A characterization of all single-integral, non-kernel divergence estimators

Soham Jana and Ayanendranath Basu

***Abstract*—Divergence measures have been used for a long time for different purposes in information theory and statistics. In particular, density-based minimum divergence estimation is a popular tool in the statistical literature. Given the sampled data and a parametric model, we estimate the model parameter by choosing the member of the model family that is closest to the data distribution in terms of the given divergence. In the absolutely continuous set up, when the distributions from the model family and the unknown data generating distribution are assumed to have densities, the application of kernel based non-parametric smoothing is sometimes unavoidable to get an estimate of the true data density. The use of kernels (or other non-parametric smoothing techniques) makes the estimation process considerably more complex, as now one has to impose necessary conditions not just on the model but also on the kernel and its bandwidth. In higher dimensions the efficiency of the kernel density estimator (KDE) often becomes too low for the minimum divergence procedure to be practically useful. It can, therefore, lead to a significant advantage to have a divergence which allows minimum divergence estimation bypassing the use of non-parametric smoothing. For the same reason, characterizing the class of such divergences would be a notable achievement. In this work, we provide a characterization of the class of divergences that bypasses the use of non-parametric smoothing in the construction of divergences, providing a solution to this very important problem.**

***Index Terms*—Bregman divergence, Characterization, Single-integral, Non-kernel divergence.**

I. Introduction

THE use of minimum distance (or, more generally, minimum divergence) methods in statistics has a long history. It is a very natural method, where one matches a suitable empirical quantity against its model counterpart and optimizes the degree of this match over the model elements. Pearson's chi-square (Pearson, 1900) and the corresponding minimum chi-square method provides an early example of the matching of the empirical and model probabilities (or frequencies); thus the history of minimum distance methods must be traced back at least that far. In the 1950s, Wolfowitz studied minimum distance methods as a class; see, e.g., Wolfowitz (1957). Of particular interest to us is the class of density-based minimum divergence procedures, of which the maximum likelihood method and the minimum chi-square method are special cases. A density-based statistical divergence between two densities is a nonnegative measure of discrepancy which equals zero only when the densities are identically equal. They are distance like measures, but need not be mathematical metrics.

The approach of minimum divergence methods got a major boost in the 1960s, with the development of the Csiszár class of divergences (Csiszár 1963, 1967; Ali and Silvey 1966), although it did take the statisticians a few more decades to get a general understanding of the scope and full potential of the procedures resulting from the minimization of these divergences. These divergences have been called by several names including ϕ-divergences, f-divergences and disparities. Within the field of statistical inference, the topic of robustness had started to grow roughly around the same time, with (Huber, 1964) providing the first formalization of the theory of robustness. Along this line of research came Beran's seminal paper on Hellinger distance (Beran, 1977), providing the first formal link between minimum distance methods and the corresponding robustness advantages. Beran's paper actually does more in that it establishes the asymptotic efficiency of the minimum Hellinger distance estimator and demonstrates, possibly for the first time, that asymptotic efficiency and robustness need not be in conflict.

Beran's work opened up a new line of research and was followed up by, among others, Tamura and Boos (1986), Donoho and Liu (1988), and Simpson (1987, 1989). During the 1980s and 1990s much effort was also spent in trying to extend the scope of robust minimum divergence inference beyond the Hellinger distance. Significant contributions during this period came from Cressie and Read (1984) and Lindsay (1994). Although the celebrated Cressie-Read family of divergences were primarily applied by the authors in the context of goodness-of-fit testing, the huge potential of these divergences in robust (and efficient) minimum divergence inference was also immediately appreciated by the research community. Lindsay, on the other hand, described the entire geometry behind the method of inference based on disparities (ϕ-divergences), and explained which characteristics made these methods robust even when the influence function approach failed to capture their robustness. He developed several other robustness measures which helped to provide a complete framework for minimum disparity inference. Construction of local divergences and local hypothesis testing has been taken up by Avlogiaris et al. (2016 a,b) in this connection. At least four books (Liese and Vajda 1987; Vajda 1989; Pardo 2006; Basu, Shioya, and Park 2011) cover different aspects of statistical inference based on ϕ-divergences or disparities.

In spite of the successful application of the minimum divergence procedures based on disparities in many domains,

S. Jana is with the Department of Statistics and Data Science, Yale University, New Haven, CT, USA, email: soham.jana@yale.edu

A. Basu is with the Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, West Bengal, India, email: ayanbasu@isical.ac.in

there is one major hindrance in the application of these methods under continuous models. The observed data are always discrete, irrespective of the model. In continuous models, therefore, it is necessary to construct a continuous estimate of the data generating density from the observed data before the disparity can be computed. This requires the use of a suitable non-parametric smoothing technique such as kernel density estimation (see, e.g., Devroye and Györfi 1985; Silverman 1986; Scott 1992; Wand and Jones 1994). Density estimation techniques are also used extensively in model selection (see, e.g., Lerasle 2012). However, this component often introduces substantial additional complication in the inference process. Given two distributions G and F, having densities g and f with respect to a common dominating measure, the disparity $\rho(G, F)$ between these two distributions is defined by

