

# Bregman divergence and density integration

Noboru Murata and Yu Fujimoto

Received on August 29, 2009 / Revised on October 4, 2009

**Abstract.** In this paper, a problem of integrating density functions is considered. First, the Bregman divergence in the space of positive finite measures is introduced, and properties of consistent subspaces and well-behaved parametric models associated with the Bregman divergence is investigated. Based on those results, a natural method for integration of estimates from the Bregman divergence is derived.

*Keywords.* Bregman divergence, consistent subspace, Pythagorean relation,  $u$ -model, density integration.

## 1. INTRODUCTION

In recent years, the Bregman divergence receives a considerable attention in machine learning and statistics [4, 6, 8, 10, 12, 13]. In the machine learning community, the Bregman divergence is widely noticed as a class of well-behaved surrogate loss functions, and various effective algorithms are proposed based on it. In the statistical community, its robust properties are intensively investigated, and various probability estimation methods under heavy contamination are proposed. Thus we have a number of methods for finding a single estimate based on the Bregman divergence.

When we have various estimates from different data sets, we sometimes need to integrate them appropriately. For example, if we have too many data to train our probability models, we take a “divide and conquer” strategy in order to reduce computational burden and time. We partition data into small portions, train multiple models with different portions of data, and concatenate trained models in some way. Another example is a case of aggregating multiple opinions of experts. Assume that experts are trained by their individual experiences and express their opinions in the form of probability distributions. To make the final decision, we have to aggregate the experts’ distributions in some way. Thus, integration of probabilities is not a strange or unusual situation in statistical inference and machine learning [2, 7, 9, 11].

In this paper, we discuss integration of multiple density functions of positive finite measures estimated from the Bregman divergence. We consider the case that each density function is estimated based on a different and independent data set. By investigating properties of the Bregman divergence and associated models, we try to derive a natural way of integrating multiple density functions. The rest of paper is organized as follows. In section 2, we give a definition of the Bregman divergence in a space of positive finite measures in two different ways, and in section

3, we define three subspaces in which consistent estimation is guaranteed based on the property of the Bregman divergence. In section 4, a parametric model associated with the Bregman divergence is introduced, and its property is investigated in terms of so-called Pythagorean relation. In section 5, two mixtures of density functions are considered and their characterization is discussed. Our main results for integrating density functions are given in section 6, and section 7 is devoted for concluding remarks.

## 2. BREGMAN DIVERGENCE

We start from a formal definition of the Bregman divergence. Let  $U$  be a monotonically increasing convex function on  $\mathbb{R}$ , and the derivative of  $U$  is denoted by  $u$ . Let  $\Xi$  be the Legendre transform of  $U$ ,

$$(1) \quad \Xi(\zeta) = \sup_{z \in \mathbb{R}} z\zeta - U(z),$$

and the derivative of  $\Xi$  is denoted by  $\xi$ .

In this paper, a transformation of  $f$  by  $\mathcal{T}$  is called  $\mathcal{T}$ -representation of  $f$ , and we mainly consider following two representations,

$$(2) \quad m\text{-representation: } f = id(f),$$

$$(3) \quad u\text{-representation: } \check{f} = \xi(f),$$

where  $id$  is the identity function. Using both representations,  $f$  is written in two different forms as

$$f = id(f) = u(\check{f}).$$

Note that the  $m$ -representation is a trivial transformation, however, a pair of the  $m$ - and  $u$ -representations plays an important role in order to understand dualistic properties of the Bregman divergence as shown in later.

Now we give a definition of the Bregman divergence.

**Definition 1.** A quadruplet  $(U, \Xi, u, \xi)$  defines the Bregman divergence between points  $f$  and  $g$  as

$$(4) \quad d_U(f, g) = \Xi(f) + U(\check{g}) - f\check{g}.$$

Observing

$$\xi(\zeta) = u^{-1}(\zeta)$$

from its definition and

$$U(z') - U(z) \geq u(z)(z' - z)$$

from the convexity of  $U$ , the positivity of the Bregman divergence

$$d_U(f, g) = U(\xi(g)) - U(\xi(f)) - f(\xi(g) - \xi(f)) \geq 0$$

is easily confirmed.

In the rest of the paper, we consider a space of density functions of positive finite measures on  $\mathcal{X} \subset \mathbb{R}^m$  under a carrier measure  $\mu$ , that is defined as

$$(5) \quad \mathcal{M} = \left\{ m(\mathbf{x}) \mid m : \mathcal{X} \rightarrow \mathbb{R}_+, \int_{\mathcal{X}} m(\mathbf{x}) d\mu(\mathbf{x}) < \infty \right\}.$$

For descriptive simplicity, we introduce a notation for the inner product of  $f$  and  $g$  under  $\mu$  as

$$(6) \quad \int f(\mathbf{x})g(\mathbf{x}) d\mu(\mathbf{x}) = \langle f, g \rangle,$$

and also we abusively use it as

$$(7) \quad \int f(\mathbf{x}) d\mu(\mathbf{x}) = \langle f, 1 \rangle = \langle f \rangle.$$

For example,  $\mathcal{M}$  is simply written as

$$(8) \quad \mathcal{M} = \left\{ m(\mathbf{x}) \mid m : \mathcal{X} \rightarrow \mathbb{R}_+, \langle m \rangle < \infty \right\}.$$

Let  $P$  and  $Q$  be positive measures on  $\mathcal{X}$ , and let  $p$  and  $q$  be density functions of  $P$  and  $Q$ , respectively. To measure the discrepancy between  $P$  and  $Q$ , we define the Bregman divergence on  $\mathcal{M} \times \mathcal{M}$  as follows.

