Unified Development of Multiplicative Algorithms for Linear and Quadratic Nonnegative Matrix Factorization

Zhirong Yang, *Member, IEEE,* and Erkki Oja, *Fellow, IEEE,*

***Abstract*—Multiplicative updates have been widely used in approximative Nonnegative Matrix Factorization (NMF) optimization because they are convenient to deploy. Their convergence proof is usually based on the minimization of an auxiliary upper-bounding function, the construction of which however remains specific and only available for limited types of dissimilarity measures. Here we make significant progress in developing convergent multiplicative algorithms for NMF. First, we propose a general approach to derive the auxiliary function for a wide variety of NMF problems, as long as the approximation objective can be expressed as a finite sum of monomials with real exponents. Multiplicative algorithms with theoretical guarantee of monotonically decreasing objective function sequence can thus be obtained. The solutions of NMF based on most commonly used dissimilarity measures such as α- and β-divergence as well as many other more comprehensive divergences can be derived by the new unified principle. Second, our method is extended to a non-separable case that includes e.g. γ-divergence and Rényi divergence. Third, we develop multiplicative algorithms for NMF using second-order approximative factorizations, in which each factorizing matrix may appear twice. Preliminary numerical experiments demonstrate that the multiplicative algorithms developed using the proposed procedure can achieve satisfactory KKT optimality. We also demonstrate NMF problems where algorithms by the conventional method fail to guarantee descent at each iteration but those by our principle are immune to such violation.**

***Index Terms*—multiplicative, nonnegative, matrix factorization, divergence, optimization.**

I. Introduction

NONNEGATIVE Matrix Factorization (NMF) has attracted a lot of research effort in the past decade. Since Lee and Seung [1], [2] advocated the use of nonnegativity constraint in approximative matrix factorization with two convenient algorithms, much progress has been made by adopting various dissimilarity measures, approximation schemes, and constraints to this problem (e.g. [3], [4], [5], [6], [7], [8], [9], [10]). NMF has also been successfully applied in many fields, including text, speech, music, bioinformatics, neuroinformatics, etc. (e.g. [11], [8], [12], [13]). See [14] for a survey.

When computing the low-rank factorization, many existing NMF algorithms employ multiplicative updates where the new estimate is obtained by element-wise product with nonnegative factors. A generic principle for forming the multiplying factors is widely used: the gradient of the approximation error with respect to one of the factorizing matrices is first computed and then the sums of positive and unsigned negative terms of the gradient are respectively placed in the numerator and denominator of a ratio. The same principle can in turn be applied to the other factorizing matrices. The resulting multiplicative update rules have some advantages over conventional additive gradient descent approach. Firstly, the rules naturally maintain the nonnegativity of the factorizing matrices without any extra projection steps. Secondly, the fixed-point algorithm that iteratively applies the update rule requires no user-specified parameters such as the learning step size, which facilitates its implementation and applications. Though recently faster additive optimization methods for particular NMF objectives, especially for the least-square errors, have been developed (e.g. [15], [13], [16], [17]), multiplicative updates as a more convenient optimization method are still widely adopted by many NMF applications.

The above heuristic principle was previously justified by using the Karush-Kuhn-Tucker (KKT) conditions [5], [6] or the natural gradients [18]. However, these justifications cannot provide theoretical guarantee that the resulting updates will monotonically decrease the approximation error. A counter-example is the NMF based on the α-divergence. On the other hand, it is known that some multiplicative update rules which do guarantee monotonicity do not use this form of multiplying with ratios, for example, the NMF based on the dual I-divergence (see e.g. [5]).

To obtain a theoretical guarantee of the monotonic decrease, a commonly used method is to construct an auxiliary function that globally upper bounds the objective, and then the multiplicative update should minimize the auxiliary function. Previously the construction of such an auxiliary function seemed to be challenging and only successful for a few types of objectives. For example, Lee and Seung employed two different inequalities for NMF based on the Euclidean distance and the I-divergence, respectively, where the first approach uses an inequality for the positive quadratic term, while the second one uses the Jensen inequality on the logarithm function.

Recently, two unifying methods have been introduced that make further use of convexity and employ the Jensen inequality beyond the logarithm. Dhillon and Sra [5] presented a general auxiliary function for Bregman divergences between the approximation $\widehat{\mathbf{X}}$ and the input $\mathbf{X}$, i.e. $D_\phi(\mathbf{X}||\widehat{\mathbf{X}})$ or $D_\phi(\widehat{\mathbf{X}}||\mathbf{X})$. Their method is nevertheless only applicable to particular Bregman divergences, especially the latter case

Z. Yang and E. Oja are with Department of Information and Computer Science, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland. e-mail: {zhirong.yang,erkki.oja}@aalto.fi

$D_\phi(\widehat{\mathbf{X}}||\mathbf{X})$. In another work [6], Cichocki et al. proposed multiplicative update rules for the family of α-divergences, except the dual I-divergence, by using convexity of the α-function.

Despite these efforts, there are a number of problems remaining. Firstly, there are many NMF objectives that are not convex, for example the β-divergence with very large (or very small) β [4], [14]. Then these two methods do not apply. Secondly, even for convex objectives, the above approaches may still fail because they require that the derivative of the underlying convex function must be decomposable. Otherwise the solution of the stationary point equation in general has no analytical form; see Eq. 3.3 in [5] for an example. Moreover, the monotonicity proofs for objectives that are non-separable over matrix elements, for example the γ-divergence [19] and Rényi divergence are still lacking.

To address such problems, we present here a novel unified procedure for developing multiplicative NMF algorithms. After this introductory part and a brief review in the next section, we propose in Section III a much more general principle for deriving multiplicative update rules that guarantee monotonically decreasing approximation errors. Our method works for any separable NMF objective as long as it can be expressed as a finite sum of monomials with real exponents. With such expressions, all we need for most commonly used NMF objectives are two well-known inequalities based on the convexity or concavity of each monomial. For comprehensive objectives that comprise more than two convex monomials, we introduce a novel inequality by which one can merge the monomials into two terms. Using this generic principle, we derive the ensuing multiplicative update rules with theoretical monotonicity guarantee for a large number of objectives with various divergence measures. Our rules match all the existing ones whose monotonicity has been earlier proven in literature using a number of specific approaches. Furthermore, novel rules for many other NMF problems can be obtained.

In Section IV, the proposed technique is extended to a non-separable case where the power operation appears over the sum of approximating matrix elements. As a result, the multiplicative update rules as well as their auxiliary functions of most existing NMF optimization problems can be derived using the proposed principle, including the α-divergence, β-divergence and all their special cases. Furthermore, many new multiplicative update rules with theoretical guarantee are given for NMF based on γ-divergence and Rényi divergence, their special cases including normalized Kullback-Leibler divergence, and some other more comprehensive approximation error measures.

In Section V, we generalize the proposed principle to NMF using quadratic factorizations where a factorizing matrix may appear twice in the approximation. This may be useful e.g. in the graph isomorphism problem. A multitude of multiplicative algorithms with monotonicity guarantee for both symmetric and asymmetric quadratic NMF problems can be obtained.

Finally, empirical results in Section VI show that (1) most multiplicative algorithms derived by using our method can asymptotically achieve the KKT optimality; (2) there exist NMF problems where multiplicative algorithms by the conventional heuristic principle can cause undesired increase of approximation error, whereas those by our method always guarantee descent. In Section VII we conclude the paper and discuss the future work.

II. Nonnegative Matrix Factorization and Multiplicative Updates

Given an input matrix $\mathbf{X} \in \mathbb{R}_+^{m\times n}$, Nonnegative Matrix Factorization (NMF) seeks its low-rank approximation in the form $\widehat{\mathbf{X}} = \mathbf{WH}$, where $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$, with $r < \min(m, n)$. Besides the linear factorization where each factorizing matrix only appears once in the approximation, the factorization can be of higher-order where each factorizing matrix may appear more than once (see Section V).

The approximation error between the input matrix and its approximation can be measured by various divergences $D(\mathbf{X}||\widehat{\mathbf{X}})$. Typical measures include the Euclidean distance (Frobenius norm) and the non-normalized Kullback-Leibler divergence (I-divergence). NMF was later generalized to other divergences such as α-divergence [6], β-divergence [4], [20], Csiszár divergences [21] and Bregman divergences [5]. More example divergences are summarized in Table IV. For brevity, in this paper we only consider minimization over the matrix $\mathbf{W}$, while exactly the same technique can be applied to other factorizing matrices in a very similar way, leading to alternating minimization algorithms. In the derivation, we use $\mathbf{W}$, $\mathbf{W}^{\text{new}}$, and $\widetilde{\mathbf{W}}$ to distinguish the current estimate, the new estimate and the variable, respectively. Accordingly, we also use $\widetilde{\mathbf{X}}$ in place of $\widehat{\mathbf{X}}$ wherever we emphasize that the approximation contains the variable.

The conventional gradient descent method does not work for the NMF problem because the updated estimate after each iteration is not necessarily nonnegative. Projecting the estimate back to the positive hyper-quadrant is therefore needed after each update. Line search by a number of objective evaluations is often needed to guarantee descent after each update, which can be expensive for NMF problems and thus only works for a few particular types of objectives (e.g. [15], [16], [17]).

Using multiplicative updates is a more convenient optimization method for NMF, because it can be easily adapted from the gradient while naturally maintaining the nonnegativity. It requires no extra efforts to tune learning parameters such as the step size.

Let ∇ be the gradient of the approximation objective function with respect to $\mathbf{W}$, and denote by ∇^+ and ∇^- the sums of positive and unsigned negative terms, i.e. $\nabla = \nabla^+ - \nabla^-$. It used to be common belief that the multiplicative update

$$W_{ik}^{\text{new}} = W_{ik}\frac{\nabla_{ik}^-}{\nabla_{ik}^+}, \tag{1}$$

can minimize the NMF objectives (see e.g. [22]). The above update rule is connected to the steepest gradient descent method $W_{ik}^{\text{new}} = W_{ik} - \mu_{ik}\left(\nabla_{ik}^+ - \nabla_{ik}^-\right)$ by assuming that each matrix element has its own learning rate $\mu_{ik} = W_{ik}/\nabla_{ik}^+$. A recent similar justification [18] interprets the multiplicative update rule as a natural gradient descent with a unitary learning rate, where the underlying Riemannian manifold is

defined by the tensor $[\mathcal{G}(\mathbf{W})]_{ijkl} = \delta_{ij}\delta_{kl}\frac{\nabla^+_{ik}}{W_{ik}}$, with δ_{ij} the Kronecker delta. Yet another way to suggest the principle (1) is by rearranging the Karush-Kuhn-Tucker (K.K.T.) conditions $(\nabla^+ - \nabla^-)_{ik} W_{ik} = 0$ [5], [6]. However, actually none of the above justifications can provide theoretical guarantee that the resulting updates will monotonically decrease the objective function. Counter-examples include $D_\alpha(\mathbf{X}||\mathbf{WH})$ and $D_{\text{EU}}(\mathbf{X}||\mathbf{WW}^T\mathbf{X})$ (see Table IV), where the updates using Eq. (1) may increase the approximation error. Furthermore, it is known that some multiplicative update rules which do guarantee monotonicity do not take the form of Eq. (1), for example, the one for $D_{\text{I}}(\mathbf{WH}||\mathbf{X})$.

