Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Kamil Dedecius

# Information Fusion with Functional Bregman Divergence

No. 2345 January 6, 2015

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: +420 266052570, Fax: +420 266052068, Url: http://www.utia.cas.cz,
E-mail: dedecius@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the institute.

# 1 Introduction

The Bregman divergences [1], originally developed in the context of the convex optimization, attained in the recent decade a significant focus of information theorists. Their use in the statistical information processing is a result of the rapid development of information geometry, in a huge extent pioneered by Amari's work [2]. The research focus has been given mainly to the clustering problem, namely the $k$-means algorithm, where the divergences were studied as the measures of the intraclusters distances between the centres of the clusters' masses – the centroids – and the points within these clusters, e.g. [3],[4]. Other application-oriented works exploit the Bregman divergences for statistical modelling with logistic regression method [5, 6], construction of Voronoi diagrams [7] etc.

It was not until 2008 when the functional version of the Bregman divergences was proposed by Frigyik et al. [8]. This approach is adopted in this paper for its immense generality allowing, unlike the previous versions, direct application of the relevant methods to a much wider range of problems, from the real vector spaces to the spaces of functions. In our scope, the application aims at the fusion of probability distributions in distributed systems. Similarly to the clustering problems, the divergence serves as the minimization criterion quantifying the dissimilarity of the consensus distribution and the group of distributions to be fused.

In order to alleviate potentially difficult understanding of the abstract mathematical machinery, explanatory "real-world" examples are provided along the way of theory exposition.

# 2 Bregman divergence

In this section, we introduce the Bregman divergence in a very general sense, over a vector space equipped with a norm, that is, a Banach space. As an example at the "highest" level, we name the $L^p$ spaces with $p \geq 1$ of probability distributions, at the "lowest" the space of parameters $\Theta$ of these functions. We conjecture, that the results generalize for arbitrary Fréchet space equipped with a seminorm in the ordinary sense.

The definition of the Bregman divergence relies on derivatives, which we need to define in an appropriate general form. The Fréchet derivative is suitable.

**Definition 1** (Fréchet derivative)**.** *Let $X$ and $Y$ be two Banach spaces and $U \subset X$ open subset of $X$. A function $f : U \to Y$ is Fréchet differentiable at $x \in U$ if there exists a bounded linear operator $Df(x) : X \to Y$ and a continuous function $\psi : B(0, \varepsilon) \to Y$ where $B(0, \varepsilon)$ is an open neighbourhood of $0 \in X$ with $\psi(0) = 0$, such that*

$$f(x + h) - f(x) = (Df(x))h + ||h||\, \psi(h)$$

*for all $h \in V$.*

We say that $f \in C^p$, $p = 1, 2, \ldots$ if the $p$th Fréchet derivative $D^p f$ exists and it is a continuous function.

The Fréchet derivative keeps most of the important properties of the ordinary function derivatives, for instance the uniqueness, linearity and the chain rule. Its main advantage is its straightforward applicability to more complicated structures than functions, for instance to functionals, where it allows differentiation of functionals $\phi$ with respect to a function $f$, i.e. $D(\phi(f))$. Indeed, it still remains suitable for differentiable functions over real fields as shows the following example.

**Example 1.** *In function spaces, the Fréchet derivative coincides with the ordinary derivative of functions $f : \mathbb{R} \to \mathbb{R}$. Assume $t > 0$, then*

$$f(x + th) = f(x) + (Df(x))(th) + ||th||\,\psi(th)$$

*that is*

$$(Df(x))h = \frac{f(x + th) - f(x) - ||th||\,\psi(th)}{t}.$$

*The result follows by taking limit $t \to 0$. This is also why $\psi$ needs to be continuous.*

**Definition 2** (Bregman divergence)**.** *Let $X$ be a Banach space and $\phi : X \to \mathbb{R}$ a strictly convex $C^2$ function. The Bregman divergence $d_\phi : X \times X \to \mathbb{R}_+$ for all admissible $x, y \in X$ is the mapping*

$$d_\phi(x, y) = \phi(x) - \phi(y) - (D\phi(y))\,(x - y). \tag{1}$$

## 2.1 Bregman Divergence Properties

The Bregman divergence has some well known properties. Some of them (for proofs see [8]) are:

**Nonnegativity** $d_\phi(x, y) \geq 0$ with 0 iff $x = y$ almost everywhere.

**Convexity in the first argument** $d_\phi(x, y)$ is convex in $x$. Generally, this does not hold for $y$.

**Linearity** both additivity $d_{\phi_1+\phi_2}(x, y) = d_{\phi_1}(x, y) + d_{\phi_2}(x, y)$ and homogeneity $d_{c\phi}(x, y) = cd_\phi(x, y)$ for $c \in \mathbb{R}$ hold.

**Generalized triangle inequality** for admissible $x, y, z \in X$ it holds $d_\phi(x, y) + d_\phi(y, z) + (Df(x))(x - y) - (Df(z))(x - y)$.

