Data-Driven Nonparametric Hypothesis Testing

Yixian Liu*, **Yingbin Liang**†, **Shuguang Cui***

*Shanghai Institute of Microsystem & Information Technology, Chinese Academy of Sciences

†Department of Electrical and Computer Engineering, The Ohio State University

*Department of Electrical and Computer Engineering, University of California at Davis

Problem formulation

There are in total M distinct discrete distributions $p_1, \ldots, p_M$ and a testing sequence $\boldsymbol{y}$, which consists of n i.i.d. samples generated by one of the M distributions.

- Which is the distribution that generates the test sequence?
- p_i's for $i > M_1$ are unknown;
- for an unknown p_i, a training sequence $\boldsymbol{t}_i$ is available.

Review of Existing Work

For parametric model, to make sequence detection between two distribution p_1 and p_2, the optimal test is the likelihood ratio test (LRT), and the optimal error exponent is the Chernoff information $C(p_1, p_2)$. [1]

For universal outlier hypothesis testing, the error exponent of GLRT is determined by the Bhattacharyya distance between the typical and outlier distributions. [2]

Testing

Test 1.

$$\sigma(\boldsymbol{y}) = \arg\min_i \begin{Bmatrix} D(\gamma(\boldsymbol{y})||p_i), & \text{if } i \leqslant M_1 \\ D(\gamma(\boldsymbol{y})||\gamma(\boldsymbol{t}_i)), & \text{if } i > M_1 \end{Bmatrix}. \quad \text{(T)}$$

- $\gamma(\boldsymbol{y})$ denotes the empirical distribution of $\boldsymbol{y}$ given by

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

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

$$D(p||q) = \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}. \quad (1)$$

Main Result

Theorem 1.

Apply test **(T)** 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 $j > M_1$,

$$e_{i,j} = \min_{q, q_j \in \Delta} \quad D(q||p_j) + \beta D(q_j||p_j) \quad \text{s.t.} \quad D(q||q_j) \geqslant D(q||p_i), \quad (2)$$

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

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

$$e_{i,j} = \min_{q_i \in \Delta} \quad C(q_i, p_j) + \beta D(q_i||p_i); \quad (3)$$

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

$$e_{i,j} = \min_{q, q_i, q_j \in \Delta} \quad D(q||p_j) + \beta D(q_i||p_i) + \beta D(q_j||p_j) \quad \text{s.t.} \quad D(q||q_j) \geqslant D(q||q_i). \quad (4)$$

Main Result(continued)

- $C(\cdot, \cdot)$ denotes the Chernoff information given by

$$C(p, q) = \max_{\lambda \in [0,1]} -\log \left(\sum_{y \in \mathcal{Y}} p(y)^{\lambda} q(y)^{1-\lambda} \right). \quad (5)$$

- β denotes the ratio between the lengths of training and testing sequences, $\beta = \lim_{n \to \infty} \frac{\tilde{n}}{n}$.

Corollary.

1. If $\beta > 0$, test **(T)** 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.
2. If $\beta = 0$ and $M_1 < M$, the test **(T)** is not exponentially consistent.

Computation of Error Exponent

- Approximation of error exponent (2)

$$\min_{q, q_j \in \Delta} \quad G(q, q_j) = D(q||p_j) + \beta D(q_j||p_j) + l\big(D(q||q_j) - D(q||p_i)\big), \quad (6)$$

where

$$l(x) = \begin{cases} 0, & x \geqslant 0 \\ \frac{1}{2}\mu x^2, & x < 0 \end{cases} \quad (7)$$

then update

$$q_j^{(k+1)} = \mathcal{P}_{\Delta}[q_j^{(k)} - s\nabla G_{q_j}(q^{(k)}, q_j^{(k)})] \quad (8)$$

$$q^{(k+1)} = \mathcal{P}_{\Delta}[q^{(k)} - s\nabla G_q(q^{(k)}, q_j^{(k)})]. \quad (9)$$

- For error exponent (3)

$$q_j^{(k+1)} = \mathcal{P}_{\Delta}[q_j^{(k)} - s\nabla F_{q_j}(q_j^{(k)}, \lambda^{(k)})] \quad (10)$$

$$\lambda^{(k+1)} = \mathcal{P}_{[0,1]}[\lambda^{(k)} + s\nabla F_{\lambda}(q_j^{(k+1)}, \lambda^{(k)})]. \quad (11)$$

Numerical Experiment

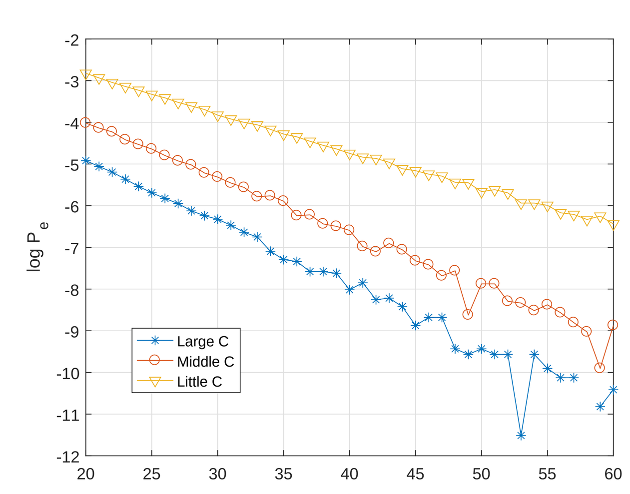

Fig.1: Impact of the Chernoff information on error decay performance.

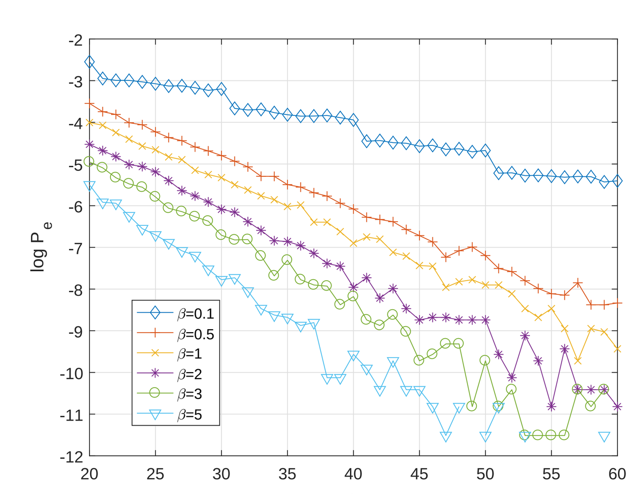

Fig.2: Impact of the ratio $\beta = \frac{\tilde{n}}{n}$ of the training and testing sequences on error decay performance.

References

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

[2] Y. Li, S. Nitinawarat, and V. V. Veeravalli. Universal outlier hypothesis testing. *IEEE Transactions on Information Theory*, 60(7):4066–4082, Apr. 2014.