

Functional Bregman Divergence

Bela A. Frigiyik

Department of Mathematics
Purdue University
West Lafayette, IN 47907, USA
Email: bfrigiyik@math.purdue.edu

Santosh Srivastava

Fred Hutchinson Cancer Research Center
Seattle, WA 98109, USA
Email: ssvrast@fhcrc.org

Maya R. Gupta

Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
Email: gupta@ee.washington.edu

Abstract—To characterize the differences between two positive functions or two distributions, a class of distortion functions has recently been defined termed the functional Bregman divergences. The class generalizes the standard Bregman divergence defined for vectors, and includes total squared difference and relative entropy. Recently a key property was discovered for the vector Bregman divergence: that the mean minimizes the average Bregman divergence for a finite set of vectors. In this paper the analog result is proven: that the mean function minimizes the average Bregman divergence for a set of positive functions that can be parameterized by a finite number of parameters. In addition, the relationship of the functional Bregman divergence to the vector Bregman divergence and pointwise Bregman divergence is stated, as well as some important properties.

I. INTRODUCTION

Bregman divergences are a class of distortion functions that include squared error, relative entropy, logistic loss, Mahalanobis distance, and the Itakura-Saito function. Bregman divergences are popular in statistical estimation and information theory. Analysis with Bregman divergences has led to recent advances in statistical learning [1]–[8], clustering [9], [10], inverse problems [11], maximum entropy estimation [12], and the applicability of the data processing theorem [13]. A key property of Bregman divergences is that the mean is the minimizer of the expected Bregman divergence for a set of d -dimensional points [9], [14].

Recently, we defined a functional Bregman divergence to characterize the distortion between two non-negative functions or two distributions rather than between two vectors [6]. The main result of this paper is showing that the mean minimizes the expected functional Bregman divergence. In addition, we state the relationship between functional Bregman divergence and the standard vector Bregman divergence and the pointwise Bregman divergence previously defined for measurable functions [8], [15]. The functional Bregman definition includes a larger class of distortion functions and enables stronger results through the use of calculus of variations machinery.

First, we review the standard vector and pointwise Bregman definitions. Then, we define the functional Bregman divergence, note how it relates to the vector and pointwise Bregman divergences, and give examples for total squared difference and squared bias. Next, we present and prove the main theorem: that the expectation of a weighted set of functions minimizes the expected Bregman divergence. Last,

we note that many of the properties of the standard Bregman divergence also hold for the functional Bregman divergence.

II. BREGMAN DIVERGENCE DEFINITIONS

Given vectors $x, y \in \mathbb{R}^n$ and strictly convex and twice-differentiable $\tilde{\phi} : \mathbb{R}^n \rightarrow \mathbb{R}$, the standard vector Bregman divergence is

$$d_{\tilde{\phi}}(x, y) = \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla \tilde{\phi}(y)^T(x - y). \quad (1)$$

By re-arranging the terms of (1), one sees that the Bregman divergence is the tail of the Taylor series expansion of $\tilde{\phi}$ around y :

$$\tilde{\phi}(x) = \tilde{\phi}(y) + \nabla \tilde{\phi}(y)^T(x - y) + d_{\tilde{\phi}}(x, y).$$

Jones and Byrne analyzed a general class of divergences between functions using a pointwise formulation [8]. Csiszár specialized the pointwise formulation to a class of divergences he termed *Bregman distances* $B_{s,\nu}$ [15], where given a σ -finite measure space (X, Ω, ν) , and non-negative measurable functions $f(x)$ and $g(x)$, $B_{s,\nu}(f, g)$ equals

$$\int s(f(x)) - s(g(x)) - s'(g(x))(f(x) - g(x))d\nu(x). \quad (2)$$

The function $s : (0, \infty) \rightarrow \mathbb{R}$ is constrained to be differentiable and strictly convex, and the limit $\lim_{x \rightarrow 0} s(x)$ and $\lim_{x \rightarrow 0} s'(x)$ must exist, but not necessarily be finite. The function s plays a role similar to the function ϕ in the functional Bregman divergence; however, s acts on the range of the functions f, g , whereas ϕ acts on the pair of functions f, g . The pointwise definition is useful [8], [15], but restrictive as the last operation is constrained to be an integration over the functions' domain.