**Example 2** (Quadratic Loss)**.** *Assume a Banach space $X$ that is an inner product space where $\phi = \langle x, x\rangle$ is the inner product $X \times X \mapsto \mathbb{R}$ in an admissible sense. Useful examples of such spaces are the Hilbert spaces, the Euclidean space $\mathbb{R}^n$ or the $L^2$ space of Lebesgue measurable functions. Suppose furthermore $x, y \in X$. Then the first derivative of $\phi$ reads*

$$\begin{aligned}\phi(x + h) - \phi(x) &= (D\phi(x))\,h + ||h||\,\psi(h)\\ &= \langle x + h, x + h\rangle - \langle x, x\rangle\\ &= 2\langle x, h\rangle + \langle h, h\rangle.\end{aligned}$$

*Hence $Df(x) = 2\langle x, h\rangle$ since $\langle h, h\rangle = ||h||^2 \to 0$ as $h \to 0$ in $X$. To prove strict convexity of $\phi$, we need the second derivative to be strictly positive. That is,*

$$Df(x + h) - Df(x) = D\,(Df(x))\,h + ||h||\,\psi(h)$$

*and after rearrangement of terms and applying limit we obtain*

$$D^2 f(x) = 2||h||^2 \tag{2}$$

*which is clearly greater than 0 by assumptions. Finally,*

$$\begin{aligned} d_\phi(x,y) &= \langle x,x\rangle - \langle y,y\rangle - 2\langle y, x-y\rangle \\ &= \langle x-y, x-y\rangle \\ &= ||x-y||^2. \end{aligned}$$

*We emphasize, that this result is applicable in any inner product space, be it an $L^2$ space of admissible continuous probability density functions or the Euclidean space of real (scalar or multivariate) parameters of these functions. The task of computing Fréchet derivatives may seem tedious, but we remind their coincidence with the ordinary notion of function derivatives in simpler spaces, recall Example 1.*

## 2.2 Bregman's Duality

**Definition 3** (Legendre-Fenchel transformation of convex functions)**.** *Let $X$ be an inner product Banach space, $f : X \to \mathbb{R} \cup \{+\infty\}$ a proper convex function. The Legendre-Fenchel transformation is the linear mapping*

$$f^*(x^*) = \sup_{x\in X} \left(\langle x^*, x\rangle - f(x)\right), \qquad x^* \in X^* \tag{3}$$

*where $X$ and $X^*$ are mutually dual spaces and (convex) $f^*$ with range space in $\mathbb{R}\cup\{+\infty\}$ is called the convex conjugate of $f$.*

From the connection between maxima and derivatives it follows that for $x$ satisfying (3), we have

$$D_x\langle x^*, x\rangle = (D_x f)(x)$$

hence

$$x^* = (Df)(x) \quad \text{and} \quad x = (Df)^{-1}(x^*). \tag{4}$$

Applying the same arguments to find $f^{**}$ shows that the convex conjugation is involutive ($f^{**} = f$), obeying the trivial ordinary differential equations $Df^* = (Df)^{-1}$ and $Df = (Df^*)^{-1}$ and using (3) it yields a straightforward solution of many tasks via[1]

$$f^*(x^*) = \left\langle x^*, (Df^{-1})(x^*)\right\rangle - f \circ (Df)^{-1}(x^*). \tag{5}$$

**Example 3** (Legendre-Fenchel transformation)**.** *For an illustration, consider $X = \mathbb{R}$ and $f(x) = x^2$. Then $\langle x^*, x\rangle = x^*x$ and $x^*$ has the meaning of the slope of the line passing through origin. Furthermore, $(Df)(x) = 2x$ and $(Df)^{-1}(x) = \frac{x}{2}$, from which follows $f^*(x^*) = \frac{(x^*)^2}{4}$, either by solving the differential equation or using relation (5). The duality follows from $(Df^*)(x^*) = \frac{x}{2}, (Df^*)^{-1}(x^*) = 2x$ and $f^{**} = f(x) = x^2$, which was the original function.*

**Example 4** (Self-duality of 2-norm)**.** *If $f = \frac{1}{2}||\cdot||^2$, then $f = f^* = f^{**}$ and $f$ is called self-dual.*

[1] Strictly speaking, the necessary and sufficient condition for the biconjugation $f = f^{**}$ is that $f$ is also proper and lower semi-continuous, which directly follows from the definition.

We leave without proof (which follows directly from Definition 3) the fundamental Fenchel-Young inequality, stating

$$f(x) + f^*(x^*) - \langle x^*, x\rangle \geq 0, \tag{6}$$

with equality if $x^* = (Df)(x)$, which is the case (5) above. By the following lemma, this equality yields the dual forms of the Bregman divergence.

**Lemma 1.** *The Bregman divergence* (1) *in an inner product space has the dual forms*

$$\begin{aligned} d_\phi(x, y) &= \phi(x) + \phi^*(y^*) - \langle x, y^*\rangle \\ &= \phi^*(x^*) + \phi(y) - \langle x^*, y\rangle \\ &= d_{\phi^*}(y^*, x^*). \end{aligned} \tag{7}$$

The proof follows simply by plugging the dual expressions (4) into Definition 2. Observe again, how the dual expressions coincide with the Fenchel-Young inequality (6) and together explain the nonnegativity of the Bregman divergence.

### 2.3 Connection with Other Divergences and the Kullback-Leibler divergence

Naturally, there are couple of other families of divergences, for instance $\alpha$-divergences by Chernoff [9] and the Tsallis $\alpha$-divergences [10, 11], $f$-divergences of Csiszár [12], Rényi $\alpha$-divergence [13], $\beta$-divergences by Basu et al. [14] popular in PCA and ICA (Principal/Independent Component Analysis), Fujisawa and Eguchi's $\gamma$-divergence recently proposed for estimation under heavy contamination by outliers [15] and others. A good message is that in many cases the divergences coincide. For instance, Amari proved in [16] that the Bregman divergence coincides with $f$-divergence and its special case $\alpha$-divergence in the important case of the Kullback-Leibler divergence, making it both information monotone and information-geometrically flat. The $\beta$-divergence is a particular case of the Bregman divergence with a specific convex function [16, 17]. Two recent comprehensive overviews of divergences are papers [17] and [18].

Let us now focus for a moment on the particular celebrated case of the Bregman divergence: the Kullback-Leibler divergence [19].[2] It is defined as follows.

