CONVERGENCE-GUARANTEED MULTIPLICATIVE ALGORITHMS FOR NONNEGATIVE MATRIX FACTORIZATION WITH β-DIVERGENCE

Masahiro Nakano[†]*, Hirokazu Kameoka*[‡]*, Jonathan Le Roux*[‡]*, Yu Kitano*[†]*, Nobutaka Ono*[†]*, Shigeki Sagayama*[†]

[†]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[‡]NTT Communication Science Laboratories, NTT Corporation,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

This paper presents a new multiplicative algorithm for nonnegative matrix factorization with β-divergence. The derived update rules have a similar form to those of the conventional multiplicative algorithm, only differing through the presence of an exponent term depending on β. The convergence is theoretically proven for any real-valued β based on the auxiliary function method. The convergence speed is experimentally investigated in comparison with previous works.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [1] has recently become a very popular technique in signal processing, and has been succesfully applied to various problems such as source separation [2, 3, 4], feature extraction, music transcription [5] or dimension reduction. Given a nonnegative matrix $\mathbf{Y}$, the goal of NMF is to find two nonnegative matrices $\mathbf{H}$ and $\mathbf{U}$ such that $\mathbf{Y} \approx \mathbf{HU}$. To measure how close $\mathbf{Y}$ and $\mathbf{HU}$ are, the Euclidean (EUC) distance, the generalized Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence are often chosen. The three of them are enclosed in the more general framework of β-divergence [6, 7]. Since the choice of an appropriate divergence depends on the application and the data [2, 8, 9], an algorithm stable for a wide range of β is desired.

Multiplicative gradient descent [7, 10] is one of the most popular approaches for NMF with β-divergence. A proof of the convergence of the algorithms for $\beta = 2$ (EUC distance) and $\beta = 1$ (KL divergence) was given in [10], and it has been extended to $1 \leq \beta \leq 2$ [11]. However, convergence for $\beta < 1$ and $\beta > 2$ remains to be proven. A generalized multiplicative algorithm, which introduces an exponent step size, has also recently been proposed in [12], discussing in particular the local and stable convergence, in the sense of Lyapunov's theory, of this algorithm when initialized in a given neighborhood of a local minimum. However, convergence is not guaranteed in general.

Another way to derive optimization algorithms for NMF is through a statistical approach. The minimization of a β-divergence can indeed be shown to be equivalent, for specific βs, to a Maximum-Likelihood (ML) estimation problem, due to the reproductive properties of some probabilistic distributions. Update equations for NMF with EUC distance ($\beta = 2$), KL divergence ($\beta = 1$) and IS divergence ($\beta = 0$) have been obtained under this framework in [13] based on the expectation maximization (EM) algorithm. Although convergence to a stationary point is then guaranteed, this approach is currently limited to the cases $\beta = 0$, 1, and 2.

This paper proposes a new multiplicative algorithm for NMF with β-divergence, for which the monotonic decrease of the objective function at each iteration is theoretically guaranteed for all β. The derivation of this algorithm is based on the careful design of a so-called auxiliary function [10] for each term of the objective function.

The remainder of this paper is organized as follows. We will first briefly review the formulation of NMF with β-divergence in Section 2, and give a survey of the previous methods in Section 3. We will then derive the proposed multiplicative algorithm in Section 4, and finally present in Section 5 basic experimental results validating our method and comparing it to previous works.

2. NONNEGATIVE MATRIX FACTORIZATION WITH β-DIVERGENCE

Given a matrix $\mathbf{Y} = (Y_{\omega,t})_{\Omega \times T} \in \mathbb{R}^{\geq 0, \Omega \times T}$ and an integer K, NMF is the problem of finding a factorization:

$$\mathbf{Y} \approx \mathbf{HU}, \tag{1}$$

where $\mathbf{H} = (H_{\omega,k})_{\Omega \times K} \in \mathbb{R}^{\geq 0, \Omega \times K}$ and $\mathbf{U} = (U_{k,t})_{K \times T} \in \mathbb{R}^{\geq 0, K \times T}$ are nonnegative matrices of dimensions $\Omega \times K$ and $K \times T$, respectively. K is usually chosen such that $\Omega K + KT \ll \Omega T$, hence reducing the data dimension. This

problem can be formulated as the minimization of an objective function

$$D(\mathbf{Y} \mid \mathbf{HU}) = \sum_{\omega,t} d\Big(Y_{\omega,t} \mid \sum_k H_{\omega,k}U_{k,t}\Big), \qquad (2)$$

where d is a scalar divergence.