**Definition 2** (Bregman divergence). For  $p, q \in \mathcal{M}$ , the Bregman divergence is defined by

$$(9) \quad D_U(p, q) = \int d_U(p(\mathbf{x}), q(\mathbf{x})) d\mu(\mathbf{x}) \\ = \langle d_U(p, q) \rangle.$$

From a viewpoint of information theory, we can give another definition of the Bregman divergence. First, we introduce an extended entropy measure as follows [10].

**Definition 3** (Bregman entropy). For  $p, q \in \mathcal{M}$ ,  $U$ -cross-entropy is defined by

$$(10) \quad H_U(p, q) = \langle U(\check{q}) \rangle - \langle p, \check{q} \rangle \\ = \langle U(\xi(q)) \rangle - \langle p, \xi(q) \rangle.$$

Also  $U$ -entropy ( $U$ -auto-entropy) is defined by

$$(11) \quad H_U(p) = H_U(p, p) = \langle U(\check{p}) \rangle - \langle p, \check{p} \rangle.$$

Using the above extended entropies, the Bregman divergence is written as the difference between those entropies.

**Definition 4** (Bregman divergence). The Bregman divergence between  $p$  and  $q$  in  $\mathcal{M}$  is defined by

$$(12) \quad D_U(p, q) = H_U(p, q) - H_U(p).$$

Both of Eqs. (9) and (12) give the same definition

$$(13) \quad D_U(p, q) = \langle U(\check{q}) \rangle - \langle U(\check{p}) \rangle - \langle p, \check{q} - \check{p} \rangle,$$

and it is easily observed that

$$D_U(p, q) = 0$$

holds if and only if  $p(\mathbf{x}) = q(\mathbf{x})$  a.s.

**Example 1.** The followings are important examples of the convex function  $U$ .

- exponential (Kullback-Leibler divergence)

$$U(z) = \exp(z), \quad \Xi(\zeta) = \zeta(\log(\zeta) - 1), \\ u(z) = \exp(z), \quad \xi(\zeta) = \log(\zeta).$$

- $\beta$ -type ( $\beta$ -divergence)

$$U(z) = \frac{1}{\beta + 1}(\beta z + 1)^{\frac{\beta + 1}{\beta}}, \quad \Xi(\zeta) = \frac{\zeta^{\beta + 1}}{\beta(\beta + 1)} - \frac{\zeta}{\beta}, \\ u(z) = (\beta z + 1)^{\frac{1}{\beta}}, \quad \xi(\zeta) = \frac{\zeta^{\beta} - 1}{\beta}.$$

- $\eta$ -type ( $\eta$ -divergence)

$$U(z) = (1 - \eta)\exp(z) + \eta z, \\ \Xi(\zeta) = (\zeta - \eta) \left( \log \frac{\zeta - \eta}{1 - \eta} - 1 \right), \\ u(z) = (1 - \eta)\exp(z) + \eta, \\ \xi(\zeta) = \log \frac{\zeta - \eta}{1 - \eta}.$$

### 3. CONSISTENT SUBSPACES

So far, the Bregman divergence is defined in the space of the positive measures, however, we sometimes need to consider certain constrained subspaces of the positive measures in the context of statistical inference or machine learning. This is because an appropriate constraint helps us to construct a simple and effective algorithm. A typical example is the AdaBoost algorithm [5, 10], in which a combined classifier is searched in a specific subspace described with a moment matching condition under empirical distribution, which will be introduced later.

A fundamental subspace of  $\mathcal{M}$  is a space of probability densities defined as

$$(14) \quad \mathcal{P} = \left\{ p(\mathbf{x}) \mid p \in \mathcal{M}, \langle p \rangle = 1 \right\},$$

where the total mass of the measure is constrained to unity. This subspace is referred as a normalized subspace or a statistical subspace.

In addition, we consider the following two different constrained subspaces. One is called a moment matching subspace [10, 13], in which the inner product between the  $u$ -representation of the measure and a certain fixed measure  $p_0$  is constant,

$$(15) \quad \mathcal{Q} = \left\{ p(\mathbf{x}) \mid p \in \mathcal{M}, \langle p_0, \check{p} \rangle = \text{const.} \right\},$$

where *const.* is typically set to 0, and the reference point  $p_0$  is indicated as  $\mathcal{Q}(p_0)$  if necessary. The other is called a constant volume subspace [13], in which the volume of the  $u$ -representation measured with respect to  $U$  is constant,

$$(16) \quad \mathcal{R} = \left\{ p(\mathbf{x}) \mid p \in \mathcal{M}, \langle U(\check{p}) \rangle = \text{const.} \right\}.$$

In the following, the closest point from a point  $p$  in terms of  $D_U(p, \cdot)$  is called  $U$ -estimate of  $p$ . The  $U$ -estimate in the subspace  $\mathcal{Q}$  or  $\mathcal{R}$  has the following good property [13].

**Lemma 1.** *Let  $p$  be in  $\mathcal{M}$  and  $q$  be the  $U$ -estimate of  $p$  in  $\mathcal{Q}(p)$  or  $\mathcal{R}$ , that is, the minimizer of the Bregman divergence,*

$$(17) \quad q = \arg \min_{r \in \mathcal{Q}(p)} D_U(p, r) \quad \text{or} \quad q = \arg \min_{r \in \mathcal{R}} D_U(p, r).$$