$$\rho(G,F) = \int C(\delta(x)) f(x) dx, \qquad (1)$$

where $\delta(x) = g(x)/f(x) - 1$, and $C(\delta)$ is a convex function on $[-1, \infty)$ with $C(0) = 0$. Consider the standard set up of parametric inference where $X_1, \ldots, X_n$ is an independently and identically distributed (i.i.d.) sample from the true distribution G, modeled by the parametric family of distributions $\{F_\theta : \theta \in \Theta\}$; let f_θ represent the density of F_θ. The minimum disparity estimator of θ is obtained by choosing the model element which is closest to the data in the sense of the given disparity. Since the true density g is unknown in practice, one needs to minimize the right hand side of the equation (1) over $\theta \in \Theta$, where one replaces the density g with a non-parametric density estimate g^*. In continuous models, the construction of this density estimate requires the use of non-parametric smoothing techniques such as kernel density estimation. As indicated, this makes the method more complicated, both from a theoretical point of view as well as from the point of view of implementation. This includes, among other things, the problem of bandwidth selection and slow convergence of kernels in higher dimensions. Yet all the methods described in the two previous paragraphs would inevitably run into this problem when dealing with continuous models. We consider two representative disparities, the Kullback-Leibler divergence and the Hellinger distance, which are frequently used in statistical methodology for different purposes, including minimum divergence estimation. The forms of these two divergences are given by

$$\rho_{KL}(G,F) = \int g(x) \log\left(\frac{g(x)}{f(x)}\right) dx,$$
$$\rho_{HD}(G,F) = \int \left(g^{1/2}(x) - f^{1/2}(x)\right)^2 dx.$$

When G represents the true distribution and $F = F_\theta$ represents the model distribution, the Kullback-Leibler divergence may be represented as

$$\begin{aligned}\rho_{KL}(G,F_\theta) &= \int g(x) \log\left(\frac{g(x)}{f_\theta(x)}\right) dx \qquad (2)\\ &= \int g(x)\log(g(x))dx - \int g(x)\log(f_\theta(x))dx\\ &= M - \int g(x)\log(f_\theta(x))dx\\ &= M - \int \log(f_\theta(x))dG(x).\end{aligned}$$

Note that the quantity M in the above equation is independent of θ. When G is replaced, based on an i.i.d. sample of size n, by the empirical cumulative distribution function, say $\hat{G}$, the quantity to be minimized becomes $-n^{-1}\sum \log f_\theta(X_i)$; minimizing this is equivalent to maximizing the log likelihood, which generates the maximum likelihood estimator. Thus one can get by here by using the empirical distribution function and not requiring the kernel density estimator or a similar construct. Notice that here we have replaced the theoretical expectation $\int \log(f_\theta(x))dG(x)$ by its empirical estimate $n^{-1}\sum \log f_\theta(X_i)$, the corresponding sample mean. However, minimizing the Hellinger distance is equivalent to maximizing $\int g^{1/2}(x) f_\theta^{1/2}(x)dx$, and given the data from the unknown distribution G, this maximization can not be performed without constructing a continuous density estimate, unlike the Kullback-Leibler divergence. In fact the Kullback-Leibler divergence is the only divergence within the class of disparities which allows minimum divergence estimation without the construction of a density estimate (Basu, Shioya, and Park, 2011).

To motivate this further, we also consider the class of density power divergences (Basu et al., 1998), where the divergences have the form

$$d_\alpha(G,F_\theta) = \int f_\theta^{1+\alpha}(x)dx \quad - \left(1+\tfrac{1}{\alpha}\right) \int g(x) f_\theta^\alpha(x)dx + \tfrac{1}{\alpha} \int g^{1+\alpha}(x)dx. \qquad (3)$$

Here α (taking values in [0, 1]) is the tuning parameter of this family of divergences. Since the last term on the right hand side of the above equation is independent of θ, the minimization of the above divergence, based on an i.i.d. random sample $X_1, \ldots, X_n$ of size n, can be achieved by the minimization of $\int f_\theta^{1+\alpha}(x)dx - \left(1+\frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^n f_\theta^\alpha(X_i)$ over θ in Θ. Here also we have replaced the theoretical expectation by the sample mean in the objective function to be minimized. Thus, while the minimization of any disparity (except the Kullback-Leibler divergence) requires a continuous density estimate, the minimization of the density power divergence can be undertaken without any continuous density estimate for any α in [0,1]. The density power divergence and similar techniques have been used heavily in the literature. In particular it has been used extensively in model selection (e.g., Mattheou et al. 2009). Mattheou and Karagrigoriou (2010) have also used it for tests of fit.

The necessity of a non-parametrically smoothed density estimate is a major issue in density-based minimum divergence estimation. The properties of the resulting estimator depend quite heavily on the properties of the kernel density estimator (KDE), both in respect of its efficiency and robustness. For example, the rate of convergence of the KDE has a direct impact on the convergence of the estimator, and slow convergence of the KDE, particularly in multivariate settings, is a definite problem; see, for example, Tamura and Boos (1986). On the other hand, the robustness of the estimator has a direct relation to the bandwidth selection problem. Very large bandwidths lead to overly smooth KDEs, leading to a loss in robustness. Very small bandwidths make the KDE spiky and relatively unsmooth, and the iterative algorithms for solving the estimating equations run into issues of slow and unstable convergence. Thus the divergences which are necessarily dependent on the KDE approach have to overcome all these difficulties to be successfully applied. Because of all these reasons, suitable non-kernel minimum divergence estimators provide alternatives which allow very substantial simplification on the method based on disparities and other KDE based divergences. Characterizing such non-kernel divergences is, therefore, an exercise which is of tremendous practical importance. In the present paper we have provided a solution to this very important problem and provide a neat and useful characterization of non-kernel, single-integral divergences.

II. General form of non-kernel divergence measures

In practice, to do minimum divergence inference, one has to get an empirical estimate of the divergence based on the sampled data which can then be optimized over the parameter space. In general there are two ways for constructing such an empirical estimate of the theoretical divergence.

