Unbiased Estimation Equation under f-Separable Bregman Distortion Measures

Masahiro Kobayashi and Kazuho Watanabe

Toyohashi University of Technology

Email: kobayashi@lisl.cs.tut.ac.jp and wkazuho@cs.tut.ac.jp

Abstract—We discuss unbiased estimation equations in a class of objective function using a monotonically increasing function f and Bregman divergence. The choice of the function f gives desirable properties such as robustness against outliers. In order to obtain unbiased estimation equations, analytically intractable integrals are generally required as bias correction terms. In this study, we clarify the combination of Bregman divergence, statistical model, and function f in which the bias correction term vanishes. Focusing on Mahalanobis and Itakura-Saito distances, we provide a generalization of fundamental existing results and characterize a class of distributions of positive reals with a scale parameter, which includes the gamma distribution as a special case. We discuss the possibility of latent bias minimization when the proportion of outliers is large, which is induced by the extinction of the bias correction term.

I. INTRODUCTION

The maximum likelihood estimation (MLE) for the statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$ estimates the parameter $\boldsymbol{\theta}$ by minimizing the negative log-likelihood. It is equivalent to empirical inference under the Kullback-Leibler (KL)-divergence. However, MLE is susceptible to outliers or mismatch of the assumed model. In robust statistics, estimation methods weakening adverse effect of outliers have been studied [1], [2]. One of the most popular methods is M-estimation which changes KL-divergence corresponding to MLE to robust divergences applicable to empirical inference. These divergences are constructed through estimation equations by weighted (negative) score function $s(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, where $l(\boldsymbol{x}, \boldsymbol{\theta}) = -\log p(\boldsymbol{x}|\boldsymbol{\theta})$. The following two types of estimation equations are well known:

$$\frac{1}{n}\sum_{i=1}^{n}\xi(l(\boldsymbol{x}_i, \boldsymbol{\theta}))s(\boldsymbol{x}_i, \boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{x}, \boldsymbol{\theta}))s(\boldsymbol{x}, \boldsymbol{\theta})\right], \quad (1)$$

$$\frac{\sum_{i=1}^{n}\xi(l(\boldsymbol{x}_i, \boldsymbol{\theta}))s(\boldsymbol{x}_i, \boldsymbol{\theta})}{\sum_{j=1}^{n}\xi(l(\boldsymbol{x}_j, \boldsymbol{\theta}))} = \frac{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{x}, \boldsymbol{\theta}))s(\boldsymbol{x}, \boldsymbol{\theta})\right]}{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\xi(l(\boldsymbol{x}, \boldsymbol{\theta}))\right]}, \quad (2)$$

where $\xi : \mathbb{R} \to \mathbb{R}$ works as the weight function. Equation (1) is called the *unnormalized estimation equation* because the summation of weights of score functions is not one. This estimation equation is obtained from minimizing β-divergence (density power divergence), U-divergence, Ψ-divergence and so on [3]–[6]. Equation (2) is called the *normalized estimation equation* because the summation of weights of score functions is one. Windham proposed the estimator using density power weight in (2) [7]. Then Jones et al. constructed corresponding divergence [8]. It was proved that this divergence, named γ-divergence, has the property that the latent bias can be minimized even when the proportion of outliers is large, and that the divergence with such a property is unique under some assumptions [9]. This property of γ-divergence was extended to the normalized estimation equation (2) with general weight ξ [10]. However, these approaches require bias correction terms, that is, the right hand sides of (1) and (2), which in general result in analytically intractable integrals.

In this paper, we consider the M-estimation under f-separable distortion measures, which were proposed to extend linear distortion such as the average distortion to non-linear distortion, and for which the rate-distortion function was studied [11]. It was also applied to the estimation problem with Bregman divergence as the base distortion measure and a simple clustering or vector quantization algorithm was constructed [12]. In this paper, we call this class of objective functions the f-separable Bregman distortion measure. As will be discussed in Section III, the M-estimation under this distortion measure can be viewed as deviance-based estimation of the regular exponential family model. On one hand, unbiasedness of the estimation equation of deviance-based methods has been studied and some sufficient conditions for it have been obtained [13], [14]. However, these results only apply to the case where the data-generating distribution is included in the assumed model. On the other hand, the M-estimation of the location family is proved to have an unbiased estimation equation for general symmetric distributions [2]. It is unknown in what cases of f-separable Bregman distortion measures the estimation equation is unbiased for such a general class of distributions. If an estimation equation is unbiased, it can be regarded as normalized and the estimator has the potential to minimize the latent bias even if the proportion of outliers is large.

In this paper, we study the conditions for bias correction terms of f-separable Bregman distortion measures to vanish and characterize the combination of Bregman divergence, the statistical model, and the function f. Focusing on Mahalanobis and Itakura-Saito (IS) distances, we specify the conditions for the general model classes and the function f to achieve unbiased estimation equations. Furthermore, we discuss if the latent bias can be minimized when the proportion of outlier is large. We compare the M-estimation under the f-separable IS distortion measure with the estimation methods minimizing β and γ divergences in terms of asymptotic efficiency.

