Bregman distances in exponential families of probability measures

Wolfgang Stummer [a,1] and *Igor Vajda* [b,2]

[a]Department of Mathematics, University of Erlangen–Nürnberg,
Bismarckstrasse $1\frac{1}{2}$, D – 91054 Erlangen, Germany,
Tel.: ++49-9131-85-22503; Fax: ++49-9131-85-26214.

[b]Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic,
Pod Vodárenskou Věží 4, 182 08 Praha 8 – Libeň, Czech Republic,
E-mail: vajda@utia.cas.cz

December 14, 2008

Keywords: Bregman distances; exponential families; $\phi-$divergences.

AMS Subject Classification: 62B10; 94A17; 62E10.

1 Bregman distances of probability measures

The concept of Bregman distances for *Euclidean space vectors* was introduced by Bregman (1967) in the context of convex programming. In this setup, the Bregman method has been widely applied and adapted, especially for the design of regularization algorithms for finding a good approximate solution of inverse problems e.g. in image processing (tomography etc.), see for instance Censor and Lent (1981), Eggermont (1993), Byrne (1999), Resmerita (2005), Silva Neto and Cella (2006), Resmerita and Scherzer (2007), Resmerita and Anderssen (2007), Xu and Osher (2007), Burger et al. (2008), Cai et al. (2008), Marquina and Osher (2008), Osher et al. (2008), Scherzer et al. (2008), and the references therein. Bregman distances for *non-negative* functions were treated in Csiszár (1995). In the context of information theory and statistical decision theory, Bregman distances were studied e.g. by Csiszár (1991, 1994) as well as Pardo and Vajda (1997, 2003) basically for *discrete probability measures* or related functional quantities; closely related contexts are also applied in machine learning, see e.g. in Lafferty et al. (1997), Kivinen and Warmuth (1999), Lafferty (1999), Collins et al. (2001), Della Pietra et al. (2002), Murata et al. (2004), as well as in Cesa-Bianchi and Lugosi (2006). Applications to statistical physics are e.g. given in Topsoe (2007).

[1]Corresponding author. E-mail:stummer@mi.uni-erlangen.de
[2]Supported by the MŠMT grant 1M0572 and the GA ČR grant 102/07/1131.

In this paper, we study Bregman distances of general probability measures, in particular of the laws belonging to an exponential family. As a by-product, we retrieve some of the results of Azoury and Warmuth (2001) as a special case. Our setup contrasts with the studies in Banerjee et al. (2005) which *pointwise* represent the densities of exponential family distributions in terms of Bregman distances of Euclidean vectors.

1.1 Divergences of probability measures and finite measures

As a preparation for the below exact definition of the Bregman distances, and for the derivation of some of their basic properties, we first introduce some notations and discuss some relevant issues on the $\phi-$divergences of measures and probability measures. Denoting by $\mathcal{P}$ respectively $\mathcal{M}$ the space of all probability respectively finite measures on a measurable space $(\mathcal{X},\mathcal{A})$, throughout this paper we shall always consider $P_1,\ P_2,\in\mathcal{P}_1$ and $Q\in\mathcal{M}$, all three of them mutually measure-theoretically equivalent and dominated by a σ-finite measure μ on $(\mathcal{X},\mathcal{A})$. Then the densities

$$p_i=\frac{\mathrm{d}P_i}{\mathrm{d}\mu},\quad i=1,2\qquad\text{and}\qquad q=\frac{\mathrm{d}Q}{\mathrm{d}\mu}$$

may be assumed as strictly positive on $\mathcal{X}$. Furthermore, let $\phi:(0,\infty)\mapsto\mathbb{R}$ be a continuous 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

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

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

For every $P\in\{P_1,\ P_2\}$ we consider the ϕ-divergence

$$D_\phi(P,Q)=\int_{\mathcal{X}}q\,\phi\left(\frac{p}{q}\right)\mathrm{d}\mu,\qquad\text{with } p=\frac{\mathrm{d}P}{\mathrm{d}\mu}\ , \tag{2}$$

which does *n*ot depend on the choice of the dominating measure μ (see e.g. Liese and Vajda (1987)). It is useful to take into account that for $s\in(0,\infty)$ one gets the bounds

$$\phi(s)+\phi'_+(s)(t-s)\le\phi(t)\le\phi(0)+t\phi^*(0),\quad\text{for all } t\in(0,\infty). \tag{3}$$

The left-hand side is the well-known support line of $\phi(t)$ at $t=s$. The right-hand inequality is trivial if $\phi(0)=\infty$, and in the opposite case it follows by taking $s\to\infty$ in the inequality

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

which is an easy consequence of the Jensen inequality between $\phi(t)$ and the extremal values $\phi(0),\phi(s)$ (with $0<t<s$). By taking the special case $s=1$ and $t=p/q$ in (3) and multiplying both sides by q, we end up at

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

Integrating this inequality, we get the ϕ-divergence bounds

$$\phi'_+(1)(1-Q(\mathcal{X})) \;\le\; D_\phi(P,Q) \;\le\; Q(\mathcal{X})\,\phi(0)+\phi^*(0) \tag{4}$$

which can be used e.g. to check the finiteness of $D_\phi(P,Q)$ (playing a role for several representation results below).

