Correspondence

Bayes Risk Error is a Bregman Divergence

Kush R. Varshney

***Abstract*—In previous work reported in these Transactions, we proposed a new distortion measure for the quantization of prior probabilities that are used in the threshold of likelihood ratio test detection: Bayes risk error. In this correspondence, we show that the Bayes risk error is a member of the class of Bregman divergences and discuss the implications of this fact.**

***Index Terms*—Bayesian hypothesis testing, Bregman divergence, quantization, signal detection.**

I. Introduction

Bayesian hypothesis testing for signal detection relies on precise knowledge of the prior probabilities of the hypotheses in setting thresholds of likelihood ratio tests [1]. Moreover, the detector requires such prior probabilities for the entire universe of objects it might observe. However, it may be that the detector can only use a finite number of prior probabilities due to memory and processing limitations, and must thus quantize the infinite set of prior probabilities across the universe to a finite set of representative priors [2]. Quantization or clustering to a finite set also guards against overfitting when prior probabilities are estimated from a small number of noisy observations per object [3]. In [2], motivated by these concerns we propose and substantiate a quantizer for probabilities with a novel distortion criterion, *Bayes risk error*, which directly incorporates detection performance measured by Bayes risk.

In reexamining the Bayes risk error distortion function herein, we come to the realization that it belongs to the class of *Bregman divergences* [4], a class that includes squared Euclidean and Mahalanobis distances, Kullback–Leibler divergence, generalized I-divergence, and Itakura–Saito divergence [5], [6]. This connection enriches the class of Bregman divergences. Importantly, it will allow us to link the theory developed for Bregman divergences to Bayes risk error, which may lead to new directions for research, and also allow us to feed advances that have been made in the context of Bayes risk error to general Bregman divergences.

The remainder of this correspondence is organized as follows. In Section II, we review the Bayes risk error distortion function. In Section III, we show that it is a Bregman divergence. In Section IV, we discuss implications of this realization. Finally, in Section V, we conclude.

II. Bayes Risk Error Distortion

In the signal detection problem, we have a noisy observation $Y \in \mathcal{Y}$ of an underlying hypothesis $H \in \{h_0, h_1\}$ that is distributed according to the likelihood functions $f_{Y|H}(y|H = h_0)$ and $f_{Y|H}(y|H = h_1)$.

Manuscript received February 25, 2011; revised June 01, 2011; accepted June 01, 2011. Date of publication June 13, 2011; date of current version August 10, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoît Champagne.

The author is with the Business Analytics and Mathematical Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598 USA (e-mail: krvarshn@us.ibm.com).

Digital Object Identifier 10.1109/TSP.2011.2159500

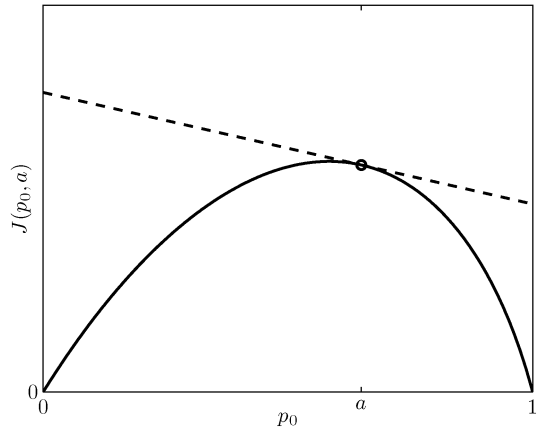

Fig. 1. Example $J(p_0)$ (solid curve) and $J(p_0, a)$ (dashed line).

The prior probabilities of the hypotheses are $p_0 = \Pr[H = h_0]$ and $p_1 = \Pr[H = h_1] = 1 - p_0$. The detection rule $\hat{h}(\cdot)$ is a mapping from $\mathcal{Y}$ to $\{h_0, h_1\}$. There are two types of errors that the detection rule may produce; it may output h_1 when the true hypothesis is h_0 (a type I error with probability $p_E^{\mathrm{I}} = \Pr[\hat{h}(Y) = h_1 | H = h_0]$) or it may output h_0 when the true hypothesis is h_1 (a type II error with probability $p_E^{\mathrm{II}} = \Pr[\hat{h}(Y) = h_0 | H = h_1]$). In the Bayesian hypothesis test, non-negative costs are associated with the two errors: c_{10} for type I errors and c_{01} for type II errors [1].