II. f-SEPARABLE BREGMAN DISTORTION MEASURES

In this section, we introduce the estimation method based on f-separable Bregman distortion measures [12]. We consider estimating the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$ of a statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$ when given the data $\boldsymbol{x}^n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i = (x_i^{(1)}, \cdots, x_i^{(d)})^{\mathrm{T}} \in \mathbb{R}^d$. We assume that $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$ is the data-generating distribution and the parameter $\boldsymbol{\theta}$ is the expected value of $\boldsymbol{x}$ under the model, that is, $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{X}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}$ if it exists. The objective function (3) is defined by a differentiable and continuous monotonically increasing function $f : \mathbb{R}_+ \to \mathbb{R}$ and Bregman divergence $d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of non-negative real numbers.

$$L_f(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n f\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right) \tag{3}$$

Bregman divergence is defined by a differentiable strictly convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ as

$$d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) \triangleq \phi(\boldsymbol{x}) - \phi(\boldsymbol{\theta}) - \langle \boldsymbol{x} - \boldsymbol{\theta}, \nabla\phi(\boldsymbol{\theta}) \rangle,$$

where $\nabla\phi$ is its gradient vector and $\langle \cdot, \cdot \rangle$ is the inner product. The estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}^*$ is given by the minimum solution of (3) as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L_f(\boldsymbol{\theta}).$$

The corresponding estimation equation is given by

$$\frac{1}{n} \sum_{i=1}^n f'(d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta})) \frac{\partial}{\partial\boldsymbol{\theta}} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \tag{4}$$

where f' is the derivative of f. This is not generally unbiased. The property of the estimator depends on the function f. For example, if the function f is concave, the estimator is robust against outliers.

The original f-separable distortion measures are defined by f-mean with respect to some base distortion d [11]. From the view point of f-mean, representative examples are the log-sum-exp function and power mean, which are given by the following functions:

$$f(z) = \frac{1 - \exp(-\alpha z)}{\alpha}, f'(z) = \exp(-\alpha z) \tag{5}$$

$$f(z) = \frac{(z+a)^\beta - 1}{\beta}, f'(z) = (z+a)^{\beta-1} \quad (a \geq 0), \tag{6}$$

respectively, where if tuning parameters satisfy $\alpha > 0$ or $\beta < 1$, the estimators become robust. When $\alpha = 0$ and $\beta = 1$, (5) and (6) become linear functions.

III. RELATION TO ROBUST DIVERGENCES

First, we show that the minimization of $L_f(\boldsymbol{\theta})$ is derived from deviance-based M-estimation of the expectation parameter under the regular exponential family,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = r_\phi(\boldsymbol{x}) \exp(-d_\phi(\boldsymbol{x}, \boldsymbol{\theta})), \tag{7}$$

where $r_\phi(\boldsymbol{x})$ is uniquely determined by the strictly convex function ϕ [15]. In fact, the deviance function [13] of this model is

$$l(\boldsymbol{x}, \boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}} l(\boldsymbol{x}, \boldsymbol{\theta}) = d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) - \min_{\boldsymbol{\theta}} d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) = d_\phi(\boldsymbol{x}, \boldsymbol{\theta}).$$

Next, we turn to empirical inference based on robust divergences under the regular exponential family (7). The negative score function is given by $s(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} d_\phi(\boldsymbol{x}, \boldsymbol{\theta})$. Suppose for a moment that the bias correction term can be ignored. In this case, the unnormalized estimation equation (1) is given by

$$\frac{1}{n} \sum_{i=1}^n \xi\left(l(\boldsymbol{x}_i, \boldsymbol{\theta})\right) \frac{\partial}{\partial\boldsymbol{\theta}} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{0}. \tag{8}$$

Compared with this estimation equation, the estimation equation (4) of f-separable Bregman distortion measures can be interpreted as a weighted score function. We focus on the arguments of the weight functions of (4) and (8). The only difference is the term $\inf_{\boldsymbol{\theta}} l(\boldsymbol{x}, \boldsymbol{\theta}) = -\log r_\phi(\boldsymbol{x})$. Specifically, if the domain of the function f' is extended to $(-\infty, \infty)$, the function f' works identically to the weight function ξ. In view of this relation, the function (5) associated with the log-sum-exp function yields the estimation methods that minimize β and γ divergences with the unnormalized and normalized estimation equations, (1) and (2), respectively. In other words, when we assume the regular exponential family and the function (5), then it is related to the estimation based on power of the statistical model.

While, in this section, we have assumed the bias correction term is exactly $\boldsymbol{0}$, it does not hold in general. With the combination of the model and Bregman divergence discussed in the next section, the estimation equation (4) becomes unbiased without any bias correction term for any function f satisfying the condition given in the main theorems.

IV. CONDITIONS FOR UNBIASED ESTIMATION EQUATION

In general, the estimator based on f-separable Bregman distortion measures introduced in Section II does not satisfy consistency because its estimation equation is not necessarily unbiased. In order to satisfy an unbiased estimation equation, we must subtract the bias correction term $b_f(\boldsymbol{\theta})$ from the objective function (3) as follows:

$$L_f(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n f\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right) - b_f(\boldsymbol{\theta}),$$

$$b_f(\boldsymbol{\theta}) = -\int \boldsymbol{\nabla\nabla}\phi(\boldsymbol{\theta}) \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)(\boldsymbol{x} - \boldsymbol{\theta})\right] d\boldsymbol{\theta},$$

where $\int \cdot d\boldsymbol{\theta}$ denotes the indefinite integral with respect to $\boldsymbol{\theta}$. Then, the unnormalized estimation equation is given by

$$\frac{1}{n} \sum_{i=1}^n f'\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)(\boldsymbol{x}_i - \boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)(\boldsymbol{x} - \boldsymbol{\theta})].$$

On the other hand, we can consider the normalized estimation equation as follows:

$$\frac{\sum_{i=1}^n f'\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)(\boldsymbol{x}_i - \boldsymbol{\theta})}{\sum_{j=1}^n f'\left(d_\phi\left(\boldsymbol{x}_j, \boldsymbol{\theta}\right)\right)} = \frac{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)(\boldsymbol{x} - \boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)]}.$$

Fujisawa has elucidated that this estimation equation can possibly minimize the latent bias even when the proportion of outliers is large [10]. In both cases, it is necessary to calculate the integral for bias correction for each combination of statistical model, Bregman divergence, and the function f. However, in many cases, the integral may not exist or be analytically intractable. In this paper, we discuss the following estimation equation:

$$\frac{1}{n} \sum_{i=1}^{n} f'\left(d_\phi\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)\left(\boldsymbol{x}_i - \boldsymbol{\theta}\right) = \boldsymbol{0}.$$

That is, the bias correction term does not depend on the parameter $\boldsymbol{\theta}$. In other words, the following equation is satisfied,

$$\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)\left(\boldsymbol{x} - \boldsymbol{\theta}\right)\right] = \boldsymbol{0}. \qquad (9)$$

Then, this estimation equation is automatically normalized. Therefore, the estimator has the possibility to minimize the latent bias even when the proportion of outliers is large. In the rest of this section, we characterize the combination of the statistical model $p(\boldsymbol{x}|\boldsymbol{\theta})$, Bregman divergence $d_\phi(\boldsymbol{x}, \boldsymbol{\theta})$ and the function f where the bias correction term vanishes. Note that the statistical model considered hereafter is generally not the regular exponential family.

In particular, we focus on Mahalanobis and IS distances. In the case of estimating the location parameter of elliptical distribution, it is known that the bias correction term vanishes and the estimator is consistent under certain conditions on the function f [2]. In the case of log-gamma regression model, it is known that the bias correction term vanishes. This is equivalent to the case where IS distance is used and the model is the gamma distribution [14]. In this paper, we derive a simple condition of the function f which induces unbiased estimation equation. In particular, in the case of IS distance, the class of the model is extended to a more general class.

A. Mahalanobis distance

When the strictly convex function is given by $\phi(\boldsymbol{x}) = \boldsymbol{x}^\mathrm{T} \boldsymbol{A} \boldsymbol{x}$, where $\boldsymbol{A}$ is a positive definite matrix. Then the corresponding Bregman divergence is given by

$$d_{\mathrm{Mah.}}(\boldsymbol{x}, \boldsymbol{\theta}) \triangleq (\boldsymbol{x} - \boldsymbol{\theta})^\mathrm{T} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\theta}).$$

If the positive definite matrix $\boldsymbol{A}$ is identity, Mahalanobis distance reduces to squared distance,

$$\|\boldsymbol{x} - \boldsymbol{\theta}\|^2 = \sum_{j=1}^{d} (x^{(j)} - \theta^{(j)})^2.$$

We assume that the statistical model is the elliptical distribution.

Definition 1 (Elliptical distribution [16]): For $\boldsymbol{x} \in \mathbb{R}^d$ and the parameter $\theta \in \Theta = \mathbb{R}^d$ and the function $g : \mathbb{R}_+ \to \mathbb{R}_+$, let $C < \infty$ be the normalization constant, and the positive definite matrix $\boldsymbol{A}$ be the inverse of a fixed covariance matrix. Then the elliptical distribution is defined by the following probability density function,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{C} g((\boldsymbol{x} - \boldsymbol{\theta})^\mathrm{T} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\theta})). \qquad (10)$$

This distribution includes Gaussian, Laplace, t distributions and so on.

Theorem 1: If the following condition holds against the combination of the function f and the statistical model (10), the estimation equation holds without a bias correction term:

$$\int_0^\infty g(t) f'(t) t^{\frac{d-1}{2}} dt < \infty.$$

Although the unbiased estimation equation in this case is intuitively trivial because of the symmetry around $\boldsymbol{\theta}$ and has been pointed out in the literature [2], the explicit condition for the unbiasedness has never been discussed.

B. IS distance

When the strictly convex function is given by $\phi(x) = -\log x$, then the corresponding Bregman divergence is given by

$$d_{\mathrm{IS}}(x, \theta) \triangleq \frac{x}{\theta} - \log \frac{x}{\theta} - 1. \qquad (11)$$

Definition 2 (IS distribution): For $x \in \mathbb{R}_+$ and the scale parameter $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$, and the function $g : \mathbb{R}_+ \to \mathbb{R}_+$, we define the following probability density function with the normalization constant $C < \infty$,

$$p(x|\theta) = \frac{1}{C} \frac{1}{x} g(d_{\mathrm{IS}}(x, \theta)). \qquad (12)$$

When the expectation exists, the scale parameter also coincides with the expectation. In particular, if $g(z) = \exp(-kz)$, the IS distribution reduces to the gamma distribution $p(x|\theta) = \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)} x^{k-1} \exp\left(-\frac{k}{\theta} x\right)$ with the known shape parameter $k > 0$. Details of the IS distribution are described in Appendix D.

Theorem 2: If the following condition holds against the combination of the function f and statistical model (12), the estimation equation holds without a bias correction term:

$$\int_0^\infty g(t) f'(t) dt < \infty \qquad (13)$$

1) Example: Gamma distribution: In the case of the function (5) and gamma distribution with the known shape parameter $k > 0$, that is, $g(z) = \exp(-kz)$, then the integral in (13) becomes as follows:

$$\int_0^\infty \exp(-kz) \exp(-\alpha z) dz = \int_0^\infty \exp(-(k+\alpha)z) dz.$$

Therefore, the condition $\alpha > -k$ must be satisfied for the integral to be bounded. In other words, the lower limit of α that satisfies the unbiased estimation equation differs for each shape parameter k. Since $k > 0$, we can see that the condition of Theorem 2 is satisfied if $\alpha > 0$, for which the estimator is robust against outliers.

In the case of the function (6) and the gamma distribution with the known shape parameter $k > 0$, then the integral in (13) becomes as follows:

$$\int_0^\infty \exp(-kz)(z+a)^{\beta-1} dz.$$

When $a > 0$, the condition of Theorem (13) holds for $\beta < \infty$. When $a = 0$, the condition of Theorem (13) holds for $0 < \beta < \infty$. However, it does not hold for $\beta \leq 0$.

C. Discussion: other Bregman divergence

When the dimension is one, the conditions of Theorems 1 and 2 are the same. A common point is that the statistical model is expressed by Bregman divergence used for estimation. Hence, the results of Theorems 1 and 2 can be generalized to a wider class of continuous distributions written by Bregman divergence. We refer for the details of the continuous Bregman distribution and its theorem to Appendix E.The elliptical and IS distributions are rare examples which have unbiased estimation equations for the corresponding f-separable Bregman distortion measures and include the corresponding regular exponential family models.

V. LATENT BIAS

In this section, we discuss the possibility of the latent bias minimization when the proportion of outliers is large. It is induced by the vanishing bias correction term. From the view point of the normalized estimation equation, the condition of latent bias minimization was shown as a theorem [10], whereas generally its condition is difficult to be examined. However, it can be easily discussed as γ-divergence when the bias correction term vanishes. The definitions of outliers are different for f-separable distortion measures and γ-divergence. We obtain, as a by-product, a solution to a drawback of γ-divergence.

A. Contaminated distribution

We assume that the data-generating distribution is given as follows:

$$\tilde{p}(\boldsymbol{x}) = (1 - \varepsilon)p(\boldsymbol{x}|\boldsymbol{\theta}^*) + \varepsilon c(\boldsymbol{x}),$$

where $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$ is the target distribution and $c(\boldsymbol{x})$ is the contamination distribution which generates outliers and ε is the proportion of outliers. Suppose the parameter $\hat{\boldsymbol{\theta}}$ estimated from the data generated from this distribution is expressed asymptotically as $\tilde{\boldsymbol{\theta}}$. That is, $\hat{\boldsymbol{\theta}} \xrightarrow{P} \tilde{\boldsymbol{\theta}}$. Here, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ is called the latent bias, which expresses the bias caused by the contamination distribution [10].

B. γ-divergence

In the estimation based on γ-divergence, it is assumed that the following quantity can be made arbitrarily small by adjusting $\gamma_0 > 0$ as an assumption regarding outliers,

$$\nu_p = \left[\mathbb{E}_{c(\boldsymbol{x})} \left[p(\boldsymbol{x}|\boldsymbol{\theta}^*)^{\gamma_0} \right] \right]^{\frac{1}{\gamma_0}}. \tag{14}$$

This assumption means that outliers are distributed over the region where the likelihood is small in the target distribution $p(\boldsymbol{x}|\boldsymbol{\theta}^*)$. Since nothing about the outlier proportion is assumed, it is also possible to deal with the case where the outlier proportion is large. Kuchibhotla et al. reported γ-divergence is adversely affected by data at the edge of the support of the target model [17]. For example, in the estimation of the

scale parameter of the exponential distribution, a wrong global solution is generated when very small inlier around $x = 0$ such as $x = 10^{-4}$ is mixed. Recently, a solution to this problem has been invented, whereas it is not fully resolved [17].