Then  $q$  satisfies

$$(18) \quad \frac{q}{\langle q \rangle} = \frac{p}{\langle p \rangle}.$$

*Proof.* For  $q \in \mathcal{Q}(p)$ , the variation in  $D_U(p, q)$  with respect to  $q$  under the constraint

$$\langle p, \check{q} \rangle = \text{const.}$$

is given by

$$\begin{aligned} \delta D_U(p, q) - \lambda \delta \langle p, \check{q} \rangle &= \langle u(\check{q}), \delta \check{q} \rangle - \langle p, \delta \check{q} \rangle - \lambda \langle p, \delta \check{q} \rangle \\ &= \langle q - (1 + \lambda)p, \delta \check{q} \rangle \\ &= 0, \end{aligned}$$

that means

$$q = (1 + \lambda)p.$$

For  $q \in \mathcal{R}$ , the variation in  $D_U(p, q)$  with respect to  $q$  under the constraint

$$\langle U(\check{q}) \rangle = \text{const.}$$

is given by

$$\begin{aligned} \delta D_U(p, q) - \lambda \delta \langle U(\check{q}) \rangle &= \langle u(\check{q}), \delta \check{q} \rangle - \langle p, \delta \check{q} \rangle - \lambda \langle u(\check{q}), \delta \check{q} \rangle \\ &= \langle (1 - \lambda)q - p, \delta \check{q} \rangle \\ &= 0, \end{aligned}$$

that means

$$q = p/(1 - \lambda). \quad \square$$

Due to the consistency of the Bregman divergence, the  $U$ -estimate of  $p \in \mathcal{P}$  in  $\mathcal{P}$  is  $p$  itself, that is,

$$(19) \quad p = \arg \min_{r \in \mathcal{P}} D_U(p, r),$$

and this is a suitable property of the  $U$ -estimate in  $\mathcal{P}$  for the statistical inference. The above lemma claims a kind of consistency of the  $U$ -estimate in  $\mathcal{Q}$  and  $\mathcal{R}$  up to constant factor, and supports for constructing algorithms in those restricted subspaces.

#### 4. PYTHAGOREAN RELATION AND $u$ -MODEL

By considering the relationship between the  $m$ - and  $u$ -representations, we can construct an important model for  $U$ -estimation. First, we show an essential theorem obtained from duality of  $m$ - and  $u$ -representations [4, 10].

**Theorem 1** (Pythagorean relation). *Let  $p, q$  and  $r$  be in  $\mathcal{M}$ . If  $p - q$  and  $\check{r} - \check{q}$  (i.e.  $\xi(r) - \xi(q)$ ) are orthogonal*

$$(20) \quad \langle p - q, \check{r} - \check{q} \rangle = 0,$$

then the Pythagorean relation

$$(21) \quad D_U(p, r) = D_U(p, q) + D_U(q, r)$$

holds.

The statement is easily obtained by calculating  $D_U(p, r) - D_U(p, q) - D_U(q, r)$  with Eq. (13), and the complete proof can be found in [4, 10].

This theorem shows two important observations. One is that flat subspaces in terms of the  $m$ - and  $u$ -representations play a key role in geometry associated with the Bregman divergence. The flat subspaces are defined by

**$m$ -flat subspace:**  $\mathcal{M}_m$

$$(22) \quad p, q \in \mathcal{M}_m \Rightarrow \alpha p + (1 - \alpha)q \in \mathcal{M}_m, \quad 0 < \forall \alpha < 1,$$

**$u$ -flat subspace:**  $\mathcal{M}_u$

$$(23) \quad p, q \in \mathcal{M}_u \Rightarrow u(\alpha \check{p} + (1 - \alpha)\check{q}) \in \mathcal{M}_u, \quad 0 < \forall \alpha < 1.$$

The other is that the  $U$ -estimate  $q$  is regarded as an orthogonal projection with the  $m$ -representation from  $p$  to the  $u$ -flat subspace. Geometrical structures of estimation with the Bregman divergence are intensively discussed in [4, 10].

We can also introduce a parametric model which suits for statistical inference as follows [4, 8].

**Theorem 2** ( $u$ -model). *Let  $\mathbf{t}(\mathbf{x})$  be a  $d$ -dimensional vector-valued function on  $\mathcal{X}$ , and consider an equal mean subspace in  $\mathcal{P}$*

$$(24) \quad \Gamma_{\boldsymbol{\tau}} = \left\{ r(\mathbf{x}) \mid r \in \mathcal{P}, E_r[\mathbf{t}] = \langle r, \mathbf{t} \rangle = \boldsymbol{\tau} \right\}.$$

Let  $p$  be the maximum  $U$ -entropy function in  $\Gamma_{\boldsymbol{\tau}}$ , which is defined by

$$(25) \quad p = \arg \max_{r \in \Gamma_{\boldsymbol{\tau}}} H_U(r),$$

then  $p$  is found in the  $u$ -model defined by

$$(26) \quad p(\mathbf{x}) = u(\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\theta})),$$

where  $\boldsymbol{\theta}$  is a parameter vector in  $\mathbb{R}^d$  and  $b$  is a normalization constant to impose  $p \in \mathcal{P}$ .

Note that if  $f$  is a probability density function,  $\langle f, g \rangle$  becomes the expectation of  $g$ , which is denoted by  $E_f[g]$  in some cases.

*Proof.* The variation in  $H_U(p)$  under the constraint

$$\langle p, \mathbf{t} \rangle = \boldsymbol{\tau} \quad \text{and} \quad \langle p \rangle = 1$$

is given by

$$\begin{aligned} & \delta H_U(p) + \boldsymbol{\theta}^T \delta \langle p, \mathbf{t} \rangle + \lambda \delta \langle p \rangle \\ &= \langle u(\check{p}), \delta \check{p} \rangle - \langle p, \delta \check{p} \rangle - \langle \delta p, \check{p} \rangle + \boldsymbol{\theta}^T \langle \delta p, \mathbf{t} \rangle + \lambda \langle \delta p \rangle \\ &= \left\langle \delta p, -\check{p} + \boldsymbol{\theta}^T \mathbf{t} + \lambda \right\rangle \\ &= 0, \end{aligned}$$

where  $\boldsymbol{\theta} \in \mathbb{R}^d$  and  $\lambda \in \mathbb{R}$ , and we know  $p$  has the form of

$$(27) \quad \check{p} = \boldsymbol{\theta}^T \mathbf{t}(\mathbf{x}) + \lambda.$$