The detection rule is designed to minimize the total cost-weighted probability of error. This weighted combination of the two error probabilities is known as the Bayes risk. The detection rule that minimizes Bayes risk is the following likelihood ratio test:

$$\frac{f_{Y|H}(y|H = h_1)}{f_{Y|H}(y|H = h_0)} \underset{\hat{h}(y)=h_0}{\overset{\hat{h}(y)=h_1}{\gtrless}} \frac{ac_{10}}{(1-a)c_{01}} \tag{1}$$

where the parameter a in the threshold on the right side of (1) equals p_0 for optimality. The two error probabilities are a function of a and consequently of p_0 for the optimal detection rule. The Bayes risk of the optimal detection rule written as a function of p_0 is

$$J(p_0) = c_{10} p_0 p_E^{\mathrm{I}}(p_0) + c_{01}(1 - p_0) p_E^{\mathrm{II}}(p_0). \tag{2}$$

This Bayes risk function (2) is strictly concave in the interval (0,1) [7], [8].

As discussed in Section I, it may be that the precise prior probability is not employed in setting the threshold of the likelihood ratio test, i.e., $a \neq p_0$ in (1). In that suboptimal case, the Bayes risk is

$$J(p_0, a) = c_{10} p_0 p_E^{\mathrm{I}}(a) + c_{01}(1 - p_0) p_E^{\mathrm{II}}(a). \tag{3}$$

The function $J(p_0, a)$ is linear and tangent to $J(p_0)$ at a [1], [7], [8]. Example $J(p_0)$ and $J(p_0, a)$ are shown in Fig. 1.

In [2], we define a new distortion criterion for quantizing prior probabilities as the difference between the mismatched Bayes risk function (3) and the optimal Bayes risk function (2), naming it the Bayes risk error:

$$d(p_0, a) = J(p_0, a) - J(p_0), \tag{4}$$

with $p_0 \in (0,1)$ and $a \in (0,1)$. We show that the Bayes risk error is non-negative and only equal to zero when $p_0 = a$, that it is strictly convex in p_0, and that it is quasi-convex in a for deterministic likelihood ratio tests [2]. In the next section, we revisit the Bayes risk error and show that it is a Bregman divergence.

III. Interpretation as Bregman Divergence

Bregman divergences are functions of two arguments that map to the non-negative real numbers. They are defined based on strictly convex real-valued loss functions $\phi(\mathbf{u})$ defined over a convex set. The Bregman divergence is

$$d(\mathbf{u},\mathbf{v}) = \phi(\mathbf{u}) - \phi(\mathbf{v}) - \langle \mathbf{u}-\mathbf{v}, \nabla\phi(\mathbf{v})\rangle \tag{5}$$

where $\langle\cdot,\cdot\rangle$ is the inner product and $\nabla\phi(\mathbf{v})$ is the gradient of ϕ evaluated at $\mathbf{v}$ [4], [5]. With scalar arguments, the Bregman divergence simplifies to

$$d(u,v) = \phi(u) - \phi(v) - (u-v)\phi'(v). \tag{6}$$

Let the loss function be the negative Bayes risk function, i.e., $\phi(u) = -J(u)$. As such, due to the concavity of the Bayes risk function, $\phi(u)$ is strictly convex and differentiable over a convex set, the interval (0,1). The Bregman divergence generated by the Bayes risk function is

$$d(p_0,a) = -J(p_0) + J(a) + (p_0-a)J'(a). \tag{7}$$

We now show that this divergence is the Bayes risk error. As stated in Section II, the mismatched Bayes risk is a linear function that is tangent to $J(p_0)$ at a. Therefore, its slope is the derivative of $J(p_0)$ evaluated at a, i.e., $J'(a)$. Based on the point-slope formula of lines,

$$J(p_0,a) = J(a) + (p_0-a)J'(a). \tag{8}$$

Substituting this form into (4), we see that (7) and (4) are indeed equivalent functions, and that Bayes risk error is the Bregman divergence generated by the negative Bayes risk function.

IV. Discussion

In this section, we discuss some implications of the acknowledgement that Bayes risk error is a Bregman divergence.

A. Properties

Several properties of the Bayes risk error are proven in [2] that follow from the fact that it is a Bregman divergence; these properties could thus have been stated directly without the proofs given in [2], although the given proofs are of independent interest as alternatives. By being a Bregman divergence, it follows that the Bayes risk error is non-negative and only equal to zero when its two arguments are equal, and that it is convex in its first argument.