C. f-separable Bregman distortion measures

In the estimation based on f-separable Bregman distortion measures, we assume that the following quantity can be made arbitrarily small by adjusting the function f as an assumption regarding outliers,

$$\nu_{d_\phi} = \mathbb{E}_{c(\boldsymbol{x})} \left[f \left(d_\phi \left(\boldsymbol{x}, \boldsymbol{\theta}^* \right) \right) \right], \tag{15}$$

under Assumption 1 described later. This assumption is corresponding to the assumption (14) of γ-divergence and means that when the random variable follows the contamination distribution, that is, $\boldsymbol{X} \sim c(\boldsymbol{X})$, an outlier is in the region where $d_\phi(\boldsymbol{X}, \boldsymbol{\theta}^*) \to \infty$ is satisfied. When estimating the location parameter of the elliptical distribution using Mahalanobis distance, the definition of outlier is same as (14). That is, $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \to \infty$ is regarded as the outlier. However, when estimating the scale parameter of the IS distribution using IS distance, the definition of outlier is not same as (14). In this case, from

$$\lim_{x \to 0} d_{\mathrm{IS}}(x, \theta) = \lim_{x \to \infty} d_{\mathrm{IS}}(x, \theta) = \infty,$$

the data near 0 or ∞ are regarded as outliers. In other words, the estimator based on f-separable IS distortion measures is robust against large outliers and very small inliers to which γ-divergence is vulnerable.

D. Condition of function f

In the following, we identify the function $f(z)$ with $f(z) + \mathrm{constant}$, because the estimator depends only on the derivative of the function f.

Assumption 1: $\forall z \in \mathbb{R}_+$, $|f(z)| < \infty$ and $\lim_{z \to \infty} f(z) = 0$

Assumption 2: Under Assumption 1, (15) can be made arbitrarily small by adjusting the function f.

Assumption 3: When $\epsilon = 0$, the estimator $\tilde{\boldsymbol{\theta}}$ is a consistent estimator, that is, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$.

Theorem 3: Under Assumptions 1-3, the latent bias can be made arbitrarily small by adjusting the function f.

1) Example: We consider the function (5), which is identified with $\exp(-\alpha z)$. Then Assumption 1 holds immediately. Assumption 2 follows from Lyapunov's inequality with sufficiently large α. Assumption 3 depends on the target distribution. In the case of the gamma distribution, we can prove the consistency of the estimator [13].

VI. ASYMPTOTIC PROPERTY

The estimation based on f-separable Bregman distortion measures, which satisfies the unbiasedness of estimation equation, can be interpreted as an M-estimation. Therefore, under appropriate assumptions, the following consistency and

asymptotic normality of the estimator follow from the asymptotic theory of M-estimation [1], [2], [18],

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^*, \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \xrightarrow{d} N(\mathbf{0}, \Sigma(\boldsymbol{\theta}^*)),$$

where $\Sigma(\boldsymbol{\theta}^*) = \boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\boldsymbol{I}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)$,

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[[f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)]^2\left(\boldsymbol{x} - \boldsymbol{\theta}\right)\left(\boldsymbol{x} - \boldsymbol{\theta}\right)^{\mathrm{T}}\right],$$

$$\boldsymbol{J}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}\left[\frac{\partial f'\left(d_\phi\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)\left(\boldsymbol{x} - \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right].$$

If the proportion of outliers is large, the asymptotic variance is given by the technique in [9], [10].

A. Gamma distribution

We assume that the statistical model is the gamma distribution $p(x|\theta) = \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)} x^{k-1} \exp\left(-\frac{k}{\theta}x\right)$, the function f is (5) and Bregman divergence is IS distance (11), then the asymptotic variance of the estimator is given by

$$V[\hat{\theta}] = \Sigma(\theta^*) = \frac{\Gamma(2\alpha + k)\Gamma(k)}{[\Gamma(\alpha + k)]^2}\frac{(\alpha + k)^{2(\alpha+1+k)}}{(2\alpha + k)^{2\alpha+1+k}}\frac{1}{k^{2+k}}\theta^{*2},$$

where $\Gamma(\cdot)$ is the gamma function and the tuning parameter satisfies $\alpha > -0.5k$. In the case of the exponential distribution ($k = 1$), we can compare the asymptotic relative efficiencies (AREs) of the estimators based on minimizing f-separable IS distortion measures and β and γ divergences. The ARE is given by $\frac{V[\hat{\theta}_{\mathrm{MLE}}]}{V[\hat{\theta}]}$, where $V[\hat{\theta}_{\mathrm{MLE}}]$ is the asymptotic variance of the maximum likelihood estimator ($\alpha = 0$). In the case of the exponential distribution, the asymptotic variance of the estimators based on β and γ divergence were derived respectively [3], [8]. Figure 1 shows their AREs, when the tuning parameter $\alpha = \beta = \gamma$. We notice that the range of tuning parameter $\alpha = \beta = \gamma > 0$ induces the robustness against outliers. From Figure 1, for the function (5) and IS distance, the ARE is generally greater than that of β-divergence in the range of tuning parameter $\alpha < 2$. The ARE is also greater than that of γ-divergence in the entire range of the tuning parameter. However, in general, the ARE and robustness have trade-off relationship. Hence, it is important to choose the tuning parameter appropriately taking into account both of them.

VII. Conclusion

In this paper, we discussed the condition for the unbiased estimation equation in the class of parameter estimation by minimizing f-separable Bregman distortion measures. Its condition consists of the statistical model, Bregman divergence and the function f. We clarified in the cases of Mahalanobis and IS distances that the condition the function f and the statistical model should satisfy is characterized by a simple integral. In the parameter estimation of the scale parameter of the gamma distribution, divergence-based estimation generally requires bias correction terms. Furthermore, we proved that the vanishing bias correction term implies the possibility of minimizing latent bias caused by the large proportion of outliers.

Fig. 1. Comparison of asymptotic relative efficiency under the exponential model ($k = 1$).

