# A New Family of Bounded Divergence Measures and Application to Signal Detection

Shivakumar Jolad[1], Ahmed Roman[2], Mahesh C. Shastry[3], Mihir Gadgil[4] and Ayanendranath Basu[5]
[1]*Department of Physics, Indian Institute of Technology Gandhinagar, Ahmedabad, Gujarat, India*
[2]*Department of Mathematics, Virginia Tech , Blacksburg, VA, U.S.A.*
[3]*Department of Physics, Indian Institute of Science Education and Research Bhopal, Bhopal, Madhya Pradesh, India*
[4]*Biomedical Engineering Department, Oregon Health & Science University, Portland, OR, U.S.A.*
[5]*Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, West Bengal-700108, India*

Keywords: Divergence Measures, Bhattacharyya Distance, Error Probability, F-divergence, Pattern Recognition, Signal Detection, Signal Classification.

Abstract: We introduce a new one-parameter family of divergence measures, called bounded Bhattacharyya distance (BBD) measures, for quantifying the dissimilarity between probability distributions. These measures are bounded, symmetric and positive semi-definite and do not require absolute continuity. In the asymptotic limit, BBD measure approaches the squared Hellinger distance. A generalized BBD measure for multiple distributions is also introduced. We prove an extension of a theorem of Bradt and Karlin for BBD relating Bayes error probability and divergence ranking. We show that BBD belongs to the class of generalized Csiszar f-divergence and derive some properties such as curvature and relation to Fisher Information. For distributions with vector valued parameters, the curvature matrix is related to the Fisher-Rao metric. We derive certain inequalities between BBD and well known measures such as Hellinger and Jensen-Shannon divergence. We also derive bounds on the Bayesian error probability. We give an application of these measures to the problem of signal detection where we compare two monochromatic signals buried in white noise and differing in frequency and amplitude.

## 1 INTRODUCTION

Divergence measures for the distance between two probability distributions are a statistical approach to comparing data and have been extensively studied in the last six decades (Kullback and Leibler, 1951; Ali and Silvey, 1966; Kapur, 1984; Kullback, 1968; Kumar et al., 1986). These measures are widely used in varied fields such as pattern recognition (Basseville, 1989; Ben-Bassat, 1978; Choi and Lee, 2003), speech recognition (Qiao and Minematsu, 2010; Lee, 1991), signal detection (Kailath, 1967; Kadota and Shepp, 1967; Poor, 1994), Bayesian model validation (Tumer and Ghosh, 1996) and quantum information theory (Nielsen and Chuang, 2000; Lamberti et al., 2008). Distance measures try to achieve two main objectives (which are not mutually exclusive): to assess (1) how "close" two distributions are compared to others and (2) how "easy" it is to distinguish between one pair than the other (Ali and Silvey, 1966).

There is a plethora of distance measures available to assess the convergence (or divergence) of probability distributions. Many of these measures are not metrics in the strict mathematical sense, as they may not satisfy either the symmetry of arguments or the triangle inequality. In applications, the choice of the measure depends on the interpretation of the metric in terms of the problem considered, its analytical properties and ease of computation (Gibbs and Su, 2002). One of the most well-known and widely used divergence measures, the Kullback-Leibler divergence (KLD)(Kullback and Leibler, 1951; Kullback, 1968), can create problems in specific applications. Specifically, it is unbounded above and requires that the distributions be *absolutely continuous* with respect to each other. Various other information theoretic measures have been introduced keeping in view ease of computation ease and utility in problems of signal selection and pattern recognition. Of these measures, Bhattacharyya distance (Bhattacharyya, 1946; Kailath, 1967; Nielsen and Boltz, 2011) and Chernoff distance (Chernoff, 1952; Basseville, 1989; Nielsen



Jolad, S., Roman, A., Shastry, M., Gadgil, M. and Basu, A.
A New Family of Bounded Divergence Measures and Application to Signal Detection.
DOI: 10.5220/0005695200720083
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 72-83
ISBN: 978-989-758-173-1

and Boltz, 2011) have been widely used in signal processing. However, these measures are again unbounded from above. Many bounded divergence measures such as Variational, Hellinger distance (Basseville, 1989; DasGupta, 2011) and Jensen-Shannon metric (Burbea and Rao, 1982; Rao, 1982b; Lin, 1991) have been studied extensively. Utility of these measures vary depending on properties such as tightness of bounds on error probabilities, information theoretic interpretations, and the ability to generalize to multiple probability distributions.

Here we introduce a new one-parameter ($\alpha$) family of bounded measures based on the Bhattacharyya coefficient, called bounded Bhattacharyya distance (BBD) measures. These measures are symmetric, positive-definite and bounded between 0 and 1. In the asymptotic limit ($\alpha \to \pm\infty$) they approach squared Hellinger divergence (Hellinger, 1909; Kakutani, 1948). Following Rao (Rao, 1982b) and Lin (Lin, 1991), a generalized BBD is introduced to capture the divergence (or convergence) between multiple distributions. We show that BBD measures belong to the generalized class of f-divergences and inherit useful properties such as curvature and its relation to Fisher Information. Bayesian inference is useful in problems where a decision has to be made on classifying an observation into one of the possible array of states, whose prior probabilities are known (Hellman and Raviv, 1970; Varshney and Varshney, 2008). Divergence measures are useful in estimating the error in such classification (Ben-Bassat, 1978; Kailath, 1967; Varshney, 2011). We prove an extension of the Bradt Karlin theorem for BBD, which proves the existence of prior probabilities relating Bayes error probabilities with ranking based on divergence measure. Bounds on the error probabilities $P_e$ can be calculated through BBD measures using certain inequalities between Bhattacharyya coefficient and $P_e$. We derive two inequalities for a special case of BBD ($\alpha = 2$) with Hellinger and Jensen-Shannon divergences. Our bounded measure with $\alpha = 2$ has been used by Sunmola (Sunmola, 2013) to calculate distance between Dirichlet distributions in the context of Markov decision process. We illustrate the applicability of BBD measures by focusing on signal detection problem that comes up in areas such as gravitational wave detection (Finn, 1992). Here we consider discriminating two monochromatic signals, differing in frequency or amplitude, and corrupted with additive white noise. We compare the Fisher Information of the BBD measures with that of KLD and Hellinger distance for these random processes, and highlight the regions where FI is insensitive large parameter deviations. We also characterize the performance of BBD for different signal to noise ratios, providing thresholds for signal separation.

Our paper is organized as follows: Section I is the current introduction. In Section II, we recall the well known Kullback-Leibler and Bhattacharyya divergence measures, and then introduce our bounded Bhattacharyya distance measures. We discuss some special cases of BBD, in particular Hellinger distance. We also introduce the generalized BBD for multiple distributions. In Section III, we show the positive semi-definiteness of BBD measure, applicability of the Bradt Karl theorem and prove that BBD belongs to generalized f-divergence class. We also derive the relation between curvature and Fisher Information, discuss the curvature metric and prove some inequalities with other measures such as Hellinger and Jensen Shannon divergence for a special case of BBD. In Section IV, we move on to discuss application to signal detection problem. Here we first briefly describe basic formulation of the problem, and then move on computing distance between random processes and comparing BBD measure with Fisher Information and KLD. In the Appendix we provide the expressions for BBD measures , with $\alpha = 2$, for some commonly used distributions. We conclude the paper with summary and outlook.

