DATA-DRIVEN NONPARAMETRIC HYPOTHESIS TESTING

Yixian Liu[⋆] *Yingbin Liang* [↑] *Shuguang Cui*[∗]

[⋆]Shanghai Institute of Microsystem & Information Technology, Chinese Academy of Sciences
[⋆]University of Chinese Academy of Sciences
[⋆]School of Information Science and Technology, ShanghaiTech University
[↑]Department of Electrical and Computer Engineering, The Ohio State University
[∗]Department of Electrical and Computer Engineering, University of California at Davis
[∗]Shenzhen Research Institute of Big Data

ABSTRACT

We investigate a nonparametric hypothesis testing problem, in which we assume a testing data stream is generated by one of a set of distributions (hypotheses), and the goal is to test which one of the multiple distributions generates the testing data stream, i.e., which hypothesis occurs. We assume that some distributions in the set are unknown with only training sequences generated by the corresponding distributions are given. We construct the generalized likelihood (GL) test, and characterize the error exponent of the maximum error probability. We show that the error exponent is captured by the Chernoff distance between each pair of distributions as well as the KL divergence between the approximated distributions (via training sequences) and the true distributions. We also show that the ratio between the lengths of training and testing sequences plays an important role in determining the error decay behavior.

***Index Terms*—** Multiple hypothesis testing, generalized likelihood test, error exponent, KL divergence

1. INTRODUCTION

In this paper, we consider a nonparametric hypothesis testing problem. We assume that there are a set of M distinct discrete distributions $p_1, ..., p_M$, and a training data stream that consists of data samples drawn from each distribution is available if the corresponding distribution is unknown. Furthermore, a testing data stream is observed, which consists of n samples drawn from one of the M distributions. The goal is to determine which distribution generates the testing data stream.

For parametric scenarios, where all distributions $p_1, ..., p_M$ are known in advance, this problem has been well studied, e.g., [1, 2]. For nonparametric scenarios, where the distributions are unknown, but instead, a training data stream generated from each distribution is given, previous studies [3–5] focused only on the Neyman-Pearson formulation, i.e., given the requirement on the error probability for some hypotheses, the error probability for the remaining hypotheses needs to be minimized. The focus of this paper is to solve the problem in the nonparametric case based on a different performance criterion: the maximum of all types of error probabilities, as a multiple hypothesis testing problem. Our focus is on the characterization of the error exponent for such an error performance metric as the sample size enlarges. Our study suggests that such a different performance criterion offers very different understanding and insights about this problem.

Our contributions are summarized as follows. 1) We construct a generalized likelihood (GL) test for the nonparametric hypothesis testing problem, and characterize the error exponent for this problem. 2) We show that the error exponent is captured by the Chernoff distance between each pair of distributions as well as the KL divergence between the approximated distributions (via training sequences) and the true distributions. 3) We show that the ratio β between the lengths of training and testing sequences plays an important role in determining the error decay behavior: If $\beta \to \infty$, the error exponent of the considered nonparametric model converges to that of the parametric problem; if $0 < \beta < \infty$, the GL test is exponentially consistent (i.e., the error exponent is positive); and finally, if $\beta \to 0$, the test is not exponentially consistent.

Our problem is related to but different from the following models recently studied. One type of anomalous sample detection problems was studied in [6–8], in which given a training set of samples generated by one (or more) typical distributions, a new sample needs to be tested whether it is generated from the typical distributions or from an anomalous distribu-

The work of Y. Liu was supported by University of Chinese Academy of Sciences under UCAS Joint PhD Training Program UCAS[2015]37.

Y. Liu would like to thank Dr. Zhi Ding for his support of Yixian Liu's visit to Syracuse University, where this work was performed.

The work of Y. Liang was supported in part by DARPA FunLoL program and by NSF grants ECCS-1818904 and CCF-1801855.

The work of S. Cui was supported in part by grant NSFC-61629101, by NSF with grants DMS-1622433, AST-1547436, ECCS-1659025, and CNS-1343155, by DoD with grant HDTRA1-13-1-0029, and by Shenzhen Fundamental Research Fund under Grant No. KQTD2015033114415450.

tion. Our study focuses on characterizing the error exponent, whereas the previous studies do not analyze the error exponent. Our problem is also related but different from a class of outlying sequence detection problems studied in [9–12]. Our model has training sequences associated with hypotheses, whereas the previous results did not consider such information.

