On the stability of multiplicative update algorithms. Application to non-negative matrix factorization

Sur la stabilité des règles de mises à jour multiplicatives. Application à la factorisation en matrices positives

Roland Badeau
Nancy Bertin
Emmanuel Vincent

2009D022

novembre 2009

Département Traitement du Signal et des Images
Groupe AAO : Audio, Acoustique et Ondes

On the stability of multiplicative update algorithms. Application to non-negative matrix factorization.

Sur la stabilité des règles de mises à jour multiplicatives. Application à la factorisation en matrices positives.

Roland Badeau [(1)], Nancy Bertin [(2)], and Emmanuel Vincent [(2)]

(1) Institut Télécom; Télécom ParisTech; CNRS LTCI, 75634 Paris Cedex 13, France

(2) INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France

Abstract

Multiplicative update algorithms have encountered a great success to solve optimization problems with non-negativity constraints, such as the famous non-negative matrix factorization and its many variants. However, despite several years of research on the topic, the understanding of their convergence properties is still to be improved. In this paper, we show that Lyapunov's stability theory provides a very enlightening viewpoint on the problem. We prove the exponential or asymptotic stability of the solutions to general optimization problems with non-negative constraints, and finally study the difficult case of non-negative matrix factorization.

Index Terms

Optimization methods, non-negative matrix factorization, multiplicative update algorithms, convergence of numerical methods, stability, Lyapunov methods.

Résumé

Les règles de mises à jour multiplicatives ont connu un grand succès pour résoudre des problèmes d'optimisation avec contraintes de positivité, tels que la célèbre factorisation en matrices positives et ses nombreuses variantes. Néanmoins, malgré plusieurs années de recherche sur le sujet, la compréhension de leurs propriétés de convergence demeure imparfaite. Dans cet article, nous prouvons que la théorie de la stabilité de Lyapunov fournit un point de vue très instructif sur le problème. Nous prouvons la stabilité exponentielle ou asymptotique des solutions de problèmes généraux d'optimisation avec contraintes de positivité, et nous étudions finalement le cas difficile de la factorisation en matrices positives.

Mots clés

Méthodes d'optimisation, factorisation en matrices positives, algorithmes de mises à jour multiplicatives, convergence des méthodes numériques, stabilité, méthodes de Lyapunov.

The research leading to this document was supported by the French GIP ANR under contract ANR-06-JCJC-0027-01, Décompositions en Éléments Sonores et Applications Musicales - DESAM.

I. INTRODUCTION

NON-NEGATIVE matrix factorization (NMF) is a powerful, unsupervised decomposition technique allowing the representation of two-dimensional non-negative data as a linear combination of meaningful elements in a basis [1]. Its computation is generally formalized as a constrained optimization problem. The objective functions most often encountered in NMF literature rely on the Euclidean (EUC) distance, the Kullback-Leibler (KL) divergence [1], [2], or the Itakura-Saito (IS) divergence [3]. The three of them are enclosed in the more general framework of β-divergences [4], [5].

Several algorithms have been proposed in order to compute a NMF. The most popular is the multiplicative update algorithm initially proposed by Lee and Seung [2] for the EUC and KL divergences, which has then been generalized to the β-divergence [5], [6]. For the interested reader, other approaches have also been proposed, such as the projected gradient method [7]–[9], alternating least squares (ALS) algorithms [10], [11], the quasi-Newton algorithm [12], [13], a multilayer technique [14], or a space-alternating expectation-maximization (SAGE) algorithm derived in a statistical framework [15]. See for instance [13] for a recent survey on the topic.

After Lee and Seung's paper, the multiplicative update philosophy has been applied to various optimization problems involving non-negativity constraints, such as some variants of the NMF. These algorithms generally aim at enhancing (or enforcing) a particular property in the decomposition, depending on the application. In the context of image representation and recognition for instance, various properties have been investigated, such as orthogonality [16], spatial localization [17], sparseness [18], or transformation-invariance [19]. In the context of multipitch and music transcription, the desired properties are rather sparseness and decorrelation [20], spectral harmonicity [20]–[23], and temporal continuity [22], [23]. In source separation, classical constraints include sparseness [24], temporal continuity [24], [25], decorrelation [25], [26], and shift-invariance [27]. Additionally, reference [28] proposes a general framework for including shift-invariances, higher order decompositions and sparseness constraints for various applications.

A curious point is that to the best of our knowledge, despite many years of research and several papers on the topic, the convergence properties of multiplicative update algorithms for standard unsupervised NMF have not been clearly identified:

- Lee and Seung proved that the objective function based on EUC and KL decreases at each iteration [2] (and the proof was later generalized to β-divergences [6], for $\beta \in [1, 2]$). However, this proves neither that the limit value of the objective function is a local minimum, nor that the successive iterates converge to a limit point.
- In constrained optimization, all the local minima of the objective function are proved to be stationary points, defined as the solutions to Karush, Kuhn and Tucker's (KKT) optimality conditions. Stationary points of NMF were studied in [29]–[31]. Some numerical examples have been presented in [32], where the KKT conditions are not fulfilled after a high (but finite) number of iterations, but this does not contradict the possible asymptotic convergence to a local minimum.
- Since the multiplicative updates involve divisions, numerical problems could be encountered if the divisor becomes arbitrarily small. In order to circumvent this problem, it is proposed in [31] to add a small positive quantity to the divisor, and it is proved that any accumulation point of the sequence of iterates computed in this way is a stationary point[1]. However there is no guarantee that such a stationary point is a local minimum, nor that the algorithm converges to this accumulation point.

Analyzing the convergence properties of NMF multiplicative rules is difficult because at each iteration, these algorithms usually switch between two updates: one for the left factor, and one for the right factor of the matrix product. Nevertheless, the convergence analysis happens to be simpler if only one of the two factors is updated, the other one being kept unchanged throughout the iterations. This is the case of *supervised NMF* for instance: the basis vectors (which form the left factor) are learned in a preprocessing stage, and only the coefficients of the decomposition (which form the right factor) are iteratively estimated from the observed data [33]–[36].

In this paper, we intend to analyze the convergence of a general form of such simplified multiplicative update algorithms, before studying the case of standard unsupervised NMF. Two important aspects of these algorithms have to be taken into account:

- *Local convergence:* Since the objective function generally admits several local minima [5], [29], there is no guarantee that the algorithm converges to the global minimum. So the best result we can prove is the local

[1]Note that this proof only stands for the Euclidean distance, and does not apply when the added quantity is zero.

convergence to a local minimum[2]. This means that if the algorithm is initialized in a given neighborhood of a local minimum called *basin of attraction*, then the algorithm will converge to this local minimum.

- *Stability:* Because of the multiplicative form of the algorithm, a zero coefficient remains zero in all subsequent iterations. Zeroing may happen because of a bad initialization or because of the finite machine precision for instance. This prevents the convergence to a local minimum whose corresponding coefficient would be non-zero. However, this problem can be very easily circumvented. Indeed, in this case, our numerical simulations have shown that the limit point of the algorithm is generally unstable: replacing the zero coefficient by any arbitrarily small quantity will make the algorithm escape from this trap and finally converge to a stable limit point. Other well-known examples of unstable stationary points of the algorithm (with non-zero entries) are saddle points [10], [29], [30].

These remarks show that an appropriate notion for analyzing the convergence of multiplicative update algorithms is the *asymptotic stability* in the sense of Lyapunov's theory [39] (or the *semi-stability* in the case of standard unsupervised NMF, as shown in section IV), which implies both local and stable convergence.

The paper is organized as follows. In section II, we present some elementary results about NMF, general optimization problems with non-negativity constraints, and multiplicative update algorithms. The convergence of these algorithms is analyzed by means of Lyapunov's stability theory in section III, and the difficult case of non-negative matrix factorization is studied in section IV. The main conclusions of this paper are summarized in section V. Finally, the mathematical proofs of the main results presented in this paper are included in the Appendix (due to the lack of room, some proofs have been moved to a separate document [40]).

II. Theoretical background

A. Multiplicative update algorithms and NMF

Given a matrix $\boldsymbol{V} \in \mathbb{R}_+^{F\times T}$ and an integer $K < \min(F,T)$, NMF consists in computing a reduced-rank approximation of $\boldsymbol{V}$ as a product $\boldsymbol{WH}$, where $\boldsymbol{W} \in \mathbb{R}_+^{F\times K}$ and $\boldsymbol{H} \in \mathbb{R}_+^{K\times T}$. This problem can be formalized as the minimization of an objective function

$$D(\boldsymbol{V}|\boldsymbol{WH}) = \sum_{f=1}^{F}\sum_{t=1}^{T} d\left(v_{ft}\middle|\sum_{k=1}^{K} w_{fk}h_{kt}\right), \tag{1}$$

where d is a scalar divergence (i.e. a function such that $\forall x, y \in \mathbb{R}_+, d(x|y) \geq 0$, and $d(x|y) = 0$ if and only if $y = x$).