A common way to measure how close $\mathbf{Y}$ and $\mathbf{HU}$ are is to use a so-called β-divergence [14], defined by

$$d_\beta(y|x) = \begin{cases} \dfrac{y^\beta}{\beta(\beta-1)} + \dfrac{x^\beta}{\beta} - \dfrac{yx^{\beta-1}}{\beta-1} & \beta \in \mathbb{R}\backslash\{0,1\} \\ y(\log y - \log x) + (x-y) & \beta = 1 \\ \dfrac{y}{x} - \log\dfrac{y}{x} - 1 & \beta = 0 \end{cases}.$$

It can be shown to be continuous in terms of β through the following identities:

$$\begin{aligned} \lim_{\beta\to 1} d_\beta(y|x) &= \lim_{\beta\to 1}\left(y\frac{y^{\beta-1}-x^{\beta-1}}{\beta-1} + \frac{x^\beta - y^\beta}{\beta}\right) \\ &= y(\log y - \log x) + (x-y), \\ \lim_{\beta\to 0} d_\beta(y|x) &= \lim_{\beta\to 0}\left(y\frac{x^{\beta-1}}{1-\beta} - \frac{y^\beta - x^\beta}{\beta}\right) + \frac{y^\beta}{\beta-1} \\ &= \frac{y}{x} - \log\frac{y}{x} - 1. \end{aligned}$$

The choice of β should be driven by the type of data being analyzed and the application considered. In the NMF-related literature, $\beta = 1$ is for example often used for sound source separation [2], while $\beta = 0.5$ is used for the estimation of time-frequency activations [8] and $\beta = 0$ for multipitch estimation [9]. How to choose β for multipitch estimation and musical source separation is discussed in [5] and [3], respectively.

3. CONVENTIONAL ALGORITHMS

3.1. Multiplicative algorithms

The multiplicative gradient descent approach [10, 7] consists in updating each parameter by multiplying its value at the previous iteration by a certain coefficient. Here, let θ denote the set of all parameters $\{(H_{\omega,k})_{\Omega\times K}, (U_{k,t})_{K\times T}\}$. The derivative with respect to $H_{\omega,k}$ of the objective function $\mathcal{J}_\beta(\theta) = \sum_{\omega,t} d_\beta(Y_{\omega,t}|\sum_k H_{\omega,k}U_{k,t})$ is

$$\begin{aligned} \frac{\partial \mathcal{J}_\beta(\theta)}{\partial H_{\omega,k}} = &\sum_t \left(\sum_k H_{\omega,k}U_{k,t}\right)^{\beta-1} U_{k,t} \\ &- \sum_t Y_{\omega,t}\left(\sum_k H_{\omega,k}U_{k,t}\right)^{\beta-2} U_{k,t}. \end{aligned} \qquad (3)$$

Considering the following simple additive update for $H_{\omega,k}$,

$$H_{\omega,k} \leftarrow H_{\omega,k} - \eta_{\omega,k}\frac{\partial \mathcal{J}_\beta(\theta)}{\partial H_{\omega,k}}, \qquad (4)$$

the objective function will be decreasing if the coefficients $\eta_{\omega,k}$ are all set equal to a sufficiently small positive number, as this corresponds to the conventional gradient descent. If we now set

$$\eta_{\omega,k} = \frac{H_{\omega,k}}{\sum_t \left(\sum_k H_{\omega,k}U_{k,t}\right)^{\beta-1} U_{k,t}}, \qquad (5)$$

we obtain the following update rule for $H_{\omega,k}$:

$$H_{\omega,k} \leftarrow H_{\omega,k}\frac{\sum_t Y_{\omega,t}\left(\sum_k H_{\omega,k}U_{k,t}\right)^{\beta-2} U_{k,t}}{\sum_t \left(\sum_k H_{\omega,k}U_{k,t}\right)^{\beta-1} U_{k,t}}. \qquad (6)$$

