Generalized Nonnegative Matrix Approximations with Bregman Divergences

Inderjit S. Dhillon   Suvrit Sra
Dept. of Computer Sciences
The Univ. of Texas at Austin
Austin, TX 78712.
{inderjit,suvrit}@cs.utexas.edu

UTCS Technical Report # TR-xx-05

1st June, 2005

Abstract

Nonnegative matrix approximation (NNMA) is a recent technique for dimensionality reduction and data analysis that yields a parts based, sparse nonnegative representation of the nonnegative input data. Due to these advantages, NNMA has found a wide variety of applications, including text analysis, document clustering, face/image recognition, language modeling, speech processing and many others. Despite these numerous applications, the algorithmic development for computing the NNMA factors has been relatively deficient. This paper makes algorithmic progress by modeling and *solving* (using multiplicative updates) new generalized NNMA problems that minimize Bregman divergences between the input matrix and its low-rank approximation. The multiplicative update formulae in the pioneering work by Lee and Seung [20] arise as a special case of our algorithms. In addition, the paper shows how to use penalty functions for incorporating constraints other than nonnegativity into the problem. Further, some interesting extensions to the use of "link" functions for modeling non-linear relationships are also discussed.

Keywords: Nonnegative matrix factorization, approximation, Bregman divergence, multiplicative updates.

1 Introduction

Nonnegative matrix approximation (NNMA) is a method for dimensionality reduction and data analysis that has gained substantial prominence over the past few years. NNMA has previously been called *positive matrix factorization* [27] and *nonnegative matrix factorization* [21]. Assume that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ are N nonnegative input (M-dimensional) vectors. We organize these vectors as the columns of a nonnegative data matrix

$$\boldsymbol{A} \triangleq \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_N \end{bmatrix}.$$

NNMA seeks a small set of K nonnegative representative vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K$ that can be nonnegatively (or conically) combined to approximate the input vectors $\boldsymbol{a}_i$. That is,

$$\boldsymbol{a}_n \approx \sum_{k=1}^{K} c_{kn} \boldsymbol{b}_k, \quad 1 \leq n \leq N,$$

where the combining coefficients c_{kn} are restricted to be nonnegative. If c_{kn} and $\boldsymbol{b}_k$ are unrestricted, and we minimize $\sum_n \|\boldsymbol{a}_n - \boldsymbol{B}\boldsymbol{c}_n\|^2$, the Truncated Singular Value Decomposition (TSVD) of $\boldsymbol{A}$ yields the optimal $\boldsymbol{b}_k$ and c_{kn} values. If the $\boldsymbol{b}_k$ are unrestricted, but the coefficient vectors $\boldsymbol{c}_n$ are restricted to be indicator vectors, then we obtain the problem of hard-clustering (See [37, Chapter 8] for related discussion regarding different constraints on $\boldsymbol{c}_n$ and $\boldsymbol{b}_k$).

In this paper we consider problems where all involved matrices are nonnegative. For many practical problems nonnegativity is a natural requirement. For example, color intensities, chemical concentrations, frequency counts etc., are all nonnegative entities, and approximating their measurements by nonnegative representations leads to greater interpretability. NNMA has found a significant number of applications, not only due to increased interpretability, but also because admitting only nonnegative combinations of the $\boldsymbol{b}_k$ leads to sparse representations.

This paper contributes to the algorithmic advancement of NNMA by generalizing the problem significantly, and deriving efficient algorithms based on multiplicative updates for the generalized problems. The scope of this paper is primarily on generic methods for NNMA, rather than on specific applications. The multiplicative update formulae in the pioneering work by Lee and Seung [20] arise as a special case of our algorithms, which seek to minimize Bregman divergences between the nonnegative input $\boldsymbol{A}$ and its approximation. In addition, we discuss the use penalty functions for incorporating constraints other than nonnegativity into the problem. Further, we illustrate an interesting extension of our algorithms for handling non-linear relationships through the use of “link” functions.

2 Problems

Given a nonnegative matrix $\boldsymbol{A}$ as input, the classical NNMA problem is to approximate it by a lower rank nonnegative matrix of the form $\boldsymbol{BC}$, where $\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_K]$ and $\boldsymbol{C} = [\boldsymbol{c}_1, ..., \boldsymbol{c}_N]$ are themselves nonnegative. That is, we seek the approximation,

$$\boldsymbol{A}_{M\times N} \approx \boldsymbol{B}_{M\times K}\boldsymbol{C}_{K\times N}, \qquad \text{where } \boldsymbol{B}, \boldsymbol{C} \geq 0. \tag{2.1}$$

We judge the goodness of the approximation in (2.1) by using a general class of distortion measures called *Bregman divergences*. For any strictly convex function $\varphi : S \subseteq \mathbb{R} \to \mathbb{R}$ that has a continuous first derivative, the corresponding **Bregman divergence** $D_\varphi : S \times \text{int}(S) \to \mathbb{R}_+$ is defined as $D_\varphi(x,\, y) \triangleq \varphi(x) - \varphi(y) - \nabla\varphi(y)(x-y)$, where $\text{int}(S)$ is the interior of set S [1, 3]. Bregman divergences are nonnegative, convex in the first argument and zero if and only if $x = y$. These divergences play an important role in convex optimization [3]. For the sequel we consider only separable Bregman divergences, i.e., $D_\varphi(\boldsymbol{X},\, \boldsymbol{Y}) = \sum_{ij} D_\varphi(x_{ij},\, y_{ij})$. We further require $x_{ij}, y_{ij} \in \text{dom}\varphi \cap \mathbb{R}_+$.

Formally, the resulting generalized nonnegative matrix approximation problems are:

$$\min_{\boldsymbol{B},\, \boldsymbol{C} \geq 0} \quad D_\varphi(\boldsymbol{BC},\, \boldsymbol{A}) + \alpha(\boldsymbol{B}) + \beta(\boldsymbol{C}), \tag{2.2}$$

$$\min_{\boldsymbol{B},\, \boldsymbol{C} \geq 0} \quad D_\varphi(\boldsymbol{A},\, \boldsymbol{BC}) + \alpha(\boldsymbol{B}) + \beta(\boldsymbol{C}). \tag{2.3}$$

The functions α and β serve as *penalty* functions, and they allow us to enforce regularization (or other constraints) on $\boldsymbol{B}$ and $\boldsymbol{C}$. We consider both (2.2) and (2.3) since Bregman divergences are generally asymmetric. Table 1 gives a small sample of NNMA problems to illustrate the breadth of our formulation.

Divergence D_φ	φ	α	β	Remarks
$\lVert\boldsymbol{A} - \boldsymbol{BC}\rVert_\text{F}^2$	$\frac{1}{2}x^2$	$\mathbf{0}$	$\mathbf{0}$	Lee and Seung [20, 21]
$\lVert\boldsymbol{A} - \boldsymbol{BC}\rVert_\text{F}^2$	$\frac{1}{2}x^2$	$\mathbf{0}$	$\lambda\mathbf{1}^T\boldsymbol{C}\mathbf{1}$	Hoyer [17]
$\lVert\boldsymbol{W} \odot (\boldsymbol{A} - \boldsymbol{BC})\rVert_\text{F}^2$	$\frac{1}{2}x^2$	$\mathbf{0}$	$\mathbf{0}$	Paatero and Tapper [27]
$\text{KL}(\boldsymbol{A}, \boldsymbol{BC})$	$x \log x - x$	$\mathbf{0}$	$\mathbf{0}$	Lee and Seung [20]
$\text{KL}(\boldsymbol{A}, \boldsymbol{WBC})$	$x \log x - x$	$\mathbf{0}$	$\mathbf{0}$	Guillamet et al. [15]
$\text{KL}(\boldsymbol{A}, \boldsymbol{BC})$	$x \log x - x$	$c_1\mathbf{1}^T\boldsymbol{B}^T\boldsymbol{B}\mathbf{1}$	$-c_2\lVert\boldsymbol{C}\rVert_\text{F}^2$	Feng et al. [10]
$D_\varphi(\boldsymbol{A},\, \boldsymbol{W}_1\boldsymbol{BCW}_2)$	$\varphi(x)$	$\alpha(\boldsymbol{B})$	$\beta(\boldsymbol{C})$	Weighted NNMA (new)

Table 1: Some example NNMA problems that may be obtained from (2.3). The corresponding asymmetric problem (2.2) has *not* been previously treated in the literature. KL$(x,\, y)$ denotes the generalized KL-Divergence = $\sum_i x_i \log \frac{x_i}{y_i} - x_i + y_i$ (also called I-divergence).