The monotonicity guarantee is very important for fixed-point algorithms in non-convex optimization problems, including the multiplicative updates and also e.g. the well-known Expectation-Maximization method. It is a basic fact that a lower-bounded monotonically decreasing sequence is convergent. Since the NMF approximation error is generally lower-bounded by zero, the proof of objective convergence in NMF is reduced to finding the theoretical guarantee of monotonicity. Unless otherwise stated, the term "convergence" in this paper generally refers to the *objective function convergence* or, equivalently, the monotonic decrease of the NMF approximation error. Numerical experiments for checking the *point convergence*, or the optimality conditions of the points of convergence, are provided in Section VI-A.

Currently the auxiliary function technique [2] is the most widely accepted approach for monotonicity proof of multiplicative updates. Given an objective function $\mathcal{J}(\mathbf{W})$ to be minimized, $G(\mathbf{W}, \mathbf{U})$ is called an auxiliary function if it is a tight upper bound of $\mathcal{J}(\mathbf{W})$, i.e. $G(\mathbf{W}, \mathbf{U}) \geq \mathcal{J}(\mathbf{W})$, and $G(\mathbf{W}, \mathbf{W}) = \mathcal{J}(\mathbf{W})$ for any $\mathbf{W}$ and $\mathbf{U}$. Define

$$\mathbf{W}^{\text{new}} = \arg\min_{\widetilde{\mathbf{W}}} G(\widetilde{\mathbf{W}}, \mathbf{W}). \tag{2}$$

By construction, $\mathcal{J}(\mathbf{W}) = G(\mathbf{W}, \mathbf{W}) \geq G(\mathbf{W}^{\text{new}}, \mathbf{W}) \geq G(\mathbf{W}^{\text{new}}, \mathbf{W}^{\text{new}}) = \mathcal{J}(\mathbf{W}^{\text{new}})$, where the first inequality is the result of minimization and the second comes from the upper bound. Iteratively applying the update rule (2) thus results in a monotonically decreasing sequence of $\mathcal{J}$. Besides the tight upper bound, it is often desired that the minimization (2) has a closed-form solution. In particular, setting $\partial G/\partial\widetilde{\mathbf{W}} = 0$ should lead to the iterative update rule in analysis. The above technique is also named *Majorization-Minimization* (MM) in the optimization literature (see e.g. [23]). Another related generic principle is *Difference of Convex functions* (DC) programming (see e.g. [24]). Especially, as a requirement for NMF problems, the factorizing matrices after applying the update rule should maintain the nonnegativity.

The majorization or construction of such an auxiliary function, however, has not been a trivial task so far. Though generic for a wide range of optimization tasks, there is, however, no specific principle in the MM or DC literature for deriving the auxiliary function by exploiting the structure and constraint of NMF problems. Improper majorization may lead to comprehensive programming steps instead of simple multiplicative update rules that are easy to implement. Another plausible approach is to construct a quadratic upper-bound by using the Lipschitz constant if the NMF objective is a C^2 function, though the resulting updates are often very slow and require extra steps to maintain the nonnegativity. Some specific methods have been used in particular NMF objectives, for example, the fourth and fifth inequalities given in Appendix A for the quadratic terms (e.g. [2], [10], [25], [9]), the Jensen inequality on the logarithm [2] or α-function [6], and the Concave-Convex Procedure (CCCP) for majorizing β-divergence [26], [20].

In the following we present a common principle that not only unifies the above proofs but also is easily generalized to other unproven objectives or higher-order factorizations. As a result, our method turns the auxiliary function construction that seemingly requires intense intelligent work into a simple symbolic manipulation procedure.

III. Multiplicative algorithms for separable divergences

Before going into details, we need generalized definitions of monomials and polynomials. In this work, a monomial with real-valued exponent, or *monomial* by short, of the scalar variable z is of the form az^b where a and b can take any real value, without restriction to nonnegative integers. A sum of a (finite) number of monomials is called a (finite) *generalized polynomial*.

NMF objectives that can be expressed in the generalized polynomial form have two nice properties: 1) individual monomials, denoted by

$$\omega_d(\widetilde{X}_{ij}) = f_{dij}\widetilde{X}_{ij}^{\tau_d}, \tag{3}$$

are either convex or concave with respect to $\widetilde{\mathbf{W}}$ and thus can easily be upper-bounded; 2) an exponential is multiplicatively decomposable, i.e. $(xy)^\tau = x^\tau y^\tau$, which is critical in deriving the multiplicative update rule. We thus apply a two-step strategy for obtaining a multiplicative algorithm with monotonic objective convergence: first, we construct individual upper-bounds for each monomial according to its convexity or concavity; second, the auxiliary function is obtained by merging individual upper-bounds to monomials of two different exponents.

The finite generalized polynomial form covers a large variety of separable dissimilarity measures used in NMF, for example, the Euclidean distance (Frobenius norm), the I-divergence (non-normalized Kullback-Leibler divergence), the dual I-divergence, the Hellinger distance, the Itakura-Saito divergence, the Log-Quad cost, as well as many other unnamed Csiszár divergences and Bregman divergences. Some example objectives and their finite generalized polynomials are shown in Table IV.

Many information-theoretic divergences involve the logarithm function. We can unify it to our generalized polynomial form by using the limit

$$\ln z = \lim_{\epsilon\to 0^+} \frac{z^\epsilon - 1}{\epsilon}. \tag{4}$$

Notice that limits in 0^+ and 0^- are the same. We use the former to remove the convexity ambiguity. In this way, the

logarithm can be decomposed into two monomials where the first contains an infinitesimal positive exponent. Because most existing divergences that contain the logarithm are smooth with respect to both the factorizing matrix $\widetilde{\mathbf{W}}$ and $\epsilon > 0$, we can safely exchange the order of the logarithm limit and the derivative with respect to $\widetilde{\mathbf{W}}$. In other words, the deriving procedure is standard, after rewriting the logarithm as monomials. Upon upper-bounding and taking derivative with respect to $\widetilde{\mathbf{W}}$, the limit operator $\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\{\cdot\}$ is applied to obtain the gradient of the auxiliary function.

Next we formally show that multiplicative algorithms with monotonicity guarantee always exist as long as the approximation objective function is separable over matrix elements and can be expressed as a sum of a finite number of monomials with real-valued exponents.

A. The auxiliary upper bounding function

Theorem 1: Denote $\widetilde{\mathbf{X}} = \widetilde{\mathbf{W}}\mathbf{H}$, $2 \le p < \infty$, $\tau_d \in \mathbb{R}$, $d = 1, \ldots, p$ and f_{dij} constants independent of $\widetilde{\mathbf{W}}$. Suppose

1) the approximation objective is separable over indices i and j, i.e.

$$D\left(\mathbf{X}||\widetilde{\mathbf{X}}\right) = \sum_{d=1}^{p}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{dij}\widetilde{X}_{ij}^{\tau_d} + \text{constant}; \quad (5)$$

2) there is at least one non-zero stationary point of D with respect to $\widetilde{\mathbf{W}}$; and
3) $\forall\ i, k,\ W_{ik} > 0$.

There are real numbers $\psi_{\max}$ and $\psi_{\min}$ ($\psi_{\max} > \psi_{\min}$) such that

$$G(\widetilde{\mathbf{W}}, \mathbf{W}) = \sum_{ik}\left[\frac{W_{ik}}{\psi_{\max}}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\max}}\nabla_{ik}^{+} - \frac{W_{ik}}{\psi_{\min}}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\min}}\nabla_{ik}^{-}\right] + \text{constant}, \quad (6)$$

is an auxiliary upper-bounding function of $D\left(\mathbf{X}||\widetilde{\mathbf{X}}\right)$, where ∇^+ and ∇^- are the sums of positive and unsigned negative terms of ∇, the gradient of the divergence with respect to $\widetilde{\mathbf{W}}$ at the current estimate $\mathbf{W}$, respectively (i.e. $\nabla = \nabla^+ - \nabla^-$).

Proof:

Step 1: Upper-bounding individual monomials

The objective function in the form (5) is a sum of either convex or concave monomials. The *concave* monomials can be upper bounded by using the first-order Taylor expansion at the current estimate:

$$\sum_{ij}\omega_d(\widetilde{X}_{ij}) \le \sum_{ik}\partial_{dik}\widetilde{W}_{ik} + \text{constant}, \quad (7)$$

where

$$\partial_{dik} \triangleq \left.\frac{\partial\sum_{ab}\omega_d(\widetilde{X}_{ab})}{\partial\widetilde{W}_{ik}}\right|_{\widetilde{\mathbf{W}}=\mathbf{W}} = \sum_j f_{dij}\tau_d\left(\mathbf{WH}\right)_{ij}^{\tau_d-1}H_{kj}. \quad (8)$$

For *convex* monomials, we introduce $\lambda_{ijk} = \frac{W_{ik}H_{kj}}{(\mathbf{WH})_{ij}}$ and then obtain their upper bound by using the Jensen inequality:

$$\sum_{ij}\omega_d(\widetilde{X}_{ij}) \le \sum_{ij} f_{dij}\sum_k \lambda_{ijk}\left(\frac{\widetilde{W}_{ik}H_{kj}}{\lambda_{ijk}}\right)^{\tau_d} = \sum_{ik}\frac{W_{ik}}{\tau_d}\partial_{dik}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\tau_d}. \quad (9)$$

Step 2: Combining individual upper-bounds

A total upper-bounding function can be obtained by summing up individual ones obtained from the previous subsection. For $p = 2$ and $\tau_1 \neq \tau_2$, their sum already forms an auxiliary function of the form given in Theorem 1. This is the case for most existing NMF algorithms based e.g. α-divergence [6], β-divergence [26], and their special cases.

However, when there are more than two monomials with distinct exponents, the summed upper-bound does not have the form given in Theorem 1. An example is the Log-Quad cost shown in Table IV. It is generally difficult to solve the equation by setting the derivative of the summed upper-bound to zero. The complexity of analytical form of the roots grows drastically when the number of such monomials increases and therefore does not lead to desired multiplicative update rules.

We therefore consider merging the individual upper-bounds into an upper-bounding function that contains monomials of only two different exponents such that the derivation can continue with the $p = 2$ case. The merging is implemented by further upper-bounding the monomials. In Appendix B we prove the following result:

Lemma 2: Assume cx^a convex. Then $cx^a \le ac\frac{x^b}{b} + c\left(1 - \frac{a}{b}\right)$ if one of the following holds:

1) $a > 1$ and $a < b$,
2) $a < 1$ and $a > b$,
3) $a = 1$, $c > 0$ and $a < b$,
4) $a = 1$, $c < 0$ and $a > b$.

The equality holds if and only if $x = 1$.

For our merging purpose, notice that all individual upper-bounding monomials are convex and have the form

$$\frac{W_{ik}\partial_{dik}}{\psi_d}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_d} \quad (10)$$

where $\psi_d = \tau_d$ for the convex upper-bounds and $\psi_d = 1$ for the concave upper-bounds. We can then further upper bound these individual monomials according to the cases in Lemma 2 as follows. Denote $\psi_{\max}$ and $\psi_{\min}$ the maximum and minimum of $\{\psi_d\}_{d=1}^p$, respectively.