Notice that $D_\phi(P,Q)$ might be negative. For probability measures P_1, P_2 the bounds (4) take on the form

$$0 \le D_\phi(P_1,P_2) \le \phi(0)+\phi^*(0)\ , \tag{5}$$

and the equalities are achieved under well-known conditions (cf. Liese and Vajda (1987), (2006)): the left equality holds *if* $P_1 = P_2$, and the right one holds *if* $P_1 \perp P_2$ (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 (4), which extends the left-hand inequality in (5) including the conditions for the equality, is given by the following theorem.

Theorem 1. For every $P \in \mathcal{P}$, $Q \in \mathcal{M}$ one gets the lower divergence bound

$$Q(\mathcal{X})\,\phi\left(\frac{1}{Q(\mathcal{X})}\right) \le D_\phi(P,Q)\ , \tag{6}$$

where the equality holds if

$$p \;=\; \frac{q}{Q(\mathcal{X})} \qquad P\text{-a.s.} \tag{7}$$

If $D_\phi(P,Q) < \infty$ and $\phi(t)$ is strictly convex at $t = 1/Q(\mathcal{X})$, the equality in (6) holds if and only if (7) holds.

Proof. By (2) and the definition (1) of the convex function ϕ^*_α,

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

Hence by Jensen's inequality

$$D_\phi(P,Q) \ge \phi^*\left(\int_{\mathcal{X}} \frac{q}{p}\,\mathrm{d}P\right) = \phi^*(Q(\mathcal{X})) \tag{8}$$

which proves the desired inequality (6). Since

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

is the condition for equality in (8), 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_1, P_2 we get directly from definition (2) the "skew symmetry" ϕ-divergence formula

$$D_{\phi^*}(P_1, P_2) = D_\phi(P_2, P_1) \ , \tag{9}$$

as well as the sufficiency of the condition

$$\phi(t) - \phi^*(t) \ \equiv \ \text{constant} \cdot (t-1) \tag{10}$$

for the ϕ-divergence symmetry

$$D_\phi(P_1, P_2) = D_\phi(P_2, P_1) \quad \text{for all } P_1, P_2 \ . \tag{11}$$

Liese and Vajda (1987) proved that if $\phi(t)$ is strictly convex at $t = 1$, then condition (10) is is not only *sufficient* but also *necessary* for the symmetry (11).

1.2 General Bregman distance

In the following, we present the basic concept of the current paper, which is a measure-theoretic version of the Bregman distance for Euclidean space vectors introduced into the literature by Bregman (1967).

Definition 1. The *Bregman distance* of probability measures $P_1, \ P_2 \in \mathcal{P}$ relative a finite measure $Q \in \mathcal{M}$ is defined by the formula

$$B_\phi\left(P_1, P_2 \,|\, Q\right) \ = \ \int_{\mathcal{X}} \left[\phi\left(\frac{p_1}{q}\right) - \phi\left(\frac{p_2}{q}\right) - \phi'_+\left(\frac{p_2}{q}\right)\left(\frac{p_1}{q} - \frac{p_2}{q}\right)\right] \mathrm{d}Q \tag{12}$$

$$= \ \int_{\mathcal{X}} \left[q\phi\left(\frac{p_1}{q}\right) - q\phi\left(\frac{p_2}{q}\right) - \phi'_+\left(\frac{p_2}{q}\right)(p_1 - p_2)\right] \mathrm{d}\mu. \tag{13}$$

Remark. By putting $t = p_1/q$ and $s = p_2/q$ in (3) we find the argument of the integral in (12) to be nonnegative. Hence the Bregman distance $B_\phi\left(P_1, P_2 \,|\, Q\right)$ is well-defined by (12) or (13) and is always nonnegative (possibly infinite).

The definition (12), (13) was formally given in Stummer (2007) within the context of probability measures, which for the case of differentiable, strictly convex "scaling function" ϕ can also be deduced from the context of Bregman divergences for nonnegative functions (rather than measures) in Csiszár (1995); see also Gao et al. (2004) for a financial version concerning equivalent martingale measures under some integrability restrictions. As it will be shown in Subsection 1.5 below, the Bregman distance $B_\phi\left(P_1, P_2 \,|\, Q\right)$ *g*enerally does depend on the choice of the reference measure Q respectively μ (in contrast to the $\phi-$divergence $D_\phi(P, Q)$). For $\mathcal{X} \subset \mathbb{R}$ (resp. $\mathbb{R}^d$), the following special choices of the reference measure Q have already been used in literature:

(a) $\mathcal{X}$ is finite or countable and $Q = \mu$ is the counting measure on $\mathcal{X}$ (which is in general $\sigma-$finite rather than finite which was assumed above). Then, for discrete probability measures P_1 and P_2 supported on $\mathcal{X}$ the densities are $p_i(x) = P_i(\{x\}) > 0$ on $\mathcal{X}$ and the Bregman distance reduces to

$$B_\phi(P_1, P_2 \,||\, Q) = \sum_{x \in \mathcal{X}} \Big(\phi(p_1(x)) - \phi(p_2(x)) - \phi'_+(p_2(x))\,(p_1(x) - p_2(x)) \Big) \,. \tag{14}$$

For finite $\mathcal{X}$, this special case coincides with the Bregman divergence definition of Csiszár (1991), (1994), and for countable $\mathcal{X}$ with that used by Pardo and Vajda (1997), (2003). Closely related definitions and results in the context of machine learning can be found e.g. in Lafferty et al. (1997), Kivinen and Warmuth (1999), Lafferty (1999), Collins et al. (2001), Della Pietra et al. (2002), Murata et al. (2004), as well as in Cesa-Bianchi and Lugosi (2006). For some special cases in the field of inverse problems, see e.g. Byrne (1999) as well as Silva Neto and Cella (2006).