2. PROBLEM FORMULATION

Suppose there are in total M distinct discrete distributions $p_1, ..., p_M$ with the support set $\mathcal{Y}$. We assume a general hypothesis testing model, in which the distributions p_i's for $1 \leqslant i \leqslant M_1$ are known, and the distributions p_i's for $M_1 < i \leqslant M$ are unknown, where $0 < M_1 \leqslant M$. Clearly, if $M_1 = M$, the problem is fully parametric with all distributions known. If $M_1 = 0$, the problem is fully nonparametric with all distributions are unknown. Thus, the model we study here unifies the parametric, semiparametric and nonparametric models; but we refer to such a model as a nonparametric model for simplicity, to which we make our main technical contributions. For each unknown distribution p_i with $i > M_1$, a training sequence $\boldsymbol{t}_i$ is available, which consists of $\bar{n}$ i.i.d. samples generated by p_i. Furthermore, a testing sequence $\boldsymbol{y}$ is observed, which consists of n i.i.d. samples generated by one of the M distributions, say, p_s. Our problem is to determine which distribution generates the testing sequence. Equivalently, this problem can be viewed as the following multiple hypothesis testing problem

$$H_i\text{: } p_s = p_i, \text{ for } i = 1, ..., M,$$

where the goal is to determine which hypothesis occurs.

We let $\sigma(\boldsymbol{y})$ denote a test rule, which maps the testing sequence $\boldsymbol{y}$ to one of the M hypotheses. Then, we take the following maximum of M error probabilities as the performance metric:

$$P_e(\sigma) = \max_{1 \leqslant i \leqslant M} P(\sigma \neq H_i | H_i). \tag{1}$$

We further define the error exponent of $P_e(\sigma)$ as

$$e(\sigma) = \lim_{n \to \infty} -\frac{1}{n} \log P_e(\sigma). \tag{2}$$

A test σ is *consistent* if $P_e(\sigma)$ converges to zero as n goes to infinity:

$$\lim_{n \to \infty} P_e(\sigma) = 0, \tag{3}$$

and a test σ is *exponentially consistent* if $P_e(\sigma)$ converges to zero exponentially fast with respect to n, i.e., $e(\sigma) > 0$.

3. MAIN RESULTS

In this section, we construct the test rule and analyze the performance of the test. We also discuss computational issues of the error exponent. All proofs are omitted due to the page limitations, and can be referred to [13].

3.1. Theoretical Analysis

The problem of the *parametric* multiple hypothesis testing problem has been well studied in the literature. It has been shown (see [1, 2]) that the following maximum likelihood test achieves the optimal exponent of the maximum error probability:

$$\sigma(\boldsymbol{y}) = \arg\max_i P(\boldsymbol{y}|H_i). \tag{4}$$

We here focus on the nonparametric case. We construct a generalized maximum likelihood test by replacing each unknown distribution p_i (for $i > M_1$) in (4), corresponding to hypothesis H_i, with the empirical distribution of the training sequence $\boldsymbol{t}_i$. The resulting test is given by:

$$\sigma(\boldsymbol{y}) = \arg\max_i \left\{ \begin{array}{ll} P(\boldsymbol{y}|p_i), & \text{if } i \leqslant M_1 \\ P(\boldsymbol{y}|\gamma(\boldsymbol{t}_i)), & \text{if } i > M_1 \end{array} \right\}, \tag{5}$$

where

$$P(\boldsymbol{y}|p_i) = \exp\left\{-nH(\gamma(\boldsymbol{y})) - nD(\gamma(\boldsymbol{y})\|p_i)\right\}, \tag{6}$$

for $1 \leqslant i \leqslant M_1$, $\gamma(\boldsymbol{y})$ denotes the empirical distribution of $\boldsymbol{y}$ given by

$$\gamma(y) \triangleq \frac{\text{number of samples } y \text{ in } \boldsymbol{y}}{\text{length of } \boldsymbol{y}},$$

$H(\cdot)$ denotes the entropy given by

$$H(p) = \sum_{y \in \mathcal{Y}} p(y) \log p(y), \tag{7}$$

and $D(\cdot\|\cdot)$ denotes the KL divergence given by

$$D(p\|q) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}. \tag{8}$$

Applying (6), test (5) is equivalent to

$$\sigma(\boldsymbol{y}) = \arg\min_i \left\{ \begin{array}{ll} D(\gamma(\boldsymbol{y})\|p_i), & \text{if } i \leqslant M_1 \\ D(\gamma(\boldsymbol{y})\|\gamma(\boldsymbol{t}_i)), & \text{if } i > M_1 \end{array} \right\}. \tag{9}$$