β-divergences [4], [5] are defined for all $\beta \in \mathbb{R}\backslash\{0,1\}$ as

$$d_\beta(x|y) = \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right). \tag{2}$$

The Euclidean distance corresponds to $\beta = 2$, and KL and IS divergences are obtained when $\beta \to 1$ and $\beta \to 0$, respectively. The generalization of Lee and Seung's multiplicative updates to the β-divergence takes the following form [5], [6]:

$$\begin{array}{rcl} \boldsymbol{W} & \leftarrow & \boldsymbol{W} \otimes \frac{(\boldsymbol{V}\otimes(\boldsymbol{WH})^{\beta-2})\boldsymbol{H}^T}{(\boldsymbol{WH})^{\beta-1}\boldsymbol{H}^T} \\ \boldsymbol{H} & \leftarrow & \boldsymbol{H} \otimes \frac{\boldsymbol{W}^T(\boldsymbol{V}\otimes(\boldsymbol{WH})^{\beta-2})}{\boldsymbol{W}^T(\boldsymbol{WH})^{\beta-1}} \end{array} \tag{3}$$

where the symbol $\otimes$ and the fraction bar denote entrywise matrix product and division respectively, and the exponentiations must also be understood entrywise. In [6], it is proved that if $\beta \in [1,2]$, then the objective function is non-increasing at each iteration of (3). As in [2], this algorithm can be interpreted as a gradient descent with an adaptive step size for each entry, defined as a function of both matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, chosen so that the successive iterates remain non-negative.

Alternately, we consider the interpretation used in [16], [23], [24]: the recursion for $\boldsymbol{W}$ in (3) can be written $\forall f, k,\ w_{fk} \leftarrow w_{fk}\,\frac{m^w_{fk}}{p^w_{fk}}$, where $p^w_{fk} \geq 0$ and $m^w_{fk} \geq 0$ are such that the partial derivative of the objective function

[2]Avoiding to be trapped in a local minimum distinct from the global minimum is not the topic of this paper (the interested reader can have a look at [37], [38]).

wrt w_{fk} is equal to $p^w_{fk} - m^w_{fk}$:

$$\frac{\partial D}{\partial w_{fk}} = \underbrace{\sum_{t=1}^{T} \hat{v}_{ft}^{\beta-1} h_{kt}}_{p^w_{fk}} - \underbrace{\sum_{t=1}^{T} v_{ft}\, \hat{v}_{ft}^{\beta-2} h_{kt}}_{m^w_{fk}}, \tag{4}$$

where

$$\hat{v}_{ft} = \sum_{k=1}^{K} w_{fk} h_{kt}. \tag{5}$$

Thus if $\frac{\partial D}{\partial w_{fk}} > 0$, $\frac{m^w_{fk}}{p^w_{fk}} < 1$ so that w_{fk} decreases, and conversely if $\frac{\partial D}{\partial w_{fk}} < 0$, $\frac{m^w_{fk}}{p^w_{fk}} > 1$ so that w_{fk} increases. This confirms that the updates (3) form a descent method.

We propose in this paper to generalize this approach by introducing an exponent step size $\eta > 0$:

$$\begin{array}{rcl} \boldsymbol{W} & \leftarrow & \boldsymbol{W} \otimes \left(\frac{(\boldsymbol{V} \otimes (\boldsymbol{W}\boldsymbol{H})^{\beta-2})\boldsymbol{H}^T}{(\boldsymbol{W}\boldsymbol{H})^{\beta-1}\boldsymbol{H}^T} \right)^{\eta} \\ \boldsymbol{H} & \leftarrow & \boldsymbol{H} \otimes \left(\frac{\boldsymbol{W}^T(\boldsymbol{V} \otimes (\boldsymbol{W}\boldsymbol{H})^{\beta-2})}{\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{H})^{\beta-1}} \right)^{\eta} \end{array} \tag{6}$$

Note that standard multiplicative updates (3) correspond to the particular case $\eta = 1$. The convergence properties of the generalized algorithm (6) will be analyzed in section IV. The following proposition proves that (6) satisfies the same decrease property as (3) for all $\eta \in]0, 1]$.

Proposition 1. *Consider the objective function $D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ defined in equation* (1)*, involving the β-divergence* (2)*, with $\beta \in [1, 2]$. If $\eta \in]0, 1]$ and if $(\boldsymbol{W}, \boldsymbol{H})$ is not a fixed point of* (6)*, then* (6) *makes the objective function strictly decrease.*

Proposition 1 is proved in Appendix A. Note that this property does not guarantee that that the limit value of the objective function is a local minimum, nor that the successive values of $\boldsymbol{W}$ and $\boldsymbol{H}$ converge to a limit point.

From now on and until section IV, in order to both simplify the mathematical developments and address optimization problems with non-negativity constraints in the most general framework, we will consider a simplified form of multiplicative update algorithm. The main difference with (6) is that this algorithm does not switch between two multiplicative updates (such as those for $\boldsymbol{W}$ and $\boldsymbol{H}$ in (6)), but rather updates all variables at once (like in supervised NMF).

B. General optimization problems with non-negativity constraints

We consider the minimization in the first orthant $\mathbb{R}^n_+$ of a twice continuously differentiable objective function $J : \mathbb{R}^n_+ \to \mathbb{R}$. For any vector $\boldsymbol{x} = [x_1 \dots x_n]^T \in \mathbb{R}^n_+$, the constraint $x_i \geq 0$ is said to be *active* if $x_i = 0$, or *inactive* if $x_i > 0$.

The following two propositions are the consequence of classical results in constrained optimization theory [41].

Proposition 2 (First order KKT optimality conditions)**.** *Let $\boldsymbol{\nabla} J(\boldsymbol{x})$ denote the gradient vector of the objective function J. Then for any local minimum $\boldsymbol{x}$ of function J in $\mathbb{R}^n_+$,*

- $\forall i \in \{1 \dots n\}$ *such that* $x_i > 0$*,* $\nabla_i J = 0$*;*
- $\forall i \in \{1 \dots n\}$ *such that* $x_i = 0$*,* $\nabla_i J \geq 0$*.*

If $x_i = 0$ and $\nabla_i J > 0$, the constraint is said to be *strictly active*.

Following these considerations, we introduce the following notations for denoting the extraction of particular sub-vectors or sub-matrices:

- $[.]_0$ is obtained by selecting the coefficients (of a vector) or the rows and columns (of a matrix) corresponding to strictly active constraints, i.e. whose index i is such that $\nabla_i J(\boldsymbol{x}) > 0$ (and $x_i = 0$).
- $[.]_+$ is obtained by selecting the coefficients, or the rows and columns, whose index i is such that $\nabla_i J(\boldsymbol{x}) = 0$ (and either $x_i > 0$ or $x_i = 0$).
- $[.]^\star_+$ is obtained by selecting the coefficients, or the rows and columns, corresponding to inactive constraints, i.e. whose index i is such that $x_i > 0$ (and $\nabla_i J(\boldsymbol{x}) = 0$).

We then have the following optimality condition at order 2:

Proposition 3 (Second order optimality condition)**.** *Let $\boldsymbol{\nabla}^2 J(\boldsymbol{x})$ denote the Hessian matrix of the objective function J. Then for any local minimum $\boldsymbol{x}$ of function J in $\mathbb{R}_+^n$, the sub-matrix $[\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+$ is positive semidefinite.*

C. Multiplicative update algorithms

Suppose that the gradient of function J can be written as the difference of two non-negative functions: $\boldsymbol{\nabla} J(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{x})$, where both functions $\boldsymbol{p} : \mathbb{R}_+^n \to \mathbb{R}_+^n$ and $\boldsymbol{m} : \mathbb{R}_+^n \to \mathbb{R}_+^n$ are continuously differentiable. For simplicity, we will assume that $p_i(\boldsymbol{x}) > 0$, $\forall i \in \{1 \ldots n\}$ and $\forall \boldsymbol{x} \in \mathbb{R}_+^n$.

Definition 1 (Multiplicative update algorithm)**.** A multiplicative update algorithm is defined by a recursion of the form $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$, where mapping $\boldsymbol{\phi} : \mathbb{R}_+^n \to \mathbb{R}_+^n$ is defined as

$$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\Lambda}(\boldsymbol{x})^\eta \, \boldsymbol{x} \tag{7}$$

with

$$\boldsymbol{\Lambda}(\boldsymbol{x}) = \operatorname{diag}\left(\frac{\boldsymbol{m}(\boldsymbol{x})}{\boldsymbol{p}(\boldsymbol{x})}\right).$$

The scalar parameter η is the step size of the recursion.

The following lemma is a corollary of Proposition 2.

Lemma 4. *If $\boldsymbol{x}$ is a local minimum of function J, then $\boldsymbol{x}$ is a fixed point of mapping $\boldsymbol{\phi}$ (i.e. $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$).*

Proof: Obviously, if $x_i = 0$, $\phi_i(\boldsymbol{x}) = \boldsymbol{x}_i$. Otherwise Proposition 2 proves that $\forall i \in \{1 \ldots n\}$ such that $x_i > 0$, $p_i(\boldsymbol{x}) = m_i(\boldsymbol{x})$, thus $\phi_i(\boldsymbol{x}) = \boldsymbol{x}_i$. ∎

Note that recursion $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$ can be seen as a descent method. Indeed, the first order expansion of function $\eta \mapsto J\left(\boldsymbol{\Lambda}(\boldsymbol{x})^\eta \, \boldsymbol{x}\right)$ in the neighborhood of $\eta = 0$ yields

$$\begin{aligned} J(\boldsymbol{\phi}(\boldsymbol{x})) - J(\boldsymbol{x}) &= \eta \boldsymbol{\nabla} J(\boldsymbol{x})^T (\ln\left(\boldsymbol{\Lambda}(\boldsymbol{x})\right) \boldsymbol{\phi}(\boldsymbol{x})) + O(\eta^2) \\ &= -\eta \sum_{i=1}^n \phi_i(\boldsymbol{x})(p_i(\boldsymbol{x}) - m_i(\boldsymbol{x})) \ln\left(\tfrac{p_i(\boldsymbol{x})}{m_i(\boldsymbol{x})}\right)) + O(\eta^2) \end{aligned}$$

This equation shows that if $\boldsymbol{\phi}(\boldsymbol{x}) \neq \boldsymbol{x}$ and if $\eta > 0$ is small enough, then $J(\boldsymbol{\phi}(\boldsymbol{x})) - J(\boldsymbol{x}) < 0$, which means that the objective function decreases.