Depending on  $\boldsymbol{\theta}$ ,  $\lambda$  is chosen so as to satisfy the normalized condition of  $p \in \mathcal{P}$ , and we obtain the  $u$ -model by rewriting  $\lambda = -b(\boldsymbol{\theta})$ .  $\square$

While this normalized  $u$ -model is suitable for conventional statistical inference, we can additionally consider different conditions for the bias function  $b(\boldsymbol{\theta})$  of the  $u$ -model

$$(28) \quad \mathcal{U} = \left\{ q(\mathbf{x}) = u(\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\theta})), \boldsymbol{\theta} \in \mathbb{R}^d \right\}$$

based on consistent subspaces  $\mathcal{Q}$  and  $\mathcal{R}$ .

For  $q \in \mathcal{U}$ , the derivative of  $b$  has to satisfy following conditions depending on constraints, respectively.

**normalized  $u$ -model:**  $\mathcal{U} \cap \mathcal{P}$

$$(29) \quad \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\langle u'(\check{q}), \mathbf{t} \rangle}{\langle u'(\check{q}) \rangle},$$

where  $u'$  is the derivative of  $u$ .

**moment matching  $u$ -model:**  $\mathcal{U} \cap \mathcal{Q}(p_0)$

$$(30) \quad \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \langle p_0, \mathbf{t} \rangle = E_{p_0}[\mathbf{t}] = \text{const.}$$

**constant volume  $u$ -model:**  $\mathcal{U} \cap \mathcal{R}$

$$(31) \quad \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\langle q, \mathbf{t} \rangle}{\langle q \rangle}.$$

Note that each restricted model is not simply flat, however we can prove modified Pythagorean relations as follows.

**Corollary 1** (normalized model). Let  $p$  be in  $\mathcal{P}$ , and  $q, r$  be in  $\mathcal{U} \cap \mathcal{P}$ . If  $q$  is the  $U$ -estimate of  $p$ , that is,

$$(32) \quad q = \arg \min_{q' \in \mathcal{U} \cap \mathcal{P}} D_U(p, q'),$$

then

$$(33) \quad D_U(p, r) = D_U(p, q) + D_U(q, r)$$

holds.

*Proof.* Since  $D_U$  is minimized at  $q$ , we see

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} D_U(p, q) &= \frac{\partial}{\partial \boldsymbol{\theta}} \{ \langle U(\check{q}) \rangle - \langle p, \check{q} \rangle \} \\ &= \langle u(\check{q}), \mathbf{t} - \mathbf{b}' \rangle - \langle p, \mathbf{t} - \mathbf{b}' \rangle \\ &= \langle q - p, \mathbf{t} - \mathbf{b}' \rangle \\ &= 0, \end{aligned}$$

where  $\mathbf{b}' = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ . Using the fact that  $\langle p \rangle = \langle q \rangle = 1$ , we have

$$\begin{aligned} \langle q - p, \mathbf{t} - \mathbf{b}' \rangle &= \langle q - p, \mathbf{t} \rangle - \mathbf{b}' \langle q - p \rangle \\ &= \langle q - p, \mathbf{t} \rangle \\ &= 0. \end{aligned}$$

Writing the  $u$ -representations of  $q$  and  $r$  with  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$  as

$$\check{q} = \boldsymbol{\theta}^T \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\theta}), \quad \check{r} = \boldsymbol{\eta}^T \mathbf{t}(\mathbf{x}) - b(\boldsymbol{\eta}),$$

we see that  $p - q$  and  $\check{r} - \check{q}$  are orthogonal as

$$\begin{aligned} & \langle p - q, \check{r} - \check{q} \rangle \\ &= \langle p - q, (\boldsymbol{\eta} - \boldsymbol{\theta})^T \mathbf{t} - (b(\boldsymbol{\eta}) - b(\boldsymbol{\theta})) \rangle \\ &= (\boldsymbol{\eta} - \boldsymbol{\theta})^T \langle p - q, \mathbf{t} \rangle - (b(\boldsymbol{\eta}) - b(\boldsymbol{\theta})) \langle p - q \rangle \\ &= 0. \end{aligned}$$

$\square$

**Corollary 2** (moment matching model). Let  $p$  be in  $\mathcal{P}$ , and  $q, r$  be in  $\mathcal{U} \cap \mathcal{Q}$ . If  $q$  is the  $U$ -estimate of  $p$ , that is,

$$(34) \quad q = \arg \min_{q' \in \mathcal{U} \cap \mathcal{Q}} D_U(p, q'),$$

then

$$(35) \quad D_U(p, r) = D_U(p, q) + D_U(q, r)$$

holds.

*Proof.* Let us define

$$\boldsymbol{\tau} = \langle p_0, \mathbf{t} \rangle = E_{p_0}[\mathbf{t}],$$

then, from the moment matching condition (30),  $b$  is written as

$$b(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\tau} + \text{const.}$$

Knowing that

$$\frac{\partial}{\partial \boldsymbol{\theta}} D_U(p, q) = \langle q - p, \mathbf{t} - \mathbf{b}' \rangle = \langle q - p, \mathbf{t} - \boldsymbol{\tau} \rangle = 0,$$

we conclude

$$\begin{aligned}
& \langle p - q, \check{r} - \check{q} \rangle \\
&= \langle p - q, (\boldsymbol{\eta} - \boldsymbol{\theta})^T \mathbf{t} - (\boldsymbol{\eta} - \boldsymbol{\theta})^T \boldsymbol{\tau} \rangle \\
&= (\boldsymbol{\eta} - \boldsymbol{\theta})^T \langle p - q, \mathbf{t} - \boldsymbol{\tau} \rangle \\
&= 0.
\end{aligned}$$

□

**Corollary 3** (constant volume model). *Let  $p$  be in  $\mathcal{P}$ , and  $q, r$  be in  $\mathcal{U} \cap \mathcal{R}$ , and let us define an auxiliary measure of  $p$  scaled by  $q$  with*

$$\tilde{p} = \langle q \rangle p.$$

If  $q$  is the  $U$ -estimate of  $p$ , that is,

$$(36) \quad q = \arg \min_{q' \in \mathcal{U} \cap \mathcal{R}} D_U(p, q'),$$

then among  $\tilde{p}$ ,  $q$  and  $r$ , the Pythagorean relation

$$(37) \quad D_U(\tilde{p}, r) = D_U(\tilde{p}, q) + D_U(q, r)$$

holds.