Our functional Bregman divergence definition requires further preliminaries. Let $(\mathbb{R}^d, \Omega, \nu)$ be a measure space, where ν is a Borel measure, d is a positive integer, and define a set of functions $\mathcal{A} = \{a \in L^p(\nu) \text{ subject to } a : \mathbb{R}^d \rightarrow \mathbb{R}, a \geq 0\}$ where $1 \leq p \leq \infty$. The gradient in (1) must be replaced by an analog of the derivative for functions. To this end, we review the definition of the Fréchet derivative [16]. Let ϕ be a real functional over the normed space $L^p(\nu)$. The bounded linear functional $\delta\phi[f; \cdot]$ is the Fréchet derivative of ϕ at $f \in L^p(\nu)$ if

$$\begin{aligned} \phi[f + a] - \phi[f] &= \Delta\phi[f; a] \\ &= \delta\phi[f; a] + \epsilon[f, a] \|a\|_{L^p(\nu)} \end{aligned} \quad (3)$$

for all $a \in L^p(\nu)$, with $\epsilon[f, a] \rightarrow 0$ as $\|a\|_{L^p(\nu)} \rightarrow 0$.

Let $\phi : L^p(\nu) \rightarrow \mathbb{R}$ be a strictly convex, twice-continuously Fréchet-differentiable functional. The functional Bregman divergence $d_\phi : \mathcal{A} \times \mathcal{A} \rightarrow [0, \infty)$ is defined for all $f, g \in \mathcal{A}$ as

$$d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g], \quad (4)$$

where $\delta\phi[g; \cdot]$ is the Fréchet derivative of ϕ at g .

The functional definition in (4) is less restrictive than the pointwise definition given in (2) and forms a clearer analog to the vector definition of (1). Two propositions establish the formal relationship of the functional Bregman divergence to the other definitions. The proofs are given in the arXiv version of this paper [17].

Proposition II.1. *The functional Bregman divergence given in (4) is a generalization of the standard vector Bregman divergence given in (1).*

Proposition II.2. *Given a pointwise Bregman divergence as per (2), an equivalent functional Bregman divergence can be defined as per (4) if the measure ν is finite. However, given a functional Bregman divergence $d_\phi[f, g]$, there is not necessarily an equivalent pointwise Bregman divergence.*

A. Functional Bregman Divergence Examples

Different choices of the functional ϕ in (4) lead to different Bregman divergences. We illustrate two functional Bregman divergences that correspond to total squared difference and squared bias. Functionals for other Bregman divergences can be derived based on these examples, from the example functions for the discrete case given in Table 1 of [14], and from the fact that ϕ is a strictly convex functional if it has the form $\phi(g) = \int \tilde{\phi}(g(t))dt$ where $\tilde{\phi} : \mathbb{R} \rightarrow \mathbb{R}$, $\tilde{\phi}$ is strictly convex and g is in some well-defined vector space of functions [18].

1) *Ex: Total Squared Difference:* Let $\phi[g] = \int g^2 d\nu$, where $\phi : L^2(\nu) \rightarrow \mathbb{R}$, and let $g, f, a \in L^2(\nu)$. Then

$$\begin{aligned} \phi[g + a] - \phi[g] &= \int (g + a)^2 d\nu - \int g^2 d\nu \\ &= 2 \int gad\nu + \int a^2 d\nu. \end{aligned}$$

Because

$$\frac{\int a^2 d\nu}{\|a\|_{L^2(\nu)}^2} = \frac{\|a\|_{L^2(\nu)}^2}{\|a\|_{L^2(\nu)}^2} = \|a\|_{L^2(\nu)} \rightarrow 0$$

as $a \rightarrow 0$ in $L^2(\nu)$,

$$\delta\phi[g; a] = 2 \int gad\nu,$$

which is a continuous linear functional in a . Then, by definition of the second Fréchet derivative,

$$\begin{aligned} \delta^2\phi[g; b, a] &= \delta\phi[g + b; a] - \delta\phi[g; a] \\ &= 2 \int (g + b)ad\nu - 2 \int gad\nu \\ &= 2 \int bad\nu. \end{aligned}$$