III. Stability analysis of multiplicative updates

A. Stability definitions

Let us recall a few classical definitions in Lyapunov's stability theory of discrete dynamical systems [39].

Definition 2 (Lyapunov stability)**.** A fixed point $\boldsymbol{x}$ of the recursion $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$ is said to be *Lyapunov stable* if $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $\forall \boldsymbol{x}^{(0)} \in \mathbb{R}_+^n$, $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| < \delta \;\Rightarrow \|\boldsymbol{x}^{(p)} - \boldsymbol{x}\| < \varepsilon \;\forall p \in \mathbb{N}$.

This property means that initializing the recursion close enough to $\boldsymbol{x}$ guarantees that the subsequent iterates remain in a given bounded domain around $\boldsymbol{x}$. However, it does not guarantee local convergence. A system which is not Lyapunov stable is called *unstable*.

Definition 3 (Asymptotic stability)**.** A fixed point $\boldsymbol{x}$ of the recursion $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$ is said to be *asymptotically stable* if it is Lyapunov stable and $\exists \delta > 0$ such that $\forall \boldsymbol{x}^{(0)} \in \mathbb{R}_+^n$, $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| < \delta \;\Rightarrow \boldsymbol{x}^{(p)} \underset{p \to +\infty}{\longrightarrow} \boldsymbol{x}$.

This property means that initializing the recursion close enough to $\boldsymbol{x}$ guarantees the convergence to $\boldsymbol{x}$. A system which is Lyapunov stable, but no asymptotically stable, is sometimes called *marginally stable*.

Definition 4 (Exponential stability)**.** A fixed point $\boldsymbol{x}$ of the recursion $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$ is said to be *exponentially stable* if $\exists \delta, \alpha, \beta > 0$ such that $\forall \boldsymbol{x}^{(0)} \in \mathbb{R}_+^n$, $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| < \delta \;\Rightarrow \|\boldsymbol{x}^{(p)} - \boldsymbol{x}\| \leq \alpha \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| \exp(-\beta p) \;\forall p \in \mathbb{N}$.

This property ensures a *linear* rate of convergence; it also implies asymptotic stability. A system which is asymptotically stable, but not exponentially stable has a *sub-linear* rate of convergence.

B. Lyapunov's first (or indirect) method

Lyapunov's first (or indirect) method is not the most powerful tool for analyzing a dynamical system, since it only permits to characterize the exponential stability. Nevertheless, contrary to the second method which involves the difficult search for a candidate Lyapunov function (see section III-C), this first method can be straightforwardly applied.

Theorem 5 (Lyapunov's first stability theorem)**.** *Let $\boldsymbol{x}$ be a fixed point of a continuously differentiable mapping $\boldsymbol{\phi} : \mathbb{R}_+^n \to \mathbb{R}_+^n$. Let $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ be the Jacobian matrix*[3] *of mapping $\boldsymbol{\phi}$ at point $\boldsymbol{x}$. Then the exponential stability (or unstability) of $\boldsymbol{x}$ is characterized by the eigenvalues of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$:*

- *$\boldsymbol{x}$ is an exponentially stable fixed point if and only if all the eigenvalues of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ have a magnitude lower than 1.*
- *If at least one eigenvalue of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ has a magnitude greater than 1, then $\boldsymbol{x}$ is unstable.*

Note that if all eigenvalues of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ have a magnitude lower than or equal to 1, and at least one of them has magnitude 1, then Theorem 5 does not permit to conclude: the system can be stable, unstable, or anything else.

In order to apply Theorem 5 to the mapping $\boldsymbol{\phi}$ defined in equation (7), we characterize the eigenvalues of matrix $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ in the following proposition.

Proposition 6. *Let $\boldsymbol{x}$ be a local minimum of function J. Let*

$$\eta^\star = \frac{2}{\|\boldsymbol{P}(\boldsymbol{x})\|_2}, \tag{8}$$

where $\|.\|_2$ denotes the matrix 2-norm (or spectral norm) and $\boldsymbol{P}(\boldsymbol{x})$ is the positive semi-definite matrix

$$\boldsymbol{P}(\boldsymbol{x}) = \boldsymbol{D}(\boldsymbol{x})\boldsymbol{\nabla}^2 J(\boldsymbol{x})\boldsymbol{D}(\boldsymbol{x}) \tag{9}$$

with $\boldsymbol{D}(\boldsymbol{x}) = \operatorname{diag}(\boldsymbol{x}/\boldsymbol{p}(\boldsymbol{x}))^{\frac{1}{2}}$.

If $\eta = 0$, all the eigenvalues of the Jacobian matrix $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ are equal to 1. Otherwise, the multiplicity of $\lambda = 1$ as an eigenvalue of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ is equal to the dimension of the kernel of matrix $[\boldsymbol{P}(\boldsymbol{x})]_+$. Moreover, the other eigenvalues of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ are as follows:

- *If $\eta \in]0, \eta^\star[$, all the other eigenvalues have a magnitude lower than 1;*
- *If $\eta < 0$, all the other eigenvalues are greater than 1;*
- *If $\eta > \eta^\star$, at least one eigenvalue is lower than -1;*
- *If $\eta = \eta^\star$, at least one eigenvalue is equal to -1.*

Proposition 6 is proved in Appendix B. Note that in the case $\eta = 0$, $\boldsymbol{\phi}$ is the identity transform. Besides, if $[\boldsymbol{P}(\boldsymbol{x})]_+$ is non-singular, 1 is *not* an eigenvalue of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$. Now we can state the stability properties of the mapping (7):

Proposition 7. *Let $\boldsymbol{x}$ be a local minimum of function J.*

- *$\boldsymbol{x}$ is an exponentially stable fixed point if and only if $\eta \in]0, \eta^\star[$ and $[\boldsymbol{P}(\boldsymbol{x})]_+$ is non-singular;*
- *if $\eta \notin [0, \eta^\star]$, $\boldsymbol{x}$ is an unstable fixed point;*
- *if $\eta = 0$, $\boldsymbol{x}$ is a marginally stable fixed point.*

Proof: The first and second assertion are a corollary of Theorem 5 and Proposition 6. The third assertion is trivial, since if $\eta = 0$, mapping $\boldsymbol{\phi}$ is the identity transform. ∎

Note that the non-singularity of matrix $[\boldsymbol{P}(\boldsymbol{x})]_+$ is equivalent to the combination of the two following properties:

$$\forall i \text{ such that } x_i = 0,\ \nabla_i J(\boldsymbol{x}) > 0, \tag{10}$$

$$\text{matrix } [\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+ \text{ is positive definite.} \tag{11}$$

If $\eta = \eta^\star$ or if $\eta \in]0, \eta^\star[$ and matrix $[\boldsymbol{P}(\boldsymbol{x})]_+$ is singular, Lyapunov's first method does not permit to conclude, since there is at least one eigenvalue of magnitude 1.

[3]For $1 \leq i, j \leq n$, the $(i, j)^{\text{th}}$ coefficient of matrix $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$ is $\frac{\partial \phi_j}{\partial x_i}$.

Finally, the following proposition completes Proposition 7, and proves the equivalence between the exponentially stable fixed points of mapping ϕ, and the local minima of function J satisfying both properties (10) and (11).

Proposition 8. *Assume that $\eta > 0$ and $\boldsymbol{x}$ is an exponentially stable fixed point of mapping ϕ. Then $\boldsymbol{x}$ is a local minimum of function J, which additionally satisfies properties* (10) *and* (11).

Proposition 8 is proved in Appendix B.

C. Lyapunov's second (direct) method

For a system which is not exponentially stable, Lyapunov's first method does not permit to conclude on the possible Lyapunov or asymptotic stability of a fixed point. In such cases, we have to use the more powerful second method.

In this section, we will prove the following result, more general than that of Proposition 7: if $\eta \in]0, \eta^\star[$, assumption (11) alone is sufficient for guaranteeing the asymptotic stability of the dynamical system. Let us first recall the principle of Lyapunov's second method, that we apply in the domain $\mathbb{R}_+^n$.

Definition 5 (Lyapunov function)**.** For any $\boldsymbol{x} \in \mathbb{R}_+^n$, a Lyapunov function $\boldsymbol{y} \mapsto V(\boldsymbol{x}, \boldsymbol{y})$ is a continuous scalar function defined on a neighborhood of $\boldsymbol{x}$ included in $\mathbb{R}_+^n$, which is positive-definite (in the sense that $V(\boldsymbol{x}, \boldsymbol{x}) = 0$, and $V(\boldsymbol{x}, \boldsymbol{y}) > 0$ for all $\boldsymbol{y} \neq \boldsymbol{x}$).

Theorem 9 (Lyapunov's second stability theorem)**.** *Let $\boldsymbol{x}$ be a fixed point of a continuous mapping ϕ.*

- *If there is a Lyapunov function V such that $V(\boldsymbol{x}, \phi(\boldsymbol{y})) \leq V(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{y}$ in a neighborhood of $\boldsymbol{x}$, then $\boldsymbol{x}$ is Lyapunov stable.*
- *If there is a Lyapunov function V such that $V(\boldsymbol{x}, \phi(\boldsymbol{y})) < V(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{y} \neq \boldsymbol{x}$ in a neighborhood of $\boldsymbol{x}$, then $\boldsymbol{x}$ is asymptotically stable.*

If $\boldsymbol{x}$ is a local minimum of a function J, a natural candidate Lyapunov function for the mapping ϕ defined in equation (7) would be $V(\boldsymbol{x}, \boldsymbol{y}) = J(\boldsymbol{y}) - J(\boldsymbol{x})$. However, this choice raises two problems:

- Formally proving that mapping ϕ makes a given function J decrease in a neighborhood of $\boldsymbol{x}$ happens to be more difficult than expected;
- More importantly, the condition that V is positive-definite may not be satisfied in a neighborhood of $\boldsymbol{x}$ [4].