References

[1] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, 2005.

[2] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, John Wiley & Sons, second edition, 2009.

[3] 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.

[4] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-boost and Bregman divergence," *Neural Computation*, vol. 16, no. 7, pp. 1437–1481, 2004.

[5] S. Eguchi and Y. Kano, "Robustifing maximum likelihood estimation by psi-divergence," ISM Research Memo 802, Institute of Statistical Mathematics, 2001.

[6] T. Mukherjee, A. Mandal, and A. Basu, "The B-exponential divergence and its generalizations with applications to parametric estimation," *Statistical Methods & Applications*, vol. 28, no. 2, pp. 241–257, 2019.

[7] M. P. Windham, "Robustifying model fitting," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 3, pp. 599–609, 1995.

[8] M. C. Jones, N. L. Hjort, I. R. Harris, and A. Basu, "A comparison of related density-based minimum divergence estimators," *Biometrika*, vol. 88, no. 3, pp. 865–873, 2001.

[9] 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.

[10] H. Fujisawa, "Normalized estimating equation for robust parameter estimation," *Electron. J. Statist.*, vol. 7, pp. 1587–1606, 2013.

[11] Y. Shkel and S. Verdú, "A coding theorem for f-separable distortion measures," *Entropy*, vol. 20, no. 2, pp. 1–16, 2018.

[12] M. Kobayashi and K. Watanabe, "Generalized Dirichlet-process-means for f-separable distortion measures," *Neurocomputing*, to appear.

[13] R. V. Lenth and P. J. Green, "Consistency of deviance-based M-estimators," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 49, no. 3, pp. 326–330, 1987.

[14] A. M. Bianco, M. G. Ben, and V. J. Yohai, "Robust estimation for linear regression with asymmetric errors," *Canadian Journal of Statistics*, vol. 33, no. 4, pp. 511–528, 2005.

[15] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[16] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, vol. 11, no. 3, pp. 368–385, 1981.

[17] A. K. Kuchibhotla, S. Mukherjee, and A. Basu, "Statistical inference based on bridge divergences," *Annals of the Institute of Statistical Mathematics*, vol. 71, no. 3, pp. 627–656, 2019.

[18] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

A. Proof of Theorem 1

We assume that eigenvalue decomposition with respect to positive definite matrix $\boldsymbol{A}$. That is, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{V}^{-1} = \boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix with eigenvalues. Then, Mahalanobis distance is rewritten by

$$
(\boldsymbol{x} - \boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\theta})
$$
$$
= (\boldsymbol{x} - \boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{\theta})
$$
$$
= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{j=1}^{d} \lambda_j y_j^2,
$$

where $\boldsymbol{y} = \boldsymbol{V}^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{\theta})$ and λ_j is the j-th element of the diagonal matrix $\boldsymbol{\Lambda}$. Further, we assume that rank factorization with respect to positive definite matrix $\boldsymbol{\Lambda}$. That is, $\boldsymbol{\Lambda} = \sqrt{\boldsymbol{\Lambda}}^{\mathrm{T}} \sqrt{\boldsymbol{\Lambda}}$. If the random vector $\boldsymbol{Y}$ follow $\boldsymbol{Y} \sim \frac{1}{C}g(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y})$, then it can be decomposed as $\boldsymbol{Y} = R\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}$, where random variable R satisfies $R \geq 0$ and d-dimensional random vector $\boldsymbol{U}$ is uniformly distributed on the unit sphere surface [16]. Then, $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$ holds. From (9), ignoring the normalization constant C, we have

$$
\int_{\mathbb{R}^d} g\left(d_{\mathrm{Mah.}}(\boldsymbol{x}, \boldsymbol{\theta})\right) f'\left(d_{\mathrm{Mah.}}(\boldsymbol{x}, \boldsymbol{\theta})\right) (\boldsymbol{x} - \boldsymbol{\theta}) d\boldsymbol{x}
$$
$$
= \int_{\mathbb{R}^d} g\left(\sum_{j=1}^{d}\lambda_j y_j^2\right) f'\left(\sum_{j=1}^{d}\lambda_j y_j^2\right) \boldsymbol{V}\boldsymbol{y}|\boldsymbol{V}|d\boldsymbol{y}
$$
$$
= |\boldsymbol{V}|\boldsymbol{V}\left[\prod_{j=1}^{d}\frac{1}{\sqrt{\lambda_j}}\right]\underbrace{\mathbb{E}[\boldsymbol{U}]}_{\boldsymbol{0}}\int_0^{\infty} g(r^2)f'(r^2)r^d dr
$$
$$
= \boldsymbol{0}.
$$

