On Reliability Hypothesis Testing for Continuous Object with Quantization

Aram Yesayan
French University in Armenia (UFAR),
Yerevan, Armenia
e-mail: aram.yesayan@pers.ufar.am

Artak Martirosyan
French University in Armenia (UFAR),
Yerevan, Armenia
e-mail: artak.martirosyan@pers.ufar.am

***Abstract*—The problem of logarithmically asymptotically optimal hypothesis testing for a continuous random variable (CRV) is considered. It is assumed that M continuous probability distributions (CPDs) are known, and the object described by the CRV follows one of them.**

This problem, in the case of discrete probability distributions, was introduced and extensively studied in [1].

In this paper, a quantization method is applied to continuous distributions. Some known results are generalized and reformulated for the continuous case.

***Keywords*—Continuous probability distribution, Probability den-sity function, Hypothesis testing, Reliabilities.**

I. INTRODUCTION

The classical problem of statistical hypothesis testing involves two hypotheses [2]. A statistical decision procedure used to test hypotheses is called a *test*. The probability of incorrectly accepting one hypothesis instead of the other is known as the *error probability*.

We consider the case of a sequence of tests in which the error probabilities decay exponentially as 2^{-NE}, where N is the number of observations (i.e., the sample size) and E is the exponent of the error probability, called *reliability*.

The aim of this line of research is to determine the optimal functional relationship between the error exponents of the first and second kinds of errors. Such optimal tests were first studied by Hoeffding [3], and later by Csiszár and Longo [4], Tusnády [5], [6], Longo and Sgarro [7], Birgé [8] (who introduced the term *logarithmically asymptotically optimal*, or LAO), among others. Several authors [6], [9], [10] have also used the term *exponentially rate optimal* (ERO) to refer to this notion of testing.

Let X be a CRV taking values in a continuous set $\mathcal{X}$. Assume that the CPDs $P_1, P_2, \ldots, P_M$ are given and let the corresponding probability density functions (PDFs) be denoted by $g_m,\ m = 1, \ldots, M$. Let $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ be a random sample. Based on this sample, the statistician must make a decision among the following hypotheses:

$$H_1 : X \sim P_1, \quad H_2 : X \sim P_2, \quad \ldots, \quad H_M : X \sim P_M.$$

We approach the solution of this problem by referring to Haroutunian's result [1], proved in the case of discrete probability distributions using the Csiszár–Körner method of types. To apply the same technique in the continuous setting, we begin by quantizing the given CPDs.

Consider the partition $-\infty < a_1 < a_2 < \cdots < a_R < +\infty$ of the set $\mathcal{X}$. Let us define the quantization points $\mathcal{X}_R = \{c_r\}_{r=0}^{R}$, where each c_r is a representative point in the interval (a_r, a_{r+1}) (for example, the midpoint).

We define the quantized probability distributions (PDs) G_m^R, $m = \overline{1, M}$, on the finite set $\mathcal{X}_R$, corresponding to the PDFs g_m, by

$$G_m^R(c_r) = \int_{a_r}^{a_{r+1}} g_m(u)\, du, \quad \text{for } r = \overline{1, R-1}, \quad \text{where}$$

$$a_r < c_r < a_{r+1},$$

$$G_m^R(c_0) = \lim_{a \to -\infty} \int_{a}^{a_1} g_m(u)\, du,$$

$$G_m^R(c_R) = \lim_{b \to +\infty} \int_{a_R}^{b} g_m(u)\, du.$$

We assume that the PDFs g_m coincide with each other only at a finite number of points. Then, by choosing the partition size R sufficiently large, we can, without loss of generality, assume that all discrete probability distributions G_m^R are distinct.

For a sample $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, we construct the corresponding quantized sample $\mathbf{x}^R = (x_1^R, x_2^R, \ldots, x_N^R)$, where each quantized observation is defined by $x_n^R = c_r$ if $a_r \le x_n < a_{r+1}$.

For each fixed R, we consider a decision-making procedure in the form of a non-randomized test $\varphi^{(N,R)}$, which is defined by a partition of the sample space $\mathcal{X}_R^N$ into M disjoint subsets.

$$\mathcal{A}_l^{(N,R)} = \{\mathbf{x}^R \in \mathcal{X}_R^N : \varphi^{(N,R)}(\mathbf{x}^R) = l, l = \overline{1,M}\}.$$

The set $\mathcal{A}_l^{(N,R)}$ contains all vectors $\mathbf{x}^R \in \mathcal{X}_R^N$ for which the hypothesis $H_l^R : G^R = G_l^R$ is adopted. Therefore,

$$G_m^{(N,R)}(\mathbf{x}^R) \triangleq \prod_{n=1}^{N} G_m^R(x_n^R),$$

So, the probability $\alpha_{l|m}^{(N,R)} = \alpha_{l|m}^{(N,R)}(\varphi^{N,R})$ of the erroneous acceptance of the hypothesis H_l^R provided that H_m^R is true is

$$\alpha_{l|m}^{(N,R)} = G_m^{(N,R)}(\mathcal{A}_l^{(N,R)})$$

$$= \sum_{\mathbf{x}^R \in \mathcal{A}_l^{(N,R)}} G_m^{(N,R)}(\mathbf{x}^R),\ l \neq m.$$