For these reasons, we propose an alternate Lyapunov function in the following lemma.

Lemma 10. *The function*

$$V(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \operatorname{diag}\left(\frac{\boldsymbol{p}(\boldsymbol{x}) + \boldsymbol{p}(\boldsymbol{y})}{\boldsymbol{x} + \boldsymbol{y}}\right)(\boldsymbol{y} - \boldsymbol{x}) \tag{12}$$

defines a symmetric Lyapunov function on $\mathbb{R}_+^n \times \mathbb{R}_+^n$.

Proof: The definition and continuity of function V on the borders of the domain $\mathbb{R}_+^n$ follow from the inequality $\left|\frac{y_i - x_i}{x_i + y_i}\right| \leq 1 \; \forall x_i, y_i \in \mathbb{R}_+^\star$. ■

We can now state our main result:

Proposition 11. *Let $\boldsymbol{x}$ be a local minimum of function J in $\mathbb{R}_+^n$, satisfying property* (11)*. If $\eta \in]0, \eta^\star[$, then mapping ϕ makes the Lyapunov function $\boldsymbol{y} \mapsto V(\boldsymbol{x}, \boldsymbol{y})$ decrease in a neighborhood of $\boldsymbol{x}$. As a consequence, $\boldsymbol{x}$ is an asymptotically stable fixed point of mapping ϕ.*

Proposition 11 is proved in Appendix C.

IV. Application to standard unsupervised NMF

In this section, we show how Lyapunov's stability theory can be applied to the particular problem of NMF with an objective function based on the β-divergence, which was introduced in section II-A.

Actually, analyzing the stability of the multiplicative updates (6) is particularly difficult for the following reasons:

[4] In the case of standard unsupervised NMF, because of the invariances of the factorization, there is a continuous set of fixed points $\boldsymbol{y}$ satisfying $J(\boldsymbol{y}) = J(\boldsymbol{x})$ (cf. section IV).

- Unsupervised NMF admits several invariances. For instance, the product $\boldsymbol{W}\boldsymbol{H}$ is unchanged by replacing matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ by the non-negative matrices $\boldsymbol{W}' = \boldsymbol{W}\boldsymbol{D}$ and $\boldsymbol{H}' = \boldsymbol{D}^{-1}\boldsymbol{H}$, where $\boldsymbol{D}$ is any diagonal matrix with positive diagonal coefficients. For this simple reason, the *local minima of the objective function are never isolated* (any local minimum is reached on a *continuum* of matrices $\boldsymbol{W}'$ and $\boldsymbol{H}'$ whose product is equal to $\boldsymbol{W}\boldsymbol{H}$). The consequence is that assumption (11) never stands. Note however that if some penalty terms are introduced into the objective function, involving $\boldsymbol{W}$ or $\boldsymbol{H}$ separately (see e.g. some constrained variants of NMF listed in the introduction [16]–[18], [20], [22]–[26], [28]), the local minima may become isolated.
- Anyway, recursion (6) cannot be implemented with a mapping of the form (7), since it switches between updates for $\boldsymbol{W}$ and $\boldsymbol{H}$.

For these reasons, the results presented in section III cannot be straightforwardly applied to (6). However, the same methodology can be used to study standard unsupervised NMF.

A. Lyapunov's first method

Because of the invariances of standard NMF, the local minima of the objective function can never be exponentially stable (the Jacobian matrix always admits $\lambda = 1$ as an eigenvalue). Thus Lyapunov's first method will not permit to conclude on the stability of multiplicative updates. Nevertheless, this approach still provides an interesting insight into the stability properties of the algorithm. We summarize below the main results that we obtained by applying this approach. Due to the lack of room, the complete proofs have been moved to a separate document [40]. Nevertheless, the key ideas are provided wherever possible in the following discussion.

Notations. In the following, the entries of the $F \times K$ matrix $\boldsymbol{W}$ and the $K \times T$ matrix $\boldsymbol{H}$ are remapped into vectors $\boldsymbol{w}$ and $\boldsymbol{h}$ of dimensions KF and KT, respectively. Vector $\boldsymbol{x}$ is formed by concatenating $\boldsymbol{w}$ and $\boldsymbol{h}$. The gradient of the function $J(\boldsymbol{x}) = D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ wrt $\boldsymbol{w}$ is decomposed as the difference of two non-negative functions $\boldsymbol{p}^w(\boldsymbol{x})$ and $\boldsymbol{m}^w(\boldsymbol{x})$ whose coefficients have been defined in equation (4), and similar notations are used for $\boldsymbol{h}$. Vector $\boldsymbol{p}(\boldsymbol{x})$ is formed by concatenating $\boldsymbol{p}^w(\boldsymbol{x})$ and $\boldsymbol{p}^h(\boldsymbol{x})$. The Hessian matrices of function J wrt $\boldsymbol{x}$, $\boldsymbol{w}$ and $\boldsymbol{h}$ are denoted $\boldsymbol{\nabla}^2_{xx}J(\boldsymbol{x})$, $\boldsymbol{\nabla}^2_{ww}J(\boldsymbol{x})$ and $\boldsymbol{\nabla}^2_{hh}J(\boldsymbol{x})$, respectively. Recursion (6) is rewritten in the form

$$\left\{\begin{array}{lcl} \boldsymbol{w}^{(p+1)} & = & \boldsymbol{\phi}^w(\boldsymbol{w}^{(p)}, \boldsymbol{h}^{(p)}) \\ \boldsymbol{h}^{(p+1)} & = & \boldsymbol{\phi}^h(\boldsymbol{w}^{(p+1)}, \boldsymbol{h}^{(p)}) \end{array}\right. \tag{13}$$

with

$$\left\{\begin{array}{lcl} \boldsymbol{\phi}^w(\boldsymbol{w}, \boldsymbol{h}) & = & \boldsymbol{\Lambda}_w(\boldsymbol{x})^\eta\, \boldsymbol{w} \\ \boldsymbol{\phi}^h(\boldsymbol{w}, \boldsymbol{h}) & = & \boldsymbol{\Lambda}_h(\boldsymbol{x})^\eta\, \boldsymbol{h} \end{array}\right. \tag{14}$$

and

$$\begin{array}{lcl} \boldsymbol{\Lambda}_w(\boldsymbol{x}) & = & \mathrm{diag}\left(\frac{\boldsymbol{m}^w(\boldsymbol{x})}{\boldsymbol{p}^w(\boldsymbol{x})}\right) \\ \boldsymbol{\Lambda}_h(\boldsymbol{x}) & = & \mathrm{diag}\left(\frac{\boldsymbol{m}^h(\boldsymbol{x})}{\boldsymbol{p}^h(\boldsymbol{x})}\right) \end{array}$$

Equivalently, we can write $\boldsymbol{x}^{(p+1)} = \boldsymbol{\phi}(\boldsymbol{x}^{(p)})$, with

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left[\boldsymbol{\phi}^w(\boldsymbol{w}, \boldsymbol{h}); \boldsymbol{\phi}^h(\boldsymbol{\phi}^w(\boldsymbol{w}, \boldsymbol{h}), \boldsymbol{h})\right]. \tag{15}$$

The parameter $\eta^\star$ introduced in the following lemma will play the same role as in section III-B.

Lemma 12. *Let $\boldsymbol{x}$ be a local minimum of the NMF objective function J defined in equation* (1)*, based on a β-divergence* (2)*. Let us define*

$$\eta^\star = \min\left(\frac{2}{\|\boldsymbol{P}^w(\boldsymbol{x})\|_2}, \frac{2}{\|\boldsymbol{P}^h(\boldsymbol{x})\|_2}\right), \tag{16}$$

where $\boldsymbol{P}^w(\boldsymbol{x})$, $\boldsymbol{P}^h(\boldsymbol{x})$ are the positive semi-definite matrices

$$\begin{array}{lcl} \boldsymbol{P}^w(\boldsymbol{x}) & = & \boldsymbol{D}^w(\boldsymbol{x})\boldsymbol{\nabla}^2_{ww}J(\boldsymbol{x})\boldsymbol{D}^w(\boldsymbol{x}); \\ \boldsymbol{P}^h(\boldsymbol{x}) & = & \boldsymbol{D}^h(\boldsymbol{x})\boldsymbol{\nabla}^2_{hh}J(\boldsymbol{x})\boldsymbol{D}^h(\boldsymbol{x}). \end{array} \tag{17}$$

with

$$\begin{array}{rcl} \boldsymbol{D}^w(\boldsymbol{x}) & = & \operatorname{diag}(\boldsymbol{w}/\boldsymbol{p}^w(\boldsymbol{x}))^{\frac{1}{2}}; \\ \boldsymbol{D}^h(\boldsymbol{x}) & = & \operatorname{diag}(\boldsymbol{h}/\boldsymbol{p}^h(\boldsymbol{x}))^{\frac{1}{2}}. \end{array} \tag{18}$$

Then $\forall \beta \in \mathbb{R}$*, we have* $0 < \eta^\star \leq 2$*, and if* $\beta \in [1, 2]$*,* $\eta^\star = 2$.