1) In the first, one directly replaces the unknown data generating density by a suitable non-parametric density estimate. Under appropriate assumptions on the bandwidth sequence, the corresponding density estimator converges to the true density g; depending on the divergence, the empirical divergence estimator becomes consistent for the theoretical divergence.
2) The method described in the previous item has the drawback that in this case the bandwidth selection problem and other related convergence issues have to be tackled as an inevitable part of kernel density estimation (or other similar non-parametric smoothing techniques). This can be avoided when the empirical divergence can be represented as an i.i.d. average over the observed data points that converges to the theoretical divergence by the weak law of large numbers. Suppose G and F are two univariate distributions having densities g and f and we want to estimate the divergence between them. Typically a density based divergence will involve the following terms; (a) an integral term (or a function of an integral term) involving both the densities g and f; (b) an integral term (or a function of an integral term) involving only the density g; (c) an integral term (or a function of an integral term) involving only the density f. Henceforth we define the integral of the divergence on the common support of the two involved distributions, G and F. Now suppose we have i.i.d. data $X_1, \ldots, X_n$ from the distribution G; let $\hat{G}$ represent the empirical distribution and let G be modeled by an absolutely continuous parametric family of distributions $\{F_\theta : \theta \in \Theta\}$. When constructing a divergence between G and F_θ, the term of the type (a) will, in general be of the form $\int A(g(x), f_\theta(x))dx$. When the relation $A(s,t) = N(t)s$ holds for some suitable univariate function $N(\cdot)$, we have

$$\begin{aligned} \int A(g(x), f_\theta(x))dx &= \int N(f_\theta(x))g(x)dx \\ &= \int N(f_\theta(x))dG(x) \end{aligned}$$

and we can estimate the divergence between G and F_θ by using the empirical distribution $\hat{G}$ in place of G to get $\int N(f_\theta(x))d\hat{G}(x) = \frac{1}{n}\sum_{i=1}^{n} N(f_\theta(X_i))$, and we do not need to estimate the density g separately. Divergence measures where the structure $A(s,t) = N(t)s$ holds and the corresponding empirical divergence may be constituted without taking recourse to non-parametric smoothing will be referred to herein as non-kernel divergence measures. Broniatowski et al. (2012) have referred to such divergences as 'decomposable'. Minimum divergence estimators obtained by minimizing decomposable divergences will be called non-kernel divergence estimators.

Remark 1. A relevant associated question is whether there could be other empirical formulations of the theoretical divergence where the structure is not of the $A(s,t) = N(t)s$ type, but the empirical version can be constructed without resorting to kernel density estimation (or a similar non-parametric construct). As of now, no such technique is known to us, and there does not seem to be any obvious way to do it, since for any structure other than $A(s,t) = N(t)s$, the sample mean representation of the empirical divergence, which converges to the theoretical divergence by the weak law of large numbers, is not possible. It is left to future research to more comprehensively explore other possible non-kernel representations, but as of now, our non-kernel decomposable divergences are defined by the structure $A(s,t) = N(t)s$.

The structure defined above takes us to the following general form of non-kernel divergence measure between G and F with corresponding densities g and f including terms of the type given in (a), (b) and (c) above as

$$D(G,F) = \int \Big[B_1(g(x)) - B_2(f(x)) - (g(x) - f(x))B_3(f(x))\Big]dx, \tag{4}$$

for some suitably chosen functions B_1, B_2, and B_3. When $F = F_\theta$, an element of our parametric family, we need to minimize the above divergence with respect to θ to estimate the parameter θ of interest. The term $\int B_1(g(x))$ has no contribution in the minimization procedure and we can exclude this term from consideration. The reduced problem requires the minimization of

$$\int \big[-B_2(f_\theta(x)) - (g(x) - f_\theta(x))B_3(f_\theta(x))\big]dx.$$

Now, by replacing G by $\hat{G}$ we get the empirical measure

$$-\frac{1}{n}\sum_{i=1}^{n} B_3(f_\theta(X_i)) + \int \big[-B_2(f_\theta(x)) + f_\theta(x)B_3(f_\theta(x))\big]dx \tag{5}$$

which can be minimized with respect to θ without any non-parametric smoothing. Note that the Kullback-Leibler divergence considered in equation (2) is decomposable in the sense described in this section, with $A(s,t) = N(t)s$, where $N(t) = \log(t)$.

Remark 2. Although the construction of the above empirical measure does not require any non-parametric smoothing, implementation of the minimization method may involve other difficulties. Calculating the integral in equation (5) may not always be easy due to the possibly complex natures of the functions B_2, B_3, and the model family F_θ. Even though the integral $\int f_\theta(x)B_3(f_\theta(x))\ dx$ can be approximated, using the Monte-Carlo idea, by $\frac{1}{m}\sum_{j=1}^{m} B_3(f_\theta(Y_j))$ where Y_j's represent an i.i.d sample from the distribution which has density $f_\theta(x)$, the integral involving B_2 might have to be approached differently. For example using importance sampling (see, eg., Smith, 1997) one can approximate $\int B_2(f_\theta(x))\ dx$ by $\frac{1}{m}\sum_{j=1}^{m} B_2(f_\theta(Y_j))/f_\theta(Y_j)$ where Y_j's are generated as before. It should be noted that approximating the integrals by different sampling procedures might slow down the convergence of the estimates to the truth, but that seems also inevitable under the circumstances. In short, the evaluation of the quantity in equation (5) may require involved computation depending on the nature of the functions involved. We hope to take up this problem in our future work.