(b) For open $\mathcal{X}$ and Lebesgue measure $Q = \mu$ (which is of course again $\sigma-$finite rather than finite), p_1 and p_2 are classical Lebesgue densities. In this case, for some particular choices of ϕ the Bregman distance $B_\phi\,(P_1, P_2\,|\,Q)$ coincides with Bregman distances used for the design of regularization techniques for inverse problems (see e.g. Jones and Trutzer (1989), Jones and Byrne (1990), as well as Resmerita and Anderssen (2007)).

By using the remark after Definition 1 and applying (2) we get

$$D_\phi(P_1, Q) \geq D_\phi(P_2, Q) + \int_{\mathcal{X}} \phi'_+\left(\frac{p_2}{q}\right)(p_1 - p_2)\mathrm{d}\mu \tag{15}$$

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

$$B_\phi\,(P_1, P_2\,|\,Q) = D_\phi(P_1, Q) - D_\phi(P_2, Q) - \int_{\mathcal{X}} \phi'_+\left(\frac{p_2}{q}\right)\mathrm{d}\mu \tag{16}$$

if at least two of the right-hand side expressions are finite (which can be checked e.g. by using (4) or (6))

The formula (12) simplifies in the important special cases $Q = P_1$ and $Q = P_2$. In the former, due to $\phi(1) = 0$ it reduces to

$$B_\phi\,(P_1, P_2\,|\,P_1) = \int_{\mathcal{X}} \left[\phi'_+\left(\frac{p_2}{p_1}\right)(p_2 - p_1) - p_1\phi\left(\frac{p_2}{p_1}\right)\right]\mathrm{d}\mu \tag{17}$$

$$= \int_{\mathcal{X}} \phi'_+\left(\frac{p_2}{p_1}\right)(p_2 - p_1)\mathrm{d}\mu - D_\phi(P_2, P_1)\;, \tag{18}$$

where the difference (18) is meaningful if and only if if $D_\phi(P_2, P_1) \equiv D_{\phi^*}(P_1, P_2)$ is finite. The nonnegative divergence measure $\mathcal{B}_\phi\,(P_1, P_2) := B_\phi\,(P_1, P_2\,|\,P_1)$ is thus the difference between the nonnegative divergence measure

$$\mathcal{D}_\phi\,(P_2, P_1) = \int_{\mathcal{X}} \phi'_+\left(\frac{p_2}{p_1}\right)(p_2 - p_1)\mathrm{d}\mu \;\geq\; D_\phi(P_2, P_1)$$

and the nonnegative $\phi-$divergence $D_\phi(P_2, P_1)$.

The other special case $Q = P_2$ is simpler, leading to

$$B_\phi(P_1, P_2 \,|\, P_2) = D_\phi(P_1, P_2) \tag{19}$$

without any restriction on $P_1, P_2 \in \mathcal{P}$ (cf. the informal formula (1) in Stummer (2007)). This shows that our concept of Bregman distance strictly generalizes the concept of $\phi-$divergence.

In the following we discuss some important special cases with respect to the "scaling function" ϕ.

1.3 Bregman logarithmic distance

Let us consider the special function $\phi(t) = t \ln t$. Then $\phi'(t) = \ln t + 1$ so that (12) implies

$$\begin{aligned} B_{t\ln t}(P_1, P_2 \,|\, Q) &= \int_{\mathcal{X}} \left[p_1 \ln \frac{p_1}{q} - p_2 \ln \frac{p_2}{q} - \left(\ln \frac{p_2}{q} + 1 \right) (p_1 - p_2) \right] \mathrm{d}\mu \\ &= \int_{\mathcal{X}} \left[p_1 \ln \frac{p_1}{q} - p_1 \ln \frac{p_2}{q} \right] \mathrm{d}\mu \\ &= \int_{\mathcal{X}} p_1 \ln \frac{p_1}{p_2} \, \mathrm{d}\mu \;=\; D_{t\ln t}(P_1, P_2) \ . \end{aligned} \tag{20}$$

Thus, for $\phi(t) = t \ln t$ the Bregman distance $B_\phi(P_1, P_2 \,|\, Q)$ does not depend on the choice of the reference measure Q resp. μ; in fact, it always leads to the Kulllback-Leibler information divergence (relative entropy) $D_{t\ln t}(P_1, P_2)$, see Stummer (2007).

