On Bregman Distances and Divergences of Probability Measures

Wolfgang Stummer and Igor Vajda, *Fellow, IEEE*

***Abstract*—This paper introduces scaled Bregman distances of probability distributions which admit nonuniform 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 3-D 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, classification, divergences, exponential distributions, exponential processes, information retrieval, machine learning, statistical decision, sufficiency.**

I. Introduction

BREGMAN [7] 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 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 $p = (p_1, \ldots, p_d)$, $q = (q_1, \ldots, q_d)$ 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., [5], [20], and [21] for adjacent random projection studies). 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

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

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 [17]–[19] and Pardo and Vajda [36], [37] 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

$$D_\phi(p,q) = \sum_{i=1}^{d} q_i \phi\left(\frac{p_i}{q_i}\right) \tag{4}$$

for functions ϕ which are convex on $[0,\infty)$, continuous on $(0,\infty)$ and strictly convex at 1 with $\phi(1) = 0$. Originating in [15], they share some properties with the Bregman distances (2), e.g., they are pseudodistances too. For example, the previously considered functions $\phi(t) = (t-1)^2$ and $\phi(t) = t\ln t$ lead, in this case, to the classical Pearson divergence

$$D_\phi(p,q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{q_i} \tag{5}$$

Manuscript received October 27, 2009; revised August 04, 2011; accepted October 04, 2011. Date of current version February 29, 2012. This work was supported by MŠMT Grant 1M0572 and the GAČR Grant 102/07/1131.

W. Stummer is with the Department of Mathematics and the School of Business and Economics, University of Erlangen–Nürnberg, 91058 Erlangen, Germany (e-mail: stummer@mi.uni-erlangen.de).

I. Vajda, deceased, was with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 18208 Prague, Czech Republic.

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2011.2178139

and the previously 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, the reader is referred to [44]).

However, there also exist some sharp differences between these two types of pseudodistances of distributions. One distinguishing property of Bregman distances is that their use as loss criterion induces the conditional expectation as outcoming unique optimal predictor from given data (cf., [2]); this is, for instance, used in [3] for designing generalizations of the k-means algorithm which deals with the special case of squared Euclidean error (3) (cf., the seminal work of Lloyd [32] reprinting a Technical Report of Bell Laboratories dated by 1957). These features are generally not shared by those of the ϕ-divergences which are not Bregman distances, e.g., by the Pearson divergence (5). On the other hand, a distinguishing 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 ([16], see in this respect also [31]). 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 $\tilde{p} = (1/2, 1/2)$ 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 the following.

Ai) *information retrieval* (see, e.g., [22] and [25]),

Aii) *optimal decision* (for *general decision*, see, e.g., [4], [6], [23], and [45]; for *speech processing*, see, e.g., [11] and [46]; for *image processing*, see, e.g., [33], [40], and [47]),

Aiii) *machine learning* (see, e.g., [1], [3], [28], [35], and [43]).

Aiv) *parallel optimization and computing* (see, e.g., [12]).

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 (4) by introducing the scaled Bregman distances which for the discrete setup reduce to

$$B_\phi(p,q|m) = \sum_{i=1}^{d} \Big[\phi(p_i/m_i) - \phi(q_i/m_i) - \phi'_+(q_i/m_i)(p_i/m_i - q_i/m_i)\Big] m_i \quad (6)$$

for arbitrary finite scale vectors $m = (m_1, \ldots, m_d)$, convex functions ϕ and right-hand derivatives ϕ'_+. Obviously, the uniform scales $m = (1, \ldots, 1)$ lead to the Bregman distances (2) and the probability distribution scales $m = q = (q_1, \ldots, q_d)$ lead to the ϕ-divergences (4). 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 nondiscrete 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 this paper, $\mathfrak{M}$ denotes the space of all finite measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and $\mathfrak{P} \subset \mathfrak{M}$ 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

$$p = \frac{dP}{d\lambda}, \quad q = \frac{dQ}{d\lambda}, \quad \text{and} \quad m = \frac{dM}{d\lambda} \quad (7)$$

have a common support which will be identified with $\mathcal{X}$ (i.e., the densities (7) are positive on $\mathcal{X}$). Unless otherwise explicitly stated, it is assumed that $P, Q \in \mathfrak{P}$, $M \in \mathfrak{M}$ and that $\phi : (0,\infty) \mapsto \mathbb{R}$ is a continuous and convex function. It is known that 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

$$\phi^*(t) = t\phi(1/t) \quad (8)$$

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 $P \in \mathfrak{P}$ and $M \in \mathfrak{M}$, we consider

$$D_\phi(P,M) = \int_{\mathcal{X}} \phi\left(\frac{p}{m}\right) dM = \int_{\mathcal{X}} m\,\phi\left(\frac{p}{m}\right) d\lambda \quad \text{(cf., (7))} \quad (9)$$

generated by the same convex functions as considered in the formula (4) for discrete P and M. An important special case is $D_\phi(P,Q)$ with $Q \in \mathfrak{P}$.

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

$$\phi'_+(1)(p-m) \le m\,\phi\left(\frac{p}{m}\right) \le m\,\phi(0) + p\,\phi^*(0) \quad (10)$$

on the integrand, leading to the ϕ-divergence bounds

$$\phi'_+(1)(1 - M(\mathcal{X})) \le D_\phi(P,M) \le M(\mathcal{X})\,\phi(0) + \phi^*(0). \quad (11)$$

The integrand bounds (10) follow by putting $s = 1$ and $t = p/m$ in the inequality

$$\phi(s) + \phi'_+(s)(t-s) \le \phi(t) \le \phi(0) + t\phi^*(0) \tag{12}$$

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

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

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 of order 1 in the sense $\psi(tp, tm) = t\psi(p, m)$ for all $t > 0$, the divergences (9) 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 (11) take on the form

$$0 \le D_\phi(P, Q) \le \phi(0) + \phi^*(0) \tag{13}$$

and the equalities are achieved under well-known conditions (cf., [30] and [31]): 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 (11), which extends the left-hand inequality in (13) 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 [42]).

