Sequential Classification with Empirically Observed Statistics

Mahdi Haghifam∗, Vincent Y. F. Tan†, and Ashish Khisti∗
∗Department of Electrical and Computer Engineering, University of Toronto.
Email: mahdi.haghifam@mail.utoronto.ca, akhisti@ece.utoronto.ca
†Department of Electrical and Computer Engineering & Department of Mathematics, National University of Singapore.
Email: vtan@nus.edu.sg

Abstract—Motivated by real-world machine learning applications, we consider the binary statistical classification task in the sequential setting where the generating distributions are unknown and only empirically sampled sequences are available to the decision maker. Then, the decision maker is tasked to classify a test sequence which is known to be generated according to either one of two distributions. The decision maker wishes to perform the classification task with minimum number of the test samples, so, at each step, it declares either “1”, “2” or “give me one more test sample”. We propose a classifier and analyze the type-I and type-II error probabilities. Also, we show the advantage of our sequential scheme compared to the existing non-sequential classifiers.

I. INTRODUCTION

Quick and accurate classification are crucial in many applications. For instance, to diagnose hematologic diseases based on blood test results, a physician wishes to detect the pattern, deviations, and relations in the blood samples of a patient as quickly as possible to make treatment plans. This example illustrates in many application there is a inherent tradeoff between speed and accuracy.

In many real-world applications, hypothesis testing is infeasible due to the fact that the probability distributions of sources are unknown. On the other hand, statistical classification also addresses the same problem with the difference that the probability distributions of the sources are not known exactly, but instead, have to be estimated from training sequences produced by the sources. These differences make statistical classification viable for many real-world problems.

The problem of classification using empirically observed statistics has been studied in prior works. Gutman in [1] considers the binary classification problem and proposes an asymptotically optimal test. The non-asymptotic performance of Gutman’s test is analyzed in [2]. Moreover, [3] studies the relationship between binary classification and universal data compression. Unnikrishnan in [4], [5] extends the Gutman’s proposed test for the case with multiple test sequence and shows its optimality. Furthermore, [6] analyzes the statistical classification when sources have finite but very large alphabets. The authors in [7] studies the problem of distributed detection in the setting that the fusion center has access to noisy training sequences. Finally, the related problem of closeness testing has been investigated [8], [9].

In this paper, we consider the binary sequential classification. Recall that in the simple sequential binary hypothesis testing a decision maker is given a varying length test sequence and knows that it is either generated in an i.i.d. fashion from one of the known distributions P1 or P2. It is well-known that sequential probability ratio test is optimal for the sequential binary hypothesis testing [10]. However, we consider a scenario that the decision maker does not know the generating distributions, i.e., P1 and P2. Instead the decision maker has two training sequences, one is drawn from P1 and the other from P2. Then, the task of the decision maker is to classify a test sequence. The decision maker observes the test sequence sequentially and may choose when to stop sampling to declare the decision. For this setup, we propose a decision rule, which is defined formally in Section III. Then, we analyze the type-I and type-II error exponent of the proposed scheme. Also, we compare our achievable type-I and type-II error exponents with that of Gutman’s and observe that our scheme significantly outperforms Gutman’s fixed-length scheme in terms of type-I and type-II error exponents.

A. Paper outline

This paper is organized as follows. The notation in this paper is explained in the next subsection. Section II presents the problem statement of binary classification in the sequential setting as well as existing results. Then, the proposed classifier for the problem is stated in Section III. In Section IV, we analytically characterize the performance of the proposed scheme. The proof sketch of our results is given in Section V.