A similar update rule can be obtained for $U_{k,t}$. Altogether, the algorithm can be summarized as follows:

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \frac{(\mathbf{Y}\cdot(\mathbf{HU})^{\beta-2})\mathbf{U}^T}{(\mathbf{HU})^{\beta-1}\mathbf{U}^T}, \qquad (7)$$

$$\mathbf{U} \leftarrow \mathbf{U} \cdot \frac{\mathbf{H}^T(\mathbf{Y}\cdot(\mathbf{HU})^{\beta-2})}{\mathbf{H}^T(\mathbf{HU})^{\beta-1}}, \qquad (8)$$

where the symbol $\cdot$ and the fraction bar denote entrywise matrix product and division respectively, and the exponentiations are also performed entrywise. Nonnegativity of the parameters is preserved through the updates, provided they are initialized with nonnegative values.

The convergence of the conventional multiplicative updates have been proven only for $1 \le \beta \le 2$, i.e, when $d_\beta(y|x)$ is convex w.r.t. x [10, 11].

3.2. EM-based algorithms

In [13], NMF with EUC distance ($\beta = 2$), KL divergence ($\beta = 1$) and IS divergence ($\beta = 0$) is shown to be implicit in the following generative model of superimposed components,

$$Y_{\omega,t} = \sum_k C_{\omega,t,k}. \qquad (9)$$

The components $C_{\omega,t,k}$ act as latent variables and may be used as complete data in the EM algorithm. We briefly review here the update rules obtained through this method successively for $\beta = 2$, $\beta = 1$ and $\beta = 0$.

3.2.1. EUC-NMF

NMF with EUC distance ($\beta = 2$) is equivalent to constrained ML estimation for the generative model (9) with

$$C_{\omega,t,k} \sim \left(\frac{2\pi\sigma^2}{K}\right)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(C_{\omega,t,k} - H_{\omega,k}U_{k,t})^2\frac{K}{\sigma^2}\Big). \qquad (10)$$

"Constrained" means here that the parameters $\mathbf{H}$ and $\mathbf{U}$ are estimated under the assumption that they are non-negative.

Thereby, the following update rules are obtained based on the EM algorithm:

$$H_{\omega,k} = \left[\frac{\sum_t U_{k,t} \left(\frac{1}{K}(Y_{\omega,t} - W_{\omega,t}) + H_{\omega,k} U_{k,t} \right)}{\sum_t U_{k,t}^2} \right]_+, \quad (11)$$

$$U_{k,t} = \left[\frac{\sum_\omega H_{\omega,k} \left(\frac{1}{K}(Y_{\omega,t} - W_{\omega,t}) + H_{\omega,k} U_{k,t} \right)}{\sum_t H_{\omega,k}^2} \right]_+, \quad (12)$$

where $[x]_+ = \max\{x, 0\}$ and

$$W_{\omega,t} = \sum_k H_{\omega,k} U_{k,t}. \quad (13)$$

3.2.2. KL-NMF

NMF with the generalized KL divergence ($\beta = 1$) is equivalent to ML estimation for the model (9) with

$$C_{\omega,t,k} \sim \exp(-H_{\omega,k} U_{k,t}) \frac{(H_{\omega,k} U_{k,t})^{C_{\omega,t,k}}}{\Gamma(C_{\omega,t,k} + 1)}, \quad (14)$$

where Γ denotes the Gamma function. Thereby, the following update rules are obtained based on the EM algorithm:

$$H_{\omega,k} = H_{\omega,k} \frac{\sum_t U_{k,t} (Y_{\omega,t} / \sum_k H_{\omega,k} U_{k,t})}{\sum_t U_{k,t}}, \quad (15)$$

$$U_{k,t} = U_{k,t} \frac{\sum_\omega H_{\omega,k} (Y_{\omega,t} / \sum_k H_{\omega,k} U_{k,t})}{\sum_t H_{\omega,k}}, \quad (16)$$

which coincide with the classical multiplicative updates.