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

$$M(\mathcal{X})\,\phi\left(\frac{1}{M(\mathcal{X})}\right) \le D_\phi(P, M) \tag{14}$$

where the equality holds if

$$p = \frac{m}{M(\mathcal{X})}\ P\text{-a.s.} \tag{15}$$

If $D_\phi(P, M) < \infty$ and $\phi(t)$ is strictly convex at $t = 1/M(\mathcal{X})$, the equality in (14) holds if and only if (15) holds.

Proof: By (9) and the definition (8) of the convex function ϕ^*

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

Hence, by Jensen's inequality

$$D_\phi(P, M) \ge \phi^*\left(\int_{\mathcal{X}} \frac{m}{p}\, dP\right) = \phi^*(M(\mathcal{X})) \tag{16}$$

which proves the desired inequality (14). Since

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

is the condition for equality in (16), 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 (9) the "skew symmetry" ϕ-divergence formula

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

as well as the sufficiency of the condition

$$\phi(t) - \phi^*(t) \equiv \text{constant} \cdot (t - 1) \tag{17}$$

for the ϕ-divergence symmetry

$$D_\phi(P, Q) = D_\phi(Q, P) \text{ for all } P, Q. \tag{18}$$

Liese and Vajda [30] proved that under the assumed strict convexity of $\phi(t)$ at $t = 1$ the condition (17) is not only *sufficient* but also *necessary* for the symmetry (18).

III. Scaled Bregman Distances

Let us now introduce the basic concept this paper, which is a measure-theoretic version of the Bregman distance (6). In this definition, it is assumed that ϕ is a finite convex function in the domain $t > 0$, continuously extended to $t = 0$. As previously, $\phi'_+(t)$ denotes the right-hand derivative which for such $\phi(t)$ exists and p, q, m are the densities defined in (7).

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

$$\begin{aligned} &B_\phi\,(P, Q \,|\, M) \\ &= \int_{\mathcal{X}} \left[\phi\left(\frac{p}{m}\right) - \phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)\left(\frac{p}{m} - \frac{q}{m}\right)\right] dM \\ &= \int_{\mathcal{X}} \left[m\phi\left(\frac{p}{m}\right) - m\phi\left(\frac{q}{m}\right) - \phi'_+\left(\frac{q}{m}\right)(p - q)\right] d\lambda. \end{aligned} \tag{19}$$

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

Remark 1:

1) By putting $t = p/m$ and $s = q/m$ in (12), we find the argument of the integral in (19) to be nonnegative. Hence, the Bregman distance $B_\phi\,(P, Q \,|\, M)$ is well defined by (19) and is always nonnegative (possibly infinite).
2) Notice that the integrand in the first (respectively, second) integral of (19) constitutes a function, say, $\widetilde{\Upsilon}(p, q, m)$ (respectively, $\Upsilon(p, q, m)$) which is homogeneous of order 0 (respectively, order 1), i.e., for all $t > 0$, there holds $\widetilde{\Upsilon}(tp, tq, tm) = \widetilde{\Upsilon}(p, q, m)$ (respectively, $\Upsilon(tp, tq, tm) = t \cdot \Upsilon(p, q, m)$). Analogously, as already partially indicated earlier, the integrand in the first (respectively, second) integral of (9) is also a function, say, $\widetilde{\psi}(p, m)$ (respectively,

$\psi(p,m)$) which is homogeneous of order 0 (respectively, order 1).

3) In our *measure-theoretic* context (19), we have incorporated the possible nondifferentiability of ϕ by using its right-hand derivative, which will be essential at several places below. For general *Banach spaces*, one typically employs various directional derivatives (see, e.g., [9] in connection with different types of convexity properties).

The special scaled Bregman distances $B_\phi(P,Q\,|\,M)$ for probability scales $M \in \mathfrak{P}$ were introduced by Stummer [41]. Let us mention some other important previously considered special cases.

1) For $\mathcal{X}$ finite or countable and counting measure $M = \lambda$, some authors were already cited earlier in connection with the formula (2) and the research areas **(Ai)–(Aiii)**. In addition to them, one can also mention [10], [13], [14], and [34].
2) For open Euclidean set $\mathcal{X}$ and Lebesgue measure $M = \lambda$ on it, one can mention [26], as well as [39].

In the rest of this paper, we restrict ourselves to the Bregman distances $B_\phi(P,Q\,|\,M)$ scaled by finite measures $M \in \mathcal{M}$ and to the same class of convex functions as considered in the ϕ-divergence formulas (4) and (9). By using the remark after Definition 1 and applying (12), we get

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

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

$$B_\phi(P,Q\,|\,M) = D_\phi(P,M) - D_\phi(Q,M) - \int_{\mathcal{X}} \phi'_+\left(\frac{q}{m}\right)d\lambda \tag{20}$$

if at least two of the right-hand side expressions are finite [which can be checked, e.g., by using (11) or (14)].

The formula (19) simplifies in the important special cases $M = P$ and $M = Q$. In the first case, due to $\phi(1) = 0$, it reduces to

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

where the difference (21) is meaningful if and only if $D_\phi(Q,P) \equiv D_{\phi^*}(P,Q)$ is finite. The nonnegative divergence measure $\mathcal{B}_\phi(P,Q) := B_\phi(P,Q\,|\,P)$ is, thus, the difference between the nonnegative dissimilarity measure

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

and the nonnegative ϕ-divergence $D_\phi(Q,P)$. Furthermore, in the second special case $M = Q$, the formula (19) leads to the equality

$$B_\phi(P,Q\,|\,Q) = D_\phi(P,Q) \tag{22}$$

without any restriction on $P,\,Q \in \mathfrak{P}$ as realized already by Stummer [41].

