Beta-Divergence as a Subclass of Bregman Divergence

Romain Hennequin, *Student Member, IEEE*, Bertrand David, *Member, IEEE*, and Roland Badeau, *Senior Member, IEEE*

***Abstract*—In this paper, we present a complete proof that the β-divergence is a particular case of Bregman divergence. This little-known result makes it possible to straightforwardly apply theorems about Bregman divergences to β-divergences. This is of interest for numerous applications since these divergences are widely used, for instance in non-negative matrix factorization (NMF).**

***Index Terms*—Beta-divergence, Bregman divergence, non-negative matrix factorization.**

I. Introduction

DIVERGENCES are distance-like functions, widely used to assess the similarity between two objects. For instance, Kullback–Liebler (KL) divergence [14] is used in information theory to compare two probability distributions, and the Itakura–Saito (IS) divergence is used as a measure of the perceptual difference between spectra [12]. Generalized classes of divergences, for instance Bregman divergences, are used in pattern classification and clustering [1]. In non-negative matrix factorization (NMF [15]), divergences are used as cost functions: NMF approximates an $F \times T$ non-negative matrix $\mathbf{V}$ with the product of two non-negative low-rank matrices:

$$\mathbf{V} \approx \mathbf{WH}$$

where the size of $\mathbf{W}$ is $F \times R$ and the size of $\mathbf{H}$ is $R \times T$ (with $R < F$ and $R < T$).

This approximation is generally quantized with a cost function to be minimized with respect to $\mathbf{W}$ and $\mathbf{H}$. This cost function is often an element-wise divergence between $\mathbf{V}$ and $\mathbf{WH}$[15].

Numerous divergences are used as cost functions in NMF. Most common divergences probably are the Euclidean (EUC) distance, the KL divergence (see [15]) and the IS divergence (see [9]).

Several authors proposed generalized divergences which encompass these classical divergences.

- Csiszar's divergence [5], which is a generalization of Amari's α-divergence [6]. Both these divergences encompass the KL divergence and its dual.
- Bregman divergence [3], [7], which encompasses the EUC distance, the KL divergence and the IS divergence.
- β-divergence, introduced in [8] and studied as a cost function for NMF in [13] which also encompasses the EUC distance, the KL divergence and the IS divergence.

NMF is widely used in numerous areas such as image processing [11], [15], text mining [17], email surveillance [2], spectroscopy [10] and audio processing [9], [18], [19].

In this paper, we present a formal proof that the beta-divergence actually is a subclass of Bregman divergence. While this result is assumed to be known in a certain community [4], [16], we provide a full demonstration of it in the wide framework of element-wise divergences, and we present some applications to illustrate its interest. This result indeed permits to immediately particularize properties derived for Bregman divergence to β-divergence.

II. Divergence

In this section, we define the concept of divergence, element-wise divergence, and the particular case of Bregman divergence and β-divergence.

A. Definition

Divergences are distance-like functions which measure the separation between two elements.

Definition 2.1: Let $\mathcal{S}$ be a set. A *divergence* on S is a function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ satisfying

$$\forall (p,q) \in \mathcal{S} \times \mathcal{S} \quad D(p\|q) \geq 0, \text{ and } D(p\|q) = 0 \text{ iff } p = q.$$

As a distance, a divergence should be non-negative and separable. However, a divergence does not necessarily satisfy the triangle inequality and the symmetry axiom of a distance. In order to avoid the confusion with distances, the notation $D(p\|q)$ is often used instead of the classical distance notation $D(p,q)$.

B. Bregman Divergence

Definition 2.2: Let $\mathcal{S}$ be a convex subset of a Hilbert space and $\Phi : \mathcal{S} \to \mathbb{R}$ a continuously differentiable strictly convex function. The *Bregman divergence* [3] $D_\Phi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$

Manuscript received October 07, 2010; revised November 15, 2010; accepted November 15, 2010. Date of publication December 03, 2010; date of current version December 20, 2010. This work was supported by the CONTINT program of the French National Research Agency (ANR), as a part of the DReaM project (ANR-09-CORD-006-03), and by the Quaero Programme, funded by OSEO, French State agency for innovation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Konstantinos N. Plataniotis.

The authors are with the Institut Telecom, Telecom ParisTech, CNRS LTCI, 75634 Paris Cedex 13, France (e-mail: romain.hennequin@telecom-paristech.fr; bertrand.david@telecom-paristech.fr; roland.badeau@telecom-paristech.fr).