**Definition 4.** *Given two pdfs $f$ and $g$ defined on a common space $X$ (e.g. the real line) and such that $f$ is absolutely continuous with respect to $g$. Their Kullback-Leibler divergence is defined by*

$$\mathcal{D}(f||g) = \mathbb{E}_f\left[\log\frac{f(x;\theta_f)}{g(x;\theta_g)}\right] = \int_X f(x;\theta_f)\log\frac{f(x;\theta_f)}{g(x;\theta_g)}dx.$$

As a Bregman divergence, the Kullback-Leibler divergence is a premetric, i.e., it is nonnegative, $D(f||g) = 0$ if and only if $f = g$ almost everywhere. Unless this equality, the symmetry property of a usual metric does not apply, $\mathcal{D}(f||g) \neq \mathcal{D}(g||f)$. Neither does the triangle inequality. Note, that the absolute continuity of $f$ with respect to $g$ is critical as it preserves (by limit) the definition for the cases $g(x) = 0$.

[2] The Kullback-Leibler divergence is generated with $F(x) = x\log(x) - x$. However, in general (even continuous) spaces, certain topology restrictions apply. This is beyond the scope of the paper; the particular case of exponential family is covered in later sections.

Natural question is why the particularly the Kullback-Leibler divergence is often the proper measure when there exist so many other classes of divergences. For instance, the Hellinger distance, which is a member of the so-called Rényi divergences family just like the Kullback-Leibler divergence, is a proper metric. In many applications, the reason lies in the adopted Bayesian paradigm, relying on conditional distributions. The conditionality is conveniently covered just by the Kullback-Leibler divergence. Clearly,

$$\begin{aligned}\mathcal{D}\left(f(x,y)||g(x,y)\right) &= \mathcal{D}\left(f(x|y)f(y)||g(x|y)g(y)\right) \\ &= \mathbb{E}_{f(x,y)}\left[\log\frac{f(x|y)}{g(x|y)} + \log\frac{f(y)}{g(y)}\right] \\ &= \mathcal{D}\left(f(y)||g(y)\right) + \mathbb{E}_{f(y)}\left[\mathcal{D}(f(x|y)||g(x|y)\right].\end{aligned}$$

Assuming the models $f(x|y)$ and $g(x|y)$ identical, the divergence is driven by the information carried by the prior distribution. The expectation of the Kullback-Leibler divergence in the right-hand side is sometimes referred to as the conditional divergence and the relation as the chain rule for the Kullback-Leibler divergence.

An important property of the Kullback-Leibler divergence, immediately arising from careful investigation the defining expectation, is the difference of its behavior due to the asymmetry. Assume, that $f$ is a pdf of a fixed distribution and we search the pdf $g$ of a distribution minimizing the divergence. Two very different cases emerge:

1. Using $\mathcal{D}(f||g) = \mathbb{E}_f[\log\frac{f}{g}]$, the approximating pdf $g(x)$ can avoid approaching to zero when $f(x)$ is close to it, while still retaining minimum contribution to the divergence.

2. With $\mathcal{D}(f||g) = \mathbb{E}_g[\log\frac{g}{f}]$, whenever $f(x)$ approaches zero, the pdf $g(x)$ must do so as well, otherwise the argument of the logarithm will rapidly grow. That is, this order of parameters makes the divergence zero-forcing.

Both these cases have found numerous applications in the estimation theory and machine learning. For instance, the former gives rise to expectation propagation [20], while the second is the cornerstone of variational Bayesian methods [21] including variational message passing [22]. In the scope of approximate inference and machine learning, both approaches are thoroughly treated in [21].

As it will be demonstrated later, the Kullback-Leibler divergence is closely tight with the important exponential family distributions.

# 3 Information Fusion with Bregman Divergence

Let $\{f_i\}_{i=1,\ldots,n}$ be a set of probability density functions assigned with weights (probabilities) $\omega_i \in [0,1]$ summing to unity. The goal is to find the probability density $g$ minimizing the weighted convex combination

$$\sum_{i=1}^{n} \omega_i d_\phi\left(f_i, g\right) \tag{8}$$

The result summarizes the following proposition.

**Proposition 1.** *The minimizer $g$ of*

$$\arg\min_g \sum_{i=1}^{n} \omega_i d_\phi\left(f_i, g\right) \tag{9}$$

*is the convex combination* $\sum_{i=1}^{n} \omega_i f_i$.

*Proof.* From the definition of the Bregman divergence follows

$$\begin{aligned}\sum_{i=1}^{n} \omega_i d_\phi\left(f_i, g\right) &= \sum_{i=1}^{n} \omega_i \left(\phi(f_i) - \phi(g) - D\phi(g)(f_i - g)\right) \\ &= \sum_{i=1}^{n} \omega_i \phi(f_i) - \phi(\bar{f}) \\ &\quad + \underbrace{\phi(\bar{f}) - \phi(g) - D\phi(g)\left(\sum_{i=1}^{n} \omega_i f_i - g\right)}_{=\Upsilon(\bar{f}, g, \sum \omega_i f_i)}. \end{aligned} \tag{10}$$

Inspection of $\Upsilon(\cdot)$ reveals the Bregman divergence if

$$\bar{f} = \sum_{i=1}^{n} \omega_i f_i,$$

hence $\Upsilon(\bar{f}, g, \sum \omega_i f_i) = d_\phi(\bar{f}, g)$. Since the first two terms in (10) are independent of $g$, the minimum is achieved by setting $g = \bar{f}$. □

We will refer to the given case as *Type-1 fusion.* Note, that the result is independent of $\phi$. The task may naturally be formulated in the other way (*Type-2 fusion*), requiring minimization of the weighted combination of the form

$$\sum_{i=1}^{n} \omega_i d_\phi\left(g, f_i\right) \tag{11}$$

The result summarizes the following proposition.