Conclusion 1: Equality (22)—together with the fact that $B_\phi(P,Q\,|\,M)$ depends in general on M (see, e.g., Section III-B)—shows that the concept of scaled Bregman distance (19) strictly generalizes the concept of ϕ-divergence $D_\phi(P,Q)$ of probability measures P, Q.

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

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

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

As demonstrated by an example in Section I, measurable transformations (statistics)

$$T : (\mathcal{X},\mathcal{A}) \mapsto (\mathcal{Y},\mathcal{B}) \tag{23}$$

which are *not* sufficient for the pair $\{P,Q\}$ can increase those of the scaled Bregman distances $B_\phi(P,Q\,|\,M)$ which are not ϕ-divergences. On the other hand, the transformations (23) which *are* sufficient for the pair $\{P,Q\}$ need not preserve these distances either. Next, we formulate conditions under which the scaled Bregman distances $B_\phi(P,Q\,|\,M)$ are preserved by transformations of observations.

Definition 2: We say that the transformation (23) 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

$$\begin{aligned} &p(x) = g_P(Tx)h(x),\; q(x) = g_Q(Tx)h(x) \\ &\text{and } m(x) = g_M(Tx)h(x). \end{aligned} \tag{24}$$

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 [29, pp. 18–19]). All transformations (23) 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 (23) sufficient for the triplet $\{P,Q,\,M\}$ preserve the scaled Bregman distances in the sense that

$$B_\phi\left(PT^{-1}, QT^{-1}\,|\,MT^{-1}\right) = B_\phi(P,Q\,|\,M). \tag{25}$$

Proof: By (19) and (24), the right-hand side of (25) is equal to

$$\int_{\mathcal{X}} [\phi_{P,M}(Tx) - \phi_{Q,M}(Tx) - \Delta_{P,Q,M}(Tx)]\,dM \tag{26}$$

for

$$\phi_{P,M}(y) = \phi\left(\frac{g_P(y)}{g_M(y)}\right),\; \phi_{Q,M}(y) = \phi\left(\frac{g_Q(y)}{g_M(y)}\right) \tag{27}$$

and

$$\Delta_{P,Q,M}(y) = \phi'_+\left(\frac{g_Q(y)}{g_M(y)}\right)(g_P(y) - g_Q(y)). \tag{28}$$

From [24, Sec. 39, Th. D], the integral (26) is equal to

$$\int_{\mathcal{Y}} [\phi_{P,M}(y) - \phi_{Q,M}(y) - \Delta_{P,Q,M}(y)]\, dMT^{-1} \tag{29}$$

and, moreover

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

and similarly for Q instead of P. Therefore

$$\frac{dPT^{-1}}{d\lambda T^{-1}} = g_P(y)\, h(T^{-1}y) \text{ and } \frac{dQT^{-1}}{d\lambda T^{-1}} = g_Q(y)\, h(T^{-1}y)$$

which together with (19), (27), and (28) implies that the integral (29) is nothing but the left-hand side of (25). This completes the proof. □

Remark 2: Notice that by means of Remark 1(2) after Definition 1, the assertion of Theorem 1 can be principally related to the preservation of ϕ-divergences by transformations which are sufficient for the pair $\{P, Q\}$.

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 (19) implies

$$\begin{aligned} &B_{t\ln t}(P,Q\,|\,M) \\ &= \int_{\mathcal{X}} \left[p\ln\frac{p}{m} - q\ln\frac{q}{m} - \left(\ln\frac{q}{m} + 1\right)(p-q)\right] d\lambda \\ &= \int_{\mathcal{X}} \left[p\ln\frac{p}{m} - p\ln\frac{q}{m}\right] d\lambda \\ &= \int_{\mathcal{X}} p\ln\frac{p}{q}\, d\lambda = D_{t\ln t}(P,Q). \end{aligned} \tag{30}$$

Thus, for $\phi(t) = t\ln t$, the Bregman distance $B_\phi(P,Q\,|\,M)$ 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) $D_{t\ln t}(P,Q)$ (cf., [41]). As a side effect, this independence gives also rise to examples for the conclusion that the validity of (25) does generally not imply that T is sufficient for the triplet $\{P, Q, M\}$.

B. Bregman Reversed Logarithmic Distance

Let now $\phi(t) = -\ln t$ so that $\phi'(t) = -1/t$. Then, (19) implies

$$\begin{aligned} &B_{-\ln t}(P,Q\,|\,M) \\ &= \int_{\mathcal{X}} \left[m\ln\frac{m}{p} - m\ln\frac{m}{q} + \frac{m}{q}(p-q)\right] d\lambda \end{aligned} \tag{31}$$

$$= D_{t\ln t}(M,P) - D_{t\ln t}(M,Q) + \int_{\mathcal{X}} \frac{mp}{q}\, d\lambda - M(\mathcal{X}) \tag{32}$$

$$= D_{-\ln t}(P,M) - D_{-\ln t}(Q,M) + \int_{\mathcal{X}} \frac{mp}{q} d\lambda - M(\mathcal{X}) \tag{33}$$

where the equalities (32) and (33) hold if at least two out of the first three expressions on the right-hand side are finite. In particular, (31) implies [consistent with (22)]

$$B_{-\ln t}(P,Q\,|\,Q) = D_{-\ln t}(P,Q) \tag{34}$$

and (32) implies for $D_{t\ln t}(P,Q) < \infty$ (consistent with (21))

$$B_{-\ln t}(P,Q\,|\,P) = \chi^2(P,Q) - D_{t\ln t}(P,Q) \tag{35}$$

where

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

is the well-known Pearson information divergence. From (34) and (35), 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 section, we restrict ourselves for simplicity to probability measures $M \in \mathfrak{P}$, i.e., we suppose $M(\mathcal{X}) = 1$. Under this assumption, we investigate the scaled Bregman distances

$$B_\alpha(P,Q\,|\,M) = B_{\phi_\alpha}(P,Q\,|\,M), \quad \alpha \in \mathbb{R},\ \alpha \neq 0,\ \alpha \neq 1 \tag{36}$$

for the family of power convex functions

$$\phi(t) \equiv \phi_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha-1)} \text{ with } \phi'_\alpha(t) = \frac{t^{\alpha-1}}{\alpha - 1}. \tag{37}$$