Since the term $H(\gamma(\boldsymbol{y}))$ does not depend on the distribution p_i, it is dropped from the maximum likelihood test. The following theorem characterizes that the performance of test (9) is determined by Chernoff information defined as

$$C(p, q) = \max_{\lambda \in [0,1]} -\log\Big(\sum_{y \in \mathcal{Y}} p(y)^{\lambda} q(y)^{1-\lambda}\Big). \tag{10}$$

Theorem 1. *Apply test* (9) *to the nonparametric multiple hypothesis testing problem. The error exponent of the maximum error probability is given by*

$$\min_{i,j: i \neq j} e_{i,j},$$

where $e_{i,j}$ is given as follows.

- *For $i \leqslant M_1$ and $j \leqslant M_1$, $e_{i,j} = C(p_i, p_j)$;*
- *For $i \leqslant M_1$ and $j > M_1$,*

$$e_{i,j} = \min_{q, q_j \in \Delta} \quad D(q\|p_j) + \beta D(q_j\|p_j) \qquad (11)$$
$$s.t. \quad D(q\|q_j) \geqslant D(q\|p_i),$$

where $\Delta = \{q : \sum_{y\in\mathcal{Y}} q(y) = 1, 0 \leqslant q(y) \leqslant 1\}$;

- *For $i > M_1$ and $j \leqslant M_1$,*

$$e_{i,j} = \min_{q_i \in \Delta} \quad C(q_i, p_j) + \beta D(q_i\|p_i); \qquad (12)$$

- *For $i > M_1$ and $j > M_1$,*

$$e_{i,j} = \min_{q, q_i, q_j \in \Delta} \quad D(q\|p_j) + \beta D(q_i\|p_i) + \beta D(q_j\|p_j) \qquad (13)$$
$$s.t. \quad D(q\|q_j) \geqslant D(q\|q_i).$$

Theorem 1 implies that the error exponent of test (9) is determined by the nearest two alternative distributions, and $e_{i,j}$ in (9) measures the distance between a pair of distributions. It is the error exponent for the case where the ground truth is H_j but $\delta(\boldsymbol{y}) = i$. If both p_i and p_j are known, $e_{i,j}$ equals the Chernoff information between p_i and p_j. Thus, if $M_1 = M$, i.e., all distributions are known, the error exponent reduces to that of the parametric hypothesis testing problem given in [2]. If p_i is known, but p_j is unknown, (11) consists of two terms: the second term captures the approximation error of the training sequence to p_j (where q_j can be viewed as the approximation of p_j), and the first term represents the detection error (where q can be viewed as the approximation of the testing distribution). Hence, if $q_j = p_j$ (i.e., p_j is perfectly learned from the training sequence), $e_{i,j}$ becomes $C(p_i, p_j)$ (the parametric case). This also implies that $e_{i,j}$ in such a case is no larger than $C(p_i, p_j)$. If p_i is unknown but p_j is known, (12) also consists of two terms: the second term captures the approximation error of the training sequence to p_i , and the first term represents the detection error. If neither p_i nor p_j is known, (13) consists of three terms: the last two terms correspond respectively to the approximation errors of p_i and p_j, and the first term represents the detection error. In this case, $e_{i,j}$ reduces to $C(p_i, p_j)$ if $q_i = p_i$ and $q_j = p_j$ (i.e., the approximations of the distributions are perfect). Thus, the error exponent in this case is no larger than $C(p_i, p_j)$.

The following corollary explains under what conditions the error exponent of test (9) approaches that for parametric hypothesis testing, which serves as an upper bound.

Corollary 1. *If $\beta > 0$, test* (9) *is exponentially consistent. Especially, if $\beta \to \infty$, the error exponent goes to $\min_{\{i,j:i\neq j\}} C(p_i, p_j)$, which is optimal, i.e., the error exponent of the nonparametric case approaches that of the parametric case if the length of the training sequences is much larger than that of the testing sequence.*

To further explain the above result, if $0 < \beta < \infty$, for a larger β, the error between $\gamma(\boldsymbol{t}_i)$ and p_i is smaller, and hence the exponent of the maximum error probability takes a larger value. The error exponent is strictly larger than 0. In the extreme case with $\beta = \infty$, the error exponent equals that of the fully parametric model, and hence achieves the optimal value. Thus, if the length of training sequences increases much faster than that of the testing sequence, the error between $\gamma(\boldsymbol{t}_i)$ and p_i can be ignored. In such a case, those unknown distributions can be accurately estimated, and hence do not affect the error exponent of the maximum error probability.