Digital Object Identifier 10.1109/LSP.2010.2096211

(where $\mathbb{R}_+$ is the set of non-negative real numbers) is defined as

$$D_\Phi(\mathbf{x}\|\mathbf{y}) = \Phi(\mathbf{x}) - \Phi(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \nabla\Phi(\mathbf{y})\rangle$$

where $\nabla\Phi(\mathbf{y})$ stands for the gradient of Φ evaluated at $\mathbf{y}$ and $\langle .,. \rangle$ is the standard Hermitian dot product.

The value of Bregman divergence $D_\Phi(\mathbf{x}\|\mathbf{y})$ can be viewed as the difference between the function $\Phi(\mathbf{x})$ and its first order Taylor series at $\mathbf{y}$. Thus, adding an affine form to Φ does not change D_Φ.

III. Element-Wise Divergence

A. Definition

In this section, $\mathcal{S} = \mathbb{R}_+^N$ or $\mathcal{S} = (\mathbb{R}_+\backslash\{0\})^N$. On such sets, one can define *element-wise divergences*: a divergence on $\mathbb{R}_+^N$ (resp. $(\mathbb{R}_+\backslash\{0\})^N$) is called element-wise if there exists a divergence d on $\mathbb{R}_+$ (resp. $\mathbb{R}_+\backslash\{0\}$) such that

$$\forall \mathbf{x}=(x_1,\ldots,x_n), \forall \mathbf{y}=(y_1,\ldots,y_n)\ D(\mathbf{x}\|\mathbf{y})=\sum_{n=1}^{N} d(x_n|y_n).$$

B. Element-Wise Bregman Divergence

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:

$$\forall \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{S} \quad \Phi(\mathbf{x}) = \sum_{n=1}^{N} \phi(x_i).$$

Then $D_\Phi(\mathbf{x}\|\mathbf{y}) = \sum_{i=1}^{N} d_\phi(x_i|y_i)$ where $d_\phi(x|y) = \phi(x) - \phi(y) - \phi'(y)(x-y)$ and thus, the divergence is element-wise. For element-wise Bregman divergences, we can equivalently denote the divergence D_Φ or D_ϕ.

C. β-Divergence

Definition 3.1: Let $\beta \in \mathbb{R}$. The β-divergence on $\mathbb{R}_+\backslash\{0\}$ is defined by:

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

One should notice that the previous definition of β-divergence is continuous with respect to β in the sense that:

$$\forall \beta_0 \in \mathbb{R}, \quad \forall x, y \in \mathbb{R}_+\backslash\{0\} \quad d_{\beta_0}(x,y) = \lim_{\beta\to\beta_0} d_\beta(x,y)$$

particularly for $\beta_0 = 0$ and $\beta_0 = 1$.

From this divergence on $\mathbb{R}_+\backslash\{0\}$, one can define an element-wise β-divergence on $(\mathbb{R}_+\backslash\{0\})^N$:

$$D_\beta(\mathbf{x}\|\mathbf{y}) = \sum_{n=1}^{N} d_\beta(x_n|y_n).$$

IV. β-Divergence as a Bregman Divergence

In this section, we show that the Bregman divergence encompasses the β-divergence in a natural way.

For $\beta \in \mathbb{R}$, let $\phi_\beta : \mathbb{R}_+\backslash\{0\} \to \mathbb{R}$ be the function defined as

$$\forall x \in \mathbb{R}_+\backslash\{0\}, \quad \phi_\beta(x) = \begin{cases} -\log x + x - 1, & \beta = 0 \\ x\log x - x + 1, & \beta = 1 \\ \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta}, & \text{otherwise.} \end{cases}$$

As shown in Appendix A, this definition is continuous with respect to β in the sense that

$$\forall \beta_0 \in \mathbb{R}, \forall x \in \mathbb{R}_+\backslash\{0\} \quad \lim_{\beta\to\beta_0} \phi_\beta(x) = \phi_{\beta_0}(x).$$

For all $\beta \in \mathbb{R}$, ϕ_β is smooth on $\mathbb{R}_+\backslash\{0\}$ and its second derivative is

$$\phi_\beta''(x) = x^{\beta-2}. \tag{1}$$

Thus ϕ_β is strictly convex and one can define the Bregman divergence D_{ϕ_β} associated to ϕ_β:

$$D_{\phi_\beta}(\mathbf{x}\|\mathbf{y}) = \sum_{n=1}^{N} \phi_\beta(x_n) - \phi_\beta(y_n) - \phi_\beta'(y_n)(x_n - y_n).$$

Straightforward calculations (see Appendix B) show that for all $\beta \in \mathbb{R}$, $D_{\phi_\beta} = D_\beta$ is a β-divergence. Thus the Bregman divergence encompasses β-divergence.