1.4 Bregman reversed logarithmic distance

Let now $\phi(t) = -\ln t$ so that $\phi'(t) = -1/t$. Then (12) implies

$$B_{-\ln t}(P_1, P_2 \,|\, Q) = \int_{\mathcal{X}} \left[q \ln \frac{q}{p_1} - q \ln \frac{q}{p_2} + \frac{q}{p_2}(p_1 - p_2) \right] \mathrm{d}\mu \tag{21}$$

$$= D_{t\ln t}(Q, P_1) - D_{t\ln t}(Q, P_2) + \int_{\mathcal{X}} \frac{q p_1}{p_2} \, \mathrm{d}\mu - Q(\mathcal{X}) \tag{22}$$

$$= D_{-\ln t}(P_1, Q) - D_{-\ln t}(P_2, Q) + \int_{\mathcal{X}} \frac{q p_1}{p_2} \mathrm{d}\mu - Q(\mathcal{X}) \tag{23}$$

where the equalities (22) and (23) hold if at least two out of the first three expressions on the right-hand side are finite. In particular, (21) implies (in consistency with (19))

$$B_{-\ln t}(P_1, P_2 \,|\, P_2) = D_{-\ln t}(P_1, P_2) \tag{24}$$

and (22) implies for $D_{t\ln t}(P_1, P_2) < \infty$ (in consistency with (18))

$$B_{-\ln t}(P_1, P_2 \,|\, P_1) = \chi^2(P_1, P_2) - D_{t\ln t}(P_1, P_2) \tag{25}$$

where

$$\chi^2(P_1, P_2) = \int_{\mathcal{X}} \frac{(p_1 - p_2)^2}{p_2} \, \mathrm{d}\mu$$

is the well-known Pearson information divergence. From (24) and (25) one can also see that the Bregman distance $B_\phi(P_1, P_2 \,|\, Q)$ does in general depend on the choice of the reference measure Q.

1.5 Bregman power distances

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

$$B_\alpha(P_1, P_2 \,|\, Q) = B_{\phi_\alpha}(P_1, P_2 \,|\, Q) \ , \qquad \alpha \in \mathbb{R},\ \alpha \neq 0,\ \alpha \neq 1 \tag{26}$$

for the family of power convex functions

$$\phi(t) \equiv \phi_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha - 1)} \quad \text{with} \quad \phi'_\alpha(t) = \frac{t^{\alpha-1}}{\alpha - 1} \ . \tag{27}$$

For comparison and representation purposes, we use for $P \in \{P_1, P_2\}$ the power divergences

$$\begin{aligned} D_\alpha(P, Q) &= D_{\phi_\alpha}(P, Q) = \frac{1}{\alpha(\alpha-1)} \left[\int_{\mathcal{X}} p^\alpha \, q^{1-\alpha} \, \mathrm{d}\mu - 1 \right] && (28) \\ &= \frac{\exp \rho_\alpha(P, Q) - 1}{\alpha(\alpha - 1)} \ , \qquad \text{with } \rho_\alpha(P, Q) = \ln \int_{\mathcal{X}} p^\alpha \, q^{1-\alpha} \, \mathrm{d}\mu && (29) \end{aligned}$$

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

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

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

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

of Rényi (1961).

Returning now to the Bregman power distances, observe that if $D_\alpha(P_1, Q) + D_\alpha(P_2, Q)$ is finite then (16), (26) and (27) imply for $\alpha \neq 0,\ \alpha \neq 1$

$$\begin{aligned} B_\alpha(P_1, P_2 \,|\, Q) &= -D_\alpha(P_2, Q) - \frac{1}{\alpha - 1} \int_{\mathcal{X}} \left(\frac{p_2}{q} \right)^{\alpha-1} (p_1 - p_2) \, \mathrm{d}\mu && (30) \\ &= D_\alpha(P_1, Q) - D_\alpha(P_2, Q) - \frac{1}{\alpha - 1} \int_{\mathcal{X}} \left[\left(\frac{p_2}{q} \right)^{\alpha-1} p_1 - \left(\frac{p_2}{q} \right)^{\alpha} q \right] \mathrm{d}\mu && (31) \\ &= D_\alpha(P_1, Q) - (1-\alpha) \, D_\alpha(P_2, Q) - \frac{1}{\alpha-1} \left[\int_{\mathcal{X}} \left(\frac{p_2}{q} \right)^{\alpha-1} p_1 \, \mathrm{d}\mu - 1 \right]. && (32) \end{aligned}$$

In particular, we get from here (in consistency with (19))

$$B_\alpha(P_1, P_2 \,|\, P_2) = D_\alpha(P_1, P_2) \tag{33}$$

and in case of $D_\alpha(P_2, P_1) \equiv D_{1-\alpha}(P_1, P_2) < \infty$ also

$$\begin{aligned} B_\alpha(P_1, P_2 \,|\, P_1) &= (\alpha - 2) \, D_{\alpha-1}(P_2, P_1) + (\alpha - 1) \, D_\alpha(P_2, P_1) && (34) \\ &\equiv (\alpha - 2) \, D_{2-\alpha}(P_1, P_2) + (\alpha - 1) \, D_{1-\alpha}(P_1, P_2). && (35) \end{aligned}$$

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

$$D_1(P,Q) = D_{t\ln t}(P,Q) \quad \text{and} \quad D_0(P,Q) = D_{-\ln t}(P,Q) \tag{36}$$

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

$$\lim_{\alpha\downarrow 0} D_\alpha(P,Q) = D_{-\ln t}(P,Q) \quad \text{and} \quad \lim_{\alpha\uparrow 1} D_\alpha(P,Q) = D_{t\ln t}(P,Q) \tag{37}$$