3.2.3. IS-NMF

NMF with IS divergence ($\beta = 0$) is equivalent to constrained ML estimation for the model (9) with

$$C_{\omega,t,k} \sim |\pi H_{\omega,k} U_{k,t}|^{-1} \exp\left(-|C_{\omega,t,k}|^2 |H_{\omega,k} U_{k,t}|^{-1} \right). \quad (17)$$

Thereby, the following update rules are obtained based on the EM algorithm:

$$H_{\omega,k} = \frac{1}{T} \sum_t \frac{\mu_{\omega,t,k}^2 + \nu_{\omega,t,k}}{U_{k,t}}, \quad (18)$$

$$U_{k,t} = \frac{1}{\Omega} \sum_\omega \frac{\mu_{\omega,t,k}^2 + \nu_{\omega,t,k}}{H_{\omega,k}}, \quad (19)$$

with

$$\mu_{\omega,t,k} = \frac{H_{\omega,k} U_{k,t}}{\sum_k H_{\omega,k} U_{k,t}} Y_{\omega,t}, \quad (20)$$

$$\nu_{\omega,t,k} = \frac{H_{\omega,k} U_{k,t}}{\sum_k H_{\omega,k} U_{k,t}} \sum_{l \neq k} H_{\omega,l} U_{l,t}. \quad (21)$$

4. NEW MULTIPLICATIVE ALGORITHMS

We consider the following optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \mathcal{J}_\beta(\theta) = \textstyle\sum_{\omega,t} d_\beta \left(Y_{\omega,t} | \sum_k H_{\omega,k} U_{k,t} \right) \\ \text{subject to} \quad & \forall \omega, k, H_{\omega,k} \geq 0, \ \forall k, t, U_{k,t} \geq 0, \end{aligned}$$

where θ denotes the set $\{(H_{\omega,k})_{\Omega \times K}, (U_{k,t})_{K \times T}\}$ of all parameters. The main result of this paper can be summarized in the following

Theorem 1. *The objective function $\mathcal{J}_\beta(\theta)$ is non-increasing under the following updates:*

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \left(\frac{(\mathbf{Y} \cdot (\mathbf{HU})^{\beta-2}) \mathbf{U}^T}{(\mathbf{HU})^{\beta-1} \mathbf{U}^T} \right)^{\varphi(\beta)}, \quad (22)$$

$$\mathbf{U} \leftarrow \mathbf{U} \cdot \left(\frac{\mathbf{H}^T (\mathbf{Y} \cdot (\mathbf{HU})^{\beta-2})}{\mathbf{H}^T (\mathbf{HU})^{\beta-1}} \right)^{\varphi(\beta)}, \quad (23)$$

with

$$\varphi(\beta) = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \leq \beta \leq 2) \\ 1/(\beta - 1) & (\beta > 2) \end{cases}. \quad (24)$$

The proof of the above theorem will be based on the auxiliary function approach, similarly to [10]. In the following, we first explain the principle of the auxiliary function method. Next, we introduce Lemma 1 and Lemma 2 which will be useful for the construction of an auxiliary function. Lemma 3 gives an auxiliary function for the objective function $\mathcal{J}_\beta(\theta)$. Finally, Theorem 1 is proven based on the principle of the auxiliary function method and Lemma 3.

Let us first briefly review the concept of this approach. Let $G(\theta)$ denote an objective function to be minimized w.r.t. a parameter θ. A function $G^+(\theta, \hat{\theta})$ which satisfies

$$G(\theta) = \min_{\hat{\theta}} G^+(\theta, \hat{\theta}) \quad (25)$$

is then called an auxiliary function for $G(\theta)$, and $\hat{\theta}$ an auxiliary variable. The function $G(\theta)$ is non-increasing through the following iterative update rules:

$$\hat{\theta}^{(s+1)} \leftarrow \mathrm{argmin}_{\hat{\theta}} G^+(\theta^{(s)}, \hat{\theta}), \quad (26)$$

$$\theta^{(s+1)} \leftarrow \mathrm{argmin}_{\theta} G^+(\theta, \hat{\theta}^{(s+1)}), \quad (27)$$

where $\hat{\theta}^{(s+1)}$ and $\theta^{(s+1)}$ denote the updated values of $\hat{\theta}$ and θ after the s-th step. Then by construction, we have

$$\begin{aligned} G(\theta^{(s)}) &= G^+(\theta^{(s)}, \hat{\theta}^{(s)}) \\ &\geq G^+(\theta^{(s)}, \hat{\theta}^{(s+1)}) \\ &\geq G^+(\theta^{(s+1)}, \hat{\theta}^{(s+1)}) = G(\theta^{(s+1)}). \end{aligned} \quad (28)$$

This proves that $G(\theta)$ is non-increasing. By iteratively updating θ and $\hat{\theta}$, $G(\theta)$ will thus converge to a stationary point.