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

$$\begin{aligned} &D_\alpha(P,M) = D_{\phi_\alpha}(P,M) \\ &= \frac{1}{\alpha(\alpha-1)}\left[\int_{\mathcal{X}} p^\alpha\, m^{1-\alpha}\, d\lambda - 1\right] \\ &= \frac{\exp\rho_\alpha(P,M) - 1}{\alpha(\alpha-1)} \text{ with } \rho_\alpha(P,M) = \ln\int_{\mathcal{X}} p^\alpha\, m^{1-\alpha}\, d\lambda \end{aligned} \tag{38}$$

of real powers α different from 0 and 1, studied for arbitrary probability measures P, M in [30]. 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 [30] 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 [38].

Returning now to the Bregman power distances, observe that if $D_\alpha(P,M) + D_\alpha(Q,M)$ is finite, then (20), (36), and (37) imply for $\alpha \neq 0$, $\alpha \neq 1$

$$\begin{aligned} & B_\alpha(P,Q\,|\,M) \\ &= -D_\alpha(Q,M) - \frac{1}{\alpha-1}\int_{\mathcal{X}} \left(\frac{q}{m}\right)^{\alpha-1}(p-q)\,d\lambda \\ &= D_\alpha(P,M) - D_\alpha(Q,M) \\ &\qquad - \frac{1}{\alpha-1}\int_{\mathcal{X}} \left[\left(\frac{q}{m}\right)^{\alpha-1} p - \left(\frac{q}{m}\right)^{\alpha} m\right] d\lambda \\ &= D_\alpha(P,M) - (1-\alpha)\,D_\alpha(Q,M) \\ &\qquad - \frac{1}{\alpha-1}\left[\int_{\mathcal{X}} \left(\frac{q}{m}\right)^{\alpha-1} p d\lambda - 1\right]. \end{aligned} \tag{39}$$

In particular, we get from here [consistent with (22)]

$$B_\alpha(P,Q\,|\,Q) = D_\alpha(P,Q)$$

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

$$\begin{aligned} B_\alpha(P,Q\,|\,P) &= (\alpha-2)\,D_{\alpha-1}(Q,P) + (\alpha-1)\,D_\alpha(Q,P) \\ &\equiv (\alpha-2)\,D_{2-\alpha}(P,Q) + (\alpha-1)\,D_{1-\alpha}(P,Q). \end{aligned}$$

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

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

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

$$\begin{aligned} &\lim_{\alpha\downarrow 0} D_\alpha(P,M) = D_{-\ln t}(P,M) \text{ and} \\ &\lim_{\alpha\uparrow 1} D_\alpha(P,M) = D_{t\ln t}(P,M) \end{aligned} \tag{40}$$

proved as [30, Prop. 2.9] 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

$$\lim_{\alpha\downarrow 0} B_\alpha(P,Q\,|\,M)$$

$$= D_0(P,M) - D_0(Q,M) + \int_{\mathcal{X}} \frac{mp}{q} d\lambda - 1 \tag{41}$$

$$= B_{-\ln t}(P,Q\,|\,M). \tag{42}$$

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

$$\begin{aligned} &\lim_{\beta\downarrow 0} \int_{\mathcal{X}} \frac{(q/m)^{-\beta} - 1}{\beta}\, dP \\ &= \int_{\mathcal{X}} \lim_{\beta\downarrow 0} \frac{(q/m)^{-\beta} - 1}{\beta}\, dP = -\int_{\mathcal{X}} \ln\frac{q}{m}\, dP \end{aligned} \tag{43}$$

then

$$\lim_{\alpha\uparrow 1} B_\alpha(P,Q\,|\,M) = D_1(P,M) - \int_{\mathcal{X}} \ln\frac{q}{m}\, dP \tag{44}$$

$$= D_1(P,Q) = B_{t\ln t}(P,Q\,|\,M). \tag{45}$$

Proof: If $0 < \alpha < 1$, then $D_\alpha(P,M), D_\alpha(Q,M)$ are finite so that (39) holds. Applying the first relation of (40) in (39), we get (41) 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 (40) and the assumption (43) in (39), we end up at (44) where the right-hand side is well defined because $D_1(P,M) + D_1(Q,M)$ is assumed to be finite. The identity (42) follows from (41), (33) and the identity (45) from (44), (30). □

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

$$B_1(P,Q\,|\,M) = B_{t\ln t}(P,Q\,|\,M) \tag{46}$$

and

$$B_0(P,Q\,|\,M) = B_{-\ln t}(P,Q\,|\,M) \tag{47}$$

and thus, (45) and (42) become

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

and

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

Furthermore, in these notations, the relations (30), (34), and (35) reformulate (under the corresponding assumptions) as follows:

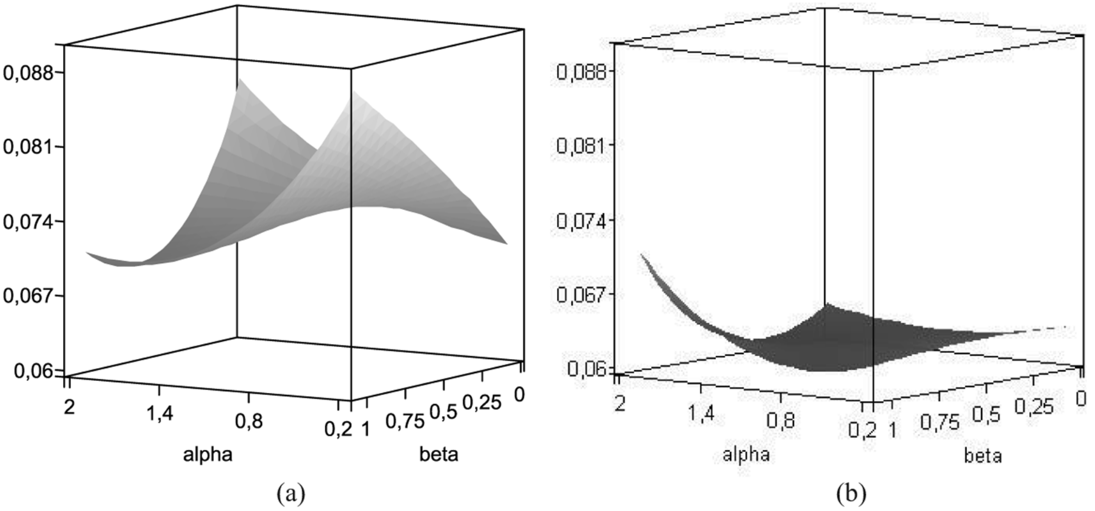
$$\begin{aligned} B_1(P,Q\,|\,M) &= D_1(P,Q) \\ B_0(P,Q\,|\,Q) &= D_0(P,Q) \end{aligned}$$

and

$$\begin{aligned} B_0(P,Q\,|\,P) &= \chi^2(P,Q) - D_1(P,Q) \\ &= 2\,D_2(P,Q) - D_1(P,Q). \end{aligned} \tag{48}$$

Remark 3: 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 2-D discrimination plots $\{[D_\alpha(P,Q);\alpha] : c \le \alpha \le d\} \subset \mathbb{R}^2$ into more informative 3-D *discrimination plots*

$$\{[B_\alpha(P,Q|\beta P + (1-\beta)Q);\ \alpha;\ \beta] : c \le \alpha,\ \beta \le d\} \subset \mathbb{R}^3 \tag{49}$$



(a)

(b)

Fig. 1. Three-dimensional discrimination plots (49) for $P = \mathrm{Bin}(10, \widetilde{p})$, $Q = \mathrm{Bin}(10, \widetilde{q})$ with $0.2 \le \alpha \le 2$ and $0 \le \beta \le 1$.

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 colored projections of the 3-D plots (49) allow much more intimate insight into the relation between data and their statistical models. Therefore this computer-aided 3-D 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 $P = \mathrm{Bin}(n, \widetilde{p})$ be a binomial distribution with parameters $n, \widetilde{p}$ (with a slight abuse of notation), and $Q = \mathrm{Bin}(n, \widetilde{q})$. Fig. 1 presents projections of the corresponding 3-D discrimination plots (49) for $0.2 \le \alpha \le 2$ and $0 \le \beta \le 1$, where the Fig. 1(a) used the parameter constellation $n = 10$, $\widetilde{p} = 0.25$, $\widetilde{q} = 0.20$ whereas the Fig. 1(b) used $n = 10$, $\widetilde{p} = 0.25$, $\widetilde{q} = 0.30$. In both cases, the ranges of $B_\alpha(P, Q | \beta P + (1 - \beta) Q)$ 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

$$b(\theta) = \ln \int_{\mathbb{R}^d} e^{x \cdot \theta} d\lambda(x), \qquad \theta \in \mathbb{R}^d \tag{50}$$

and the convex set

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

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

$$p_\theta(x) \equiv \frac{dP_\theta}{d\lambda}(x) = \exp\{x \cdot \theta - b(\theta)\}, \quad x \in \mathbb{R}^d, \quad \theta \in \Theta. \tag{51}$$

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

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

Note that (51) 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., [8]).

The formula

$$\int_{\mathbb{R}^d} e^{x \cdot \theta} \, d\lambda(x) = e^{b(\theta)}, \quad \theta \in \Theta \tag{52}$$

follows from (50) and implies

$$\int_{\mathbb{R}^d} x \, e^{x \cdot \theta} \, d\lambda(x) = e^{b(\theta)} \nabla b(\theta), \theta \in \mathring{\Theta}. \tag{53}$$

Both formulas (52) and (53) will be useful in the sequel.

We are interested in the scaled Bregman power distances

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

Here, P_{θ_1}, P_{θ_2}, and P_{θ_0} are measure-theoretically equivalent probability measures, so that we can turn attention to the formulas (39), (30), (33), and (46) to (48), 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 (30), (33), and (39). To begin with, let us mention the following well-established representation.

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

$$\frac{\exp\left\{ b(\alpha \theta_1 + (1 - \alpha)\, \theta_2) - \alpha b(\theta_1) - (1 - \alpha)\, b(\theta_2) \right\} - 1}{\alpha(\alpha - 1)}. \tag{54}$$

In particular, 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.

This can be easily seen by slightly extending (38) to get for arbitrary $\alpha \in \mathbb{R}$ and $\theta_1, \theta_2 \in \Theta$

$$1 + \alpha \cdot (\alpha - 1) \cdot D_\alpha(P_{\theta_1}, P_{\theta_2}) = \int_{\mathbb{R}^d} p_{\theta_1}^{\alpha}\, p_{\theta_2}^{1-\alpha}\, d\lambda$$
$$= \frac{\int_{\mathbb{R}^d} \exp\{x \cdot [\alpha\theta_1 + (1-\alpha)\,\theta_2]\}\, d\lambda(x)}{\exp\{\alpha b(\theta_1) + (1-\alpha)\, b(\theta_2)\}}$$

which together with (52) gives the desired result.

The skew symmetry as well as the remaining power divergences $D_0(P_{\theta_1}, P_{\theta_2})$ and $D_1(P_{\theta_1}, P_{\theta_2})$ is given in the next, straightforward theorem.