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

Theorem 2. If $D_0(P_1,Q) + D_0(P_2,Q) < \infty$ then

$$\lim_{\alpha\downarrow 0} B_\alpha(P_1,P_2\,|\,Q) = D_0(P_1,Q) - D_0(P_2,Q) + \int_{\mathcal{X}} \frac{qp_1}{p_2}\,\mathrm{d}\mu - 1 \tag{38}$$

$$\equiv B_{-\ln t}(P_1,P_2\,|\,Q). \tag{39}$$

If $D_1(P_1,Q) + D_1(P_2,Q) < \infty$ and

$$\lim_{\beta\downarrow 0} \int_{\mathcal{X}} \frac{(p_2/q)^{-\beta} - 1}{\beta}\,\mathrm{d}P_1 = \int_{\mathcal{X}} \lim_{\beta\downarrow 0} \frac{(p_2/q)^{-\beta} - 1}{\beta}\,\mathrm{d}P_1 = -\int_{\mathcal{X}} \ln\frac{p_2}{q}\,\mathrm{d}P_1 \tag{40}$$

then

$$\lim_{\alpha\uparrow 1} B_\alpha(P_1,P_2\,|\,Q) = D_1(P_1,Q) - \int_{\mathcal{X}} \ln\frac{p_2}{q}\,\mathrm{d}P_1 \tag{41}$$

$$= D_1(P_1,P_2) = B_{t\ln t}(P_1,P_2\,|\,Q)\ . \tag{42}$$

Proof. If $0 < \alpha < 1$ then $D_\alpha(P_1,Q)$, $D_\alpha(P_2,Q)$ are finite so that (32) holds. Applying the first relation of (37) in (32) we get (38) where the right hand side is well defined because $D_\alpha(P_1,Q) + D_\alpha(P_2,Q)$ is by assumption finite. Simlarly, by using the second relation of (37) and the assumption (40) in (32) we end up at (41) where the right-hand side is well defined because $D_1(P_1,Q)+D_1(P_2,Q)$ is assumed to be finite. The identity (39) follows from (38), (23) and the identity (42) from (41), (20). □

Motivated by the above Theorem 2, we introduce for all probability measures P_1, P_2, Q under consideration the simplified notations

$$B_1(P_1,P_2\,|\,Q) = B_{t\ln t}(P_1,P_2\,|\,Q) \tag{43}$$

and

$$B_0(P_1,P_2\,|\,Q) = B_{-\ln t}(P_1,P_2\,|\,Q)\ , \tag{44}$$

and thus, (42) and (39) become

$$B_1(P_1,P_2\,|\,Q) = \lim_{\alpha\uparrow 1} B_\alpha(P_1,P_2\,|\,Q) \tag{45}$$

and

$$B_0(P_1, P_2 \,|\, Q) \;\;=\;\; \lim_{\alpha \downarrow 0} B_\alpha(P_1, P_2 \,|\, Q). \tag{46}$$

Furthermore, in these notations the relations (20), (24) and (25) reformulate (under the corresponding assumptions) as follows

$$B_1(P_1, P_2 \,|\, Q) \;=\; D_1(P_1, P_2)\;, \tag{47}$$

$$B_0(P_1, P_2 \,|\, P_2) = D_0(P_1, P_2) \tag{48}$$

and

$$\begin{aligned} B_0(P_1, P_2 \,|\, P_1) \;&=\; \chi^2(P_1, P_2) - D_1(P_1, P_2) \\ &=\; 2\, D_2(P_1, P_2) - D_1(P_1, P_2)\;. \end{aligned} \tag{49}$$

2 Bregman power distances in exponential families

In this section we show that the Bregman power distances can be *explicitly* evaluated for P_1, P_2, Q from exponential families. Let μ be a finite measure on $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}^d)$, and let $y \cdot \theta$ denote the scalar product of the Euclidean vectors $y, \theta \in \mathbb{R}^d$. The extended real valued function

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

and the parameter space

$$\Theta = \{\theta \in \mathbb{R}^d : b(\theta) < \infty\} \tag{51}$$

define on $(\mathbb{R}^d, \mathcal{B}^d)$ an *exponential family of probability measures* $\{P_\theta : \theta \in \Theta\}$ with the densities

$$p_\theta(x) \equiv \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x) = e^{x\cdot\theta - b(\theta)}, \qquad x \in \mathbb{R}^d, \quad \theta \in \Theta. \tag{52}$$

The function $b(\theta)$ is convex on $\mathbb{R}^d$, the parameter space Θ is a convex subset of $\mathbb{R}^d$ containing $0 \in \mathbb{R}^d$, and the function $b(\theta)$ is infinitely differentiable in the interior $\mathring{\Theta}$ of Θ with the gradient

$$\triangledown b(\theta_0) = ((\partial/\partial\theta_1, ..., \partial/\partial\theta_r) \cdot b(\theta))_{\theta=\theta_0}\,, \quad \theta_0 \in \mathring{\Theta}. \tag{53}$$

The formula

$$\int_{\mathbb{R}^d} e^{x\cdot\theta}\, \mathrm{d}\mu(x) \;\;=\;\; e^{b(\theta)}, \quad \theta \in \Theta \tag{54}$$

useful in the sequel follows from (52) and implies

$$\int_{\mathbb{R}^d} x\, e^{x\cdot\theta}\, \mathrm{d}\mu(x) \;\;=\;\; e^{b(\theta)} \nabla b(\theta), \quad \theta_0 \in \mathring{\Theta}. \tag{55}$$

We are interested in the Bregman power distances

$$B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \quad \text{for } \theta_0,\, \theta_1,\, \theta_2 \in \Theta,\;\, \alpha \in \mathbb{R}. \tag{56}$$

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 (32), (20), (23), and (43) to (46), promising to mainly 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 applicability of the formulas (32), (20) and (23).