B. Notation

The set of all discrete distributions on alphabet X is denoted as \( \mathcal{P}(\mathcal{X}) \). Notation concerning the method of types follows [11, Chapter 11] and [12]. Given a vector \( x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n \), the type or empirical distribution is denoted as \( \mu_n (a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{ x_i = a \} \), \( \forall a \in \mathcal{X} \). The set of all sequences of length n with type Q is denoted by \( \mathcal{Q}_n^Q \) (we sometimes omit n if it is clear from the context). If \( Q \) is a distribution on \( \mathcal{X} \) then \( Q^n \) is the n-fold i.i.d product measure on \( \mathcal{X}^n \), i.e.,

\[
Q^n (x^n) = \prod_{i=1}^{n} Q (x_i)
\]
II. PROBLEM STATEMENT

We assume that a decision maker has two training sequences of length \( N \) which are generated according to two unknown but fixed distributions \((P_1, P_2) \in \mathcal{P}(\mathcal{X})^2\). The training sequences are denoted by \( X_1^N \) and \( X_2^N \). Then, at each time \( n \in \mathbb{N} \) a test sample \( Y_n \), as being independently generated from one of two distinct distributions, i.e., \( P_1 \) and \( P_2 \), is given to the decision maker. The objective of the decision maker is to decide between the following two hypotheses:

- \( H_1 \): The test sequence and the first training sequence are generated according to the same distribution.
- \( H_2 \): The test sequence and the second training sequence are generated according to the same distribution.

To achieve this goal, the decision maker at each time \( n \) can take three actions:

- Stop drawing a new test sample and declare the test sequence and the first training sequence are generated according to the same distribution.
- Stop drawing a new test sample and declare the test sequence and the second training sequence are generated according to the same distribution.
- Continue drawing a new test sample.

As opposed to sequential hypothesis testing [13, Section 15.3] where two distributions are known, in this setup the decision maker does not know the distributions. Instead, the only information decision maker has about \( P_1 \) and \( P_2 \) is through two training sequences \( X_1^N \) and \( X_2^N \) generated in an i.i.d. fashion according to \( P_1 \) and \( P_2 \) respectively. Moreover, the problem considered here is different from [1]–[3], [5] where the classification is studied for the cases that the length of the test sequence is fixed prior to the decision making. In our setup, we let the size of the test sequence be a stopping time determined by the decision maker’s action. Next, we provide a precise formulation of this problem. We begin with the definition of the test for the aforementioned setup.

**Definition 1 (Test).** A test is a pair \( \Phi = (T, d) \) where

- \( T \in \mathbb{N} \) is a stopping time with respect to the filtration \( \mathcal{F}_n = \sigma\{X_1^N, X_2^N, Y_1, \ldots, Y_n\} \).
- \( d : (X_1^N, X_2^N, Y^T) \to \{1, 2\} \) is the terminal decision rule.

**Definition 2 (Type-I and Type-II Error Probabilities).** For a test \( \Phi \), the type-I and type-II error probabilities are defined as

\[
\mathbb{P}_\Phi(E|H_1) = \mathbb{P}(d(X_1^N, X_2^N, Y^T) = 2|H_1), \quad (1)
\]

\[
\mathbb{P}_\Phi(E|H_2) = \mathbb{P}(d(X_1^N, X_2^N, Y^T) = 1|H_2), \quad (2)
\]

respectively.

**Definition 3 (Error Exponents).** For a test \( \Phi \) such that \( \mathbb{P}_\Phi(E|H_i) \to 0 \) as \( N \to \infty \) for each \( i \in \{1, 2\} \), we define

\[
E_i(\Phi) = \liminf_{N \to \infty} -\frac{\log \mathbb{P}_\Phi(E|H_i)}{\mathbb{E}_\Phi[T|H_i]}, \quad (3)
\]

where \( \mathbb{E}_\Phi[T|H_i] \) represents the expected value of the stopping time under \( H_i \).

**Remark 1.** Note that \( \mathbb{E}_\Phi[T|H_i] \) also indicates the average number of the test samples under \( H_i \) before the decision is made.

