ and $\{t_j\}_{j=1}^{\infty}$ are all even. Moreover, there exists a subsequence of $\left\{\mathbf{Z}^{t_j - 1}\right\}_{j=1}^{\infty}$, denoted as $\left\{\mathbf{Z}^{t_{j_i} - 1}\right\}_{i=1}^{\infty}$, such that $\mathbf{Z}^{t_{j_i} - 1}$ converges to (possibly) some other limit point $\mathring{\mathbf{Z}}' \triangleq (\mathring{\mathbf{W}}', \mathring{\mathbf{H}}')$ as $i \to \infty$.

Next we show $\mathring{\mathbf{Z}} = \mathring{\mathbf{Z}}'$. By the update rule (12), we have

$$\mathbf{H}^{t_{j_i}/2} = \underset{\mathbf{H} \in \mathbb{R}_+^{K \times N}}{\arg\min}\ G_2\left(\mathbf{H} | \mathbf{Z}^{t_{j_i} - 1}\right), \ \forall\, i \in \mathbb{N}. \tag{19}$$

Thus for any $i \in \mathbb{N}$,

$$G_2(\mathbf{H}^{t_{j_i}/2} | \mathbf{Z}^{t_{j_i} - 1}) \leq G_2(\mathbf{H} | \mathbf{Z}^{t_{j_i} - 1}), \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \tag{20}$$

By **(P2)**, we also have for any $i \in \mathbb{N}$,

$$\ell(\mathbf{Z}^{t_{j_i}/2}) \triangleq \ell(\mathbf{W}^{t_{j_i}/2}, \mathbf{H}^{t_{j_i}/2}) \leq G_2(\mathbf{H}^{t_{j_i}/2} | \mathbf{Z}^{t_{j_i} - 1}). \tag{21}$$

Taking $i \to \infty$ on both sides of (20) and (21), we have

$$\ell(\mathring{\mathbf{Z}}) \leq G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}') \leq G_2(\mathbf{H} | \mathring{\mathbf{Z}}'), \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}, \tag{22}$$

by the joint continuity of $G_2(\cdot|\cdot)$ in both arguments in **(P3)**. Thus

$$\mathring{\mathbf{H}} = \underset{\mathbf{H} \in \mathbb{R}_+^{K \times N}}{\arg\min}\ G_2(\mathbf{H} | \mathring{\mathbf{Z}}'). \tag{23}$$

Taking $\mathbf{H} = \mathring{\mathbf{H}}'$ in (22), we have

$$\ell(\mathring{\mathbf{Z}}) \leq G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}') \leq G_2(\mathring{\mathbf{H}}' | \mathring{\mathbf{Z}}') \triangleq \ell(\mathring{\mathbf{Z}}'). \tag{24}$$

Since $\{\ell(\mathbf{Z}^t)\}_{t=1}^{\infty}$ converges (to a unique limit point), we have $\ell(\mathring{\mathbf{Z}}) = \ell(\mathring{\mathbf{Z}}')$. This implies that $\ell(\mathring{\mathbf{Z}}) = G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}')$. Then for any $\mathbf{H} \in \mathbb{R}_+^{K \times N}$,

$$G_2(\mathring{\mathbf{H}}' | \mathring{\mathbf{Z}}') = \ell(\mathring{\mathbf{Z}}') = \ell(\mathring{\mathbf{Z}}) = G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}') \leq G_2(\mathbf{H} | \mathring{\mathbf{Z}}'). \tag{25}$$

This implies that

$$\mathring{\mathbf{H}}' = \underset{\mathbf{H} \in \mathbb{R}_+^{K \times N}}{\arg\min}\ G_2(\mathbf{H} | \mathring{\mathbf{Z}}'). \tag{26}$$