Proof: This lemma is proved by exhibiting eigenvectors of the positive semi-definite matrices $\boldsymbol{P}^w(\boldsymbol{x})$ and $\boldsymbol{P}^h(\boldsymbol{x})$, whose eigenvalue is equal to 1. If, additionally, $\beta \in [1, 2]$, the convexity of function J wrt $\boldsymbol{w}$ and $\boldsymbol{h}$ permits to prove that all the eigenvalues of $\boldsymbol{P}^w(\boldsymbol{x})$ and $\boldsymbol{P}^h(\boldsymbol{x})$ are lower than or equal to 1. ■

Lemma 12 confirms that if $\beta \in [1, 2]$, any step size $\eta \in]0, 1]$, for which the objective function was proved to decrease (cf. Proposition 1), is upper bounded by $\eta^\star = 2$. The following proposition characterizes the eigenvalues of the Jacobian matrix $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$.

Proposition 13. *Let* $\boldsymbol{x}$ *be a local minimum of function* J*. If* $\eta = 0$*, all the eigenvalues of the Jacobian matrix* $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$ *are equal to* 1*. Otherwise,* $\lambda = 1$ *is always an eigenvalue of* $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$*, whose multiplicity is equal to the dimension of the kernel of matrix* $[\boldsymbol{P}(\boldsymbol{x})]_+$*, where*

$$\boldsymbol{P}(\boldsymbol{x}) = \boldsymbol{D}(\boldsymbol{x}) \boldsymbol{\nabla}^2_{xx} J(\boldsymbol{x}) \boldsymbol{D}(\boldsymbol{x}) \tag{19}$$

and $\boldsymbol{D}(\boldsymbol{x}) = \operatorname{diag}(\boldsymbol{x}/\boldsymbol{p}(\boldsymbol{x}))^{\frac{1}{2}}$*. Moreover, the other eigenvalues of* $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$ *are as follows:*

- *If* $\eta \notin [0, 2]$*, there is at least one eigenvalue greater than* 1*;*
- *If* $\eta \in]0, \eta^\star[$*, all other eigenvalues have a magnitude lower than* 1.

Proof: The proof of Proposition 13 follows the same outline as the proof of Proposition 6. It additionally relies on the observation that $\lambda = 1$ and $\lambda = (1 - \eta)^2$ are always eigenvalues of the Jacobian matrix $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$ (which is proved by exhibiting the corresponding eigenvectors). ■

As mentioned above, Proposition 13 never permits to conclude on the stability of multiplicative updates. Nevertheless, it permits to formally prove the first and second assertions in the following proposition (the third one is trivial):

Proposition 14. *Let* $\boldsymbol{x}$ *be a local minimum of function* J.

- $\forall \eta \in \mathbb{R}$*,* $\boldsymbol{x}$ *is not an exponentially stable fixed point;*
- *if* $\eta \notin [0, 2]$*,* $\boldsymbol{x}$ *is an unstable fixed point;*
- *if* $\eta = 0$*,* $\boldsymbol{x}$ *is a marginally stable fixed point.*

Proposition 14 does not tell us what happens if $\eta \in]0, 2]$. Is the rate of convergence sub-linear? Is $\boldsymbol{x}$ at least a marginally stable fixed point ? This question is discussed in the next section.

B. Lyapunov's second method

As proved in Proposition 1, the objective function J based on the β-divergence with $\beta \in [1, 2]$ is decreasing when applying the multiplicative update rules (6) with $\eta \in]0, 1]$. More precisely, $J(\boldsymbol{\phi}(\boldsymbol{y})) < J(\boldsymbol{y})$, if $\boldsymbol{y}$ is not a fixed point of the recursion. Following the idea of Lyapunov's second method applied to $V(\boldsymbol{x}, \boldsymbol{y}) = J(\boldsymbol{y}) - J(\boldsymbol{x})$, a superficial interpretation of Theorem 9 could wrongly suggested that if $\boldsymbol{x}$ is a local minimum of function J, the decrease of V implies the local convergence to $\boldsymbol{x}$. Indeed, the decrease of this function does not prove the asymptotic stability, because V is not positive-definite (the local minima of J are never isolated). Besides, the decrease of function J may not be verified if $\eta \in]1, 2[$, whereas Proposition 13 suggests that there is no real difference in terms of exponential convergence[5].

Actually, asymptotic stability is not the appropriate notion for analyzing the convergence properties of standard unsupervised NMF multiplicative updates, since in practice initializing the algorithm in a neighborhood of a local minimum $(\boldsymbol{W}, \boldsymbol{H})$ does not make the algorithm converge to $(\boldsymbol{W}, \boldsymbol{H})$, but rather to an other pair $(\boldsymbol{W}', \boldsymbol{H}')$ such that $\boldsymbol{W}'\boldsymbol{H}' = \boldsymbol{W}\boldsymbol{H}$.

[5]Our numerical tests provided some examples where $\eta \in]1, 2[$ and J is not globally decreasing, whereas the algorithm still converges to a local minimum.

The appropriate notion is rather that of *semi-stability*, which was introduced to analyze the stability of continuous-time dynamical systems having a continuum of equilibria [42]. Semi-stability can be seen as a generalization of asymptotic stability to non-isolated fixed points.

Definition 6 (Semi-stability)**.** A fixed point $\boldsymbol{x}$ of a continuous mapping $\boldsymbol{\phi}$ is semi-stable, if there is a neighborhood $\mathcal{U}_x$ of $\boldsymbol{x}$ in $\mathbb{R}_+^n$ such that any fixed point of $\boldsymbol{\phi}$ in $\mathcal{U}_x$ is Lyapunov stable, and if there is $\delta > 0$ such that $\forall \boldsymbol{x}^{(0)} \in \mathcal{U}_x$, $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| < \delta \;\Rightarrow \boldsymbol{x}^{(p)} \underset{p \to +\infty}{\longrightarrow} \boldsymbol{x}^\star$, where $\boldsymbol{x}^\star$ is a fixed point of $\boldsymbol{\phi}$ in $\mathcal{U}_x$.

In particular, it can be noticed that an *isolated*, semi-stable fixed point is asymptotically stable, since $\boldsymbol{x}^\star$ is necessary equal to $\boldsymbol{x}$. In order to prove the semi-stability of a continuous mapping $\boldsymbol{\phi}$ having a continuum of fixed points, we introduce the following theorem.

Theorem 15. *Let $\boldsymbol{x}$ be a fixed point of a continuous mapping $\boldsymbol{\phi}$. Suppose that there is a neighborhood $\mathcal{U}_x$ of $\boldsymbol{x}$ in $\mathbb{R}_+^n$ and a Lyapunov function $V(\boldsymbol{x}, .)$ defined on $\mathcal{U}_x$, such that $\forall \boldsymbol{y} \in \mathcal{U}_x$, $\boldsymbol{\phi}(\boldsymbol{y}) \neq \boldsymbol{y} \;\Rightarrow\; V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{y})) < V(\boldsymbol{x}, \boldsymbol{y})$. Also suppose that any fixed point $\boldsymbol{x}^\star$ in $\mathcal{U}_x$ satisfies a similar condition (involving a Lyapunov function $V(\boldsymbol{x}^\star, .)$ defined on a neighborhood $\mathcal{U}_{x^\star}$). Then $\boldsymbol{x}$ is a semi-stable fixed point.*

Theorem 15 is proved in Appendix D. Unfortunately, the semi-stability (or possible unstability) of (6) for $\eta \in]0, \eta^\star[$ is still an open question, since we have not yet been able to apply Theorem 15 to this algorithm. The candidate-Lyapunov function proposed in section III-C might be a suitable choice, but this is to be proved.

V. Conclusions

In this paper, we analyzed the convergence properties of general multiplicative update algorithms, designed to solve optimization problems with non-negativity constraints. We have applied Lyapunov's first and second methods to find some criteria which guarantee the exponential or asymptotic stability of the local minima of the objective function, by either analyzing the eigenvalues of the Jacobian matrix of the mapping, or introducing a suitable Lyapunov function. This study can be straightforwardly applied to supervised NMF algorithms, which update all variables at once, instead of switching between two multiplicative updates.

We then applied the same methodology to study the more complex case of standard unsupervised NMF based on β-divergences. Analyzing the stability properties of the multiplicative updates happened to be particularly difficult because of the invariances of the factorization, which make the local minima of the objective function non-isolated, thus non-asymptotically stable. Actually the appropriate notion to describe the behavior of this algorithm in the neighborhood of a local minimum seems to be that of semi-stability. Although this semi-stability is still to be proved for $\eta \in]0, \eta^\star[$, Lyapunov's first method has already provided some interesting insights into the convergence properties of standard NMF's generalized multiplicative update rules.

A possible extension of this work would focus on the hybrid case of constrained unsupervised NMF. If the constraint is expressed via a modification of the underlying model like in [19], [21], [27], the convergence analysis may be similar to that of standard unconstrained NMF. If the constraint is expressed via additional penaly terms in the objective function like in [16]–[18], [20], [24]–[26], the convergence analysis may paradoxically be simpler than in unconstrained NMF: the algorithm still switches between two multiplicative updates, but the modified objective function may have isolated, thus asymptotically stable, local minima.

