HHS Public Access

Author manuscript

IEEE Trans Inf Theory. Author manuscript; available in PMC 2019 October 11.

Published in final edited form as:

IEEE Trans Inf Theory. 2008 July ; 54(7): 3327–3329. doi:10.1109/TIT.2008.924656.

Asymptotic Geometry of Multiple Hypothesis Testing

M. Brandon Westover

Department of Neurology, Massachusetts General Hospital, Boston, MA 02114-2622 USA

Abstract

We present a simple geometrical interpretation for the solution to the multiple hypothesis testing problem in the asymptotic limit. Under this interpretation, the optimal decision rule is a nearest neighbor classifier on the probability simplex.

Index Terms—

Geometry; hypothesis testing; large deviations; pattern recognition

I. Introduction

In the binary hypothesis testing problem with independent and identically distributed (i.i.d.) observations, it is well known that the error probability for the optimal decision rule decays with a constant exponential rate equal to the Chernoff distance between the two hypothesis distributions. The generalization to multiple hypothesis testing for i.i.d. observations was derived by Leang and Johnson [1], and extended for observations modeled as stationary ergodic processes by Schmid and O'Sullivan [2]. In this correspondence, we focus on the i.i.d. case. For M-ary hypothesis testing, the error probability decays exponentially with a rate equal to the minimum Chernoff distance between all distinct pairs of hypothesis distributions. In this correspondence, we describe a simple geometrical interpretation of this result, illustrated in Fig. 2. We first review the binary case.

II. Binary Hypothesis Testing

Let π_i, $i = 1, 2$ be the prior probabilities for the hypotheses H_1 and H_2; and let $Y^n = (Y_1, \ldots, Y_n)$ be an i.i.d. sequence of observations from a finite alphabet $\mathcal{Y}$. We must decide between

$$H_1 : Y^n \sim P_1$$

and

$$H_2 : Y^n \sim P_2 .$$

(mb.westover@gmail.com).

A. Optimal Decision Rule and Error Exponent

The following results are standard (see, e.g., [3, Ch. 12]). Letting $\pi_1 = \pi_2 = 1/2$, the optimal decision $\widehat{H} = \widehat{H}(Y^n)$ rule can be expressed in terms of the Kullback–Leibler (KL)-divergence as

$$\widehat{H} = \underset{i \in \{1,2\}}{\operatorname{argmin}} D\left(P_{Y^n} \middle\| P_i\right) \quad (1)$$

where P_{Y^n} is the type (empirical histogram) of Y^n. This decision rule partitions the probability simplex $\mathcal{P}$ over $\mathcal{Y}$ into two disjoint decision regions A_1, A_2. Denoting their respective complements $A_1^c = A_2$, $A_2^c = A_1$, an application of Sanov's theorem shows that the error probabilities are

$$\begin{aligned} P_1^n(A_1^c) &\doteq 2^{-nD\left(P_1^* \| P_1\right)} \\ P_2^n(A_2^c) &\doteq 2^{-nD\left(P_2^* \| P_2\right)} \end{aligned} \quad (2)$$

where

$$P_i^* \triangleq \underset{\boldsymbol{p} \in A_i^c}{\operatorname{argmin}} D\left(\boldsymbol{p} \| P_i\right), \quad i = 1, 2$$

and $\doteq$ denotes "equality to first order in the exponent," i.e., $a_n \doteq b_n$ means $\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$ (see [3, p. 55]). For the total probability of error P_e^n we have

$$\begin{aligned} P_e^n &= \pi_1 P_1^n(A_1^c) + \pi_2 P_2^n(A_2^c) \\ &\doteq \pi_1 2^{-nD\left(P_1^* \| P_1\right)} + \pi_2 2^{-nD\left(P_2^* \| P_2\right)} \\ &\doteq 2^{-n \min\left\{D\left(P_1^* \| P_1\right), D\left(P_2^* \| P_2\right)\right\}} \\ &\triangleq 2^{-nC\left(P_1, P_2\right)}. \end{aligned} \quad (3)$$

The exponent

$$C\left(P_1, P_2\right) = \min\left\{D\left(P_1^* \| P_1\right), D\left(P_2^* \| P_2\right)\right\}$$

is the Chernoff distance between P_1 and P_2, usually written in the alternative but equivalent form

$$C(P_1, P_2) = \min_{\lambda \in [0,1]} \log\left(\sum_y P_1^\lambda(y) P_2^{\lambda-1}(y)\right).$$