We will use the following definition for the probability of rejection of the hypothesis H_m^R, when it is true, as

$$\alpha_{m|m}^{(N,R)} \triangleq \sum_{l \neq m} \alpha_{l|m}^{(N,R)},$$

and will consider the error probability exponents, which are also called reliabilities.

$$-\overline{\lim_{N\to\infty}} \frac{1}{N} \log \alpha_{l|m}^{(N,R)} \triangleq E_{l|m}^R \geq 0,\ m, l = \overline{1,M}.$$

It is known that

$$E_{m|m}^R = \min_{l \neq m} E_{l|m}^R.$$

The matrix $E^R = \{E_{l|m}^R(\varphi^R)\}$ is called the reliability matrix of the sequence of tests φ^R.

If we consider a discrete random variable defined on finite sets, we will use the same notation as above, omitting the index R for all corresponding quantities.

II. Generalization of Some Results

We generalize the concept of reliability for the continuous case in the following way:

$$E_{l|m} \triangleq \lim_{R\to\infty} E_{l|m}^R, m, l = \overline{1,M}.$$

We will use the (Kullback - Leibler) divergence $D(G\|Q)$ for discrete probability distributions (PDs) G and Q, as usual [12], [13]:

$$D(G\|Q) = \sum_x G(x) \log \frac{G(x)}{Q(x)},$$

and for PDFs f and g, the notation $K(f \parallel g) = \int_{\mathcal{X}} f log \frac{f}{g}$.

We now formulate and prove the main theorem of this article, the application of which will solve the proposed problem.

Theorem: Let F^R and G^R be the R-quantizations of the PDFs f and g, respectively, and suppose that the Kullback–Leibler divergence $K(f \parallel g) < \infty$. Then

$$D(F^R \parallel G^R) \to K(f \parallel g), \quad \text{as} \quad R \to \infty.$$

Proof: By the mean value theorem, there exist values c_r^f and c_r^g such that

$$F^R(c_r) = f(c_r^f)\Delta_R \quad \text{and} \quad G^R(c_r) = g(c_r^g)\Delta_R.$$

The divergence of the quantized versions is

$$D(F^R \parallel G^R) = \sum_{r=0}^{R} F^R(c_r) \log \frac{F^R(c_r)}{G^R(c_r)}$$

$$= \sum_{r=0}^{R} f(c_r^f)\Delta_R \log \frac{f(c_r^f)\Delta_R}{g(c_r^g)\Delta_R} = \sum_{r=0}^{R} f(c_r^f)\Delta_R \log \frac{f(c_r^f)}{g(c_r^g)}.$$

Then, from the inequality

$$\min_{c\in\{c_r^f, c_r^g\}} \sum_{r=0}^{R} f(c)\Delta_R \log \frac{f(c)}{g(c)} \leq D(F^R \parallel G^R)$$

$$\leq \max_{c\in\{c_r^f, c_r^g\}} \sum_{r=0}^{R} f(c)\Delta_R \log \frac{f(c)}{g(c)},$$

and since $K(f \parallel g) < \infty$, the function under the sum is Riemann integrable, so

$$\sum_{r=0}^{R} f(c)\Delta_R \log \frac{f(c)}{g(c)}$$

is a Riemann sum that approximates $K(f \parallel g)$. Hence, it follows that

$$D(F^R \parallel G^R) \to K(f \parallel g), \quad \text{as} \quad R \to \infty.$$

So, the theorem is proved.

III. LAO Testing of Many Hypotheses for Continuous Object

In this section, we consider the LAO hypothesis testing problem in the context of R-quantization.

A sequence of tests φ^{R*} is said to be logarithmically asymptotically optimal (LAO) if, for the given positive values of the first $M-1$ diagonal elements of the matrix $\mathbf{E}(\varphi^{R*})$, the remaining elements achieve their maximal possible values.

Let $N(x|\mathbf{x})$ be the number of repetitions of the element $x \in \mathcal{X}_R$ in the vector $\mathbf{x} \in \mathcal{X}_R^N$, and let

$$Q_{\mathbf{x}^R} = \{N(x|\mathbf{x})/N, x \in \mathcal{X}_R\}$$

be the empirical distribution (type) of the sample $\mathbf{x}$.