Therefore, if the following integral exists, then the unbiased estimation equation holds without any bias correction term

$$
\int_0^{\infty} g(r^2)f'(r^2)r^d dr
$$
$$
= \int_0^{\infty} g(t)f'(t)t^{\frac{d-1}{2}} dt,
$$

where we used integration by substitution as $t = r^2$. □

B. Proof of Theorem 2

From (9), ignoring the normalization constant C, we have

$$
\int_0^{\infty} \frac{1}{x}g(d_{\mathrm{IS}}(x, \theta))f'(d_{\mathrm{IS}}(x, \theta))(x - \theta)dx
$$
$$
= \int_0^{\theta} \frac{1}{x}g(d_{\mathrm{IS}}(x, \theta))f'(d_{\mathrm{IS}}(x, \theta))(x - \theta)dx
$$
$$
+ \int_{\theta}^{\infty} \frac{1}{x}g(d_{\mathrm{IS}}(x, \theta))f'(d_{\mathrm{IS}}(x, \theta))(x - \theta)dx
$$
$$
= \theta\int_{\infty}^{0} g(t)f'(t)dt + \theta\int_0^{\infty} g(t)f'(t)dt = 0.
$$

We used integration by substitution as $t = d_{\mathrm{IS}}(x, \theta)$. Therefore, if the following integral exists, then the unbiased estimation equation holds without any bias correction term

$$
\int_0^{\infty} g(t)f'(t)dt < \infty.
$$

□

C. Proof of Theorem 3

We take the expectation of the objective function (3) by $\tilde{p}(\boldsymbol{x}) = (1 - \varepsilon)p(\boldsymbol{x}|\boldsymbol{\theta}^*) + \varepsilon c(\boldsymbol{x})$ as $\int \tilde{p}(\boldsymbol{x})f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x}$. We have

$$
\int \tilde{p}(\boldsymbol{x})f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x}
$$
$$
= (1 - \varepsilon)\int p(\boldsymbol{x}|\boldsymbol{\theta}^*)f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x} + \varepsilon\int c(\boldsymbol{x})f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x}
$$
$$
= (1 - \varepsilon)\int p(\boldsymbol{x}|\boldsymbol{\theta}^*)f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x} + O(\varepsilon\nu_{d_{\phi}}).
$$

Here, we consider $\varepsilon\nu_{d_{\phi}} = \int c(\boldsymbol{x})f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x}$ as $\boldsymbol{\theta} \approx \boldsymbol{\theta}^*$. From Assumptions 1, 2, we can ignore $O(\varepsilon\nu_{d_{\phi}})$,

$$
\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \int \tilde{p}(\boldsymbol{x})f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x}
$$
$$
= \arg\min_{\boldsymbol{\theta}} \int p(\boldsymbol{x}|\boldsymbol{\theta}^*)f\left(d_{\phi}(\boldsymbol{x}, \boldsymbol{\theta})\right)d\boldsymbol{x} = \boldsymbol{\theta}^*,
$$

where we have used Assumption 3. Therefore, the latent bias can be made sufficiently small, that is, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx \boldsymbol{0}$ by adjusting the function f. □

D. Detail of IS distribution

In Section IV-B, we defined a new distribution and named the IS distribution because it is characterized by the IS distance. In this appendix, we explain properties of IS distribution. IS distribution is defined by

$$
p(x|\theta) = \frac{1}{C}\frac{1}{x}g(d_{\mathrm{IS}}(x, \theta)),
$$

where C is the normalization constant. The normalization constant C is expressed as follows without depending on θ,

$$
C = \int_0^{\infty} \frac{1}{x}g(d_{\mathrm{IS}}(x, \theta))dx = \int_0^{\infty} \frac{1}{t}g(d_{\mathrm{IS}}(t, 1))dt.
$$

We used integration by substitution $t = x/\theta$. Specifically, if the expected value exists, $\mathbb{E}[X] < \infty$, $\mathbb{E}[X] = \theta$ holds from the estimation equation (9) with $f(z) = z$. The condition for the unbiased estimation equation is given by (13). This condition with $f(z) = z$ reduces to $\int_0^{\infty} g(t)dt < \infty$, that is, $g \in L^1(\mathbb{R}_+)$. Thus, the following relation holds with respect to the expectation and the function g,

$$
g \in L^1(\mathbb{R}_+) \Leftrightarrow \mathbb{E}[X] = \theta. \tag{16}
$$

In other words, the existence of the expectation depends only on the function g. This property holds in the general

continuous Bregman distribution described later. Then, the normalization constant is expressed as

$$C = \int_0^\infty g(d_{\mathrm{IS}}(x,1))dx. \tag{17}$$

Because we have

$$\theta = \mathbb{E}[X] = \int_0^\infty \frac{1}{C}\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))x\,dx$$
$$= \frac{1}{C}\int_0^\infty g(d_{\mathrm{IS}}(x,\theta))dx = \theta\frac{1}{C}\int_0^\infty g(d_{\mathrm{IS}}(x,1))dx,$$

the normalization constant C must satisfy (17).