Gutman [1] considers the setup in which the decision maker has a test sequence \( Y^n \) of fixed length \( n \) which is independently generated from \( X_1^N \) and \( X_2^N \). In [1] it is assumed that \( N = n\alpha \) for some \( \alpha \in \mathbb{R}_+ \). Then, Gutman proposes a test which enjoys the optimal type-I and type-II error exponents’ tradeoff. To present Gutman’s results, we need the following definition.

**Definition 4 (Generalized Jensen-Shannon Divergence).** Given \( \alpha \in \mathbb{R}_+ \) and \( (P_1, P_2) \in \mathcal{P}(\mathcal{X})^2 \), the generalized Jensen-Shannon divergence is defined as

\[
GJS(P_1, P_2, \alpha) = \alpha D\left( \frac{P_1 + P_2}{2} \bigg| \bigg| \frac{P_1 + P_2}{2} \right) + \left( \frac{\alpha P_1 + P_2}{1 + \alpha} \right) + \left( \frac{P_1 + \alpha P_2}{1 + \alpha} \right). \quad (4)
\]

**Theorem 1.** ( [1, Thm. 1] ) Gutman’s decision rule, \( \Phi_G \), has the following type-I and type-II error exponents:

\[
E_1(\Phi_G) = \liminf_{N \to \infty} -\frac{\log \mathbb{P}_{\Phi_G}(E|H_1)}{n} \geq \lambda \quad (5)
\]

\[
E_2(\Phi_G) = \liminf_{N \to \infty} -\frac{\log \mathbb{P}_{\Phi_G}(E|H_2)}{n} \geq G(\alpha, \lambda), \quad (6)
\]

where

\[
G(\alpha, \lambda) = \min_{(Q_1, Q_2) \in \mathcal{P}(\mathcal{X})^2} \alpha D(Q_1||P_1) + D(Q_2||P_2)
\]

subject to \( GJS(Q_1, Q_2, \alpha) \leq \lambda \).

III. PROPOSED TEST

Consider \( \gamma \in \mathbb{R}_+ \). The proposed test for the sequential classification is \( \Phi_{seq} = (T_{seq}, d_{seq}) \) where \( T_{seq} \) and \( d_{seq} \) are defined as

\[
T_{seq} = \inf \left\{ n \geq 1 : \exists i \in \{1, 2\} \right. \left. \text{ such that } nGJS\left( T_{X_1^N, Y^n, \frac{N}{n}} \right) \geq \gamma N \right\}, \quad (8)
\]

and

\[
d_{seq} = \begin{cases} 1 & \text{if } T_{seq}GJS\left( T_{X_1^N, Y^n, \frac{N}{n}} \right) \geq \gamma N, \\ 2 & \text{if } T_{seq}GJS\left( T_{X_1^N, Y^n, \frac{N}{n}} \right) \geq \gamma N \end{cases}, \quad (9)
\]

respectively.

As an illustrative, Figure 1 shows a realization of \( \Phi_{seq} \) for \( P_1 = [0.2, 0.5, 0.3] \) and \( P_2 = [0.4, 0.4, 0.2] \) where the test sequence drawn from \( P_0 \) and the length of the training sequence is \( N = 400 \). Note that the stopping time defined in (8) is \( T_{seq} = 305 \) since at \( n = 305 \), \( nGJS\left( T_{X_1^N, Y^n, \frac{N}{n}} \right) \) exceeds the threshold.

Then, based on (9), we declare 2 as the final decision.

To provide some intuition behind choosing \( \Phi_{seq} \), we present the following lemma.
Lemma 1. Assume \(v^n\) and \(w^n\) are sequences which consist of, respectively, \(N\) and \(n\) i.i.d samples from unknown distributions, possibly the same distribution, over the same alphabet. Consider the following optimization problem.

\[
\min_{P \in \mathcal{P}(\mathcal{X})} -\log P^{N+n} (v^n, w^n). \tag{10}
\]

Then, the optimal value is

\[
n \text{GJS} \left( T_{v^n}, T_{w^n}, \frac{N}{n} \right) + NH (T_{w^n}) + nH (T_{v^n}). \tag{11}\]

Also, the optimal solution is given by \(P^* = \frac{n T_{v^n} + T_{w^n}}{1 + \frac{n}{N}}\).