Corollary 2. *If $\beta = 0$ and $M_1 < M$ (at least one distribution is unknown), the error exponent of the maximum error probability for test* (9) *equals zero, i.e., the test is not exponentially consistent.*

The above result implies that if the length of training sequences is small compared with that of the testing sequence, then the unknown distributions cannot be well estimated, which consequently causes the inconsistency of the test.

3.2. Computation of Error Exponent

It is clear in our analysis that the error exponent in various cases is expressed as the minimum value of an optimization problem, which is nonconvex and difficult to solve. We next discuss how to obtain solutions to these optimization problems.

First, problem (12) is a min-max problem, which can be written as

$$\min_{q_j\in\Delta} \max_{\lambda\in[0,1]} F(q_j, \lambda) = -\log\Big(\sum_{y\in\mathcal{Y}} p_j(y)^{\lambda} q_i(y)^{1-\lambda}\Big) + \beta \sum_{y\in\mathcal{Y}} q_i(y) \log \frac{q_i(y)}{p_j(y)}. \qquad (14)$$

It is easy to prove that the objective function is convex over q_j and concave over λ, and for every saddle point $(\hat{q_j}, \hat{\lambda})$, we have

$$\inf_{q_j\in\Delta} \sup_{\lambda\in[0,1]} F(q_j, \lambda) \leqslant F(\hat{q_j}, \hat{\lambda}) \leqslant \sup_{\lambda\in[0,1]} \inf_{q_j\in\Delta} F(q_j, \lambda). \qquad (15)$$

Thus, with the following lemma, all saddle points of problem (14) share the same function value for $F(\cdot, \cdot)$.

Lemma 1. (Corollary 37.3.2 in [14]) *Let C and D be nonempty closed convex sets in R^m and R^n, respectively, and let K be a continuous finite concave-convex function on $C \times D$. If either C or D is bounded, we have*

$$\inf_{v\in D} \sup_{u\in C} K(u, v) = \sup_{u\in C} \inf_{v\in D} K(u, v). \qquad (16)$$

In addition, following [14], the optimal point of the min-max problem is one of the saddle points of the objective function. Thus, with the discussion above, the problem is solved at a satisfactory level as long as we find one saddle point. A sub-gradient method [15] can be utilized to find such a point. First initialize $q_j^{(0)}$ and $\lambda^{(0)}$ randomly, and then perform sub-gradient decent steps alternatively over q_j and λ as

$$q_j^{(k+1)} = \mathcal{P}_{\Delta}[q_j^{(k)} - s\nabla F_{q_j}(q_j^{(k)}, \lambda^{(k)})] \quad (17)$$

$$\lambda^{(k+1)} = \mathcal{P}_{[0,1]}[\lambda^{(k)} + s\nabla F_{\lambda}(q_j^{(k+1)}, \lambda^{(k)})], \quad (18)$$

where s is the step size, $\mathcal{P}_{\Delta}$ and $\mathcal{P}_{[0,1]}$ denote the projections onto the simplex sets Δ and $[0,1]$, respectively; ∇F_{q_j} and ∇F_{λ} denote the sub-gradients of F with respect to q_j and λ, respectively. Then by choosing an appropriate step size s, the above algorithm can be shown to converge.

The problem in (11) is also a nonconvex problem since the constraint is nonconvex. Hence, it is difficult to make the projection onto the constraint set. In this case, we incorporate the constraint set into the objective function as

$$\min_{q,q_j\in\Delta} \quad G(q,q_j) = D(q\|p_j) + \beta D(q_j\|p_j) + l\big(D(q\|q_j) - D(q\|p_i)\big), \quad (19)$$

where

$$l(x) = \begin{cases} 0, & x \geqslant 0 \\ \frac{1}{2}\mu x^2, & x < 0 \end{cases} \quad (20)$$

for some $\mu > 0$. To minimize the difference between (11) and (19), we need to set a large value for μ. It can be shown that (11) is a Kurdyka-Lojasiewicz (KL) function [16] and Lipschitz continuous near the critical point. Then, we apply the following gradient projection method [16, 17]

$$q_j^{(k+1)} = \mathcal{P}_{\Delta}[q_j^{(k)} - s\nabla G_{q_j}(q^{(k)}, q_j^{(k)})] \quad (21)$$

$$q^{(k+1)} = \mathcal{P}_{\Delta}[q^{(k)} - s\nabla G_{q}(q^{(k)}, q_j^{(k)})], \quad (22)$$

where s is the step size, $\mathcal{P}_{\Delta}$ denotes the projection onto the simplex set Δ, and ∇G_q and ∇G_{q_j} denote the sub-gradients of G with respect to q and q_j, respectively. By choosing a $q_j^{(0)}$ to be close to p_j (e.g., let $q_j^{(0)}$ takes the empirical distribution of $\boldsymbol{t}_j$, and $q^{(0)}$ be in the middle of $q_j^{(0)}$ and p_i), the iteration can be shown to converge to a local minimizer of (19).