# 2 DIVERGENCE MEASURES

In the following subsection we consider a measurable space $\Omega$ with $\sigma$-algebra $\mathcal{B}$ and the set of all probability measures $\mathcal{M}$ on $(\Omega, \mathcal{B})$. Let $P$ and $Q$ denote probability measures on $(\Omega, \mathcal{B})$ with $p$ and $q$ denoting their densities with respect to a common measure $\lambda$. We recall the definition of absolute continuity (Royden, 1986):

**Absolute Continuity:** A measure $P$ on the Borel subsets of the real line is absolutely continuous with respect to Lebesgue measure $Q$, if $P(A) = 0$, for every Borel subset $A \in \mathcal{B}$ for which $Q(A) = 0$, and is denoted by $P << Q$.

## 2.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence (KLD) (or relative entropy) (Kullback and Leibler, 1951; Kullback, 1968) between two distributions $P, Q$ with densities $p$ and $q$ is given by:

$$I(P,Q) \equiv \int p \log\left(\frac{p}{q}\right) d\lambda. \tag{1}$$

The symmetrized version is given by

$$J(P,Q) \equiv (I(P,Q) + I(Q,P))/2$$

(Kailath, 1967), $I(P,Q) \in [0,\infty]$. It diverges if $\exists \quad x_0 : q(x_0) = 0$ and $p(x_0) \neq 0$.

KLD is defined only when $P$ is absolutely continuous w.r.t. $Q$. This feature can be problematic in numerical computations when the measured distribution has zero values.

## 2.2 Bhattacharyya Distance

Bhattacharyya distance is a widely used measure in signal selection and pattern recognition (Kailath, 1967). It is defined as:

$$B(P,Q) \equiv -\ln\left(\int \sqrt{pq}d\lambda\right) = -\ln(\rho), \quad (2)$$

where the term in parenthesis $\rho(P,Q) \equiv \int \sqrt{pq}d\lambda$ is called Bhattacharyya coefficient (Bhattacharyya, 1943; Bhattacharyya, 1946) in pattern recognition, affinity in theoretical statistics, and fidelity in quantum information theory. Unlike in the case of KLD, the Bhattacharyya distance avoids the requirement of absolute continuity. It is a special case of Chernoff distance

$$C_\alpha(P,Q) \equiv -\ln\left(\int p^\alpha(x)q^{1-\alpha}(x)dx\right),$$

with $\alpha = 1/2$. For discrete probability distributions, $\rho \in [0,1]$ is interpreted as a scalar product of the probability vectors $\mathbf{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n})$ and $\mathbf{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_n})$. Bhattacharyya distance is symmetric, positive-semidefinite, and unbounded ($0 \leq B \leq \infty$). It is finite as long as there exists some region $S \subset X$ such that whenever $x \in S : p(x)q(x) \neq 0$.

## 2.3 Bounded Bhattacharyya Distance Measures

In many applications, in addition to the desirable properties of the Bhattacharyya distance, boundedness is required. We propose a new family of bounded measure of Bhattacharyya distance as below,

$$B_{\psi,b}(P,Q) \equiv -\log_b(\psi(\rho)) \quad (3)$$

where, $\rho = \rho(P,Q)$ is the Bhattacharyya coefficient, $\psi_b(\rho)$ satisfies $\psi(0) = b^{-1}$ , $\psi(1) = 1$. In particular we choose the following form :

$$\begin{aligned} \psi(\rho) &= \left[1 - \frac{(1-\rho)}{\alpha}\right]^\alpha \\ b &= \left(\frac{\alpha}{\alpha-1}\right)^\alpha, \end{aligned} \quad (4)$$

where $\alpha \in [-\infty, 0) \cup (1, \infty]$. This gives the measure

$$B_\alpha(\rho(P,Q)) \equiv -\log_{\left(1-\frac{1}{\alpha}\right)^{-\alpha}}\left[1 - \frac{(1-\rho)}{\alpha}\right]^\alpha, \quad (5)$$

which can be simplified to

$$B_\alpha(\rho) = \frac{\log\left[1 - \frac{(1-\rho)}{\alpha}\right]}{\log\left[1 - \frac{1}{\alpha}\right]}. \quad (6)$$

It is easy to see that $B_\alpha(0) = 1$, $B_\alpha(1) = 0$.

## 2.4 Special Cases

1. For $\alpha = 2$ we get,
$$B_2(\rho) = -\log_{2^2}\left[\frac{1+\rho}{2}\right]^2 = -\log_2\left(\frac{1+\rho}{2}\right). \quad (7)$$
We study some of its special properties in Sec.3.7.
2. $\alpha \to \infty$
$$B_\infty(\rho) = -\log_e e^{-(1-\rho)} = 1 - \rho = H^2(\rho), \quad (8)$$
where $H(\rho)$ is the Hellinger distance (Basseville, 1989; Kailath, 1967; Hellinger, 1909; Kakutani, 1948)
$$H(\rho) \equiv \sqrt{1 - \rho(P,Q)}. \quad (9)$$
3. $\alpha = -1$
$$B_{-1}(\rho) = -\log_2\left(\frac{1}{2-\rho}\right). \quad (10)$$
4. $\alpha \to -\infty$
$$B_{-\infty}(\rho) = \log_e e^{(1-\rho)} = 1 - \rho = H^2(\rho). \quad (11)$$

We note that BBD measures approach squared Hellinger distance when $\alpha \to \pm\infty$. In general, they are convex (concave) when $\alpha > 1$ ($\alpha < 0$) in $\rho$, as seen by evaluating second derivative

$$\begin{aligned} \frac{\partial^2 B_\alpha(\rho)}{\partial \rho^2} &= \frac{-1}{\alpha^2 \log\left(1-\frac{1}{\alpha}\right)\left(1-\frac{1-\rho}{\alpha}\right)^2} = \\ &= \begin{cases} > 0 & \alpha > 1 \\ < 0 & \alpha < 0 \end{cases}. \end{aligned} \quad (12)$$

From this we deduce $B_{\alpha>1}(\rho) \leq H^2(\rho) \leq B_{\alpha<0}(\rho)$ for $\rho \in [0,1]$. A comparison between Hellinger and BBD measures for different values of $\alpha$ are shown in Fig. 1.

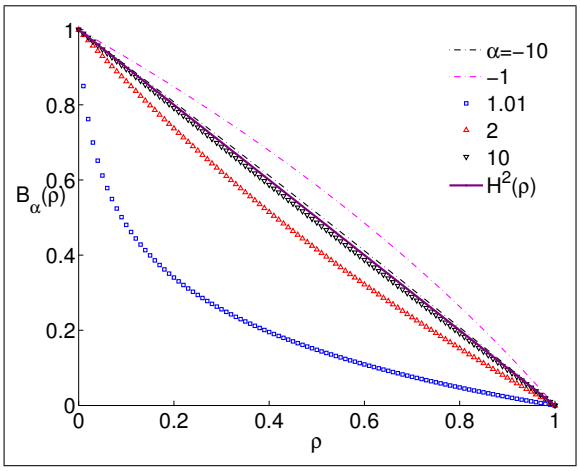


Figure 1: [Color Online] Comparison of Hellinger and bounded Bhattacharyya distance measures for different values of $\alpha$.