B. Geometric Interpretation

Fig. 1 illustrates the preceding results for alphabet size $|\mathcal{Y}| = 3$, with corresponding probability simplex

$$\mathcal{P} = \{\boldsymbol{p} : p_1, p_2, p_3 \geq 0, p_1 + p_2 + p_3 = 1\}. \quad (4)$$

In this illustration, each probability distribution $\boldsymbol{p} \in \mathcal{P}$ is mapped to a point $\boldsymbol{r} \in \mathbb{R}^2$ in an equilateral triangle with vertices at points $\boldsymbol{r} = (x, y) \in \mathbb{R}^2 : (0, 1), (\sqrt{3}/2, -1/2)$, and $(-\sqrt{3}/2, -1/2)$, using the transformation $\boldsymbol{r} = \boldsymbol{T}\boldsymbol{p} + \boldsymbol{b}$, where

$$\boldsymbol{T} = \begin{pmatrix} -1/\sqrt{3} & -1/3 \\ 1/\sqrt{3} & -1/3 \\ 0 & 2/3 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

Given an observation sequence Y^n, the decision rule (1) can be interpreted geometrically in terms of the following algorithm:

- Locate the observation's type P_{Y^n} in the simplex $\mathcal{P}$;
- Compute the distance to each hypothesis distribution $D\left(P_{Y^n} \| P_1\right)$, $D\left(P_{Y^n} \| P_2\right)$;
- Assign Y^n to the closest hypothesis.

$\widehat{H}$ thus divides $\mathcal{P}$ into two disjoint cells,[1] with centroids P_1 and P_2, and each point $\boldsymbol{p} \in \mathcal{P}$ assigned to the nearest centroid, with "distance" measured in KL-divergence.

Next consider (2). Under hypothesis $H_1 : Y^n \sim P_1$, the probability of incorrectly deciding $Y^n \sim \mathcal{P}_2$, that is, $P_1^n(A_1^c)$, depends on the distance from P_1 to the closest distribution outside of A_1. This distribution P_1^* can be expressed as a point along the geodesic in $\mathcal{P}$ joining P_1 and P_2

$$\boldsymbol{p}(\lambda) = \frac{1}{Z(\lambda)} P_1^{1-\lambda} P_2^\lambda, \quad \lambda \in [0, 1] \quad (5)$$

where exponentiation is defined pointwise (e.g., if $|\mathcal{Y}| = 3$, then $\boldsymbol{p}^\lambda = \left(p_1^\lambda, p_2^\lambda, p_3^\lambda\right)$. Conceptually, the value λ^* for which $\boldsymbol{p}(\lambda^*) = P_1^*$ can be found by setting $\lambda = 0$, then

[1] It is tempting to call these "Voronoi" cells, but strictly speaking they are not, because the KL-divergence is not a true distance metric.

continuously increasing λ until the point $\boldsymbol{p}(\lambda)$ reaches the border separating regions A_1 and A_2. Symmetric arguments apply for P_2^* and, by symmetry, we must have $P_1^* = P_2^*$.

Finally, interpreting (3), the overall probability of error P_e^n is determined by measuring the distance along the geodesic from either centroid P_1 or P_2 to its border. The shortest of the two paths asymptotically dominates the overall probability of error.

III. *M*-Ary Hypothesis Testing

Now consider the M-ary hypothesis testing problem. Let $\boldsymbol{\pi}_i$ be the prior probability for hypothesis H_i. We must decide among

$$H_i : Y^n \sim P_i, \quad i = 1, \ldots, M .$$

As above, we assume equal prior probabilities, $\boldsymbol{\pi}_i = 1/M$, $i = 1, \ldots, M$.

An important special case is the following formulation of the general statistical pattern recognition problem [4]: Given a set of M template patterns $\mathscr{C} = \{x^n(1), \ldots, x^n(M)\}$, suppose Nature selects one of the pattern templates $w \in \{1, \ldots, M\}$ at random with probability $p(w) = 1/M$, $w = 1, \ldots, M$. The observation data Y^n is generated from the template $x^n(w)$ according to $P_w = p\left(y^n \mid x^n(w)\right) = \prod_{j=1}^{n} p\left(y_j \mid x_j(w)\right)$. In this case, the optimal hypothesis test $\widehat{w}$ is a classifier, i.e., a rule that infers the pattern class underlying the observation Y^n.