3 Algorithms

In this section we present algorithms that seek to optimize (2.2) and (2.3). Our algorithms are iterative in nature, and are directly inspired by the efficient algorithms of Lee and Seung [20]. Appealing properties include ease of implementation and computational efficiency.

Note that the problems (2.2) and (2.3) are not jointly convex in $\boldsymbol{B}$ and $\boldsymbol{C}$, so it is not easy to obtain globally optimal solutions in polynomial time. Our iterative procedures start by initializing $\boldsymbol{B}$ and $\boldsymbol{C}$ randomly or otherwise. Then, $\boldsymbol{B}$ and $\boldsymbol{C}$ are alternately updated until there is no further appreciable change in the objective function value.

3.1 Algorithms for (2.2)

We utilize the concept of auxiliary functions [20] for our derivations. It is sufficient to illustrate our methods using a single column of $\boldsymbol{C}$ (or row of $\boldsymbol{B}$), since our divergences are separable.

Definition 3.1 (Auxiliary function)**.** A function $G(\boldsymbol{c}, \boldsymbol{c}')$ is called an auxiliary function for $F(\boldsymbol{c})$ if:

1. $G(\boldsymbol{c}, \boldsymbol{c}) = F(\boldsymbol{c})$, and
2. $G(\boldsymbol{c}, \boldsymbol{c}') \geq F(\boldsymbol{c})$ for all $\boldsymbol{c}'$.

Auxiliary functions turn out to be useful due to the following lemma.

Lemma 3.2 (Iterative minimization)**.** *If $G(\boldsymbol{c}, \boldsymbol{c}')$ is an auxiliary function for $F(\boldsymbol{c})$, then F is non-increasing under the update*

$$\boldsymbol{c}^{t+1} = \operatorname{argmin}_{\boldsymbol{c}} G(\boldsymbol{c}, \boldsymbol{c}^t).$$

Proof. $F(\boldsymbol{c}^{t+1}) \leq G(\boldsymbol{c}^{t+1}, \boldsymbol{c}^t) \leq G(\boldsymbol{c}^t, \boldsymbol{c}^t) = F(\boldsymbol{c}^t)$. □

As can be observed, the sequence formed by the iterative application of Lemma 3.2 leads to a monotonic decrease in the objective function value $F(\boldsymbol{c})$. For an algorithm that iteratively updates $\boldsymbol{c}$ in its quest to minimize $F(\boldsymbol{c})$, the method for proving convergence boils down to the construction of an appropriate auxiliary function. Auxiliary functions have been used in many places before, see for example [5, 20].

We now construct simple auxiliary functions for (2.2) that yield multiplicative updates. To avoid clutter we drop the functions α and β from (2.2), noting that our methods can easily be extended to incorporate these functions.

Suppose $\boldsymbol{B}$ is fixed and we wish to compute an updated column of $\boldsymbol{C}$. We wish to minimize

$$F(\boldsymbol{c}) = D_\varphi(\boldsymbol{Bc},\, \boldsymbol{a}), \tag{3.1}$$

where $\boldsymbol{a}$ is the column of $\boldsymbol{A}$ corresponding to the column $\boldsymbol{c}$ of $\boldsymbol{C}$. The lemma below shows how to construct an auxiliary function for (3.1). For convenience of notation we use ψ to denote $\nabla\varphi$ for the rest of this section.

Lemma 3.3 (Auxiliary function)**.** *The function*

$$G(\boldsymbol{c}, \boldsymbol{c}') = \sum_{ij} \lambda_{ij} \varphi\left(\frac{b_{ij}c_j}{\lambda_{ij}}\right) - \sum_i \varphi(a_i) - \psi(a_i)\big((\boldsymbol{Bc})_i - a_i\big), \tag{3.2}$$

with $\lambda_{ij} = (b_{ij}c'_j)/(\sum_l b_{il}c'_l)$, is an auxiliary function for (3.1)*. Note that by definition $\sum_j \lambda_{ij} = 1$, and as both b_{ij} and c'_j are nonnegative, $\lambda_{ij} \geq 0$.*

Proof. It is easy to verify that $G(\boldsymbol{c}, \boldsymbol{c}) = F(\boldsymbol{c})$, since $\sum_j \lambda_{ij} = 1$. Using the convexity of φ, we conclude that if $\sum_j \lambda_{ij} = 1$ and $\lambda_{ij} \geq 0$, then

$$\begin{aligned}
F(\boldsymbol{c}) &= \sum_i \varphi\left(\sum_j b_{ij}c_j\right) - \varphi(a_i) - \psi(a_i)\big((\boldsymbol{Bc})_i - a_i\big) \\
&\leq \sum_{ij} \lambda_{ij} \varphi\left(\frac{b_{ij}c_j}{\lambda_{ij}}\right) - \sum_i \varphi(a_i) - \psi(a_i)\big((\boldsymbol{Bc})_i - a_i\big) \\
&= G(\boldsymbol{c}, \boldsymbol{c}').
\end{aligned}$$

□

To obtain the update, we minimize $G(\boldsymbol{c}, \boldsymbol{c}')$ w.r.t. $\boldsymbol{c}$. Let $\psi(\boldsymbol{x})$ denote the vector $[\psi(x_1), \ldots, \psi(x_n)]^T$. We compute the partial derivative

$$\begin{aligned}\frac{\partial G}{\partial c_p} &= \sum_i \lambda_{ip} \psi\left(\frac{b_{ip} c_p}{\lambda_{ip}}\right) \frac{b_{ip}}{\lambda_{ip}} - \sum_i b_{ip} \psi(a_i) \\ &= \sum_i b_{ip} \psi\left(\frac{c_p}{c'_p} (\boldsymbol{B}\boldsymbol{c}')_i\right) - (\boldsymbol{B}^T \psi(\boldsymbol{a}))_p. \qquad (3.3)\end{aligned}$$

We need to solve (3.3) for c_p by setting $\partial G / \partial c_p = 0$. Solving this equation analytically is not always possible. However, for a broad class of functions, we can obtain an analytic solution. For example, if ψ is multiplicative (i.e., $\psi(xy) = \psi(x)\psi(y)$) we obtain a multiplicative update relation for $\boldsymbol{c}$ as follows:

$$\frac{\partial G}{\partial c_p} = \sum_i b_{ip} \psi\left(\frac{c_p{}'}{c_p}\right) \psi\big((\boldsymbol{B}\boldsymbol{c}')_i\big) - (\boldsymbol{B}^T \psi(\boldsymbol{a}))_p \quad = \quad \psi\left(\frac{c_p{}'}{c_p}\right) [\boldsymbol{B}^T \psi(\boldsymbol{B}\boldsymbol{c}')]_p - [\boldsymbol{B}^T \psi(\boldsymbol{a})]_p.$$