Combining (23) and (26), by the strictly convexity of $G_2(\cdot|\mathring{\mathbf{Z}}')$ in **(P5)**, $\mathring{\mathbf{H}} = \mathring{\mathbf{H}}'$. By symmetry, we can show $\mathring{\mathbf{W}} = \mathring{\mathbf{W}}'$, hence $\mathring{\mathbf{Z}} = \mathring{\mathbf{Z}}'$. Thus (25) becomes

$$G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}) \leq G_2(\mathbf{H} | \mathring{\mathbf{Z}}), \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \tag{27}$$

Now, the convexity of $G_2(\cdot|\mathring{\mathbf{Z}})$ implies that

$$\left\langle \nabla_{\mathbf{H}} G_2(\mathring{\mathbf{H}} | \mathring{\mathbf{Z}}), \mathbf{H} - \mathring{\mathbf{H}} \right\rangle \geq 0, \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \tag{28}$$

From the first-order property of $G_2(\cdot|\mathring{\mathbf{Z}})$ in **(P4)**, we have

$$\left\langle \nabla_{\mathbf{H}} \ell(\mathring{\mathbf{W}}, \mathring{\mathbf{H}}), \mathbf{H} - \mathring{\mathbf{H}} \right\rangle \geq 0, \ \forall\, \mathbf{H} \in \mathbb{R}_+^{K \times N}. \tag{29}$$

Similarly, we also have

$$\left\langle \nabla_{\mathbf{W}} \ell(\mathring{\mathbf{W}}, \mathring{\mathbf{H}}), \mathbf{W} - \mathring{\mathbf{W}} \right\rangle \geq 0, \ \forall\, \mathbf{W} \in \mathbb{R}_+^{F \times K}. \tag{30}$$

The variational inequalities (29) and (30) together show that $(\mathring{\mathbf{W}}, \mathring{\mathbf{H}})$ is a stationary point of (4). □

Remark 6. We now provide some intuitions of the proof. We first use the positivity of $\lambda_1, \widetilde{\lambda}_1$ to assert that $\mathcal{S}_0$ is compact. This allows us to extract convergent subsequences. The most crucial step (27) states that at an arbitrary limit point of $\{\mathbf{Z}^t\}_{t=1}^{\infty}$, denoted as $\mathring{\mathbf{Z}} = (\mathring{\mathbf{W}}, \mathring{\mathbf{H}})$, $\mathring{\mathbf{H}}$ serves as a minimizer of $G_2(\cdot|\mathring{\mathbf{Z}})$ over $\mathbb{R}_+^{K \times N}$. By symmetry, $\mathring{\mathbf{W}}$ also serves as a minimizer of $G_1(\cdot|\mathring{\mathbf{Z}})$ over $\mathbb{R}_+^{F \times K}$. In the single-block case, this idea is fairly intuitive. However, to prove (27) in the double-block case, we consider two subsequences $\{\mathbf{Z}^{t_{j_i}}\}_{i=1}^{\infty}$ and $\{\mathbf{Z}^{t_{j_i} - 1}\}_{i=1}^{\infty}$. In each sequence, only $\mathbf{W}$ or $\mathbf{H}$ is updated. Then we show these two sequences converge to the same limit point. This implies the Gauss-Seidel minimization procedure [33] in the double-block case is essentially the same as the minimization in the single-block case. The claim then follows immediately.

Acknowledgements: The authors would like to sincerely thank Zhirong Yang for many useful comments on the manuscript.

[9] Here we use $\nabla_{\mathbf{W}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}})$ and $\nabla_{\mathbf{H}} \ell(\overline{\mathbf{W}}, \overline{\mathbf{H}})$ to denote $\nabla_{\mathbf{W}} \ell(\mathbf{W}, \overline{\mathbf{H}})\big|_{\mathbf{W} = \overline{\mathbf{W}}}$ and $\nabla_{\mathbf{H}} \ell(\overline{\mathbf{W}}, \mathbf{H})\big|_{\mathbf{H} = \overline{\mathbf{H}}}$ respectively.

5. REFERENCES

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

[2] Daniel D. Lee and H. Sebastian Seung, "Algorithms for nonnegative matrix factorization," in *Proc. NIPS*, Denver, USA, Dec. 2000, pp. 556–562.