## 2.5 Generalized BBD Measure

In decision problems involving more than two random variables, it is very useful to have divergence measures involving more than two distributions (Lin, 1991; Rao, 1982a; Rao, 1982b). We use the generalized geometric mean ($G$) concept to define bounded Bhattacharyya measure for more than two distributions. The $G_\beta(\{p_i\})$ of $n$ variables $p_1, p_2, \ldots, p_n$ with weights $\beta_1, \beta_2, \ldots, \beta_n$, such that $\beta_i \geq 0$, $\sum_i \beta_i = 1$, is given by

$$G_\beta(\{p_i\}) = \prod_{i=1}^{n} p_i^{\beta_i}.$$

For $n$ probability distributions $P_1, P_2, \ldots, P_n$ , with densities $p_1, p_2, \ldots, p_n$ , we define a generalized Bhattacharyya coefficient, also called Matusita measure of affinity (Matusita, 1967; Toussaint, 1974):

$$\rho_\beta(P_1, P_2, \ldots, P_n) = \int_\Omega \prod_{i=1}^{n} p_i^{\beta_i} d\lambda. \qquad (13)$$

where $\beta_i \geq 0$, $\sum_i \beta_i = 1$. Based on this, we define the generalized bounded Bhattacharyya measures as:

$$B_\alpha^\beta(\rho_\beta(P_1, P_2, \ldots, P_n)) \equiv \frac{\log(1 - \frac{1-\rho_\beta}{\alpha})}{\log(1 - 1/\alpha)} \qquad (14)$$

where $\alpha \in [-\infty, 0) \cup (1, \infty]$. For brevity we denote it as $B_\alpha^\beta(\rho)$. Note that, $0 \leq \rho_\beta \leq 1$ and $0 \leq B_\alpha^\beta(\rho) \leq 1$, since the weighted geometric mean is maximized when all the $p_i$'s are the same, and minimized when any two of the probability densities $p_i$'s are perpendicular to each other.

# 3 PROPERTIES

## 3.1 Symmetry, Boundedness and Positive Semi-definiteness

**Theorem 3.1.** *$B_\alpha(\rho(P,Q))$ is symmetric, positive semi-definite and bounded in the interval $[0,1]$ for $\alpha \in [-\infty, 0) \cup (1, \infty]$.*

*Proof.* Symmetry: Since $\rho(P,Q) = \rho(Q,P)$, it follows that

$$B_\alpha(\rho(P,Q)) = B_\alpha(\rho(Q,P)).$$

Positive-semidefinite and boundedness: Since $B_\alpha(0) = 1$, $B_\alpha(1) = 0$ and

$$\frac{\partial B_\alpha(\rho)}{\partial \rho} = \frac{1}{\alpha \log(1 - 1/\alpha)\,[1 - (1-\rho)/\alpha]} < 0$$

for $0 \leq \rho \leq 1$ and $\alpha \in [-\infty, 0) \cup (1, \infty]$, it follows that

$$0 \leq B_\alpha(\rho) \leq 1. \qquad (15)$$

□

## 3.2 Error Probability and Divergence Ranking

Here we recap the definition of error probability and prove the applicability of Bradt and Karlin (Bradt and Karlin, 1956) theorem to BBD measure.

**Error Probability:** The optimal Bayes error probabilities (see eg: (Ben-Bassat, 1978; Hellman and Raviv, 1970; Toussaint, 1978)) for classifying two events $P_1, P_2$ with densities $p_1(x)$ and $p_2(x)$ with prior probabilities $\Gamma = \{\pi_1, \pi_2\}$ is given by

$$P_e = \int \min[\pi_1 p_1(x), \pi_2 p_2(x)] dx. \qquad (16)$$

**Error Comparison:** Let $p_i^\beta(x)$ $(i = 1, 2)$ be parameterized by $\beta$ (Eg: in case of Normal distribution $\beta = \{\mu_1, \sigma_1; \mu_2, \sigma_2\}$ ). In signal detection literature, a signal set $\beta$ is considered better than set $\beta'$ for the densities $p_i(x)$ , when the error probability is less for $\beta$ than for $\beta'$ (i.e. $P_e(\beta) < P_e(\beta')$) (Kailath, 1967).

**Divergence Ranking:** We can also rank the parameters by means of some divergence $D$. The signal set $\beta$ is better (in the divergence sense) than $\beta'$, if $D_\beta(P_1, P_2) > D_{\beta'}(P_1, P_2)$.

In general it is *not* true that $D_\beta(P_1, P_2) > D_{\beta'}(P_1, P_2) \implies P_e(\beta) < P_e(\beta')$. Bradt and Karlin proved the following theorem relating error probabilities and divergence ranking for *symmetric* Kullback Leibler divergence **J**:

**Theorem 3.2 (Bradt and Karlin (Bradt and Karlin, 1956)).** *If $J_\beta(P_1,P_2) > J_{\beta'}(P_1,P_2)$, then $\exists$ a set of prior probabilities $\Gamma = \{\pi_1,\pi_2\}$ for two hypothesis $g_1,g_2$, for which*

$$P_e(\beta,\Gamma) < P_e(\beta',\Gamma) \tag{17}$$

*where $P_e(\beta,\Gamma)$ is the error probability with parameter $\beta$ and prior probability $\Gamma$.*

It is clear that the theorem asserts existence, but no method of finding these prior probabilities. Kailath (Kailath, 1967) proved the applicability of the Bradt Karlin Theorem for Bhattacharyya distance measure. We follow the same route and show that the $B_\alpha(\rho)$ measure satisfies a similar property using the following theorem by Blackwell.

**Theorem 3.3 (Blackwell (Blackwell, 1951)).** *$P_e(\beta',\Gamma) \leq P_e(\beta,\Gamma)$ for all prior probabilities $\Gamma$ if and only if*

$$\mathbb{E}_{\beta'}[\Phi(L_{\beta'})|g] \leq \mathbb{E}_\beta[\Phi(L_\beta)|g],$$

*$\forall$ continuous concave functions $\Phi(L)$, where $L_\beta = p_1(x,\beta)/p_2(x,\beta)$ is the likelihood ratio and $\mathbb{E}_\omega[\Phi(L_\omega)|g]$ is the expectation of $\Phi(L_\omega)$ under the hypothesis $g = P_2$.*

**Theorem 3.4.** *If $B_\alpha(\rho(\beta)) > B_\alpha(\rho(\beta'))$, or equivalently $\rho(\beta) < \rho(\beta')$ then $\exists$ a set of prior probabilities $\Gamma = \{\pi_1,\pi_2\}$ for two hypothesis $g_1,g_2$, for which*

$$P_e(\beta,\Gamma) < P_e(\beta',\Gamma). \tag{18}$$

*Proof.* The proof closely follows Kailath (Kailath, 1967). First note that $\sqrt{L}$ is a concave function of $L$ (likelihood ratio) , and

$$\begin{aligned}\rho(\beta) &= \sum_{x\in X}\sqrt{p_1(x,\beta)p_2(x,\beta)} \\ &= \sum_{x\in X}\sqrt{\frac{p_1(x,\beta)}{p_2(x,\beta)}}p_2(x,\beta) \\ &= \mathbb{E}_\beta\left[\sqrt{L_\beta}|g_2\right]. \end{aligned}\tag{19}$$

Similarly

$$\rho(\beta') = \mathbb{E}_{\beta'}\left[\sqrt{L_{\beta'}}|g_2\right] \tag{20}$$

Hence, $\rho(\beta) < \rho(\beta') \Rightarrow$

$$\mathbb{E}_\beta\left[\sqrt{L_\beta}|g_2\right] < \mathbb{E}_{\beta'}\left[\sqrt{L_{\beta'}}|g_2\right]. \tag{21}$$