- *Up-Merging* (UM): for 1) $\psi_d > 1$ or 3) $\psi_d = 1$ and $W_{ik}\partial_{dik}/\psi_d > 0$, up to some additive constant,

$$\frac{W_{ik}\partial_{dik}}{\psi_d}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_d} \le \frac{W_{ik}\partial_{dik}}{\psi_{\max}}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\max}}. \quad (11)$$

- *Down-Merging* (DM): for 2) $\psi_d < 1$ or 4) $\psi_d = 1$ and $W_{ik}\partial_{dik}/\psi_d < 0$, up to some additive constant,

$$\frac{W_{ik}\partial_{dik}}{\psi_d}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_d} \le \frac{W_{ik}\partial_{dik}}{\psi_{\min}}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\min}}. \quad (12)$$

Because the individual upper-bounding monomials are convex, we have $\partial_{dik} > 0$ if $d \in$ UM and $\partial_{dik} < 0$ if $d \in$ DM. Moreover, by such assignment, UM and DM cases must both be present. Otherwise the gradient $\nabla_{ik} = \sum_{dik} \partial_{dik}$ would be always positive (or always negative) for all $\mathbf{W}$ and (i,k), which makes the origin the only stationary point and therefore contradicts the second theorem assumption. We thus obtain the auxiliary function (6) with $\nabla^+_{ik} = \sum_{d\in\text{UM}} \partial_{dik}$ and $\nabla^-_{ik} = -\sum_{d\in\text{DM}} \partial_{dik}$.

Finally, all upper bounds used in the above steps come from one of the five inequalities given in Appendix A. It is worth to notice that all these upper-bounds are tight at $\widetilde{\mathbf{W}} = \mathbf{W}$, i.e. $G(\mathbf{W},\mathbf{W}) = \mathcal{J}(\mathbf{W})$. Therefore, the ultimate upper bound does form an auxiliary function. ■

B. The multiplicative update rule

The derivative of the auxiliary function (6) with respect to $\widetilde{W}_{ik}$ is given by

$$\frac{\partial G(\widetilde{\mathbf{W}},\mathbf{W})}{\partial \widetilde{W}_{ik}} = \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\text{max}}-1} \nabla^+_{ik} - \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\text{min}}-1} \nabla^-_{ik}. \tag{13}$$

For most cases, setting this derivative to zero yields the update rule

$$W^{\text{new}}_{ik} = W_{ik}\left(\frac{\nabla^-_{ik}}{\nabla^+_{ik}}\right)^{\eta} \tag{14}$$

where $\eta = 1/(\psi_{\text{max}} - \psi_{\text{min}})$.

An exception is when the logarithm limit (4) is applied and $\lim_{\epsilon\to 0+} (\psi_{\text{max}} - \psi_{\text{min}}) = 0$. In this case the limit of the derivative (13) has the form $\frac{0}{0}$. We thus apply the L'Hôpital's rule [27, pages 201-202] to obtain the limit before setting it to zero. See Section D-B for an example on the dual I-divergence.

The multiplicative update rule (14) also guarantees strictly positive descent as long as the current estimate is not a stationary point. Otherwise $\mathbf{W}$ will remain unchanged due to $W_{ik} = 0$ or $\nabla_{ik} = 0$ for all (i,k).

Corollary 3: If $\mathbf{W}$ is not a stationary point of $\mathcal{J}(\mathbf{W})$, then for Eq. (14), $\mathcal{J}(\mathbf{W}^{\text{new}}) < \mathcal{J}(\mathbf{W})$.

Proof: Because G is tangent to $\mathcal{J}$ at $\mathbf{W}$, which is not a stationary point of $\mathcal{J}$, $\mathbf{W}$ is not a stationary point of G either. Meanwhile, $\mathbf{W}^{\text{new}}$ as the stationary point also achieves the minimum of G because G is a convex function in $\widetilde{\mathbf{W}}$. Then by Theorem 1, $\mathcal{J}(\mathbf{W}) = G(\mathbf{W},\mathbf{W}) > G(\mathbf{W}^{\text{new}},\mathbf{W}) \geq G(\mathbf{W}^{\text{new}},\mathbf{W}^{\text{new}}) = \mathcal{J}(\mathbf{W}^{\text{new}})$. ■

Besides the generalized monomials, there are two mild assumptions in Theorem 1: one requires that the objective has at least one non-trivial local minimum and the other constrains that $\mathbf{W}$ contains no zero entries such that G is well-defined. The latter assumption is commonly needed by all multiplicative algorithms because multiplicative updates make no changes to zero entries. This requirement in practice can be easily fulfilled by positively initializing factorizing matrices such that everything on the right-hand side of Eq. (14) generally remains positive, and so does the updated $\mathbf{W}$.

C. Summary of the derivation procedure

The principle of deriving a multiplicative update rule and the corresponding auxiliary function of a given separable objective is summarized as the following procedure.

1) Transform the objective function into the form of finite generalized polynomials. Use the limit form (4) wherever the objective comprises the logarithm.
2) Upper bound each monomial according to their concavity or convexity by using their first-order Taylor expansion (7) or the Jensen inequality (9), respectively.
3) If there are three or more individual upper-bounds, combine them into two monomials using (11) or (12) according to their exponents. Form the auxiliary function.
4) Take the derivative of the auxiliary function with respect to the factorizing matrix variable.
5) Apply the logarithm limit if needed. Employ L'Hôpital's rule when the limit has the form $\frac{0}{0}$.
6) Obtain the multiplicative update rule by setting the derivative to zero.

In Appendix D, we present the derivation details of five example multiplicative update rules for readers' better understanding of the deriving procedure.

It is interesting to see that the derived rules using our generic principle *match all existing ones* whose objective monotonicity has been theoretically proven using other specific approaches [2], [5], [6], [4], [20]. For other separable divergences listed in Table IV, the multiplicative update rules take the form in Eq. (14) with $\eta = 1$ for Log-Quad cost and $\eta = 1/(\alpha-\beta+1)$ for $\alpha\beta$-Bregman divergence. The corresponding update rules for the examples is given in Table V.

A couple of remarks should be addressed for the resulting exponent η in Eq. (14). Firstly, although η plays a role similar to step size, it has two distinguishing properties from the conventional methods such as exponentiated gradient descent (see e.g. [28]): (1) users need no extra effort to tune η, as it is uniquely determined by the NMF objective and our proof procedure; (2) η is not required to approach zero for monotonicity guarantee, as shown by the examples in Appendix D. This is a major difference between our multiplicative algorithms and the conventional exponential gradient descent method because in the latter method, to our knowledge, there is no means to obtain a constant learning rate that guarantees monotonic objective decrease and one has to choose a very small step size to avoid monotonicity violation. This advantage also provides the base for an acceleration strategy using adaptive exponents, for example, to use more aggressive exponents and switch back to the safe choice whenever ascent occurs (see e.g. [29]).

Secondly, η actually defines the upper bound of a safe interval of exponents that guarantee monotonicity given our mild assumption and majorization steps based on convexity/concavity. That is, the update rule (14) still makes the objective function converge when η is replaced with a smaller positive η^*. To see this, one can apply an even further upper

bound

$$\frac{W_{ik}\partial_{dik}}{\psi_{\max}}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi_{\max}} \leq \frac{W_{ik}\partial_{dik}}{\psi^*}\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\psi^*} \tag{15}$$

with $\psi^* > \psi_{\max}$. This leads to a multiplicative update rule similar to (14) except the exponent $\eta = \frac{1}{\psi_{\max}-\psi_{\min}}$ changes to $\eta^* = \frac{1}{\psi^*-\psi_{\min}}$, which is still an update rule with monotonicity guarantee. Similar looser bounding can be applied by replacing $\psi_{\min}$ to obtain the same result. Nevertheless, a smaller exponent corresponds to more conservative multiplicative learning steps, which often leads to slower descent speed in practice. In this sense, η is the most aggressive and safe constant exponent provided by our method.

Note that unlike the heuristic rule (1), there is no ambiguity when decomposing the gradient into positive and negative parts in our method. The decomposition is uniquely determined by the proposed procedure. If one adds and subtracts the same positive constant to the gradient, the resulting multiplicative update rule essentially requires a looser upper bound and thus leads to slower convergence.

IV. A NON-SEPARABLE CASE

We also consider a non-separable case

$$D(\mathbf{X}||\widetilde{\mathbf{X}}) = \sum_d \Omega_d\left(\sum_{ij} g_{dij}\widetilde{X}_{ij}^{\phi_d}\right) \tag{16}$$

where $\Omega_d(z) = \nu_d \cdot z^{\tau_d}$ with ν_d, τ_d, g_{dij} and ϕ_d constants independent of $\widetilde{\mathbf{W}}$. A typical example is the family of γ-divergences, with the original or normalized Kullback-Leibler divergence as its special case when $\gamma \to 0$.

When $\Omega_d(z)$ is concave with respect to z, the term can be upper bounded by its first-order Taylor expansion at $\sum_{ab} g_{dab}\widehat{X}_{ab}^{\phi_d}$: up to some additive constant,

$$\begin{aligned}\Omega_d\left(\sum_{ij} g_{dij}\widetilde{X}_{ij}^{\phi_d}\right) &\leq \left(\sum_{ij} g_{dij}\widetilde{X}_{ij}^{\phi_d}\right)\Omega_d'\left(\sum_{ab} g_{dab}\widehat{X}_{ab}^{\phi_d}\right)\\ &= \sum_{ij} f_{dij}\widetilde{X}_{ij}^{\phi_d},\end{aligned} \tag{17}$$

where $f_{dij} \triangleq \nu_d\tau_d g_{dij}\cdot\left(\sum_{ab} g_{dab}\widehat{X}_{ab}^{\phi_d}\right)^{\tau_d-1}$.

When $\Omega_d(z)$ is convex with respect to z, we can write $\theta_{dij} = \frac{g_{dij}\widehat{X}_{ij}^{\phi_d}}{\sum_{ab} g_{dab}\widehat{X}_{ab}^{\phi_d}}$ and apply Jensen's inequality to obtain the upper bound

$$\begin{aligned}\Omega_d\left(\sum_{ij} g_{dij}\widetilde{X}_{ij}^{\phi_d}\right) &\leq \sum_{ij}\theta_{dij}\Omega_d\left(\frac{g_{dij}\widetilde{X}_{ij}^{\phi_d}}{\theta_{dij}}\right)\\ &= \sum_{ij} f_{dij}\widetilde{X}_{ij}^{\phi_d\tau_d},\end{aligned} \tag{18}$$

where $f_{dij} \triangleq \nu_d g_{dij}\widehat{X}_{ij}^{\phi_d(1-\tau_d)}\left(\sum_{ab} g_{dab}\widehat{X}_{ab}^{\phi_d}\right)^{\tau_d-1}$.

We can see that both convex and concave cases reduce to the same form for the separable objectives and thus continue with the same procedure in Section III. For the non-separable divergences in Table (IV), we obtain the multiplicative update rules in the form Eq. (14) with $\eta = 1$ for Kullback-Leibler divergence, $\eta = 1/(1+\gamma)$ for γ-divergence when $\gamma > 0$, $\eta = 1/(1-\gamma)$ when $\gamma < 0$, $\eta = 1/\rho$ for Rényi divergence when $\rho > 1$, and $\eta = 1$ for $0 < \rho < 1$.

Here we only present exponentials on the matrix sum for notational brevity, while the same technique can easily be applied to the cases of row-wise or column-wise sums. Objectives that are not separable over d in Eq. (5) can be handled in a similar way.

V. QUADRATIC FACTORIZATIONS

The previous discussion focused on the linear factorization $\mathbf{X} \approx \mathbf{WH}$, where the factorizing matrix $\mathbf{W}$ (or $\mathbf{H}$) appears only once in the factorizing expression. There exist other NMF problems in which a factorizing matrix may appear twice, which we call *Quadratic Nonnegative Matrix Factorization* (QNMF).