The auxiliary function will be suitably designed depending on the value of β. For β such that $1 \leq \beta \leq 2$ [10, 11], such an auxiliary function can be constructed thanks to the following lemma, refered to as Jensen's inequality:

Lemma 1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. If $\lambda_k (k = 1, 2, \cdots, K)$ satisfies $\forall k, \lambda_k > 0$ and $\sum_k \lambda_k = 1$, then for $x_k (k = 1, 2, \cdots, K) \in \mathbb{R}$,*

$$f\big(\sum_k x_k\big) \leq \sum_k \lambda_k f\big(\frac{x_k}{\lambda_k}\big). \tag{29}$$

Equality holds when $\lambda_k = x_k / \sum_k x_k$.

Note that minimizing the auxiliary function obtained through the above lemma then leads to update equations which are none other than the classical multiplicative update equations. However, Jensen's inequality cannot be applied to $\beta < 1$ and $\beta > 2$, because $d_\beta(y|x)$ is then not convex w.r.t. x. We are going to alleviate this problem by decomposing the objective function into several terms which are going to be either convex or concave depending on the value of β, and then use adequate inequalities to build an auxiliary function. Indeed, if we write the objective function as

$$\mathcal{J}_\beta(\theta) = \frac{1}{\beta(\beta-1)} \sum_{\omega,t} Y_{\omega,t} + \frac{1}{\beta} \sum_{\omega,t} \left(\sum_k H_{\omega,k} U_{k,t} \right)^\beta - \frac{1}{\beta - 1} \sum_{\omega,t} Y_{\omega,t} \left(\sum_k H_{\omega,k} U_{k,t} \right)^{\beta-1}, \tag{30}$$

we can see that, with respect to each parameter, the second term is convex for $\beta \geq 1$ and concave for $\beta < 1$, while the third term is convex for $\beta \leq 2$ and concave for $\beta > 2$. To cope with concave terms, as in [15, 4], we shall use the following lemma:

Lemma 2. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a continuously differentiable and concave function. Then, for any point z,*

$$f(x) \leq f'(z)(x - z) + f(z). \tag{31}$$

If β satisfies $\beta \geq 1$, Lemma 1 leads to the following inequality for the second term:

$$\frac{1}{\beta} \left(\sum_k H_{\omega,k} U_{k,t} \right)^\beta \leq \frac{1}{\beta} \sum_k \lambda_{\omega,t,k} \left(\frac{H_{\omega,k} U_{k,t}}{\lambda_{\omega,t,k}} \right)^\beta, \tag{32}$$

where $\forall k, \lambda_{\omega,t,k} \geq 0$ and $\sum_k \lambda_{\omega,t,k} = 1$. Let $\hat{\theta}$ denote the set of auxiliary variables $\{(\lambda_{\omega,t,k})_{\Omega \times T \times K}, (Z_{\omega,t})_{\Omega \times T}\}$, where $Z_{\omega,t} \in \mathbb{R}$ will be used later on. We define $\mathcal{Q}^{(\beta)}_{\omega,t}(\theta, \hat{\theta})$ as the right-hand side of Eq. (32). The equality holds when

$$\lambda_{\omega,t,k} = \frac{H_{\omega,k} U_{k,t}}{\sum_k H_{\omega,k} U_{k,t}}. \tag{33}$$

If β now satisfies $\beta \leq 1$, we apply Lemma 2 and obtain the following inequality for the second term:

$$\frac{1}{\beta} \Big(\sum_k H_{\omega,k} U_{k,t} \Big)^\beta \leq Z_{\omega,t}^{\beta-1} \Big(\sum_k H_{\omega,k} U_{k,t} - Z_{\omega,t} \Big) + \frac{Z_{\omega,t}^\beta}{\beta}. \tag{34}$$

We define $\mathcal{R}^{(\beta)}_{\omega,t}(\theta, \hat{\theta})$ as the right-hand side of Eq. (34). The equality holds when

$$Z_{\omega,t} = \sum_k H_{\omega,k} U_{k,t}. \tag{35}$$

Note that $\mathcal{Q}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}) = \mathcal{R}^{(\beta)}_{\omega,t}(\theta, \hat{\theta})$ when $\beta = 1$.