Suppose assertion of the stated theorem is not true, then for all $\Gamma$, $P_e(\beta',\Gamma) \leq P_e(\beta,\Gamma)$. Then by Theorem 3.3, $\mathbb{E}_{\beta'}[\Phi(L_{\beta'})|g_2] \leq \mathbb{E}_\beta[\Phi(L_\beta)|g_2]$ which contradicts our result in Eq. 21. □

## 3.3 Bounds on Error Probability

Error probabilities are hard to calculate in general. Tight bounds on $P_e$ are often extremely useful in practice. Kailath (Kailath, 1967) has shown bounds on $P_e$ in terms of the Bhattacharyya coefficient $\rho$:

$$\tfrac{1}{2}\left[2\pi_1 - \sqrt{1-4\pi_1\pi_2\rho^2}\right] \leq P_e \leq \left(\pi_1 - \tfrac{1}{2}\right) + \sqrt{\pi_1\pi_2}\rho, \tag{22}$$

with $\pi_1+\pi_2 = 1$. If the priors are equal $\pi_1 = \pi_2 = \frac{1}{2}$, the expression simplifies to

$$\frac{1}{2}\left[1-\sqrt{1-\rho^2}\right] \leq P_e \leq \frac{1}{2}\rho. \tag{23}$$

Inverting relation in Eq. 6 for $\rho(B_\alpha)$, we can get the bounds in terms of $B_\alpha(\rho)$ measure. For the equal prior probabilities case, Bhattacharyya coefficient gives a tight upper bound for large systems when $\rho \to 0$ (zero overlap) and the observations are independent and identically distributed. These bounds are also useful to discriminate between two processes with arbitrarily low error probability (Kailath, 1967). We suppose that tighter upper bounds on error probability can be derived through Matusita's measure of affinity (Bhattacharya and Toussaint, 1982; Toussaint, 1977; Toussaint, 1975), but is beyond the scope of present work.

## 3.4 F-divergence

A class of divergence measures called f-divergences were introduced by Csiszar (Csiszar, 1967; Csiszar, 1975) and independently by Ali and Silvey (Ali and Silvey, 1966) (see (Basseville, 1989) for review). It encompasses many well known divergence measures including KLD, variational, Bhattacharyya and Hellinger distance. In this section, we show that $B_\alpha(\rho)$ measure for $\alpha \in (1,\infty]$, belongs to the generic class of f-divergences defined by Basseville (Basseville, 1989).

**F-divergence (Basseville, 1989).** Consider a measurable space $\Omega$ with $\sigma$-algebra $\mathcal{B}$. Let $\lambda$ be a measure on $(\Omega,\mathcal{B})$ such that any probability laws $P$ and $Q$ are absolutely continuous with respect to $\lambda$, with densities $p$ and $q$. Let $f$ be a continuous *convex* real function on $\mathbb{R}^+$, and $g$ be an *increasing* function on $\mathbb{R}$. The class of divergence coefficients between two probabilities:

$$d(P,Q) = g\left(\int_\Omega f\left(\frac{p}{q}\right)qd\lambda\right) \tag{24}$$

are called the f-divergence measure w.r.t. functions $(f,g)$ . Here $p/q = L$ is the likelihood ratio. The term in the parenthesis of $g$ gives the Csiszar's (Csiszar, 1967; Csiszar, 1975) definition of f-divergence.

The $B_\alpha(\rho(P,Q))$ , for $\alpha \in (1,\infty]$ measure can be written as the following $f$ divergence:

$$f(x) = -1 + \frac{1-\sqrt{x}}{\alpha}, \;\; g(F) = \frac{\log(-F)}{\log(1-1/\alpha)}, \quad (25)$$

where,

$$\begin{aligned} F &= \int_\Omega \left[ -1 + \frac{1}{\alpha}\left(1 - \sqrt{\frac{p}{q}}\right)\right] q d\lambda \\ &= \int_\Omega \left[ q\left(-1+\frac{1}{\alpha}\right) - \frac{1}{\alpha}\sqrt{pq}\right] d\lambda \\ &= -1 + \frac{1-\rho}{\alpha}. \end{aligned} \quad (26)$$

and

$$g(F) = \frac{\log(1-\frac{1-\rho}{\alpha})}{\log(1-1/\alpha)} = B_\alpha(\rho(P,Q)). \quad (27)$$

## 3.5 Curvature and Fisher Information

In statistics, the information that an observable random variable $X$ carries about an unknown parameter $\theta$ (on which it depends) is given by the Fisher information. One of the important properties of f-divergence of two distributions of the same parametric family is that their curvature measures the Fisher information. Following the approach pioneered by Rao (Rao, 1945), we relate the curvature of BBD measures to the Fisher information and derive the differential curvature metric. The discussions below closely follow (DasGupta, 2011).

**Definition.** Let $\{f(x|\theta); \theta \in \Theta \subseteq \mathbb{R}\}$, be a family of densities indexed by real parameter $\theta$, with some regularity conditions ($f(x|\theta)$ is absolutely continuous).

$$Z_\theta(\phi) \equiv B_\alpha(\theta,\phi) = \frac{\log(1-\frac{1-\rho(\theta,\phi)}{\alpha})}{\log(1-1/\alpha)} \quad (28)$$

where $\rho(\theta,\phi) = \int \sqrt{f(x|\theta) f(x|\phi)} dx$

**Theorem 3.5.** *Curvature of $Z_\theta(\phi)|_{\phi=\theta}$ is the Fisher information of $f(x|\theta)$ up to a multiplicative constant.*

*Proof.* Expand $Z_\theta(\phi)$ around theta

$$Z_\theta(\phi) = Z_\theta(\theta) + (\phi-\theta)\frac{dZ_\theta(\phi)}{d\phi} + \frac{(\phi-\theta)^2}{2}\frac{d^2 Z_\theta(\phi)}{d\phi^2} + \dots \quad (29)$$

Let us observe some properties of Bhattacharyya coefficient : $\rho(\theta,\phi) = \rho(\phi,\theta)$, $\rho(\theta,\theta) = 1$, and its derivatives:

$$\frac{\partial \rho(\theta,\phi)}{\partial \phi}\Big|_{\phi=\theta} = \frac{1}{2}\frac{\partial}{\partial\theta}\int f(x|\theta) dx = 0, \quad (30)$$

$$\begin{aligned} \frac{\partial^2 \rho(\theta,\phi)}{\partial \phi^2}\Big|_{\phi=\theta} &= -\frac{1}{4}\int \frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2 dx + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\int f dx \\ &= -\frac{1}{4}\int f(x|\theta)\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)^2 dx \\ &= -\frac{1}{4} I_f(\theta). \end{aligned} \quad (31)$$

where $I_f(\theta)$ is the Fisher Information of distribution $f(x|\theta)$

$$I_f(\theta) = \int f(x|\theta)\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)^2 dx. \quad (32)$$

Using the above relationships, we can write down the terms in the expansion of Eq. 29 $Z_\theta(\theta) = 0$ , $\frac{\partial Z_\theta(\phi)}{\partial\phi}\Big|_{\phi=\theta} = 0$, and

$$\frac{\partial^2 Z_\theta(\phi)}{\partial\phi^2}\Big|_{\phi=\theta} = C(\alpha) I_f(\theta) > 0, \quad (33)$$

where $C(\alpha) = \frac{-1}{4\alpha\log(1-1/\alpha)} > 0$ □

The leading term of $B_\alpha(\theta,\phi)$ is given by

$$B_\alpha(\theta,\phi) \sim \frac{(\phi-\theta)^2}{2} C(\alpha) I_f(\theta). \quad (34)$$

## 3.6 Differential Metrics

Rao (Rao, 1987) generalized the Fisher information to multivariate densities with vector valued parameters to obtain a "geodesic" distance between two parametric distributions $P_\theta, P_\phi$ of the same family. The Fisher-Rao metric has found applications in many areas such as image structure and shape analysis (Maybank, 2004; Peter and Rangarajan, 2006) , quantum statistical inference (Brody and Hughston, 1998) and Blackhole thermodynamics (Quevedo, 2008). We derive such a metric for BBD measure using property of f-divergence.