**Proposition 2.** *The minimizer* $g$ *of*

$$\arg\min_{g} \sum_{i=1}^{n} \omega_i d_\phi\left(g, f_i\right) \tag{12}$$

*has the form*

$$g = (D\phi)^{-1}\left(\sum_{i=1}^{n} \omega_i f_i^*\right). \tag{13}$$

*Proof.* From the dual forms of the Bregman divergence (Lemma 7) it follows

$$\sum_{i=1}^{n} \omega_i d_\phi(g, f_i) = \sum_{i=1}^{n} \omega_i d_{\phi^*}(f_i^*, g^*).$$

By Proposition 1 we directly have

$$g^* = \bar{f}^* = \sum_{i=1}^{n} \omega_i f_i^*$$

hence by convex conjugation (4)

$$\begin{aligned} g &= (D\phi)^{-1}(g^*) \\ &= (D\phi)^{-1}\left(\sum_{i=1}^{n} \omega_i f_i^*\right) \\ &= (D\phi)^{-1}\left(\sum_{i=1}^{n} \omega_i (D\phi)(f_i)\right). \end{aligned}$$

□

For comparison, let us derive *Type-1* and *Type-2* fusions optimal in the Kullback-Leibler sense as examples. Recall, that they will correspond to the zero-avoiding and zero-forcing forms.

**Example 5** (Kullback-Leibler optimal Type-1 fusion)**.** *Let us find the pdf $g$ closest to the set $\{f_i; i = 1, \ldots, n\}$ weighted by $\omega_i$ and satisfying*

$$\sum_{j=1}^{n} \omega_i \mathcal{D}(f_i || g) \to \min. \tag{14}$$

*By the definition of the Kullback-Leibler divergence,*

$$\begin{aligned} &\sum_{j=1}^{n} \omega_i \int_y f_i(y|\cdot) \log \frac{f_i(y|\cdot)}{g(y|\cdot)} \\ &\qquad = \int_y \sum_{j=1}^{n} \omega_i f_i(y|\cdot) \log \frac{\sum_{j=1}^{n} \omega_i f_i(y|\cdot)}{g(y|\cdot)} + \kappa_t \\ &\qquad = \mathcal{D}\left(\sum_{j=1}^{n} \omega_i f_i \middle|\middle| g\right) + \kappa_t, \end{aligned} \tag{15}$$

*where $\kappa$ denotes the terms independent of $g$. Indeed,*

$$g = \sum_{j=1}^{n} \omega_i f_i, \tag{16}$$

*hence the resulting $g$ is a mixture density.*

**Example 6** (Kullback-Leibler Divergence, Type-2 fusion)**.** *In this case, we seek the pdf $g$ satisfying*

$$\sum_{i=1}^{n} \omega_i \mathcal{D}(g || f_i) \to \min. \tag{17}$$

*We proceed similarly to the previous example,*

$$\begin{aligned} \sum_{j=1}^{n} \omega_i \int_y g(y|\cdot) \log \frac{g(y|\cdot)}{f_i(y|\cdot)} dy &= \int_y g(y|\cdot) \log \frac{g(y|\cdot)}{\prod_{j\in\mathcal{N}_i} f_i(y|\cdot)^{\omega_i}} dy \\ &= \mathcal{D}\left(g \,\middle|\middle|\, \prod_{j\in\mathcal{N}_i} f_i^{\omega_i}\right). \end{aligned}$$

*That is, the Kullback-Leibler optimal pdf $g$ has the form of the weighted geometric mean*

$$g = \prod_{i=1}^{n} f_i^{\omega_i}. \tag{18}$$

## 4 Exponential Family

The Bregman divergence has a close relation to the exponential family of distributions, which in turn are fundamental in Bayesian modelling and its dynamic branch in particular.

**Definition 5.** *A regular exponential family distribution of a random variable $X$ with parameters $\theta$ in open parameter set $\Theta$ is characterized by the pdf of the form*

$$p(x;\theta) = \exp\left(\langle \eta(\theta), T(x)\rangle - A(\eta(\theta)) + B(x)\right), \tag{19}$$

*where $\eta \equiv \eta(\theta)$ is the natural parameter, $T(x)$ is the sufficient statistics, $A(\theta) \in C^{\infty}$ is the log-partition function and $B(x)$ is a link function.*

We assume compatible dimensions of the variables in the definition. The role of the dimension preserving term $T(x)$ is to accumulate all information about $\theta$ contained in data $x$. The log-partition function $A(\eta) = \log \int_X \exp\left(\langle T(x), \eta(\theta)\rangle + B(x)\right) dx$ has the role of the normalizing term, assuring

$$\int_X p(x;\theta) dx = 1.$$

If $\eta(\theta) = \theta$, the exponential family is canonical.

**Lemma 2** (Barndorff-Nielsen [23])**.** *If $A$ is the log-partition function or a regular exponential family with natural parameter space $\eta(\Theta)$, then $A$ is a proper convex function.*

**Corollary 1.** *There exists a convex conjugate function $A^*$ fulfilling Definition 3.*

If $T$ and $\eta$ are identity mappings, the exponential family is called natural.

The first and second moments of the random variable $X$ can be estimated by relations

$$\begin{aligned} \mathbb{E}[T(x)] &= DA(\eta) = \frac{\partial A(\eta)}{\partial \eta} \\ \operatorname{cov}(T(x)) &= D^2 A(\eta) = \frac{\partial^2 A(\eta)}{\partial \eta^2}, \end{aligned} \tag{20}$$

where the derivatives may be in the multivariate sense. Notice, the Legendre-Fenchel duality $\mathbb{E}[T(x)] = \eta^*$.

