On Bregman Distances and Divergences of Probability Measures

Wolfgang Stummer and Igor Vajda, *Fellow, IEEE*

Abstract

The paper introduces scaled Bregman distances of probability distributions which admit non-uniform contributions of observed events. They are introduced in a general form covering not only the distances of discrete and continuous stochastic observations, but also the distances of random processes and signals. It is shown that the scaled Bregman distances extend not only the classical ones studied in the previous literature, but also the information divergence and the related wider class of convex divergences of probability measures. An information processing theorem is established too, but only in the sense of invariance w.r.t. statistically sufficient transformations and not in the sense of universal monotonicity. Pathological situations where coding can increase the classical Bregman distance are illustrated by a concrete example. In addition to the classical areas of application of the Bregman distances and convex divergences such as recognition, classification, learning and evaluation of proximity of various features and signals, the paper mentions a new application in 3D-exploratory data analysis. Explicit expressions for the scaled Bregman distances are obtained in general exponential families, with concrete applications in the binomial, Poisson and Rayleigh families, and in the families of exponential processes such as the Poisson and diffusion processes including the classical examples of the Wiener process and geometric Brownian motion.

***Index Terms* — Bregman distances, divergences, sufficiency, exponential distributions, exponential processes, classification, statistical decision, information retrieval, machine learning.**

I. Introduction

Bregman (1967) introduced for convex functions $\phi : \mathbb{R}^d \to \mathbb{R}$ with gradient $\nabla\phi$ the ϕ-depending nonnegative measure of dissimilarity

$$B_\phi(p, q) = \phi(p) - \phi(q) - \nabla\phi(q)(p - q) \tag{1}$$

of d-dimensional vectors $p, q \in \mathbb{R}^d$. His motivation was the problem of convex programming, but in the subsequent literature it became widely applied in many other problems under the name *Bregman distance* in spite of that it is not in general the usual metric distance (it is a pseudodistance which is reflexive but neither

W. Stummer is with the Department of Mathematics, University of Erlangen–Nürnberg, Bismarckstrasse $1\frac{1}{2}$, 91054 Erlangen, Germany (e-mail: stummer@mi.uni-erlangen.de), as well as with the School of Business and Economics, University of Erlangen–Nürnberg.

I. Vajda is with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou Věží 4, 182 08 Prague, Czech Republic (e-mail: vajda@utia.cas.cz).

Manuscript received October 26, 2009. This work was supported by the MŠMT grant 1M0572 and the GA ČR grant 102/07/1131.

symmetric nor satisfying the triangle inequality). The most important feature is the special *separable form* defined by

$$B_\phi(p,q) = \sum_{i=1}^{d} [\phi(p_i) - \phi(q_i) - \phi'(q_i)(p_i - q_i)] \tag{2}$$

for vectors $p = (p_1, ..., p_d), q = (q_1, ..., q_d)$ and convex differentiable functions $\phi : \mathbb{R} \to \mathbb{R}$. For example, the function $\phi(t) = (t-1)^2$ leads to the classical squared Euclidean distance

$$B_\phi(p,q) = \sum_{i=1}^{d} (p_i - q_i)^2 . \tag{3}$$

In the optimization-theoretic context the Bregman distances are usually studied in the general form (1), see e.g. Csiszár and Matúš (2008, 2009). In the information-theoretic or statistical context they are typically used in the separable form (2) for vectors p, q with nonnegative coordinates representing generalized distributions (finite *discrete* measures) and functions $\phi : [0,\infty) \to \mathbb{R}$ differentiable on $(0,\infty)$ (the problem with $q_i = 0$ is solved by resorting to the right-hand derivative $\phi'_+(0)$). The concrete example $\phi(t) = t \ln t$ leads to the well-known Kullback divergence

$$B_\phi(p,q) = \sum_{i=1}^{d} p_i \ln \frac{p_i}{q_i}. \tag{4}$$

Of course, the most common context are discrete probability distributions p, q since vectors of hypothetical or observed frequencies p, q are easily transformed to the relative frequencies normed to 1. For example, Csiszár (1991, 1994, 1995) or Pardo and Vajda (1997, 2003) used the Bregman distances of probability distributions in the context of information theory and asymptotic statistics.

Important alternatives to the Bregman distances (2) are the *ϕ-divergences* defined by

$$D_\phi(p,q) = \sum_{i=1}^{d} q_i \phi\left(\frac{p_i}{q_i}\right) \tag{5}$$

for functions ϕ which are convex on $[0,\infty)$, continuous on $(0,\infty)$ and strictly convex at 1 with $\phi(1) = 0$. Originating in the paper of Csiszár (1963), they share some properties with the Bregman distances (2), e.g. they are pseudodistances too. For example, the above considered functions $\phi(t) = (t-1)^2$ and $\phi(t) = t \ln t$ lead in this case to the classical Pearson divergence

$$D_\phi(p,q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{q_i} \tag{6}$$

and the above mentioned Kullback divergence $D_\phi(p,q) \equiv B_\phi(p,q)$ which are asymmetric in p, q and contradict the triangle inequality. On the other hand, $\phi(t) = |t-1|$ leads to the L_1-norm $||p-q||$ which is a metric distance and $\phi(t) = (t-1)^2/(t+1)$ defines the LeCam divergence

$$D_\phi(p,q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{p_i + q_i}$$

which is a squared metric distance (for more about the metricity of ϕ-divergences see Vajda (2009)).

However, there exist also some sharp differences between these two types of pseudodistances of distributions. A distinguished property of the Bregman distances is the *optimality of the k-means algorithm* for them. For the squared Euclidean error (3) this optimality was known long ago (see in this respect the seminal work of

Lloyd (1982) reprinting a Technical Report of Bell Laboratories dated by 1957). For all Bregman distances (1) it was established relatively recently by Banerjee et al. (2005). This property is not shared by those of the ϕ-divergences which are not Bregman distances, e.g. by the Pearson divergence (6). A distinguished property of ϕ-divergences is the *information processing property*, i.e. the impossibility to increase the value $D_\phi(p,q)$ by transformations of the observations distributed by p, q and preservation of this value by the statistically sufficient transformations (Csiszár (1967), see in this respect also Liese and Vajda (2006)). This property is not shared by the Bregman distances which are not ϕ-divergences. For example, the distributions $p = (1/2, 1/4, 1/4)$ and $q = (1, 0, 0)$ are mutually closer (less discernible) in the Euclidean sense (3) than their reductions $\tilde{p} = (1/2, 1/4)$ and $\tilde{q} = (1, 0)$ obtained by merging the second and third observation outcomes into one.

Depending on the need to exploit one or the other of these distinguished properties, the Bregman distances or Csiszár divergences are preferred, and both of them are widely applied in important areas of information theory, statistics and computer science, for example in

(Ai) *information retrieval* (see e.g. Do and Vetterli (2002), Hertz at al. (2004)),

(Aii) *optimal decision* (for *general decision* see e.g. Boratynska (1997), Freund et al. (1997), Bartlett et al. (2006), Vajda and Zvárová (2007), for *speech processing* see e.g. Carlson and Clements (1991), Veldhuis and Klabers (2002), for *image processing* see e.g. Xu and Osher (2007), Marquina and Osher (2008), Scherzer et al. (2008)), and

(Aiii) *machine learning* (see e.g. Laferty (1999), Banerjee et al. (2005), Amari (2007), Teboulle (2007), Nock and Nielsen (2009)).

In this context it is obvious the importance of the functionals of distributions which are simultaneously divergences in both the Csiszár and Bregman sense or, more broadly, of the research of relations between the Csiszár and Bregman divergences. This paper is devoted to this research. It generalizes the separable Bregman distances (2) as well as the ϕ-divergences (5) by introducing the *scaled Bregman distances* which for the discrete setup reduce to

$$\begin{aligned} B_\phi(p,q|m) &= \sum_{i=1}^{d} \Big[\phi(p_i/m_i) - \phi(q_i/m_i) \\ &\quad -\phi'_+(q_i/m_i)(p_i/m_i - q_i/m_i)\Big] m_i \end{aligned} \tag{7}$$

for arbitrary finite scale vectors $m = (m_1, ..., m_d)$, convex functions ϕ and right-hand derivatives ϕ'_+. Obviously, the uniform scales $m = (1, ..., 1)$ lead to the Bregman distances (2) and the probability distribution scales $m = q = (q_1, ..., q_d)$ lead to the ϕ-divergences (5). We shall work out further interesting relations of the $B_\phi(p,q|m)$ distances to the ϕ-divergences $D_\phi(p,q)$ and $D_\phi(p,m)$ and evaluate explicit formulas for the stochastically scaled Bregman distances in arbitrary exponential families of distributions, including also the non-discrete setup.

Section II defines the ϕ-divergences $D_\phi(P,M)$ of general probability measures P and arbitrary finite measures M and briefly reviews their basic properties. Section III introduces scaled Bregman distances $B_\phi(P,Q|M)$ and investigates their relations to the ϕ-divergences $D_\phi(P,Q)$ and $D_\phi(P,M)$. Section IV studies in detail the situation where all three measures P, Q, M are from the family of general exponential distributions. Finally, Section V illustrates the results by investigating concrete examples of P, Q, M from classical statistical families as well as from a family of important random processes.