Remark 3. In constructing the divergence $D(G, F)$, we have assumed that the components of (a), (b) and (c) are additive and therefore can be expressed as a single-integral. The density power divergence introduced in equation (3) is a ready example. So are all the members of the class of the disparities. We will refer to such divergences as single-integral divergences. However there may be legitimate non-kernel divergence measures which are not single-integral divergences. An example of this is provided by the class of logarithmic density power divergence (LDPD) measures (Jones et al., 2001). The LDPD between distributions G and F_θ with densities g and f_θ is given by

$$\begin{aligned} Q(G, F_\theta) = &\log\left(\int f_\theta^{1+\beta}(x)dx\right) \\ &- \left(1+\frac{1}{\beta}\right)\log\left(\int g(x)f_\theta^{\beta}(x)dx\right) \\ &+ \frac{1}{\beta}\log\left(\int g^{1+\beta}(x)dx\right), \end{aligned}$$

where β is a suitably chosen tuning constant. This is clearly not a single integral divergence, nor the logarithm of one.

We will refer to all minimum divergence estimators obtained by minimizing single-integral, non-kernel divergence measures as “single-integral, non-kernel divergence estimators”.

III. Bregman divergence

A prominent class of non-kernel divergence measures used to summarize the statistical discrepancy between two distributions is the class of “Bregman Divergences”. See, e.g., Stummer and Vajda (2012). Suppose G and F are two distributions having corresponding densities g and f. Then the Bregman divergence between these distributions is expressed as

$$\begin{aligned} &D_B(G, F) \\ &= \int \big[B(g(x)) - B(f(x)) - (g(x) - f(x))B'(f(x))\big]dx \end{aligned} \tag{6}$$

where B is some suitable smooth convex function and B' represents the derivative of B with respect to its argument. It is easy to check that it possesses all the properties of a valid statistical divergence. In the next section we will systematically demonstrate that the form of divergence presented in (4) reduces to that of the Bregman divergence under some very general assumptions including the convexity of B_1 and very elementary smoothness properties of B_1, B_2, B_3. This provides a characterization of single-integral non-kernel divergences and shows that under very general conditions, all such divergences must be Bregman divergences.

Remark 4. Continuing the above argument, one can construct examples of known divergences that can be written as Bregman divergences. A few of them are provided in Table I. The L^2 distance is a very common loss function that is used for density estimation under different circumstances and do not need further explanation. The DPD or the BHHJ family of divergences, first introduced in (Basu et al., 1998), is a well known distance function that is used under the density power divergence set up (Basu, Shioya, and Park, 2011). The parameter estimates based on BHHJ family exihibit asymptotic normality as well as robustness under basic assumptions on the model family, as demonstrated in Basu, Shioya, and Park (2011). The limiting divergence for the BHHJ family as α reaches 0 is actually the Kullback-Leibler, which is a fundamental distance measure in the literature of information theory. The Itakura-Saito divergence (Itakura and Saito, 1968) is used in the literature for non-negative matrix factorization (Févotte et al., 2009). The Bregman exponential divergence gives another instance of a simple Bregman divergence generated by a common convex function (Mukherjee et al., 2018).

The Bregman divergence is one of the most studied families of divergence measures. Apart from the usual benefits the divergences possess due to convexity of the underlying B function, which is helpful in the context of the various optimization problems, they also have links to the exponential family of distributions. Given any member of the regular exponential family one can write the density uniquely in terms of exponential of a Bregman divergence with a properly chosen B function, as mentioned in Banerjee et al. (2005,

TABLE I
BREGMAN DIVERGENCES WITH DIFFERENT B FUNCTIONS

Divergence	$D_B(G, F_\theta)$	$B(x)$
Kullback-Leibler	$\int g(x) \log \frac{g(x)}{f_\theta(x)} dx$	$x \log(x)$
L^2 distance	$\int (g(x) - f_\theta(x))^2 \, dx$	x^2
DPD/ BHHJ Family (Generalization of L^2)	$\int \left\{ f_\theta^{1+\alpha}(x) - (1 + \frac{1}{\alpha}) g(x) f^\alpha(x) + \frac{1}{\alpha} g^{1+\alpha}(x) \right\} dx$ $\alpha > 0$	$\frac{x^{(1+\alpha)}}{\alpha}$
Itakura-Saito distance	$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \frac{g(x)}{f_\theta(x)} - \log \frac{g(x)}{f_\theta(x)} + 1 \right\} dx$	$-\frac{\log(x)}{2\pi}$
Bregman Exponential Divergence	$\frac{2}{\alpha} \int \left\{ e^{\alpha f(x)} \left(f(x) - \frac{1}{\alpha} \right) - e^{\alpha f(x)} g(x) + \frac{1}{\alpha} e^{\alpha g(x)} \right\} dx$	$\frac{2(e^{\alpha x} - \alpha x - 1)}{\alpha^2}$

Theorem 4). Other then that, in case of discrete state spaces the Bregman divergence based on the empirical mass function is asymptotically distributed as χ^2 with suitably chosen degrees of freedom under general regularity conditions (Pardo and Vajda, 2001), which is often utilized in hypothesis testing. Bregman divergences also generalize the Pythagorean identity in functional spaces, which provides better understanding of the underlying mathematical structure (Frigyik et al., 2008).

IV. RESULTS

Here we prove the claim about how the divergence presented in (4) can be reduced to some element in the class of Bregman divergences. We start with a subclass of the divergences presented in (4), which relates to the case where the integrand of the general form is nonnegative. In the following theorem we show that the aforementioned nonnegativity forces the integral to be a member of the class of Bregman divergence.