V. Applications

In this section, we present examples showing how our result can particularize properties of the Bregman divergence to the β-divergence, in order to illustrate its potential fields of application.

A. Non-Negative Matrix Factorization

Non-negative matrix factorization generally consists in minimizing an element-wise divergence between an $F \times T$ non-negative matrix $\mathbf{V}$ and its low-rank approximation $\mathbf{WH}$ (where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices respectively of dimension $F \times R$ and $R \times T$, with $R \ll F, T$):

$$D(\mathbf{V}\|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{t=1}^{T} d([\mathbf{V}]_{ft}|[\mathbf{WH}]_{ft}). \tag{2}$$

To perform this minimization, multiplicative update algorithms are widely used [5], [7], [15]. In such algorithms, the multiplicative update rule of $\mathbf{H}$ for minimizing (2) with an element-wise Bregman divergence D_ϕ cost function given in [7] is:

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T(\phi''(\mathbf{WH}).\mathbf{V})}{\mathbf{W}^T(\phi''(\mathbf{WH}).(\mathbf{WH}))}.$$

The product ".", the fraction bar, and ϕ'' are element-wise operations on the corresponding matrices. We can directly derive the (already well-known [5]) update rule of $\mathbf{H}$ for a β-divergence D_β cost function using (1):

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T((\mathbf{WH})^{.(\beta-2)}.\mathbf{V})}{\mathbf{W}^T((\mathbf{WH})^{.(\beta-1)})}.$$

This illustrates the interest of deriving general properties about the Bregman divergence instead of the β-divergence.

B. Right Type Centroid

The right type centroid is used in clustering as a "center" of a point cloud with respect to an asymmetric divergence: the right type centroid can thus be thought as an average typical point of a set.

Definition 5.1: Given a divergence D, the right type centroid of a finite set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\} \subset \mathcal{S}$ is defined as:

$$\mathbf{c}_{\text{right}}^{D} = \underset{\mathbf{c}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} D(\mathbf{x}_i \| \mathbf{c}).$$

Theorem 5.1: When $D = D_\beta$ is a β divergence, $\mathbf{c}_{\text{right}}^{D_\beta}$ is unique, independent of β and is equal to $\boldsymbol{\mu} = (1/n)\sum_{i=1}^{n} \mathbf{x}_i$.

Proof: It was shown in [1] that, when $D = D_\Phi$ is a Bregman divergence, $\mathbf{c}_{\text{right}}^{D_\Phi}$ is unique, independent of Φ and is equal to $\boldsymbol{\mu} = (1/n)\sum_{i=1}^{n} \mathbf{x}_i$. As β-divergence is a subclass of Bregman divergence, the proof is straightforward. □

VI. Conclusion

In this letter, we presented a proof that the general class of Bregman divergence encompasses the β-divergence in a natural way. This results permits to straightforwardly apply theorems about the Bregman divergence to the β-divergence. As the latter is widely used in methods such as NMF, which has applications in numerous areas (signal processing, clustering, data mining, spectroscopy), the field of application of this result is quite wide.

Appendix A
Continuity of ϕ_β With Respect to β

With the little-o notation, one can write as $\beta \to 0$:

$$\begin{aligned}\phi_\beta(x) &= \frac{e^{\beta \log x}}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{1}{\beta} \\ &= \frac{1+\beta\log x + o(\beta)}{\beta(\beta-1)} - \frac{x}{\beta-1} + \frac{\beta-1}{\beta(\beta-1)} \\ &= \frac{1+\log x - x}{\beta-1} + o(1).\end{aligned}$$