Lemma 1 shows that a large GJS between two between types formed from two sequences implies that the probability that they are generated from the same distribution is low. This observation leads us to define \(\Phi_{\text{seq}}\) in (8) and (9).

IV. ANALYSIS OF THE PROPOSED SCHEME

In the next theorem, we present the main result of this paper which concerns the achievable type-I and type-II error exponents of the proposed test in Section III. The proof sketch of the following theorem is presented in Section V.

Theorem 2. Fix pair \((P_1, P_2) \in \mathcal{P}(\mathcal{X})^2\). Then, for any \(\gamma \in \Lambda\) where

\[
\Lambda = \left\{ \gamma \mid \min_{V \in \mathcal{P}(\mathcal{X}): D(V||P_1) \leq \gamma} D(V||P_1) \geq \gamma, \min_{V \in \mathcal{P}(\mathcal{X}): D(V||P_2) \leq \gamma} D(V||P_2) \geq \gamma \right\},
\]

the proposed test in Section III achieves

\[
E_1 (\Phi_{\text{seq}}) = \liminf_{N \to \infty} \frac{-\log \mathbb{P}_{\Phi_{\text{seq}}} (E|H_1)}{\mathbb{E}_{\Phi_{\text{seq}}} [T_{\text{seq}}|H_1]} \geq \text{GJS} (P_2, P_1, \beta^*), \tag{12}\]

\[
E_2 (\Phi_{\text{seq}}) = \liminf_{N \to \infty} \frac{-\log \mathbb{P}_{\Phi_{\text{seq}}} (E|H_2)}{\mathbb{E}_{\Phi_{\text{seq}}} [T_{\text{seq}}|H_2]} \geq \text{GJS} (P_1, P_2, \theta^*). \tag{13}\]

Here, in (12), \(\beta^*\) is the solution of

\[
\text{GJS} (P_2, P_1, \beta^*) = \gamma \beta^*. \tag{14}\]

Similarly, \(\theta^*\) in (13) is the solution of

\[
\text{GJS} (P_1, P_2, \theta^*) = \gamma \theta^*. \tag{15}\]

Here, we present a numerical example to illustrate the performance of Gutman’s test versus the test proposed in Section III. Consider the alphabet \(\mathcal{X} = \{1, 2, 3, 4\}\). Also, the generating distributions are \(P_1 = [0.25, 0.25, 0.25, 0.25]\) and \(P_2 = [0.4, 0.5, 0.05, 0.05]\). The performance measure we study is the minimum of type-I and type-II error exponents, i.e., \(\min \{E_1 (\Phi), E_2 (\Phi)\}\). In the sequential test, the average number of test samples under hypothesis \(H_1\) and \(H_2\) are \(\frac{N}{\beta}\) and \(\frac{N}{\theta}\), respectively. On the other hand, Gutman’s test is devised for a fixed number of the test samples. To ensure a fair comparison, in Fig. 2 we set \(\alpha\) in (6) to be equal to \(\beta^*\). Also, in Fig. 3, we set \(\alpha = \theta^*\). Moreover, we set the value of \(\lambda\) in (5) and (6) to

\[
\lambda^* = \arg \max_{\lambda} \{\min \{E_1 (\Phi_{\text{Gut}}), E_2 (\Phi_{\text{Gut}})\}\}. \tag{13}\]

As observed in Figures 2 and 3, the sequential test significantly outperforms the Gutman’s test in terms of minimum of two error exponents.

V. PROOF SKETCH OF THEOREM 2

This section has three subsections. First, we start with upper bounding the error probability. Then, we provide our results on the expected value of the stopping time. Finally, we conclude with the derivation of the achievable exponents in (12) and (13).