Problem (13) can be computed similarly as (11).

4. NUMERICAL RESULTS

In this section, we provide numerical results to validate the theoretical analysis. More experiment results can be found in the extended version [13].

We first study how the Chernoff information affects the error decay performance. In this experiment, we set $M = 2$ and study the following three cases. We set $p_1 = [0.9, 0.05, 0.05]$ and $p_2 = [0.05, 0.05, 0.9]$ for Case 1, in which $C(p_1, p_2) = 0.746$. We set $p_1 = [0.8, 0.1, 0.1]$ and $p_2 = [0.1, 0.1, 0.8]$ for Case 2, in which $C(p_1, p_2) = 0.4069$. We set $p_1 = [0.6, 0.2, 0.2]$ and $p_2 = [0.2, 0.2, 0.6]$ for Case 3, in which $C(p_1, p_2) = 0.1134$. We also set $\bar{n} = n$. All distributions are assumed to be unknown in the test. Fig. 1 plots the performance for all cases, and it can be seen that larger Chernoff distances result in larger error exponents, which corroborates our result.

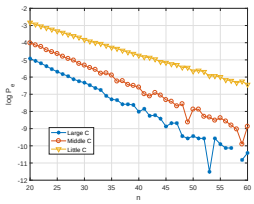

Fig. 1. Impact of the Chernoff information on error decay performance

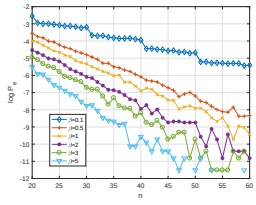

Fig. 2. Impact of the ratio $\beta = \frac{\bar{n}}{n}$ of the training and testing sequences on error decay performance

We next study how the ratio $\beta = \frac{\bar{n}}{n}$ affects the error decay performance. We set $p_1 = [0.1, 0.1, 0.8]$ and $p_2 = [0.8, 0.1, 0.1]$. Fig. 2 plots the error decay performance for the cases with $\beta = 0.1, 0.5, 1, 2, 3, 5$, respectively. All distributions are assumed unknown in the test. It can be seen that a larger β yields a larger error exponent for test (9), which corroborates Corollary 1.

5. CONCLUSION

In this paper, we have studied the nonparametric hypothesis testing problem. Our focus has been on the characterization of the error exponent for the maximum error probability. We have showed that the GL test is exponentially consistent as long as the number of training samples in each sequence scales no slower than the number of testing samples. As future work, it will be interesting to study the regime where the number of hypotheses also goes to infinity, and explore how the number of samples should scale accordingly in order to guarantee the exponential consistency of the GL test.

6. REFERENCES

[1] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*, Springer Science & Business Media, 2008.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2012.

[3] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans-*

actions on Information Theory, vol. 34, no. 2, pp. 278–286, Mar. 1988.

[4] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[5] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?," *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, Sept. 1992.

[6] A. O. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, Dec. 2006, pp. 585–592.

[7] A. O. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1921–1938, Sept. 1999.

[8] M. Zhao and V. Saligrama, "Anomaly detection with score functions based on nearest neighbor graphs," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, Dec. 2009, pp. 2250–2258.

[9] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5375–5386, July 2011.

[10] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4066–4082, Apr. 2014.

[11] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Transactions on Signal Processing*, vol. 65, no. 21, 2017.

[12] Y Bu, S. Zou, and V. V. Veeravalli, "Linear complexity exponentially consistent tests for outlying sequence detection," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 2017, pp. 988–992.

[13] Y. Liu, Y. Liang, and S. Cui, "Data-driven nonparametric existence and association problems," *Arxiv preprint*.

[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 2015.

[15] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, Mar. 2009.

[16] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, Aug. 2013.

[17] J. C. Dunn, "On the convergence of projected gradient processes to singular critical points," *Journal of Optimization Theory and Applications*, vol. 55, no. 2, pp. 203–216, Nov. 1987.