Theorem 3. If $\alpha \in \mathbb{R}$ differs from 0 and 1, then for arbitrary $\theta_1, \theta_2 \in \Theta$ one gets the representation formula

$$D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) \;=\; \frac{\exp\left\{b(\alpha\theta_1 + (1-\alpha)\,\theta_2) - \alpha b(\theta_1) - (1-\alpha)\,b(\theta_2)\right\} - 1}{\alpha(\alpha-1)}\,. \tag{57}$$

Consequently $D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right)$ is finite for all $0 < \alpha < 1$.

Proof. As a slight extension of (29), put for arbitrary $\alpha \in \mathbb{R}$ and $\theta_1, \theta_2 \in \Theta$

$$\rho_\alpha(\theta_1, \theta_2) \;=\; \ln \int_{\mathbb{R}^d} p_{\theta_1}^\alpha p_{\theta_2}^{1-\alpha}\,\mathrm{d}\mu \tag{58}$$

$$= \ln \int_{\mathbb{R}^d} \exp\Big\{\alpha[x\cdot\theta_1 - b(\theta_1)] + (1-\alpha)\,[x\cdot\theta_2 - b(\theta_2)]\Big\}\,\mathrm{d}\mu(x)$$

$$= \ln \frac{\int_{\mathbb{R}^d} e^{x\cdot[\alpha\theta_1 + (1-\alpha)\,\theta_2]}\,\mathrm{d}\mu(x)}{e^{\alpha b(\theta_1) + (1-\alpha)\,b(\theta_2)}}$$

$$= \ln \frac{e^{b(\alpha\theta_1 + (1-\alpha)\,\theta_2)}}{e^{\alpha b(\theta_1) + (1-\alpha)\,b(\theta_2)}} \qquad \text{(cf. (54))}.$$

Hence

$$\rho_\alpha(\theta_1, \theta_2) = b\Big(\alpha\theta_1 + (1-\alpha)\,\theta_2\Big) - \alpha b(\theta_1) - (1-\alpha)\,b(\theta_2)\;, \tag{59}$$

where the right hand side is finite if $0 \le \alpha \le 1$. Furthermore, (29) implies for $\alpha \in \mathbb{R}\backslash\{0,1\}$

$$D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) = \frac{\exp \rho_\alpha(\theta_1, \theta_2) - 1}{\alpha(\alpha-1)} \tag{60}$$

Thus, (57) follows from (59) and (60). The declared finitness of $D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right)$ is immediately clear. □

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

$$D_\alpha\left(P_{\theta_2}, P_{\theta_1}\right) \;=\; D_{1-\alpha}\left(P_{\theta_1}, P_{\theta_2}\right) \tag{61}$$

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

$$D_{-\ln t}\left(P_{\theta_1}, P_{\theta_2}\right) \;=\; D_0\left(P_{\theta_1}, P_{\theta_2}\right) \;=\; \lim_{\alpha\downarrow 0} D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) \tag{62}$$

$$= \; b(\theta_1) - b(\theta_2) - \nabla b(\theta_2)\,(\theta_1 - \theta_2) \tag{63}$$

$$= \; \lim_{\alpha\uparrow 1} D_\alpha\left(P_{\theta_2}, P_{\theta_1}\right) \;=\; D_1\left(P_{\theta_2}, P_{\theta_1}\right) \;=\; D_{t\ln t}\left(P_{\theta_2}, P_{\theta_1}\right) \tag{64}$$

Proof. (I) Let $\alpha(\alpha-1)\neq 0$ and $\theta_1,\,\theta_2\in\Theta$. By (1) and (27)

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

Hence, from the definitions (2) and (28) 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 (61) follows from the relations

$$\begin{aligned} D_{1-\alpha}\left(P_{\theta_2},P_{\theta_1}\right) &\equiv D_{\phi_\alpha^*}\left(P_{\theta_2},P_{\theta_1}\right) \\ &= D_{\phi_\alpha}\left(P_{\theta_1},P_{\theta_2}\right)\equiv D_\alpha\left(P_{\theta_1},P_{\theta_2}\right) \quad \text{(cf. (9))}. \end{aligned}$$

Alternatively, (61) follows from (60) using the skew symmetry

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

which is evident from (59).