Appendix

A. Multiplicative update algorithms for NMF

Proof of Proposition 1: We remark that $d_\beta(x|y)$ defined in equation (2) can be written in the form $d_\beta(x|y) = x^\beta \delta_\beta(\frac{y}{x})$, where

$$\begin{aligned}
\delta_\beta(u) &= \tfrac{1}{\beta(\beta-1)}(1 + (\beta-1)u^\beta - \beta u^{\beta-1}), \\
\tfrac{d\delta_\beta}{du} &= u^{\beta-1} - u^{\beta-2}, \\
\tfrac{d^2\delta_\beta}{du^2} &= (\beta-1)u^{\beta-2} + (2-\beta)u^{\beta-3}.
\end{aligned}$$

Function δ_β is strictly convex in $\mathbb{R}_+$ if and only if $\forall u > 0$, $\frac{d^2\delta_\beta}{du^2} > 0$, which is equivalent to $1 \leq \beta \leq 2$. Then we can write

$$D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) = \sum_{f=1}^{F}\sum_{t=1}^{T} v_{ft}^\beta \delta_\beta\left(\frac{\hat{v}_{ft}}{v_{ft}}\right)$$

where $\hat{v}_{ft}$ was defined in equation (5). Because of the strict convexity of δ_β, we have $\forall \boldsymbol{h}'_t \in \mathbb{R}_+^K$

$$\begin{aligned}\delta_\beta\left(\frac{\sum_{k=1}^{K} w_{fk}h'_{kt}}{v_{ft}}\right) &= \delta_\beta\left(\sum_{k=1}^{K} \frac{w_{fk}h_{kt}}{\hat{v}_{ft}} \times \frac{\hat{v}_{ft}h'_{kt}}{v_{ft}h_{kt}}\right)\\ &\leq \sum_{k=1}^{K} \frac{w_{fk}h_{kt}}{\hat{v}_{ft}} \delta_\beta\left(\frac{\hat{v}_{ft}h'_{kt}}{v_{ft}h_{kt}}\right)\end{aligned}$$

with equality if and only if $\forall k \in \{1 \dots K\}$, $h'_{kt} = h_{kt}$. As a consequence, $\forall \boldsymbol{H}' \in \mathbb{R}_+^{K\times T}$, $D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}') \leq G^{W,H}(\boldsymbol{H}')$, where

$$\begin{aligned}G^{W,H}(\boldsymbol{H}') &= \sum_{k=1}^{K}\sum_{t=1}^{T} G_{kt}^{W,H}(h'_{kt})\\ G_{kt}^{W,H}(h'_{kt}) &= \sum_{f=1}^{F} \frac{w_{fk}h_{kt}}{\hat{v}_{ft}} v_{ft}^{\beta} \delta_\beta\left(\frac{\hat{v}_{ft}}{v_{ft}}\frac{h'_{kt}}{h_{kt}}\right)\end{aligned}$$

with equality if and only if $\boldsymbol{H}' = \boldsymbol{H}$.

Then let $h'_{kt}(\eta) = h_{kt}\left(\frac{m_{kt}^h}{p_{kt}^h}\right)^\eta$ where p_{kt}^h and m_{kt}^h are defined similarly to p_{fk}^w and m_{fk}^w in equation (4):

$$\begin{aligned}p_{kt}^h &= \sum_{f=1}^{F} w_{fk}\hat{v}_{ft}^{\beta-1},\\ m_{kt}^h &= \sum_{f=1}^{F} w_{fk}v_{ft}\hat{v}_{ft}^{\beta-2}.\end{aligned} \tag{20}$$

Define the function $F_{kt}(\eta) = G_{kt}^{W,H}(h'_{kt}(\eta))$. Then

$$\begin{aligned}\frac{dF_{kt}}{d\eta} &= \frac{dG_{kt}^{W,H}}{dh'_{kt}}\frac{dh'_{kt}}{d\eta}\\ &= h_{kt}p_{kt}^h \ln\left(\frac{m_{kt}^h}{p_{kt}^h}\right)\left(\frac{m_{kt}^h}{p_{kt}^h}\right)^{\eta\beta}\left[1-\left(\frac{m_{kt}^h}{p_{kt}^h}\right)^{1-\eta}\right].\end{aligned}$$

If $h_{kt} \neq 0$ and $m_{kt}^h \neq p_{kt}^h$, then

- $\frac{dF_{kt}}{d\eta} < 0$ for all $\eta \in]-\infty, 1[$,
- $\frac{dF_{kt}}{d\eta} = 0$ for $\eta = 1$,
- $\frac{dF_{kt}}{d\eta} > 0$ for all $\eta \in]1, +\infty[$.

In particular, $\forall \eta \in]0,1]$, $F_{kt}(\eta) < F_{kt}(0)$.

Finally, let $F(\eta) = \sum_{k=1}^{K}\sum_{t=1}^{T} F_{kt}(\eta) = G^{W,H}(\boldsymbol{H}'(\eta))$.

If $\boldsymbol{H}'(\eta) \neq \boldsymbol{H}$, then $\forall \eta \in]0,1]$, $F(\eta) < F(0)$. Consequently, $\forall \eta \in]0,1]$, $D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}'(\eta)) \leq F(\eta) < F(0) = D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$. Thus $\boldsymbol{H}'(\eta) \neq \boldsymbol{H} \Rightarrow D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}'(\eta)) < D(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$.

The same proof can be applied to the update of $\boldsymbol{W}$. ∎

B. Lyapunov's first method

Proof of Proposition 6: By differentiating equation (7), we obtain the expression of the Jacobian matrix $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$:

$$\begin{aligned}&\nabla\boldsymbol{\phi}^T(\boldsymbol{x}) = \boldsymbol{\Lambda}(\boldsymbol{x})^\eta +\\ &\eta(\nabla\boldsymbol{m}^T\mathrm{diag}(1/\boldsymbol{m}(\boldsymbol{x})) - \nabla\boldsymbol{p}^T\mathrm{diag}(1/\boldsymbol{p}(\boldsymbol{x}))\mathrm{diag}(\boldsymbol{\phi}(\boldsymbol{x})).\end{aligned} \tag{21}$$

If $\boldsymbol{x}$ is a local minimum of the objective function J, it is a fixed point of $\boldsymbol{\phi}$, thus equation (21) yields

$$\nabla\boldsymbol{\phi}^T(\boldsymbol{x}) = \boldsymbol{\Lambda}(\boldsymbol{x})^\eta - \eta\boldsymbol{\nabla}^2 J(\boldsymbol{x})\,\mathrm{diag}(\boldsymbol{x}/\boldsymbol{p}(\boldsymbol{x})) \tag{22}$$

Now let us have a look at the eigenvalues of matrix $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$.

- For all i such that $x_i = 0$, it is easy to see that the i^{th} column of the identity matrix, that we denote by $\boldsymbol{u}_i$, is a right eigenvector of $\nabla\boldsymbol{\phi}^T(\boldsymbol{x})$, associated to the eigenvalue $\lambda_i = \left(\frac{m_i(\boldsymbol{x})}{p_i(\boldsymbol{x})}\right)^\eta$ (since the product of the last term in equation (22) and vector $\boldsymbol{u}_i$ is zero). We can conclude that if $\nabla_i J(\boldsymbol{x}) = 0$ or $\eta = 0$, then $\lambda_i = 1$; otherwise $\lambda_i \in [0, 1[$ if $\eta > 0$, and $\lambda_i > 1$ if $\eta < 0$.

- Let $\boldsymbol{u}$ be a right eigenvector of $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$ which does not belong to the subspace spanned by the previous ones, associated to an eigenvalue λ. Then

$$\boldsymbol{\Lambda}(\boldsymbol{x})^\eta \boldsymbol{u} - \eta \boldsymbol{\nabla}^2 J(\boldsymbol{x}) \mathrm{diag}(\boldsymbol{x}/\boldsymbol{p}(\boldsymbol{x}))\boldsymbol{u} = \lambda \boldsymbol{u}. \tag{23}$$

Let $\boldsymbol{v} = \boldsymbol{D}(\boldsymbol{x})\,\boldsymbol{u}$; this vector is non-zero, otherwise $\boldsymbol{u}$ would belong the subspace spanned by the previous set of eigenvectors $\boldsymbol{u}_i$. Left multiplying equation (23) by $\boldsymbol{D}(\boldsymbol{x})$ yields $\boldsymbol{\Lambda}(\boldsymbol{x})^\eta \boldsymbol{v} - \eta \boldsymbol{P}(\boldsymbol{x})\boldsymbol{v} = \lambda \boldsymbol{v}$, where the positive semi-definite matrix $\boldsymbol{P}(\boldsymbol{x})$ was defined in equation (9). Then noting that $\boldsymbol{\Lambda}(\boldsymbol{x})^\eta \boldsymbol{v} = \boldsymbol{v}$, we obtain $(\boldsymbol{I}_n - \eta\, \boldsymbol{P}(\boldsymbol{x}))\,\boldsymbol{v} = \lambda \boldsymbol{v}$, where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix. This proves that λ is an eigenvalue of $\boldsymbol{I}_n - \eta\, \boldsymbol{P}(\boldsymbol{x})$. It is easy to see that the previous set of vectors $\boldsymbol{u}_i$ are also eigenvectors of $\boldsymbol{P}(\boldsymbol{x})$, associated to the eigenvalue 0, but they cannot be colinear to $\boldsymbol{v}$, since $v_i = 0$ for all i such that $x_i = 0$. Thus $\lambda = 1 - \eta\mu$, where μ is an eigenvalue of $[\boldsymbol{P}(\boldsymbol{x})]_+^\star$. We can conclude that if $\mu = 0$, then $\lambda = 1$. Otherwise we note that $\eta^\star$ defined in equation (8) is equal to $\frac{2}{\|[\boldsymbol{P}(\boldsymbol{x})]_+^\star\|_2}$, and

  - if $\eta = 0$, all the other eigenvalues are also equal 1;
  - if $0 < \eta < \eta^\star$, all the other eigenvalues λ belong to $]-1, 1[$;
  - if $\eta < 0$, all the other eigenvalues λ are greater than 1;
  - if $\eta > \eta^\star$, there is at least one eigenvalue $\lambda < -1$;
  - if $\eta = \eta^\star$, there is at least one eigenvalue $\lambda = -1$.