Thus upon setting $\partial G / \partial c_p = 0$, we obtain

$$\psi\left(\frac{c_p{}'}{c_p}\right) = \frac{[\boldsymbol{B}^T \psi(\boldsymbol{a})]_p}{[\boldsymbol{B}^T \psi(\boldsymbol{B}\boldsymbol{c}')]_p},$$

which yields the update

$$c_p \leftarrow \quad c_p \cdot \psi^{-1}\Big(\frac{[\boldsymbol{B}^T \psi(\boldsymbol{a})]_p}{[\boldsymbol{B}^T \psi(\boldsymbol{B}\boldsymbol{c})]_p}\Big). \qquad (3.4)$$

We can compute updates for $\boldsymbol{B}$ one row at a time. Let this row be denoted by $\boldsymbol{b}^T$, and the corresponding row of matrix $\boldsymbol{A}$ by $\boldsymbol{a}^T$. The objective function for a row is

$$H(\boldsymbol{b}) = D_\varphi(\boldsymbol{b}^T \boldsymbol{C},\, \boldsymbol{a}^T) = \sum_j D_\varphi(\boldsymbol{b}^T \boldsymbol{c}_j,\, a_j),$$

where $\boldsymbol{c}_j$ denotes the j-th column of $\boldsymbol{C}$, and a_j denotes the j-th component of the row vector $\boldsymbol{a}^T$. Once again, using the convexity of φ, we define an appropriate auxiliary function $K(\boldsymbol{b}, \boldsymbol{b}')$ for $H(\boldsymbol{b})$, where

$$K(\boldsymbol{b}, \boldsymbol{b}') = \sum_{jk} \mu_{kj} \varphi\big(\frac{c_{kj} b_k}{\mu_{kj}}\big) - \sum_j \varphi(a_j) - \psi(a_j)(\boldsymbol{b}^T \boldsymbol{c}_j - a_j),$$

and $\mu_{kj} = c_{kj} b'_k / (\sum_l c_{lj} b'_l)$, and $\mu_{kl} \geq 0$.
We now compute $\partial K / \partial b_p$ to obtain

$$\frac{\partial K}{\partial b_p} = \sum_j c_{pj} \psi\left(\frac{b_p}{b'_p} \boldsymbol{b}'^T \boldsymbol{c}_j\right) - \sum_j c_{pj} \psi(a_j).$$

Once again, assuming multiplicative ψ, and setting $\partial K / \partial b_p = 0$, we obtain the update

$$b_p \leftarrow b_p \cdot \psi^{-1}\Big(\frac{[\psi(\boldsymbol{a}^T) \boldsymbol{C}^T]_p}{[\psi(\boldsymbol{b}^T \boldsymbol{C}) \boldsymbol{C}^T]_p}\Big). \qquad (3.5)$$

It turns out that when φ is a convex function of Legendre type, then ψ^{-1} can be obtained by the derivative of the conjugate function φ^* of φ, i.e., $\psi^{-1} = \nabla \varphi^*$ [31].
Note. (3.5) & (3.4) coincide with updates derived by Lee and Seung [20], if $\varphi(x) = \frac{1}{2} x^2$.

3.2 Examples of New NNMA Problems

We illustrate the power of our generic auxiliary functions given above for deriving algorithms with multiplicative updates for some specific interesting problems.

3.2.1 New KL-Divergence NNMA

First we consider the problem that seeks to minimize the divergence,

$$\mathrm{KL}(\boldsymbol{Bc}, \boldsymbol{a}) \;=\; \sum_i (\boldsymbol{Bc})_i \log \frac{(\boldsymbol{Bc})_i}{a_i} - (\boldsymbol{Bc})_i + a_i, \qquad \boldsymbol{B}, \boldsymbol{c} \geq 0. \tag{3.6}$$

Let $\varphi(x) = x \log x - x$. Then, $\psi(x) = \log x$, and as $\psi(xy) = \psi(x) + \psi(y)$, upon substituting in (3.3), and setting the resultant to zero we obtain

$$\begin{aligned}
\frac{\partial G}{\partial c_p} &= \sum_i b_{ip} \log(c_p(\boldsymbol{Bc}')_i / c'_p) - \sum_i b_{ip} \log a_i = 0, \\
\implies (\boldsymbol{B}^T\boldsymbol{1})_p \log \frac{c_p}{c'_p} &= [\boldsymbol{B}^T \log \boldsymbol{a} - \boldsymbol{B}^T \log(\boldsymbol{Bc}')]_p \\
\implies c_p = c'_p \cdot \exp&\left(\frac{[\boldsymbol{B}^T \log\big(\boldsymbol{a}/(\boldsymbol{Bc}')\big)]_p}{[\boldsymbol{B}^T\boldsymbol{1}]_p} \right).
\end{aligned}$$

The update for $\boldsymbol{b}$ is derived to be

$$b_p = b'_p \cdot \exp\left(\frac{[\big(\log(\boldsymbol{a}/\boldsymbol{b}'^T\boldsymbol{C})\big)^T \boldsymbol{C}^T]_p}{[\boldsymbol{1}^T\boldsymbol{C}^T]_p} \right).$$

3.2.2 Constrained NNMA

Next we consider NNMA problems that have additional constraints. We illustrate our ideas on a problem with linear constraints.

$$\begin{aligned}
\min_{\boldsymbol{c}} \quad & D_\varphi(\boldsymbol{Bc},\, \boldsymbol{a}) \\
\text{s.t.} \quad & \boldsymbol{Pc} \leq \boldsymbol{0}, \quad \boldsymbol{c} \geq \boldsymbol{0}.
\end{aligned} \tag{3.7}$$

We can solve (3.7) problem using our method by making use of an appropriate (differentiable) penalty function that enforces $\boldsymbol{Pc} \leq \boldsymbol{0}$. We consider,

$$F(\boldsymbol{c}) = D_\varphi(\boldsymbol{Bc},\, \boldsymbol{a}) + \rho \|\max(0, \boldsymbol{Pc})\|^2, \tag{3.8}$$

where $\rho > 0$ is some penalty constant. Assuming multiplicative ψ and following the auxiliary function technique described above, we obtain the following updates for $\boldsymbol{c}$,

$$c_p \leftarrow c_p \cdot \psi^{-1}\left(\frac{[\boldsymbol{B}^T\psi(\boldsymbol{a})]_p - \rho[\boldsymbol{P}^T(\boldsymbol{Pc})^+]_p}{[\boldsymbol{B}^T\psi(\boldsymbol{Bc})]_p} \right),$$

where $(\boldsymbol{Pc})^+ = \max(\boldsymbol{0}, \boldsymbol{Pc})$. Note that care must be taken to ensure that the addition of this penalty term does not violate the nonnegativity of $\boldsymbol{c}$, and to ensure that the argument of ψ^{-1} lies in its domain.

Remarks. Incorporating additional constraints into (3.6) is however easier, since the exponential updates ensure nonnegativity. Given $\boldsymbol{a} = \boldsymbol{1}$, with appropriate penalty functions, our solution to (3.6) can be utilized for maximizing entropy of $\boldsymbol{Bc}$ subject to linear or non-linear constraints on $\boldsymbol{c}$. That is, for the maximum entropy problem

$$\max_{\boldsymbol{c}} \quad \text{ent}(\boldsymbol{Bc}) \quad \text{s.t. } \boldsymbol{Pc} \leq 0, \boldsymbol{c} \geq 0,$$

we solve the corresponding problem (with appropriate normalization)

$$\min \mathrm{KL}(\boldsymbol{Bc}, \boldsymbol{1}) \qquad \text{s.t. } \boldsymbol{Pc} \leq 0, \boldsymbol{c} \geq 0.$$

Using the penalty function as described above, and employing the iterative update relation derived in Section 3.2 we obtain the following update scheme for $\boldsymbol{c}$

$$c_p \quad \leftarrow \quad c_p \cdot \exp\left(\frac{[-\boldsymbol{B}^T \log(\boldsymbol{Bc}) - \rho \boldsymbol{P}^T(\boldsymbol{Pc})^+]_p}{[\boldsymbol{B}^T\boldsymbol{1}]_p} \right)$$

3.2.3 Nonlinear models with “link” functions

If $\boldsymbol{A} \approx h(\boldsymbol{BC})$, where h is a “link” function that models a nonlinear relationship between $\boldsymbol{A}$ and the approximant $\boldsymbol{BC}$, we may wish to minimize $D_\varphi(h(\boldsymbol{BC}),\, \boldsymbol{A})$. We can easily extend our methods to handle this case for appropriate h. Recall that the auxiliary function that we used, depended upon the convexity of φ. Thus, if $(\varphi \circ h)$ is a convex function, whose derivative $(\varphi \circ h)'(x)$ is “factorizable,” then we can easily derive algorithms for this problem with link functions.

For example, if h is convex (concave) and φ is an increasing (decreasing) function then, $\varphi \circ h$ is also convex, since the second derivative of the composition $\varphi \circ h$ is given by

$$(\varphi \circ h)''(x) = h''(x)\psi(h(x)) + \psi'(h(x))(h'(x))^2, \tag{3.9}$$

and it is nonnegative for such h and φ.

3.3 Algorithms using KKT conditions

We now derive efficient multiplicative update relations for (2.3), and these updates turn out to be simpler than those for (2.2). To avoid clutter, we describe our methods with $\alpha \equiv 0$, and $\beta \equiv 0$, noting that if α and β are differentiable, then it is easy to incorporate them in our derivations. For convenience we use $\zeta(x)$ to denote $\nabla^2(x)$ for the rest of this section.

We now compute the gradient $\nabla_{\boldsymbol{B}} D_\varphi(\boldsymbol{A},\, \boldsymbol{BC})$. Using the fact that $\partial(\boldsymbol{BC})_{ij}/\partial b_{pq} = c_{qj}$, we see that $\partial D_\varphi(\boldsymbol{A},\, \boldsymbol{BC})/\partial b_{pq}$ is given by

$$\begin{aligned}
&\frac{\partial}{\partial b_{pq}}\Big\{\sum_{ij}\varphi(a_{ij}) - \varphi((\boldsymbol{BC})_{ij}) - \psi((\boldsymbol{BC})_{ij})(a_{ij} - (\boldsymbol{BC})_{ij})\Big\}\\
=&\sum_j -\psi((\boldsymbol{BC})_{pj})c_{qj} - \zeta((\boldsymbol{BC})_{pj})c_{qj}(\boldsymbol{BC})_{pj} + c_{qj}\psi((\boldsymbol{BC})_{pj}) - \zeta((\boldsymbol{BC})_{pj})c_{qj}a_{pj}\\
=&\sum_j \zeta((\boldsymbol{BC})_{pj})((\boldsymbol{BC})_{pj} - a_{pj})c_{qj}\\
=&\big[\big(\zeta(\boldsymbol{BC}) \odot (\boldsymbol{BC} - \boldsymbol{A})\big)\boldsymbol{C}^T\big]_{pq}.
\end{aligned}$$

In a similar way, using the fact that $\partial(\boldsymbol{BC})_{ij}/\partial c_{pq} = b_{ip}$, we see that $\partial D_\varphi(\boldsymbol{A},\, \boldsymbol{BC})/\partial c_{pq}$ is given by

$$\big[\boldsymbol{B}^T\big(\zeta(\boldsymbol{BC}) \odot (\boldsymbol{BC} - \boldsymbol{A})\big)\big]_{pq}.$$

According to the KKT conditions, there exist Lagrange multiplier matrices $\boldsymbol{\Lambda} \geq 0$ and $\boldsymbol{\Omega} \geq 0$ such that

$$[\nabla_{\boldsymbol{B}} D_\varphi(\boldsymbol{A},\, \boldsymbol{BC})]_{mk} = \lambda_{mk}, \qquad [\nabla_{\boldsymbol{C}} D_\varphi(\boldsymbol{A},\, \boldsymbol{BC})]_{kn} = \omega_{kn}, \tag{3.10a}$$

$$\lambda_{mk} b_{mk} = \omega_{kn} c_{kn} = 0. \tag{3.10b}$$

Multiplying (3.10a)a by b_{mk}, and using (3.10b, we obtain

$$\big[\big(\zeta(\boldsymbol{BC}) \odot (\boldsymbol{BC} - \boldsymbol{A})\big)\boldsymbol{C}^T\big]_{mk} b_{mk} = \lambda_{mk} b_{mk} = 0,$$

which suggests the iterative scheme

$$b_{mk} \leftarrow b_{mk}\frac{\big[\big(\zeta(\boldsymbol{BC}) \odot \boldsymbol{A}\big)\boldsymbol{C}^T\big]_{mk}}{\big[\big(\zeta(\boldsymbol{BC}) \odot \boldsymbol{BC}\big)\boldsymbol{C}^T\big]_{mk}}. \tag{3.11}$$

Proceeding in a similar fashion we obtain a similar iterative scheme for c_{kn}, which is

$$c_{kn} \leftarrow c_{kn}\frac{[\boldsymbol{B}^T\big(\zeta(\boldsymbol{BC}) \odot \boldsymbol{A}\big)]_{kn}}{[\boldsymbol{B}^T\big(\zeta(\boldsymbol{BC}) \odot \boldsymbol{BC}\big)]_{kn}}. \tag{3.12}$$

3.4 Examples of New and Old NNMA Problems as Special Cases

We now illustrate the power of our approach by showing how one can easily obtain iterative update relations for many NNMA problems, including known and new problems.

3.4.1 Lee and Seung's Algorithms.

Let $\alpha \equiv 0$, $\beta \equiv 0$. Now if we set $\varphi(x) = \frac{1}{2}x^2$, then (3.11) and (3.12) reduce to

$$b_{mk} \leftarrow b_{mk} \frac{(\boldsymbol{A}\boldsymbol{C}^T)_{mk}}{(\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^T)_{mk}}, \qquad c_{kn} \leftarrow c_{kn} \frac{(\boldsymbol{B}^T\boldsymbol{A})_{kn}}{(\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{C})_{kn}},$$

and these correspond to the Frobenius norm update rules originally derived by Lee and Seung [20].

If $\varphi(x) = x \log x$, then $\zeta(x) = 1/x$. With $\alpha \equiv 0$ and $\beta \equiv 0$, (3.11) and (3.12) reduce to

$$\begin{aligned} b_{mk} &\leftarrow b_{mk} \left\{ \frac{\left([\frac{\boldsymbol{A}}{\boldsymbol{BC}}]\boldsymbol{C}^T\right)_{mk}}{(\boldsymbol{1}_M\boldsymbol{1}_N^T\boldsymbol{C}^T)_{mk}} = \frac{\sum_s c_{ks} a_{ms}/(\boldsymbol{BC})_{ms}}{\sum_n c_{kn}} \right\}, \\ c_{kn} &\leftarrow c_{kn} \left\{ \frac{\left(\boldsymbol{B}^T[\frac{\boldsymbol{A}}{\boldsymbol{BC}}]\right)_{kn}}{(\boldsymbol{B}^T\boldsymbol{1}_M\boldsymbol{1}_N^T)_{kn}} = \frac{\sum_t b_{tk} a_{tn}/(\boldsymbol{BC})_{tn}}{\sum_m b_{mk}} \right\}. \end{aligned}$$

These updates are the same as the ones originally derived by Lee and Seung [20].

3.4.2 Elementwise weighted distortion.

Here we wish to minimize $\|\boldsymbol{W} \odot (\boldsymbol{A} - \boldsymbol{BC})\|_F^2$. Using $\boldsymbol{X} \leftarrow \sqrt{\boldsymbol{W}} \odot \boldsymbol{X}$, and $\boldsymbol{A} \leftarrow \sqrt{\boldsymbol{W}} \odot \boldsymbol{A}$ in (3.11) and (3.12) one obtains

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \odot \frac{(\boldsymbol{W} \odot \boldsymbol{A})\boldsymbol{C}^T}{(\boldsymbol{W} \odot (\boldsymbol{BC}))\boldsymbol{C}^T}, \qquad \boldsymbol{C} \leftarrow \boldsymbol{C} \odot \frac{\boldsymbol{B}^T(\boldsymbol{W} \odot \boldsymbol{A})}{\boldsymbol{B}^T(\boldsymbol{W} \odot (\boldsymbol{BC}))}.$$

These iterative updates are significantly simpler than the PMF algorithms of [27], and can be used as an alternate way for obtaining elementwise weighted approximations.

3.4.3 The Multifactor NNMA Problem (new).

The above ideas can be extended to the multifactor NNMA problem that seeks to minimize the following divergence

$$D_\varphi(\boldsymbol{A}, \boldsymbol{B}_1\boldsymbol{B}_2 \ldots \boldsymbol{B}_R),$$

where all matrices involved are nonnegative. We compute the gradient of the distortion w.r.t. each $\boldsymbol{B}_r$. Let $\widehat{\boldsymbol{B}} = \boldsymbol{B}_1\boldsymbol{B}_2 \ldots \boldsymbol{B}_{r-1}$, $\widehat{\boldsymbol{C}} = \boldsymbol{B}_{r+1}\boldsymbol{B}_{r+2} \ldots \boldsymbol{B}_R$, $\boldsymbol{H} = \boldsymbol{B}_1\boldsymbol{B}_2 \ldots \boldsymbol{B}_R$, $\boldsymbol{Z} =$. We have

$$\begin{aligned} \frac{\partial D_\varphi}{\partial b_{pq}^r} &= \sum_{ij} \zeta(h_{ij})(h_{ij} - a_{ij}) \frac{\partial}{\partial b_{pq}^r} \sum_{kl} \widehat{\boldsymbol{B}}_{ik}(\boldsymbol{B}_r)_{kl}\widehat{\boldsymbol{C}}_{lj} \\ &= \sum_{ij} \zeta(h_{ij})(h_{ij} - a_{ij})\hat{b}_{ip}\hat{c}_{qj} \\ &= \left[\widehat{\boldsymbol{B}}^T(\zeta(\boldsymbol{H}) \odot (\boldsymbol{H} - \boldsymbol{A}))\widehat{\boldsymbol{C}}^T\right]_{pq}. \end{aligned}$$

Thus the update formula for $\boldsymbol{B}_r$ is given by

$$\boldsymbol{B}_r \leftarrow \boldsymbol{B}_r \odot \frac{\widehat{\boldsymbol{B}}^T(\zeta(\boldsymbol{H}) \odot \boldsymbol{A})\widehat{\boldsymbol{C}}^T}{\widehat{\boldsymbol{B}}^T(\zeta(\boldsymbol{H}) \odot \boldsymbol{H})\widehat{\boldsymbol{C}}^T}. \tag{3.13}$$

A typical usage of multifactor NNMA problem would be to obtain a three-factor NNMA, namely $\boldsymbol{A} \approx \boldsymbol{RBC}$. Such an approximation is closely tied to the problem of co-clustering [4], and can be used to produce relaxed co-clustering solutions.

3.4.4 Weighted NNMA Problems

There are two main ways in which weighting can be incorporated into the NNMA model. First is the model that uses elementwise weighting, and the second is the a generalized weighting scheme. Formally, these problems are

$$\begin{aligned} \min \quad & D_\varphi(\boldsymbol{A},\, \boldsymbol{W} \odot (\boldsymbol{B}\boldsymbol{C})), \\ \min \quad & D_\varphi(\boldsymbol{A},\, \boldsymbol{W}_1 \boldsymbol{B}\boldsymbol{C}\boldsymbol{W}_2), \end{aligned}$$

and their corresponding asymmetric versions. The first of these problems is solved easily using the KKT techniques described above. The second version of the problem is solved likewise. For reference, we provide the updates below, where we use $\boldsymbol{Z}$ as a shorthand for $\boldsymbol{W}_1\boldsymbol{B}\boldsymbol{C}\boldsymbol{W}_2$,

$$\begin{aligned} \boldsymbol{B} \leftarrow \boldsymbol{B} \odot \frac{\boldsymbol{W}_1^T\big(\zeta(\boldsymbol{Z}) \odot \boldsymbol{A}\big)\boldsymbol{W}_2^T\boldsymbol{C}^T}{\boldsymbol{W}_1^T\big(\zeta(\boldsymbol{Z}) \odot (\boldsymbol{Z})\big)\boldsymbol{W}_2^T\boldsymbol{C}^T} \\ \boldsymbol{C} \leftarrow \boldsymbol{C} \odot \frac{\boldsymbol{B}^T\boldsymbol{W}_1^T\big(\zeta(\boldsymbol{Z}) \odot \boldsymbol{A}\big)\boldsymbol{W}_2^T}{\boldsymbol{B}^T\boldsymbol{W}_1^T\big(\zeta(\boldsymbol{Z}) \odot (\boldsymbol{Z})\big)\boldsymbol{W}_2^T}. \end{aligned}$$

The weighting matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are assumed to be nonnegative. Further study is needed to determine what restrictions are necessary on these matrices to ensure monotonic convergence of the update relations.

Weighted Bregman Divergences. If we seek to minimize a weighted sum of Bregman divergences, that is yet another type of weighting. Here we aim to

$$\underset{\boldsymbol{B},\boldsymbol{C}\geq 0}{\text{Minimize}}\; D_\varphi(\boldsymbol{A},\, \boldsymbol{B}\boldsymbol{C}) = \sum_{ij} w_{ij} D_\varphi(a_{ij},\, (\boldsymbol{B}\boldsymbol{C})_{ij}), \tag{3.14}$$

where $\boldsymbol{W} = [w_{ij}]$ is a weight matrix.

The problem (3.14) may be solved in a manner analogous to the solutions in the previous section. After computing appropriate gradients, and using the KKT conditions, the following update relations give solutions to problem (3.14),

$$\begin{aligned} \boldsymbol{B} \leftarrow & \boldsymbol{B} \odot \frac{\big(\zeta(\boldsymbol{B}\boldsymbol{C}) \odot \boldsymbol{A} \odot \boldsymbol{W}\big)\boldsymbol{C}^T}{\big(\zeta(\boldsymbol{B}\boldsymbol{C}) \odot \boldsymbol{W} \odot (\boldsymbol{B}\boldsymbol{C})\big)\boldsymbol{C}^T} \\ \boldsymbol{C} \leftarrow & \boldsymbol{C} \odot \frac{\boldsymbol{B}^T\big(\zeta(\boldsymbol{B}\boldsymbol{C}) \odot \boldsymbol{W} \odot \boldsymbol{A}\big)}{\boldsymbol{B}^T\big(\zeta(\boldsymbol{B}\boldsymbol{C}) \odot \boldsymbol{W} \odot \boldsymbol{B}\boldsymbol{C}\big)}. \end{aligned}$$

We remark that these solutions appear to be the same as (3.11) and (3.12) except for the addition of a $\boldsymbol{W}$ term in both the numerator and the denonimator. These solutions are also closely tied to the solution of the elementwise weighted Bregman divergence NNMA problems.

3.5 Convergence

In this section we study convergence for only the updates derived in Section 3.3. We have verified (by implementing) that the updates derived in Section 3.3 converge empirically (see Figure 1) for a large number of divergence measures (including squared Euclidean and Generalized KL-Divergence). However, we have not yet formalized a unified convergence proof for the general case. As an illustration, we offer here proofs of two important special cases. These proofs do not make use of auxiliary functions, and are the first known direct proofs of convergence of the NNMA algorithms for the Frobenius norm and KL-Divergence based problems (proofs for these cases were furnished by [20] in their paper).

We wish to show that the updates (3.11) and (3.12) are non-increasing for their corresponding objective functions. For simplicity, we first assume that $h(\boldsymbol{X}) \equiv \boldsymbol{X}$. Once again, let $F(\boldsymbol{c})$ denote the objective function value contributed by column $\boldsymbol{c}$ of $\boldsymbol{C}$. That is,

$$F(\boldsymbol{c}) = D_\varphi(\boldsymbol{a},\, \boldsymbol{B}\boldsymbol{c}),$$

for a given column $\boldsymbol{c}$ of $\boldsymbol{C}$, and the corresponding column $\boldsymbol{a}$ of $\boldsymbol{A}$. Let $\boldsymbol{d}$ denote the updated value of $\boldsymbol{c}$ resulting from (3.12). To prove monotonicity we should prove $F(\boldsymbol{c}) \geq F(\boldsymbol{d})$.

Initial approach. We use the strict convexity of φ to arrive at a condition that can prove to be crucial for monotonicity of many of our NNMA problems. The change in objective function value is,

$$\begin{aligned}F(d_i)-F(c_i) &= \varphi((\boldsymbol{Bd})_i) - \varphi((\boldsymbol{Bc})_i) - \psi((\boldsymbol{Bc})_i)(a_i - (\boldsymbol{Bc})_i) + \psi((\boldsymbol{Bd})_i)(a_i - (\boldsymbol{Bd})_i)\\ &\geq\ ((\boldsymbol{Bd})_i - (\boldsymbol{Bc})_i)\psi((\boldsymbol{Bc})_i) - \psi((\boldsymbol{Bc})_i)(a_i - (\boldsymbol{Bc})_i) + \psi((\boldsymbol{Bd})_i)(a_i - (\boldsymbol{Bd})_i)\\ &\geq\ (a_i - (\boldsymbol{Bd})_i)(\psi((\boldsymbol{Bd})_i) - \psi((\boldsymbol{Bc})_i)),\end{aligned}$$

where for the first inequality we utilized the fact that

$$\varphi(x) \geq \varphi(y) + (x-y)\psi(y).$$

Since $F(\boldsymbol{c}) = \sum_i F(c_i)$, we conclude[1]

$$F(\boldsymbol{c}) - F(\boldsymbol{d}) \geq (\boldsymbol{a} - \boldsymbol{Bd})^T\big(\psi(\boldsymbol{Bd}) - \psi(\boldsymbol{Bc})\big).$$

If we now show the latter quantity above to be nonnegative for $\boldsymbol{d}$ given by update (3.12), we will have our monotonicity proof.

We illustrate these ideas by providing new proofs of convergence for the the Frobenius norm and the KL-Divergence NNMA problems.

3.5.1 Frobenius norm NNMA

Let $\Delta(\boldsymbol{d}) = (\boldsymbol{a} - \boldsymbol{Bd})^T\big(\psi(\boldsymbol{Bd}) - \psi(\boldsymbol{Bc})\big)$. For the Frobenius norm NNMA problem, we have

$$d_i = c_i\frac{(\boldsymbol{B}^T\boldsymbol{a})_i}{(\boldsymbol{B}^T\boldsymbol{Bc})_i}. \tag{3.15}$$

We can write $\Delta(\boldsymbol{d}) = \boldsymbol{d}^T\boldsymbol{B}^T\boldsymbol{a} - \boldsymbol{d}^T\boldsymbol{B}^t\boldsymbol{Bd}$. To prove that $\Delta(\boldsymbol{d}) \geq 0$, we show that $\min_{\boldsymbol{d}} \Delta(\boldsymbol{d}) = 0$.

From (3.15) we know that $(\boldsymbol{B}^T\boldsymbol{a})_i = d_i(\boldsymbol{B}^T\boldsymbol{Bc})_i/c_i$, thus

$$\Delta(\boldsymbol{d}) = \sum_i \frac{d_i^2}{c_i}(\boldsymbol{B}^T\boldsymbol{Bc})_i - d_i(\boldsymbol{B}^T\boldsymbol{Bd})_i.$$

The first derivative of Δ w.r.t. d_p is

$$\frac{\partial\Delta}{\partial d_p} = 2\frac{d_p}{c_p}(\boldsymbol{B}^T\boldsymbol{Bc})_p - 2(\boldsymbol{B}^T\boldsymbol{Bd})_p,$$

which tells us that an optimum $\boldsymbol{d}$ must satisfy

$$(\boldsymbol{B}^T\boldsymbol{a})_p = (\boldsymbol{B}^T\boldsymbol{Bd})_p.$$

Therefore, for such a $\boldsymbol{d}$ we see that $\Delta(\boldsymbol{d}) = 0$. It now remains to verify that the Hessian of Δ is positive semi-definite.

A quick calculation reveals that Hessian of Δ is given by (δ_{ij} is the Kronecker delta function),

$$H_{ij} = \delta_{ij}\frac{(\boldsymbol{B}^T\boldsymbol{Bc})_i}{c_i} - (\boldsymbol{B}^T\boldsymbol{B})_{ij}. \tag{3.16}$$

Lemma 3.4 (Hessian is SPD)**.** *The Hessian matrix given by* (3.16) *is positive semidefinite*[2].

Proof. See Lee and Seung [20]. □

[1] In fact $F(\boldsymbol{c}) - F(\boldsymbol{d}) = (\boldsymbol{a} - \boldsymbol{Bd})^T\big(\psi(\boldsymbol{Bd}) - \psi(\boldsymbol{Bc})\big) + D_\varphi(\boldsymbol{Bd}, \boldsymbol{Bc})$, whereby, we might be better off showing that $\Delta(\boldsymbol{d}) + D_\varphi(\boldsymbol{Bd}, \boldsymbol{Bc}) \geq 0$.

[2] In their paper Lee and Seung [20] also hinge their proof upon the positive semi-definiteness of matrix $\boldsymbol{H}$, however their arrived at this matrix by another approach.