Notational conventions: Throughout the paper, $\mathcal{M}$ denotes the space of all finite measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and $\mathcal{P} \subset \mathcal{M}$ the subspace of all probability measures. Unless otherwise explicitly stated P, Q, M are mutually measure-theoretically equivalent measures on $(\mathcal{X}, \mathcal{A})$ dominated by a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$. Then the densities

$$p = \frac{\mathrm{d}P}{\mathrm{d}\lambda}, \quad q = \frac{\mathrm{d}Q}{\mathrm{d}\lambda} \qquad \text{and} \qquad m = \frac{\mathrm{d}M}{\mathrm{d}\lambda} \tag{8}$$

have a common support which will be identified with $\mathcal{X}$ (i.e. the densities (8) are positive on $\mathcal{X}$). Unless otherwise explicitly stated, it is assumed that $P, Q \in \mathcal{P}, \quad M \in \mathcal{M}$ and that $\phi : (0, \infty) \mapsto \mathbb{R}$ is a continuous and convex function. It is known that then the possibly infinite extension $\phi(0) = \lim_{t \downarrow 0} \phi(t)$ and the right-hand derivatives $\phi'_+(t)$ for $t \in [0, \infty)$ exist, and that the adjoint function

$$\phi^*(t) = t\phi(1/t) \tag{9}$$

is continuous and convex on $(0, \infty)$ with possibly infinite extension $\phi^*(0)$. We shall assume that $\phi(1) \equiv \phi^*(1) = 0$.

II. DIVERGENCES

For $P, Q, \in \mathcal{P}$ and $M \in \mathcal{M}$ we consider

$$D_\phi(P, M) = \int_{\mathcal{X}} m\, \phi\left(\frac{p}{m}\right) \mathrm{d}\lambda \qquad (cf.\ (8)) \tag{10}$$

generated by the same convex functions as considered in the formula (5) for discrete P and M. $D_\phi(P, Q)$ is a special case.

The existence (but possible infinity) of the ϕ-divergences follows from the bounds

$$\phi'_+(1)(p - m) \;\leq\; m\, \phi\left(\frac{p}{m}\right) \;\leq\; m\, \phi(0) + p\, \phi^*(0) \tag{11}$$

on the integrand, leading to the ϕ-divergence bounds

$$\phi'_+(1)(1 - M(\mathcal{X})) \;\leq\; D_\phi(P, M) \;\leq\; M(\mathcal{X})\, \phi(0) + \phi^*(0). \tag{12}$$

The integrand bounds (11) follow by putting $s = 1$ and $t = p/m$ in the inequality

$$\phi(s) + \phi'_+(s)(t - s) \leq \phi(t) \leq \phi(0) + t\phi^*(0), \tag{13}$$

where the left-hand side is the well-known support line of $\phi(t)$ at $t = s$. The right-hand inequality is obvious for $\phi(0) = \infty$. If $\phi(0) < \infty$ then it follows by taking $s \to \infty$ in the inequality

$$\phi(t) \leq \phi(0) \,+\, t\, \frac{\phi(s) - \phi(0)}{s}\ ,$$

obtained from the Jensen inequality for $\phi(t)$ situated between $\phi(0)$ and $\phi(s)$. Since the function $\psi(p, m) = m\phi(p/m)$ is homogeneous in the sense $\psi(tp, tm) = t\psi(p, m)$ for all $t > 0$, the divergences (10) do not depend on the choice of the dominating measure λ.

Notice that $D_\phi(P,M)$ might be negative. For probability measures P,Q the bounds (12) take on the form

$$0 \le D_\phi(P,Q) \le \phi(0) + \phi^*(0), \tag{14}$$

and the equalities are achieved under well-known conditions (cf. Liese and Vajda (1987), (2006)): the left equality holds *if* $P = Q$, and the right one holds *if* $P \perp Q$ (singularity). Moreover, if $\phi(t)$ is strictly convex at $t = 1$, the first *if* can be replaced by *iff*, and in the case $\phi(0) + \phi^*(0) < \infty$ also the second *if* can be replaced by *iff*.

An alternative to the left-hand inequality in (12), which extends the left-hand inequality in (14) including the conditions for the equality, is given by the following statement (for a systematic theory of ϕ-divergences of finite measures we refer to the recent paper of Stummer and Vajda (2009)).

Lemma 1: For every $P \in \mathcal{P}$, $M \in \mathcal{M}$ one gets the lower divergence bound

$$M(\mathcal{X})\,\phi\left(\frac{1}{M(\mathcal{X})}\right) \le D_\phi(P,M)\ , \tag{15}$$

where the equality holds if

$$p \;=\; \frac{m}{M(\mathcal{X})} \qquad P\text{-a.s.} \tag{16}$$

If $D_\phi(P,M) < \infty$ and $\phi(t)$ is strictly convex at $t = 1/M(\mathcal{X})$, the equality in (15) holds if and only if (16) holds.

Proof: By (10) and the definition (9) of the convex function ϕ^*

$$D_\phi(P,M) = \int_{\mathcal{X}} \phi^*\left(\frac{m}{p}\right) \mathrm{d}P.$$

Hence by Jensen's inequality

$$D_\phi(P,M) \ge \phi^*\left(\int_{\mathcal{X}} \frac{m}{p}\,\mathrm{d}P\right) = \phi^*(M(\mathcal{X})) \tag{17}$$

which proves the desired inequality (15). Since

$$\frac{m}{p} \;=\; M(\mathcal{X}) \quad P\text{-a. s.}$$

is the condition for equality in (17), the rest is clear from the easily verifiable fact that $\phi^*(t)$ is strictly convex at $t = s$ if and only if $\phi(t)$ is strictly convex at $t = 1/s$. □

For some of the representation investigations below, it will also be useful to take into account that for probability measures P,Q we get directly from definition (10) the "skew symmetry" ϕ-divergence formula

$$D_{\phi^*}(P,Q) = D_\phi(Q,P)\ , \tag{18}$$

as well as the sufficiency of the condition

$$\phi(t) - \phi^*(t) \;\equiv\; \text{constant} \cdot (t-1) \tag{19}$$

for the ϕ-divergence symmetry

$$D_\phi(P,Q) = D_\phi(Q,P) \quad \text{for all } P,Q\ . \tag{20}$$

Liese and Vajda (1987) proved that under the assumed strict convexity of $\phi(t)$ at $t = 1$ the condition (19) is is not only *sufficient* but also *necessary* for the symmetry (20).

III. SCALED BREGMAN DISTANCES

Let us now introduce the basic concept of the current paper, which is a measure-theoretic version of the Bregman distance (7). In this definition it is assumed that ϕ is a finite convex function in the domain $t > 0$, continuously extended to $t = 0$. As before, $\phi'_+(t)$ denotes the right-hand derivative which for such $\phi(t)$ exists and p, q, m are the densities defined in (8).

Definition 1: The *Bregman distance* of probability measures P, Q *scaled* by an arbitrary measure M on $(\mathcal{X}, \mathcal{A})$ measure-theoretically equivalent with P, Q is defined by the formula

$$\begin{aligned} &B_\phi\left(P, Q \,|\, M\right) \\ &= \int_{\mathcal{X}} \left[\phi\left(\frac{p}{m}\right) - \phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)\left(\frac{p}{m} - \frac{q}{m}\right)\right] \mathrm{d}M \\ &= \int_{\mathcal{X}} \left[m\phi\left(\frac{p}{m}\right) - m\phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)(p-q)\right] \mathrm{d}\lambda. \end{aligned} \tag{21}$$

The convex ϕ under consideration can be interpreted as a generating function of the distance.

Remark: By putting $t = p/m$ and $s = q/m$ in (13) we find the argument of the integral in (21) to be nonnegative. Hence the Bregman distance $B_\phi\left(P, Q \,|\, M\right)$ is well-defined by (21) and is always nonnegative (possibly infinite).

The special scaled Bregman distances $B_\phi\left(P, Q \,|\, M\right)$ for probability scales $M \in \mathcal{P}$ were introduced by Stummer (2007). Let us mention some other important previously considered special cases.

(a) For $\mathcal{X}$ finite or countable and counting measure $M = \lambda$ some authors were already cited above in connection with the formula (2) and the research areas **(Ai) - (Aiii)**. In addition to them, one can mention also Byrne (1999), Collins et al. (2002), Murata et al. (2004), Cesa-Bianchi and Lugosi (2006).

(b) For open Euclidean set $\mathcal{X}$ and Lebesgue measure $M = \lambda$ on it one can mention Jones and Byrne (1990), as well as Resmerita and Anderssen (2007).

In the rest of this paper, we restrict ourselves to the Bregman distances $B_\phi\left(P, Q \,|\, M\right)$ scaled by finite measures $M \in \mathcal{M}$ and to the same class of convex functions as considered in the ϕ-divergence formulas (5) and (10). By using the remark after Definition 1 and applying (13) we get

$$D_\phi(P, M) \geq D_\phi(Q, M) + \int_{\mathcal{X}} \phi'_+\left(\frac{q}{m}\right)(p-q)\mathrm{d}\lambda \tag{22}$$

if at least one of the right-hand side expressions is finite. Similarly,

$$B_\phi\left(P, Q \,|\, M\right) = D_\phi(P, M) - D_\phi(Q, M) - \int_{\mathcal{X}} \phi'_+\left(\frac{q}{m}\right)\mathrm{d}\lambda \tag{23}$$

if at least two of the right-hand side expressions are finite (which can be checked e.g. by using (12) or (15)).

The formula (21) simplifies in the important special cases $M = P$ and $M = Q$. In the first case, due to $\phi(1) = 0$ it reduces to

$$B_\phi\left(P,Q\,|\,P\right) = \int_{\mathcal{X}} \left[\phi'_+\left(\frac{q}{p}\right)(q-p) - p\phi\left(\frac{q}{p}\right)\right] \mathrm{d}\lambda \tag{24}$$

$$= \int_{\mathcal{X}} \phi'_+\left(\frac{q}{p}\right)(q-p)\mathrm{d}\lambda - D_\phi(Q,P)\ , \tag{25}$$

where the difference (25) is meaningful if and only if $D_\phi(Q,P) \equiv D_{\phi^*}(P,Q)$ is finite. The nonnegative divergence measure $\mathcal{B}_\phi\left(P,Q\right) := B_\phi\left(P,Q\,|\,P\right)$ is thus the difference between the nonnegative dissimilarity measure

$$\mathcal{D}_\phi\left(Q,P\right) = \int_{\mathcal{X}} \phi'_+\left(\frac{q}{p}\right)(q-p)\,\mathrm{d}\lambda \ \geq\ D_\phi(Q,P)$$

and the nonnegative $\phi-$divergence $D_\phi(Q,P)$. Furthermore, in the second special case $M = Q$ the formula (21) leads to the equality

$$B_\phi\left(P,Q\,|\,Q\right) = D_\phi(P,Q) \tag{26}$$

without any restriction on $P,Q \in \mathcal{P}$ as realized already by Stummer (2007).