Let $\theta,\phi \in \Theta \subseteq \mathbb{R}^p$, then using the fact that $\frac{\partial Z(\theta,\phi)}{\partial\theta_i}\Big|_{\phi=\theta} = 0$, we can easily show that

$$dZ_\theta = \sum_{i,j=1}^{p} \frac{\partial^2 Z_\theta}{\partial\theta_i \partial\theta_j} d\theta_i d\theta_j + \dots = \sum_{i,j=1}^{p} g_{ij} d\theta_i d\theta_j + \dots. \quad (35)$$

The curvature metric $g_{ij}$ can be used to find the geodesic on the curve $\eta(t)$, $t \in [0,1]$ with

$$\mathcal{C} = \eta(t): \;\; \eta(0) = \theta \;\; \eta(1) = \phi. \quad (36)$$

Details of the geodesic equation are given in many standard differential geometry books. In the context of probability distance measures reader is referred to (see 15.4.2 in A DasGupta (DasGupta, 2011)

for details) The curvature metric of all Csiszar f-divergences are just scalar multiple KLD measure (DasGupta, 2011; Basseville, 1989) given by:

$$g^f_{ij}(\theta) = f''(1) g_{ij}(\theta). \tag{37}$$

For our BBD measure

$$\begin{aligned} f''(x) &= \left(-1 + \frac{1-\sqrt{x}}{\alpha}\right)'' = \frac{1}{4\alpha x^{3/2}} \\ \tilde{f}''(1) &= 1/4\alpha. \end{aligned} \tag{38}$$

Apart from the $-1/\log(1-\frac{1}{\alpha})$, this is same as $C(\alpha)$ in Eq. 34. It follows that the geodesic distance for our metric is same KLD geodesic distance up to a multiplicative factor. KLD geodesic distances are tabulated in DasGupta (DasGupta, 2011).

## 3.7 Relation to Other Measures

Here we focus on the special case $\alpha = 2$, i.e. $B_2(\rho)$

**Theorem 3.6.**

$$B_2 \le H^2 \le \log 4 \; B_2 \tag{39}$$

*where 1 and* $\log 4$ *are sharp.*

*Proof.* Sharpest upper bound is achieved via taking $\sup_{\rho\in[0,1)} \frac{H^2(\rho)}{B_2(\rho)}$. Define

$$g(\rho) \equiv \frac{1-\rho}{-\log_2(1+\rho)/2}. \tag{40}$$

We note that $g(\rho)$ is continuous and has no singularities whenever $\rho \in [0,1)$. Hence

$$g'(\rho) = \frac{\frac{1-\rho}{1+\rho} + \log(\frac{1+\rho}{2})}{\log^2 \frac{\rho+1}{2}} \log 2 \ge 0.$$

It follows that $g(\rho)$ is non-decreasing and hence $\sup_{\rho\in[0,1)} g(\rho) = \lim_{\rho\to 1} g(\rho) = \log(4)$. Thus

$$H^2/B_2 \le \log 4. \tag{41}$$

Combining this with convexity property of $B_\alpha(\rho)$ for $\alpha > 1$, we get

$$B_2 \le H^2 \le \log 4 \; B_2$$

Using the same procedure we can prove a generic version of this inequality for $\alpha \in (1,\infty]$ , given by

$$B_\alpha(\rho) \le H^2 \le -\alpha \log\left(1 - \frac{1}{\alpha}\right) B_\alpha(\rho) \tag{42}$$

□

**Jensen-Shannon Divergence:** The Jensen difference between two distributions $P, Q$, with densities $p, q$ and weights $(\lambda_1, \lambda_2)$; $\lambda_1 + \lambda_2 = 1$, is defined as,

$$\mathcal{J}_{\lambda_1,\lambda_2}(P,Q) = H_s(\lambda_1 p + \lambda_2 q) - \lambda_1 H_s(p) - \lambda_2 H_s(q), \tag{43}$$

where $H_s$ is the Shannon entropy. Jensen-Shannon divergence (JSD) (Burbea and Rao, 1982; Rao, 1982b; Lin, 1991) is based on the Jensen difference and is given by:

$$\begin{aligned} JS(P,Q) =& \mathcal{J}_{1/2,1/2}(P,Q) \\ =& \frac{1}{2}\int \Big[ p \log\left(\frac{2p}{p+q}\right) \\ &+ q \log\left(\frac{2q}{p+q}\right)\Big] d\lambda \end{aligned} \tag{44}$$

The structure and goals of JSD and BBD measures are similar. The following theorem compares the two metrics using Jensen's inequality.

**Lemma 3.7** ***Jensen's Inequality:*** *For a convex function* $\psi$, $\mathbb{E}[\psi(X)] \ge \psi(\mathbb{E}[X])$.

**Theorem 3.8 (Relation to Jensen-Shannon measure).** $JS(P,Q) \ge \frac{2}{\log 2} B_2(P,Q) - \log 2$

We use the un-symmetrized Jensen-Shannon metric for the proof.

*Proof.*

$$\begin{aligned} JS(P,Q) =& \int p(x) \log \frac{2p(x)}{p(x)+q(x)} dx \\ =& -2\int p(x) \log \frac{\sqrt{p(x)+q(x)}}{\sqrt{2p(x)}} dx \\ \ge& -2\int p(x) \log \frac{\sqrt{p(x)}+\sqrt{q(x)}}{\sqrt{2p(x)}} dx \\ & (\text{since } \sqrt{p+q} \le \sqrt{p}+\sqrt{q}) \\ =& \mathbb{E}_P\left[-2\log \frac{\sqrt{p(X)}+\sqrt{q(X)}}{\sqrt{2p(X)}}\right] \end{aligned}$$

By Jensen's inequality
$\mathbb{E}[-\log f(X)] \ge -\log \mathbb{E}[f(X)]$, we have

$$\begin{aligned} &\mathbb{E}_P\left[-2\log \frac{\sqrt{p(X)}+\sqrt{q(X)}}{\sqrt{2p(X)}}\right] \ge \\ &\quad -2\log \mathbb{E}_P\left[\frac{\sqrt{p(X)}+\sqrt{q(X)}}{\sqrt{2p(X)}}\right]. \end{aligned}$$

Hence,

$$\begin{aligned} JS(P,Q) &\geq -2\log\int p(x)\frac{\left(\sqrt{p(x)}+\sqrt{q(x)}\right)}{\sqrt{2p(x)}}dx \\ &= -2\log\left(\frac{1+\int\sqrt{p(x)q(x)}}{2}\right)-\log 2 \\ &= 2\left(\frac{B_2(p(x),q(x))}{\log 2}\right)-\log 2 \\ &= \frac{2}{\log 2}B_2(P,Q)-\log 2. \end{aligned} \tag{45}$$

□

# 4 APPLICATION TO SIGNAL DETECTION

Signal detection is a common problem occurring in many fields such as communication engineering, pattern recognition, and Gravitational wave detection (Poor, 1994). In this section, we briefly describe the problem and terminology used in signal detection. We illustrate though simple cases how divergence measures, in particular BBD can be used for discriminating and detecting signals buried in white noise of correlator receivers (matched filter). For greater details of the formalism used we refer the reader to review articles in the context of Gravitational wave detection by Jaranowski and Królak(Jaranowski and Królak, 2007) and Sam Finn (Finn, 1992).