The Quadratic NMF has a wide range of applications. For example, the *Projective Nonnegative Matrix Factorization* (PNMF) is able to achieve high sparseness in both feature extraction and sample clustering [30], [31], [32]. Other applications include the graph isomorphism problem if the sparsity (or orthogonality) of $\mathbf{W}$ is enforced [33], as well as parameter estimation of hidden Markov chains [34].

Here we focus on two typical factorization forms of QNMF: (1) the asymmetric form (AQNMF) $\mathbf{X} \approx \mathbf{WW}^T\mathbf{Y}$ and (2) the symmetric form (SQNMF) $\mathbf{X} \approx \mathbf{WYW}^T$. Note that other forms of quadratic factorization, e.g. $\mathbf{X} \approx \mathbf{YWW}^T$, can equivalently be transformed to AQNMF or SQNMF. Here $\mathbf{Y}$ can be (i) a constant, for instance $\mathbf{Y} = \mathbf{X}$ that leads to PNMF and $\mathbf{Y} = \mathbf{I}$ that leads to *Symmetric NMF* (see e.g. [10]), (ii) an abbreviation for the product of other factorizing matrices that only appear once in the approximation, or (iii) shorthand for recursively defined quadratic factorizations, for instance $\mathbf{Y} = \mathbf{XUU}^T$ that leads to $\mathbf{X} \approx \mathbf{WW}^T\mathbf{XUU}^T$.

The quadratic factorizations usually give rise to more difficult optimization problems. As an example, consider the PNMF based on Euclidean distance:

$$\underset{\mathbf{W}\geq 0}{\text{minimize}} \quad \|\mathbf{X} - \mathbf{WW}^T\mathbf{X}\|_F^2, \tag{19}$$

where the objective function is quartic with respect to $\mathbf{W}$ and therefore cannot be upper-bounded by the quadratic functions that were earlier proposed in [2], [10]. One must find an auxiliary function of at least the fourth power. To our knowledge, there has been no general principle of developing multiplicative algorithms with monotonicity guarantee for QNMF problems such as (19).

Our novel method now provides a straightforward way to include the second-order factorizations. Here we only show the derivation procedure for the symmetric and separable case, which can similarly be applied to the asymmetric and/or non-separable cases. Denote

$$\partial_{dik} = \left.\frac{\partial\sum_{ab}\omega_d\left(\widetilde{X}_{ab}\right)}{\partial\widetilde{W}_{ik}}\right|_{\widetilde{\mathbf{W}}=\mathbf{W}} = \sum_{jl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)W_{jl}, \tag{20}$$

where $S^{(d)}_{ijkl} = \tau_d f_{dij} \left(\mathbf{W}\mathbf{Y}\mathbf{W}^T\right)^{\tau_d - 1}_{ij} Y_{kl}$. We divide the monomials into the following four categories, each of which can be tightly upper-bounded (see Appendix C for derivation details).

(Case 1) For concave monomials and $S^{(d)}_{ijkl} > 0$

$$\sum_{ij} \omega_d \left(\widetilde{X}_{ij}\right) \leq \sum_{ik} \frac{W_{ik}\partial_{dik}}{2} \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^2 + \text{constant}. \quad (21)$$

(Case 2) For concave monomials and $S^{(d)}_{ijkl} < 0$,

$$\sum_{ij} \omega_d \left(\widetilde{X}_{ij}\right) \leq \lim_{\varepsilon \to 0^+} \sum_{ik} \frac{W_{ik}\partial_{dik}}{\varepsilon} \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\varepsilon} + \text{constant}. \quad (22)$$

(Case 3) For convex monomials and $f_{dij} > 0$,

$$\sum_{ij} \omega_d(\widetilde{X}_{ij}) \leq \sum_{ik} \frac{W_{ik}\partial_{dik}}{2\tau_d} \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{2\tau_d}. \quad (23)$$

(Case 4) For convex monomials and $f_{dij} < 0$

$$\sum_{ij} \omega_d \left(\widetilde{X}_{ij}\right) \leq \lim_{\varepsilon \to 0^+} \sum_{ik} \frac{W_{ik}\partial_{dik}}{\varepsilon} \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\varepsilon} + \text{constant}. \quad (24)$$

Particularly, when $S^{(d)}_{ijkl} + S^{(d)}_{jilk}$ is negative semi-definite, i.e. $\sum_{ijkl}\left(S^{(d)}_{ijkl} + S^{(d)}_{jilk}\right) U_{ik}U_{jl} \leq 0$ for any $\mathbf{U} \in \mathbb{R}^{n\times n}$, we have a tighter upper bound using the first-order Taylor expansion:

(Case 2a) For concave monomials and $S^{(d)}_{ijkl} + S^{(d)}_{jilk}$ is negative semi-definite,

$$\sum_{ij} \omega_d \left(\widetilde{X}_{ij}\right) \leq \sum_{ik} \partial_{dik}\widetilde{W}_{ik} + \text{constant}. \quad (25)$$

(Case 4a) For convex monomials and $S^{(d)}_{ijkl} + S^{(d)}_{jilk}$ is negative semi-definite,

$$\sum_{ij} \omega_d \left(\widetilde{X}_{ij}\right) \leq \sum_{ik} \frac{W_{ik}\partial_{dik}}{\tau_d} \left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\tau_d} + \text{constant}. \quad (26)$$

The resulting individual upper bounds of all the above cases are in the form of Eq. (10). The derivation can therefore proceed to the combining step in Section III-A and onwards in the same way as for the linear factorization.

As examples, we present the multiplicative update rules for QNMF based on α-divergence, β-divergence, γ-divergence and Rényi divergence. These families of divergences cover most commonly used dissimilarity measures, for example, the squared Euclidean distance ($\beta = 1$), Hellinger distance ($\alpha = 0.5$), χ^2-divergence ($\alpha = 2$), I-divergence ($\alpha \to 1$ or $\beta \to 0$), dual I-divergence ($\alpha \to 0$), Itakura-Saito divergence ($\beta \to -1$) and Kullback-Leibler divergence ($\gamma \to 0$ or $\rho \to 1$).

In general, the multiplicative update rules take the following forms:

- for AQNMF

$$W^{\text{new}}_{ik} = W_{ik}\left[\frac{\left(\mathbf{Q}\mathbf{Y}^T\mathbf{W} + \mathbf{Y}\mathbf{Q}^T\mathbf{W}\right)_{ik}}{\left(\mathbf{P}\mathbf{Y}^T\mathbf{W} + \mathbf{Y}\mathbf{P}^T\mathbf{W}\right)_{ik}} \cdot \theta\right]^{\eta}, \quad (27)$$

TABLE I
NOTATIONS IN THE MULTIPLICATIVE UPDATE RULES OF QNMF BASED ON α-, β-, γ-, AND ρ- (RÉNYI) DIVERGENCES, WHERE $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T\mathbf{Y}$ FOR AQNMF AND $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{Y}\mathbf{W}^T$ FOR SQNMF.

	P_{ij}	Q_{ij}	θ	η
α	1	$\frac{X^{\alpha}_{ij}}{\widehat{X}^{\alpha}_{ij}}$	1	$\frac{1}{2\alpha}$ $\alpha \in (1,+\infty)$ $\frac{1}{2}$ $\alpha \in (0,1)$ $\frac{1}{2\alpha-2}$ $\alpha \in (-\infty,0)$
β	$\widehat{X}^{\beta}_{ij}$	$\frac{X_{ij}}{\widehat{X}^{1-\beta}_{ij}}$	1	$\frac{1}{2+2\beta}$ $\beta \in (0,+\infty)$ $\frac{1}{2-2\beta}$ $\beta \in (-\infty,0)$
γ	$\widehat{X}^{\gamma}_{ij}$	$\frac{X_{ij}}{\widehat{X}^{1-\gamma}_{ij}}$	$\frac{\sum_{ab}\widehat{X}^{\gamma+1}_{ab}}{\sum_{ab}X_{ab}\widehat{X}^{\gamma}_{ab}}$	$\frac{1}{2+2\gamma}$ $\gamma \in (0,+\infty)$ $\frac{1}{2-2\gamma}$ $\gamma \in (-\infty,0)$
ρ	1	$\frac{X^{\rho}_{ij}}{\widehat{X}^{\rho}_{ij}}$	$\frac{\sum_{ab}\widehat{X}_{ab}}{\sum_{ab}X^{\rho}_{ab}\widehat{X}^{1-\rho}_{ab}}$	$\frac{1}{2\rho}$ $\rho \in (1,\infty)$ $\frac{1}{2}$ $\rho \in (0,1)$

- for SQNMF

$$W^{\text{new}}_{ik} = W_{ik}\left[\frac{\left(\mathbf{Q}\mathbf{W}\mathbf{Y}^T + \mathbf{Q}^T\mathbf{W}\mathbf{Y}\right)_{ik}}{\left(\mathbf{P}\mathbf{W}\mathbf{Y}^T + \mathbf{P}^T\mathbf{W}\mathbf{Y}\right)_{ik}} \cdot \theta\right]^{\eta}, \quad (28)$$

where $\mathbf{P}$, $\mathbf{Q}$, θ, and η are specified in Table I. For example, the rule for SQNMF based on the squared Euclidean distance ($\beta \to 1$) reads

$$W^{\text{new}}_{ik} = W_{ik}\left[\frac{\left(\mathbf{X}\mathbf{W}\mathbf{Y}^T + \mathbf{X}^T\mathbf{W}\mathbf{Y}\right)_{ik}}{\left(\mathbf{W}\mathbf{Y}\mathbf{W}^T\mathbf{W}\mathbf{Y}^T + \mathbf{W}\mathbf{Y}^T\mathbf{W}^T\mathbf{W}\mathbf{Y}\right)_{ik}}\right]^{1/4}. \quad (29)$$

As an exception, the update rules for the dual I-divergence take a different form

$$W^{\text{new}}_{ik} = W_{ik}\exp\left[\frac{1}{2}\frac{\left(\mathbf{Q}\mathbf{Y}^T\mathbf{W} + \mathbf{Y}\mathbf{Q}^T\mathbf{W}\right)_{ik}}{\left(\mathbf{P}\mathbf{Y}^T\mathbf{W} + \mathbf{Y}\mathbf{P}^T\mathbf{W}\right)_{ik}}\right], \quad (30)$$

$$W^{\text{new}}_{ik} = W_{ik}\exp\left[\frac{1}{2}\frac{\left(\mathbf{Q}\mathbf{W}\mathbf{Y}^T + \mathbf{Q}^T\mathbf{W}\mathbf{Y}\right)_{ik}}{\left(\mathbf{P}\mathbf{W}\mathbf{Y}^T + \mathbf{P}^T\mathbf{W}\mathbf{Y}\right)_{ik}}\right] \quad (31)$$

for AQNMF and SQNMF respectively, with $P_{ij} = 1$ and $Q_{ij} = \ln\left(X_{ij}/\widehat{X}_{ij}\right)$.

VI. EXPERIMENTS

A. KKT optimality of the multiplicative algorithms

In the above we have rigorously proven that the objective is monotonically decreasing in NMF. Because the approximation error is lower-bounded by zero, the monotonicity directly implies that the objective evolution is convergent.