Conclusion 1: Equality (26) – together with the fact that $B_\phi\left(P,Q\,|\,M\right)$ depends in general on M (see e.g. Subsection B below) – shows that the concept of scaled Bregman distance (21) strictly generalizes the concept of $\phi-$divergence $D_\phi(P,Q)$ of probability measures P,Q.

Example 1: As an illustration not considered earlier we can take the non-differentiable function $\phi(t) = 2|t-1|$ for which

$$B_\phi\left(P,Q\,|\,Q\right) = V(P,Q)$$

i.e. this particular scaled Bregman distance reduces to the well known total variation.

As demonstrated by an example in the Introduction, measurable transformations (statistics)

$$T : (\mathcal{X},\mathcal{A}) \mapsto (\mathcal{Y},\mathcal{B}) \tag{27}$$

which are *not* sufficient for $\{P,Q\}$ can increase those of the scaled Bregman distances $B_\phi\left(P,Q\,|\,M\right)$ which are not ϕ-divergences. On the other hand, the transformations (27) which *are* sufficient for $\{P,Q\}$ need not preserve these distances either. Next we formulate conditions under which the scaled Bregman distances $B_\phi\left(P,Q\,|\,M\right)$ are preserved by transformations of observations.

Definition 2: We say that the transformation (27) is sufficient for the triplet $\{P,\,Q,\,M\}$ if there exist measurable functions $g_P, g_Q, g_M : \mathcal{Y} \mapsto \mathbb{R}$ and $h : \mathcal{X} \mapsto \mathbb{R}$ such that

$$p(x) = g_P(Tx)h(x), \quad q(x) = g_Q(Tx)h(x)$$

$$\text{and} \quad m(x) = g_M(Tx)h(x). \tag{28}$$

If M is probability measure then our definition reduces to the classical statistical sufficiency of the statistic T for the family $\{P,\,Q,\,M\}$ (see pp. 18-19 in Lehman (2005)). All transformations (27) induce the probability measures PT^{-1}, QT^{-1} and the finite measure MT^{-1} on $(\mathcal{Y},\mathcal{B})$. We prove that the scaled Bregman distances of induced probability measures PT^{-1}, QT^{-1} scaled by MT^{-1} are preserved by sufficient transformations T.

Theorem 1: The transformations (27) sufficient for the triplet $\{P, Q,\, M\}$ preserve the scaled Bregman distances in the sense that

$$B_\phi\left(PT^{-1}, QT^{-1} \,|\, MT^{-1}\right) = B_\phi\left(P, Q \,|\, M\right). \tag{29}$$

Proof.: By (21) and (28), the right-hand side of (29) is equal to

$$\int_{\mathcal{X}} \left[\phi_{P,M}\left(Tx\right) - \phi_{Q,M}\left(Tx\right) - \Delta_{P,Q,M}\left(Tx\right)\right] \mathrm{d}M \tag{30}$$

for

$$\phi_{P,M}\left(y\right) = \phi\left(\frac{g_P(y)}{g_M(y)}\right), \quad \phi_{Q,M}\left(y\right) = \phi\left(\frac{g_Q(y)}{g_M(y)}\right) \tag{31}$$

and

$$\Delta_{P,Q,M}\left(y\right) = \phi_+^{'}\left(\frac{g_Q(y)}{g_M(y)}\right)\left(g_P(y) - g_Q(y)\right). \tag{32}$$

By Theorem D in Section 39 of Halmos (1964), the integral (30) is equal to

$$\int_{\mathcal{Y}} \left[\phi_{P,M}\left(y\right) - \phi_{Q,M}\left(y\right) - \Delta_{P,Q,M}\left(y\right)\right] \mathrm{d}MT^{-1} \tag{33}$$

and, moreover,

$$P(T^{-1}B) = \int_B g_P(y)\, h(T^{-1}y)\, \mathrm{d}\lambda T^{-1}$$

and similarly for Q instead of P. Therefore

$$\frac{\mathrm{d}PT^{-1}}{\mathrm{d}\lambda T^{-1}} = g_P(y)\, h(T^{-1}y) \quad \text{and} \quad \frac{\mathrm{d}QT^{-1}}{\mathrm{d}\lambda T^{-1}} = g_Q(y)\, h(T^{-1}y)$$

which together with (31), (32) and (21) implies that the integral (33) is nothing but the left-hand side of (29). This completes the proof. □

In the rest of this section we discuss some important special classes of scaled Bregman distances obtained for special distance-generating functions ϕ.

A. Bregman logarithmic distance

Let us consider the special function $\phi(t) = t \ln t$. Then $\phi'(t) = \ln t + 1$ so that (21) implies

$$\begin{aligned}
& B_{t\ln t}\left(P, Q \,|\, M\right) \\
&= \int_{\mathcal{X}} \left[p \ln \frac{p}{m} - q \ln \frac{q}{m} - \left(\ln \frac{q}{m} + 1\right)(p - q)\right] \mathrm{d}\lambda \\
&= \int_{\mathcal{X}} \left[p \ln \frac{p}{m} - p \ln \frac{q}{m}\right] \mathrm{d}\lambda \\
&= \int_{\mathcal{X}} p \ln \frac{p}{q}\, \mathrm{d}\lambda \;=\; D_{t\ln t}\left(P, Q\right) \;.
\end{aligned} \tag{34}$$

Thus, for $\phi(t) = t \ln t$ the Bregman distance $B_\phi\left(P, Q \,|\, M\right)$ exceptionally does not depend on the choice of the scaling and reference measures M and λ; in fact, it always leads to the Kulllback-Leibler information divergence (relative entropy) $D_{t\ln t}(P, Q)$, see Stummer (2007).

B. Bregman reversed logarithmic distance

Let now $\phi(t) = -\ln t$ so that $\phi'(t) = -1/t$. Then (21) implies

$$
\begin{aligned}
& B_{-\ln t}\,(P,Q\,|\,M) \\
&= \int_{\mathcal{X}} \left[m \ln \frac{m}{p} - m \ln \frac{m}{q} + \frac{m}{q}(p-q) \right] \mathrm{d}\lambda && (35) \\
&= D_{t\ln t}(M,P) - D_{t\ln t}(M,Q) + \int_{\mathcal{X}} \frac{mp}{q}\,\mathrm{d}\lambda - M(\mathcal{X}) && (36) \\
&= D_{-\ln t}(P_1,M) - D_{-\ln t}(q,M) + \int_{\mathcal{X}} \frac{mp}{q}\mathrm{d}\lambda - M(\mathcal{X}) && (37)
\end{aligned}
$$

where the equalities (36) and (37) hold if at least two out of the first three expressions on the right-hand side are finite. In particular, (35) implies (in consistency with (26))

$$
B_{-\ln t}\,(P,Q\,|\,Q) = D_{-\ln t}(P,Q) \tag{38}
$$

and (36) implies for $D_{t\ln t}(P,Q) < \infty$ (in consistency with (25))

$$
B_{-\ln t}\,(P,Q\,|\,P) = \chi^2(P,Q) - D_{t\ln t}(P,Q) \tag{39}
$$

where

$$
\chi^2(P,Q) = \int_{\mathcal{X}} \frac{(p-q)^2}{q}\,\mathrm{d}\lambda
$$

is the well-known Pearson information divergence. From (38) and (39) one can also see that the Bregman distance $B_\phi\,(P,Q\,|\,M)$ does in general depend on the choice of the reference measure M.

C. Bregman power distances

In this subsection we restrict ourselves for simplicity to probability measures $M \in \mathcal{P}$, i.e. we suppose $M(\mathcal{X}) = 1$. Under this assumption we investigate the scaled Bregman distances

$$
B_\alpha\,(P,Q\,|\,M) = B_{\phi_\alpha}\,(P,Q\,|\,M)\;\;, \quad \alpha \in \mathbb{R},\; \alpha \neq 0,\; \alpha \neq 1 \tag{40}
$$

for the family of power convex functions

$$
\phi(t) \equiv \phi_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha-1)} \quad \text{with} \quad \phi'_\alpha(t) = \frac{t^{\alpha-1}}{\alpha - 1}\;. \tag{41}
$$

For comparison and representation purposes, we use for P (and analogously for Q instead of P) the power divergences

$$
\begin{aligned}
& D_\alpha(P,M) = D_{\phi_\alpha}(P,M) \\
&= \frac{1}{\alpha(\alpha-1)} \left[\int_{\mathcal{X}} p^\alpha\, m^{1-\alpha}\,\mathrm{d}\lambda - 1 \right] && (42) \\
&= \frac{\exp \rho_\alpha(P,M) - 1}{\alpha(\alpha-1)} \quad \text{with } \rho_\alpha(P,M) = \ln \int_{\mathcal{X}} p^\alpha\, m^{1-\alpha}\,\mathrm{d}\lambda && (43)
\end{aligned}
$$

of real powers α different from 0 and 1, studied for arbitrary probability measures P, M in Liese and Vajda (1987). They are one-one related to the Rényi divergences

$$
R_\alpha(P,M) = \frac{\rho_\alpha(P,M)}{\alpha(\alpha-1)}, \quad \alpha \in \mathbb{R},\; \alpha \neq 0,\; \alpha \neq 1,
$$

introduced in Liese and Vajda (1987) as an extension of the original narrower class of the divergences

$$R_\alpha(P,M) = \frac{\rho_\alpha(P,M)}{\alpha - 1}, \quad \alpha > 0,\ \alpha \neq 1$$

of Rényi (1961).