A. Optimal Decision Rule and Error Exponent

The natural generalizations of (1)–(3) to the M-ary case, derived by Leang and Johnson [1], are as follows. The optimal decision rule is

$$\widehat{H} = \underset{i \in \{1, \ldots, M\}}{\operatorname{argmin}} \; D\left(P_{Y^n} \middle\| P_i\right). \quad (6)$$

Denoting the optimal decision regions by A_i, and their complements by A_i^c, applying Sanov's theorem yields for the error probabilities under each hypothesis H_i, $i = 1, \ldots, M$

$$P_i^n\left(A_i^c\right) \doteq 2^{-n\, D\left(P_i^* \| P_i\right)} \quad (7)$$

$$P_i^* \triangleq \underset{\boldsymbol{p} \in A_i^c}{\operatorname{argmin}} D\left(\boldsymbol{p} \middle\| P_i\right).$$

Hence, for the total probability of error P_e^n we have

$$P_e^n = \sum_{i=1}^{M} \pi_i P_i^n\left(A_i^c\right) \qquad (8)$$
$$\doteq \sum_{i=1}^{M} \pi_i 2^{-nD\left(P_i^* \| P_i\right)}$$
$$\doteq 2^{-n \min_i D\left(P_i^* \| P_i\right)}$$
$$= 2^{-nC\left(P_w, P_{w'}\right)}$$

where

$$(w, w') = \underset{i \neq j}{\operatorname{argmin}}\, C\left(P_i, P_j\right)$$

i.e., P_w, $P_{w'}$ is the closest pair of distributions, measured in Chernoff distance.

B. Geometric Interpretation

Now consider the geometric interpretation of the preceding results for M-ary hypothesis testing. Fig. 2 represents the M-ary generalization of Fig. 1.

Given an observation sequence Y^n, the optimal decision rule (6) can be interpreted geometrically in terms of the following algorithm:

- Locate the observation's type P_{Y^n} in the simplex $\mathcal{P}$;
- Compute the distance to each hypothesis distribution $D\left(P_{Y^n} \| P_i\right), i = 1, \ldots, M$;
- Assign Y^n to the closest hypothesis.

The rule $\widehat{H}$ thus divides $\mathcal{P}$ into M cells, with centroids P_i, $i = 1, \ldots, M$, with each point $p \in \mathcal{P}$ assigned to the nearest centroid.

These cells are convex: Given a cell A_i with centroid P_i and two points $Q_1, Q_2 \in A_i$, let $Q = \lambda Q_1 + (1 - \lambda) Q_2$, $\lambda \in [0, 1]$, let $P_j \in A_j$ be any other centroid, and define

$$\Delta\left(Q_k\right) \triangleq D\left(Q_k \| P_i\right) - D\left(Q_k \| P_j\right)$$
$$= E_{Q_k} \log \frac{P_j}{P_i}, k = 1, 2.$$

The condition $Q_1, Q_2 \in A_i$ is equivalent to requiring $\Delta(Q_k) < 0$, $k = 1, 2$. Hence, noting that $\Delta(Q_k)$ is linear in Q_k, we have

$$\Delta(Q) = \lambda\Delta\left(Q_1\right) + (1-\lambda)\Delta\left(Q_2\right) < 0$$

whence $Q \in A_i$, establishing the convexity of A_i. A similar argument shows that the borders of the decision cells are composed of straight lines: Q_1, Q_2 are on the boundary between A_i, A_j, if and only if $\Delta(Q_1) = \Delta(Q_2) = 0$, hence, $\Delta(Q) = \lambda\Delta(Q_1) + (1 - \lambda)\Delta(Q_2) = 0$.

Turning to (7), under each hypothesis H_i the corresponding probability of error $P_i\left(A_i^c\right)$ is again determined by the distance from the centroid P_i of A_i to the nearest point outside A_i. However, in the M-ary case, each complement A_i^c consists of $M - 1$ other decision regions, and P_1^* lies on the geodesic joining P_i with the closest neighboring region to A_i. Note that, unlike the binary case, for two neighboring cells A_i, A_j, it is not necessarily the case that $P_i^* = P_j^*$, although this is the case when geodesics intersect a shared border.

Finally, consider (8). To determine the overall probability of error P_e^n, we compare the geodesic distance from each centroid to its nearest border, $D\left(P_i^* \| P_i\right)$. The overall probability of error is dominated by the shortest such path, of length $D(P_w \| P_{w'})$.