A. Error Probability

The next lemma provides an upper bound on the probability that the empirical GJS of the test sequence with the training sequence with the same generating distribution exceeds the threshold in (8) at some time \(n\).
Lemma 2. Assume \( u^N \) and \( w^m \) are two sequences drawn i.i.d. from the same probability distribution \( Q \). Then, we have
\[
P \left( n \text{GJS} \left(T_{u^N}, T_{w^m}, \frac{N}{n} \right) \geq \gamma N \right) \leq \exp \left( -\gamma N \right) \left( m + N + 1 \right)^{\left| X^N \right|} \tag{16}\]

Assume that the test sequence generated according to \( P_2 \). Consider any sequence \( \delta_N \) with the following two properties:

1. \( \lim_{N \to \infty} \delta_N = 0 \)
2. \( \frac{\delta_N}{n} \log \frac{1}{\delta_N} = o(1) \)

For instance one can consider \( \delta_N = \frac{1}{\sqrt{2} N} \). Then, we define test \( \Phi_{\text{trunc}} \) as a truncated version of \( \Phi_{\text{seq}} \). Using \( \Phi_{\text{trunc}} \), the decision maker follows the same decision rule as \( \Phi_{\text{seq}} \) in the interval \( \left[ 1, \left\lfloor \frac{N}{\delta_N} \right\rfloor \right] \). But, provided that in the interval \( \left[ 1, \left\lfloor \frac{N}{\delta_N} \right\rfloor \right] \) stopping time \( T_{\text{seq}} \) has not occurred, the decision maker declares error. It is easy to see that the error probability of \( \Phi_{\text{trunc}} \) is larger than \( \Phi_{\text{seq}} \). Hence, we can write
\[
P_{\Phi_{\text{trunc}}} \left( \mathcal{E} | H_2 \right) \leq P_{\Phi_{\text{seq}}} \left( \mathcal{E} | H_2 \right) \tag{17}\]
\[
\leq P \left( \text{Wrong Decision} | H_2 \right) + P \left( \text{No Decision} | H_2 \right) \tag{18}\]

Theorem 3. Let \( \theta^* \) denote the solution of
\[
\text{GJS} \left( P_1, P_2, \theta^* \right) = \gamma \theta^* \tag{22}\]

Similarly, define \( \beta^* \) as the solution of
\[
\text{GJS} \left( P_1, P_2, \beta^* \right) = \gamma \beta^* \tag{23}\]

Then, we have
\[
\mathbb{E}_{\Phi_{\text{seq}}} \left[ T_{\text{seq}} | H_1 \right] = \frac{N}{\beta^*} \left( 1 + o(1) \right) \tag{24}\]
\[
\mathbb{E}_{\Phi_{\text{seq}}} \left[ T_{\text{seq}} | H_2 \right] = \frac{N}{\theta^*} \left( 1 + o(1) \right) \tag{25}\]

Proof. We prove Theorem 3 by providing upper and lower bounds on the expected value of the stopping time. First of all we present two lemmas that will be used in the proof.

Lemma 3. The stopping time defined in (8) is almost surely greater than \( \left( \frac{N}{2 \log 2} \right)^2 \).

Lemma 4. Let \( X_1^N \) denote a sequence consisting of \( N \) i.i.d. samples drawn according to \( P_1 \). Also, \( \theta_N^* \) denote the solution of \( \text{GJS} \left( T_{X_1^N}, P_2, \theta_N^* \right) = \lambda \theta_N^* \). Then, as \( N \) goes to infinity \( \theta_N^* \) converges in probability to \( \theta^* \), defined by the solution of \( \text{GJS} \left( P_1, P_2, \theta^* \right) = \gamma \theta^* \).