Returning now to the Bregman power distances, observe that if $D_\alpha(P,M) + D_\alpha(Q,M)$ is finite then (23), (40) and (41) imply for $\alpha \neq 0,\ \alpha \neq 1$

$$\begin{aligned} & B_\alpha(P,Q\,|\,M) \\ &= -D_\alpha(Q,M) - \frac{1}{\alpha-1}\int_{\mathcal{X}} \left(\frac{q}{m}\right)^{\alpha-1}(p-q)\,\mathrm{d}\lambda && (44) \\ &= D_\alpha(P,M) - D_\alpha(Q,M) \\ &\qquad - \frac{1}{\alpha-1}\int_{\mathcal{X}} \left[\left(\frac{q}{m}\right)^{\alpha-1} p - \left(\frac{q}{m}\right)^{\alpha} m\right]\mathrm{d}\lambda && (45) \\ &= D_\alpha(P,M) - (1-\alpha)\,D_\alpha(Q,M) \\ &\qquad - \frac{1}{\alpha-1}\left[\int_{\mathcal{X}} \left(\frac{q}{m}\right)^{\alpha-1} p\,\mathrm{d}\lambda - 1\right]. && (46) \end{aligned}$$

In particular, we get from here (in consistency with (26))

$$B_\alpha(P,Q\,|\,Q) = D_\alpha(P,Q) \tag{47}$$

and in case of $D_\alpha(Q,P) \equiv D_{1-\alpha}(P,Q) < \infty$ also

$$B_\alpha(P,Q\,|\,P) = (\alpha-2)\,D_{\alpha-1}(Q,P) + (\alpha-1)\,D_\alpha(Q,P) \tag{48}$$

$$\equiv (\alpha-2)\,D_{2-\alpha}(P,Q) + (\alpha-1)\,D_{1-\alpha}(P,Q). \tag{49}$$

In the following theorem, and elsewhere in the sequel, we use the simplified notation

$$D_1(P,M) = D_{t\ln t}(P,M) \ \text{ and } \ D_0(P,M) = D_{-\ln t}(P,M) \tag{50}$$

for the probability measures P, M under consideration (and also later on where M is only a finite measure). This step is motivated by the limit relations

$$\begin{aligned} \lim_{\alpha\downarrow 0} D_\alpha(P,M) &= D_{-\ln t}(P,M) \quad \text{and} \\ \lim_{\alpha\uparrow 1} D_\alpha(P,M) &= D_{t\ln t}(P,M) && (51) \end{aligned}$$

proved as Proposition 2.9 in Liese and Vajda (1987) for arbitrary probability measures P, M. Applying these relations to the Bregman distances, we obtain

Theorem 2: If $D_0(P,M) + D_0(Q,M) < \infty$ then

$$\begin{aligned} & \lim_{\alpha\downarrow 0} B_\alpha(P,Q\,|\,M) \\ &= D_0(P,M) - D_0(Q,M) + \int_{\mathcal{X}} \frac{mp}{q}\,\mathrm{d}\lambda - 1 && (52) \\ &= B_{-\ln t}(P,Q\,|\,M). && (53) \end{aligned}$$

If $D_1(P,M) + D_1(Q,M) < \infty$ and

$$\lim_{\beta \downarrow 0} \int_{\mathcal{X}} \frac{(q/m)^{-\beta} - 1}{\beta} \, \mathrm{d}P$$
$$= \int_{\mathcal{X}} \lim_{\beta \downarrow 0} \frac{(q/m)^{-\beta} - 1}{\beta} \, \mathrm{d}P = - \int_{\mathcal{X}} \ln \frac{q}{m} \, \mathrm{d}P \tag{54}$$

then

$$\lim_{\alpha \uparrow 1} B_\alpha(P,Q \,|\, M) = D_1(P,M) - \int_{\mathcal{X}} \ln \frac{q}{m} \, \mathrm{d}P \tag{55}$$

$$= D_1(P,Q) \;=\; B_{t \ln t}(P,Q \,|\, M) \;. \tag{56}$$

Proof: If $0 < \alpha < 1$ then $D_\alpha(P,M)$, $D_\alpha(Q,M)$ are finite so that (46) holds. Applying the first relation of (51) in (46) we get (52) where the right hand side is well defined because $D_0(P,M) + D_0(Q,M)$ is by assumption finite. Similarly, by using the second relation of (51) and the assumption (54) in (46) we end up at (55) where the right-hand side is well defined because $D_1(P,M)+D_1(Q,M)$ is assumed to be finite. The identity (53) follows from (52), (37) and the identity (56) from (55), (34). □

Motivated by this theorem, we introduce for all probability measures P, Q, M under consideration the simplified notations

$$B_1(P,Q \,|\, M) \;=\; B_{t \ln t}(P,Q \,|\, M) \tag{57}$$

and

$$B_0(P,Q \,|\, M) \;=\; B_{-\ln t}(P,Q \,|\, M) \;, \tag{58}$$

and thus, (56) and (53) become

$$B_1(P,Q \,|\, M) = \lim_{\alpha \uparrow 1} B_\alpha(P,Q \,|\, M) \tag{59}$$

and

$$B_0(P,Q \,|\, M) = \lim_{\alpha \downarrow 0} B_\alpha(P,Q \,|\, M). \tag{60}$$

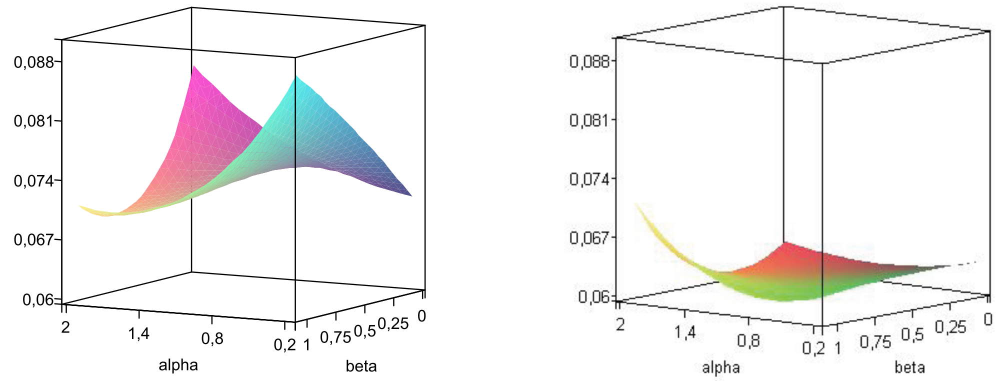
Furthermore, in these notations the relations (34), (38) and (39) reformulate (under the corresponding assumptions) as follows

$$B_1(P,Q \,|\, M) \;=\; D_1(P,Q) \;, \tag{61}$$

$$B_0(P,Q \,|\, Q) = D_0(P,Q) \tag{62}$$

and

$$\begin{aligned} B_0(P,Q \,|\, P) \;&=\; \chi^2(P,Q) - D_1(P,Q) \\ &=\; 2\, D_2(P,Q) - D_1(P,Q). \end{aligned} \tag{63}$$



(a) p=0.25, q=0.20 (b) p=0.25, q=0.30

Fig. 1. 3D-discrimination plots (64) for $P = \mathrm{B}(10, p)$, $Q = \mathrm{B}(10, q)$ with $0.2 \leq \alpha \leq 2$ and $0 \leq \beta \leq 1$.

Remark 1: The power divergences $D_\alpha(P, Q)$ are usually applied in the statistics as criteria of discrimination or goodness-of-fit between the distributions P and Q. The scaled Bregman distances $B_\alpha(P, Q \,|\, M)$ as generalizations of the power divergences $D_\alpha(P, Q) \equiv B_\alpha(P, Q \,|\, Q)$ allow to extend the 2D-discrimination plots $\{[D_\alpha(P, Q);\ \alpha] : c \leq \alpha \leq d\} \subset \mathbb{R}^2$ into more informative 3D-*discrimination plots*

$$\{[B_\alpha(P, Q \mid \beta P + (1-\beta)Q);\ \alpha;\ \beta] : c \leq \alpha,\, \beta \leq d\} \subset \mathbb{R}^3 \tag{64}$$

reducing to the former ones for $\beta = 0$. The simpler 2D-plots known under the name Q–Q-plots are famous tools for the exploratory data analysis. It is easy to consider that the computer-aided appropriately coloured projections of the 3D-plots (64) allow much more intimate insight into the relation between data and their statistical models. Therefore this computer-aided 3D-exploratory analysis deserves a deeper attention and research. The next example presents projections of two such plots obtained for a binomial model P and its data based binomial alternative Q.

Example 2: Let $P = \mathrm{B}(n, p)$ be a binomial distribution with parameters n, p (with a slight abuse of notation), and $Q = \mathrm{B}(n, q)$. Figure 1 presents projections of the corresponding 3D-discrimination plots (64) for $0.2 \leq \alpha \leq 2$ and $0 \leq \beta \leq 1$, where the Subfigure (a) used the parameter constellation $n = 10$, $p = 0.25$, $q = 0.20$ whereas the Subfigure (b) used $n = 10$, $p = 0.25$, $q = 0.30$. In both cases, the ranges of $B_\alpha(P, Q \mid \beta P + (1-\beta)Q)$ are subsets of the interval $[0.06, 0.088]$.