The following inequalities for the third term can be derived similarly:

$$-\frac{1}{\beta - 1} \Big(\sum_k H_{\omega,k} U_{k,t} \Big)^{\beta-1} \leq \begin{cases} -Q^{(\beta-1)}_{\omega,t}(\theta, \hat{\theta}) & (\beta \leq 2) \\ -R^{(\beta-1)}_{\omega,t}(\theta, \hat{\theta}) & (\beta \geq 2) \end{cases}. \tag{36}$$

The equality holds when $\lambda_{\omega,t,k}$ and $Z_{\omega,t}$ satisfy Eq. (33) and Eq. (35).

We can deduce the following lemma from the above.

Lemma 3. *The function*

$$\mathcal{J}^+_\beta(\theta, \hat{\theta}) = \sum_{\omega,t} \frac{Y_{\omega,t}}{\beta(\beta-1)} + \sum_{\omega,t} \mathcal{S}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}), \tag{37}$$

where

$$\mathcal{S}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}) = \begin{cases} \mathcal{R}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}) - Y_{\omega,t} \mathcal{Q}^{(\beta-1)}_{\omega,t}(\theta, \hat{\theta}) & (\beta < 1) \\ \mathcal{Q}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}) - Y_{\omega,t} \mathcal{Q}^{(\beta-1)}_{\omega,t}(\theta, \hat{\theta}) & (1 \leq \beta \leq 2) \\ \mathcal{Q}^{(\beta)}_{\omega,t}(\theta, \hat{\theta}) - Y_{\omega,t} \mathcal{R}^{(\beta-1)}_{\omega,t}(\theta, \hat{\theta}) & (\beta > 2) \end{cases}, \tag{38}$$

is an auxiliary function for $\mathcal{J}_\beta(\theta)$. $\mathcal{J}^+_\beta(\theta, \hat{\theta})$ is minimized w.r.t. $\hat{\theta}$ when $\hat{\theta}$ satisfies Eq. (33) and Eq. (35).

Proof of Lemma 3. Eq. (32) and Eq. (34) show that

$$\mathcal{J}_\beta(\theta) \leq \mathcal{J}^+_\beta(\theta, \hat{\theta}). \tag{39}$$

The equality holds when $\hat{\theta}$ satisfies Eq. (33) and Eq. (35). Thus, Eq. (33) and Eq. (35) minimizes $\mathcal{J}^+_\beta(\theta, \hat{\theta})$ w.r.t. $\hat{\theta}$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Lemma 3 gives us an auxiliary function of $\mathcal{J}_\beta(\theta)$. According to the principle of the auxiliary

function method, we need to prove that minimizing $\mathcal{J}_\beta^+(\theta,\hat{\theta})$ w.r.t. θ and $\hat{\theta}$ iteratively lead to the update rules, Eq. (22) and Eq. (23).

First, we focus on minimizing $\mathcal{J}_\beta^+(\theta,\hat{\theta})$ w.r.t. θ. The derivative of $\mathcal{J}_\beta^+(\theta,\hat{\theta})$ w.r.t. $H_{\omega,k}$ is

$$\frac{\partial \mathcal{J}_\beta^+(\theta,\hat{\theta})}{\partial H_{\omega,k}} = \mathcal{V}_\beta(\theta) - \mathcal{W}_\beta(\theta), \tag{40}$$

where

$$\mathcal{V}_\beta(\theta) = \begin{cases} \sum_t Z_{\omega,t}^{\beta-1} U_{k,t} & (\beta < 1) \\ H_{\omega,k}^{\beta-1} \sum_t \lambda_{\omega,t,k}^{1-\beta} U_{k,t}^{\beta} & (\beta \geq 1) \end{cases}, \tag{41}$$

$$\mathcal{W}_\beta(\theta) = \begin{cases} H_{\omega,k}^{\beta-2} \sum_t \lambda_{\omega,t,k}^{2-\beta} Y_{\omega,t} U_{k,t}^{\beta-1} & (\beta \leq 2) \\ \sum_t Z_{\omega,t}^{\beta-2} Y_{\omega,t} U_{k,t} & (\beta > 2) \end{cases}. \tag{42}$$