Thus $\delta^2\phi[g; b, a]$ is a quadratic form, where $\delta^2\phi$ is actually independent of g and strongly positive since

$$\delta^2\phi[g; a, a] = 2 \int a^2 d\nu = 2\|a\|_{L^2(\nu)}^2$$

for all $a \in L^2(\nu)$, which implies that ϕ is strictly convex and

$$\begin{aligned} d_\phi[f, g] &= \int f^2 d\nu - \int g^2 d\nu - 2 \int g(f - g)d\nu \\ &= \int (f - g)^2 d\nu \\ &= \|f - g\|_{L^2(\nu)}^2. \end{aligned}$$

2) *Ex: Squared Bias:* Although we have not found it previously noted that squared bias is a Bregman divergence for vectors, it is easy to show and may be useful in estimation and statistics. However, squared bias between functions cannot be expressed using the pointwise Bregman definition of (2). Squared bias is a functional Bregman divergence though, as we now illustrate.

Let $\phi[g] = (\int g d\nu)^2$, where $\phi : L^1(\nu) \rightarrow \mathbb{R}$. In this case

$$\begin{aligned} \phi[g + a] - \phi[g] &= \left(\int g d\nu + \int a d\nu \right)^2 - \left(\int g d\nu \right)^2 \\ &= 2 \int g d\nu \int a d\nu + \left(\int a d\nu \right)^2. \end{aligned} \quad (5)$$

Note that $2 \int g d\nu \int a d\nu$ is a continuous linear functional on $L^1(\nu)$ and $(\int a d\nu)^2 \leq \|a\|_{L^1(\nu)}^2$, so that

$$0 \leq \frac{(\int a d\nu)^2}{\|a\|_{L^1(\nu)}} \leq \frac{\|a\|_{L^1(\nu)}^2}{\|a\|_{L^1(\nu)}} = \|a\|_{L^1(\nu)}.$$

Thus from (5) and the definition of the Fréchet derivative,

$$\delta\phi[g; a] = 2 \int g d\nu \int a d\nu.$$

By the definition of the second Fréchet derivative,

$$\begin{aligned} \delta^2\phi[g; b, a] &= \delta\phi[g + b; a] - \delta\phi[g; a] \\ &= 2 \int (g + b)d\nu \int a d\nu - 2 \int g d\nu \int a d\nu \\ &= 2 \int b d\nu \int a d\nu \end{aligned}$$

is another quadratic form, and $\delta^2\phi$ is independent of g .

Because the functions in \mathcal{A} are positive, $\delta^2\phi$ is strongly positive on \mathcal{A} (which again implies that ϕ is strictly convex):

$$\delta^2\phi[g; a, a] = 2 \left(\int a d\nu \right)^2 = 2\|a\|_{L^1(\nu)}^2 \geq 0$$

for $a \in \mathcal{A}$. The Bregman divergence is thus

$$\begin{aligned} d_\phi[f, g] &= \left(\int f d\nu \right)^2 - \left(\int g d\nu \right)^2 - 2 \int g d\nu \int (f - g)d\nu \\ &= \left(\int f d\nu \right)^2 + \left(\int g d\nu \right)^2 - 2 \int g d\nu \int f d\nu \\ &= \left(\int (f - g)d\nu \right)^2. \end{aligned}$$

III. THE MEAN MINIMIZES THE EXPECTED FUNCTIONAL BREGMAN DIVERGENCE

Consider a set of functions (or distributions), \mathcal{M} . Let $F \in \mathcal{M}$ be a random function with realization f . Suppose there exists a probability distribution P_F over the set \mathcal{M} , such that $P_F(f)$ is the probability of $f \in \mathcal{M}$. For example, consider the set of Gaussian distributions, and given samples drawn independently and identically from a randomly selected Gaussian distribution N , the data imply a posterior probability $P_N(\mathcal{N})$ for each possible generating realization of a Gaussian distribution \mathcal{N} . The goal is to find the function g^* that has minimum expected functional Bregman divergence with respect to the random function F . The following theorem shows that if the mean function $E_F[F]$ exists, then $g^* = E_F[F]$ minimizes the expected functional Bregman divergence for any choice of functional Bregman divergence.