Banerjee *et al.* show that the centroid condition of Bregman divergences does not depend on the specific Bregman divergence that is used; all Bregman divergences have the same centroid condition: the expectation of the variable being quantized within the quantization cell is the unique minimizer [5], [6]. That is, within a fixed quantization cell $\mathcal{R}_k$, the optimal representation point is

$$a_k = \arg\min_a \int_{\mathcal{R}_k} d(p_0,a) f_{P_0}(p_0)dp_0 = \int_{\mathcal{R}_k} p_0 f_{P_0}(p_0)dp_0. \tag{9}$$

We found the mean Bayes risk error centroid condition for a specific additive Gaussian noise example [2, eq. (25)], but did not simplify the expression completely to the expectation. We also did not give this simply stated centroid condition generally for Bayes risk error, which follows from it being a Bregman divergence. As the centroid is the mean value of the distribution of prior probabilities within the quantization cell, it is not a function of the Bayes costs or the signal measurement model; they only affect the nearest neighbor condition.

Additionally, the Lloyd–Max algorithm is guaranteed to find a local optimum for Bregman divergences [6], which we proved for Bayes risk error based on [9]. The Bayes risk error inherits other properties of Bregman divergences that were not stated in [2]. For example, it is a linear operator and satisfies a generalized Pythagorean theorem [6]. The centroid condition given in [2] is for the right-sided centroid, but following [10], left-sided and symmetrized centroid conditions can be stated using the Legendre transform, although their interpretation from the Bayesian hypothesis testing perspective is unclear. (The Legendre transform of $-J(u)$ is

$$\psi(p) = -pJ'^{-1}(p) + J\left(-J'^{-1}(p)\right) \tag{10}$$

where J'^{-1} is the inverse function of the derivative of J.)

B. Bregman Information

Banerjee *et al.* define the concept of Bregman information for empirical or discrete distributions in [6], which corresponds to the mean Bayes risk error of the optimal representation point (when there is a single quantization level and we are working with an empirical distribution). The Bregman information is sample variance for squared Euclidean distance and is mutual information for Kullback–Leibler divergence, in both cases indicating a level of uncertainty. The Bregman information for Bayes risk error can be interpreted similarly; it represents the uncertainty of the decision rule with respect to an unknown prior. As such, mean Bayes risk error is the appropriate quantification of distortion for a rate-distortion characterization; it is an information radius [11].

C. M-ary Hypothesis Test

The signal detection problem considered in [2] and to this point in this correspondence is the *binary* Bayesian hypothesis test. It is mentioned in [2] that M-ary hypothesis testing for $M > 2$ can also be considered for quantization. However this extension has not been pursued heretofore because proofs in [2], especially for unique minimization, rely on properties of receiver operating characteristics; the operating characteristic is not straightforwardly defined for $M > 2$ hypotheses. With the recognition that Bayes risk error is a Bregman divergence generated by the negative Bayes risk function, no such proofs are necessary. All that is required is to present a strictly convex function defined over a convex domain, and all relevant properties and conditions for quantization follow.

With M hypotheses, we have M prior probabilities $p_i > 0$, $i = 0,\ldots,M-1$ such that $\sum_i p_i = 1$. Let us write the collection of priors as the vector $\mathbf{p}$. We also have an $M \times M$ matrix of costs c_{ij}. The detection rule in the M-ary case uses ratios of priors and costs analogously to the likelihood ratio test (1). The domain over which we are working is the simplex, which is convex. The Bayes risk function

$$J(\mathbf{p}) = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} c_{ij}p_j \Pr\left[\hat{h}(Y,\mathbf{p}) = h_i | H = h_j\right] \tag{11}$$

is strictly concave [7], [8], and thus we can define an M-ary Bayes risk error

$$d(\mathbf{p},\mathbf{a}) = -J(\mathbf{p}) + J(\mathbf{a}) + (\mathbf{p}-\mathbf{a})^T\nabla J(\mathbf{a}) \tag{12}$$

with all of the attendant properties of Bregman divergences mentioned above, and others such as the fact that partitions induced by Bregman divergences have linear separators [6].

Going even further, one may consider Bayesian estimation rather than M-ary Bayesian detection and define a functional Bregman divergence [12] as the distortion for quantization of regression prior functions.