Finally, the total number of eigenvalues equal to 1 (if $\eta \neq 0$) is the number of coefficients i such that $x_i = 0$ and $\nabla_i J(\boldsymbol{x}) = 0$, plus the dimension of the kernel of matrix $[\boldsymbol{P}(\boldsymbol{x})]_+^\star$. In other words, it is equal to the dimension of the kernel of matrix $[\boldsymbol{P}(\boldsymbol{x})]_+$. ■

Proof of Proposition 8:

Since $\boldsymbol{x}$ is an exponentially stable fixed point of mapping $\boldsymbol{\phi}$, all the eigenvalues of $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$ have magnitude lower than 1. Moreover, $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$ and $\forall i$, either $x_i = 0$ or $m_i(\boldsymbol{x}) = p_i(\boldsymbol{x})$. Thus equation (21) still yields equation (22).

Again, let us have a look at the eigenvalues of matrix $\nabla \boldsymbol{\phi}^T(\boldsymbol{x})$:

- For all i such that $x_i = 0$, the eigenvalue $\lambda_i = \left(\frac{m_i(\boldsymbol{x})}{p_i(\boldsymbol{x})}\right)^\eta$ associated to the eigenvector $\boldsymbol{u}_i$ is lower than 1. Thus $m_i(\boldsymbol{x}) < p_i(\boldsymbol{x})$ and $\nabla_i J(\boldsymbol{x}) > 0$.
- Previous developments show that the others eigenvalues λ can be written in the form $\lambda = 1 - \eta\,\mu$, where μ is an eigenvalue of $[\boldsymbol{P}(\boldsymbol{x})]_+^\star$. Since, $\lambda < 1$, we conclude that $\mu > 0$, thus $[\boldsymbol{P}(\boldsymbol{x})]_+^\star$ is a positive definite matrix, and so is $[\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+^\star = [\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+$.

We have thus proved that properties (10) and (11) stand. This proves that $\boldsymbol{x}$ is a local minimum of function J. ■

C. Lyapunov's second method

Proof of Proposition 11: Function $V(\boldsymbol{x}, \boldsymbol{y})$ defined in equation (12) can be decomposed as follows:

$$\begin{aligned} V(\boldsymbol{x}, \boldsymbol{y}) = & \sum_{i/x_i=0} y_i \,\frac{p_i(\boldsymbol{x}) + p_i(\boldsymbol{y})}{2} \\ & + \sum_{i/x_i>0} \tfrac{1}{2}(y_i - x_i)^2\, \frac{p_i(\boldsymbol{x}) + p_i(\boldsymbol{y})}{x_i + y_i} \\ = & \sum_{i/x_i=0} y_i \,(p_i(\boldsymbol{x}) + O(\|\boldsymbol{y} - \boldsymbol{x}\|)) \\ & + \sum_{i/x_i>0} \tfrac{1}{2}(y_i - x_i)^2 \left(\frac{p_i(\boldsymbol{x})}{x_i} + O(\|\boldsymbol{y} - \boldsymbol{x}\|)\right). \end{aligned} \tag{24}$$

Note that, since $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{\phi}$ is continuously differentiable at $\boldsymbol{x}$, $\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\| = O(\|\boldsymbol{y} - \boldsymbol{x}\|)$. Therefore replacing $\boldsymbol{y}$ by $\boldsymbol{\phi}(\boldsymbol{y})$ in equation (24) yields

$$\begin{aligned} V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{y})) = & \sum_{i/x_i=0} \phi_i(\boldsymbol{y})\,(p_i(\boldsymbol{x}) + O(\|\boldsymbol{y} - \boldsymbol{x}\|)) \\ & + \sum_{i/x_i>0} \tfrac{1}{2}(\phi_i(\boldsymbol{y}) - x_i)^2 \left(\frac{p_i(\boldsymbol{x})}{x_i} + O(\|\boldsymbol{y} - \boldsymbol{x}\|)\right) \end{aligned} \tag{25}$$

Then subtracting equation (24) to equation (25) yields

$$\begin{array}{l} V(\boldsymbol{x},\boldsymbol{\phi}(\boldsymbol{y})) - V(\boldsymbol{x},\boldsymbol{y}) \\ = \sum\limits_{i/x_i=0} (\phi_i(\boldsymbol{y}) - y_i)\,(p_i(\boldsymbol{x}) + O(\|\boldsymbol{y}-\boldsymbol{x}\|)) + \\ \sum\limits_{i/x_i>0} (\phi_i(\boldsymbol{y}) - y_i)(y_i - x_i + \frac{\phi_i(\boldsymbol{y})-y_i}{2})(\frac{p_i(\boldsymbol{x})}{x_i} + O(\|\boldsymbol{y}-\boldsymbol{x}\|)) \end{array} \tag{26}$$

However, equation (7) proves that $\phi_i(\boldsymbol{y}) - y_i = y_i\left(\left(\frac{m_i(\boldsymbol{y})}{p_i(\boldsymbol{y})}\right)^\eta - 1\right)$; in particular:

- if $x_i = 0$ and $\nabla_i J(\boldsymbol{x}) > 0$,
$$\phi_i(\boldsymbol{y}) - y_i = -y_i\left(1 - \left(\frac{m_i(\boldsymbol{x})}{p_i(\boldsymbol{x})}\right)^\eta + O(\|\boldsymbol{y}-\boldsymbol{x}\|)\right)$$
- if $x_i = 0$ and $\nabla_i J(\boldsymbol{x}) = 0$,
$$\phi_i(\boldsymbol{y}) - y_i = -\frac{\eta y_i}{p_i(\boldsymbol{x})}\left([\boldsymbol{\nabla}^2 J(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})]_i + O(\|\boldsymbol{y}-\boldsymbol{x}\|^2)\right)$$
- if $x_i > 0$ (and $\nabla_i J(\boldsymbol{x}) = 0$),
$$\phi_i(\boldsymbol{y}) - y_i = -\frac{\eta\, x_i}{p_i(\boldsymbol{x})}[\boldsymbol{\nabla}^2 J(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})]_i + O(\|\boldsymbol{y}-\boldsymbol{x}\|^2).$$

Substituting these three equalities into equation (26), plus a few manipulations show that (with the notations $[.]_0$ and $[.]_+$ introduced in section II-B)

$$\begin{array}{l} V(\boldsymbol{x},\boldsymbol{\phi}(\boldsymbol{y})) - V(\boldsymbol{x},\boldsymbol{y}) = -[\boldsymbol{y}]_0^T\,([\boldsymbol{v}(\boldsymbol{x})]_0 + O(\|\boldsymbol{y}-\boldsymbol{x}\|)) \\ -\eta\,[\boldsymbol{y}-\boldsymbol{x}]_+^T[M(\boldsymbol{x})]_+[\boldsymbol{y}-\boldsymbol{x}]_+ + O(\|\boldsymbol{y}-\boldsymbol{x}\|^3) \end{array} \tag{27}$$

where

$$[\boldsymbol{v}(\boldsymbol{x})]_0 = [\boldsymbol{I}_n - \boldsymbol{\Lambda}(\boldsymbol{x})^\eta]_0\,[\boldsymbol{p}(\boldsymbol{x})]_0$$

is a vector with (strictly) positive coefficients since $\eta > 0$, and

$$\begin{array}{ll} [M(\boldsymbol{x})]_+ & = [\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+ \\ & -\frac{\eta}{2}[\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+\left[\mathrm{diag}\left(\frac{\boldsymbol{x}}{\boldsymbol{p}(\boldsymbol{x})}\right)\right]_+[\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+ \end{array} \tag{28}$$

is a positive definite matrix since $\eta < \eta^\star$ (cf. lemma 16 below). Equation (27) finally proves that there is a neighborhood of $\boldsymbol{x}$ such that $\forall \boldsymbol{y} \neq \boldsymbol{x}$, $V(\boldsymbol{x},\boldsymbol{\phi}(\boldsymbol{y})) - V(\boldsymbol{x},\boldsymbol{y}) < 0$.

■

Lemma 16. *The matrix* $[\boldsymbol{M}(\boldsymbol{x})]_+$ *defined in equation* (28) *is positive definite if and only if* $\eta < \eta^\star$.