The theorem applies only to a set of functions \mathcal{M} that lie on a finite-dimensional manifold M for which a differential element dM can be defined. For example, the set \mathcal{M} could be a parametric distribution that is parameterized by a finite number of parameters, or could be a set of functions that can be decomposed into a finite set of d basis functions $\{\psi_1, \psi_2, \dots, \psi_d\}$ such that each f can be expressed as

$$f = \sum_{j=1}^d c_j \psi_j,$$

where $c_j \in \mathbb{R}$ for all j . The theorem requires slightly stronger conditions on ϕ than the definition of the functional Bregman divergence (4) requires.

Theorem III.1 (Minimizer of the Expected Bregman Divergence). *Let $\delta^2\phi[f; a, a]$ be a strongly positive quadratic form, and let $\phi \in \mathcal{C}^3(L^1(\nu); \mathbb{R})$ be a three-times continuously Fréchet-differentiable functional on $L^1(\nu)$. Let \mathcal{M} be a set of functions f that lie on a finite-dimensional manifold M , and have associated differential element dM . Suppose there is a probability distribution P_F defined over the set \mathcal{M} . Let g^* be defined by*

$$g^* = \arg \inf_{g \in \mathcal{A}} E_F[d_\phi(F, g)].$$

Then, if g^* exists, it is given by

$$g^* = \int_M f P(f) dM = E_F[F]. \quad (6)$$

Proof: To prove the theorem we must review the functional optimality conditions and some related facts [16]. For a functional J to have an extremum (minimum) at $f = \hat{f}$, it is necessary that

$$\delta J[\hat{f}; a] = 0 \quad \text{and} \quad \delta^2 J[\hat{f}; a, a] \geq 0,$$

for all admissible functions $a \in \mathcal{A}$, where for this proof $p = 1$ in the definition of \mathcal{A} . A sufficient condition for a functional $J[f]$ to have a minimum for $f = \hat{f}$ is that the first variation $\delta J[f; a]$ must vanish for $f = \hat{f}$, and its second variation $\delta^2 J[f; a, a]$ must be **strongly positive** for $f = \hat{f}$; we define

strongly positive shortly. When the second variation $\delta^2\phi$ and the third variation $\delta^3\phi$ exist, they are described by

$$\begin{aligned} \Delta\phi[f; a] &= \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + \epsilon[f, a] \|a\|_{L^p(\nu)}^2 \\ &= \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + \frac{1}{6}\delta^3\phi[f; a, a, a] \\ &\quad + \epsilon[f, a] \|a\|_{L^p(\nu)}^3, \end{aligned} \quad (7)$$

where $\epsilon[f, a] \rightarrow 0$ as $\|a\|_{L^p(\nu)} \rightarrow 0$. The term $\delta^2\phi[f; a, b]$ is bilinear with respect to arguments a and b , and $\delta^3\phi[f; a, b, c]$ is trilinear with respect to a, b , and c . The quadratic functional $\delta^2\phi[f; a, a]$ defined on normed linear space $L^p(\nu)$ is **strongly positive** if there exists a constant $k > 0$ such that $\delta^2\phi[f; a, a] \geq k \|a\|_{L^p(\nu)}^2$ for all $a \in \mathcal{A}$. In a finite-dimensional space, strong positivity of a quadratic form is equivalent to the quadratic form being positive definite. Also,

$$\begin{aligned} \phi[f + a] &= \phi[f] + \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + o(\|a\|^2), \\ \phi[f] &= \phi[f + a] - \delta\phi[f + a; a] + \\ &\quad \frac{1}{2}\delta^2\phi[f + a; a, a] + o(\|a\|^2), \end{aligned}$$

where $o(\|a\|^2)$ stands for a function that goes to zero as $\|a\|$ goes to zero, even if it is divided by $\|a\|^2$. Adding the above two equations yields

$$\begin{aligned} 0 &= \delta\phi[f; a] - \delta\phi[f + a; a] + \frac{1}{2}\delta^2\phi[f; a, a] \\ &\quad + \frac{1}{2}\delta^2\phi[f + a; a, a] + o(\|a\|^2), \end{aligned}$$

which is equivalent to

$$\delta\phi[f + a; a] - \delta\phi[f; a] = \delta^2\phi[f; a, a] + o(\|a\|^2), \quad (8)$$

because

$$\begin{aligned} &|\delta^2\phi[f + a; a, a] - \delta^2\phi[f; a, a]| \\ &\leq \|\delta^2\phi[f + a; \cdot, \cdot] - \delta^2\phi[f; \cdot, \cdot]\| \|a\|^2, \end{aligned}$$

and we assumed $\phi \in \mathcal{C}^2$, so $\delta^2\phi[f + a; a, a] - \delta^2\phi[f; a, a]$ is of order $o(\|a\|^2)$. This shows that the variation of the first variation of ϕ is the second variation of ϕ . A procedure like the above can be used to prove that analogous statements hold for higher variations if they exist.