[3] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, 2011.

[4] Jingu Kim and Haesun Park, "Toward faster nonnegative matrix factorization: A new algorithm and comparisons," in *Proc. ICDM*, Pisa, Italy, Dec. 2008, pp. 353–362.

[5] Chih-Jen Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, Oct. 2007.

[6] Hyunsoo Kim and Haesun Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. A.*, vol. 30, no. 2, pp. 713–730, 2008.

[7] Yangyang Xu, Wotao Yin, Zaiwen Wen, and Yin Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers Math. China*, vol. 7, pp. 365–384, 2012.

[8] A Cichocki, H-K Lee, Y-D Kim, and S Choi, "Nonnegative matrix factorization with alpha-divergence," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1433–1440, 2008.

[9] A Cichocki and S Amari, "Families of Alpha-Beta-and Gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[10] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, 2011.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[12] Stephen A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. Optim.*, vol. 20, no. 3, pp. 1364–1377, 2009.

[13] David Donoho and Victoria Stodden, "When does nonnegative matrix factorization give correct decomposition into parts?," in *Proc. NIPS*, Vancouver, Canada, Dec. 2004, pp. 1141–1148.

[14] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra, "Computing a nonnegative matrix factorization – provably," in *Proc. STOC*, New York, New York, USA, May 2012, pp. 145–162.

[15] Chih-Jen Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, 2007.

[16] Nicolas Gillis and François Glineur, "Nonnegative factorization and the maximum edge biclique problem," arXiv:0810.4225, 2008.

[17] Ryota Hibi and Norikazu Takahashi, "A modified multiplicative update algorithm for euclidean distance-based nonnegative matrix factorization and its global convergence," in *Proc. ICONIP*, Shanghai, China, Nov. 2011, pp. 655–662.

[18] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 21, no. 12, pp. 1869–1881, Dec. 2010.

[19] Hyunsoo Kim and Haesun Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, pp. 1495–1502, 2007.

[20] D. Hajinezhad, T. H. Chang, X. Wang, Q. Shi, and M. Hong, "Nonnegative matrix factorization using admm: Algorithm and convergence analysis," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 4742–4746.

[21] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorizations as probabilistic inference in composite models," in *Proc. EUSIPCO*, Glasgow, UK, Aug. 2009, pp. 1913–1917.

[22] Renbo Zhao, Vincent Y. F. Tan, and Huan Xu, "Online nonnegative matrix factorization with general divergences," arXiv:1608.00075, 2016.

[23] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.

[24] M. Hong, M. Razaviyayn, Z. Q. Luo, and J. S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.

[25] Z. Yang and E. Oja, "Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1878–1891, Dec. 2011.

[26] Renbo Zhao and Vincent Y. F. Tan, "A unified convergence analysis of the multiplicative update algorithm for nonnegative matrix factorization," arXiv:1609.00951, 2016.

[27] Karthik Devarajan, Guoli Wang, and Nader Ebrahimi, "A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing," *Mach. Learn.*, vol. 99, no. 1, pp. 137–163, 2015.

[28] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, vol. 67, no. 2, pp. 301–320, 2005.

[29] P. O. Hoyer, "Non-negative sparse coding," in *Proc. NNSP*, Valais, Switzerland, Sep. 2002, pp. 557–565.

[30] V. Paul Pauca, J. Piper, and Robert J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," *Linear Algebra Appl.*, vol. 416, no. 1, pp. 29 – 47, 2006.

[31] Ivan Ginchev and Denitza Gintcheva, "Characterization and recognition of D.C. functions," *J. Glob. Optim.*, vol. 57, no. 3, pp. 633–647, 2013.

[32] Julien Mairal, "Incremental majorization-minimization optimization with application to large-scale machine learning," *SIAM J. Optim.*, vol. 25, no. 2, pp. 829–855, 2015.

[33] Dimitri P. Bertsekas, *Nonlinear Programming*, Athena Scitific, 1999.