Another type of convergence, concerning the KKT optimality of the converged points, is of interest in the optimization field. Particularly, the KKT optimality in linear NMF refers to the satisfaction of the following conditions: for all i, k

1) (*feasibility*): $W_{ik} \geq 0$ and $H_{ik} \geq 0$,
2) (*stationarity*): $\nabla_{\mathbf{W},ik} \geq 0$, and $\nabla_{\mathbf{H},ik} \geq 0$
3) (*complementary slackness*) $W_{ik}\nabla_{\mathbf{W},ik} = 0$, and $H_{ik}\nabla_{\mathbf{H},ik} = 0$.

TABLE II
VIOLATION EXTENTS OF THE SLACKNESS CONDITION IN LINEAR NMF, (A) FOR $\mathbf{W}$ AND (B) FOR $\mathbf{H}$, AND IN PNMF (C).

(A)

	iris	ecoli5	swimmer
Euclidean	9e-11±1e-26	2e-10±0	6e-07±0
I	2e-10±0	9e-16±1e-31	1e-11±0
dual-I	1e+01±0	3e+00±5e-16	1e+02±0
Itakura-Saito	7e-14±1e-29	2e-15±0	2e-09±5e-09
Log-Quad	3e-07±6e-23	8e-12±2e-27	0±0
KL	7e-15±8e-31	2e-15±4e-31	0±0

(B)

	iris	ecoli5	swimmer
Euclidean	3e-09±0	9e-09±2e-24	4e-07±6e-23
I	1e-09±0	2e-14±3e-30	1e-12±0
dual-I	4e+02±6e-14	1e+02±3e-14	4e+01±7e-15
Itakura-Saito	1e-12±2e-28	4e-14±0	9e-10±2e-09
Log-Quad	5e-06±0	3e-10±0	0±0
KL	0±0	4e-13±5e-29	0±0

(C)

	iris	ecoli5	swimmer
Euclidean	1e-03±0	2e-06±0	1e-13±1e-29
I	2e-05±4e-21	1e-06±0	2e-13±0
dual-I	4e+01±7e-15	1e+01±2e-15	9e+02±1e-13
Itakura-Saito	3e-07±6e-23	4e-04±6e-20	0±0
Log-Quad	3e-04±6e-20	7e-06±2e-21	0±0
KL	2e-05±0	1e-06±2e-22	0±0

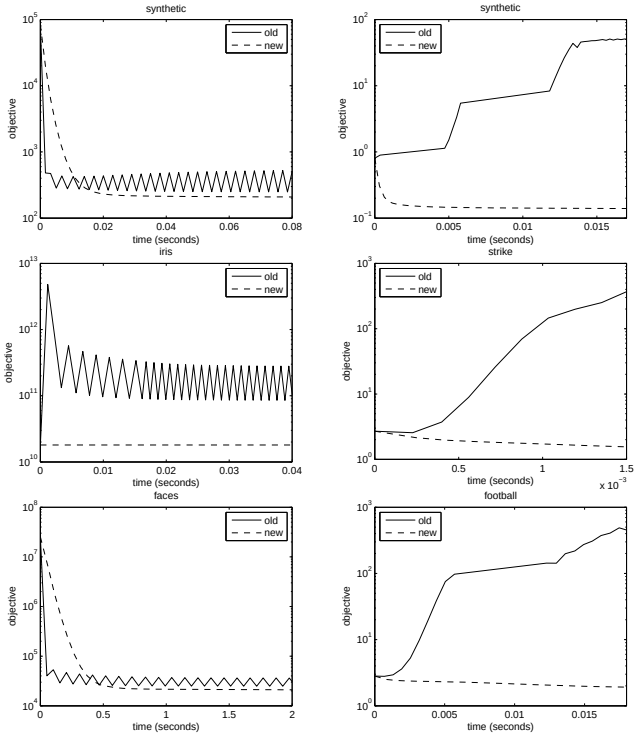

Fig. 1. Objective evolutions of (left) NMF and (right) SNMF using multiplicative update rules by the old and new principles.

Similar optimality conditions apply to the quadratic NMF, e.g. the matrix $\mathbf{W}$ and its associated gradient in PNMF.

The first condition obviously holds for multiplicative updates as long as the factorizing matrices are nonnegatively initialized. In our practice, the stationarity is easily checked and also valid for all multiplicative algorithms after sufficiently long runs. The major question that remains is the complementary slackness.

Usually multiplicative algorithms operate on positive factorizing matrices. In this setting, the complementary slackness can only hold in an asymptotic manner. The rigorous proof of such asymptotic convergence is difficult and only available for some particular divergences (see e.g. [35]).

In this section we provide the numerical results for checking the complementary slackness. The extent of violation of the condition is measured by $\sum_{ik} |W_{ik}\nabla_{\mathbf{W},ik}|$ and $\sum_{ik} |H_{ik}\nabla_{\mathbf{H},ik}|$ for $\mathbf{W}$ and $\mathbf{H}$, respectively.

We have used three publicly available datasets: two of them, *iris* and *ecoli* are collected from the UCI repository[1], and the *swimmer* dataset [36] consists of 256 binary images depicting moving parts of swimmers. The dimensions of the datasets are 150×4, 327×7, and 1024×256, respectively. To avoid numerical errors, we add a small positive number (e.g. 10^{-16}) to each denominator and to each logarithm in the update rules. Each multiplicative algorithm has been run at least 10^6 iterations and repeated 100 times. The resulting means and standard deviations of the violation extent are shown in Table II (A)-(B). We also did the same experiments for PNMF and the results are shown in Table II (C).

From these results, we can see that most multiplicative algorithms achieve zero or nearly zero violation of complementary slackness, compared to the averages of non-zero entries of the input matrix, which are 3.46, 0.48, and 0.06, respectively. One exception is the dual-I divergence, where the converged points do not fulfill the complementary slackness condition for all datasets. The violation can be explained by the fixed-points of its multiplicative algorithm, which are given by $W_{ik} \exp\left(\frac{\nabla^-_{\mathbf{W},ik}}{\nabla^+_{\mathbf{W},ik}}\right) = W_{ik}$ for the matrix $\mathbf{W}$ in linear NMF, i.e. $W_{ik} \to 0$ or $\frac{\nabla^-_{\mathbf{W},ik}}{\nabla^+_{\mathbf{W},ik}} = 0$ with positive initialization. The latter condition has no connection to the complementary slackness. By contrast, other multiplicative algorithms using Eq. (14), have fixed-points $W_{ik} \to 0$ or $\frac{\nabla^-_{\mathbf{W},ik}}{\nabla^+_{\mathbf{W},ik}} = 1$, which is consistent to the complementary slackness if $\nabla^+_{\mathbf{W},ik} > 0$.

B. Comparison to the conventional principle

Conventionally, multiplying factors in the multiplicative update rules are formed by putting the positive part of the gradient to the denominator and the unsigned negative part to the numerator. However, such multiplicative updates cannot guarantee monotonic decrease of the approximation error after each iteration. In this section we show descent violation examples for both linear and quadratic NMF problems, whereas the multiplicative algorithms obtained by our new principle can completely avoid such violations.

We perform the comparison on NMF based on $\alpha\beta$-Bregman divergence ($\alpha = 10$ and $\beta = 0.5$) and Symmetric NMF (SNMF) based on Rényi divergence ($\rho = 2$). We selected these two problems because they cover both linear and quadratic factorization cases. In addition, these problems have not been solved by any previously existing methods, which demon-

[1]http://archive.ics.uci.edu/ml/

TABLE III
PERCENTAGE OF DESCENT VIOLATION IN (A) NMF AND (B) SNMF.

(A)

method	synthetic	iris	faces
old	49.98±0.12	1.75±8.68	50.00±0.02
old (alternation)	0.70±4.72	0.06±0.05	0.13±0.08
new	0.00±0.00	0.00±0.00	0.00±0.00

(B)

method	synthetic	strike	football
old	68.14±6.70	10.70±1.27	20.64±4.97
new	0.00±0.00	0.00±0.00	0.00±0.00

strates that our principle can easily be extended to new NMF problems.

For each of the above problems, we have tested the compared multiplicative algorithms on three datasets. For linear NMF, we have used (*syn*) a synthetic nonnegative matrix generated by the Matlab command *rand(50,30)*, (*iris*) a dataset containing 4 nonnegative measurements of 150 iris plants, and (*faces*) the inner part of the ORL facial images [37] (400×644). For SNMF, we have used (*syn*) a synthetic nonnegative matrix generated by symmetrizing *rand(50,50)*, (*strike*) and (*football*)[2], two real-world undirected graph adjacency matrices of sizes 24 and 115.

The objective function evolutions are shown in Figure 1. In order to illustrate the descent violation more clearly, we only present the curves for linear NMF iterations on $\mathbf{W}$ with fixed $\mathbf{H}$. It can be seen that for both NMF problems, the multiplicative algorithms by the conventional principle may cause undesired ascent after each update. The resulting objectives are far away from the local minimum in four out of six selected datasets. In contrast, the multiplicative update rules obtained by our new principle produce decreasing curves in all demonstrated experiments.

We ran each experiment 100 times with different starting points and recorded the numbers of objective increase. We then obtained a violation percentage by dividing the numbers with the length of the run. The resulting means and standard deviations across 100 runs are shown in Table III, from which we can see that it is quite common for the conventional method to yield a violation. Alternating the updates for $\mathbf{W}$ and $\mathbf{H}$ can reduce the violation times, though objective increasing still happens for the multiplicative algorithms obtained by the conventional principle. One may even alternate among updates of individual entries of $\mathbf{W}$ or $\mathbf{H}$ to achieve descent in practice, though it is slow and still lacks theoretical guarantee. By contrast, the multiplicative algorithms by our new principle, as expected, produce zero violation in all experiments.

VII. DISCUSSION

We have presented a generic principle for deriving multiplicative update rules, as well a proof of the convergence of their objective function, that applies for a large variety of linear and quadratic nonnegative matrix factorization problems. The proposed principle only requires that the NMF approximation objective function can be written as a sum of a finite number of monomials, a mild assumption that holds for many commonly used approximation error measures, as shown by the many examples given here. As a result, our method turns the derivation that seemingly requires intense mathematical work into a routine exercise that could be even readily automated using symbolic mathematics software.

There exist divergences that are not covered by our method. Some approximation objectives cannot be expressed as a sum of a finite number of monomials, for example the Relative Jensen-Shannon divergence and the Arimoto distance (see e.g. Chapter 2 of [14]). For such objectives, the resulting exponent of the multiplying factor will become infinitesimal when applying the proposed procedure. This problem is however much alleviated in practice by using approximations with finite expansions. Usually the resulting multiplicative updates can still effectively decrease the original objective. Another type of excluded divergences are some constrained ones such as the Bit entropic loss [14, page 99]. The fixed-point optimization of such comprehensive objectives is still an open problem.

We have presented a unified principle with mild generic assumptions. In practice, in any specific case, the update rules can be further improved by ad hoc methods with more properties of the specific NMF objective considered. For example, when finding individual upper-bounds for the PNMF based on the squared Euclidean distance, one can make use of the positive semi-definiteness of $\mathbf{X}\mathbf{X}^T$ and thus apply Case 2a instead of Case 2 in Step 2 (see Appendix D-D). This consequently leads to a multiplicative update rule with exponent $1/3$ instead of $1/4$. Another speedup example for NMF based on the squared Euclidean distance can be found in [38].