Proof of Lemma 16: Matrix $[\boldsymbol{M}(\boldsymbol{x})]_+$ is positive definite if and only if $[\boldsymbol{I}_n]_+ - \frac{\eta}{2}[\boldsymbol{P}'(\boldsymbol{x})]_+$ is positive definite, where

$$[\boldsymbol{P}'(\boldsymbol{x})]_+ = \left([\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+\right)^{\frac{1}{2}}\left[\mathrm{diag}\left(\frac{\boldsymbol{x}}{\boldsymbol{p}(\boldsymbol{x})}\right)\right]_+\left([\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+\right)^{\frac{1}{2}}$$

which is equivalent to $\eta < \frac{2}{\|[\boldsymbol{P}'(\boldsymbol{x})]_+\|_2}$. However, it is easy to prove that the eigenvalues of $[\boldsymbol{P}'(\boldsymbol{x})]_+$ are equal to those of

$$[\boldsymbol{P}(\boldsymbol{x})]_+ = [\boldsymbol{D}(\boldsymbol{x})]_+\,[\boldsymbol{\nabla}^2 J(\boldsymbol{x})]_+\,[\boldsymbol{D}(\boldsymbol{x})]_+\,.$$

Consequently, $\|[\boldsymbol{P}'(\boldsymbol{x})]_+\|_2 = \|[\boldsymbol{P}(\boldsymbol{x})]_+\|_2 = \|\boldsymbol{P}(\boldsymbol{x})\|_2$. ■

D. Application to standard unsupervised NMF

Proof of Theorem 15: First, Theorem 9 proves that any fixed point in $\mathcal{U}_x$ is Lyapunov stable. In particular, $\boldsymbol{x}$ is a Lyapunov stable fixed point. Let $\varepsilon > 0$ such that the closed ball of radius ε centered at $\boldsymbol{x}$ is included in $\mathcal{U}_x$. Then $\exists \delta > 0$ such that $\forall \boldsymbol{x}^{(0)} \in \mathcal{U}_x$, $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| < \delta \Rightarrow \|\boldsymbol{x}^{(p)} - \boldsymbol{x}\| < \varepsilon \; \forall p \in \mathbb{N}$. Since the sequence $\boldsymbol{x}^{(p)}$ is bounded, it admits an accumulation point: there is a subsequence $\boldsymbol{x}^{(p_q)}$ which converges to $\boldsymbol{x}^\star \in \mathcal{U}_x$. Since the whole sequence $V(\boldsymbol{x}, \boldsymbol{x}^{(p)})$ is non-increasing and lower-bounded, it converges. Thus $V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}^{(p_q)})) - V(\boldsymbol{x}, \boldsymbol{x}^{(p_q)}) \underset{q \to +\infty}{\longrightarrow} 0$. However, since functions V and $\boldsymbol{\phi}$ are continuous, $V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}^{(p_q)})) - V(\boldsymbol{x}, \boldsymbol{x}^{(p_q)}) \underset{q \to +\infty}{\longrightarrow} V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}^\star)) - V(\boldsymbol{x}, \boldsymbol{x}^\star)$. Therefore $V(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{x}^\star)) = V(\boldsymbol{x}, \boldsymbol{x}^\star)$, which proves that $\boldsymbol{\phi}(\boldsymbol{x}^\star) = \boldsymbol{x}^\star$. We have thus proved that $\boldsymbol{x}^\star$ is a fixed point of mapping $\boldsymbol{\phi}$ in $\mathcal{U}_x$.

Now let us prove that the whole sequence $\boldsymbol{x}^{(p)}$ converges to $\boldsymbol{x}^\star$. Let $\varepsilon > 0$: since $\boldsymbol{x}^\star$ is Lyapunov stable, $\exists \delta > 0$ such that $\forall \boldsymbol{y}^{(0)}$, $\|\boldsymbol{y}^{(0)} - \boldsymbol{x}^\star\| < \delta \Rightarrow \|\boldsymbol{y}^{(p)} - \boldsymbol{x}^\star\| < \varepsilon \; \forall p \in \mathbb{N}$. However, since $\boldsymbol{x}^\star$ is the limit of the sequence $\boldsymbol{x}^{(p_q)}$, $\exists q$ such that $\|\boldsymbol{x}^{(p_q)} - \boldsymbol{x}^\star\| < \delta$. Applying the previous statement to $\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(p_q)}$ proves that $\forall p \geq p_q$, $\|\boldsymbol{x}^{(p)} - \boldsymbol{x}^\star\| < \varepsilon$. This finally proves that $\boldsymbol{x}^{(p)} \underset{p \to +\infty}{\longrightarrow} \boldsymbol{x}^\star$. ■

REFERENCES

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

[2] ——, "Algorithms for non-negative matrix factorization," in *Proc. of the Conference on Advances in Neural Information Processing Systems*, vol. 13. Vancouver, British Columbia, Canada: MIT Press, Dec. 2001, pp. 556Ű–562.

[3] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. of 6th International Congress on Acoustics*, Tokyo, Japan, Aug. 1968, pp. C17–C20.

[4] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep., 2001. [Online]. Available: http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf

[5] A. Cichocki, R. Zdunek, and S. Amari, "CsiszarŠs divergences for non-negative matrix factorization : Family of new algorithms," in *Proc. of 6th International Conference on Independent Component Analysis and Blind Signal Separation*, Charleston, SC, USA, Mar. 2006, pp. 32–39.

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

[7] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[8] R. Zdunek and A. Cichocki, "Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems," *Computational Intelligence and Neuroscience*, vol. 2008, pp. 1–13, 2008, article ID 939567.

[9] W. Wang and X. Zou, "Non-negative matrix factorization based on projected nonlinear conjugate gradient algorithm," in *Proc. of ICA Research Network International Workshop (ICARN 2008)*, Liverpool, UK, Sept. 2008, pp. 5–8.

[10] L. Finesso and P. Spreij, "Approximate nonnegative matrix factorization via alternating minimization," in *Proc. of the 16th International Symposium on Mathematical Theory of Networks and Systems*, Louvain, Belgium, July 2004.

[11] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating non-negativity-constrained least squares and the active set method," *SIAM Journal in Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, 2008.

[12] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Processing*, vol. 87, no. 8, pp. 1904–1916, 2007.

[13] A. Cichocki, R. Zdunek, and S. Amari, "Nonnegative matrix and tensor factorization," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 142–145, Jan. 2008.

[14] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended SMART algorithms for non-negative matrix factorization," *Springer LNAI*, vol. 4029, pp. 548–562, 2006.

[15] 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, no. 3, pp. 793–830, Mar. 2009.

[16] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. of IEEE International Joint Conference on Neural Networks*, Orlando, Florida, USA, June 2008, pp. 1828–1832.

[17] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, Dec. 2001, pp. 207–Ű212.

[18] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. of IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, July 2004, pp. 2529Ű–2533.

[19] J. Eggert, H. Wersing, and E. Korner, "Transformation-invariant representation and NMF," in *Proc. of IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, July 2004, pp. 2535–2539.

[20] S. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, Sept. 2007.

[21] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, 2009, to be pusblished. [Online]. Available: http://hal.inria.fr/inria-00350163/fr/

[22] 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*, 2009, to be published. [Online]. Available: http://www.tsi.enst.fr/publications/enst/techreport-2009-8812.pdf

[23] ——, “Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription,” in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, NY, USA, Oct. 2009, pp. 29–32.

[24] 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, no. 3, pp. 1066Ű–1074, Mar. 2007.

[25] Z. Chen, A. Cichocki, and T. M. Rutkowski, “Constrained non-negative matrix factorization method for EEG analysis in early detection of AlzheimerŠs disease,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006, pp. 893Ű–896.

[26] Y. Zhang and Y. Fang, “A NMF algorithm for blind separation of uncorrelated signals,” in *Proc. of IEEE International Conference on Wavelet Analysis and Pattern Recognition*, Beijing, China, Nov. 2007, pp. 999–1003.

[27] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Proc. of International Conference on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, Sept. 2004, pp. 494Ű–499.

[28] M. Shashanka, B. Raj, and P. Smaragdis, “Probabilistic latent variable models as nonnegative factorizations,” *Computational Intelligence and Neuroscience*, vol. 2008, pp. 1–8, 2008, article ID 947438.

[29] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, “Algorithms and applications for approximate nonnegative matrix factorization,” *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[30] M. Chu, F. Diele, R. Plemmons, and S. Ragni, “Optimality, computation, and interpretation of nonnegative matrix factorizations,” Wake Forest University, North Carolina, USA, Tech. Rep., Oct. 2004.

[31] C.-J. Lin, “On the convergence of multiplicative update algorithms for nonnegative matrix factorization,” *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–Ű1596, Nov. 2007.

[32] E. Gonzalez and Y. Zhang, “Accelerating the Lee-Seung algorithm for nonnegative matrix factorization,” TR-05-02, Rice University, Houston, Texas, USA, Tech. Rep., Mar. 2005.

[33] E. Vincent, “Musical source separation using time-frequency source priors,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 91–98, Jan. 2006.

[34] A. Cont, “Realtime multiple pitch observation using sparse non-negative constraints,” in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2006, pp. 206Ű–212.

[35] G. Peyré, “Non-negative sparse modeling of textures,” in *Proc. Scale Space and Variational Methods in Computer Vision (SSVM)*, 2007, pp. 628Ű–639.

[36] D. Guillamet, J. Vitrià, and B. Schiele, “Introducing a weighted non-negative matrix factorization for image classification,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447–2454, 2003.

[37] N. Bertin and R. Badeau, “Initialization, distances and local minima in audio applications of the non-negative matrix factorization,” in *Proc. of Acoustics’08*. Paris, FRANCE: JASA, July 2008. [Online]. Available: http://perso.telecom-paristech.fr/nbertin/publis/acoustics08.pdf

[38] N. Bertin, C. Févotte, and R. Badeau, “A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 1545–1548.

[39] J. Lasalle, *The stability and control of discrete processes*. New York, NY, USA: Springer-Verlag, 1986.

[40] R. Badeau, N. Bertin, and E. Vincent, “Supporting document for the paper "On the stability of multiplicative update algorithms. Application to non-negative matrix factorization.",” Télécom ParisTech, Tech. Rep., Nov. 2009.

[41] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., ser. Springer Series in Operations Research and Financial Engineering. New York Inc.: Springer-Verlag, Aug. 2006.

[42] S. Bhat and D. Bernstein, “Lyapunov analysis of semistability,” in *Proc. of American Control Conference*, vol. 3, San Diego, California, USA, June 1999, pp. 1608–1612.

Dépôt légal : 2009 – 4ème trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI