J. Korean Soc. Ind. Appl. Math. Vol.25, No.3, 107–116, 2021 https://doi.org/10.12941/jksiam.2021.25.107

UNDERSTANDING NON-NEGATIVE MATRIX FACTORIZATION IN THE FRAMEWORK OF BREGMAN DIVERGENCE

KYUNGSUP KIM[1,†]

[1]DEPARTMENT OF COMPUTER ENGINEERING, CHUNGNAM NATIONAL UNIVERSITY, SOUTH KOREA
E-mail address: †sclkim@cnu.ac.kr

ABSTRACT. We introduce optimization algorithms using Bregman Divergence for solving non-negative matrix factorization (NMF) problems. Bregman divergence is known a generalization of some divergences such as Frobenius norm and KL divergence and etc. Some algorithms can be applicable to not only NMF with Frobenius norm but also NMF with more general Bregman divergence. Matrix Factorization is a popular non-convex optimization problem, for which alternating minimization schemes are mostly used. We develop the Bregman proximal gradient method applicable for all NMF formulated in any Bregman divergences. In the derivation of NMF algorithm for Bregman divergence, we need to use majorization/minimization(MM) for a proper auxiliary function. We present algorithmic aspects of NMF for Bregman divergence by using MM of auxiliary function.

1. INTRODUCTION

Non-negative matrix factorisation (NMF) has popular application for machine learning, computer vision, Bio-informatics and many others. Matrix Factorization is a popular non-convex optimization problem, for which alternating minimization schemes are mostly used. We review and compare various optimization algorithms for non-negative matrix factorization. The computation of NMF remains challenging and expensive due to the constraints.The most frequently used techniques for solving matrix factorization problems involve alternating updates similar to GaussSeidel type methods.

We consider a wide family of iterative algorithms for non-negative matrix factorization (NMF) and related problems, subject to additional constraints such as sparsity and/or smoothness. We consider a wide class of cost functions or divergences leading to generalized multiplicative algorithms with regularization and/or penalty terms.

Some algorithms can be applicable to not only NMF with Frobenius norm but also NMF with more general Bregman divergence. We develop one united algorithm applicable for all NMF formulated in any Bregman divergences. Bregman divergence is known a generalization

Received April 30 2021; Revised September 23 2021; Accepted in revised form September 23 2021; Published online September 25 2021.

Key words and phrases. Non-negative matrix factorisation (NMF), Bregman Distance, Auxiliary function, majorization/minimization(MM), Bregman proximal gradient, Block coordinate descent(BCD).

† Corresponding author.

of some divergences. The range of cost functions in Bregman divergence includes large number of generalized divergences, such as the squared weighted Euclidean distance, relative entropy, Kullback-Leibler I-divergence, α- and β-divergences and Csiszar f-divergence [4, 6, 7, 9, 10, 19, 20]. Various Bregman divergences, such as Frobenius norm and KL divergence, have been used in a wide range of applications, including text clustering, signal processing, image processing, and music analysis[9, 20]. NMF with the β-divergence has been widely used in music signal processing in particular, for transcription and source separation[8, 9, 20].

Bregman divergences have been used to derive an exact characterization of the difference between the two sides of Jensens inequality. In the derivation of NMF algorithm for Bregman divergence, we need to use surrogate majorization/minimization(MM) concept. Surrogate maximization (or minimization) (SM) algorithms are a family of algorithms that can be regarded as a generalization of expectation-maximization (EM) algorithms [22].

We discuss one united framework applicable for all NMF using any Bregman divergences. We extend our methods in the framework of Bregman divergence so that they are more general hence admit potentially more applications. We develop the inertial version of the Bregman proximal gradient method applicable for all NMF formulated in any Bregman divergences. In the derivation of NMF algorithm for Bregman divergence, we need to know how to choose a proper auxiliary (surrogate) function for MM step. We will present and review algorithmic aspects of NMF for Bregman divergence by using MM algorithm of auxiliary (surrogate) function.

2. Related Works and problem definition

In NMF, a data matrix V with non-negative entries of dimension $M \times N$ is factorized into a matrix W of dimension $M \times K$ and H of dimension $K \times N$ such that WH approximately equals V . The entries in W and H must be non-negative, and we assume K is small relative to M and N (i.e., $V \approx WH$). The various divergences can be used as cost function for NMF.

In order to measure the discrepancy between the input data and the low rank approximation, the Kullback-Leibler (KL) divergence is one of the most widely used objective function for NMF [11]. The quality of the approximation is measured using an objective function, which typically has the form

$$D(V|WH) = \sum_{i=1} \sum_{j=1} d(V_{ij}|(WH)_{ij}), \tag{2.1}$$

where $d(x|y)$ is a scalar cost function and the entries in W and H must be non-negative. We will define the cost function in terms of Bregman divergence.

2.1. **Bregman divergence.** Bregman divergences are a general class of distortion functions, which include squared Euclidean distance, KL-divergence, Itakura-Saito distance, etc., as special cases [6, 7, 9, 11]. Bregman divergences may be considered a generalization of squared Euclidean distance because of many shared properties. Bregman divergences are not symmetric, and do not satisfy the triangle inequality.

What we intended by cost function is a positive-valued function with a single minimum. A popular cost function is the β divergence [4, 9]. Divergences are considered as (dis)similarity measures. The divergence has the following properties: $d(x|y) \geq 0$ for all $x, y \geq 0$ and $d(x|y) = 0$ if and only if $x = y$. However, a divergence does not necessarily satisfy the triangle inequality and the symmetry axiom of the distance or metric definition [4, 10]. For $\beta \in R - \{0, 1\}$, the β- divergence $d_\beta(x|y)$ is defined by

$$d_\beta(x|y) = \frac{x^\beta}{\beta(\beta-1)} + \frac{y^\beta}{\beta} - \frac{xy^{\beta-1}}{\beta-1}$$

This definition is extended to $\beta \in \{0, 1\}$ in the obvious way, by taking limits. The three divergence functions most commonly used with NMF are special cases of the β-divergence:

(1) $\beta = 2$ (Euclidean): $d(x|y) = 1/2(x-y)^2$
(2) $\beta = 1$ (Kullback-Leibler): $d(x|y) = x \log \frac{x}{y} - x + y$
(3) $\beta = 0$ (Itakura-Saito): $d(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1$.

In many applications, such as image analysis, pattern recognition and statistical machine learning we use the information-theoretic divergences rather than Euclidean squared or l_p-norm distances [4]. The Bregman divergence encompasses the β-divergence in a natural way by defining the kernel function Φ [10]. The kernel function $\Phi : \mathcal{S} \to R$ is a continuously differentiable strictly convex function where $\mathcal{S}$ be a convex subset of a Hilbert space. The Bregman divergence $D : \mathcal{S} \times \mathcal{S} \to R_+$ (where R_+ is the set of non-negative real numbers) is defined by

$$D_\phi(x|y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y) \rangle,$$

where $\nabla\phi(y)$ stands for the gradient of ϕ evaluated at y and $\langle \cdot, \cdot \rangle$ is the standard Hermitian dot product.

Element-wise Bregman divergences are a subclass of Bregman divergences for which Φ is the sum of N scalar, continuously differentiable and strictly convex element-wise functions: $\Phi(\mathbf{x}) = \sum_i \phi(x_i)$ for $\mathbf{x} = (x_1, \cdots, x_n)$. Then we have

$$D_\Phi(x|y) = \sum_{i=1} d_\phi(x_i|y_i).$$

where $d_\phi(x|y) = \phi(x) - \phi(y) - \phi'(y)(x-y)$ and thus, the divergence is element-wise. When ϕ is strictly convex and differentiable. A differentiable function of one variable is convex on an interval if and only if its graph lies above all of its tangents. It is easy to show that $d_\phi(x,y) \geq 0$ and $d_\phi(x,y) = 0$ if and only if $x = y$.

We show that the Bregman divergence encompasses the β-divergence and Amari's alpha divergences in a natural way by defining the the kernel function[4, 5, 6, 7]. Let the kernel function $\phi : R_+ - \{0\} \to R$ be the function defined as:

$$\phi_\beta(x) = \begin{cases} -\log x + x - 1, & \beta = 0 \\ x \log x - x + 1, & \beta = 1 \\ \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta}, & \text{otherwise.} \end{cases}$$

If we rewrite it as $\phi(x) = \frac{x(x^{\beta-1}-1)}{\beta(\beta-1)} + \frac{(1-x)}{\beta}$, then we have a family of Amari's alpha divergences ($\beta = (1+\alpha)/2$).

2.2. **Block coordinate descent(BCD).**

Block coordinate descent. Block coordinate descent (BCD) (more precisely, block coordinate update) is very general and widely used for solving both convex and nonconvex problems with multiple blocks of variables. We consider the optimization problem

$$\min_{x\in\mathcal{X}} F(x_1,\cdots,x_s) \equiv f(x_1,\cdots,x_s) + \sum_{i=1}^{s} r_i(x_i)$$

where variable x is decomposed into s blocks $x_1,\cdots,x_s$, the set $\mathcal{X}$ of feasible points is assumed to be a closed and block multiconvex subset. f is assumed to be a differentiable and block multiconvex function. We call a function f block multiconvex if, for each i, f is a convex function of x_i while all the other blocks are fixed. The block coordinate descent (BCD) method of GaussSeidel type minimizes F cyclically over each of $x_1,\cdots,x_s$.

The projected coordinate descent algorithm is an iterative method for optimization problems. In each iteration, one coordinate is updated, while the other coordinates remain fixed such as

$$\begin{aligned} h_k^{t+1} &= \arg\min_{h_k\geq 0} f(h_1^t,\cdots,h_k,\cdots,h_K^t) \\ h_l^{t+1} &= h_l^t \quad \text{for } l \neq k. \end{aligned}$$

Typically, there are three main types of BCD methods: classical BCD, proximal BCD and proximal gradient BCD [13]. The classical BCD methods alternatively minimize the block i functions of the objective. The proximal BCD methods improve the classical BCD methods by coupling the block i objective functions with a proximal term. The proximal gradient BCD methods minimize a standard proximal linearization of the objective function. These BCD methods belong to a more general framework, named the block successive upper-bound minimization algorithm. It is closely related to the majorization/minimization(MM) algorithm [13].

Inertial Method. Incorporating inertial force is a popular and efficient method to accelerate the convergence of first-order methods. The inertial method adds to the new direction a momentum term equal to the difference of the two previous iterates; this is also known as extrapolation.

3. Optimization algorithm in the framework of Bregman divergence

We extend our methods in the framework of Bregman divergence so that they are more general hence admit potentially more applications.

3.1. **Proximal distance algorithms.**

Proximal distance algorithms combine the classical penalty method of constrained minimization with distance majorization[14]. For convex problems, proximal distance algorithms reduce to proximal gradient algorithms and therefore enjoy well understood convergence properties.

The inertial versions for the proximal and proximal gradient BCD methods in the framework of Bregman divergence have been introduced in [1, 13]. The proximal method minimizes sums of differentiable and non-differentiable convex functions $f = g + h$ such as

$$\min_{x \in R^n} \{h(x) + g(x)\} \tag{3.1}$$

where g is differentiable and convex, and h is closed and convex [12, 13]. Here nondifferentiable term h can be simple but has inexpensive prox-operator.

Let $Prox_t$ be the following proximal map

$$prox_t(x) = \arg\min_{z \in R^n} \{h(z) + \frac{1}{2t}\|z - x\|^2\}.$$

We consider the use of Bregman distances in constrained optimization through the proximal minimization method. We generalize the proximal minimization algorithm by replacing the quadratic term by Bregman divergence.

For a given $v \in E_i$, and a positive number t, the Bregman proximal map of a function ϕ is defined by

$$prox_t(v) = \arg\min_{u \in R^n} \{h(u) + \frac{1}{2t} D_\phi(u|v)\}.$$

Let us consider how to construct proximal point algorithms with Bregman functions. At the kth step of a proximal minimization algorithm (PMA), we minimize the function

$$G_k(x) = h(x) + \frac{1}{2t_k}\phi(x, x^{k-1}).$$

We review the well-know proximal gradient method and its variant, namely, the Bregman proximal gradient method. The Bregman proximal mapping is defined by replacing the Euclidean distance with the Bregman distance. Now we consider the proximal gradient problem such as

$$x_{k+1} \;=\; \arg\min_{x \in C} \left\{ \langle x, \nabla f(x_k)\rangle + \frac{1}{\alpha_k}\frac{\|x - x_k\|_2^2}{2} \right\}.$$

One employs a majorize-minimize approach. By definition of the Euclidean projection, this can also be written as

$$x_{k+1} \;=\; \arg\min_{x \in C} \; \|(x - x_k) + \alpha_k \nabla f(x_k)\|_2^2.$$

Let $L = 1/\alpha_k$. If we still wish to apply a gradient type method to minimize such a function, L-smoothness can sometimes be forced upon f. This is sufficient in principle as the theory for constrained first order methods only requires the gradients to be L-Lipschitz. On the projected gradient descent (PGD), we showed that a way to solve the constrained minimization problem with a differentiable f is to follow the iteration

$$x_{k+1} \;=\; \pi_C(x_k - \alpha_k \nabla f(x_k)),$$

where π_C denotes the Euclidean projection onto the constrain domain C.

One reason why one might want to consider another distance-like function to penalize how much we move in a particular direction is that doing so can better reflect what we may know about the geometry of C which can make steps easier to compute or convergence to a minimizer faster. In order to broaden the applicability of this method, we consider a class of Bregman proximal gradient methods. This method adopts the more general Bregman regularization instead of the Euclidean distance. The inertial version of the Bregman proximal gradient method for relative-smooth nonconvex optimization was studied [21]. Proximal gradient BCD methods update using a linearization of f. The Bregman proximal gradient map of a pair of functions (g is continuously differentiable) is defined by

$$Gprox(u_1; \nabla g(u_2)) = \arg\min_u h(u) + \langle \nabla g(u_2), u \rangle + \frac{1}{\beta} D_\phi(u, u_1)$$

For convex h and Bregman kernel ϕ, Bregman proximal mapping is defined by

$$\begin{aligned} prox_h(y, a) &= \arg\min_x \big(h(x) + a^T x + d_\phi(x, y)\big) \\ &= \arg\min_x \big(h(x) + (a - \nabla\phi(y))^T x + \phi(x)\big) \\ &= prox_h(x - a) \end{aligned}$$

where first argument y must be in $int(dom(\phi))$ and second argument a can take any value. Well use this only if for every y and a, a unique minimizer $x \in int(dom(\phi))$ exists.

Let us now consider a μ-strongly convex and differentiable function ϕ and the associated Bregman divergence D_ϕ. If we use this as divergence, the GPGD iteration is

$$x_{k+1} \quad \in \quad \arg\min_{x \in C} \left\{ \langle x, \nabla f(x_k) \rangle + \frac{1}{\alpha_k} D_\phi(x, x_k) \right\} \tag{3.2}$$

for $\alpha_k > 0$.

3.2. **Majorization/minimization(MM) approach for Bregman divergence.** The principle of iteratively minimizing a majorizing surrogate of an objective function is often called majorization/minimization (MM)[18, 22]. We note that the proximal gradient method can be interpreted as an example of MM algorithms which includes the gradient method, Newtons method, and the EM algorithm. The EM algorithm from statistics is a special case [22]. Surrogate maximization (or minimization) (SM) algorithms are a family of algorithms that can be regarded as a generalization of expectation-maximization (EM) algorithms [22].

We call any algorithm based on this iterative method an MM algorithm. The MM principle is not an algorithm, but a prescription or principle for constructing optimization algorithms. The MM framework is a two-step approach:

(1) Majorization: Find a majorizer, that is, a function that is equal to the objective function at the current iterate while being larger everywhere else on the feasible domain, and
(2) Minimization: minimize the majorizer to obtain the next iterate.

Let us consider the majorization-minimization scheme and its variants. This procedure relies on the concept of an auxiliary (surrogate) functions A function $g(x, \tilde{x})$ is said to be an auxiliary (surrogate) function) for $f(x)$ if the following two conditions are satisfied:

$$g(x, x) = f(x), \tag{3.3}$$

$$g(x, \tilde{x}) \geq f(x) \tag{3.4}$$

for all $\tilde{x}$. If g is an auxiliary function, then f is non-increasing under the update

$$x_{t+1} = \arg\min_{h} g(x, x_t).$$

We can directly prove that

$$f(x_{t+1}) \leq g(x_{t+1}, x_t) \leq g(x_t, x_t) = f(x_t). \tag{3.5}$$

Construction of the auxiliary function is key to the algorithms in turning an otherwise intractable optimization problem into a tractable one. A good auxiliary function should preferably have a closed-form solution. The closer is the auxiliary function to f, the more efficient is the algorithm.

If x_t is a local minimum of $g(x, x_t)$, then $f(x_{t+1}) = f(x_t)$. We can obtain a sequence of estimates that converge to a local minimum $x_{\min} = \arg\min_x f(x)$. By defining the appropriate auxiliary functions $g(x, x_t)$, the update rules follow from (3.5). We note that the derivation of various proximal distance algorithms introduced in subsection 3.1 can be interpreted as the majorization/minimization (MM) by defining proper auxiliary functions [18]. When f is twice differentiable, a quadratic auxiliary function is defined by

$$g(x, u) = f(u) + (x-u)^T \nabla f(u) + \frac{1}{2}(x-u)^T H (x-u).$$

where H is a positive semi-definite Hessian. It has the auxiliary properties. Such auxiliary functions appear frequently in the statistics and machine learning literature. When f is differentiable and ∇f is L-Lipschitz, f admits the following auxiliary function

$$g(x, u) = f(u) + \langle x, \nabla f(u) \rangle + L D_{\phi}(x, u). \tag{3.6}$$

By finding optimal solution minimizing $g(x, u)$ with respect to x, we can obtain the GPGD iteration (3.2).

4. NMF METHODS USING BREGMAN DIVERGENCE

In this section, we consider a wider class of NMF problems under Bregman divergence. Various algorithms for nonnegative matrix factorization (NMF) with variant Bregman-divergence have been introduced [1, 5, 9, 11]. Some algorithms can be applicable to not only NMF with Frobenius norm but also NMF with more general Bregman divergence. They used Taylor series expansion to derive the element-wise problem [16]. Column-wise update algorithms for solving Bregman divergence NMF was introduced by [15]. We will propose a unified method that globally converges to a stationary point of the optimization problem in Eq. (2.1) in a block mirror descent method with Bregman divergence. In this section, we present algorithmic aspects of NMF for Bregman divergence by using majorization/minimization(MM) of auxiliary (surrogate) function. We can see that the auxiliary function for an objective function is not unique. A main issue is how to choose a proper the auxiliary function for NMF with more general Bregman divergence.

We note that $D(V|WH)$ does not possess the Lipschitz smoothness property [11]. The notion of relative smoothness is a generalization of the Lipschitz smoothness [17]. We use the notion of smooth adaptable functions introduced in [3]. A pair (f, κ) is called L-smooth adaptable on $C = int\, dom(\kappa)$ if there exists $L > 0$ such that $L\kappa - f$ and $L\kappa + f$ are convex on C. Then we have $D_f(x, y) \leq LD_\kappa(x, y)$, even though f needs not be convex. Let us define the objective function $f(h) = \sum_k d(v_k|[Wh]_k)$ in Eq (3.1). For KL divergence, it is known that the KL objective function is a relative smooth function[11]. It is explained how to find a function $\kappa(x)$ to which the function $f(x)$ is L-smooth relative [2, 3, 17]. It is $\kappa(h) = \sum_{j=1} \log h_j$. Then there is a scalar L for which

$$f(x) \leq f(u) + \langle x, \nabla f(u) \rangle + LD_\kappa(x, u).$$

Then we can derive a proper auxiliary function $g(x, u)$ for Bregman proximal gradient as Eq. (3.6). The bound L can be computed by $L = \|v\|_1$.

Jensens inequality provides a natural mechanism to obtain auxiliary function for convex functions. Suppose a function $f : R \to R$ is convex . We can define an auxiliary function such as

$$g(u, w) = \sum_{i=1}^{N} \frac{c_i w_i}{c^T w} f\left(\frac{c^T w}{w_i} u_i\right)$$

for $u, w \in R^n$. We can see that it has auxiliary properties (3.3) and (3.4) by using Jensens inequality.

In order extend the optimization algorithm using auxiliary function in a block coordinate descent (BCD) framework, the auxiliary function g are separable into n components such as

$$g(u, v) = \sum_{i=1}^{n} g_i(u_i, v_i) + C$$

. for $u, v \in R^n$ where C is constant with respect to u.

We consider how to choose a proper kernel function κ generating Bregman distances to handle optimization problem when f is not convex. First, when f is concave and also differentiable on its domain, it can be linearized by first-order Taylor approximation, as

$$f(u) \leq f(v) + \nabla f(v)^T (u - v). \tag{4.1}$$

Then we construct an auxiliary function by using Eq. (4.1). Since most continuous functions can be expressed as the difference of two convex functions, we can often use this trick to construct an auxiliary function. If for any $f(u) = f_1(u) - h(u)$ where both $f_1(u)$ and $h(u)$ are convex, we can write

$$f(u) \leq f_1(u) - h(v) - \nabla h(v)^T (u - v).$$

The use of differences of convex functions is a very important strategy in convex optimization and has received much attention recently in machine learning.

We will explain how an separable auxiliary function $f(h)$ for the specific case of the β-divergence is constructed by modifying the results in [9]. We have the β- divergence $d_\beta(x|y)$ decomposed into $d_\beta(x|y) = \check{d}_\beta(x|y) + \hat{d}_\beta(x|y) + \bar{d}(x|y)$, where $\check{d}$ is convex function of y, $\hat{d}$ is a concave function of y and $\bar{d}$ is a constant of y. Let $\tilde{v} = \tilde{w}H$ and $\tilde{w}$ be such that $\tilde{v}_n \geq 0$ for all n and $\tilde{w}_k > 0$ for all k. Then the function

$$g(w|\hat{w}) = \sum_n \left\{ \sum_k \frac{\tilde{w}_k h_{kn}}{\tilde{v}_n} \check{d}(v_n|\tilde{v}_n \frac{w_k}{\tilde{w}_k}) + \hat{d}(v_n|\tilde{w}_n) + \hat{d}'(v_n|\tilde{v}_n) \sum_k (w_k - \tilde{w}_k) h_{kn} + \bar{d}(v_n) \right\}$$

is an auxiliary function to $f(h)$. The auxiliary function $g(x|\tilde{x})$ is by construction separable in functions of the individual coefficients x_k of x, which allows to decouple the optimization. The Hessian matrix $\nabla^2_{h_k} g(h|\tilde{h})$ is a diagonal matrix with positive value.

We may write

$$g(x|\tilde{x}) = \sum_k g_k(x_k|\tilde{x}) + C$$

where C is a constant with respect to h. The gradient of the auxiliary function is given by $\nabla_{x_k} g(x|\tilde{x}) = \nabla_{x_k} g_k(x|\tilde{x})$.

The derivative of the criterion $D(V|WH|)$ with respect to θ can be expressed as the difference of two nonnegative function $\nabla_\theta D(\theta) = \nabla^+_\theta D(\theta) - \nabla^-_\theta D(\theta)$. Then, a multiplicative algorithm simply writes

$$\theta \leftarrow \theta \frac{\nabla^-_\theta D(\theta)}{\nabla^+_\theta D(\theta)}$$

which ensures nonnegativity of the parameter updates, provided initialization with a nonnegative value.