*Proof.* According to the constant volume condition (31) and  $\langle p \rangle = 1$ ,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} D_U(p, q) &= \langle q - p, \mathbf{t} - \mathbf{b}' \rangle \\
&= \langle q - p, \mathbf{t} \rangle - \frac{\langle q, \mathbf{t} \rangle}{\langle q \rangle} \langle q - p \rangle \\
&= \frac{\langle q, \mathbf{t} \rangle}{\langle q \rangle} - \langle p, \mathbf{t} \rangle \\
&= 0,
\end{aligned}$$

therefore we obtain

$$\begin{aligned}
\langle \tilde{p} - q, \check{r} - \check{q} \rangle &= \langle \langle q \rangle p - q, (\boldsymbol{\eta} - \boldsymbol{\theta})^T \mathbf{t} - (b(\boldsymbol{\eta}) - b(\boldsymbol{\theta})) \rangle \\
&= (\boldsymbol{\eta} - \boldsymbol{\theta})^T (\langle q \rangle \langle p, \mathbf{t} \rangle - \langle q, \mathbf{t} \rangle) \\
&\quad - (b(\boldsymbol{\eta}) - b(\boldsymbol{\theta})) (\langle q \rangle \langle p \rangle - \langle q \rangle) \\
&= 0.
\end{aligned}$$

□

## 5. CHARACTERIZATION OF MIXTURE

Aside from parametric models, we can consider mixtures of density functions in  $m$ - and  $u$ -representations, which are regarded as flat subspaces spanned by finite density functions. The following two theorems characterize the  $m$ -mixture and  $u$ -mixture of density functions respectively.

**Theorem 3.** *Let  $p_i, i = 1, \dots, n$  be probability density functions and  $w_i, i = 1, \dots, n$  be associated weights which satisfy*

$$w_i \geq 0, \sum_{i=1}^n w_i = 1.$$

We define the  $m$ -mixture of  $p_i$  with

$$(38) \quad p_m(\mathbf{x}) = \sum_{i=1}^n w_i p_i(\mathbf{x}),$$

then the minimizer of the weighted Bregman divergence  $\sum_{i=1}^n w_i D_U(p_i, q)$  in  $\mathcal{P}$  is given by the  $m$ -mixture of  $p_i$ 's as

$$(39) \quad \arg \min_{q \in \mathcal{P}} \sum_{i=1}^n w_i D_U(p_i, q) = p_m.$$

*Proof.* From the definition of the Bregman divergence, we can rewrite the weighted Bregman divergence as

$$\begin{aligned}
& \sum_{i=1}^n w_i D_U(p_i, q) \\
&= \sum_{i=1}^n w_i \{ \langle U(\check{q}) \rangle - \langle U(\check{p}_i) \rangle - \langle p_i, \check{q} - \check{p}_i \rangle \} \\
&= \langle U(\check{q}) \rangle - \langle \sum_{i=1}^n w_i p_i, \check{q} \rangle \\
&\quad - \sum_{i=1}^n w_i \langle U(\check{p}_i) \rangle + \sum_{i=1}^n w_i \langle p_i, \check{p}_i \rangle \\
&= H_U(p_m, q) - \sum_{i=1}^n w_i H_U(p_i).
\end{aligned}$$

Therefore, the optimization objective becomes

$$\begin{aligned}
\arg \min_q \sum_{i=1}^n w_i D_U(p_i, q) &= \arg \min_q H_U(p_m, q) \\
&= \arg \min_q D_U(p_m, q),
\end{aligned}$$

because of

$$\begin{aligned}
\arg \min_q D_U(p, q) &= \arg \min_q H_U(p, q) - H_U(p) \\
&= \arg \min_q H_U(p, q).
\end{aligned}$$

From the consistency of the Bregman divergence in  $\mathcal{P}$ , we conclude

$$\arg \min_{q \in \mathcal{P}} \sum_{i=1}^n w_i D_U(p_i, q) = \arg \min_{q \in \mathcal{P}} D_U(p_m, q) = p_m.$$

□

Since the Bregman divergence  $D_U(p, q)$  is not symmetric with respect to  $p$  and  $q$ , the reversal usage of the Bregman divergence gives a slightly different result.