The second derivative is

$$\frac{\partial^2 \mathcal{J}_\beta^+(\theta,\hat{\theta})}{\partial H_{\omega,k} \partial H_{\omega',k'}} = \{\mathcal{V}'_\beta(\theta) - \mathcal{W}'_\beta(\theta)\} \delta_{\omega,\omega'} \delta_{k,k'}, \tag{43}$$

where $\delta_{i,j}$ is 1 if $i = j$, otherwise 0 and

$$\mathcal{V}'_\beta(\theta) = \begin{cases} 0 & (\beta < 1) \\ (\beta-1) H_{\omega,k}^{\beta-2} \sum_t \lambda_{\omega,t,k}^{1-\beta} U_{k,t}^{\beta} & (\beta \geq 1) \end{cases},$$

$$\mathcal{W}'_\beta(\theta) = \begin{cases} (\beta-2) H_{\omega,k}^{\beta-3} \sum_t \lambda_{\omega,t,k}^{2-\beta} Y_{\omega,t} U_{k,t}^{\beta-1} & (\beta \leq 2) \\ 0 & (\beta > 2) \end{cases}.$$

Thus, $\mathcal{J}_\beta^+(\theta,\hat{\theta})$ is a convex function in $\mathbf{H}$. Setting the first derivative to zero, we obtain the update rule for $H_{\omega,k}$:

$$H_{\omega,k} = \begin{cases} \left(\dfrac{\sum_t \lambda_{\omega,t,k}^{2-\beta} Y_{\omega,t} U_{k,t}^{\beta-1}}{\sum_t Z_{\omega,t}^{\beta-1} U_{k,t}} \right)^{\frac{1}{2-\beta}} & (\beta < 1) \\ \dfrac{\sum_t \lambda_{\omega,t,k}^{2-\beta} Y_{\omega,t} U_{k,t}^{\beta-1}}{\sum_t \lambda_{\omega,t,k}^{1-\beta} U_{k,t}^{\beta}} & (1 \leq \beta \leq 2) \\ \left(\dfrac{\sum_t Z_{\omega,t}^{\beta-2} Y_{\omega,t} U_{k,t}}{\sum_t \lambda_{\omega,t,k}^{1-\beta} U_{k,t}^{\beta}} \right)^{\frac{1}{\beta-1}} & (\beta > 2) \end{cases}. \tag{44}$$

$\mathbf{U}$ can be discussed similarly.

Next, we consider the auxiliary variables $\hat{\theta}$. Eq. (33) and Eq. (35) minimize $\mathcal{J}_\beta^+(\theta,\hat{\theta})$ w.r.t. $\hat{\theta}$. Thus, minimizing Eq. (33) and Eq. (35) into Eq. (44) gives the following update rule:

$$H_{\omega,k} \leftarrow H_{\omega,k} \left(\frac{\sum_t Y_{\omega,t} \left(\sum_k H_{\omega,k} U_{k,t} \right)^{\beta-2} U_{k,t}}{\sum_t \left(\sum_k H_{\omega,k} U_{k,t} \right)^{\beta-1} U_{k,t}} \right)^{\varphi(\beta)}.$$

The update rule for $U_{k,t}$ can be obtained similarly. The update rules for $\mathbf{H}$ and $\mathbf{U}$ can be simply rewritten as Eq. (22) and Eq. (23). □

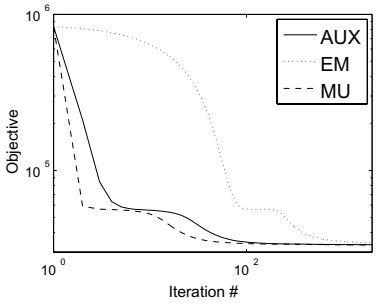

Fig. 1. Evolution in log-log scale of the objective function with $\beta = 0$.