One of the central problem in signal detection is to detect whether a deterministic signal $s(t)$ is embedded in an observed data $x(t)$, corrupted by noise $n(t)$. This can be posed as a hypothesis testing problem where the null hypothesis is absence of signal and alternative is its presence. We take the noise to be additive , so that $x(t) = n(t) + s(t)$. We define the following terms used in signal detection: *Correlation G* (also called matched filter) between $x$ and $s$, and *signal to noise ratio* $\rho$ (Finn, 1992; Budzyński et al., 2008).

$$G = (x|s), \quad \rho = \sqrt{(s,s)}, \tag{46}$$

where the scalar product $(.|.)$ is defined by

$$(x|y) := 4\Re\int_0^\infty \frac{\tilde{x}(f)\tilde{y}^*(f)}{\tilde{N}(f)}\,\mathrm{d}f. \tag{47}$$

$\Re$ denotes the real part of a complex expression, tilde denotes the Fourier transform and the asterisk * denotes complex conjugation. $\tilde{N}$ is the *one-sided spectral density of the noise*.
For white noise, the probability densities of $G$ when respectively signal is present and absent are given by (Budzyński et al., 2008)

$$p_1(G) = \frac{1}{\sqrt{2\pi}\rho}\exp\left(-\frac{(G-\rho^2)^2}{2\rho^2}\right), \tag{48}$$

$$p_0(G) = \frac{1}{\sqrt{2\pi}\rho}\exp\left(-\frac{G^2}{2\rho^2}\right) \tag{49}$$

## 4.1 Distance between Gaussian Processes

Consider a stationary Gaussian random process $\mathbf{X}$, which has signals $s_1$ or $s_2$ with probability densities $p_1$ and $p_2$ respectively of being present in it. These densities follow the form Eq. 48 with signal to noise ratios $\rho_1^2$ and $\rho_2^2$ respectively. The probability density $p(\mathbf{X})$ of Gaussian process can modeled as limit of multivariate Gaussian distributions. The divergence measures between these processes $d(s_1,s_2)$ are in general functions of the correlator $(s_1-s_2|s_1-s_2)$ (Budzyński et al., 2008). Here we focus on distinguishing monochromatic signal $s(t) = A\cos(\omega t+\phi)$ and filter $s_F(t) = A_F\cos(\omega_F t+\phi)$ (both buried in noise), separated in frequency or amplitude.

The Kullback-Leibler divergence between the signal and filter $I(s,s_F)$ is given by the correlation $(s-s_F|s-s_F)$:

$$\begin{aligned} I(s,s_F) =& (s-s_F|s-s_F) = (s|s)+(s_F|s_F)-2(s|s_F) \\ =& \rho^2+\rho_F^2-2\rho\rho_F[\langle\cos(\Delta\omega t)\rangle\cos(\Delta\phi) \\ & -\langle\sin(\Delta\omega t)\rangle\sin(\Delta\phi)], \end{aligned} \tag{50}$$

where $\langle\rangle$ is the average over observation time $[0,T]$. Here we have assumed that noise spectral density $N(f) = N_0$ is constant over the frequencies $[\omega,\omega_F]$. The SNRs are given by

$$\rho^2 = \frac{A^2T}{N_0}, \quad \rho_F^2 = \frac{A_F^2T}{N_0}. \tag{51}$$

(for detailed discussions we refer the reader to Budzynksi *et. al* (Budzyński et al., 2008)).

The Bhattacharyya distance between Gaussian processors with signals of same energy is ( Eq 14 in (Kailath, 1967)) just a multiple of the KLD $B = I/8$. We use this result to extract the Bhattacharyya coefficient :

$$\rho(s,s_F) = \exp\left(-\frac{(s-s_F|s-s_F)}{8}\right) \tag{52}$$

### 4.1.1 Frequency Difference

Let us consider the case when the SNRs of signal and filter are equal, phase difference is zero, but frequencies differ by $\Delta\omega$. The KL divergence is obtained by

evaluating the correlator in Eq. 50

$$I(\Delta\omega) = (s - s_F|s - s_F) = 2\rho^2\left(1 - \frac{\sin(\Delta\omega T)}{\Delta\omega T}\right). \tag{53}$$

by noting $\langle\cos(\Delta\omega t)\rangle = \frac{\sin(\Delta\omega T)}{\Delta\omega T}$ and $\langle\sin(\Delta\omega t)\rangle = \frac{1-\cos(\Delta\omega T)}{\Delta\omega T}$. Using this, the expression for BBD family can be written as

$$B_\alpha(\Delta\omega) = \frac{\log\left(1 - \frac{1}{\alpha}\left[1 - e^{-\frac{\rho^2}{4}\left(1-\frac{\sin(\Delta\omega T)}{\Delta\omega T}\right)}\right]\right)}{\log\left(1 - \frac{1}{\alpha}\right)}. \tag{54}$$

As we have seen in section 3.4, both BBD and KLD belong to the f-divergence family. Their curvature for distributions belonging to same parametric family is a constant times the Fisher information (FI) (see Theorem: 3.5). Here we discuss where the BBD and KLD deviates from FI, when we account for higher terms in the expansion of these measures.

The Fisher matrix element for frequency $g_{\omega,\omega} = E\left[\left(\frac{\partial \log\Lambda}{\partial\omega}\right)^2\right] = \rho^2 T^2/3$ (Budzyński et al., 2008), where $\Lambda$ is the likelihood ratio. Using the relation for line element $ds^2 = \sum_{i,j} g_{ij} d\theta_i d\theta_j$ and noting that only frequency is varied, we get

$$\mathrm{d}s = \frac{\rho T \Delta\omega}{\sqrt{3}}. \tag{55}$$

Using the relation between curvature of BBD measure and Fisher's Information in Eq.34, we can see that for low frequency differences the line element varies as:

$$\sqrt{\frac{2B_\alpha(\Delta\omega)}{C(\alpha)}} \sim \mathrm{d}s.$$

Similarly $\sqrt{d_{KL}} \sim ds$ at low frequencies. However, at higher frequencies both KLD and BBD deviate from the Fisher information metric. In Fig. 2, we have plotted d$s$, $\sqrt{d_{KL}}$ and $\sqrt{2B_\alpha(\Delta\omega)/C(\alpha)}$ with $\alpha = 2$ and Hellinger distance ($\alpha \to \infty$) for $\Delta\omega \in (0, 0.1)$. We observe that till $\Delta\omega = 0.01$ (i.e. $\Delta\omega T \sim 1$), KLD and BBD follows Fisher Information and after that they start to deviate. This suggests that Fisher Information is not sensitive to large deviations. There is not much difference between KLD, BBD and Hellinger for large frequencies due to the correlator $G$ becoming essentially a constant over a wide range of frequencies.

#### 4.1.2 Amplitude Difference

We now consider the case where the frequency and phase of the signal and the filter are same but they differ in amplitude $\Delta A$ (which reflects in differing SNR).