(II) The equalities (62) and (64) follow from the already proved skew symmetry (61) and from the definition of the α-divergences of orders $\alpha=0$ and $\alpha=1$ in (37), (36). It remains to prove that the limit in (62) equals (63). 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 (60), 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

$$\begin{aligned} -\lim_{\alpha\to 0}\frac{\rho_\alpha(\theta_1,\,\theta_2)}{\alpha} &\equiv \lim_{\alpha\to 0}\frac{\alpha\,b(\theta_1)+(1-\alpha)\,b(\theta_2)-b(\alpha\,\theta_1+(1-\alpha)\,\theta_2)}{\alpha} \quad \text{(cf. (59))} \\ &= b(\theta_1)-b(\theta_2)+\nabla b(\theta_2)\,(\theta_2-\theta_1). \end{aligned}$$

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 (59) 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 by the formula

$$\sigma_\alpha(\theta_0,\,\theta_1,\,\theta_2)=\sigma_\alpha^I(\theta_0,\,\theta_1,\,\theta_2)-\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2) \tag{65}$$

with the nonnegative (possibly infinite)

$$\sigma_\alpha^I(\theta_0,\,\theta_1,\,\theta_2)=b\Big(\alpha\,\theta_1+(1-\alpha)\ [\theta_1-\theta_2+\theta_0]\Big) \tag{66}$$

and with the finite

$$\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2)=\alpha\,b(\theta_1)+(1-\alpha)\,\Big[b(\theta_1)-b(\theta_2)+b(\theta_0)\Big]\ . \tag{67}$$

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} relative to P_{θ_0} is given by the formula

$$B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) = \frac{\exp \rho_\alpha(\theta_1, \theta_0)}{\alpha(\alpha-1)} + \frac{\exp \rho_\alpha(\theta_2, \theta_0)}{\alpha} + \frac{\exp \sigma_\alpha(\theta_0, \theta_1, \theta_2)}{1-\alpha}\,. \tag{68}$$

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

$$B_0\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \;=\; b(\theta_1) - b(\theta_2) - \nabla b(\theta_0)\,(\theta_1 - \theta_2) + \exp \sigma_0(\theta_0, \theta_1, \theta_2) - 1 \tag{69}$$

respectively

$$B_1\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) = b(\theta_2) - b(\theta_1) - \nabla b(\theta_1)\,(\theta_2 - \theta_1)\,. \tag{70}$$

Proof. (I) By (52) it holds for every $\alpha \in \mathbb{R}$ and $\theta_0, \theta_1,\, \theta_2 \in \Theta$

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

with $\sigma_\alpha^{II}(\theta_0,\,\theta_1,\,\theta_2)$ from (67). Since (54) leads to

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

for $\sigma_\alpha^{I}(\theta_0,\,\theta_1,\,\theta_2)$ given by (66), it holds

$$\int_{\mathcal{X}} \left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\,\mathrm{d}\mu \;=\; \exp \sigma_\alpha(\theta_0, \theta_1, \theta_2) \tag{71}$$

where $\sigma_\alpha(\theta_0, \theta_1, \theta_2)$ was defined in (65). Now, by taking in (32) the exponential family distributions

$$P_1 = P_{\theta_1}, \quad P_2 = P_{\theta_2}, \quad Q = P_{\theta_0} \quad \text{(cf. (52))},$$

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

$$\begin{aligned}
B_\alpha\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \;&=\; D_\alpha\left(P_{\theta_1}, P_{\theta_2}\right) - (1-\alpha)\, D_\alpha\left(P_{\theta_2}, P_{\theta_0}\right) \\
&\quad + \frac{1}{1-\alpha}\left[\int_{\mathcal{X}} \left(\frac{p_{\theta_2}}{p_{\theta_0}}\right)^{\alpha-1} p_{\theta_1}\,\mathrm{d}\mu - 1\right].
\end{aligned} \tag{72}$$

Applying the power divergence formula (60) together with (71) to (72), one obtains the desired formula (68).

(II) By the representation of $B_0(P_1, P_2 \,|\, Q)$ in (46) and by (38)

$$B_0\left(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0}\right) \;=\; D_0\left(P_{\theta_1}, P_{\theta_0}\right) - D_0\left(P_{\theta_2}, P_{\theta_0}\right) + \int_{\mathcal{X}} \frac{p_{\theta_0} p_{\theta_1}}{p_{\theta_2}}\,\mathrm{d}\mu - 1$$

where

$$\int_{\mathcal{X}} \frac{p_{\theta_0}\, p_{\theta_1}}{p_{\theta_2}}\, \mathrm{d}\mu \;=\; \exp \sigma_0(\theta_0,\, \theta_1,\, \theta_2) \qquad \text{(cf. (71))}.$$

For $\theta_0 \in \mathring{\Theta}$ the desired assertion (69) follows from here and from the formulas

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

obtained from (63).

(III) The desired formula (70) follows immediately from (45), (41), (42), (63) and (64). □

Remarks. (i) Since $\theta_1,\, \theta_2$ lie in the convex subset Θ of $\mathbb{R}^d$ and the function $b(\cdot)$ is convex as well as differentiable in $\mathring{\Theta}$, formula (70) suggests that $B_1\,(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0})$ can be interpreted as the original classical definition of a Bregman distance (with respect to the scaling function b) on the multidimensional Euclidean space. But this also means that the formulas (69) and (68) give direct alternatives for classical Bregman distances where (68) tends to the classical definition as α tends to 1.
(ii) We see from Theorems 4 and 5 that – in consistency with (20), (42) – for arbitrary interior parameters $\theta_0,\, \theta_1,\, \theta_2 \in \mathring{\Theta}$

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

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 "background distribution" P_{θ_0}. The distance of order $\alpha = 0$ satisfies the relation