IV. Conclusion

We have described a simple, easily remembered geometrical interpretation of the results of [1] for the multiple hypothesis testing problem in the asymptotic regime. Under this interpretation, the optimal decision rule is a nearest neighbor classifier, with “distance” measured in terms of the KL-divergence between the observation data’s type and each hypothesis distribution. That is, the optimal decision rule splits the probability simplex into M cells, and assigns each observation to the nearest centroid. The error probability under each hypothesis is determined by the geodesic distance from the centroid of its cell to the nearest cell border. The overall probability of error is determined by the smallest such distance.

Acknowledgment

The author wishes to thank Joseph A. O’Sullivan for insightful comments.

References

[1]. Leang C and Johnson DH, “On the asymptotics of M-hypothesis Bayesian detection,” IEEE Trans. Inf. Theory, vol. 43, no. 1, pp. 280–282, Jan. 1997.

[2]. Schmid NA and O’Sullivan JA, “Performance prediction methodology for biometric systems using a large deviations approach,” IEEE Trans. Signal Process, vol. 52, no. 10, pp. 3036–3045, Oct. 2004.

[3]. Cover TM and Thomas J, Elements of Information Theory. New York: Wiley, 1991.

[4]. Westover MB and O’Sullivan JA, “Achievable rates for pattern recognition,” IEEE Trans. Inf. Theory, vol. 54, pp. 299–320, Jan. 2008.

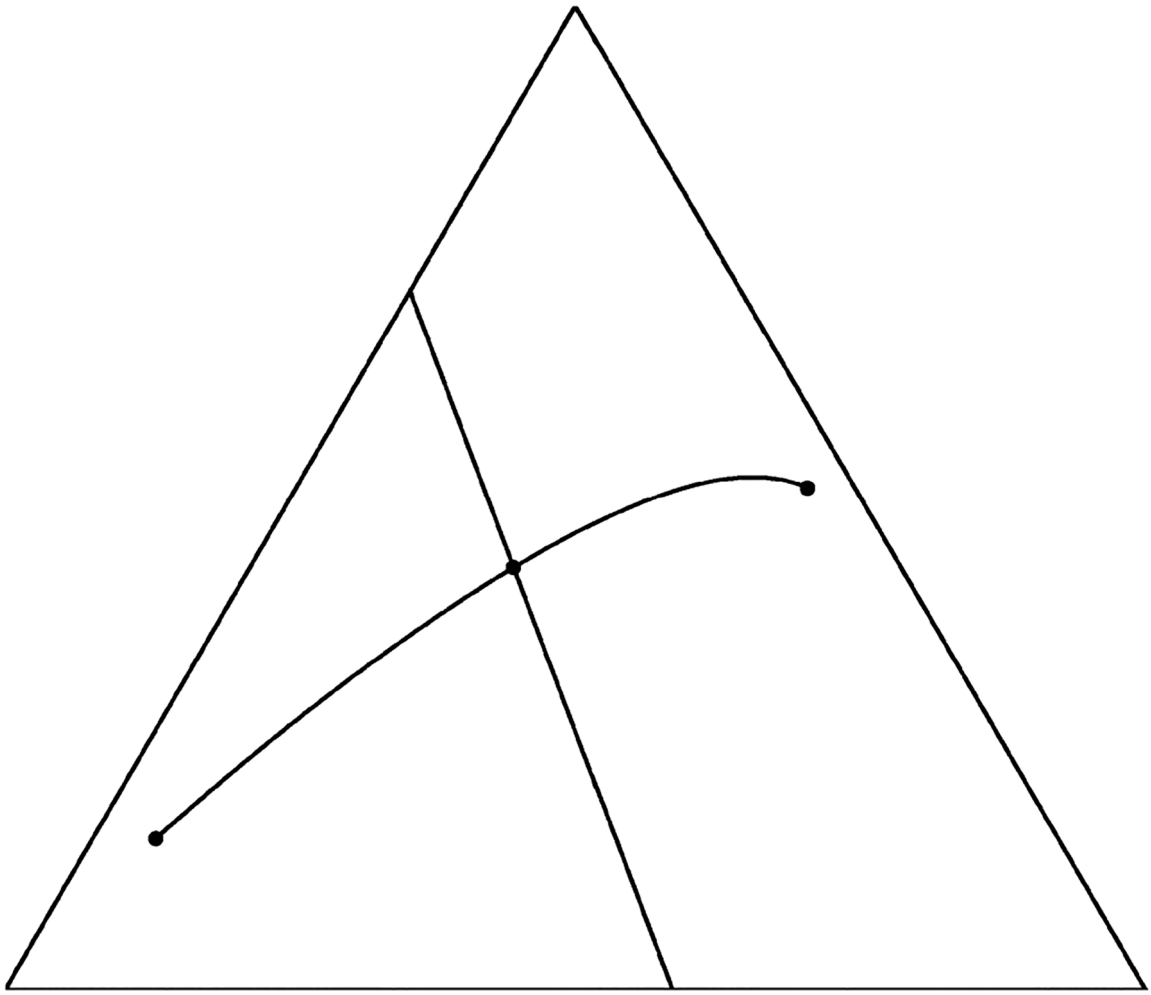

Fig. 1.
Binary hypothesis testing.