For a fixed pair of the training sequences \( x_1^N \) and \( x_2^N \), define \( \theta_N^* \) as the solution of
\[
\text{GJS} \left( T_{x_1^N}, P_2, \theta_N^* \right) = \gamma \theta_N^* \tag{26}\]

Moreover, for each \( N \) we construct \( \theta_N^* \) which satisfies the following three properties:

1. \( \theta_N^+ > \theta_N^* \)
2. \( \lim_{N \to \infty} \left( \theta_N^+ - \theta_N^* \right) = 0 \)
3. \( \mathbb{P} \left( \left( \frac{N}{2 \log 2} \right)^2 \leq T_{\text{seq}} \leq \frac{N}{\theta_N^*} | H_2, X_1^N \right) = o(1) \)

Due to space limitations, the explicit construction of \( \theta_N^* \) is omitted. First we obtain
\[
\mathbb{E}[T_{\text{seq}} | H_2, X_1^N] = \sum_{k \geq 1} \mathbb{P}(T_{\text{seq}} \geq k | H_2, X_1^N) \tag{27}\]
\[
\geq \sum_{k \geq 1} \mathbb{P}(T_{\text{seq}} \geq k | H_2, X_1^N) \tag{28}\]

\[
\geq \frac{N}{\theta_N^*} \left( 1 - \mathbb{P} \left( 1 \leq T_{\text{seq}} \leq \frac{N}{\theta_N^*} | H_2, X_1^N \right) \right) \tag{29}\]

\[
= \frac{N}{\theta_N^*} \left( 1 - \mathbb{P} \left( \left( \frac{N}{2 \log 2} \right)^2 \leq T_{\text{seq}} \leq \frac{N}{\theta_N^*} | H_2, X_1^N \right) \right), \tag{30}\]
where in (30) we have used Lemma 3 which shows there is a almost surely lower bound for $T_{seq}$. Therefore, we conclude

$$\mathbb{E}[T_{seq} | H_2, X_1^N] \geq \frac{N}{\theta_N} (1 - o(1)) \tag{31}$$

Then, to lower bound the expected value, we define $\theta_N$ for each $N$ which satisfies

1. $\theta_N < \theta_N$
2. $\lim_{N \to \infty} (\theta_N - \theta_N) = 0$
3. $\min_{V \in \mathcal{P}(X)} D(V\|P_2) \geq \frac{1}{\sqrt{N}}$

Due to space limitations, the explicit construction of $\theta_N$ is omitted. Then, We can write

$$\mathbb{P} \left( T_{seq} > \frac{N}{\theta_N} | H_2, X_1^N \right) = \sum_{k \geq \frac{N}{\theta_N}} \mathbb{P}(T_{seq} = k + 1 | H_2, X_1^N) \tag{32}$$

$$\leq \sum_{k \geq \frac{N}{\theta_N}} \mathbb{P} \left( k \text{GJS} \left( T_{x_1^N}, y, k\frac{N}{k} \right) \leq \gamma N | H_2, X_1^N \right) \tag{33}$$

$$\leq \sum_{k \geq \frac{N}{\theta_N}} (k + 1)^{|\mathcal{X}|} \exp \left( -k \min_{V \in \mathcal{P}(X)} D(V\|P_2) \right) \tag{34}$$

$$\leq \sum_{k \geq \frac{N}{\theta_N}} (k + 1)^{|\mathcal{X}|} \exp \left( -k \min_{V \in \mathcal{P}(X)} D(V\|P_2) \right) \tag{35}$$

$$\leq \sum_{k \geq \frac{N}{\theta_N}} \exp \left( \frac{|\mathcal{X}|}{2} \log(k + 1) \right) \exp \left( -k \frac{1}{\theta_N} \right) \tag{36}$$

$$\leq \exp \left( -\frac{\sqrt{N}}{\theta_N} (1 + o(1)) \right). \tag{37}$$

Here, (35) follows using the fact that $k \text{GJS} \left( P, Q, \frac{N}{k} \right)$ is decreasing in $k$ for fixed $P$, $Q$, and $N$. Then, (36) is obtained using the properties of $\theta_N$. Also, the last step follows from some manipulations. Then, from (32)-(37), we deduce that