**Example 7** (Univariate normal distribution)**.** *The normal distribution of the random variable $X$ has two parameters $\theta = [\mu, \sigma^2]$. Its pdf*

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

*can be rewritten into the form* (19) *as*

$$p(x;\mu,\sigma^2) = \exp\left\{\left\langle \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \begin{bmatrix} x \\ x^2 \end{bmatrix} \right\rangle - \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2 \right\}$$

*hence* $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-\frac{\eta_2}{\pi}\right)$. *The first and (the diagonal of) the second moments of* $T(x)$ *via* (20)

$$\begin{aligned} DA(\eta) &= [\hat{x}, \hat{x^2}]^\intercal = \left[-\frac{\eta_1}{2\eta_2}, \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2}\right]^\intercal = [\mu, \mu^2 + \sigma^2]^\intercal \\ D^2A(\eta) &= [\mathrm{var}(x), \mathrm{var}(x^2)]^\intercal = \left[-\frac{1}{2\eta_2}, -\frac{\eta_1^2}{2\eta_2^3} + \frac{1}{\eta_2}\right]^\intercal \\ &= [\sigma^2, \sigma^2(4\mu^2 + 1)]^\intercal. \end{aligned} \tag{21}$$

Note that this links the mean and variance of observations with the mean and variance of the distribution.

The following fundamental lemma associates the Bregman divergence of two exponential family distributions directly with the Kullback-Leibler divergence. Its immediate impact will become clear shortly.

**Lemma 3.** *The Bregman divergence* $d(\eta(\theta_q), \eta(\theta_p))$ *of two distributions* $p(x; \eta(\theta_p) = p(x; \theta_p)$ *and* $q(x; \eta(\theta_q)) = q(x; \theta_q)$ *from the same exponential family generated by* $A$ *coincides with the Kullback-Leibler divergence* $\mathcal{D}(p||q)$.

*Proof.* Denote $\eta_p = \eta(\theta_p)$ and $\eta_q = \eta(\theta_q)$. Then

$$\begin{aligned} d_A(\eta_q, \eta_p) &= A(\eta_q) - A(\eta_p) - \langle \eta_q - \eta_p, D(\eta_p) \rangle \\ &= \log \frac{\exp\left(\langle T(x), \eta_p \rangle - A(\eta_p) - B(x)\right)}{\exp\left(\langle T(x), \eta_q \rangle - A(\eta_q) - B(x)\right)} \\ &+ \langle \eta_p - \eta_q, DA(\eta_p) - T(x) \rangle \\ &= \log \frac{p(x, \theta_p)}{q(x, \theta_q)} + \langle \eta_p - \eta_q, DA(\eta_p) - T(x) \rangle \end{aligned}$$

By (20) and taking expectations of both sides with respect to $p(x; \theta_p)$ yields the result

$$d_A(\eta_q, \eta_p) = \mathbb{E}\left[\log \frac{p(x; \theta_p)}{q(x; \theta_q)}\right] = \mathcal{D}(p||q).$$

□

**Corollary 2.** *Assume the set* $\{f_i\}_{i=1,\dots,n}$ *of probability density functions from the same exponential family assigned with weights* $\omega_i \in [0, 1]$ *summing to unity. Then the distribution* $g$ *of the same family, whose parameter* $\theta_g$ *(identically* $\eta(\theta_g)$*) minimizes the Bregman divergences generated by the log-partition function* $A$ *in the sense of the Type-1 fusion is given by*

$$\eta_g = \sum_{i=1}^{n} \omega_i \eta(\theta_{f_i}). \tag{22}$$

*Identically, this pdf minimizes the criterion*

$$\sum_{i=1}^{n} \omega_i \mathcal{D}(g||f_i)$$

*yielding*

$$g = \prod_{i=1}^{n} f_i^{\omega_i}, \tag{23}$$

*which coincides with the Kullback-Leibler optimal Type-2 fusion.*

Note that in the identical exponential family, the sum of exponents (hence natural parameters) obviously results from the weighted geometric mean. Let us also draw attention to an interesting peculiarity: the Type-1 fusion (Prop. 1) of whole pdfs yields a convex combination, i.e. a mixture pdf with the individual merged pdfs as components. On the other hand, the same fusion of exponential family pdfs in terms of parameters yields the geometric mean of pdfs resulting in a single pdf $g$ of the same family.

## 5 Examples

The ongoing examples consider distributed Bayesian inference of model parameters with conjugate priors. Generally, this type of inference obeys the following lemma.

**Lemma 4.** *Assume modelling of an observed random variable $y_t$ based on an observed variable $x_t$ and a latent parameter $\theta$. Let the model and the conjugate prior distributions have the forms*

$$f(y_t|x_t, \theta) = \exp\left\{\langle \eta(\theta), T(x_t, y_t)\rangle - A\left(\eta(\theta)\right) + B(x_t, y_t)\right\}$$

*and*

$$\pi(\theta|\xi_{t-1}, \nu_{t-1}) = \exp\left\{\langle \eta(\theta), \xi_{t-1}\rangle - \nu_{t-1} A\left(\eta(\theta)\right) + C(\xi_{t-1}, \nu_{t-1})\right\},$$

*respectively ($\nu_{t-1} \in \mathbb{R}_+$). Then, the Bayesian update*

$$\pi(\theta|\xi_t, \nu_t) = \frac{f(y|x,\theta)\pi(\theta|\xi,\nu)}{\int f(y|x,\theta)\pi(\theta|\xi,\nu)d\theta} \tag{24}$$

*reduces to the linear update of the conjugate prior hyperparameters $\xi$ and $\nu$*

$$\begin{aligned} \xi_t &= \xi_{t-1} + T(x_t, y_t) \\ \nu_t &= \nu_{t-1} + 1. \end{aligned} \tag{25}$$