**Theorem 4.** *Let  $p_i, i = 1, \dots, n$  be probability density functions and  $w_i, i = 1, \dots, n$  be associated weights which satisfy*

$$w_i \geq 0, \sum_{i=1}^n w_i = 1.$$

We define the  $u$ -mixture of  $p_i$  with

$$(40) \quad p_u(\mathbf{x}) = u(\sum_{i=1}^n w_i \check{p}_i(\mathbf{x}) - b),$$

where  $b$  is a normalization constant for imposing  $p_u \in \mathcal{P}$ , then the minimizer of the weighted Bregman divergence  $\sum_{i=1}^n w_i D_U(q, p_i)$  in  $\mathcal{P}$  is given by the  $u$ -mixture of  $q_i$ 's as

$$(41) \quad \arg \min_{q \in \mathcal{P}} \sum_{i=1}^n w_i D_U(q, p_i) = p_u.$$

*Proof.* From the definition of the Bregman divergence, we see that

$$\begin{aligned} & \sum_{i=1}^n w_i D_U(q, p_i) \\ &= \sum_{i=1}^n w_i \{ \langle U(\check{p}_i) \rangle - \langle U(\check{q}) \rangle - \langle q, \check{p}_i - \check{q} \rangle \} \\ &= \sum_{i=1}^n w_i \langle U(\check{p}_i) \rangle - \langle U(\check{q}) \rangle - \langle q, \sum_{i=1}^n w_i \check{p}_i - \check{q} \rangle \\ &= \langle U(\check{p}_u) \rangle - \langle U(\check{q}) \rangle - \langle q, \check{p}_u - \check{q} \rangle \\ &\quad - \langle U(\check{p}_u) \rangle + \sum_{i=1}^n w_i \langle U(\check{p}_i) \rangle - b \langle q \rangle \\ &= D_U(q, p_u) - b \langle q \rangle + \underbrace{\langle U(\check{p}_u) \rangle + \sum_{i=1}^n w_i \langle U(\check{p}_i) \rangle}_{\text{not depend on } q}. \end{aligned}$$

Using the fact that  $\langle q \rangle = 1$  because  $q \in \mathcal{P}$  and the consistency of the Bregman divergence, we conclude

$$\arg \min_{q \in \mathcal{P}} \sum_{i=1}^n w_i D_U(q, p_i) = \arg \min_{q \in \mathcal{P}} D_U(q, p_u) = p_u.$$

These mixtures are employed in various methods of machine learning and statistics, and provide important applications. For example, the  $m$ -mixture is used in the bagging algorithm [3], and the  $u$ -mixture is implicitly used in the boosting algorithm [5, 10].

## 6. DENSITY INTEGRATION

Now we are ready for stating our main theorem for integrating the probabilities estimated from different data sets.

**Theorem 5.** Let  $p_i, i = 1, \dots, n$  be probability density functions and  $w_i, i = 1, \dots, n$  be associated weights which satisfy

$$w_i \geq 0, \sum_{i=1}^n w_i = 1.$$

Let  $q_i, i = 1, \dots, n$  be the  $U$ -estimates of  $p_i$  in the normalized  $u$ -model  $\mathcal{U} \cap \mathcal{P}$ ,

$$(42) \quad q_i = \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(p_i, q).$$

For the  $m$ -mixture of  $p_i$ 's, the  $U$ -estimate in  $\mathcal{U} \cap \mathcal{P}$  is given by the  $U$ -estimate of the  $m$ -mixture of  $q_i$ 's as

$$(43) \quad \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i p_i, q) = \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i q_i, q).$$

*Proof.* Following the same discussion with the  $m$ -mixture and using the Pythagorean relation, we see that

$$\begin{aligned} & \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} H_U(\sum_{i=1}^n w_i p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \langle U(\check{q}) \rangle - \langle \sum_{i=1}^n w_i p_i, \check{q} \rangle \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w_i \{ \langle U(\check{q}) \rangle - \langle p_i, \check{q} \rangle \} \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w_i H_U(p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w_i D_U(p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w_i \{ D_U(p_i, q_i) + D_U(q_i, q) \} \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w_i D_U(q_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i q_i, q). \end{aligned}$$

□

We should note that although  $q_i$ 's are in the  $u$ -model, the optimal  $q$  with respect to  $D_U(\sum_{i=1}^n w_i q_i, q)$  is not always found in the  $u$ -mixture of  $q_i$ 's, that is,

$$q(\mathbf{x}) = u(\sum_{i=1}^n w'_i \check{q}_i(\mathbf{x}) - b).$$

This is easily understood, for example, by considering mixtures of two distributions under the Kullback-Leibler divergence where  $U(z) = \exp(z)$ . Suppose  $q_1$  and  $q_2$  are densities of normal distributions with the same variance and different means, then  $u$ -mixtures (exponential mixtures) of  $q_1$  and  $q_2$  have the same variance, while  $m$ -mixtures of  $q_1$  and  $q_2$  have different variances. This means that the projection of  $m$ -mixtures onto normal distributions, which is the  $u$ -model in this case, can not be included in the  $u$ -mixtures. Even though  $q$  is restricted in the  $u$ -mixture of  $q_i$ 's, the relationship between weights of the  $m$ - and  $u$ -mixtures,  $w_i$  and  $w'_i$ , is not explicitly written in general.

For  $q$  in the moment matching  $u$ -model  $\mathcal{U} \cap \mathcal{Q}$ , we can make the exactly same argument with  $q$  in the normalized  $u$ -model, that is,

$$(44) \quad \arg \min_{q \in \mathcal{U} \cap \mathcal{Q}} D_U(\sum_{i=1}^n w_i p_i, q) = \arg \min_{q \in \mathcal{U} \cap \mathcal{Q}} D_U(\sum_{i=1}^n w_i q_i, q).$$

A similar statement also holds for the constant volume  $u$ -model  $\mathcal{U} \cap \mathcal{R}$  with a slight modification of the mixture weights as follows.