Theorem 1. *Suppose that the quantity $D(G, F)$ defined in (4) is a valid statistical divergence, $B_1(\cdot)$ is continuously differentiable and $B_3(\cdot)$ is continuous in their respective arguments. Then if the integrand in the expression of $D(G, F)$ in equation (4) is nonnegative for all absolutely continuous distributions G and F, then $B_1(\cdot)$ is identically equal to $B_2(\cdot)$ and $B_3(\cdot)$ is identically equal to $B_1'(\cdot)$, where $B_1'(\cdot)$ is the derivative of $B_1(\cdot)$.*

Proof. We first show that $B_1(x) = B_2(x)$ for all x in $\mathbb{R}^+$. As the functions $B_1(\cdot)$, $B_2(\cdot)$ and $B_3(\cdot)$ will be applied to densities, we can assume their domains to be $\mathbb{R}^+$. Under the assumption that $D(G, F)$ is a valid divergence between any two absolutely continuous distributions G and F, we have $D(G, F) = 0$ if and only if G is identically equal to F. Given any $x > 0$, consider the distribution H with density h given by

$$h(y) = \begin{cases} x, & y \in \left(0, \frac{1}{x}\right) \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Now take $G = F = H$. Then $D(G, F) = 0$ implies

$$\int_0^{1/x} [B_1(g(y)) - B_2(f(y))] dy = \frac{B_1(x) - B_2(x)}{x} = 0, \tag{8}$$

which implies that

$$B_1(x) = B_2(x).$$

As x is arbitrary, we get $B_1(x) = B_2(x)$ for all x in $\mathbb{R}^+$.

Henceforth we replace B_2 by B_1 in the expansion of the divergence in (4). To prove the remaining part, again fix $x > 0$, arbitrarily. Now consider the sequence of distributions G_n given by the corresponding sequence of densities

$$g_n(y) = \begin{cases} x - \frac{1}{n}, & y \in \left(0, \frac{1}{2x}\right) \\ x + \frac{1}{n}, & y \in \left(\frac{1}{2x}, \frac{1}{x}\right) \end{cases}.$$

Take F to be identically equal to H as defined in (7). As the integrand of $D(G_n, F)$ is always nonnegative by assumption, we get for $y \in \left(\frac{1}{2x}, \frac{1}{x}\right)$,

$$\begin{aligned} & B_1(g_n(y)) - B_1(f(y)) - (g_n(y) - f(y)) B_3(f(y)) \\ & = B_1\left(x + \frac{1}{n}\right) - B_1(x) - \frac{1}{n} B_3(x) \geq 0 \end{aligned} \tag{9}$$

for all n. But this implies

$$\frac{B_1(x + 1/n) - B_1(x)}{1/n} \geq B_3(x), \tag{10}$$

for all n. Taking limit as n goes to infinity, continuity of $B_1'(\cdot)$ gives us

$$B_1'(x) \geq B_3(x). \tag{11}$$

Now we interchange the role of F and G_n. As the integrand of $D(F, G_n)$ is always nonnegative by assumption, we get for $z \in \left(\frac{1}{2x}, \frac{1}{x}\right)$,

$$\begin{aligned} &B_1(f(z)) - B_1(g_n(z)) - (f(z) - g_n(z))B_3(g_n(z)) \\ &= B_1(x) - B_1\left(x + \frac{1}{n}\right) + \frac{1}{n}B_3\left(x + \frac{1}{n}\right) \geq 0 \end{aligned}$$

for all n. This implies

$$\frac{B_1(x) - B_1(x + 1/n)}{-1/n} \leq B_3\left(x + \frac{1}{n}\right),$$

for all n. Taking limit as n goes to infinity, the above equation yields

$$B_1'(x) \leq B_3(x), \tag{12}$$

by continuity of both $B_1'(\cdot)$ and $B_3(\cdot)$ in their arguments. Combining equations (11) and (12) we get $B_1'(x) = B_3(x)$. But as $x > 0$ is arbitrary, we get $B_1'(\cdot)$ is identically equal to $B_3(\cdot)$. This concludes the proof. □

The result proved in Theorem 1, useful as it is, is still of limited interest. It does not rule out the possibility of the existence of a valid single integral divergence with the integrand being negative on a positive measure set; consider the Kullback-Leibler divergence given by $\rho_{\text{KL}}(G, F) = \int g(x) \log(g(x)/f(x))dx$, between distributions G and F with densities g and f, as an example. One should therefore, find out the properties of B_1, B_2, B_3 that are required to cover the case where the only constraint is the non-negativity of the whole integral and its equality to zero if and only if $G = F$, without requiring the non-negativity of the integrand itself. These are, in fact the only conditions needed to make $D(G, F)$ a valid divergence. However it is remarkable that only additional necessity for this purpose is the strict convexity of the function B_1. This gives rise to the following theorem.

Theorem 2. *If the function $D(G, F)$ defined in (4) is a valid divergence between any two absolutely continuous distributions G and F and B_1 is strictly convex and differentiable on the positive real line, then $B_2(x) = B_1(x)$ as in Theorem 1 and we have $B_3(x) = \frac{d}{dx}B_1(x)$ for all x in $\mathbb{R}^+$ up to an additive constant independent of x. Hence $D(G, F)$ is identical to the $D_B(G, F)$ divergence generated by the function B_1.*

Proof. The proof of $B_1(\cdot) \equiv B_2(\cdot)$ follows exactly as in Theorem 4.1 so we do not repeat it. Henceforth we take $B_1(x) = B_2(x) = B(x)$ for all x in $\mathbb{R}^+$. The divergence can then be rewritten as

$$\begin{aligned} &D(G, F) \\ &= \int \left[(g(x) - f(x))\left\{\frac{B(g(x)) - B(f(x))}{g(x) - f(x)} - B'(f(x))\right\}\right] dx \\ &\quad + \int [(g(x) - f(x))(B'(f(x)) - B_3(f(x)))]dx \\ &= \int \left[(g(x) - f(x))\left\{\frac{B(g(x)) - B(f(x))}{g(x) - f(x)} - B'(f(x))\right\}\right] dx \\ &\quad + \int [(g(x) - f(x))M(f(x))]dx \end{aligned}$$

where $M := B' - B_3$. If $M(\cdot)$ is constant function then we are done. If not, then there exists two points x_1, x_2 in $\mathbb{R}^+$ such that $M(x_1) > M(x_2)$. Now, consider the distribution F with density f given by

$$f(y) = \begin{cases} x_1, & y \in \left(0, \dfrac{1}{x_1 + x_2}\right) \\ x_2, & y \in \left(\dfrac{1}{x_1 + x_2}, \dfrac{2}{x_1 + x_2}\right) \end{cases}.$$

Also consider the sequence of distributions G_n given by the corresponding sequence of densities g_n defined as

$$g_n(y) = \begin{cases} x_1 - \dfrac{1}{n}, & y \in \left(0, \dfrac{1}{x_1 + x_2}\right) \\ x_2 + \dfrac{1}{n}, & y \in \left(\dfrac{1}{x_1 + x_2}, \dfrac{2}{x_1 + x_2}\right) \end{cases}.$$

Note that $g_n(\cdot)$ is a valid probability density function for all $n > \frac{1}{x_1}$. From the definition of g_n, it follows that

$$\begin{aligned} &n(x_1 + x_2)D(G_n, F) \\ &= \left\{-\frac{B(x_1 - 1/n) - B(x_1)}{-1/n} + B'(x_1) - M(x_1)\right\} \\ &\quad + \left\{\frac{B(x_2 + 1/n) - B(x_2)}{1/n} - B'(x_2) + M(x_2)\right\}. \end{aligned}$$

Now, for each x in $\mathbb{R}^+$,

$$\frac{B(x - 1/n) - B(x)}{-1/n} - B'(x) \to 0$$

and

$$\frac{B(x + 1/n) - B(x)}{1/n} - B'(x) \to 0$$

as n goes to infinity. Hence there exists $n_1 > \frac{1}{x_1}$ such that

$$\begin{aligned} &-\frac{B(x_1 - 1/n_1) - B(x_1)}{-1/n_1} + B'(x_1) \\ &+ \frac{B(x_2 + 1/n_1) - B(x_2)}{1/n_1} - B'(x_2) < M(x_1) - M(x_2), \end{aligned} \tag{13}$$

as $M(x_1) - M(x_2)$ is a strictly positive quantity. This implies $D(G_{n_1}, F) < 0$, a contradiction as by our assumption, $D(G, F)$ is a statistical divergence for all distributions G and F. This concludes the proof of the theorem. □

One point that should be mentioned here is that the strict convexity of B_1 is never used here for the proof itself, but it is required for $D_B(G,F)$, and hence $D(G,F)$, to be valid divergences. In that context, we present our final result.

Theorem 3. *$D_B(G,F)$ based on the function B (which we assume here to be at least first order continuously differentiable) is a valid statistical divergence for all absolutely continuous distributions F and G if and only if B is a strictly convex function.*

Proof. First we present the proof of the 'if' part. This is the easier part of the theorem. One can check that by convexity of $B(\cdot)$ in its argument, the integrand of $D_B(\cdot,\cdot)$ is always nonnegative and equal to zero only when $f(\cdot)$ is identically equal to $g(\cdot)$ outside a set of measure zero under the model. Hence $D_B(G,F)$ is a valid statistical divergence for all absolutely continuous distributions G and F.

Before proving the only if part, it should be pointed out that there exists related results on non-negativity of Bregman divergence between two points; see, e.g., Boyd and Vandenberghe (2004). In the result proved below, we only assume non-negativity of the concerned integral (this allows the integrand to be negative in some cases as well), providing much more generality in that context.

Now we begin proving the 'only if' part. We proceed via contradiction. Suppose that $B'(\cdot)$ is not strictly increasing on some open intervals in $\mathbb{R}^+$. Then by continuity there exists an interval such that $B'(\cdot)$ is constant in that interval or strictly decreasing in that interval. We treat the two cases separately.

First we consider the constant case. Suppose that there exists $x > 0$ and $r > 0$ such that $B'(y)$ is constant for all y in $V = (x-r, x+r)$, with $x-r > 0$. Take x_1, x_2, r_1, r_2 such that $V_1 = (x_1-r_1, x_1+r_1) \subset V$ and $V_2 = (x_2-r_2, x_2+r_2) \subset V$ with $V_1 \cap V_2 = \Phi$. Based on this notation we construct $f(\cdot)$ and $g(\cdot)$ same as before. Then we have

$$\begin{aligned} & n(x_1+x_2)D_B(G_n,F) \\ &= \left\{-\frac{B(x_1-1/n)-B(x_1)}{-1/n} + B'(x_1)\right\} \\ &\quad + \left\{\frac{B(x_2+1/n)-B(x_2)}{1/n} - B'(x_2)\right\}. \end{aligned} \tag{14}$$

Take $n_0 > \max(\frac{1}{r_1}, \frac{1}{r_2})$. Then $D_B(G_{n_0}, F) = 0$ in spite of G_{n_0} and F being two different distributions. This is a contradiction as a valid statistical divergence between two distributions is zero if and only if they are unequal on a measure zero set.

On the contrary let us assume that there exists an interval such that $B'(\cdot)$ is strictly decreasing in that interval. Under the previous notations, assume that $B'(\cdot)$ is strictly decreasing in V. By Lagrange's Mean Value Theorem, there exists ξ_1, ξ_2 such that

$$\frac{B(x_1-1/n_0)-B(x_1)}{-1/n_0} = B'(\xi_1), \quad x_1 - \frac{1}{n_0} < \xi_1 < x_1$$

and

$$\frac{B(x_2+1/n_0)-B(x_2)}{1/n_0} = B'(\xi_2), \quad x_2 < \xi_2 < x_2 + \frac{1}{n_0}.$$

By our assumption of strictly decreasing $B'(\cdot)$ in V we have $B'(\xi_1) > B'(x_1)$ and $B'(\xi_2) < B'(x_2)$. Hence we get

$$\begin{aligned} & n_0(x_1+x_2)D_B(G_{n_0},F) \\ &= \left\{-\frac{B(x_1-1/n_0)-B(x_1)}{-1/n_0} + B'(x_1)\right\} \\ &\quad + \left\{\frac{B(x_2+1/n_0)-B(x_2)}{1/n_0} - B'(x_2)\right\} \\ &= \{-B'(\xi_1)+B'(x_1)\} + \{B'(\xi_2)-B'(x_2)\} \\ &< 0. \end{aligned}$$

This provides the necessary contradiction. Hence on all open intervals in $\mathbb{R}^+$, $B'(\cdot)$ is strictly increasing and hence $B(\cdot)$ is a convex function in its argument. □

Combining Theorem 2 and Theorem 3 we conclude that all single integral, non-kernel divergence measures for absolutely continuous models belong to the class of Bregman divergences. Since every Bregman divergence is a single-integral, non-kernel divergence measure, this indicates that all single-integral minimum divergence estimators that can be computed without any non-parametric smoothing under absolutely continuous models are minimum Bregman divergence estimators, and the reverse relation also holds. This provides a characterization of all single-integral, non-kernel divergence estimators under absolutely continuous models. We conclude the paper with the following remarks.

Remark 5. As the paper demonstrates, all single-integral, non-kernel divergences are Bregman divergences with possibly strictly convex B functions. The objective function therefore turns out to be an i.i.d. average over the observed data points and the corresponding minimum divergence estimator turns out to be an M-estimator. Thus the well established theory and the asymptotic and convergence results about M-estimators automatically apply to these non-kernel minimum divergence estimators. Yet these estimators are not in the classical spirit of M-estimators where the primary ψ functions are of the location-scale type. Since here the model densities are the arguments of the convex generating function B, these estimators make explicit use of the form of the model density in the structure of its ψ function. More research is necessary in the future to get a better idea about the impact of the nature of the B function on the corresponding inference; the known properties of the minimum divergence estimators based on the density power divergence make such explorations eminently sensible.

Remark 6. The use of non parametric methods in order to smooth the data and find a density estimate has a rich literature, but often runs into issues related to convergence. Other than an assumption of a large enough sample size, one has to be really concerned about the rate in which the associated bandwidth decreases to zero and a choice of the kernel. These factors are important for the theoretical justification of convergence to the truth. In the multivariate set up, inefficiency of the kernel density estimate increases sharply with dimension (Silverman, 1986), which results in unreliability of the inference in dealling with small datasets. Even with a large volume of data, use of common multivariate kernels introduces a bias in the density estimate, which necessitates the consideration of kernels with complex moment conditions (Lu et al., 2009), often making asymptotic consistency of the estimator difficult to verify. In addition, the proper choice of the bandwidth relative to the sample size is also important for the robustness of the estimator and the computational complexity of the procedure. In repeated simulations it is the experience of the authors that a relatively large bandwidth can make the kernel density estimate overly smooth and push the solution towards the non-robust maximum likelihood estimator. On the other hand, a very small bandwidth will make the kernel density estimate extremely spiky, leading to slower (often, significantly slower) convergence of the iterative root solving procedure.

Remark 7. The characterization provided in this paper asserts that one needs to focus exclusively on Bregman divergences to avoid the usual drawbacks of non-parametric smoothing in minimum divergence inference. For illustration, consider $\theta_0 \in \Theta$ such that $F_{\theta_0} \in \{F_\theta : \theta \in \Theta\}$ is closest to the data generating distribution G in the Bregman divergence sense. Under the notation $v_\theta(X_i) = -B'(f_\theta(X_i)) + \int \big[-B(f_\theta(x)) + f_\theta(x)B'(f_\theta(x))\big] dx$ and the results in our paper, the minimization of the expression in (5) translates to the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n} v_\theta(X_i)$$

over $\theta \in \Theta$. Then we can invoke Theorem 2 and Theorem 4 from Yuan and Jennrich (1998) to show that with probability one there exists roots $\hat{\theta}_n$ of the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\theta} v_\theta(X_i) = 0$$

such that

$$\hat{\theta}_n \to \theta_0.$$

Furthermore this sequence of roots also exhibits

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\mathcal{L}}{\to} N(0, W)$$

where the last convergence is in distribution for some positive definite matrix W, depending possibly on θ_0. The quantity $(\hat{\theta}_n - \theta_0)$ is therefore a $O_p(n^{-1/2})$ term. It may be noted that the assumptions that need to be verified for the above convergence are standard, and in terms of the defining B function only some smoothness conditions on the derivatives of the B function, of order at most 3, are needed. On the other hand the convergence rates of kernel based estimators involve extra dependency on bandwidths of the smoothing method in addition to similar conditions on B_1, B_2, and B_3. See Tamura and Boos (1986) for a description of the bias term generated in estimation due to the slow convergence of the kernel.

Remark 8. The consistency of the estimators generated in the process depends only on the B function and proper choice of it gives rise to important statistical results (e.g., in case of Kullback-Leibler, L^2, BHHJ, etc.). Moreover the characterization also implies that given any model family of densities, one only needs to focus on the estimators based on the Bregman divergences to get a good understanding of the underlying parameters. The structure of divergence involved here can also be analyzed in the regime of convex optimization which possesses well behaved ways of dealing with such expressions. One should mention that there are examples of divergences that can be constructed with the help of convex functions (for example the class of disparities considered in Basu, Shioya, and Park 2011) but the class considered here enjoys the additional benefit of simplicity of the divergence which makes it easier to analyze the properties of the estimator. This opens up consideration of optimization problems of similar type that might be of statistical significance. However, we do not discuss this in any more detail as that is not the focus of this paper.