$$\mathbb{E} \left[ T_{seq} | H_2, X_1^N \right] \leq \frac{N}{\theta_N} \mathbb{P} \left( T_{seq} \leq \frac{N}{\theta_N} | H_2, X_1^N \right)$$

$$+ \sum_{k \geq \frac{N}{\theta_N}} (k + 1) \mathbb{P}(T_{seq} = k + 1 | H_2, X_1^N) \leq \frac{N}{\theta_N} (1 + o(1)), \tag{38}$$

Thus, we conclude that

$$\mathbb{E} \left[ T_{seq} | H_2, X_1^N \right] \leq \frac{N}{\theta_N} (1 + o(1)), \tag{39}$$

Using (31), (39), and Lemma 4 we can conclude (25).

C. Error Exponent

First we present the following lemma which help us characterize the error exponents.

Lemma 5. ([7, Lemma 2]) We have

$$\lim_{N \to \infty} \min_{P \in \mathcal{P}(X)} D \left( \frac{1}{\delta_N} D (V_1 | P_1) + \frac{1}{\delta_N^2} D (V_2 || P_2) \right) \tag{40}$$

For the type-II error exponent we have

$$\mathbb{E} \left( \Phi_{seq} \right) = \mathbb{E} \left[ T_{seq} | H_2 \right] \tag{41}$$

$$\geq \theta^* \lim_{N \to \infty} \min_{P \in \mathcal{P}(X)} D \left( V_1 | P_1 \right) \tag{42}$$

$$= \theta^* \min_{P \in \mathcal{P}(X)} D \left( V_1 | P_1 \right) \tag{43}$$

$$= \theta^* \min_{P \in \mathcal{P}(X)} D \left( V_1 | P_1 \right) \tag{44}$$

Here, (42) and (43) is obtained using (21) and Lemma 5. Note that the extension of the results here to the type-I error exponent can be readily done by substituting $P_1$ by $P_2$ which leads to the statement in Theorem 2.

REFERENCES

[1] Michael Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. IEEE Transactions on Information Theory, 35(2):401–408, 1989.

[2] Lin Zhou, Vincent Y F Tan, and Mehul Motani. Second-order asymptotically optimal statistical classification. Information and Inference: A Journal of the IMA, 01 2019.

[3] Jacob Ziv. On classification with empirically observed statistics and universal data compression. IEEE Transactions on Information Theory, 34(2):278–286, 1988.

[4] Jayakrishnan Unnikrishnan and Farid Movahedi Naini. De-anonymizing multiple sequences to source distributions and training sequences. IEEE Transactions on Information Theory, 61(1):452–468, 2015.

[5] Jayakrishnan Unnikrishnan. Asymptotically optimal matching of multiple sequences to source distributions and training sequences. IEEE Transactions on Information Theory, 61(1):452–468, 2015.

[6] Benjamin G Kelly, Aaron B Wagner, Thitidej Tularak, and Pramod Viswanath. Classification of homogeneous data with large alphabets. IEEE Transactions on Information Theory, 59(2):782–795, 2013.

[7] Haiyun He, Lin Zhou, and Vincent YF Tan. Distributed detection with empirically observed statistics. arXiv preprint arXiv:1903.05819, 2019.

[8] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In Proceedings of the 25th Annual Conference on Learning Theory, Edinburgh, Scotland, Jun 2012.

[9] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. SIAM Journal on Computing, 46(1):429–455, 2017.

[10] Abraham Wald. Sequential tests of statistical hypotheses. The annals of mathematical statistics, 16(2):117–186, 1945.

[11] Thomas M Cover and Joy A Thomas. Elements of Information Theory. John Wiley & Sons, 2012.

[12] Imre Csiszár. The method of types. IEEE Transactions on Information Theory, 44(6):2505–2523, 1998.

[13] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2017.