Here we mainly focus on the convergence or monotonic decrease of the objective function. Rigorous proof on convergence to stationary points has recently been investigated by other researchers. Lin presented a proof that a slightly modified multiplicative algorithm for NMF based on the squared Euclidean distance can achieve KKT optimality conditions [35]. More recently, Badeau et al. proposed to examine the asymptotic convergence behavior of multiplicative updates by using the Lyapunov stability theory [39]. Their result shows that if η in our generic algorithm is small enough, multiplicative updates are asymptotically stable, namely, can arrive at a local minimum after sufficiently many iterations. However, the upper-bound of η with such stability is difficult to compute in general. Furthermore, little is known about the stability when alternation among factorizing matrices is applied.

NMF optimization problems are usually non-convex. Global optimum is not guaranteed unless extra constraints are imposed. Considerable initialization or pre-training can effectively avoid poor local minima. Typical initialization methods include PCA/SVD, clustering, and Gabor wavelets (see e.g. [40], [41], [42, Section 9.2.1]).

Another important and still open problem is how to select among various divergences. This strongly depends on the nature of the data to be analyzed, for instance, noise, outliers, and data types. For particular parameterized families of divergence, usually there is some trade-off when adjusting the divergence

[2]http://www.cise.ufl.edu/research/sparse/matrices/index.html

parameter. For example, a large β or γ leads to more robust but less efficient estimation. More details on these trade-offs can be found in [43], [44] and references therein. Automatic selection of the divergence parameters generally requires extra information or criterion, for example, ground truth data [45] or cross-validations with a fixed reference parameter [46].

ACKNOWLEDGMENT

We would like to thank support for the project *Finnish Centre of Excellence in Adaptive Informatics Research* from the Academy of Finland.

APPENDIX A
UPPER-BOUNDS USED IN THE DERIVATIONS

1) For a real convex function f and positive constants $a_1, a_2, \ldots, a_m$,

$$f\left(\mathbf{a}^T\mathbf{x}\right) \leq \sum_i \frac{a_i y_i}{\mathbf{a}^T\mathbf{y}} f\left(\frac{x_i}{y_i}\mathbf{a}^T\mathbf{y}\right).$$

This upper-bound is due to the Jensen's inequality.

2) A real concave function f is upper-bounded by its linear or first-order Taylor expansion:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + (\mathbf{x}-\mathbf{y})^T\nabla = \mathbf{x}^T\nabla + \text{constant},$$

where $\nabla = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{y}}$.

3) For $x > 0$, $y > 0$, $a < b$, $z = x/y$, we have $\frac{z^a-1}{a} \leq \frac{z^b-1}{b}$, because $h(t) = \frac{z^t-1}{t}$ is monotonically increasing for $z > 0$.

4) For $\mathbf{A} \in \mathbb{R}_+^{m\times m}$, $\mathbf{x} \in \mathbb{R}_+^m$, $\mathbf{y} \in \mathbb{R}_+^m$

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \leq \sum_i \frac{x_i^2}{2y_i}\left(\mathbf{A}\mathbf{y} + \mathbf{A}^T\mathbf{y}\right)_i$$

This upper-bound has been proven in [2], [10] by writing $x_i = y_i u_i$.

5) For $\mathbf{A} \in \mathbb{R}_+^{m\times m}$, $\mathbf{x} \in \mathbb{R}_+^m$, $\mathbf{y} \in \mathbb{R}_+^m$

$$-\mathbf{x}^T\mathbf{A}\mathbf{x} \leq -\frac{1}{2}\sum_{ij}(A_{ij}+A_{ji})y_i y_j\left(1+\log\frac{x_i x_j}{y_i y_j}\right)$$

This upper-bound is due to the inequality $1+\log z \leq z$ for $z > 0$ and was first used in [47].

Remarks:

- For notational brevity we only write the upper-bounds in the vectorial form. The same bounds also hold for the matrix case.
- All the above upper-bounds are tight where $x = y$ or $\mathbf{x} = \mathbf{y}$.
- The third upper-bound is new and unseen in the previous literature for NMF proofs. Notice that it includes the inequality $1+\log z < z$ for $z > 0$ as its a special case by writing the logarithm into the limit form Eq. (4).

APPENDIX B
PROOF OF LEMMA 2

Proof: The inequality obviously holds when $a = 0$ or $c = 0$. Otherwise, by $\frac{x^a-1}{a} \leq \frac{x^b-1}{b}$ we have the following results: 1) When $a > 1$, we have $c \geq 0$ because cx^a is convex. With $a < b$, $x^a \leq a\frac{x^b}{b} + 1 - \frac{a}{b} \Rightarrow cx^a \leq ac\frac{x^b}{b} + c\left(1-\frac{a}{b}\right)$. 2) When $0 < a < 1$, we have $c \leq 0$ because cx^a is convex. With $a > b$, $x^a \geq a\frac{x^b}{b} + 1 - \frac{a}{b} \Rightarrow cx^a \leq ac\frac{x^b}{b} + c\left(1-\frac{a}{b}\right)$. When $a < 0$, we have $c \geq 0$ because cx^a is convex. Then with $a > b$, $x^a \leq a\frac{x^b}{b} + 1 - \frac{a}{b} \Rightarrow cx^a \leq ac\frac{x^b}{b} + c\left(1-\frac{a}{b}\right)$. 3) and 4) are complementary cases for linear monomials which are similarly proven according to the sign of c. ■

APPENDIX C
UPPER BOUNDING QUADRATIC MONOMIALS

For concave monomials,

$$\begin{aligned}\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) &\leq \sum_{ij}\widetilde{X}_{ij}\frac{\partial\sum_{ab}f_{dab}\widetilde{X}_{ab}^{\tau_d}}{\partial\widetilde{X}_{ij}}\bigg|_{\widetilde{\mathbf{X}}=\widehat{\mathbf{X}}} + \text{constant}\\ &= \sum_{ijkl}\widetilde{W}_{ik}\widetilde{W}_{jl}S_{ijkl}^{(d)} + \text{constant}\\ &= \frac{1}{2}\sum_{ijkl}\widetilde{W}_{ik}\widetilde{W}_{jl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right) + \text{constant}.\end{aligned}$$

The last step is due to the exchanges of summation indices: $i \leftrightarrow j$ and $k \leftrightarrow l$. We do not assume symmetry on $\mathbf{S}^{(d)}$.
(Case 1) up to some additive constant

$$\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) \leq \sum_{ik}\frac{\widetilde{W}_{ik}^2}{2W_{ik}}\sum_{jl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)W_{jl}.$$

(Case 2) up to some additive constant

$$\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) \leq \frac{1}{2}\sum_{ijkl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)W_{ik}W_{jl}\ln\frac{\widetilde{W}_{ik}\widetilde{W}_{jl}}{W_{ik}W_{jl}}.$$

(Case 2a)

$$\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) \leq \sum_{ik}\widetilde{W}_{ik}\sum_{jl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)W_{jl} + \text{constant}.$$

For convex monomials, with $\zeta_{ijkl} = \frac{W_{ik}W_{jl}Y_{kl}}{(\mathbf{W}\mathbf{Y}\mathbf{W}^T)_{ij}}$

$$\begin{aligned}\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) &\leq \sum_{ijkl}\zeta_{ijkl}\cdot\omega_d\left(\frac{\widetilde{W}_{ik}\widetilde{W}_{jl}Y_{kl}}{\zeta_{ijkl}}\right)\\ &= \frac{1}{2\tau_d}\sum_{ijkl}\widetilde{V}_{dik}\widetilde{V}_{djl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right),\end{aligned}$$

where $\widetilde{V}_{dik} = \widetilde{W}_{ik}^{\tau_d}W_{ik}^{1-\tau_d}$ and $V_{dik} = W_{ik}$.
(Case 3)

$$\sum_{ij}\omega_d(\widetilde{X}_{ij}) \leq \sum_{ik}\frac{\widetilde{V}_{dik}^2}{2\tau_d V_{dik}}\sum_{jl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)V_{djl}.$$

(Case 4) Up to some additive constant,

$$\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right) \leq \frac{1}{2\tau_d}\sum_{ijkl}\left(S_{ijkl}^{(d)} + S_{jilk}^{(d)}\right)V_{dik}V_{djl}\ln\frac{\widetilde{V}_{dik}\widetilde{V}_{djl}}{V_{dik}V_{djl}}.$$

(Case 4a) Up to some additive constant,

$$\sum_{ij}\omega_d\left(\widetilde{X}_{ij}\right)\leq\sum_{ik}\widetilde{V}_{dik}\tau_d^{-1}\sum_{jl}\left(S_{ijkl}^{(d)}+S_{jilk}^{(d)}\right)V_{djl}.$$

APPENDIX D
EXAMPLE DERIVATIONS OF MULTIPLICATIVE ALGORITHMS

Here we present the concrete derivations of the rules and auxiliary functions for five selected approximation objectives, three for NMF and two for PNMF. Note that the main purpose of the derivations here is not to generate new algorithms. Instead, they respectively demonstrate the following aspects of the principle proposed in Sections III-V: (1) the use of concavity or convexity of individual monomials, (2) how to handle objectives that comprise logarithm, (3) merging individual upper-bounds, (4) quadratic factorization, and (5) non-separable objectives.

A. NMF based on β-divergence

A similar procedure can be found in [26], [20]. Denote $D_{ij}^{(1)}=\frac{1}{1+\beta}\widehat{X}_{ij}^{\beta+1}$ and $D_{ij}^{(2)}=-\frac{1}{\beta}X_{ij}\widehat{X}_{ij}^{\beta}$. There are three cases: (1) When $\beta>1$, $D_{ij}^{(1)}$ is convex with respect to $\widehat{X}_{ij}$ while $D_{ij}^{(2)}$ is concave. Therefore the auxiliary function is

$$\sum_{ik}\left[\frac{\widetilde{W}_{ik}^{\beta+1}}{W_{ik}^{\beta}}\frac{1}{1+\beta}\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\widetilde{W}_{ik}\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}\right]+\text{constant},$$

whose derivative with respect to $\widetilde{W}_{ik}$ is

$$\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\beta}\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}.$$

Setting the above derivative to zero leads to

$$W_{ik}^{\text{new}}=W_{ik}\left(\frac{\sum_j X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}}{\sum_j\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}}\right)^{\frac{1}{\beta}}.$$

(2) When $0\leq\beta\leq1$, both $D_{ij}^{(1)}$ and $D_{ij}^{(2)}$ are convex with respect to $\widehat{X}_{ij}$. Therefore the auxiliary function is

$$\sum_{ik}\left[\frac{\widetilde{W}_{ik}^{\beta+1}}{W_{ik}^{\beta}}\frac{1}{1+\beta}\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\frac{\widetilde{W}_{ik}^{\beta}}{W_{ik}^{\beta-1}}\frac{1}{\beta}\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}\right]+\text{constant},$$

whose derivative with respect to $\widetilde{W}_{ik}$ is

$$\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\beta}\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\beta-1}\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}.$$

Setting the above derivative to zero leads to

$$W_{ik}^{\text{new}}=W_{ik}\frac{\sum_j X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}}{\sum_j\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}},$$

which is identical to the update rule derived in [4].