Now we begin the main body of the proof. Let

$$\begin{aligned} J[g] &= E_F[d_\phi(F, g)] = \int_M d_\phi[f, g] P(f) dM \\ &= \int_M (\phi[f] - \phi[g] - \delta\phi[g; f - g]) P(f) dM, \end{aligned} \quad (9)$$

where (9) follows by substituting the definition of functional Bregman divergence given in (4). Consider the increment

$$\Delta J[g; a] = J[g + a] - J[g] \quad (10)$$

$$\begin{aligned} &= - \int_M (\phi[g + a] - \phi[g]) P(f) dM \\ &\quad - \int_M (\delta\phi[g + a; f - g - a] \\ &\quad - \delta\phi[g; f - g]) P(f) dM, \end{aligned} \quad (11)$$

where (11) follows from substituting (9) into (10). Using the definition of the differential of a functional, the first integrand in (11) can be written as

$$\phi[g + a] - \phi[g] = \delta\phi[g; a] + \epsilon[g, a] \|a\|_{L^1(\nu)}. \quad (12)$$

Take the second integrand of (11), and subtract and add $\delta\phi[g; f - g - a]$,

$$\begin{aligned} &\delta\phi[g + a; f - g - a] - \delta\phi[g; f - g] \\ &= \delta\phi[g + a; f - g - a] - \delta\phi[g; f - g - a] \\ &\quad + \delta\phi[g; f - g - a] - \delta\phi[g; f - g] \\ &\stackrel{(a)}{=} \delta^2\phi[g; f - g - a, a] + \epsilon[g, a] \|a\|_{L^1(\nu)} + \delta\phi[g; f - g] \\ &\quad - \delta\phi[g; a] - \delta\phi[g; f - g] \\ &\stackrel{(b)}{=} \delta^2\phi[g; f - g, a] - \delta^2\phi[g; a, a] + \epsilon[g, a] \|a\|_{L^1(\nu)} \\ &\quad - \delta\phi[g; a] \end{aligned} \quad (13)$$

where (a) follows from (8) and the linearity of the third term, and (b) follows from the linearity of the first term. Substitute (12) and (13) into (11),

$$\begin{aligned} \Delta J[g; a] &= - \int_M \left(\delta^2\phi[g; f - g, a] - \delta^2\phi[g; a, a] \right. \\ &\quad \left. + \epsilon[g, a] \|a\|_{L^1(\nu)} \right) P(f) dM. \end{aligned}$$

Note that the term $\delta^2\phi[g; a, a]$ is of order $\|a\|_{L^1(\nu)}^2$, that is, $\|\delta^2\phi[g; a, a]\|_{L^1(\nu)} \leq m \|a\|_{L^1(\nu)}^2$ for some constant m . Therefore,

$$\lim_{\|a\|_{L^1(\nu)} \rightarrow 0} \frac{\|J[g + a] - J[g] - \Delta J[g; a]\|_{L^1(\nu)}}{\|a\|_{L^1(\nu)}} = 0,$$

where,

$$\delta J[g; a] = - \int_M \delta^2\phi[g; f - g, a] P(f) dM. \quad (14)$$

For fixed a , $\delta^2\phi[g; \cdot, a]$ is a bounded linear functional in the second argument, so the integration and the functional can be interchanged in (14), which becomes

$$\delta J[g; a] = -\delta^2\phi \left[g; \int_M (f - g) P(f) dM, a \right]. \quad (15)$$