For the given positive diagonal elements $E^R_{1|1}, E^R_{2|2}, \ldots, E^R_{M-1|M-1}$ of the reliability matrix, we consider sets of PDs

$$\mathcal{R}^R_l \stackrel{\triangle}{=} \{Q^R : D(Q^R||G^R_l) \leq E^R_{l|l}\}, \quad l = \overline{1, M-1}, \qquad (1)$$

$$\mathcal{R}^R_M \stackrel{\triangle}{=} \{Q^R : D(Q^R||G^R_l) > E^R_{l|l}, \quad l = \overline{1, M-1}\} \qquad (2)$$

and define the values for the elements of the future reliability matrix of the LAO tests sequence as follows:

$$E^{R*}_{l|l} = E^{R*}_{l|l}(E^R_{l|l}) \stackrel{\triangle}{=} E^R_{l|l}, \quad l = \overline{M-1}, \qquad (3)$$

$$E^{R*}_{l|m} = E^{R*}_{l|m}(E^R_{l|l}) \stackrel{\triangle}{=} \inf_{Q^R \in \mathcal{R}^R_l} D(Q^R||G^R_m), \; m = \overline{1, M}, \quad m \neq l, \; l = \overline{1, M-1}, \qquad (4)$$

$$E^{R*}_{M|m} = E^{R*}_{M|m}(E^R_{1|1}, \ldots, E^R_{M-1|M-1}) \stackrel{\triangle}{=} \inf_{Q^R \in \mathcal{R}^R_M} D(Q^R||G^R_m), \; m = \overline{1, M-1}, \qquad (5)$$

$$E^{R*}_{M|M} = E^{R*}_{M|M}(E^R_{1|1}, \ldots, E^R_{M-1|M-1}) \stackrel{\triangle}{=} \min_{l=\overline{1,M-1}} E^{R*}_{M|l}. \qquad (6)$$

Thus, we can reformulate Haroutunian's theorem, the result of which provides the optimal relationships between the reliabilities.

Theorem [1]: *If the distributions G^R_m are distinct, that is, all Kullback–Leibler divergences $D(G^R_l \parallel G^R_m)$ for $l \neq m$, $l, m = \overline{1, M}$, are strictly positive, then the following two statements hold:*

a) If the given numbers $E^R_{1|1}, E^R_{2|2}, \ldots, E^R_{M-1|M-1}$ satisfy the conditions

$$0 < E^R_{1|1} < \min_{l=\overline{2,M}} D(G^R_l||G^R_1), \qquad (7)$$

$$0 < E^R_{m|m} < \min[\min_{l=\overline{1,m-1}} E^{R*}_{l|m}(E^R_{l|l}), \min_{l=\overline{m+1,M}} D(G^R_l||G^R_m)], \quad m = \overline{2, M-1}, \qquad (8)$$

for all R, then there exists an LAO sequence of tests φ^{R*} such that the elements of its reliability matrix $\mathbf{E}(\varphi^{R*}) = \{E^{R*}_{l|m}\}$, defined in (3)–(6), are all strictly positive.

b) Conversely, if even one of the conditions (7) or (8) is violated, then the reliability matrix of any such test contains at least one zero element.

IV. Conclusion

In this paper, a suitable hypothesis testing strategy is discussed for the model of one object with M known CPDs. The optimal relationships among the reliabilities are established.

References

[1] E. Haroutunian, "Logarithmically asymptotically optimal testing of multiple statistical hypotheses," *Problems of Control and Information Theory*, vol. 19, no. 5–6, pp. 413-421, 1990.

[2] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, Third Edition. Springer, New York, 2005.

[3] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions", *The Annals of Mathematical Statistics*. vol. 36, pp. 369–401, 1965.

[4] I. Csiszár and G. Longo, "On the error exponent for source coding and for testing simple statistical hypotheses", *Studia Sc. Math. Hungarica*, vol. 6, pp. 181–191, 1971.

[5] G. Tusnády, "On asymptotically optimal tests", *Annals of Statatistics*, vol. 5, no. 2, pp. 385-393, 1977.

[6] G. Tusnády, "Testing statistical hypotheses (an information theoretic approach)", *Preprint of the Mathematical Institute of the Hungarian Academy of Sciences*, Budapest (Part 1, 1979, Part 2, 1982).

[7] G. Longo and A. Sgarro, "The error exponent for the testing of simple statistical hypotheses: A combinatorial approach", *Journal of Combinatorics, Information and System Sciences*, vol. 5, no. 1, pp. 58-67, 1980.

[8] L. Birgé, "Vitesses maximals de dćroissence des erreurs et tests optimaux associeś". *Z. Wahrsch. Verw. Gebiete*, vol. 55, pp. 261–273, 1981.

[9] M. Feder and N. Merhav, "Universal composit hypothesis testing: a competititve minimax approach", *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1504-1517, 2002.

[10] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications", *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2215-2229, 2002.

[11] R. Ahlswede and E. Haroutunian, "On logarithmically asymptotically optimal testing of hypotheses and identification", *Lecture Notes in Computer Science*, vol. 4123, "General Theory of Information Transfer and Combinatorics", Springer, pp. 462– 478, 2006.

[12] E. Haroutunian, M. Haroutunian, A. Harutyunyan, Reliability criteria in information theory and in statistical hypothesis testing, *Foundations and Trends in Communications and Information Theory*, vol. 4, Issue 2, pp. 97-263, February 2007.

[13] T. Cover, J. Thomas, *Elements of information theory*, Second Edition, New Jersey, 2006.