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

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

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

$$\begin{aligned} D_{-\ln t}\left(P_{\theta_1}, P_{\theta_2}\right) &= D_0\left(P_{\theta_1}, P_{\theta_2}\right) = \lim_{\alpha \downarrow 0} D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) \\ &= b(\theta_1) - b(\theta_2) - \nabla b(\theta_2)\,(\theta_1 - \theta_2) \qquad (55) \\ &= \lim_{\alpha \uparrow 1} D_\alpha\left(P_{\theta_2}, P_{\theta_1}\right) = D_1\left(P_{\theta_2}, P_{\theta_1}\right) = D_{t \ln t}\left(P_{\theta_2}, P_{\theta_1}\right). \end{aligned} \tag{56}$$

The *main* result of this section is the following representation theorem for Bregman distances in exponential families. We formulate this in terms of the functions

$$\rho_\alpha(\theta_1, \theta_2) = b\Big(\alpha\theta_1 + (1-\alpha)\,\theta_2\Big) - \alpha b(\theta_1) - (1-\alpha)\, b(\theta_2) \tag{57}$$

(where the right-hand side is finite if $0 \le \alpha \le 1$), as well as 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

$$\sigma_\alpha(\theta_0, \theta_1, \theta_2) = \sigma_\alpha^{I}(\theta_0, \theta_1, \theta_2) - \sigma_\alpha^{II}(\theta_0, \theta_1, \theta_2) \tag{58}$$

of the nonnegative (possibly infinite)

$$\sigma_\alpha^{I}(\theta_0, \theta_1, \theta_2) = b\Big(\alpha\,\theta_1 + (1-\alpha)\;[\theta_1 - \theta_2 + \theta_0]\Big) \tag{59}$$

and the finite

$$\sigma_\alpha^{II}(\theta_0, \theta_1, \theta_2) = \alpha\, b(\theta_1) + (1-\alpha)\,\Big[b(\theta_1) - b(\theta_2) + b(\theta_0)\Big]. \tag{60}$$

Alternatively

$$\begin{aligned} \sigma_\alpha(\theta_0, \theta_1, \theta_2) &= \rho_\alpha(\theta_1, \theta_0 + \theta_1 - \theta_2) \\ &\quad + (1-\alpha)\,[b(\theta_0 + \theta_1 - \theta_2) - b(\theta_0) - b(\theta_1) + b(\theta_2)]. \end{aligned} \tag{61}$$

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

$$\begin{aligned} &B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \\ &= \frac{\exp \rho_\alpha(\theta_1, \theta_0)}{\alpha(\alpha-1)} + \frac{\exp \rho_\alpha(\theta_2, \theta_0)}{\alpha} + \frac{\exp \sigma_\alpha(\theta_0, \theta_1, \theta_2)}{1-\alpha}. \end{aligned} \tag{62}$$

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

$$\begin{aligned} B_0\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) &= b(\theta_1) - b(\theta_2) - \nabla b(\theta_0)\,(\theta_1 - \theta_2) \\ &\quad + \exp \sigma_0(\theta_0, \theta_1, \theta_2) - 1 \end{aligned} \tag{63}$$

respectively,

$$B_1\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) = b(\theta_2) - b(\theta_1) - \nabla b(\theta_1)\,(\theta_2 - \theta_1). \tag{64}$$

In particular, all scaled Bregman distances (62)–(64) 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 is a d-vector.

Proof:

1) By (51), it holds for every $\alpha \in \mathbb{R}$ and $\theta_0, \theta_1, \theta_2 \in \Theta$

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

with $\sigma_\alpha^{II}(\theta_0, \theta_1, \theta_2)$ from (60). Since (52) leads to

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

for $\sigma_\alpha^{I}(\theta_0, \theta_1, \theta_2)$ given by (59), it holds

$$\int_{\mathcal{X}} \left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\, d\lambda = \exp \sigma_\alpha(\theta_0, \theta_1, \theta_2) \tag{65}$$

where $\sigma_\alpha(\theta_0, \theta_1, \theta_2)$ was defined in (58). Now, by plugging

$$P = P_{\theta_1}, \quad Q = P_{\theta_2}, \quad M = P_{\theta_0} \quad [\text{cf., (52)}]$$

in (39), we get for $\alpha(\alpha-1)\neq 0$ the Bregman distances

$$\begin{aligned} &B_\alpha\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right)\\ &\quad= D_\alpha\left(P_{\theta_1},P_{\theta_2}\right) - (1-\alpha)\,D_\alpha\left(P_{\theta_2},P_{\theta_0}\right)\\ &\qquad+\frac{1}{1-\alpha}\left[\int_{\mathcal{X}}\left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\,d\lambda - 1\right]. \end{aligned}\tag{66}$$

By combining the power divergence formula (54) with (57), one ends up with $D_\alpha\left(P_{\theta_1},P_{\theta_2}\right) = \frac{\exp\{\rho_\alpha(\theta_1,\theta_2)\}-1}{\alpha(\alpha-1)}$ which together with (65) and (66) leads to the desired representation (62).

2) By the definition of $B_0(P,Q\,|\,M)$ in (47) and by (41)

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

where

$$\int_{\mathcal{X}}\frac{p_{\theta_0}\,p_{\theta_1}}{p_{\theta_2}}\,d\lambda = \exp\sigma_0(\theta_0,\,\theta_1,\,\theta_2) \quad (\text{cf., (65)}).$$

For $\theta_0\in\mathring{\Theta}$, the desired assertion (63) 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 (55).

3) The desired formula (64) follows immediately from the definition (46) and from the formulas (44) and (45), (55) and (56).

4) The finally stated invariance is immediate. □

The Conclusion 1 of Section III 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 $\mathbb{P}=\left\{P_\theta:\theta\in\mathbb{R}^d\right\}$ 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, $P_z\in\mathbb{P}$

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

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 $P_z\in\mathbb{P}$. 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 4: We see from Theorems 4 and 5 that—consistent with (30), (45)—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_{θ_1}, P_{θ_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 5: We see from the formulas (54)–(64) 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 (50), and *not* directly on the reference measure λ used in the definition formulas (50), (51).