Acknowledgments

The authors gratefully thank the Associate Editor and two anonymous referees whose comments led to a substantially improved version of the manuscript.

References

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, **B 28**, 131-142.

Avlogiaris, G., Micheas, A. and Zografos, K. (2016). On local divergences between two probability measures. *Metrika*, **79(3)**, pp.303-333.

Avlogiaris, G., Micheas, A. and Zografos, K. (2016). On testing local hypotheses via local divergence. *Statistical Methodology*, **31**, pp.20-42.

Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of machine learning research*, **6(Oct)**, pp.1705-1749.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85(3)**, 549-559.

Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. CRC Press, Boca Raton, Florida.

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics*, **5**, 445-463.

Boyd, S., and Vandenberghe, L. (2004). Convex optimization. *Cambridge university press*.

Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, **142(9)**, 2574-2585.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, **B 46**, 440-464.

Csiszár, I. (1963). Eine informations theoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, **3**, 85-107.

CSISZÁR, I. (1967). Information-type measures of diverence of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, **2**, 299-318.

DEVROYE, L. AND GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York.

DONOHO, D. L. AND LIU, R. C. (1988). The automatic robustness of minimum distance functionals. *Annals of Statistics*, **16**, 552-586.

FÉVOTTE, C., BERTIN, N. AND DURRIEU, J.L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural computation*, **21(3)**, pp.793-830.

FRIGYIK, B.A., SRIVASTAVA, S. AND GUPTA, M.R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, **54(11)**, pp.5130-5139.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73-101.

ITAKURA, F. AND SAITO, S. (1968). Analysis Synthesis Telephony based on the Maximum Likelihood, *6th ICA*. C-5-5.

JONES, M. C., HJORT, N. L., HARRIS, I. R. BASU, A. (2001). A comparison of related density based minimum divergence estimators. *Biometrika*, **88**, 865-873.

LERASLE, M. (2012). Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **48**, 884-908.

LIESE, F. AND VAJDA, I. (1987). *Convex Statistical Distances*. Teubner.

LINDSAY, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, **22**, 1081-1114.

LU, X., LIAN, H. AND LIU, W. (2012). Semiparametric estimation for inverse density weighted expectations when responses are missing at random. *Journal of Nonparametric Statistics*, **24(1)**, pp.139-152.

MATTHEOU, K., LEE, S. AND KARAGRIGORIOU, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, **139**, 228-235.

MATTHEOU, K. AND KARAGRIGORIOU, A. (2010). A new family of divergence measures for tests of fit. *Australian & New Zealand Journal of Statistics*, **52**, 187-200.

MUKHERJEE, T., MANDAL, A. AND BASU, A. (2018). The B-Exponential Divergence and its Generalizations with Applications to Parametric Inference. *Technical Report, Interdisciplinary Statistical Research Unit, Indian Statistical Institute, India*, ISRU/2018/3.

PARDO, L. (2006). *Statistical Inference based on Divergence Measures*. Chapman & Hall/CRC.

PARDO, M.C. AND VAJDA, I. (2003). On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, **49(7)**, pp.1860-1867.

PEARSON, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157-175.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley, New York.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

SIMPSON, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, **82**, 802-807.

SIMPSON, D. G. (1989). Hellinger deviance test: Efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, **84**, 107-113.

SMITH, P.J., SHAFI, M. AND GAO, H. (1997). Quick simulation: A review of importance sampling techniques in communications systems. *IEEE Journal on Selected Areas in Communications* **15(4)**, 597-613.

STUMMER, W. AND VAJDA, I. (2012). On Bregman distances and divergences of probability measures. *IEEE Trans Inf Theory*, **58(3)**, 1277-1288.

TAMURA, R. N. AND BOOS, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association* **81**, 223-229.

VAJDA, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publishers.

WAND, M. P. AND JONES, M. C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC, Boca Raton, Florida.

WOLFOWITZ, J. (1957). The minimum distance method. *Annals of Mathematical Statistics*, **28**, 7588.

YUAN, K.H. AND JENNRICH, R.I.(1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, **65(2)**, 245-260.

Soham Jana is a PhD candidate in the Department of Statistics and Data Science at Yale University, CT, USA. He received his B.Stat (Hons.) and M.Stat (Specialization: Theoretical Statistics) degrees with distinction in the years 2015 and 2017, respectively, from Indian Statistical Institute, Kolkata, India. He has previously received INSPIRE scholarship from the govt. of India for outstanding academic performance. His research interests involve robust statistical methods, minimum distance inference, small sample minmax optimization, and information theory.

Ayanendranath Basu received his Ph.D. in Statistics from the Pennsylvania State University in 1991, working under the supervision of Professor Bruce G. Lindsay, Distinguished Professor in Statistics (and later Eberly Chair). He was an Assistant Professor at the Department of Mathematics, University of Texas at Austin, from 1991 to 1995. He moved to India in 1995 and has since been with the Indian Statistical Institute, where he currently holds the post of a Professor (Higher Academic Grade). He has more than 100 journal papers, two books, and several edited volumes to his credit. He is a former editor of Sankhya, Series B, and a fellow of the National Academy of Sciences, India and West Bengal Academy of Science and Technology. He is a recipient of the C. R. Rao National Award in Statistics. His research interests include robust statistical methods, minimum distance inference and categorical data analysis.