**Theorem 6.** The  $U$ -estimate of the  $m$ -mixture of  $p_i$ 's in  $\mathcal{U} \cap \mathcal{R}$  is given by

$$(45) \quad \arg \min_{q \in \mathcal{U} \cap \mathcal{R}} D_U(\sum_{i=1}^n w_i p_i, q) = \arg \min_{q \in \mathcal{U} \cap \mathcal{R}} D_U(\sum_{i=1}^n w_i q_i / \langle q_i \rangle, q).$$

*Proof.* Let us define

$$w'_i = \frac{w_i}{\langle q_i \rangle} \quad \text{and} \quad \tilde{p}_i = \langle q_i \rangle p_i.$$

Because of the constant volume condition,  $\langle U(\tilde{q}) \rangle$  can be neglected in minimization, therefore

$$\begin{aligned} & \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} H_U(\sum_{i=1}^n w_i p_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \langle \sum_{i=1}^n w_i p_i, \tilde{q} \rangle \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \langle \sum_{i=1}^n w'_i \tilde{p}_i, \tilde{q} \rangle \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w'_i H_U(\tilde{p}_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w'_i D_U(\tilde{p}_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w'_i \{D_U(\tilde{p}_i, q_i) + D_U(q_i, q)\} \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} \sum_{i=1}^n w'_i D_U(q_i, q) \\ &= \arg \min_{q \in \mathcal{U} \cap \mathcal{P}} D_U(\sum_{i=1}^n w_i q_i / \langle q_i \rangle, q). \end{aligned}$$

□

The following two examples show applications of the theorems.

**Example 2.** Let  $\mathcal{D}_i$ ,  $i = 1, \dots, n$  be independent data sets and  $q_i$ ,  $i = 1, \dots, n$  be the  $U$ -estimates in  $\mathcal{U} \cap \mathcal{P}$  based on  $\mathcal{D}_i$ . That is, we construct an empirical distribution from a data set  $\mathcal{D}_i = \{\mathbf{x}_j^{(i)}, j = 1, \dots, |\mathcal{D}_i|\}$ , where  $|\mathcal{D}_i|$  denotes the cardinality of  $\mathcal{D}_i$ , as

$$(46) \quad p_i(\mathbf{x}) = \frac{1}{|\mathcal{D}_i|} \sum_j \delta(\mathbf{x} - \mathbf{x}_j^{(i)}),$$

and obtain the  $U$ -estimates by solving

$$(47) \quad q_i = \arg \min_{r \in \mathcal{U} \cap \mathcal{P}} D_U(p_i, r) = \arg \min_{r \in \mathcal{U} \cap \mathcal{P}} H_U(p_i, r).$$

Let  $\mathcal{D}$  be a combined data set of  $\mathcal{D}_i$ ,  $i = 1, \dots, n$

$$(48) \quad \mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i.$$

Then its empirical distribution is written in the form of the  $m$ -mixture of  $p_i$ 's as

$$(49) \quad p_m(\mathbf{x}) = \frac{\sum_{i=1}^n |\mathcal{D}_i| p_i(\mathbf{x})}{\sum_{i=1}^n |\mathcal{D}_i|}.$$

Also we define the  $m$ -mixture of  $q_i$ 's as

$$(50) \quad q_m(\mathbf{x}) = \frac{\sum_{i=1}^n |\mathcal{D}_i| q_i(\mathbf{x})}{\sum_{i=1}^n |\mathcal{D}_i|}.$$

From the theorem, the  $U$ -estimate of  $p_m$  for the combined data set  $\mathcal{D}$  is given by

$$(51) \quad q = \arg \min_{r \in \mathcal{U} \cap \mathcal{P}} D_U(q_m, r),$$

instead of solving

$$(52) \quad q = \arg \min_{r \in \mathcal{U} \cap \mathcal{P}} D_U(p_m, r).$$

This means that to estimate an integrated model from a combined data set, we simply keep estimated  $q_i$ 's as representatives of  $\mathcal{D}_i$ 's and do not have to keep all the data sets.

**Example 3.** Let  $\tilde{\mathcal{D}}_i$ ,  $i = 1, \dots, n$  be data sets with missing values. For example,  $\mathbf{x} = (x_1, x_2, x_3)$  is the 3-dimensional variable and  $\tilde{\mathcal{D}}_1$  consists of  $\{(x_1, x_2)_j, j = 1, \dots, |\tilde{\mathcal{D}}_1|\}$  where  $x_3$ 's are missing,  $\tilde{\mathcal{D}}_2$  consists of  $\{(x_2, x_3)_k, k = 1, \dots, |\tilde{\mathcal{D}}_2|\}$  where  $x_1$ 's are missing, and so on. Let  $q_i$ ,  $i = 1, \dots, n$  be the  $U$ -estimates from incomplete data sets  $\tilde{\mathcal{D}}_i$ . When data include missing values, we can apply, for example, an extended EM algorithm [6] to obtain an estimate for each data set.

To integrate  $q_i$ 's, we need to calibrate weights, because each element of the parameter vector  $\boldsymbol{\theta}$  is not equally influenced by missing elements of  $\mathbf{x}$  and the number of partially observed data is not a good weight in this case. Here we consider the effective number of incomplete data with a similar argument found in the AIC (Akaike's Information Criterion) literature [1, 10].

Let  $q$  be the  $U$ -estimate in  $\mathcal{U} \cap \mathcal{P}$  based on data  $\mathcal{D}$  generated from a distribution  $p$ , then the expected difference of the following divergences is asymptotically given by

$$(53) \quad E[D_U(p, q) - D_U(p, p_*)] = \frac{\text{tr} GQ^{-1}}{2|\mathcal{D}|},$$

where  $E$  is the expectation with respect to data  $\mathcal{D}$ ,  $p_*$  is the optimal  $U$ -estimate of  $p$  in  $\mathcal{U} \cap \mathcal{P}$ , and the  $ij$ -elements of matrices  $G$  and  $Q$  are defined by

$$(54) \quad G_{ij} = \left\langle p, \frac{\partial}{\partial \theta_i} \tilde{q} \frac{\partial}{\partial \theta_j} \tilde{q} \right\rangle = E_p \left[ \frac{\partial}{\partial \theta_i} \tilde{q} \frac{\partial}{\partial \theta_j} \tilde{q} \right],$$

$$(55) \quad Q_{ij} = \left\langle p, \frac{\partial^2}{\partial \theta_i \partial \theta_j} \tilde{q} \right\rangle = E_p \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \tilde{q} \right].$$

Note that estimates of  $G$  and  $Q$  are calculated by replacing the true distribution  $p$  with the empirical distribution in practice.