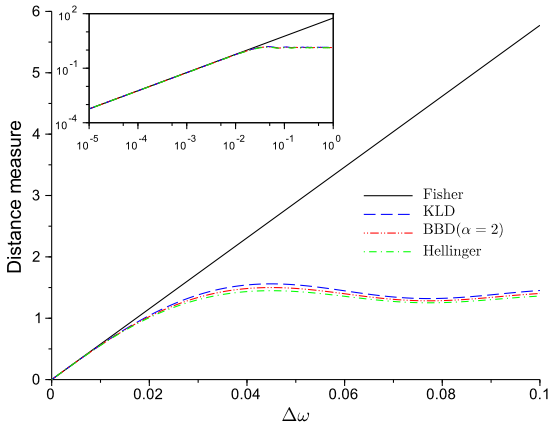


Figure 2: Comparison of Fisher Information, KLD, BBD and Hellinger distance for two monochromatic signals differing by frequency $\Delta\omega$, buried in white noise. Inset shows wider range $\Delta\omega \in (0, 1)$ . We have set $\rho = 1$ and chosen parameters $T = 100$ and $N_0 = 10^4$.

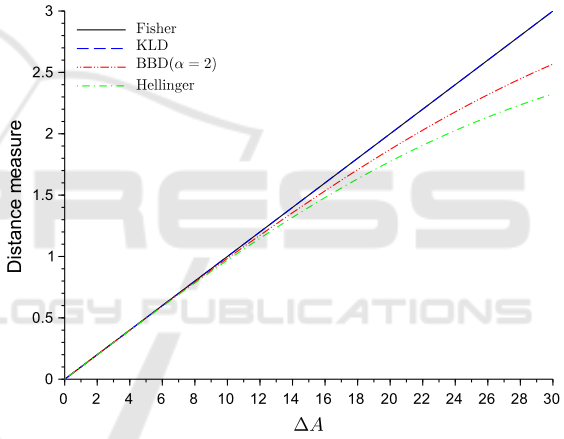


Figure 3: Comparison of Fisher information line element with KLD, BBD and Hellinger distance for signals differing in amplitude and buried in white noise. We have set $A = 1$, $T = 100$ and $N_0 = 10^4$.

The correlation reduces to

$$\begin{aligned}(s - s_F|s - s_F) &= \frac{A^2 T}{N_0} + \frac{(A+\Delta A)^2 T}{N_0} - 2\frac{A(A+\Delta A)T}{N_0} \\ &= \frac{(\Delta A)^2 T}{N_0}.\end{aligned} \tag{56}$$

This gives us $I(\Delta A) = \frac{(\Delta A)^2 T}{N_0}$, which is the same as the line element $ds^2$ with Fisher metric $ds = \sqrt{T/2N_0}\Delta A$. In Fig. 3, we have plotted d$s$, $\sqrt{d_{KL}}$ and $\sqrt{2B_\alpha(\Delta\omega)/C(\alpha)}$ for $\Delta A \in (0, 40)$. KLD and FI line element are the same. Deviations of BBD and Hellinger can be observed only for $\Delta A > 10$.

Discriminating between two signals $s_1, s_2$ requires minimizing the error probability between them. By

Theorem 3.4, there exists priors for which the problem translates into maximizing the divergence for BBD measures. For the monochromatic signals discussed above, the distance depends on parameters $(\rho_1, \rho_2, \Delta\omega, \Delta\phi)$. We can maximize the distance for a given frequency difference by differentiating with respect to phase difference $\Delta\phi$ (Budzyński et al., 2008). In Fig. 4, we show the variation of maximized BBD for different signal to noise ratios $(\rho_1, \rho_2)$, for a fixed frequency difference $\Delta\omega = 0.01$. The intensity map shows different bands which can be used for setting the threshold for signal separation.

Detecting signal of known form involves minimizing the distance measure over the parameter space of the signal. A threshold on the maximum "distance" between the signal and filter can be put so that a detection is said to occur whenever the measures fall within this threshold. Based on a series of tests, Receiver Operating Characteristic (ROC) curves can be drawn to study the effectiveness of the distance measure in signal detection. We leave such details for future work.

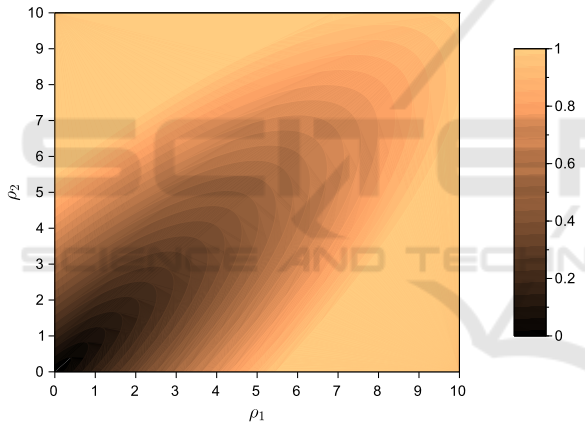


Figure 4: BBD with different signal to noise ratio for a fixed. We have set $T = 100$ and $\Delta\omega = 0.01$.

## 5 SUMMARY AND OUTLOOK

In this work we have introduced a new family of bounded divergence measures based on the Bhattacharyya distance, called bounded Bhattacharyya distance measures. We have shown that it belongs to the class of generalized f-divergences and inherits all its properties, such as those relating Fishers Information and curvature metric. We have discussed several special cases of our measure, in particular squared Hellinger distance, and studied relation with other measures such as Jensen-Shannon divergence. We have also extended the Bradt Karlin theorem on error probabilities to BBD measure. Tight bounds on Bayes error probabilities can be put by using properties of Bhattacharyya coefficient.

Although many bounded divergence measures have been studied and used in various applications, no single measure is useful in all types of problems studied. Here we have illustrated an application to signal detection problem by considering "distance" between monochromatic signal and filter buried in white Gaussian noise with differing frequency or amplitude, and comparing it to Fishers Information and Kullback-Leibler divergence.

A detailed study with chirp like signal and colored noise occurring in Gravitational wave detection will be taken up in a future study. Although our measures have a tunable parameter $\alpha$, here we have focused on a special case with $\alpha = 2$. In many practical applications where extremum values are desired such as minimal error, minimal false acceptance/rejection ratio etc, exploring the BBD measure by varying $\alpha$ may be desirable. Further, the utility of BBD measures is to be explored in parameter estimation based on minimal disparity estimators and Divergence information criterion in Bayesian model selection (Basu and Lindsay, 1994). However, since the focus of the current paper is introducing a new measure and studying its basic properties, we leave such applications to statistical inference and data processing to future studies.

## ACKNOWLEDGEMENTS

One of us (S.J) thanks Rahul Kulkarni for insightful discussions, Anand Sengupta for discussions on application to signal detection, and acknowledge the financial support in part by grants DMR-0705152 and DMR-1005417 from the US National Science Foundation. M.S. would like to thank the Penn State Electrical Engineering Department for support.

## 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. Series B (Methodological)*, 28(1):131–142.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal processing*, 18:349–369.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705.

Ben-Bassat, M. (1978). f-entropies, probability of er-

ror, and feature selection. *Information and Control*, 39(3):227–242.

Bhattacharya, B. K. and Toussaint, G. T. (1982). An upper bound on the probability of misclassification in terms of matusita's measure of affinity. *Annals of the Institute of Statistical Mathematics*, 34(1):161–165.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc*, 35(99-109):4.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406.

Blackwell, D. (1951). Comparison of experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 93–102.

Bradt, R. and Karlin, S. (1956). On the design and comparison of certain dichotomous experiments. *The Annals of mathematical statistics*, pages 390–409.