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 some important discrete and continuous exponential families, and also for exponentially distributed random processes.

Binomial Model: Consider for fixed $n\geq 2$ on the observation space $\mathcal{X}=\{0,\ldots,n\}$ the binomial distribution P_θ determined by

$$P_\theta[\{x\}] = \lambda[\{x\}]\cdot\exp\{x\cdot\theta-b(\theta)\} = \binom{n}{x}p^x(1-p)^{n-x}$$

for $x\in\{0,\ldots,n\}$, 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 (57) and (61)

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

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 achieves an explicit formula for the binomial Bregman distances $B_\alpha\left(P_{\theta_1},P_{\theta_2}\,|\,P_{\theta_0}\right)$ from here.

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

$$p_\theta(x) = \theta x \exp\left\{-\frac{\theta x^2}{2}\right\}, \quad \theta \in \Theta = (0,\infty) \tag{67}$$

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 (67) into the family of Rayleigh densities

$$p_\theta(x) = \theta \exp\{\theta x\} = \exp\{\theta x - b(\theta)\} \quad \text{for } b(\theta) = -\ln\theta,\ \theta > 0$$

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 (51). After some calculations one derives from (57)

$$\rho_\alpha(\theta_1,\theta_2) = \ln\frac{\theta_1^\alpha\,\theta_2^{1-\alpha}}{\alpha\theta_1 + (1-\alpha)\theta_2} \tag{68}$$

and

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

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

Exponentially Distributed Signals: Most of the random processes modeling physical, social, and economic phenomena are exponentially distributed. Important among them are the real valued Lévy processes $\boldsymbol{X}_t = (X_s : 0 \le s \le t)$ with trajectories $\boldsymbol{x}_t = (x_s : 0 \le s \le t)$ 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

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

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., [27]) 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

$$\frac{dP_{t,\theta}}{dP_{t,0}}(\boldsymbol{x}_t) = \exp\{\theta x_t - b_t(\theta)\} \tag{69}$$

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

$$b_t(\theta) = t\left(\delta\theta + \frac{1}{2}\sigma^2\theta^2 + \gamma(\theta)\right) \tag{70}$$

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

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

The formula (69) implies that the family $\mathbb{P}_t = \{P_{t,\theta} : \theta \in \Theta\}$ 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

$$B(P_{t,\theta_1}, P_{t,\theta_2} | P_{t,0}) = B(Q_{t,\theta_1}, Q_{t,\theta_2} | Q_{t,0}) \tag{71}$$

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 section, 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 $\mathrm{Poi}(\tau)$ with parameter $\tau = t\eta = te^\theta$ and probabilities

$$Q_{t,\theta}[\{x\}] = \frac{e^{-\tau}\,(\tau)^x}{x!} = \lambda[\{x\}] \cdot \exp\left\{x\vartheta - e^\vartheta\right\} \quad \text{for } \vartheta = \ln\tau = \theta + \ln t,\ \ \lambda[\{x\}] = \frac{1}{x!}.$$

The exponential structure is similar as earlier, so that by applying (57) 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 (61) and Theorem 5, we obtain an explicit formula for the scaled Bregman distance (71) 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

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

which are Lévy processes with $\delta = 0$, $\sigma = 1$ and $c(\theta) \equiv 0$ so that (70) 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

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

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 (57)

$$\rho_\alpha(\theta_1, \theta_2) = -\ln\frac{\theta_1^\alpha\,\theta_2^{1-\alpha}}{\alpha\theta_1 + (1-\alpha)\theta_2} \qquad [\text{cf., (68)}].$$

This together with (61) and Theorem 5 leads to the explicit formula for the scaled Bregman distance (71) of the Wiener processes under consideration.

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

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

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 (71), 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 $\mathrm{LN}(\theta_1 t, \sigma^2 t)$, $\mathrm{LN}(\theta_2 t, \sigma^2 t)$. As said previously, it coincides with the scaled Bregman distance of two normal distributions $\mathrm{N}(\theta_1 t, \sigma^2 t)$, $\mathrm{N}(\theta_2 t, \sigma^2 t)$. 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 $\mathrm{N}(\mu, v^2)$ and $\mathrm{LN}(\mu, v^2)$ 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

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

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 earlier, we get from (57)

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

Expression (61) can be automatically evaluated using this. Applying both these results in Theorem 5, one obtains explicit formula for the scaled Bregman distance (71) of these geometric Brownian motions.

ACKNOWLEDGMENT

We are grateful to all three referees for useful suggestions.

REFERENCES

[1] S.-I. Amari, "Integration of stochastic models by minimizing α -divergence," *Neural Comput.*, vol. 19, no. 10, pp. 2780–2796, 2007.

[2] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669, Jul. 2005.

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

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, pp. 138–156, 2006.

[5] H. H. Bauschke and J. M. Borwein, "Legendre functions and the method of random Bregman projections," *J. Convex Anal.*, vol. 4, no. 1, pp. 27–67, 1997.

[6] A. Boratynska, "Stability of Bayesian inference in exponential families," *Statist. Probabil. Lett.*, vol. 36, pp. 173–178, 1997.

[7] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217, 1967.

[8] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. Math. Statist., 1986.

[9] D. Butnariu and E. Resmerita, "Bregman distances, totally convex functions, and a method for solving operator equations in Banach spaces," *Abstr. Appl. Anal.*, vol. 2006, pp. 39–39, 2006.

[10] C. Byrne, "Iterative projection onto convex sets using multiple Bregman distances," *Inv. Probl.*, vol. 15, pp. 1295–1313, 1999.

[11] B. A. Carlson and M. A. Clements, "A computationally compact divergence measure for speech processing," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 13, no. 12, pp. 1255–1260, Dec. 1991.