The proof follows directly from putting the pdfs into the Bayes' rule (24). Observe that the $\pi(\theta|\xi_{t-1}, \nu_{t-1})$ above is not in the exponential family form by Def. 5, as $\eta(\theta)$ is a function of the modelled parameter, not hyperparameters $\xi$ and $\nu$. In static inference, the hyperparameters express the a priori available knowledge in terms of the number $\nu_{t-1}$ of pseudoobservations put into the statistic $\xi_{t-1}$, corresponding to the model's sufficient statistic $T$. In dynamic cases, $\nu$ and $\xi$ contain both the pseudoobservations and the incorporated real observations. In practice, the updating of the form (25) is rather rare in favour of expressions dealing with the original parameter $\theta$.

**Corollary 3** (of Lemma 4)**.** *Bregman Type-1 fusion of exponential family posterior distributions can be achieved either by rewriting the posterior into the exponential family form and using* (22)*, or directly by convex combination of hyperparameters $\xi$ and $\nu$ by* (23)*. These ways agree.*

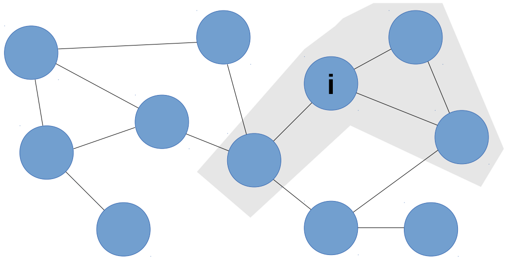

Figure 1: Diffusion network: all nodes locally process the obtained data and share information within a neighborhood ($N^{(i)}$ depicted in grey).

### 5.1 Problem Formulation

The underlying problem is the diffusion estimation of an unknown parameter $\theta$ with a fully decentralized network of cooperating nodes (estimators) $i = 1, \ldots, n$, Fig. 1. These nodes exchange either their observations $x_t^{(i)}, y_t^{(i)}$, or their posterior distributions $\pi^{(i)}(\theta|\xi_t^{(i)}, \nu_t^{(i)})$, $i = 1, \ldots, n$ or both. For simplicity, we will consider only the exchange of the posterior distribution; the identical rules apply to the fusion of observations. For a node $i$, the communication is restricted to the neighborhood $N^{(i)}$, consisting of the nodes within 1-hop distance, including $i$ itself. The node weights the information from the neighbors $j \in N^{(i)}$ by weights $\omega_j^{(i)}$ summing to unity. The generality of this setting allows application of the achieved methods to a broad variety of simpler problems, e.g. the estimation in a fusion center, where we consider only a single node with a neighborhood of all other (non-cooperating) nodes.

Below, we first describe the basic linear regression model and then consider two cases: (a) prior distribution for non-natural model parametrization and (b) prior distribution for natural model parametrization. We highlight that (a) is commonplace. For each case, we give the Bregman Type-1 optimal fusion.

## 6 Diffusion Estimation

### 6.1 Model

We consider modelling of an observed random variable $y_t \in \mathbb{R}$ determined by an observed regressor $x_t \in \mathbb{R}^p$ and latent vector of regression coefficients $\beta \in \mathbb{R}^p$,

$$y_t = x_t^\intercal \beta + \varepsilon_t, \tag{26}$$

where the iid scalar variables $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ represent white noise and $\theta = \{\beta, \sigma^2\}$. This model can be written in the form $y_t|x_t, \beta, \sigma^2 \sim \mathcal{N}(x_t^\intercal \beta, \sigma^2)$ with a pdf

$$f(y_t|x_t, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_t - x_t^\intercal \beta)^\intercal (y_t - x_t^\intercal \beta)\right). \tag{27}$$

Three most popular prior distributions inference of $\beta$ and $\sigma^2$ (or its reciprocal) are the normal inverse-gamma, normal scaled-inverse-$\chi^2$ and normal gamma distributions. We stick with the first one.

Table 1: Normal inverse-gamma parameters.

| **Hyperparameter** | **Natural hyperparameter** | **Sufficient stat.** |
|---|---|---|
| $-a - \frac{p}{2} - 1$ | $-\eta_1 - \frac{p}{2} - 1$ | $\log \sigma^2$ |
| $-b - \frac{1}{2}\mu_\beta^\intercal V_\beta^{-1}\mu_\beta$ | $-\eta_2 + \frac{1}{8}\eta_3^\intercal \eta_4^{-1}\eta_3$ | $\sigma^{-2}$ |
| $V_\beta^{-1}\mu_\beta$ | $-\frac{1}{2}\eta_{4^{-1}\eta_3}$ | $\sigma^{-2}\beta$ |
| $-\frac{1}{2}V_\beta^{-1}$ | $-\frac{1}{2}\eta_4^{-1}$ | $\sigma^{-2}\beta\beta^\intercal$ |