From the functional optimality conditions reviewed at the beginning of this proof, $J[g]$ has an extremum for $g = \hat{g}$ if $\delta J[\hat{g}; a] = 0$, and thus from (15) if,

$$\delta^2\phi \left[\hat{g}; \int_M (f - \hat{g}) P(f) dM, a \right] = 0. \quad (16)$$

Set $a = \int_M (f - \hat{g}) P(f) dM$ in (16) and use the assumption that the quadratic functional $\delta^2\phi[g; a, a]$ is strongly positive, which implies that the above functional can be zero if and only if $a = 0$, that is,

$$0 = \int_M (f - \hat{g}) P(f) dM, \quad (17)$$

$$\hat{g} = E_F[F], \quad (18)$$

where the last line holds if the expectation exists (i.e. if the measure is well-defined and the expectation is finite). Because a Bregman divergence is not necessarily convex in its second argument, it is not yet established that the above unique extremum is a minimum. To see that (18) is in fact a minimum of $J[g]$, from the functional optimality conditions it is enough to show that $\delta^2 J[\hat{g}; a, a]$ is strongly positive. To show this, for $b \in \mathcal{A}$, consider

$$\begin{aligned} &\delta J[g + b; a] - \delta J[g; a] \\ &\stackrel{(c)}{=} - \int_M (\delta^2\phi[g + b; f - g - b, a] \\ &\quad - \delta^2\phi[g; f - g, a]) P(f) dM \\ &\stackrel{(d)}{=} - \int_M (\delta^2\phi[g + b; f - g - b, a] - \delta^2\phi[g; f - g - b, a] \\ &\quad + \delta^2\phi[g; f - g - b, a] - \delta^2\phi[g; f - g, a]) P(f) dM \\ &\stackrel{(e)}{=} - \int_M (\delta^3\phi[g; f - g - b, a, b] + \epsilon[g, a, b] \|b\|_{L^1(\nu)} \\ &\quad + \delta^2\phi[g; f - g, a] - \delta^2\phi[g; b, a] \\ &\quad - \delta^2\phi[g; f - g, a]) P(f) dM \\ &\stackrel{(f)}{=} - \int_M (\delta^3\phi[g; f - g, a, b] - \delta^3\phi[g; b, a, b] \\ &\quad + \epsilon[g, a, b] \|b\|_{L^1(\nu)} - \delta^2\phi[g; b, a]) P(f) dM, \end{aligned} \quad (19)$$

where (c) follows from using integral (14); (d) from subtracting and adding $\delta^2\phi[g; f - g - b, a]$; (e) from the fact that the variation of the second variation of ϕ is the third variation of ϕ [19]; and (f) from the linearity of the first term and cancellation of the third and fifth terms. Note that in (19) for fixed a , the term $\delta^3\phi[g; b, a, b]$ is of order $\|b\|_{L^1(\nu)}^2$, while the first and the last terms are of order $\|b\|_{L^1(\nu)}$. Therefore,

$$\lim_{\|b\|_{L^1(\nu)} \rightarrow 0} \frac{\|\delta J[g + b; a] - \delta J[g; a] - \delta^2 J[g; a, b]\|_{L^1(\nu)}}{\|b\|_{L^1(\nu)}} = 0,$$

where

$$\begin{aligned} \delta^2 J[g; a, b] &= - \int_M \delta^3\phi[g; f - g, a, b] P(f) dM \\ &\quad + \int_M \delta^2\phi[g; a, b] P(f) dM. \end{aligned} \quad (20)$$

Substitute $b = a$, $g = \hat{g}$ and interchange integration and the

continuous functional $\delta^3\phi$ in the first integral of (20), then

$$\begin{aligned}\delta^2 J[\hat{g}; a, a] &= -\delta^3\phi\left[\hat{g}; \int_M (f - \hat{g})P(f)dM, a, a\right] \\ &\quad + \int_M \delta^2\phi[\hat{g}; a, a]P(f)dM \\ &= \int_M \delta^2\phi[\hat{g}; a, a]P(f)dM\end{aligned}\quad (21)$$

$$\begin{aligned}&\geq \int_M k \|a\|_{L^1(\nu)}^2 P(f)dM \\ &= k \|a\|_{L^1(\nu)}^2 > 0,\end{aligned}\quad (22)$$

where (21) follows from (17), and (22) follows from the strong positivity of $\delta^2\phi[\hat{g}; a, a]$. Therefore, from (22) and the functional optimality conditions, \hat{g} is the minimum.