IV. EXPONENTIAL FAMILIES

In this section we show that the scaled Bregman power distances $B_\alpha(P, Q \,|\, M)$ can be *explicitly evaluated* for probability measures P, Q, M from exponential families. Let us restrict ourselves to the Euclidean observation spaces $(\mathcal{X}, \mathcal{A}) \subseteq (\mathbb{R}^d, \mathcal{B}^d)$ and denote by $x \cdot \theta$ the scalar product of $x, \theta \in \mathbb{R}^d$. The convex extended real valued function

$$b(\theta) = \ln \int_{\mathbb{R}^d} e^{x \cdot \theta} \mathrm{d}\lambda(x), \qquad \theta \in \mathbb{R}^d\,, \tag{65}$$

and the convex set

$$\Theta = \{\theta \in \mathbb{R}^d : b(\theta) < \infty\} \tag{66}$$

define on $(\mathcal{X}, \mathcal{A})$ an *exponential family of probability measures* $\{P_\theta : \theta \in \Theta\}$ with the densities

$$p_\theta(x) \equiv \frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda}(x) = \exp\{x \cdot \theta - b(\theta)\}, \quad x \in \mathbb{R}^d, \quad \theta \in \Theta. \tag{67}$$

The cumulant function $b(\theta)$ is infinitely differentiable on the interior $\mathring{\Theta}$ with the gradient

$$\triangledown b(\theta) = \left(\frac{\partial}{\partial\theta_1}, ..., \frac{\partial}{\partial\theta_d}\right) b(\theta), \quad \theta \in \mathring{\Theta}. \tag{68}$$

Note that (67) are exponential type densities in the *natural form*. All exponential type distributions such as Poisson, normal etc. can be transformed to into this form (cf. e.g. Brown (1986)).

The formula

$$\int_{\mathbb{R}^d} e^{x\cdot\theta}\, \mathrm{d}\lambda(x) \;=\; e^{b(\theta)}, \quad \theta \in \Theta \tag{69}$$

follows from (65) and implies

$$\int_{\mathbb{R}^d} x\, e^{x\cdot\theta}\, \mathrm{d}\lambda(x) \;=\; e^{b(\theta)} \nabla b(\theta), \quad \theta \in \mathring{\Theta}. \tag{70}$$

Both formulas (69) and (70) will be useful in the sequel.

We are interested in the scaled Bregman power distances

$$B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \quad \text{for } \theta_0,\, \theta_1,\, \theta_2 \in \Theta,\ \alpha \in \mathbb{R}. \tag{71}$$

Here $P_{\theta_1}, P_{\theta_2}, P_{\theta_0}$ are measure-theoretically equivalent probability measures, so that we can turn attention to the formulas (46), (34), (37), and (57) to (63), promising to reduce the evaluation of $B_\alpha(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0})$ to the evaluation of the power divergences $D_\alpha(P_{\theta_1}, P_{\theta_2})$. Therefore we first study these divergences and in particular verify their finiteness, which was a sufficient condition for the applicability of the formulas (46), (34) and (37).

Theorem 3: If $\alpha \in \mathbb{R}$ differs from 0 and 1, then the power divergence $D_\alpha(P_{\theta_1}, P_{\theta_2})$ is for all $\theta_1,\, \theta_2 \in \Theta$ finite and given by the expression

$$\frac{\exp\left\{b(\alpha\theta_1 + (1-\alpha)\,\theta_2) - \alpha b(\theta_1) - (1-\alpha)\, b(\theta_2)\right\} - 1}{\alpha(\alpha-1)}\,. \tag{72}$$

In particluar, it is invariant with respect to the shifts of the cumulant function linear in $\theta \in \Theta$ in the sense that it coincides with the power divergence $D_\alpha(\tilde{P}_{\theta_1}, \tilde{P}_{\theta_2})$ in the exponential family with the cumulant function $\tilde{b}(\theta) = b(\theta) + c + v \cdot \theta$ where c is a real number and v a $d-$vector.

Proof: As a slight extension of (43), put for arbitrary $\alpha \in \mathbb{R}$ and $\theta_1,\, \theta_2 \in \Theta$

$$\begin{aligned} \rho_\alpha(\theta_1, \theta_2) \;&=\; \ln \int_{\mathbb{R}^d} p_{\theta_1}^\alpha\, p_{\theta_2}^{1-\alpha}\, \mathrm{d}\lambda \\ &= \ln \int_{\mathbb{R}^d} \exp\Big\{\alpha[x\cdot\theta_1 - b(\theta_1)] + (1-\alpha)\,[x\cdot\theta_2 - b(\theta_2)]\Big\}\, \mathrm{d}\lambda(x) \\ &= \ln \frac{\int_{\mathbb{R}^d} \exp\{x\cdot[\alpha\theta_1 + (1-\alpha)\,\theta_2]\}\, \mathrm{d}\lambda(x)}{\exp\{\alpha b(\theta_1) + (1-\alpha)\, b(\theta_2)\}} \\ &= \ln \frac{\exp\{b(\alpha\theta_1 + (1-\alpha)\,\theta_2)\}}{\exp\{\alpha b(\theta_1) + (1-\alpha)\, b(\theta_2)\}} \qquad \text{(cf. (69))}. \end{aligned} \tag{73}$$

Hence

$$\rho_\alpha(\theta_1, \theta_2) = b\Big(\alpha\theta_1 + (1-\alpha)\,\theta_2\Big) - \alpha b(\theta_1) - (1-\alpha)\, b(\theta_2)\ , \tag{74}$$

where the right hand side is finite if $0 \le \alpha \le 1$. Furthermore, (43) implies for $\alpha \in \mathbb{R}\backslash\{0,1\}$

$$D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) = \frac{\exp \rho_\alpha(\theta_1, \theta_2) - 1}{\alpha(\alpha-1)}\ . \tag{75}$$

Thus, (72) follows from (74) and (75). The declared finiteness of $D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right)$ is immediately clear, also the invariance. □

The remaining power divergences $D_0(P_{\theta_1}, P_{\theta_2})$ and $D_1(P_{\theta_1}, P_{\theta_2})$ are evaluated in the next theorem.

Theorem 4: For all $\theta_1,\, \theta_2 \in \Theta$ and $\alpha \in \mathbb{R}$ different from 0 and 1 there holds

$$D_\alpha\left(P_{\theta_2}, P_{\theta_1}\right) \ = \ D_{1-\alpha}\left(P_{\theta_1}, P_{\theta_2}\right) \tag{76}$$

and for $\theta_2 \in \mathring{\Theta}$

$$D_{-\ln t}\left(P_{\theta_1}, P_{\theta_2}\right) = D_0\left(P_{\theta_1}, P_{\theta_2}\right) = \lim_{\alpha \downarrow 0} D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) \tag{77}$$

$$= b(\theta_1) - b(\theta_2) - \nabla b(\theta_2)\,(\theta_1 - \theta_2) \tag{78}$$

$$= \lim_{\alpha \uparrow 1} D_\alpha\left(P_{\theta_2}, P_{\theta_1}\right) = D_1\left(P_{\theta_2}, P_{\theta_1}\right) = D_{t \ln t}\left(P_{\theta_2}, P_{\theta_1}\right). \tag{79}$$

Proof: (a) Let $\alpha(\alpha-1) \neq 0$ and $\theta_1,\, \theta_2 \in \Theta$. By (9) and (41)

$$\phi^*_\alpha(t) = \frac{t^{1-\alpha} - t}{\alpha(\alpha-1)}\ .$$

Hence, from the definitions (10) and (42) one can see that $D_{\phi^*_\alpha}(P_{\theta_2}, P_{\theta_1})$ coincides with the power divergence $D_{1-\alpha}(P_{\theta_2}, P_{\theta_1})$. Therefore (76) follows from the relations

$$D_{1-\alpha}\left(P_{\theta_2}, P_{\theta_1}\right) \equiv D_{\phi^*_\alpha}\left(P_{\theta_2}, P_{\theta_1}\right)$$

$$= D_{\phi_\alpha}\left(P_{\theta_1}, P_{\theta_2}\right) \equiv D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) \quad \text{(cf. (18))}.$$

Alternatively, (76) follows from (75) using the skew symmetry

$$\rho_\alpha(\theta_1, \theta_2) = \rho_{1-\alpha}(\theta_2, \theta_1)$$

which is evident from (74).

(b) The equalities (77) and (79) follow from the already proved skew symmetry (76) and from the definition of the α-divergences of orders $\alpha = 0$ and $\alpha = 1$ in (51), (50). It remains to prove that the limit in (77) equals (78). For this, let us first observe that for every real valued function $\rho(\alpha)$ defined in the open set $(-\varepsilon, \varepsilon)\backslash\{0\}$ ($\varepsilon > 0$) it holds

$$\lim_{\alpha \to 0} \frac{e^{\rho(\alpha)} - 1}{\alpha(\alpha-1)} \ = \ -\lim_{\alpha \to 0} \frac{\rho(\alpha)}{\alpha}$$

in the sense that one of the limits exists if and only if the other does so, and then the two are equal. With the help of (75), for $\rho(\alpha) = \rho_\alpha(\theta_1, \theta_2)$ this is the equivalent to

$$\lim_{\alpha \to 0} \frac{D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right)}{\alpha(\alpha-1)} \ = \ -\lim_{\alpha \to 0} \frac{\rho_\alpha(\theta_1, \theta_2)}{\alpha}\ ,$$

and the proof is completed by the easy verification of the relation

$$
\begin{aligned}
&-\lim_{\alpha\to 0}\frac{\rho_\alpha(\theta_1,\,\theta_2)}{\alpha}\\
&\equiv \lim_{\alpha\to 0}\frac{\alpha\, b(\theta_1)+(1-\alpha)\, b(\theta_2)-b(\alpha\,\theta_1+(1-\alpha)\,\theta_2)}{\alpha}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(cf. (74))}\\
&= b(\theta_1)-b(\theta_2)+\nabla b(\theta_2)\,(\theta_2-\theta_1).
\end{aligned}
$$

for θ_2 from the interior $\mathring{\Theta}$. □

The main result of this section is the following representation theorem for Bregman distances in exponential families, where in addition to the functions $\rho_\alpha(\theta_1,\theta_2)$ of (74) we also use the functions $\sigma_\alpha(\theta_0,\theta_1,\theta_2)$ ($\alpha\in\mathbb{R}$, $\theta_0,\theta_1,\theta_2\in\Theta$) defined as the difference

$$
\sigma_\alpha(\theta_0,\,\theta_1,\,\theta_2)=\sigma_\alpha^{I}(\theta_0,\,\theta_1,\,\theta_2)-\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2) \tag{80}
$$

of the nonnegative (possibly infinite)

$$
\sigma_\alpha^{I}(\theta_0,\,\theta_1,\,\theta_2)=b\Big(\alpha\,\theta_1+(1-\alpha)\;[\theta_1-\theta_2+\theta_0]\Big) \tag{81}
$$

and the finite

$$
\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2)=\alpha\, b(\theta_1)+(1-\alpha)\,\Big[b(\theta_1)-b(\theta_2)+b(\theta_0)\Big]\;. \tag{82}
$$