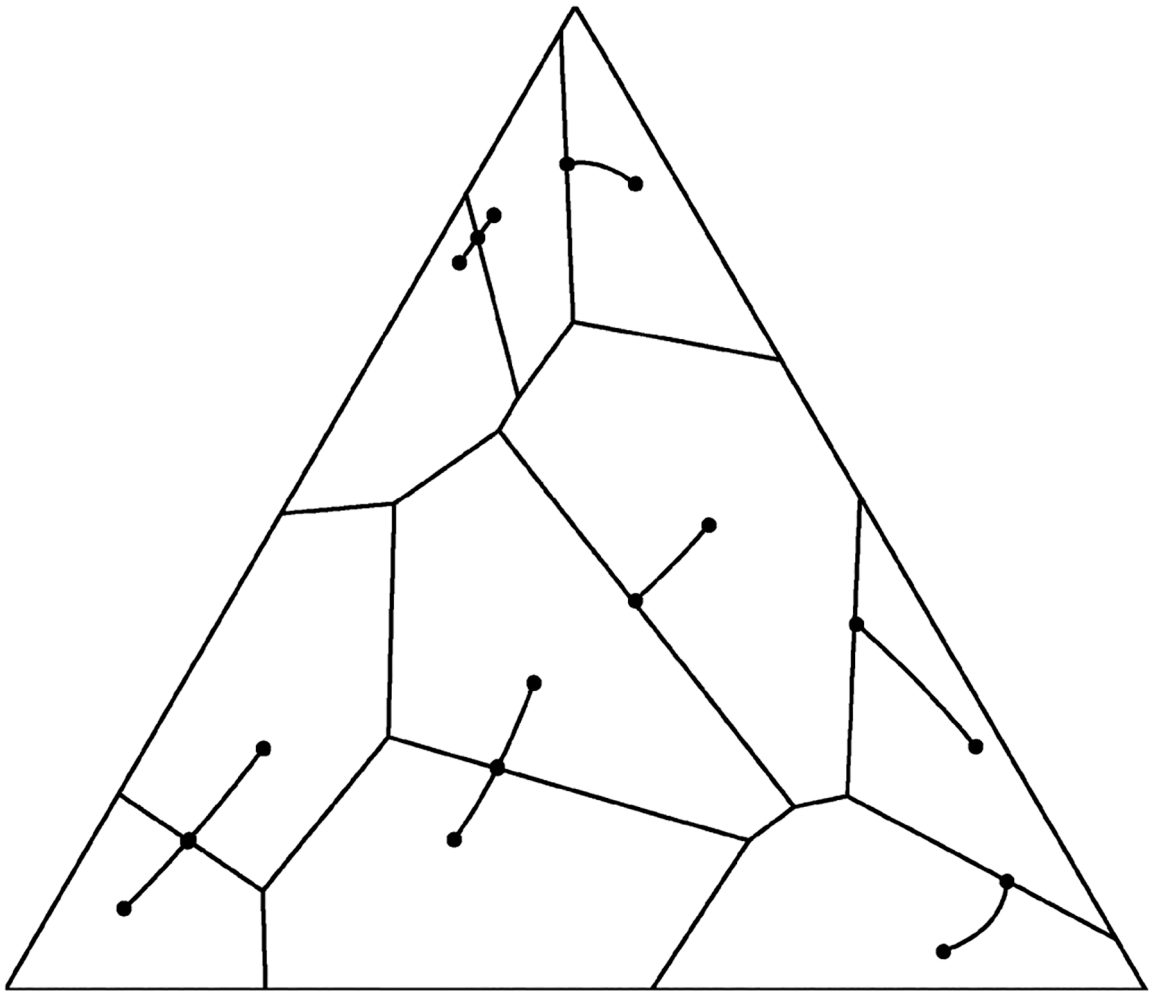

Fig. 2.
M-ary hypothesis testing.