(3) When $\beta<0,\beta\neq-1$, $D_{ij}^{(1)}$ is concave with respect to $\widehat{X}_{ij}$ while $D_{ij}^{(2)}$ is convex. Therefore the auxiliary function is

$$\sum_{ik}\left[\widetilde{W}_{ik}\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\frac{\widetilde{W}_{ik}^{\beta}}{W_{ik}^{\beta-1}}\frac{1}{\beta}\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}\right],$$

whose derivative with respect to $\widetilde{W}_{ik}$ is

$$\sum_{j}\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}-\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\beta-1}\sum_{j}X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}.$$

Setting the above derivative to zero leads to

$$W_{ik}^{\text{new}}=W_{ik}\left(\frac{\sum_j X_{ij}\left(\mathbf{WH}\right)_{ij}^{\beta-1}H_{kj}}{\sum_j\left(\mathbf{WH}\right)_{ij}^{\beta}H_{kj}}\right)^{\frac{1}{1-\beta}}.$$

B. NMF based on dual I-divergence

The monomials are convex and linear respectively. Therefore the auxiliary function according to Eqs. (7) and (9) is

$$\lim_{\epsilon\to0^+}\frac{1}{\epsilon}\sum_{ik}\left[\widetilde{W}_{ik}^{1+\epsilon}W_{ik}^{-\epsilon}\sum_{j}X_{ij}^{-\epsilon}\left(\mathbf{WH}\right)_{ij}^{\epsilon}H_{kj}-(1+\epsilon)\widetilde{W}_{ik}\sum_{j}H_{kj}\right]+\text{constant},$$

whose derivative with respect to $\widetilde{W}_{ik}$

$$\lim_{\epsilon\to0^+}\frac{1}{\epsilon}\left[(1+\epsilon)\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\epsilon}\sum_{j}Z_{ij}^{-\epsilon}H_{kj}-(1+\epsilon)\sum_{j}H_{kj}\right].$$

with $Z_{ij}=X_{ij}/\left(\mathbf{WH}\right)_{ij}$. Because the derivative has the form $\frac{0}{0}$, we apply L'Hôpital's rule to obtain the limit (calculate derivatives of both numerator and denominator w.r.t ϵ):

$$\begin{aligned}&\lim_{\epsilon\to0^+}\left[\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\epsilon}\sum_{j}Z_{ij}^{-\epsilon}H_{kj}\right.\\&+(1+\epsilon)\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\epsilon}\ln\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)\sum_{j}Z_{ij}^{-\epsilon}H_{kj}\\&\left.-(1+\epsilon)\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)^{\epsilon}\sum_{j}Z_{ij}^{-\epsilon}\ln Z_{ij}H_{kj}-\sum_{j}H_{kj}\right]\\=&\ln\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)\sum_{j}H_{kj}-\sum_{j}\ln Z_{ij}H_{kj}.\end{aligned}$$

Setting the above derivative to zero, we obtain the update rule:

$$W_{ik}^{\text{new}}=W_{ik}\exp\left(\frac{\sum_j\ln Z_{ij}H_{kj}}{\sum_j H_{kj}}\right),$$

which is identical to the one proven in [5].

C. NMF based on Log-Quad cost

The monomials are convex, linear, linear, and convex, respectively. Therefore the auxiliary function before merging terms is

$$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ik}\left[-W_{ik}^{1-\epsilon}\widetilde{W}_{ik}^{\epsilon}\sum_j Z_{ij}\left(\mathbf{WH}\right)_{ij}^{\epsilon}H_{kj}\right] + \sum_{ik}\widetilde{W}_{ik}\sum_j H_{kj} - 2\sum_{ik}\widetilde{W}_{ik}\left(\mathbf{XH}^T\right)_{ik} + \sum_{ik}\frac{\widetilde{W}_{ik}^2}{W_{ik}}\left(\mathbf{WHH}^T\right)_{ik} + \text{constant.}$$

and after merging it becomes

$$\sum_{ik}\frac{\widetilde{W}_{ik}^2}{2W_{ik}}\left[\sum_j H_{kj} + 2\left(\mathbf{WHH}^T\right)_{ik}\right] - \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ik}\frac{\widetilde{W}_{ik}^{\epsilon}}{W_{ik}^{\epsilon-1}}\left[\sum_j Z_{ij}\left(\mathbf{WH}\right)_{ij}^{\epsilon}H_{kj} + 2\epsilon\left(\mathbf{XH}^T\right)_{ik}\right]$$

up to some additive constant. The derivative of the above auxiliary function with respect to $\widetilde{W}_{ik}$ is

$$\frac{\widetilde{W}_{ik}}{W_{ik}}\left[\sum_j H_{kj} + 2\left(\mathbf{WHH}^T\right)_{ik}\right] - \frac{W_{ik}}{\widetilde{W}_{ik}}\left[\sum_j Z_{ij}H_{kj} + 2\left(\mathbf{XH}^T\right)_{ik}\right].$$

Setting the above derivative to zero leads to the multiplicative update rule

$$W_{ik}^{\text{new}} = W_{ik}\sqrt{\frac{\left(\mathbf{ZH}^T + 2\mathbf{XH}^T\right)_{ik}}{\sum_j H_{kj} + 2\left(\mathbf{WHH}^T\right)_{ik}}}.$$

D. PNMF based on Euclidean distance

The first monomial belongs to Case 2a and the second to Case 3. Therefore the auxiliary function is

$$-\sum_{ik}\widetilde{W}_{ik}\left(2\mathbf{XX}^T\mathbf{W}\right)_{ik} + \sum_{ik}\frac{\widetilde{W}_{ik}^4}{4W_{ik}^3}\left(\mathbf{WW}^T\mathbf{XX}^T\mathbf{W} + \mathbf{XX}^T\mathbf{WW}^T\mathbf{W}\right)_{ik}$$

up to some additive constant. Setting its derivative to zero leads to the multiplicative update rule

$$W_{ik}^{\text{new}} = W_{ik}\left[\frac{2\left(\mathbf{XX}^T\mathbf{W}\right)_{ik}}{\left(\mathbf{WW}^T\mathbf{XX}^T\mathbf{W} + \mathbf{XX}^T\mathbf{WW}^T\mathbf{W}\right)_{ik}}\right]^{1/3},$$

which resembles the one given in [30], [32] except the cubic root. Notice that the old rule itself does not necessarily minimize the objective and must be accompanied with an extra normalization [30] or stabilization step [32] while the new update rule can guarantee monotonic decrease.

E. PNMF based on Kullback-Leibler divergence

The first term is separable, while the second term is non-separable and concave with respect to $\sum_{ij}\widetilde{X}_{ij}$. Therefore the upper-bounding separable function is

$$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ij}\left[-X_{ij}\widehat{X}_{ij}^{\epsilon} + \epsilon\left(\sum_{ab}\widehat{X}_{ab}\right)^{\epsilon-1}\widehat{X}_{ij}\right] + \text{constant.}$$

Regardless the constant, the two term of the separable upper-bounding function belong to Cases 4 and 1. Therefore the auxiliary function is

$$-\frac{1}{2}\sum_{aik}B_{ai}W_{ak}W_{ik}\ln\widetilde{W}_{ak}\widetilde{W}_{ik} + \left(\sum_{ab}\left(\mathbf{WW}^T\mathbf{X}\right)_{ab}\right)^{-1}\sum_{ik}\frac{\widetilde{W}_{ik}^2}{2W_{ik}}\sum_a C_{ia}W_{ak} + \text{constant},$$

where $\mathbf{B} = \mathbf{XX}^T$ and $\mathbf{C} = \mathbf{BWW}^T + \mathbf{WW}^T\mathbf{B}$. Setting its derivative to zero leads to the update rule

$$W_{ik}^{\text{new}} = W_{ik}\sqrt{\frac{\left(\mathbf{BW}\right)_{ik}}{\left(\mathbf{CW}\right)_{ik}}\sum_{ab}\left(\mathbf{WH}\right)_{ab}}.$$

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, 1999.

[2] ——, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[3] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[4] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2006.

[5] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Advances in Neural Information Processing Systems*, vol. 18, 2006, pp. 283–290.

[6] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recognition Letters*, vol. 29, pp. 1433–1440, 2008.

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

[8] 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, 2009.

[9] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[10] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 126–135.

[11] S. Behnke, "Discovering hierarchical speech features using convolutional non-negative matrix factorization," in *Proc. of the International Joint Conference on Neural Networks*, vol. 4, 2003, pp. 2758–2763.

[12] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[13] A. Cichocki and R. Zdunek, "Multilayer nonnegative matrix factorization using projected gradient approaches," *International Journal of Neural Systems*, vol. 17, no. 6, pp. 431–446, 2007.

[14] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis*. John Wiley, 2009.

[15] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," *Neural Computation*, vol. 19, pp. 2756–2779, 2007.

[16] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with quadratic programming," *Neurocomputing*, vol. 71, no. 10-12, pp. 2309–2008, 2008.