## 6.2 Non-natural parameters

The normal inverse-gamma distribution is a compound of the form (time indices omitted)

$$\mathcal{N}i\mathcal{G}(\mu_\beta, \sigma^2 V_\beta, a, b) = \mathcal{N}(\mu_\beta, V_\beta) \times i\mathcal{G}(a, b)$$

with a pdf

$$\begin{aligned} &\pi(\beta, \sigma^2 | \mu_\beta, V_\beta, a, b) \\ &\quad = b^a (2\pi)^{-\frac{p}{2}} |V_\beta|^{-\frac{1}{2}} \Gamma^{-1}(a) \left(\sigma^2\right)^{-a-\frac{p}{2}-1} \\ &\quad \times \exp\left\{ -\frac{1}{\sigma^2} \left[ b + \frac{1}{2}(\beta - \mu_\beta)^\intercal V_\beta^{-1} (\beta - \mu_\beta) \right] \right\} \end{aligned} \tag{28}$$

where $V_\beta \in \mathbb{R}^{p \times p}$ is a symmetric positive definite scaling matrix. The posterior pdf

$$\pi(\beta, \sigma^2 | \mu_{\beta,t}, V_{\beta,t}, a_t, b_t) \propto f(y_t | x_t, \beta, \sigma^2) \times \pi(\beta, \sigma^2 | \mu_{\beta,t-1}, V_{\beta,t-1}, a_{t-1}, b_{t-1})$$

is given by updated hyperparameters [24]

$$\begin{aligned} \mu_{\beta,t} &= V_{\beta,t} \left( V_{\beta,t-1}^{-1} \mu_{\beta,t-1} + x_t y_t \right) \\ V_{\beta,t} &= \left( V_{\beta,t-1}^{-1} + x_t x_t^\intercal \right)^{-1} \\ a_t &= a_{t-1} + \frac{1}{2} \\ b_t &= \frac{1}{2} \left[ y_t^2 + \mu_{\beta,t-1}^\intercal V_{\beta,t-1}^{-1} \mu_{\beta,t-1}^\intercal - \mu_{\beta,t-1}^\intercal V_{\beta,t}^{-1} \mu_{\beta,t-1}^\intercal \right] + b_{t-1}. \end{aligned}$$

Consider now the aforementioned diffusion estimation problem with pdfs $\pi^{(i)}(\beta, \sigma^2|\cdot)$ and weights $\omega_j^{(i)} \in [0, 1]$ summing to unity. The goal is to find a Bregman Type-1 optimal pdf of the same type ($\mathcal{N}i\mathcal{G}$), as close to the particular $\pi^{(i)}$s as possible. In order to exploit Proposition 1, we need to rewrite (28) into the exponential family form, Def. 5. Using simple algebraic operations one arrives at the hyperparameters, their natural form and the connected sufficient statistics, summarized in Table 1. By direct use of (22), we arrive at the optimal fused posterior hyperparameters

$$\begin{aligned}
\tilde{\mu}_{\beta,t}^{(i)} &= \tilde{V}_{\beta,t}^{(i)} \cdot \sum_{j\in N^{(i)}} \omega_j^{(i)} V_{\beta,t}^{(j)-1} \mu_{\beta,t}^{(j)} \\
\tilde{V}_{\beta,t}^{(i)} &= \left( \sum_{j\in N^{(i)}} \omega_j^{(i)} V_{\beta,t}^{(j)-1} \right)^{-1} \\
\tilde{a}_t^{(i)} &= \sum_{j\in N^{(i)}} \omega_j^{(i)} a_t^{(j)} \\
\tilde{b}_t^{(i)} &= \sum_{j\in N^{(i)}} \omega_j^{(i)} b_t^{(j)}.
\end{aligned} \tag{29}$$

### 6.3 Natural Parameterization

Peterka [25] has shown that the model (27) can be identically rewritten into the form

$$f(y_t|x_t,\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \mathrm{Tr}\left( \underbrace{\frac{-1}{2\sigma^2} \begin{bmatrix} -1 \\ \beta \end{bmatrix} \begin{bmatrix} -1 \\ \beta \end{bmatrix}^\intercal}_{\eta^\intercal} \underbrace{\begin{bmatrix} y_t \\ x_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix}^\intercal}_{T(x_t,y_t)} \right) \right\}.$$

The prior normal inverse-gamma distribution $\mathcal{N}i\mathcal{G}(\xi,\nu)$ for the natural model parameterization has the Peterka's form

$$\pi(\beta,\sigma^2|\xi_{t-1},\nu_{t-1}) = \exp\left\{ \mathrm{Tr}\left( \underbrace{\frac{-1}{2\sigma^2} \begin{bmatrix} -1 \\ \beta \end{bmatrix} \begin{bmatrix} -1 \\ \beta \end{bmatrix}^\intercal}_{\eta^\intercal} \xi_{t-1} \right) + \nu A\left(\eta(\beta,\sigma^2)\right) + C(\xi,\nu) \right\}.$$

The counterparts of parameters $\mu$ and $V_\beta$ are

$$\begin{aligned}
\mu_\beta &= \xi_{[2:p+1,2:p+1]}^{-1} \xi_{2:p+1,1} \\
V_\beta &= \xi_{[2:p+1,2:p+1]}^{-1}.
\end{aligned}$$

This becomes obvious from the Bayesian update (24) taking exactly the form (25),

$$\begin{aligned}
\xi_t &= \xi_{t-1} + \begin{bmatrix} y_t \\ x_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix}^\intercal \\
\nu_t &= \nu_{t-1} + 1,
\end{aligned}$$

simply by recognizing the updates of the relevant blocks of $\xi_{t-1}$.

The Bregman Type-1 fusion of posteriors from the neighborhood $N^{(i)}$ is hence simply

$$\begin{aligned}
\xi_t^{(i)} &= \sum_{j\in N^{(i)}} \omega_j^{(i)} \xi_t^{(j)} \\
\nu_t^{(i)} &= \sum_{j\in N^{(i)}} \omega_j^{(i)} \nu_t^{(j)},
\end{aligned} \tag{30}$$

which coincides with the Kullback-Leibler optimal Type-2 fusion. In the scope of the diffusion Bayesian linear regression, this yields the whole-pdf combine step proposed by the author [26].

If this method is applied to observation, the adapt step of the diffusion recursive least squares proposed by Cattiveli *et al.* [27] is achieved in the Bayesian form, c.f. Supplement of [26].