5. Conclusion

We have presented and reviewed algorithmic aspects of NMF with using Bregman Divergence. We considered optimization algorithms using Bregman Divergence for solving nonnegative matrix factorization (NMF) problems. We introduced one united algorithm applicable for all NMF formulated in any Bregman divergences. We discussed the unified inertial version of the Bregman proximal gradient method applicable for all NMF formulated in any Bregman divergences. We proposed NMF algorithm for Bregman divergence by using majorization/minimization(MM) of auxiliary (surrogate) function. We need to resolve algorithmic problems for NMF. The related issues are the investigation into the presence of local minima in the cost functions, and ways to avoid them. We need to develop some efficient algorithms for the Bregman proximal gradient method applicable for all NMF and analyze the convergence and performance.

Acknowledgments

This work was supported by research fund of Chungnam National University.

References

[1] Masoud Ahookhosh, Le Thi Khanh Hien., Nicolas Gillis, and Panagiotis Patrinos. Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization. 30468160(30468160).

[2] Heinz H. Bauschke, Jerome Bolte, and Marc Teboulle. A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *INFORMS*, 79(0):21, 2007.

[3] Jerome Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[4] Andrzej Cichocki and Shun ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of Similarities. *Entropy*, 12(6):1532–1568, 2010.

[5] Andrzej Cichocki, Hyekyoung Lee, Yong-deok Kim, and Seungjin Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29:1433–1440, 2008.

[6] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi AMARI. Csiszár's Divergences for Non-negative Matrix Factorization : Family of New Algorithms. In *Independent Component Analysis and Blind Signal Separation. ICA 2006. Lecture Notes in Computer Science*. 2006.

[7] Andrzej Cichocki, Rafal Zdunek, Seungjin Choi, Robert Plemmons, and Shun-ichi Amari. Non-negative tensor factorization using alpha and beta divergences. In *In IEEE International Conference on Acoustics, Speech,*, 2007.

[8] Cédric Févotte, Nancy Bertin, and Jean Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[9] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[10] Romain Hennequin, Bertrand David, and Roland Badeau. Beta-divergence as a subclass of Bregman divergence. *IEEE Signal Processing Letters*, 18(2):83–86, 2011.

[11] Le Thi Khanh Hien and Nicolas Gillis. Algorithms for Nonnegative Matrix Factorization with the Kullback-Leibler Divergence. *Accepted in the Journal of Scientific Computing*, oct 2020.

[12] Le Thi Khanh Hien, Nicolas Gillis, and Panagiotis Patrinos. Inertial Block Proximal Methods for Non-Convex Non-Smooth Optimization. 2019.

[13] Le Thi Khanh Hien, Duy Nhat Phan, and Nicolas Gillis. An Inertial Block Majorization Minimization Framework for Nonsmooth Nonconvex Optimization. 2020.

[14] Kevin L. Keys, Hua Zhou, and Kenneth Lange. Proximal distance algorithms: Theory and practice. *Journal of Machine Learning Research*, 20:1–38, 2019.

[15] Keigo Kimura, Mineichi Kudo1, and Yuzuru Tanaka1. A column-wise update algorithm for nonnegative matrix factorization in Bregman divergence with an orthogonal constraint. *Machine Learning*, 103(2):285–306, 2016.

[16] Liangda Li, Guy Lebanon, and Haesun Park. Fast bregman divergence NMF using taylor expansion and coordinate descent. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 307–315, 2012.

[17] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[18] Julien Mairal. Optimization with first-order surrogate functions. *30th International Conference on Machine Learning, ICML 2013*, 28(PART 3):1820–1828, 2013.

[19] Macoumba Ndour, Mactar Ndaw, and Papa Ngom. Relationship between the Bregman divergence and beta-divergence and their Applications. *arXiv*, (November), 2018.

[20] Vincent Y F Tan and Ce dric Fe Votte. Automatic Relevance Determination in Nonnegative Matrix Factorization with the beta-Divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2013.

[21] Zhongming Wu, Chongshou Li, Min Li, and Andrew Lim. Inertial proximal gradient methods with Bregman regularization for a class of nonconvex optimization problems. *Journal of Global Optimization*, 79(3):617–644, 2021.

[22] Zhihua Zhang, James T. Kwok, and Dit Yan Yeung. Surrogate maximization/minimization algorithms and extensions. *Machine Learning*, 69(1):1–33, 2007.