IV. PROPERTIES OF THE FUNCTIONAL BREGMAN DIVERGENCE

The vector Bregman divergence has some useful properties, as reviewed in [9, Appendix A]. Here, we note some of these properties that hold for the functional Bregman divergence (4). A more complete list of properties and their proofs can be found in the arXiv version of this paper [17].

1. Non-negativity

The functional Bregman divergence is non-negative.

2. Convexity

The Bregman divergence $d_\phi[f, g]$ is always convex with respect to f .

3. Linearity

The functional Bregman divergence is linear in the sense that

$$\begin{aligned}d_{(c_1\phi_1+c_2\phi_2)}[f, g] &= (c_1\phi_1 + c_2\phi_2)[f] - (c_1\phi_1 + c_2\phi_2)[g] - \\ &\quad \delta(c_1\phi_1 + c_2\phi_2)[g; f - g], \\ &= c_1d_{\phi_1}[f, g] + c_2d_{\phi_2}[f, g].\end{aligned}$$

3. Dual Divergence

Given a pair (g, ϕ) where $g \in L^p(\nu)$ and ϕ is a strictly convex twice-continuously Fréchet-differentiable functional, then the function-functional pair (G, ψ) is the Legendre transform of (g, ϕ) [16], if

$$\phi[g] = -\psi[G] + \int g(x)G(x)d\nu(x), \quad (23)$$

$$\delta\phi[g; a] = \int G(x)a(x)d\nu(x), \quad (24)$$

where ψ is a strictly convex twice-continuously Fréchet-differentiable functional, and $G \in L^q(\nu)$, where $\frac{1}{p} + \frac{1}{q} = 1$.

Given Legendre transformation pairs $f, g \in L^p(\nu)$ and $F, G \in L^q(\nu)$,

$$d_\phi[f, g] = d_\psi[G, F].$$

4. Generalized Pythagorean Inequality

For any $f, g, h \in \mathcal{A}$,

$$d_\phi[f, h] = d_\phi[f, g] + d_\phi[g, h] + \delta\phi[g; f - g] - \delta\phi[h; f - g].$$

REFERENCES

- [1] B. Taskar, S. Lacoste-Julien, and M. I. Jordan, "Structured prediction, dual extragradient and Bregman projections," *Journal of Machine Learning Research*, vol. 7, pp. 1627–1653, 2006.
- [2] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-Boost and Bregman divergence," *Neural Computation*, vol. 16, pp. 1437–1481, 2004.
- [3] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, vol. 48, pp. 253–285, 2002.
- [4] J. Kivinen and M. Warmuth, "Relative loss bounds for multidimensional regression problems," *Machine Learning*, vol. 45, no. 3, pp. 301–329, 2001.
- [5] J. Lafferty, "Additive models, boosting, and inference for generalized divergences," *Proc. of Conf. on Learning Theory (COLT)*, 1999.
- [6] S. Srivastava, M. R. Gupta, and B. A. Frigiyik, "Bayesian quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1287–1314, 2007.
- [7] S. Srivastava and M. R. Gupta, "Distribution-based Bayesian minimum expected risk for discriminant analysis," *Proc. of the IEEE Intl. Symposium on Information Theory*, 2006.
- [8] 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. on Information Theory*, vol. 36, pp. 23–30, 1990.
- [9] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.
- [10] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1223–1235, 2006.
- [11] G. LeBesenerais, J. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. on Information Theory*, vol. 45, pp. 1565–1577, 1999.
- [12] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," *Proc. of Conf. on Learning Theory (COLT)*, 2006.
- [13] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. on Information Theory*, vol. 43, no. 4, pp. 1288–1293, 1997.
- [14] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.
- [15] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, pp. 161–185, 1995.
- [16] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. USA: Dover, 2000.
- [17] B. A. Frigiyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence and Bayesian estimation of distributions," *In review for journal publication*, available on arXiv and at idl.ee.washington.edu.
- [18] T. Rockafellar, "Integrals which are convex functionals," *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 525–539, 1968.
- [19] C. H. Edwards, *Advanced Calculus of Several Variables*. New York: Dover, 1995.