Alternatively,

$$
\begin{aligned}
\sigma_\alpha(\theta_0,\,\theta_1,\,\theta_2) &= \rho_\alpha(\theta_1,\theta_0+\theta_1-\theta_2)\\
&\quad +(1-\alpha)\,[b(\theta_0+\theta_1-\,\theta_2)-b(\theta_0)-b(\theta_1)+b(\theta_2)]\,.
\end{aligned} \tag{83}
$$

Theorem 5: Let $\theta_0,\,\theta_1,\,\theta_2\in\Theta$ be arbitrary. If $\alpha(\alpha-1)\neq 0$ then the Bregman distance of the exponential family distributions P_{θ_1} and P_{θ_2} scaled by P_{θ_0} is given by the formula

$$
\begin{aligned}
&B_\alpha\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right)\\
&=\frac{\exp\rho_\alpha(\theta_1,\theta_0)}{\alpha(\alpha-1)}+\frac{\exp\rho_\alpha(\theta_2,\theta_0)}{\alpha}+\frac{\exp\sigma_\alpha(\theta_0,\theta_1,\theta_2)}{1-\alpha}.
\end{aligned} \tag{84}
$$

If θ_0 respectively θ_1 is from the interior $\mathring{\Theta}$, then the limiting Bregman power distances are

$$
\begin{aligned}
&B_0\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right)\\
&\qquad = \; b(\theta_1)-b(\theta_2)-\nabla b(\theta_0)\,(\theta_1-\theta_2)\\
&\qquad\quad +\exp\sigma_0(\theta_0,\theta_1,\theta_2)-1
\end{aligned} \tag{85}
$$

respectively

$$
B_1\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right)=b(\theta_2)-b(\theta_1)-\nabla b(\theta_1)\,(\theta_2-\theta_1)\;. \tag{86}
$$

In particluar, all scaled Bregman distances (84) - (86) are invariant with respect to the shifts of the cumulant function linear in $\theta\in\Theta$ in the sense that they coincide with the scaled Bregman distances $B_\alpha\left(\tilde{P}_{\theta_1},\tilde{P}_{\theta_2}\,|\,\tilde{P}_{\theta_0}\right)$ in the exponential family with the cumulant function $\tilde{b}(\theta)=b(\theta)+c+v\cdot\theta$ where c is a real number and v a $d-$vector.

Proof: (a) By (67) it holds for every $\alpha \in \mathbb{R}$ and $\theta_0, \theta_1, \theta_2 \in \Theta$

$$\begin{aligned}
&\left(\frac{p_{\theta_2}(x)}{p_{\theta_0}(x)}\right)^{\alpha-1} p_{\theta_1}(x) \\
&= \exp\Big\{(\alpha-1)\big[x\cdot(\theta_2-\theta_0)-(b(\theta_2)-b(\theta_0))\big] \\
&\qquad\qquad +x\cdot\theta_1-b(\theta_1)\Big\} \\
&= \exp\Big\{x\cdot\big(\alpha\,\theta_1+(1-\alpha)\;[\theta_1-\theta_2+\theta_0]\big) \\
&\qquad\qquad -\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2)\Big\}
\end{aligned}$$

with $\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2)$ from (82). Since (69) leads to

$$\begin{aligned}
&\int_{\mathbb{R}^d}\exp\Big\{x\cdot\Big(\alpha\,\theta_1+(1-\alpha)\;[\theta_1-\theta_2+\theta_0]\Big)\Big\}\,\mathrm{d}\lambda \\
&= \exp\sigma_\alpha^{I}(\theta_0,\,\theta_1,\,\theta_2)
\end{aligned}$$

for $\sigma_\alpha^{I}(\theta_0,\,\theta_1,\,\theta_2)$ given by (81), it holds

$$\int_{\mathcal{X}}\left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\,\mathrm{d}\lambda \;=\; \exp\sigma_\alpha(\theta_0,\theta_1,\theta_2) \tag{87}$$

where $\sigma_\alpha(\theta_0,\theta_1,\theta_2)$ was defined in (80). Now, by plugging

$$P=P_{\theta_1},\quad Q=P_{\theta_2},\quad M=P_{\theta_0}\quad \text{(cf. (67))}$$

in (46), we get for $\alpha(\alpha-1)\neq 0$ the Bregman distances

$$\begin{aligned}
&B_\alpha\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right) \\
&= D_\alpha\left(P_{\theta_1},P_{\theta_2}\right)-(1-\alpha)\,D_\alpha\left(P_{\theta_2},P_{\theta_0}\right) \\
&\qquad +\frac{1}{1-\alpha}\left[\int_{\mathcal{X}}\left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\,\mathrm{d}\lambda-1\right].
\end{aligned} \tag{88}$$

Applying the power divergence formula (75) together with (87) to (88), one obtains the desired formula (84).

(b) By the definition of $B_0(P,Q\,|\,M)$ in (58) and by (52)

$$\begin{aligned}
&B_0\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right) \\
&= D_0\left(P_{\theta_1},P_{\theta_0}\right)-D_0\left(P_{\theta_2},P_{\theta_0}\right)+\int_{\mathcal{X}}\frac{p_{\theta_0}\,p_{\theta_1}}{p_{\theta_2}}\,\mathrm{d}\lambda-1
\end{aligned}$$

where

$$\int_{\mathcal{X}}\frac{p_{\theta_0}\,p_{\theta_1}}{p_{\theta_2}}\,\mathrm{d}\lambda \;=\; \exp\sigma_0(\theta_0,\,\theta_1,\,\theta_2)\quad\text{(cf. (87)).}$$

For $\theta_0\in\mathring{\Theta}$ the desired assertion (85) follows from here and from the formulas

$$D_0\left(P_{\theta_i},\,P_{\theta_0}\right)\;=\;b(\theta_i)-b(\theta_0)-\nabla b(\theta_0)\,(\theta_i-\theta_0)\quad\text{for } i=1,2$$

obtained from (78).

(c) The desired formula (86) follows immediately from the definition (57) and from the formulas (55), (56), (78) and (79).

(d) The finally stated invariance is immediate. □

The Conclusion 1 of Section 3 about the relation between scaled Bregman distances and ϕ-divergences can be completed by the following relation between both of them and the classical Bregman distances (1).

Conclusion 2: Let $B_\phi(x,y)$ be the classical Bregman distance (1) of $x,y \in \mathbb{R}^d$ and $\mathcal{P} = \left\{P_\theta : \theta \in \mathbb{R}^d\right\}$ the exponential family with cumulant function ϕ, i.e. with densities $p_\theta(s) = \exp\{s \cdot \theta - \phi(\theta)\}$, $s \in \mathbb{R}^d$. Then for all $P_x, P_y, P_z \in \mathcal{P}$

$$B_\phi(x,y) = B_1(P_y, P_x|P_z) = D_1(P_y, P_x) \ ,$$

i.e. there is a one-to-one relation between the classical Bregman distance $B_\phi(x,y)$ and the scaled Bregman distances $B_1(P_y, P_x|P_z)$ and power divergences $D_1(P_y, P_x)$ of the exponential probability measures generated by the cumulant function ϕ. This means that the family $\left\{B_\alpha(P_y, P_x|P_z) : \alpha \in \mathbb{R},\ z \in \mathbb{R}^d\right\}$ of scaled Bregman power distances and the family $\{D_\alpha(P_y, P_x) : \alpha \in \mathbb{R}\}$ of power divergences extend the classical Bregman distances $B_\phi(x,y)$ to which they reduce at $\alpha = 1$ and arbitrary $P_z \in \mathcal{P}$. In fact, we meet here the extension of the classical Bregman distances in three different directions: the first represented by various power parameters $\alpha \in \mathbb{R}$, the second represented by various possible exponential distributions parametrized by $\theta \in \mathbb{R}^d$, and the third represented by the exponential distribution parameters $z \in \mathbb{R}^d$ which are relevant when $\alpha \neq 1$.

Remark 2: We see from Theorems 4 and 5 that – in consistency with (34), (56) – for arbitrary interior parameters $\theta_0,\, \theta_1,\, \theta_2 \in \mathring{\Theta}$

$$B_1\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) = D_1\left(P_{\theta_1}, P_{\theta_2}\right),$$

i. e. that the Bregman distance of order $\alpha = 1$ of exponential family distributions $P_{\theta_1},\, P_{\theta_2}$ does not depend on the scaling distribution P_{θ_0}. The distance of order $\alpha = 0$ satisfies the relation

$$\begin{aligned} &B_0\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) = D_0\left(P_{\theta_1}, P_{\theta_2}\right) + \exp \sigma_0(\theta_0,\, \theta_1,\, \theta_2) - 1 \\ &=\ B_1\left(P_{\theta_2}, P_{\theta_1} \,|\, P_{\theta_0}\right) + \Delta(\theta_0,\, \theta_1,\, \theta_2)\ , \end{aligned}$$

where

$$\Delta(\theta_0,\, \theta_1,\, \theta_2) = \exp \sigma_0(\theta_0,\, \theta_1,\, \theta_2) - 1$$

represents a deviation from the skew-symmetry of the Bregman distances $B_0\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right)$ and $B_1\left(P_{\theta_2}, P_{\theta_1} \,|\, P_{\theta_0}\right)$ of P_{θ_1} and P_{θ_2}. This deviation is zero if (for strictly convex $b(\theta)$ if and only if) $\theta_0 = \theta_2$.