3.5.2 KL-Divergence NNMA

Once again we define $\Delta(\boldsymbol{d}) = (\boldsymbol{a} - \boldsymbol{B}\boldsymbol{d})^T(\psi(\boldsymbol{B}\boldsymbol{d}) - \psi(\boldsymbol{B}\boldsymbol{c})$. For the KL-Divergence problem, $\psi(x) = 1 + \log x$. Thus we need to prove that

$$0 \leq \Delta(\boldsymbol{d}) = \sum_i (a_i - (\boldsymbol{B}\boldsymbol{d})_i) \log \frac{(\boldsymbol{B}\boldsymbol{d})_i}{(\boldsymbol{B}\boldsymbol{c})_i}.$$

We proceed by analyzing individual terms in the summation above. We have

$$\begin{aligned}\log \frac{(\boldsymbol{B}\boldsymbol{d})_i}{(\boldsymbol{B}\boldsymbol{c})_i} &= \frac{1}{(\boldsymbol{B}\boldsymbol{c})_i}\Big((\boldsymbol{B}\boldsymbol{c})_i \log \frac{(\boldsymbol{B}\boldsymbol{d})_i}{(\boldsymbol{B}\boldsymbol{c})_i}\Big) \\ &= \frac{1}{(\boldsymbol{B}\boldsymbol{c})_i}\Big(\sum_j b_{ij}c_j \log \frac{\sum_l b_{il}d_l}{\sum_l b_{il}c_l}\Big) \\ &\geq \frac{1}{(\boldsymbol{B}\boldsymbol{c})_i} \sum_j b_{ij}c_j \log \frac{d_j}{c_j}, \end{aligned} \tag{3.17}$$

where the latter inequality follows from the log-sum inequality that says $\sum_i x_i \log \frac{x_i}{y_i} \geq (\sum_i x_i) \log \frac{\sum_i x_i}{\sum_i y_i}$. A second application of this log-sum inequality allows us to conclude that

$$-(\boldsymbol{B}\boldsymbol{d})_i \log \frac{(\boldsymbol{B}\boldsymbol{d})_i}{(\boldsymbol{B}\boldsymbol{c})_i} \geq -\sum_j b_{ij}d_j \log \frac{d_j}{c_j}. \tag{3.18}$$

Using (3.17) and (3.18) we conclude that

$$\begin{aligned}\Delta(\boldsymbol{d}) &\geq \sum_{ij} b_{ij} \log \frac{d_j}{c_j}\Big(\frac{a_i}{(\boldsymbol{B}\boldsymbol{c})_i} c_j - d_j\Big) \\ &= \sum_j \log \frac{d_j}{c_j}\Big(\big(c_j \sum_i b_{ij} \frac{a_i}{(\boldsymbol{B}\boldsymbol{c})_i}\big) - \sum_i b_{ij}d_j\Big) \\ &= 0,\end{aligned}$$

where the last equality follows from the update for d_j given by

$$d_j = c_j \frac{\sum_i b_{ij}a_i/(\boldsymbol{B}\boldsymbol{c})_i}{\sum_i b_{ij}}.$$

In a similar manner we can conclude that the objective function does not increase after an update to $\boldsymbol{B}$, which concludes the proof of monotonicity.

3.5.3 Other cases

It is easy to generalize the above two proofs to weighted versions of their respective problems. Our experiments suggest that $\Delta(\boldsymbol{d})$ is not always non-negative for all choices of φ. Thus, to prove monotonicity of our algorithms in the general case we might need to derive other conditions.

4 Experiments and Discussion

We have looked at generic algorithms for minimizing Bregman divergences between the input and its approximation. One important question arises: Which Bregman divergence should one use for a given problem? Consider the following factor analytic model

$$\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C} + \boldsymbol{N},$$

where $\boldsymbol{N}$ represents some additive noise present in the measurements $\boldsymbol{A}$, and the aim is to recover $\boldsymbol{B}$ and $\boldsymbol{C}$. If we assume that the noise is distributed according to some member of the exponential family, then minimizing the corresponding Bregman divergence [1] is appropriate. For e.g., if the noise is modeled as i.i.d. Gaussian noise, then the Frobenius norm based problem is natural.

Another question is: Which version of the problem we should use, (2.2) or (2.3)? For $\varphi(x) = \frac{1}{2}x^2$, both problems coincide. For other φ, the choice between (2.2) and (2.3) can be guided by computation issues or sparsity patterns of $\boldsymbol{A}$. Clearly, further work is needed for answering this question in more detail.

Some other open problems involve looking at the class of minimization problems to which the iterative methods of Section 3.3 may be applied. For example, determining the class of functions h, for which these methods may be used to minimize $D_\varphi(\boldsymbol{A},\, h(\boldsymbol{BC}))$. Other possible methods for solving both (2.2) and (2.3), such as the use of alternating projections (AP) for NNMA, also merit a study.

Our methods for (2.2) decreased the objective function monotonically (by construction). However, we did not demonstrate such a guarantee for the updates (3.11) & (3.12). Figure 1 offers encouraging empirical evidence in favor of a monotonic behavior of these updates. It is still an open problem to formally prove this monotonic decrease. Preliminary results that yielded *new* monotonicity proofs were discussed in Section 3.5.

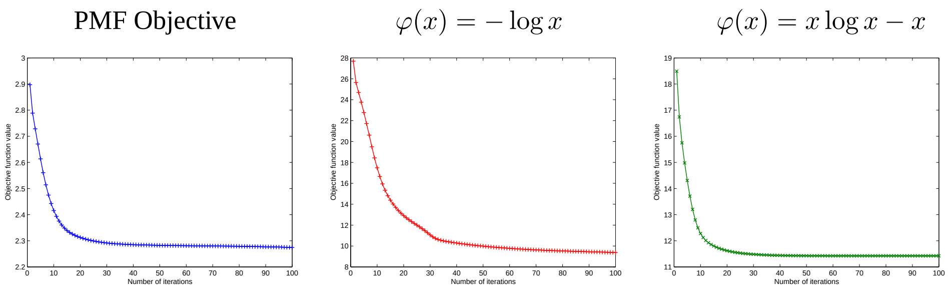

Figure 1: Objective function values over 100 iterations for different NNMA problems. The input matrix $\boldsymbol{A}$ was random 20×8 nonnegative matrix. Matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ were 20×4, 4×8, respectively.

NNMA has been used in a large number of applications, a fact that attests to its importance and appeal. We believe that special cases of our generalized problems will prove to be useful for applications in data mining and machine learning.

5 Related Work

Paatero and Tapper [27] introduced NNMA as positive matrix factorization, and they aimed to minimize $\|\boldsymbol{W} \odot (\boldsymbol{A} - \boldsymbol{BC})\|_{\mathrm{F}}$, where $\boldsymbol{W}$ was a fixed nonnegative matrix of weights. NNMA remained confined to applications in Environmetrics and Chemometrics before pioneering papers of Lee and Seung [20, 21] popularized the problem. Lee and Seung [20] provided simple and efficient algorithms for the NNMA problems that sought to minimize $\|\boldsymbol{A} - \boldsymbol{BC}\|_{\mathrm{F}}$ and $\mathrm{KL}(\boldsymbol{A}, \boldsymbol{BC})$. Lee & Seung called these problems *nonnegative matrix factorization* (NNMF), and their algorithms have inspired our generalizations.

5.1 Quick pointer to applications of NNMA

Below we provide a quick pointer to some of the literature that makes use of either Paatero's algorithms or Lee & Seung's algorithms.

Frenich et al. [11] apply OPA, PMF and ALS techniques to chromatographic data. Welling and Weber [38] generalizes the Lee/Seung NNMA algorithm from matrices to tensors. Novak and Mammone [23] apply NNMA for language modeling. Lee et al. [22] and J-H. Ahn and Choi [19] look at some Dynamic Positron Emission Tomography applications. Hoyer [17] introduces additional sparsity constraints to the NNMA problem. A collection of research articles dealing with weighting, image and face classification and some other issues has been produced by Guillamet and Vitrià [12, 13, 14], Guillamet et al. [16]. Cooper and Foote [7] applies NNMA to the task of summarizing video. Szatmáry et al. [35] extend NNMA by Sparse code shrinkage (SCS) and weight sparsification. Ramadan et al. [30] compare the methods of Paatero [25], Paatero and Tapper [27] and Paatero [26] on pollution data. Wild et al. [39] have written a brief article trying to motivate the use of NNMA along with references to means of initializing NNMA using the clustering results of spherical k-means [8]. Sajda et al. [32] apply NNMA to the recovery of constituent spectra in chemical shift imaging. An somewhat offbeat application to Polyphonic music transcription is presented by Smaragdis and Brown [33]. Donoho and Stodden [9] mull over criteria that enable one to determine when does NNMA give a correct decomposition into parts for the original data. An application to the discovery of hierarchical speech features appears in a paper by Behnke [2]. Xu et al. [40] present a simple application to clustering text data. Szatmáry et al. [36] look at robust hierarchical image representation, augmenting NNMA with SCS preprocessing. Hoyer [18] uses NNMA to model receptive fields of the visual cortex. Further related work in nonnegative Independent Components Analysis (ICA) and nonnegative Principal Components Analysis (PCA) has also been conducted [24, 28, 29].

Srebro and Jaakola [34] discuss elementwise weighted low-rank approximations without any nonnegativity constraints. Collins et al. [6] discuss algorithms for obtaining a low rank approximation of the form $\boldsymbol{A} \approx \boldsymbol{BC}$, where the loss functions are Bregman divergences, however, there is no restriction on $\boldsymbol{B}$ and $\boldsymbol{C}$. Our methods are tailored to nonnegative data, and they offer the advantages of computational efficiency and ease of implementation.

Acknowledgments

This research was supported by NSF grant CCF-0431257, NSF Career Award ACI-0093404, and NSF-ITR award IIS-0325116.

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman Divergences. In *SIAM International Conf. on Data Mining*, Lake Buena Vista, Florida, April 2004. SIAM.

[2] S. Behnke. Discovering hierarchical speech features using convolutional nonnegative matrix factorization. In *International Joint Conference on Neural Networks*, volume 4, pages 2758–2763, Portland, OR, 2003.

[3] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Numerical Mathematics and Scientific Computation. Oxford University Press, 1997.

[4] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum Sum Squared Residue based Co-clustering of Gene Expression data. In *Proc. 4th SIAM International Conference on Data Mining (SDM)*, pages 114–125, Florida, 2004. SIAM.

[5] M. Collins, R. Schapire, and Y. Singer. Logistic regression, adaBoost, and Bregman distances. In *Thirteenth annual conference on COLT*, 2000.

[6] M. Collins, S. Dasgupta, and R. E. Schapire. A Generalization of Principal Components Analysis to the Exponential Family. In *NIPS 2001*, 2001.

[7] M. Cooper and J. Foote. Summarizing video using nonnegative similarity matrix factorization. In *IEEE Multimedia Signal Processing Workshop*, St. Thomas, USVI, December 2002.

[8] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, January 2001.

[9] D. Donoho and V. Stodden. When does nonnegative matrix factorization give a correct decomposition into parts? In *Neural Information Processing Systems*, 2003.

[10] T. Feng, S. Z. Li, H-Y. Shum, and H. Zhang. Local nonnegative matrix factorization as a visual representation. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 178–193, Cambridge, MA, June 2002.

[11] A. G. Frenich, M. M. Galera, J. L. M. Vidal, D. L. Massart, J.R. Torres-Lapasió, K. De Braekeleer, J-H. Wang, and P. K. Hopke. Resolution of multicomponent peaks by orthogonal projection approach, positive matrix factorization and alternating least squares. *Analytica Chimica Acta*, 411:145–155, 2000.

[12] D. Guillamet and J. Vitrià. Determining a suitable metric when using nonnegative matrix factorization. In *16th International Conference on Pattern Recognition*. IEEE Computer Society, 2002.

[13] D. Guillamet and J. Vitrià. Classifying faces with nonnegative matrix faces. In *CCIA*, Castelló de la Plana, Spain, 2002.

[14] D. Guillamet and J. Vitrià. Analyzing non-negative matrix factorization for image classification. In *IEEE International Conference on Pattern Recognition*, volume 2, pages 116–119, 2002.

[15] D. Guillamet, M. Bressan, and J. Vitrià. A weighted nonnegative matrix factorization for local representations. In *CVPR*. IEEE, 2001.

[16] D. Guillamet, J. Vitrià, and B. Schiele. Introducing a weighted nonnegative matrix factorization for image classification. *Pattern Recognition Letters*, 24(14):2447–2454, October 2003. ISSN 0167-8655.

[17] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

[18] P. O. Hoyer. Modeling receptive fields with nonnegative sparse coding. *Neurocomputing*, 52–54:547–552, 2003.

[19] J-H. Oh J-H. Ahn, S. Kim and S. Choi. Multiple nonnegative-matrix factorization of dynamic pet images. In *ACCV*, 2004.

[20] D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562, 2000.

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

[22] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee. Application of nonnegative matrix factorization to dynamic positron emission tomography. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, CA, December 2001.

[23] M. Novak and R. Mammone. Use of nonnegative matrix factorization for language model adaptation in a lecture transcription task. In *Proceedings of the 2001 IEEE Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 541–544, Salt Lake City, UT, May 2001.

[24] E. Oja and M. Plumbley. Blind separation of positive sources using nonnegative PCA. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 2003.

[25] P. Paatero. Least-squares formulation of robust nonnegative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37:23–35, 1997.

[26] P. Paatero. The multilinear engine—a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, 8(4):854–888, December 1999.

[27] P. Paatero and U. Tapper. Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(111–126), 1994.

[28] M. D. Plumbley. Conditions for nonnegative independent component analysis. *IEEE Signal Processing letters*, 9 (6):177–180, June 2002.

[29] M. D. Plumbley. Algorithms for nonnegative independent component analysis. In *Unpublished*, 2002.

[30] Z. Ramadan, B. Eickhout, X-H. Song, L. M. C. Buydents, and P. K. Hopke. Comparison of positive matrix factorization and multilinear engine for the source apportionment of particulate pollutants. *Chemometrics and Intelligent Laboratory Systems*, 66:15–28, 2003.

[31] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[32] P. Sajda, S. Du, T. Brown, L. Parra, and R. Stoyanova. Recovery of Constituent Spectra in 3D Chemical Shift Imaging using Nonnegative Matrix Factorization. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 71–76, Nara, Japan, April 2003.

[33] P. Smaragdis and J. C. Brown. Nonnegative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, New Paltz, NY, October 2003.

[34] N. Srebro and T. Jaakola. Weighted low-rank approximations. In *Proc. of twentieth ICML*, 2003.

[35] B. Szatmáry, B. Póczos, J. Eggert, E. Körner, and A. Lőrincz. Nonnegative matrix factorization extended by sparse code shrinkage and weight sparsification algorithms. In *ECAI 2002, Proceedings of the 15th European Conference on Artificial Intelligence*, pages 503–507, Amsterdam, 2002. IOS Press.

[36] B. Szatmáry, G. Szirtes, A. Lőrincz, J. Eggert, and E. Körner. Robust hierarchical image representation using nonnegative matrix factorization with sparse code shrinkage preprocessing. *Pattern Analysis and Applications*, 2003. Accepted.

[37] J. A. Tropp. *Topics in Sparse Approximation*. PhD thesis, The University of Texas at Austin, 2004.

[38] M. Welling and M. Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22:1255–1261, 2001.

[39] S. Wild, J. Curry, and A. Dougherty. Motivating nonnegative matrix factorizations. SIAM Linear Algebra Meeting, July 2003.

[40] W. Xu, X. Liu, and Y. Gong. Document clustering based on nonnegative matrix factorization. In *SIGIR'03*, pages 267–273, Toronto, 2003. ACM.