[12] Y. Censor and S. A. Zenios, *Parallel Optimization—Theory, Algorithms, and Applications*. New York: Oxford Univ. Press, 1997.

[13] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[14] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Mach. Learn.*, vol. 48, pp. 253–285, 2002.

[15] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci., A*, vol. 8, pp. 85–108, 1963.

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

[17] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[18] I. Csiszár, "Maximum entropy and related methods," in *Proc. 12th Prague Conf. Inf. Theory, Statist. Decis. Funct. Random Process.*, Prague, Czech Republic, 1994, pp. 58–62.

[19] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, pp. 161–186, 1995.

[20] I. Csiszár and F. Matúš, "On minimization of entropy functionals under moment constraints," in *Proc. Int. Symp. Inf. Theory*, Toronto, ON, Canada, 2008, pp. 2101–2105.

[21] I. Csiszár and F. Matúš, "On minimization of multivariate entropy functionals," in *Proc. IEEE Int. Theory Workshop Netw. Inf. Theory*, Volos, Greece, 2009, pp. 96–100.

[22] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

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

[24] P. R. Halmos, *Measure Theory*. New York: Van Nostrand, 1964.

[25] T. Hertz, A. Bar-Hillel, and D. Weinshall, “Learning distance functions for information retrieval,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. II-570–II-577.

[26] L. K. Jones and C. L. Byrne, “General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis,” *IEEE Trans. Inf. Theory*, vol. 36, no. 1, pp. 23–30, Jan. 1990.

[27] U. Küchler and M. Sorensen, “Exponential families of stochastic processes and Lévy processes,” *J. Statist. Plann. Infer.*, vol. 39, pp. 211–237, 1994.

[28] J. D. Lafferty, “Additive models, boosting, and inference for generalized divergences,” in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. New York: ACM, 1999, pp. 125–133.

[29] E. L. Lehman and J. P. Romano*, Testing Statistical Hypotheses*. Berlin, Germany: Springer-Verlag, 2005.

[30] F. Liese and I. Vajda*, Convex Statistical Distances*. Leipzig, Germany: Teubner, 1987.

[31] F. Liese and I. Vajda, “On divergences and informations in statistics and information theory,” *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[32] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[33] A. Marquina and S. J. Osher, “Image super-resolution by TV-regularization and Bregman iteration,” *J. Sci. Comput.*, vol. 37, pp. 367–382, 2008.

[34] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, “Information geometry of $\mathcal{U}$-boost and Bregman divergence,” *Neural Comput.*, vol. 16, no. 7, pp. 1437–1481, 2004.

[35] R. Nock and F. Nielsen, “Bregman divergences and surrogates for learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2048–2059, Nov. 2009.

[36] M. C. Pardo and I. Vajda, “About distances of discrete distributions satisfying the data processing theorem of information theory,” *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1288–1293, Jul. 1997.

[37] M. C. Pardo and I. Vajda, “On asymptotic properties of information-theoretic divergences,” *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1860–1868, Jul. 2003.

[38] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, Berkeley, CA, 1961, vol. 1, pp. 547–561.

[39] E. Resmerita and R. S. Anderssen, “Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems,” *Math. Methods Appl. Sci.*, vol. 30, no. 13, pp. 1527–1544, 2007.

[40] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen*, Variational Methods in Imaging*. New York: Springer-Verlag, 2008.

[41] W. Stummer, “Some Bregman distances between financial diffusion processes,” *Proc. Appl. Math. Mech.*, vol. 7, no. 1, pp. 1050503–1050504, 2007.

[42] W. Stummer and I. Vajda, “On divergences of finite measures and their applicability in statistics and information theory,” *Statistics*, vol. 44, pp. 169–187, 2010.

[43] M. Teboulle, “A unified continuous optimization framework for center-based clustering methiods,” *J. Mach. Learn. Res.*, vol. 8, pp. 65–102, 2007.

[44] I. Vajda, “On metric divergences of probability measures,” *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.

[45] I. Vajda and J. Zvárová, “On generalized entropies, Bayesian decisions and statistical diversity,” *Kybernetika*, vol. 43, no. 5, pp. 675–696, 2007.

[46] R. N. J. Veldhuis, “The centroid of the Kullback–Leibler distance,” *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 96–99, Mar. 2002.

[47] J. Xu and S. Osher, “Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising,” *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 534–544, Feb. 2007.

Wolfgang Stummer graduated from the Johannes Kepler University Linz, Austria, in 1987 and received the Ph.D. degree in 1991 from the University of Zurich, Switzerland.

From 1993 to 1995, he worked as Research Assistant at the University of London and the University of Bath (UK). From 1995 to 2001 he was Assistant Professor at the University of Ulm (Germany). From 2001 to 2003 he held a Term Position as a Full Professor at the University of Karlsruhe (now KIT; Germany) where he continued as Associate Professor until 2005. Since then, he is affiliated as Full Professor at the Department of Mathematics, University of Erlangen-Nürnberg FAU (Germany); at the latter, he is also a Member of the School of Business and Economics.

Igor Vajda (M’90–F’01) was born in 1942 and passed away suddenly after a short illness on May 2, 2010. He graduated from the Czech Technical University, Czech Republic, in 1965 and received the Ph.D. degree in 1968 from Charles University, Prague, Czech Republic.

He worked at UTIA (Institute of Information Theory and Automation, Czech Academy of Sciences) from his graduation until his death, and became a member of the Board of UTIA in 1990. He was a visiting professor at the Katholieke Universiteit Leuven, Belgium; the Universidad Complutense Madrid, Spain; the Université de Montpellier, France; and the Universidad Miguel Hérnandez, Alicante, Spain. He published four monographs and more than 100 journal publications.

Dr. Vajda received the Prize of the Academy of Sciences, the Jacob Wolfowitz Prize, the Medal of merits of Czech Technical University, several Annual prizes from UTIA, and, posthumously, the Bolzano Medal from the Czech Academy of Sciences.