Then:

$$\lim_{\beta\to 0} \phi_\beta(x) = -\log x + x - 1.$$

In a similar way, one can write as $\beta \to 1$:

$$\begin{aligned}\phi_\beta(x) &= \frac{xe^{(\beta-1)\log x}}{\beta(\beta-1)} - \frac{\beta x}{\beta(\beta-1)} + \frac{1}{\beta} \\ &= \frac{x(-\beta+1+(\beta-1)\log x + o(\beta-1))}{\beta(\beta-1)} + \frac{1}{\beta} \\ &= \frac{-x + x\log x + 1}{\beta} + o(1).\end{aligned}$$

Then

$$\lim_{\beta\to 1} \phi_\beta(x) = x\log x - x + 1.$$

Appendix B
Equivalence Between the Bregman Divergence and the β-Divergence

For $\beta \in \mathbb{R}\backslash\{0,1\}$:

$$\begin{aligned}d_{\phi_\beta}(x|y) &= \frac{x^\beta}{\beta(\beta-1)} - \frac{x}{\beta-1} - \frac{y^\beta}{\beta(\beta-1)} \\ &\quad + \frac{y}{\beta-1} - \left(\frac{y^{\beta-1}}{\beta-1} - \frac{1}{\beta-1}\right)(x-y) \\ &= \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) = d_\beta(x|y).\end{aligned}$$

It is straightforward to check that the equality $d_{\phi_\beta}(x|y) = d_\beta(x|y)$ also holds for $\beta \in \{0,1\}$:

$$\begin{aligned}d_{\phi_0}(x|y) &= -\log x + x - (-\log y + y) - \left(-\frac{1}{y}+1\right)(x-y) \\ &= -\log x + \log y + (x-y) + \frac{x}{y} - 1 - (x-y) \\ &= -\log\frac{x}{y} + \frac{x}{y} - 1 = d_0(x|y), \\ d_{\phi_1}(x|y) &= x\log x - x + 1 - (y\log y - y + 1) - \log y(x-y) \\ &= x(\log x - \log y) + (y-x) = d_1(x|y).\end{aligned}$$

References

[1] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Oct. 2005.

[2] M. W. Berry and M. Browne, "Email surveillance using nonnegative matrix factorization," *Comput. Math. Organiz. Theory*, vol. 11, no. 3, pp. 249–264, Feb. 2005.

[3] L. M. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. Math. Phys*, vol. 7, no. 3, pp. 210–217, 1967.

[4] A. Cichocki and S. i. Amari, "Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, Jun. 2010.

[5] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Conf. Independent Component Analysis and Blind Source Separation (ICA)*, Charleston, SC, Mar. 2006, pp. 32–39.

[6] A. Cichocki, R. Zdunek, S. Choi, R. J. Plemmons, and S.-I. Amari, "Non-negative tensor factorization using alpha and beta divergences," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Honolulu, HI, Apr. 2007, vol. 3, pp. 1393–1396.

[7] I. S. Dhillon and S. Sra, Y. Weiss, B. Schölkopf, and J. Platt, Eds., “Generalized nonnegative matrix approximations with Bregman divergences,” in *Neural Information Processing Systems Conf. (NIPS)*, Cambridge, MA, Dec. 2006, pp. 283–290.

[8] S. Eguchi and Y. Kano*, Robustifying Maximum Likelihood Estimation. Technical Report*. Tokyo, Japan: Inst. Statist. Math., 2001.

[9] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis,” *Neural Comput.*, vol. 11, no. 3, pp. 793–830, Mar. 2009.

[10] C. Gobinet, E. Perrin, and R. Huez, “Application of non-negative matrix factorization to fluorescence spectroscopy,” in *Eur. Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sep. 2004.

[11] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[12] F. Itakura and S. Saito, “Analysis synthesis telephony based on the maximum likelihood method,” in *6th Int. Congr. Acoustics*, Tokyo, Japan, 1968, pp. C–17–C–20.

[13] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural Comput.*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[14] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[15] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[16] F. Nielsen and R. Nock, “The dual voronoi diagrams with respect to representational bregman divergences,” in *Int. Symp. Voronoi Diagrams*, Copenhagen, Denmark, Jun. 2009, pp. 71–78.

[17] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, “Text mining using non-negative matrix factorizations,” in *SIAM Int. Conf. Data Mining*, Lake Buena Vista, FL, Jan. 2004, pp. 452–456.

[18] J. Paulus and T. Virtanen, “Drum transcription with non-negative spectrogram factorization,” in *Eur. Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sep. 2005.

[19] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.