5. EXPERIMENTS

The convergence speed is compared with existing algorithms. The classical multiplicative algorithm for NMF will be denoted as "MU", the EM-based algorithm as "EM", and the proposed multiplicative algorithm based on the auxiliary function method as "AUX". NMF is often applied to analysis of audio signals. Here, we use as input data matrix the magnitude spectrogram of 8 second length music signal (generated from RWC-MDB-P-2001 No.25 [16]) downmixed to monaural and downsampled to 16kHz. It was computed using the short time Fourier transform with a 32 ms long Hanning window and with 16 ms overlap.

We compared the performances of all the algorithms for three different values of β, namely $\beta = 0, 0.5, 2$. Fig. 1 shows the results for $\beta = 0$. In this case, EM is the slowest and MU the fastest, while AUX is slightly slower than MU. As shown in Fig. 2, for $\beta = 2$, AUX (which is then equivalent to MU) is again faster than EM. Finally, the results for $\beta = 0.5$ are shown in Fig. 3. In all cases, MU is slightly faster than AUX, however, the convergence of our algorithm is theoretically proven.

6. CONCLUSIONS

In this paper, we proposed a convergence-guaranteed multiplicative algorithm for NMF with β-divergence. The form of the updates is similar to that of the conventional multiplicative algorithm but with a different exponent term. We confirmed through basic experiments that the proposed algorithms converge faster than EM algorithms. Future work will include the extension of our auxiliary function approach to the derivation of convergence-guaranteed algorithms for constrained NMF methods which involve an objective function as well as penalty terms, for example to promote sparsity or smoothness.

ACKNOWLEDGMENTS

We thank the reviewers for very helpful suggestions and comments.

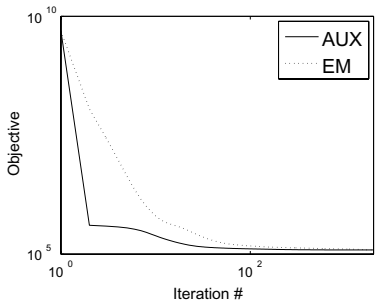

Fig. 2. Evolution in log-log scale of the objective function with $\beta = 2$. MU is equivalent to AUX.

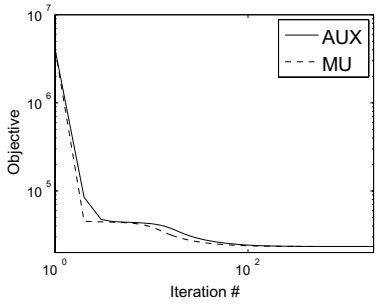

Fig. 3. Evolution in log-log scale of the objective function with $\beta = 0.5$.

7. REFERENCES

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

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1066–1074, Mar. 2007.

[3] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. of Irish Signals and Systems Conference*, 2009.

[4] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3437–3440.

[5] S. A. Raczynski, N. Ono, and S. Sagayama, "Extending nonnegative matrix factorization – a discussion in the context of multipitch frequency estimation of musical signals," in *Proc. of European Signal Processing Conference*, Aug. 2009, pp. 934–938.

[6] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Tokyo Institute of Statistical Mathematics, Tech. Rep., 2001.

[7] A. Cichocki, R. Zdunek, and S. Amari, "Csiszárs divergences for non-negative matrix factorization : Family of new algorithms," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, Mar. 2006, pp. 32–39.

[8] R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 445–448.

[9] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 538–549, 2010.

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

[11] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[12] R. Badeau, N. Bertin, and E. Vincent, "On the stability of multiplicative update algorithms. application to non-negative matrix factorization," in *Telecom ParisTech, Technical report*, 2009.

[13] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorizations as probabilistic inference in composite models," in *Proc. European Signal Processing Conference*, vol. 47, 2009, pp. 1913–1917.

[14] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[15] H. Kameoka, N. Ono, and S. Sagayama, "Auxiliary function approach to parameter estimation of constrained sinusoidal model," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, Apr. 2008, pp. 29–32.

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. International Conference on Music Information Retrieval*, 2002, pp. 287–288.