Brody, D. C. and Hughston, L. P. (1998). Statistical geometry in quantum mechanics. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1977):2445–2475.

Budzyński, R. J., Kondracki, W., and Królak, A. (2008). Applications of distance between probability distributions to gravitational wave data analysis. *Classical and Quantum Gravity*, 25(1):015005.

Burbea, J. and Rao, C. R. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489 – 495.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):pp. 493–507.

Choi, E. and Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709.

Csiszar, I. (1967). Information-type distance measures and indirect observations. *Stud. Sci. Math. Hungar*, 2:299–318.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):pp. 146–158.

DasGupta, A. (2011). *Probability for Statistics and Machine Learning*. Springer Texts in Statistics. Springer New York.

Finn, L. S. (1992). Detection, measurement, and gravitational radiation. *Physical Review D*, 46(12):5236.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, (136):210–271.

Hellman, M. E. and Raviv, J. (1970). Probability of Error, Equivocation, and the Chernoff Bound. *IEEE Transactions on Information Theory*, 16(4):368–372.

Jaranowski, P. and Królak, A. (2007). Gravitational-wave data analysis. formalism and sample applications: the gaussian case. *arXiv preprint arXiv:0711.1115*.

Kadota, T. and Shepp, L. (1967). On the best finite set of linear observables for discriminating two gaussian signals. *IEEE Transactions on Information Theory*, 13(2):278–284.

Kailath, T. (1967). The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communications*, 15(1):52–60.

Kakutani, S. (1948). On equivalence of infinite product measures. *The Annals of Mathematics*, 49(1):214–224.

Kapur, J. (1984). A comparative assessment of various measures of directed divergence. *Advances in Management Studies*, 3(1):1–16.

Kullback, S. (1968). Information theory and statistics. *New York: Dover, 1968, 2nd ed.*, 1.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):pp. 79–86.

Kumar, U., Kumar, V., and Kapur, J. N. (1986). Some normalized measures of directed divergence. *International Journal of General Systems*, 13(1):5–16.

Lamberti, P. W., Majtey, A. P., Borras, A., Casas, M., and Plastino, A. (2008). Metric character of the quantum Jensen-Shannon divergence . *Physical Review A*, 77:052311.

Lee, Y.-T. (1991). Information-theoretic distortion measures for speech recognition. *Signal Processing, IEEE Transactions on*, 39(2):330–335.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145 –151.

Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, 19(1):181–192.

Maybank, S. J. (2004). Detection of image structures using the fisher information and the rao metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1579–1589.

Nielsen, F. and Boltz, S. (2011). The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466.

Nielsen, M. and Chuang, I. (2000). Quantum computation and information. *Cambridge University Press, Cambridge, UK*, 3(8):9.

Peter, A. and Rangarajan, A. (2006). Shape analysis using the fisher-rao riemannian metric: Unifying shape representation and deformation. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 1164–1167. IEEE.

Poor, H. V. (1994). *An introduction to signal detection and estimation*. Springer.

Qiao, Y. and Minematsu, N. (2010). A study on invariance of-divergence and its application to speech recognition. *Signal Processing, IEEE Transactions on*, 58(7):3884–3890.

Quevedo, H. (2008). Geometrothermodynamics of black holes. *General Relativity and Gravitation*, 40(5):971–984.

Rao, C. (1982a). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya: The Indian Journal of Statistics, Series A*, pages 1–22.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91.

Rao, C. R. (1982b). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24 – 43.

Rao, C. R. (1987). Differential metrics in probability spaces. *Differential geometry in statistical inference*, 10:217–240.

Royden, H. (1986). *Real analysis*. Macmillan Publishing Company, New York.

Sunmola, F. T. (2013). *Optimising learning with transferable prior information*. PhD thesis, University of Birmingham.

Toussaint, G. T. (1974). Some properties of matusita's measure of affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 26(1):389–394.

Toussaint, G. T. (1975). Sharper lower bounds for discrimination information in terms of variation (corresp.). *Information Theory, IEEE Transactions on*, 21(1):99–100.

Toussaint, G. T. (1977). An upper bound on the probability of misclassification in terms of the affinity. *Proceedings of the IEEE*, 65(2):275–276.

Toussaint, G. T. (1978). Probability of error, expected divergence and the affinity of several distributions. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):482–485.

Tumer, K. and Ghosh, J. (1996). Estimating the Bayes error rate through classifier combining. *Proceedings of 13th International Conference on Pattern Recognition*, pages 695–699.

Varshney, K. R. (2011). Bayes risk error is a bregman divergence. *IEEE Transactions on Signal Processing*, 59(9):4470–4472.

Varshney, K. R. and Varshney, L. R. (2008). Quantization of prior probabilities for hypothesis testing. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 56(10):4553.

# APPENDIX

**BBD Measures of Some Common Distributions.** Here we provide explicit expressions for BBD $B_2$, for some common distributions. For brevity we denote $\zeta \equiv B_2$.

- **Binomial** :

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k},\ Q(k) = \binom{n}{k} q^k (1-q)^{n-k}.$$

$$\zeta_{bin}(P,Q) = -\log_2 \left( \frac{1 + [\sqrt{pq} + \sqrt{(1-p)(1-q)}]^n}{2} \right). \quad (57)$$

- **Poisson** :

$$P(k) = \frac{\lambda_p^k e^{-\lambda_p}}{k!},\ Q(k) = \frac{\lambda_q^k e^{-\lambda_q}}{k!}.$$

$$\zeta_{poisson}(P,Q) = -\log_2 \left( \frac{1 + e^{-(\sqrt{\lambda_p} - \sqrt{\lambda_q})^2/2}}{2} \right). \quad (58)$$

- **Gaussian** :

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left( -\frac{(x-x_p)^2}{2\sigma_p^2} \right),$$
$$Q(x) = \frac{1}{\sqrt{2\pi}\sigma_q} \exp\left( -\frac{(x-x_q)^2}{2\sigma_q^2} \right).$$

$$\zeta_{Gauss}(P,Q) = 1 - \log_2 \left[ 1 + \frac{2\sigma_p\sigma_q}{\sigma_p^2 + \sigma_q^2} \exp\left( -\frac{(x_p - x_q)^2}{4(\sigma_p^2 + \sigma_q^2)} \right) \right]. \quad (59)$$

- **Exponential** : $P(x) = \lambda_p e^{-\lambda_p x}$, $Q(x) = \lambda_q e^{-\lambda_q x}$.

$$\zeta_{exp}(P,Q) = -\log_2 \left[ \frac{(\sqrt{\lambda_p} + \sqrt{\lambda_q})^2}{2(\lambda_p + \lambda_q)} \right]. \quad (60)$$

- **Pareto** : Assuming the same cut off $x_m$,

$$P(x) = \begin{cases} \alpha_p \frac{x_m^{\alpha_p}}{x^{\alpha_p+1}} & \text{for } x \geq x_m \\ 0 & \text{for } x < x_m, \end{cases} \quad (61)$$

$$Q(x) = \begin{cases} \alpha_q \frac{x_m^{\alpha_q}}{x^{\alpha_q+1}} & \text{for } x \geq x_m \\ 0 & \text{for } x < x_m. \end{cases} \quad (62)$$

$$\zeta_{pareto}(P,Q) = -\log_2 \left[ \frac{(\sqrt{\alpha_p} + \sqrt{\alpha_q})^2}{2(\alpha_p + \alpha_q)} \right]. \quad (63)$$