1) Example: Gamma distribution: When we choose the function $g(z) = \exp(-kz)$, IS distribution becomes the gamma distribution with the known shape parameter $k > 0$. Then, $\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta))$ is expressed as

$$\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta)) = \frac{1}{x}\exp\left(-kd_{\mathrm{IS}}(x,\theta)\right)$$
$$= \frac{1}{x}\exp(-\frac{k}{\theta}x)\left(\frac{e}{\theta}\right)^k x^k = \left(\frac{e}{\theta}\right)^k x^{k-1}\exp\left(-\frac{k}{\theta}x\right).$$

The normalization constant C is given by

$$C = \int_0^\infty \frac{1}{x}\exp\left(-kd_{\mathrm{IS}}(x,\theta)\right)dx$$
$$= \left(\frac{e}{\theta}\right)^k \int_0^\infty x^{k-1}\exp\left(-\frac{k}{\theta}x\right)dx$$
$$= \left(\frac{e}{\theta}\right)^k \left(\frac{\theta}{k}\right)^k \Gamma(k)$$
$$= \left(\frac{e}{k}\right)\Gamma(k),$$

where $\Gamma(\cdot)$ is the gamma function. Therefore, the gamma distribution is obtained

$$p(x|\theta) = \frac{1}{C}\frac{1}{x}\exp(-kd_{\mathrm{IS}}(x,\theta))$$
$$= \left(\frac{k}{e}\right)^k \frac{1}{\Gamma(k)}\left(\frac{e}{\theta}\right)^k x^{k-1}\exp\left(-\frac{k}{\theta}x\right) \tag{18}$$
$$= \left(\frac{k}{\theta}\right)^k \frac{1}{\Gamma(k)}x^{k-1}\exp\left(-\frac{k}{\theta}x\right).$$

The gamma distribution is also expressed as

$$p(x|\beta,k) = \frac{x^{k-1}}{\Gamma(k)\beta^k}\exp\left(-\frac{x}{\beta}\right).$$

The parameters β and k are called scale and shape parameters, respectively. This model is corresponding to (18) by the transformation $\theta = k\beta$. Notice that the parameter θ is also the scale parameter and the expectation parameter.

E. Detail of Continuous Bregman distribution

Definition 3 (Continuous Bregman distribution): For $x \in (a,b) \subseteq \mathbb{R}$, the parameter $\theta \in \Theta \subseteq \mathbb{R}$, and the function $g : \mathbb{R}_+ \to \mathbb{R}_+$, we define the following probability density function with the normalization constant satisfying $C(\theta) < \infty$,

$$p(x|\theta) = \frac{1}{C(\theta)}\frac{\phi'(x) - \phi'(\theta)}{x - \theta}g(d_\phi(x,\theta)). \tag{19}$$

Specifically, if (20) holds and the expected value exists, $\mathbb{E}[X] < \infty$, $\mathbb{E}[X] = \theta$ holds from the estimation equation (9) with $f(z) = z$ and the condition for it is given by (21). For the same reason, the relationship (16) holds for the expectation and the function g as in the case of the IS distribution. Note that the existence of the expectation depends only on the function g regardless of the choice of Bregman divergence as long as the normalization constant $C(\theta)$ exists and (20) holds. Note that in general, the normalization constant $C(\theta)$ depends on the parameter θ.

Assumption 4:
1) Bregman divergence satisfies the following for any θ and a positive constant ζ (including ∞):

$$\lim_{x \to a} d_\phi(x,\theta) = \lim_{x \to b} d_\phi(x,\theta) = \zeta. \tag{20}$$

2) Bregman divergence used for estimation is corresponding to that of the model (19).

Under these assumptions, the unbiased estimation equation (9) holds.

Theorem 4: If the following condition holds against the combination of the function f and statistical model (19), the estimation equation holds without a bias correction term:

$$\int_0^\infty g(t)f'(t)dt < \infty. \tag{21}$$

Proof 1: From (9), ignoring the normalization constant $C(\theta)$, we have

$$\int_\mathbb{R} \frac{\phi'(x) - \phi'(\theta)}{x - \theta}g(d_\phi(x,\theta))f'(d_\phi(x,\theta))(x - \theta)dx$$
$$= \int_a^\theta (\phi'(x) - \phi'(\theta))g(d_\phi(x,\theta))f'(d_\phi(x,\theta))dx$$
$$+ \int_\theta^b (\phi'(x) - \phi'(\theta))g(d_\phi(x,\theta))f'(d_\phi(x,\theta))dx$$
$$= \int_\zeta^0 g(t)f'(t)dt + \int_0^\zeta g(t)f'(t)dt = 0.$$

We used integration by substitution as $t = d_\phi(x,\theta)$ and (20). Therefore, if integral (21) exists, then the unbiased estimation equation holds without any bias correction term. $\square$

The following models are the examples of the continuous Bregman distribution.

1) symmetric (one dimensional elliptical) distribution: We set $\phi(x) = x^2$. Then, (19) becomes the symmetric (one dimensional elliptical) distribution as follows:

$$p(x|\theta) = \frac{1}{C}g((x - \theta)^2).$$

2) IS distribution: We set $\phi(x) = -\log x$. Then, (19) becomes the IS distribution as follows:

$$p(x|\theta) = \frac{1}{C}\frac{1}{x}g(d_{\mathrm{IS}}(x,\theta)).$$

Finally, we consider the relation between the continuous Bregman distribution (19) and the regular exponential family (7). Let $g(z) = \exp(-z)$. If the factor

$$\frac{1}{C(\theta)}\frac{\phi'(x) - \phi'(\theta)}{x - \theta}$$

does not depend on the parameter θ, (19) becomes one dimensional regular exponential family as follows:

$$p(x|\theta) = r_\phi(x) \exp(-d_\phi(x, \theta)),$$

where $r_\phi(x)$ is uniquely determined by the strictly convex function ϕ [15]. The Gaussian and gamma distributions provide rare examples included in both the class of continuous Bregman distributions and the regular exponential family.