TABLE IV
EXAMPLE APPROXIMATION OBJECTIVES AND THEIR GENERALIZED POLYNOMIAL FORM

name	definition	generalized polynomial form (+constant)
Euclidean distance	$D_{\mathrm{EU}}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}\left(X_{ij}-\widehat{X}_{ij}\right)^2$	$\sum_{ij}\left(-2X_{ij}\widehat{X}_{ij}+\widehat{X}_{ij}^2\right)$
I-divergence	$D_{\mathrm{I}}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}\left(X_{ij}\ln\frac{X_{ij}}{\widehat{X}_{ij}}-X_{ij}+\widehat{X}_{ij}\right)$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ij}\left(-X_{ij}\widehat{X}_{ij}^{\epsilon}+\epsilon\widehat{X}_{ij}\right)$
Dual I-divergence	$D_{\mathrm{I}}(\widehat{\mathbf{X}}\|\|\mathbf{X}) = \sum_{ij}\left(\widehat{X}_{ij}\ln\frac{\widehat{X}_{ij}}{X_{ij}}-\widehat{X}_{ij}+X_{ij}\right)$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ij}\left(X_{ij}^{-\epsilon}\widehat{X}_{ij}^{1+\epsilon}-(1+\epsilon)\,\widehat{X}_{ij}\right)$
Itakura-Saito divergence	$D_{\mathrm{IS}}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}\left(-\ln\left(\frac{X_{ij}}{\widehat{X}_{ij}}\right)+\frac{X_{ij}}{\widehat{X}_{ij}}-1\right)$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ij}\left(-X_{ij}^{\epsilon}\widehat{X}_{ij}^{-\epsilon}+\epsilon X_{ij}\widehat{X}_{ij}^{-1}\right)$
α-divergence	$D_{\alpha}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \frac{1}{\alpha(1-\alpha)}\sum_{ij}\left(\alpha X_{ij}+(1-\alpha)\widehat{X}_{ij}-X_{ij}^{\alpha}\hat{X}_{ij}^{1-\alpha}\right)$	$\sum_{ij}\left(-\frac{1}{\alpha(1-\alpha)}X_{ij}^{\alpha}\hat{X}_{ij}^{1-\alpha}+\frac{1}{\alpha}\widehat{X}_{ij}\right)$
β-divergence	$D_{\beta}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}\left(X_{ij}\frac{X_{ij}^{\beta}-\widehat{X}_{ij}^{\beta}}{\beta}-\frac{X_{ij}^{\beta+1}-\widehat{X}_{ij}^{\beta+1}}{\beta+1}\right)$	$\sum_{ij}\left(\frac{1}{1+\beta}\widehat{X}_{ij}^{\beta+1}-\frac{1}{\beta}X_{ij}\widehat{X}_{ij}^{\beta}\right)$
Log-Quad cost	$D_{\mathrm{LQ}}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = D_{\mathrm{EU}}(\mathbf{X}\|\|\widehat{\mathbf{X}})+D_{\mathrm{I}}(\mathbf{X}\|\|\widehat{\mathbf{X}})$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\sum_{ij}\left(-X_{ij}\widehat{X}_{ij}^{\epsilon}+\epsilon\widehat{X}_{ij}-2\epsilon X_{ij}\widehat{X}_{ij}+\epsilon\widehat{X}_{ij}^2\right)$
$\alpha\beta$-Bregman divergence	$D_{\phi\alpha\beta}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}\phi(X_{ij})-\phi(\widehat{X}_{ij})-\phi'(\widehat{X}_{ij})(X_{ij}-\widehat{X}_{ij})$, where $\phi(z)=z^{\alpha}-z^{\beta}$, $\alpha\geq 1$, and $0<\beta<1$	$\sum_{ij}\Big[(\alpha-1)\widehat{X}_{ij}^{\alpha}-(\beta-1)\widehat{X}_{ij}^{\beta}-\alpha X_{ij}\widehat{X}_{ij}^{\alpha-1}+\beta X_{ij}\widehat{X}_{ij}^{\beta-1}\Big]$
Kullback-Leibler divergence	$D_{\mathrm{KL}}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \sum_{ij}X_{ij}\ln\frac{X_{ij}}{\widehat{X}_{ij}/\sum_{ab}\widehat{X}_{ab}}$, where $\sum_{ij}X_{ij}=1$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\left[-\sum_{ij}X_{ij}\widehat{X}_{ij}^{\epsilon}+\left(\sum_{ij}\widehat{X}_{ij}\right)^{\epsilon}\right]$
γ-divergence	$D_{\gamma}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \frac{1}{\gamma(1+\gamma)}\left[\ln\left(\sum_{ij}X_{ij}^{1+\gamma}\right)+\gamma\ln\left(\sum_{ij}\widehat{X}_{ij}^{1+\gamma}\right)-(1+\gamma)\ln\left(\sum_{ij}X_{ij}\widehat{X}_{ij}^{\gamma}\right)\right]$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\left[\frac{\left(\sum_{ij}\widehat{X}_{ij}^{1+\gamma}\right)^{\epsilon}}{1+\gamma}-\frac{\left(\sum_{ij}X_{ij}\widehat{X}_{ij}^{\gamma}\right)^{\epsilon}}{\gamma}\right]$
Rényi divergence	$D_{\rho}(\mathbf{X}\|\|\widehat{\mathbf{X}}) = \frac{1}{\rho-1}\ln\left[\sum_{ij}\left(\frac{X_{ij}}{\sum_{ab}X_{ab}}\right)^{\rho}\left(\frac{\widehat{X}_{ij}}{\sum_{ab}\widehat{X}_{ab}}\right)^{1-\rho}\right]$, where $\rho>0$	$\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\left[\frac{\left(\sum_{ij}X_{ij}^{\rho}\widehat{X}_{ij}^{1-\rho}\right)^{\epsilon}}{\rho-1}+\left(\sum_{ij}\widehat{X}_{ij}\right)^{\epsilon}\right]$

TABLE V
MULTIPLICATIVE UPDATE RULES FOR NMF OBJECTIVES IN TABLE IV OF THE PAPER USING THE PROPOSED PRINCIPLE, WHERE $Z_{ij} = X_{ij}/\left(\mathbf{WH}\right)_{ij}$

objective	update rule for $\mathbf{W}$
Euclidean distance	$W_{ik}^{\text{new}} = W_{ik} \frac{\left(\mathbf{XH}^T\right)_{ik}}{\left(\mathbf{WHH}^T\right)_{ik}}$
I-divergence	$W_{ik}^{\text{new}} = W_{ik} \frac{\left(\mathbf{ZH}^T\right)_{ik}}{\sum_j H_{kj}}$
dual I-divergence	$W_{ik}^{\text{new}} = W_{ik} \exp\left(\frac{\sum_j (\ln Z_{ij}) H_{kj}}{\sum_j H_{kj}}\right)$
Itakura-Saito divergence	$W_{ik}^{\text{new}} = W_{ik} \sqrt{\frac{\sum_j X_{ij} (\mathbf{WH})_{ij}^{-2} H_{kj}}{\sum_j (\mathbf{WH})_{ij}^{-1} H_{kj}}}$
α-divergence	$W_{ik}^{\text{new}} = \begin{cases} W_{ik} \left(\frac{\sum_j Z_{ij}^{\alpha} H_{kj}}{\sum_j H_{kj}}\right)^{\frac{1}{\alpha}} & \text{for } \alpha \neq 0 \\ W_{ik} \exp\left(\frac{\sum_j (\ln Z_{ij}) H_{kj}}{\sum_j H_{kj}}\right) & \text{for } \alpha \to 0 \end{cases}$
β-divergence	$W_{ik}^{\text{new}} = W_{ik} \left[\frac{\sum_j X_{ij} (\mathbf{WH})_{ij}^{\beta-1} H_{kj}}{\sum_j (\mathbf{WH})_{ij}^{\beta} H_{kj}}\right]^{\eta}$, where $\eta = \begin{cases} 1/\beta & \text{for } \beta > 1 \\ 1 & \text{for } 0 < \beta \leq 1 \\ 1/(1-\beta) & \text{for } \beta < 0 \end{cases}$
Log-Quad cost	$W_{ik}^{\text{new}} = W_{ik} \sqrt{\frac{(\mathbf{ZH}^T + 2\mathbf{XH}^T)_{ik}}{\sum_j H_{kj} + 2(\mathbf{WHH}^T)_{ik}}}$
$\alpha\beta$-Bregman divergence	$W_{ik}^{\text{new}} = W_{ik} \left[\frac{\alpha(\alpha-1) \sum_j X_{ij} (\mathbf{WH})_{ij}^{\alpha-2} H_{kj} + \beta(1-\beta) \sum_j X_{ij} (\mathbf{WH})_{ij}^{\beta-2} H_{kj}}{\alpha(\alpha-1) \sum_j (\mathbf{WH})_{ij}^{\alpha-1} H_{kj} + \beta(1-\beta) \sum_j (\mathbf{WH})_{ij}^{\beta-1} H_{kj}}\right]^{\frac{1}{\alpha-\beta+1}}$
Kullback-Leibler divergence	$W_{ik}^{\text{new}} = W_{ik} \frac{\left(\mathbf{ZH}^T\right)_{ik}}{\sum_j H_{kj}} \sum_{ab} \left(\mathbf{WH}\right)_{ab}$
γ-divergence	$W_{ik}^{\text{new}} = W_{ik} \left[\frac{\sum_j X_{ij} (\mathbf{WH})_{ij}^{\gamma-1} H_{kj}}{\sum_j (\mathbf{WH})_{ij}^{\gamma} H_{kj}} \frac{\sum_{ab} (\mathbf{WH})_{ab}^{1+\gamma}}{\sum_{ab} X_{ab} (\mathbf{WH})_{ab}^{\gamma}}\right]^{\eta}$, where $\eta = \begin{cases} 1/(1+\gamma) & \text{for } \gamma > 0 \\ 1/(1-\gamma) & \text{for } \gamma < 0 \end{cases}$
Rényi divergence	$W_{ik}^{\text{new}} = W_{ik} \left[\frac{\sum_j Z_{ij}^{r} H_{kj}}{\sum_j H_{kj}} \frac{\sum_{ab} (\mathbf{WH})_{ab}}{\sum_{ab} X_{ab}^{r} (\mathbf{WH})_{ab}^{1-r}}\right]^{\eta}$, where $\eta = \begin{cases} 1/r & \text{for } r > 1 \\ 1 & \text{for } 0 < r < 1 \end{cases}$

[17] D. Kim, S. Sra, and I. S. Dhillon, "Fast projection-based methods for the least squares nonnegative matrix approximation problem," *Statistical Analysis and Data Mining*, vol. 1, no. 1, pp. 38–51, 2008.
[18] Z. Yang and J. Laaksonen, "Multiplicative updates for non-negative projections," *Neurocomputing*, vol. 71, no. 1-3, pp. 363–373, 2007.
[19] H. Fujisawa and S. Eguchi, "Robust paramater estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, pp. 2053–2081, 2008.
[20] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, in press. [Online]. Available: http://arxiv.org/abs/1010.1763
[21] A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. of International Conference on Independent Component Analysis and Blind Signal Separation*, 2006, pp. 32–39.
[22] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010.
[23] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.
[24] L. T. H. An and P. D. Tao, "The DC (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems," *Annals of Operations Research*, vol. 133, pp. 23–46, 2005.
[25] F. Sha, Y. Lin, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Computation*, vol. 19, no. 8, pp. 2004–2031, 2007.
[26] M. Nakano, H. Kameoka, J. L. Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *Proc. of IEEE International Workshop on Machine Learning For Signal Processing*, 2010, pp. 283–288.
[27] M. Spivak, *Calculus*. Houston, Texas: Publish or Perish, 1994.
[28] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.
[29] R. Salakhutdinov and S. Roweis, "Adaptive overrelaxed bound optimization methods," in *Proc. of the Twentieth International Conference on Machine Learning*, 2003, pp. 664–671.
[30] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," in *Proc. of 14th Scandinavian Conference on Image Analysis*, Joensuu, Finland, June 2005, pp. 333–342.
[31] Z. Yang, Z. Yuan, and J. Laaksonen, "Projective non-negative matrix factorization with applications to facial image processing," *International Journal on Pattern Recognition and Artificial Intelligence*, vol. 21, no. 8, pp. 1353–1362, Dec. 2007.
[32] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Transaction on Neural Networks*, vol. 21, no. 5, pp. 734–749, 2010.
[33] C. Ding, T. Li, and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding," in *Proc. of the 8th IEEE International Conference on Data Mining*, 2008, pp. 183–192.
[34] B. Lakshminarayanan and R. Raich, "Non-negative matrix factorization for parameter estimation in hidden Markov models," in *Proc. of IEEE International Workshop on Machine Learning For Signal Processing*, 2010, pp. 89–94.
[35] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions On Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.
[36] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, 2003, pp. 1141–1148.
[37] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. of the Second IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.
[38] E. F. Gonzalez and Y. Zhang, "Accelerating the Lee-Seung algorithm for nonnegative matrix factorization," Rice University, Tech. Rep., 2005. [Online]. Available: www.caam.rice.edu/tech_reports/2005/TR05-02.ps
[39] R. Badeau, N. Bertin, and E. Vincent, "Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 21, no. 12, pp. 1869–1881, 2010.
[40] Z. Zheng, J. Yang, and Y. Zhu, "Initialization enhancer for non-negative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 1, pp. 101–110, 2007.
[41] Y. Kim and S. Choi, "A method of initialization for nonnegative matrix factorization," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 537–540.
[42] L. Eldén, *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial and Applied Mathematics, 2007.
[43] A. Cichocki and S.-I. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, pp. 1532–1568, 2010.
[44] A. Cichocki, S. Cruces, and S.-I. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, pp. 134–170, 2011.
[45] H. Choi, S. Choi, A. Katake, and Y. Choe, "Learning alpha-integration with partially-labeled data," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 14–19.
[46] M. Mollah, N. Sultana, and M. Minami, "Robust extraction of local structures by the minimum of beta-divergence method," *Neural Networks*, vol. 23, pp. 226–238, 2010.
[47] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for large margin classifiers," in *Proc. of 16th Annual Conference on Computational Learning Theory*, 2003, pp. 188–202.