Let  $\mathbf{y}$  be the visible or observable part of  $\mathbf{x}$ , and  $\mathbf{z}$  be the hidden part of  $\mathbf{x}$ . The marginal distribution of  $\mathbf{y}$  is obtained by integrating  $q(\mathbf{x}) = q(\mathbf{y}, \mathbf{z})$  with respect to  $\mathbf{z}$

$$(56) \quad \tilde{q}(\mathbf{y}) = \int q(\mathbf{y}, \mathbf{z}) d\mathbf{z}.$$

Let  $\tilde{G}$  and  $\tilde{Q}$  be the above defined two matrices calculated from the marginalized density function  $\tilde{q}$ . Then Eq. (53) for the estimate from incomplete data set  $\tilde{\mathcal{D}}$  is given by

$$(57) \quad E[D_U(p, q) - D_U(p, p_*)] = \frac{\text{tr} \tilde{G} \tilde{Q}^{-1}}{2|\tilde{\mathcal{D}}|}.$$

Suppose we have a data set  $\mathcal{D}$  without missing values, and a data set  $\tilde{\mathcal{D}}$  with missing values. If two estimates from

$\mathcal{D}$  and  $\tilde{\mathcal{D}}$  satisfy

$$(58) \quad \frac{\text{tr} GQ^{-1}}{|\mathcal{D}|} = \frac{\text{tr} \tilde{G}\tilde{Q}^{-1}}{|\tilde{\mathcal{D}}|},$$

we can expect those two estimates have the same effectiveness in estimation. Therefore,

$$(59) \quad |\mathcal{D}| = \frac{\text{tr} GQ^{-1}}{\text{tr} \tilde{G}\tilde{Q}^{-1}} |\tilde{\mathcal{D}}|$$

can be regarded as the effective number of data  $\tilde{\mathcal{D}}$ , which corresponds to the number of the complete data giving the same accuracy with the estimate from the incomplete data. Using the effective number of data  $|\mathcal{D}|$  instead of the original number of data  $|\tilde{\mathcal{D}}|$ , we define the  $m$ -mixture of  $q_i$ 's as

$$(60) \quad q_m(\mathbf{x}) = \frac{\sum_{i=1}^n |\mathcal{D}_i| q_i(\mathbf{x})}{\sum_{i=1}^n |\mathcal{D}_i|},$$

a probability integration considering information loss of partial observation is given by

$$(61) \quad q = \arg \min_{r \in \mathcal{U} \cap \mathcal{P}} D_U(q_m, r).$$

In this example, an effective number of incomplete data is discussed along the lines of AIC, however, there are many possibilities for estimating efficiency of data. Basically, efficiency is determined by accuracy of estimates, therefore any confidence assessments of estimates can be applied, such as cross-validation and bootstrap methods. Moreover, those methods are applicable to evaluate accuracy of moment matching model and the constant volume model, and hence enable us to integrate those models also.

## 7. CONCLUSION

In this paper, we have investigated properties of consistent subspaces and well-behaved parametric models associated with the Bregman divergence, and derived a natural method for integration of multiple density functions. While we have mainly focused on joint density functions, theories can be easily extended to conditional density functions in straight-forward manner, and can be applied to problems of aggregating multiple estimates in regression and classification. Also, in this paper, only simple parametric models are considered, but it would be interesting to extend to more complicated models and situations such as non-parametric models, estimates from multiple divergences, and dependent data sets.

## REFERENCES

- [1] H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974), 716–723.
- [2] S. Amari, *Integration of stochastic models by minimizing  $\alpha$ -divergence*, Neural Computation **19** (2007), 2780–2796.

- [3] L. Breiman, *Arcing classifiers*, Machine Learning **26** (1996), 123–140.
- [4] S. Eguchi, *Information geometry and statistical pattern recognition*, Sugaku Expositions **19** (2006), 197–216.
- [5] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences **55** (1997), 119–139.
- [6] Y. Fujimoto and N. Murata, *A modified EM algorithm for mixture models based on Bregman divergence*, Annals of the Institute of Statistical Mathematics **59** (2007), 57–75.
- [7] C. Genest and J. V. Zidek, *Combining probability distributions: A critique and an annotated bibliography*, Statistical Science **1** (1986), no. 1, 114–148.
- [8] P. D. Grünwald and A. P. Dawid, *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory*, The Annals of Statistics **32** (2004), no. 4, 1367–1433.
- [9] R. A. Jacobs, *Methods for combining experts' probability assessments*, Neural Computation **7** (1995), 867–888.
- [10] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, *Information geometry of U-boost and Bregman divergence*, Neural Computation **16** (2004), 1437–1481.
- [11] I. J. Myung, S. Ramamoorti, and A. D. Bailey Jr., *Maximum entropy aggregation of expert predictions*, Management Science **42** (1996), no. 10, 1420–1436.
- [12] T. Takenouchi and S. Eguchi, *Robustifying AdaBoost by adding the naive error rate*, Neural Computation **16** (2004), no. 4, 767–787.
- [13] T. Takenouchi, S. Eguchi, N. Murata, and T. Kanamori, *Robust boosting algorithm against mislabeling in multiclass problems*, Neural Computation **20** (2008), 1596–1630.

Noboru Murata

Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan.

E-mail: noboru.murata(at)eb.waseda.ac.jp

Yu Fujimoto

Aoyama Gakuin University, 5-10-1 Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan.

E-mail: yu.fujimoto(at)it.aoyama.ac.jp