Remark 3: We see from the formulas (72) – (86) that for all $\alpha \in \mathbb{R}$ the quantities $D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right)$, $\rho_\alpha(\theta_1,\, \theta_2)$, $\sigma_\alpha(\theta_0,\, \theta_1,\, \theta_2)$ and $B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right)$ only depend on the cumulant function $b(\theta)$ defined in (65), and *not* directly on the reference measure λ used in the definition formulas (65), (67).

V. EXPONENTIAL APPLICATIONS

In this section we illustrate the evaluation of scaled Bregman divergences $B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right)$ for discrete and continuous exponential families, and also for exponentially distributed random processes.

Binomial model: Consider for fixed $n \geq 2$ on the observation space $\mathcal{X} = \{0, ..., n\}$ the binomial probabilities

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} = \lambda(x) \exp\{x \cdot \theta - b(\theta)\} \tag{89}$$

where

$$\lambda(x) = \binom{n}{x}, \;\; \theta = \ln \frac{p}{1-p} \in \Theta = \mathbb{R} \;\; \text{and} \;\; b(\theta) = n \ln(1 + e^{\theta}).$$

After some calculations one obtains from (74) and (83)

$$\rho_\alpha(\theta_1, \theta_2) = n \ln \frac{1 + e^{\alpha\theta_1 + (1-\alpha)\theta_2}}{(1 + e^{\theta_1})^{\alpha} (1 + e^{\theta_2})^{1-\alpha}}$$

and

$$\sigma_\alpha(\theta_0, \theta_1, \theta_2) = n \ln \frac{\left(1 + e^{\theta_1 + (1-\alpha)(\theta_0 + \theta_1 - \theta_2)}\right) (1 + e^{\theta_2})^{1-\alpha}}{(1 + e^{\theta_0})^{\alpha} (1 + e^{\theta_1})}.$$

Applying Theorem 5 one obtains an explicit formula for the binomial Bregman distances $B_\alpha\left(P_{\theta_1}, P_{\theta_2} \mid P_{\theta_0}\right)$ from here.

Rayleigh model: An important role in communication theory play the Rayleigh distributions defined by the probability densities

$$p_\theta(x) = \theta x \exp\left\{-\frac{\theta x^2}{2}\right\}, \quad \theta \in \Theta = (0, \infty) \tag{90}$$

with respect to the restriction λ_+ of the Lebesgue measure λ on the observation space $\mathcal{X} = (0, \infty)$. The mapping

$$T(x) = -\sqrt{2x}$$

from the positive halfline $(0, \infty)$ to the negative halfline $(-\infty, 0)$ transforms (90) into the family of Rayleigh densities

$$p_\theta(x) = \theta \exp\left\{\theta x\right\} = \exp\left\{\theta x - b(\theta)\right\}$$

$$\text{for } b(\theta) = -\ln\theta, \;\; \theta > 0$$

with respect to the restriction λ_- of the Lebesgue measure λ on the observation space $\mathcal{X} = (-\infty, 0)$. These are the Rayleigh densities in the natural form assumed in (67). After some calculations one derives from (74)

$$\rho_\alpha(\theta_1, \theta_2) = \ln \frac{\theta_1^{\alpha} \, \theta_2^{1-\alpha}}{\alpha\theta_1 + (1-\alpha)\theta_2} \tag{91}$$

and

$$\sigma_\alpha(\theta_0, \theta_1, \theta_2) = \ln \frac{\theta_1 \;\, \theta_0^{1-\alpha}}{(\alpha\theta_1 + (1-\alpha)(\theta_0 + \theta_1 - \theta_2)) \;\; \theta_2^{1-\alpha}}. \tag{92}$$

Applying Theorem 5 one obtains the Rayleigh-Bregman distances $B_\alpha\left(P_{\theta_1}, P_{\theta_2} \mid P_{\theta_0}\right)$ from here.

Theorem 1 about the preservation of the scaled Bregman distances by statistically sufficient transformations is useful for the evaluation of these distances in exponential families. It implies for example that these distances in the normal and lognormal families coincide. The next two examples dealing with distances of stochastic processes make use of this theorem too.

Exponentialy distributed signals: Most of the random processes modelling physical, social and economic phenomena are exponentially distributed. Important among them are the real valued Lévy processes $\boldsymbol{X}_t = (X_s : 0 \leq s \leq t)$ with trajectories $\boldsymbol{x}_t = (x_s : 0 \leq s \leq t)$ from the Skorokchod observation spaces $(\mathcal{X}_t, \mathcal{A}_t)$ and parameters from the set

$$\Theta = \{\theta \in \mathbb{R} : c(\theta) < \infty\}$$

defined by means of the function

$$c(\theta) = \int_{\mathbb{R}\setminus\{0\}} x^2 e^{\theta x}/(1+x^2)\,\mathrm{d}\nu(x)$$

where ν is a Lévy measure which determines the probability distribution of the size of jumps of the process and the intensity with which jumps occur. It is assumed that 0 belongs to Θ and it is known (cf. e.g. Küchler and Sorensen (1994)) that the probability distributions $P_{t,\theta}$ induced by these processes on $(\mathcal{X}_t, \mathcal{A}_t)$ are mutually measure-theoretically equivalent with the relative densities

$$\frac{\mathrm{d}P_{t,\theta}}{\mathrm{d}P_{t,0}}(\boldsymbol{x}_t) = \exp\{\theta\; x_t - b_t(\theta)\} \tag{93}$$

for the end x_t of the trajectory $\boldsymbol{x}_t$. The cumulant function appearing here is

$$b_t(\theta) = t\left(\delta\theta + \frac{1}{2}\sigma^2\theta^2 + \gamma(\theta)\right) \tag{94}$$

for two genuine parameters $\delta \in \mathbb{R}$ respectively $\sigma > 0$ of the process which determine its intensity of drift respectively its volatility, and for the function

$$\gamma(\theta) = \int_{\mathbb{R}\setminus\{0\}} [e^{\theta x} - 1 - \theta x/(1+x^2)]\,\mathrm{d}\nu(x).$$

The formula (93) implies that the family $\mathcal{P}_t = \{P_{t,\theta} : \theta \in \Theta\}$ is exponential on $(\mathcal{X}_t, \mathcal{A}_t)$ for which the "extremally reduced" observation $T(\boldsymbol{x}_t) = x_t$ is statistically sufficient. Thus, by Theorem 1,

$$B(P_{t,\theta_1}, P_{t,\theta_2}|P_{t,0}) = B(Q_{t,\theta_1}, Q_{t,\theta_2}|Q_{t,0}) \tag{95}$$

where $Q_{t,\theta}$ is a probability distribution on the real line governing the marginal distribution of the last observed value X_t of the process $\boldsymbol{X}_t$.

Queueing processes and Brownian motions: For illustration of the general result of the previous subsection we can take the family of *Poisson processes* with initial value $X_0 = 0$ and intensities $\eta = e^\theta,\ \theta \in \Theta = \mathbb{R}$ for which $\delta = \sigma = 0$ and $c(\theta) = e^\theta - 1$ so that $b_t(\theta) = t\left(e^\theta - 1\right)$. Then $Q_{t,\theta}$ is the Poisson distribution $\mathrm{Po}(\tau)$ with parameter $\tau = t\eta = te^\theta$ and probabilities

$$\frac{e^{-\tau}\,(\tau)^x}{x!} = \lambda(x)\,\exp\{x\vartheta - e^\vartheta\}$$

$$\text{for}\quad \vartheta = \ln\tau = \theta + \ln t, \quad \lambda(x) = \frac{1}{x!}\ .$$

The exponential structure is similar as above, so that by applying (74) to the cumulant function $b(\vartheta) = e^\vartheta = te^\theta$ we get for the Poisson processes with parameters θ_1 and θ_2

$$\rho_\alpha(\theta_1, \theta_2) = t\left[e^{\alpha\theta_1 + (1-\alpha)\theta_2} - \alpha e^{\theta_1} - (1-\alpha)e^{\theta_2}\right].$$

Combining this with (83) and Theorem 5 we obtain an explicit formula for the scaled Bregman distance (95) of these Poisson processes.

To give another illustration of the result of the previous subsection, let us first introduce the standard Wiener process $\widetilde{X}_t$ which is the Lévy process with $\nu \equiv 0$, $\delta = 0$, $\sigma = 1$ and $\theta = 1$. It defines the *family of Wiener processes*

$$X_s = \theta \, \widetilde{X}_s, \quad 0 \le s \le t, \quad \theta \in (0, \infty), \tag{96}$$

which are Lévy processes with $\delta = 0$, $\sigma = 1$ and $c(\theta) \equiv 0$ so that (94) implies $b_t(\theta) = \theta^2/2$. They are well-known models of the random fluctuations called Brownian motions. If the initial value X_0 is zero then $Q_{t,\theta}$ is the normal distribution with mean zero and variance $v^2 = t\theta^2$. The corresponding Lebesgue densities

$$\frac{1}{\sqrt{2\pi v^2}} \exp\left\{-\frac{x^2}{2v^2}\right\} = \sqrt{\frac{\vartheta}{\pi}} \exp\left\{-\vartheta x^2\right\} \quad \text{for } \vartheta = \frac{1}{2v^2}$$

are transformed by the mapping $x \longmapsto -\sqrt{|x|}$ of $\mathbb{R}$ on the negative halfline $(-\infty, 0)$ into the natural exponential densities $\exp\{\vartheta x - b(\vartheta)\}$ with respect to the dominating density $1/\sqrt{\pi|x|}$ where $b(\vartheta) = -\frac{1}{2}\ln\vartheta = -\ln\frac{1}{\theta} + \frac{1}{2}\ln 2t$. Thus by (74)

$$\rho_\alpha(\theta_1, \theta_2) = -\ln \frac{\theta_1^\alpha \, \theta_2^{1-\alpha}}{\alpha\theta_1 + (1-\alpha)\theta_2} \qquad (cf.\ (91)).$$

This together with (83) and Theorem 5 leads to the explicit formula for the scaled Bregman distance (95) of the Wiener processes under consideration.