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

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 $B_0\,(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0})$ and $B_1\,(P_{\theta_2}, P_{\theta_1} \,|\, P_{\theta_0})$ of P_{θ_1} and P_{θ_2}. This deviation is zero if (for strictly convex $b(\theta)$ if and only if) $\theta_0 = \theta_2$.
(iii) From the formulas (57), (58), (63), (65), (66), (67), (68), (69) and (70) one can see immediately that for all $\alpha \in \mathbb{R}$ the quantities $D_\alpha\,(P_{\theta_1}, P_{\theta_2})$, $\rho_\alpha(\theta_1,\, \theta_2)$, $\sigma_\alpha(\theta_0,\, \theta_1,\, \theta_2)$ and $B_\alpha\,(P_{\theta_1}, P_{\theta_2} \,|\, P_{\theta_0})$ only depend on the function $b(\cdot)$ defined in (50), and *not* directly on the reference measure μ used in the definition formulas (50), (52).

References

Azoury, K.S., and Warmuth, M.K. (2001): Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, **43**, pp. 211-246.

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

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* **7**, No. 3, pp. 200-217.

Burger, M., Resmerita, E. and He, L. (2008): Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing*, **81**, No. 2-3, pp. 109-135.

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

Cai, J-F., Osher, S., and Shen, Z. (2008): Linearized Bregman iterations for frame-based image deblurring. *UCLA Computational and Applied Mathematics Reports 08-57.* Los Angeles, UCLA, pp. 1-26.

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

Censor, Y. and Lent, A. (1981): An iterative row-action method for interval convex programming. *J. Optimiz. Theory Applic.*, **34**, No. 3, pp. 321-353.

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

Csiszár, I. (1991): Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics*, **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*, **68**, pp. 161-185.

Della Pietra, S., Della Pietra, V., and Lafferty, J.D. (2002): Duality and auxiliary functions for Bregman distances. *Technical Report CMU-CS-01-109R.* Pittsburgh, Carnegie Mellon University, pp. 1-13.

Eggermont, P.P.B. (1993): Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM J. Math. Anal.*, **24**, No.6, pp. 1557-1576.

Gao, Y., Lim, K.G. and Ng, K.H. (2004): An approximation pricing algorithm in an incomplete market: a differential geometric approach. *Finance and Stochastics* **8**, pp. 501-523.

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* **36**, No.1, pp. 23-30.

Jones, L.K. and Trutzer, V. (1989): Computationally feasible high-resolution min imum-distance procedures which extend the maximum-entropy method. *Inverse Problems* **5**, pp. 749-766.

Kivinen, J. and Warmuth, M.K. (1999): Boosting as entropy projection. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. New York, ACM Press, pp. 134-144.

Lafferty, J.D. (1999): Additive models, boosting, and inference for generalized divergences. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. New York, ACM Press, pp. 125-133.

Lafferty, J.D. Della Pietra, S. and Della Pietra, V. (1997): Statistical learning algorithms based on Bregman distances. *Proceedings of the 1997 Canadian Workshop on Information Theory*. Toronto, Fields Institute, pp. 77-80.

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* **52**, No. 10, pp. 4394-4412.

Marquina, A. and Osher, S.J. (2008): Image super-resolution by TV-regularization and Bregman iteration. *J. Sci. Comput.*, **375**, pp. 367-382.

Murata, N., Takenouchi, T., Kanamori, T., and Eguchi, S. (2004): Information geometry of $\mathcal{U}$-Boost and Bregman divergence. *Neural Computation* **16**, pp. 1437-1481.

Osher, S., Mao, Y., Dong, B. and Yin, W. (2008): Fast linearized Bregman iteration for compressive sensing and sparse denoising. *UCLA Computational and Applied Mathematics Reports 08-37*. Los Angeles, UCLA, pp. 1-21.

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* **43**, No. 4, pp. 1288-1293.

Pardo, M.C. and Vajda, I. (1997): On asymptotic properties of information-theoretic divergences. *IEEE Transaction on Information theory* **49**, No. 7, pp. 1860-1868.

Rényi, A. (1961): On measures of entropy and information. *Proceedings of 4-th Berekeley Symposium on Probab. Theory and Math. Statistics* **2**, pp. 547-561. Univ. of California Press, Berkeley, CA.

Resmerita, E. (2005): Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems*, **21**, pp. 1303-1314.

Resmerita, E. and Anderssen R.S. (2007): Joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems. *Math. Meth. Appl. Sci.*, **30**, pp. 1527-1544.

Resmerita, E. and Scherzer, O. (2006): Error estimates for non-quadratic regularization and the relation to enhancement. *Inverse Problems*, **22**, pp. 801-814.

Scherzer, O., Grasmair, M.,Grossauer, H., Haltmeier, M. and Lenzen, F. (2008): *Variational methods of imaging*. Berlin, Springer.

Silva Neto, A.J. and Cella, N. (2006): A regularized solution with weighted Bregman distances for the inverse problem of photoacoustic spectroscopy. *Computational and Applied Mathematics*, **25**, No. 2-3, pp. 139-165.

Stummer, W. (2007): Some Bregman distances between financial diffusion processes. *Proceedings in Applied Mathematics and Mechanics*, to appear.

Topsoe, F. (2007): Exponential families and MaxEnt calculations for entropy measures of statistical physics. *Preprint*. Copenhagen, University of Copenhagen.

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