A particularly interesting aspect of (30) (and indeed (29)) is that it can be seen as a sequence of several Bayesian updates, where the $i$th node's posterior pdf with hyperparameters $\omega_i \xi_t^{(i)}$ and $\omega_i \nu_t^{(i)}$ is updated by the corresponding discounted "sufficient statistics" (or pseudoobservations) from the neighborhood (excluding $i$). In this respect, the Bregman Type-1 fusion of exponential family pdfs is in certain sense Bayes-optimal and inherits the properties of the traditional Bayes' estimators. This deserves further research yet.

The choice of weights $\omega_j^{(i)}$ drives the values of the estimates between two extremes – the "best" estimate and the "worst" estimate within $N^{(i)}$ (in the sense of the variance/bias due to the potential presence of the additional noise connected with the nodes). The distributed estimator is hence a shrinkage estimator.

### 6.4 Functional Fusion

The functional Bregman Type-1 fusion of whole pdfs yields the mixture density

$$\pi^{(i)}(\beta, \sigma^2 | \cdot) = \sum_{j \in N^{(i)}} \omega_j^{(i)} \pi^{(j)}(\beta, \sigma^2 | \cdot).$$

It coincides with the Kullback-Leibler optimal Type-1 fusion in terms of hyperparameters. The mixture is not an exponential family distribution. Its use in dynamic problems is complicated, making its approximation by a single pdf is practically inevitable. The development of a method allowing functional Type-1 fusion within a preselected class of distributions is hence quite attractive.

## 7 Conclusion

The paper shows the potential of the Bregman divergences for information fusion in distributed systems. The functional form is shown to yield a very general methodology, applicable to a quite wide range of problems, from the viewpoint of functions classes to the parameters of these functions. The Bregman divergences generated by the exponential family log-partition are shown to coincide with the Kullback-Leibler divergence. The Bregman-optimal Type-1 fusion of exponential family distributions in terms of hyperparameters particularizes to the Kullback-Leibler optimal Type-2 fusion of probability densities, resulting in a single distribution of the identical class.

The future work comprises further exploration of the functional Bregman divergences for information fusion. For instance, the related information-geometric scope is a very promising field. Another challenge is the restricted functional fusion within an a priori specified class of distributions. In other words, the Bregman-optimal fusion of a set of general pdfs $\{f_i\}$ yielding a single pdf $g$ of the same type, evaluated in the (significantly more general) function rather than parameter space.

## Acknowledgement

The work is supported by the Czech Science Foundation, grant no. 14-06678P.

# References

[1] L. M. Bregman, “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,” *USSR Computational Mathematics and Mathematical Physics*, vol. 7, pp. 200–217, 1967.

[2] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, 2000, vol. 191.

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

[4] R. Nock and F. Nielsen, “On weighting clustering.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1223–1235, 2006.

[5] M. Collins, R. E. Schapire, and Y. Singer, “Logistic regression, AdaBoost and Bregman distances,” *Machine Learning*, vol. 48, pp. 253 – 285, 2002.

[6] A. Banerjee, “An analysis of logistic models: Exponential family connections and online performance,” in *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007, pp. 204–215.

[7] J.-D. Boissonnat, F. Nielsen, and R. Nock, “Bregman Voronoi diagrams,” *Discrete & Computational Geometry*, vol. 44, no. 2, pp. 281–307, 2010.

[8] B. A. Frigyik, S. Srivastava, and M. R. Gupta, “Functional Bregman divergence and Bayesian estimation of distributions,” *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 5130–5139, Nov. 2008.

[9] H. Chernoff *et al.*, “A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations,” *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[10] J. Havrda and F. Charvát, “Quantification method of classification processes.,” *Kybernetika*, vol. 3, no. 1, pp. 30–35, 1967.

[11] C. Tsallis, “Possible generalization of Boltzmann-Gibbs statistics,”, *Journal of Statistical Physics*, vol. 52, no. 1–2, pp. 479–487, 1988.

[12] I. Csiszár, “Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten,” *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 85–108, 1963.

[13] A. Rényi, “On measures of entropy and information,” in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 547, 1961, pp. 547–561.

[14] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, “Robust and efficient estimation by minimising a density power divergence,” *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[15] H. Fujisawa and S. Eguchi, “Robust parameter estimation with a small bias against heavy contamination,” *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053–2081, 2008.

[16] S. I. Amari, "$\alpha$-divergence is unique, belonging to both f-divergence and Bregman divergence classes," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 4925–4931, Nov. 2009.

[17] A. Cichocki and S.-i. Amari, "Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, pp. 1532–1568, 2010.

[18] M. Basseville, "Divergence measures for statistical data processing – An annotated bibliography," *Signal Processing*, vol. 93, no. 4, pp. 621–633, Apr. 2013.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[20] T. Minka, "Expectation Propagation for approximate Bayesian inference," in *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[21] C. Bishop, *Pattern Recognition and Machine Learning.*NJ, USA: Springer-Verlag New York, Inc., 2006.

[22] J. Winn and C. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[23] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. John Wiley & Sons Ltd, 1978.

[24] L. Fahrmeir, T. Kneib, S. Lang and B. Marx, "Regression: Models, Methods and Applications," Springer-Verlag New York, Inc., 2013.

[25] V. Peterka, "Bayesian approach to system identification," In *Trends and Progress in System Identification*, P. Eykhoff, Ed. Oxford, U.K.: Pergamon, pp. 239–304, 1981.

[26] K. Dedecius and V. Sečkárová, "Dynamic diffusion estimation in exponential family models," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1114–1117, Nov. 2013.

[27] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Processing*, vol. 56, no. 5, pp. 1865–1877, May 2008.