Geometric Brownian motions: From the abovementioned standard Wiener process one can also build up the *family of geometric Brownian motions* (geometric Wiener processes)

$$Y_s = \exp\{\sigma \widetilde{X}_s + \theta s\}, \quad 0 \le s \le t, \quad \theta \in \mathbb{R}, \tag{97}$$

where the family-generating θ can be interpreted as drift parameters, and the volatility parameter $\sigma > 0$ is assumed to be constant all over the family. Then, $\sigma \widetilde{X}_t + \theta t$ is normally distributed with mean $m = \theta t$ and variance $v^2 = \sigma^2 t$, and Y_t is lognormally distributed with the same parameters m and v^2. By (95), the scaled Bregman distance of two geometric Brownian motions with parameters θ_1, θ_2 reduces to the scaled Bregman distance of two lognormal distributions LN$(\theta_1 t, \sigma^2 t)$, LN$(\theta_2 t, \sigma^2 t)$. As said above, it coincides with the scaled Bregman distance of two normal distributions N$(\theta_1 t, \sigma^2 t)$, N$(\theta_2 t, \sigma^2 t)$. This is seen also from the fact that the reparametrization

$$\vartheta = \frac{\mu}{v^2}, \quad \tau = \frac{1}{2v^2}$$

and transformations $\mathbb{R} \longmapsto \mathbb{R}^2$ similar to that from the previous example lead in both distributions N(μ, v^2) and LN(μ, v^2) to the same natural exponential density

$$p_{\vartheta,\tau}(x_1, x_2) = \exp\{x_1\vartheta + x_2\tau - b(\vartheta, \tau)\}$$

with

$$b(\vartheta, \tau) = \frac{1}{2}\ln\tau + \frac{\vartheta^2}{4\tau}.$$

These two distributions differ just in the dominating measures on the transformed observation space $\mathcal{X} = \mathbb{R}^2$. For $(\mu_1, v_1^2) = (\theta_1 t, \sigma^2 t)$ and $(\mu_2, v_2^2) = (\theta_2 t, \sigma^2 t)$ we get

$$(\vartheta_1, \tau_1) = \left(\frac{\theta_1}{\sigma^2}, \frac{1}{2\sigma^2 t}\right) \quad \text{and} \quad (\vartheta_2, \tau_2) = \left(\frac{\theta_2}{\sigma^2}, \frac{1}{2\sigma^2 t}\right)$$

and thus

$$
\begin{aligned}
&b(\alpha(\vartheta_1,\tau_1)+(1-\alpha)(\vartheta_2,\tau_2))-\alpha b(\vartheta_1,\tau_1)-(1-\alpha)b(\vartheta_2,\tau_2)\\
&=\frac{(\alpha\theta_1+(1-\alpha)\theta_2)^2-\alpha\theta_1^2+(1-\alpha)\theta_2^2}{2\sigma^2}\ t\ .
\end{aligned}
$$

Hence, for distributions P_{t,θ_1}, P_{t,θ_2} of the geometric Brownian motions considered above we get from (74)

$$
\rho_\alpha(\theta_1,\theta_2)=\frac{\left[(\alpha\theta_1+(1-\alpha)\theta_2)^2-\alpha\theta_1^2+(1-\alpha)\theta_2^2\right]}{2\sigma^2}\ t\ .
$$

The expression (83) can be automatically evaluated using this. Applying both these results in Theorem 5 one obtains explicit formula for the scaled Bregman distance (95) of these geometric Brownian motions.

References

Amari S.-I. (2007), "Integration of stochastic models by minimizing α-divergence," *Neural Computation*, vol. 19, no. 10, pp. 2780-2796.

Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005), "Clustering with Bregman divergences," *J. Machine Learning Research*, vol. 6, pp. 1705-1749.

Bartlett, P.L., Jordan M.I. and McAuliffe, J.D. (2006), "Convexity, classification and risk bounds," *JASA*, vol. 101,pp. 138-156.

Boratynska, A. (1997), "Stability of Bayesian inference in exponential families," *Statist. & Probab. Letters*, vol. 36, pp. 173-178.

Bregman, L.M. (1967), "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200-217.

Brown, L.D. (1986), *Fundamentals of Statistical Exponential Families.* Hayward, California: Inst. of Math. Statistics.

Byrne, C. (1999), "Iterative projection onto convex sets using multiple Bregman distances," *Inverse Problems*, vol. 15, pp. 1295-1313.

Carlson, B.A. and Clements, M.A. (1991), "A computationally compact divergence measure for speech processing. *IEEE Transactions on PAMI*, vol. 13, pp. 1255-1260.

Cesa-Bianchi, N. and Lugosi, G. (2006), *Prediction, Learning, Games.* Cambridge: Cambridge University Press.

Collins, M., Schapire, R.E. and Singer, Y. (2002), "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, vol. 48, pp. 253-285.

Csiszár, I. (1963), "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci., ser. A*, vol. 8, pp. 85-108.

Csiszár, I. (1967), "Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, vol. 2, pp. 299-318.

Csiszár, I. (1991), "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals of Statistics*, vol. 19, no. 4, pp. 2032-2066.

Csiszár, I. (1994), "Maximum entropy and related methods," *Trans. 12th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes.* Prague, Czech Acad. Sci., pp. 58-62.

Csiszár, I. (1995), "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, pp. 161-186.

Csiszár, I. and Matúš, F. (2008), "On minimization of entropy functionals under moment constraints," *Proceedings of ISIT 2008,* Toronto, Canada, pp. 2101-2105.

Csiszár, I. and Matúš, F. (2009), "On minimization of multivariate entropy functionals," *Proceedings of ITW 2009,* Volos, Greece, pp. 96-100.

Do, M.N. and Vetterli, M. (2002), "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, pp. 146-158.

Freund, Y. and Schapire, R.E. (1997), "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119-139.

Halmos, P.R. (1964), *Measure Theory*. New York: D. Van Nostrand.

Hertz, T., Bar-Hillel, A. and Weinshall, D. (2004), "Learning distance functions for information retrieval," in *Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Rec. CVPR*, vol. 2, II-570 - II-577.

Jones, L.K. and Byrne, C.L. (1990), "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Trans. Inform. Theory* vol. 36, no. 1, pp. 23-30.

Küchler, U. and Sorensen, M. (1994), "Exponential families of stochastic processes and Lévy processes," *J. of Statist. Planning and Inference*, vol. 39, pp. 211-237.

Lehman, E.L. and Romano J.P. (2005), *Testing Statistical Hypotheses.* Berlin: Springer.

Liese, F. and Vajda, I. (1987), *Convex Statistical Distances.* Leipzig: Teubner.

Liese, F. and Vajda, I. (2006), "On divergences and informations in statistics and information theory," *IEEE Transaction on Information theory*, vol. 52, no. 10, pp. 4394-4412.

Lloyd, S.P. (1982), "Least squares quantization in PCM," *IEEE Transactions on Inform. Theory*, vol. 28, no. 2, pp. 129-137.

Marquina, A. and Osher, S.J. (2008), "Image super-resolution by TV-regularization and Bregman iteration," *J. Sci. Comput.*, vol. 37, pp. 367-382.

Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004), "Information geometry of $\mathcal{U}$-Boost and Bregman divergence," *Neural Computation*, vol. 16, no. 7, pp. 1437-1481.

Nock, R. and Nielsen, F. (2009), "Bregman divergences and surrogates for learning," *IEEE Transactions on PAMI*, vol. 31, no. 11, pp. 2048 - 2059.

Pardo, M.C. and Vajda, I. (1997), "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Transaction on Information theory*, vol. 43, no. 4, pp. 1288-1293.

Pardo, M.C. and Vajda, I. (2003), "On asymptotic properties of information-theoretic divergences," *IEEE Transaction on Information theory*, vol. 49, no. 7, pp. 1860-1868.

Rényi, A. (1961), "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. Probab.*, vol. 1, pp. 547-561. Berkeley, CA: Univ. of California Press.

Resmerita, E. and Anderssen R.S. (2007), "Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems," *Math. Meth. Appl. Sci.*, vol. 30, no. 13, pp. 1527-1544.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M. and Lenzen, F. (2008); *Variational methods in imaging.* New York: Springer.

Stummer, W. (2007), "Some Bregman distances between financial diffusion processes," *Proc. Appl. Math. Mech.*, vol. 7, no. 1, pp. 1050503 - 1050504.

Stummer, W. and Vajda, I. (2009), "On divergences of finite measures and their applicability in statistics and information theory," *Statistics*, vol. 43 (in print).

Teboulle, M. (2007), “A unified continuous optimization framework for center-based clustering methiods,” *Journal of Machine Learning Research*, vol. 8, pp. 65-102.

Vajda, I. (2009), “ On metric divergences of probability measures,” *Kybernetika*, vol. 45, no. 5 (in print).

Vajda, I. and Zvárová, J. (2007), “On generalized entropies, Bayesian decisions and statistical diversity,” *Kybernetika*, vol. 43, no. 5, pp. 675-696.

Veldhuis, R.N.J. (2002), “The centroid of the Kullback-Leibler distance,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 96-99.

Xu, J. and Osher, S. (2007), “Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising,” *IEEE Transaction on Image Processing*, vol. 16, no. 2, pp. 534-544.