D. Sipping From the Fountain

A primary motivation for considering the quantization of prior probabilities is as a model for decision making by humans. Also, high-rate quantization analysis of Bayes risk error in [2] follows that of perceptual distortion measures [13]; the potential connection between Bregman divergences and perceptual distortion measures is also pointed out in [5]. Thus, to further explore and model human information processing, study of Bregman divergences seems like a promising avenue.

Known results and theories regarding Bregman divergences can be specialized to human decision making and perception. Examples include connections to exponential family distributions [6], boosting [14], and information geometry [15]. (The exponential family distribution corresponding to Bayes risk error is

$$f_U(u) \propto \exp\left(-J(u,a) + J(u)\right), \tag{13}$$

with parameter a.) Wide-ranging connections among Bregman divergences, Ali–Silvey divergences, regret bounds, signal detection theory, and other topics in statistical learning and decision theories may be found in [16]. Considering the human aspect may lead to new directions of research that apply generally to Bregman divergences as well.

E. Paying It Forward

Work on quantization that builds upon [2] within the Bayes risk context may be expanded to include the entire class of Bregman divergences. For example, it seems possible to take the theory of team-theoretic quantization of prior probabilities for distributed detection by agents [17] and apply it to general team-theoretic quantization by agents to optimize a Bregman divergence-related objective. Similarly, it seems possible to transfer ideas of intermediate levels of minimax robustness for Bayes risk error [18] to general Bregman divergences. The concept of "price of segregation" developed in [19] and [20] might be applicable in non-Bayesian hypothesis testing scenarios. It may or may not be possible to generalize game-theoretic quantization of prior probabilities for distributed Bayesian hypothesis testing [21] because game-theoretic considerations arise in that context due to differences in Bayes costs c_{ij}, which have no direct analogue in other Bregman divergences.

V. Conclusion

In this correspondence, we have seen that Bayes risk error, a distortion criterion introduced for the quantization of prior probabilities in Bayesian hypothesis testing, is a Bregman divergence. Due to this nicety, several properties of the Bayes risk error can be stated immediately, new research directions exploring the role of general Bregman divergences in human information processing can be initiated, and existing analyses involving Bayes risk error can be generalized. As a divergence, Bayes risk error can be used to quantify the closeness of prior probability vectors not only for quantization, but in other informational contexts involving the signal detection task as well.

References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968.

[2] K. R. Varshney and L. R. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4553–4562, Oct. 2008.

[3] K. R. Varshney, "Frugal hypothesis testing and classification," Ph.D. thesis, Mass. Inst. Technol., Cambridge, MA, 2010.

[4] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comp. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217, 1967.

[5] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669, Jul. 2005.

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

[7] R. A. Wijsman, "Continuity of the Bayes risk," *Ann. Math. Statist.*, vol. 41, no. 3, pp. 1083–1085, Jun. 1970.

[8] M. H. DeGroot, *Optimal Statistical Decisions*. Hoboken, NJ: Wiley-Interscience, 2004.

[9] A. V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 187–198, Mar. 1982.

[10] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2882–2904, Jun. 2009.

[11] R. Sibson, "Information radius," *Probab. Theory Rel.*, vol. 14, no. 2, pp. 149–160, Jun. 1969.

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

[13] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1082–1091, May 1999.

[14] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, Adaboost and Bregman distances," *Mach. Learn.*, vol. 48, no. 1–3, pp. 253–285, Jul. 2002.

[15] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-boost and Bregman divergence," *Neural Comput.*, vol. 16, no. 5, pp. 1437–1481, Jul. 2004.

[16] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *J. Mach. Learn. Res.*, vol. 12, pp. 731–817, Mar. 2011.

[17] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Collaboration in distributed hypothesis testing with quantized prior probabilities," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2011, pp. 303–312.

[18] K. R. Varshney and L. R. Varshney, "Multilevel minimax hypothesis testing," in *Proc. IEEE Stat. Signal Process. Workshop*, Nice, France, Jun. 2011, pp. 109–112.

[19] L. R. Varshney, "Unreliable and resource-constrained decoding," Ph.D. thesis, Mass. Inst. Technol., Cambridge, MA, 2010.

[20] L. R. Varshney, J. B. Rhim, K. R. Varshney, and V. K. Goyal, "Categorical decision making by people, committees, and crowds," in *Proc. Inf. Theory Appl. Workshop*, La Jolla, CA, Feb. 2011.

[21] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Conflict in distributed hypothesis testing with quantized prior probabilities," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2011, pp. 313–322.