On the Adversarial Robustness of Hypothesis Testing

Yulu Jin, *Student Member, IEEE* and Lifeng Lai, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate the adversarial robustness of hypothesis testing rules. In the considered model, after a sample is generated, it will be modified by an adversary before being observed by the decision maker. The decision maker needs to decide the underlying hypothesis that generates the sample from the adversarially-modified data. We formulate this problem as a minimax hypothesis testing problem, in which the goal of the adversary is to design attack strategy to maximize the error probability while the decision maker aims to design decision rules so as to minimize the error probability. We consider both hypothesis-aware case, in which the attacker knows the true underlying hypothesis, and hypothesis-unaware case, in which the attacker does not know the true underlying hypothesis. We solve this minimax problem and characterize the corresponding optimal strategies for both cases.**

***Index Terms*—minimax problem, hypothesis testing, adversarial robustness**

I. Introduction

Motivated by growing applications of various signal processing and statistical inference algorithms in safety and security-related applications [2] [3], there is an increasing interest in the study of adversary robustness of statistical inference algorithms [4]–[11]. The purpose of these studies is to understand the robustness of these algorithms in the adversarial setup, so as to properly design systems that are safe and secure even under adversarial attacks. The investigation of adversary robustness of statistical algorithms is related to but different from the large volume of work on classic robust statistics [12]–[17]. The classic robust statistical inference mainly focuses on distributional robustness, in which the true distributions of data lie in the neighborhood of nominal distributions [15], [18], [19]. On the other hand, the attack in the adversary robustness model is more powerful. In particular, in the adversarial robustness models, an adversary is typically assumed to have access to the data sample and can make data-dependent changes. The decision maker then has to make statistical inference based on the adversarially-modified data [20]. For example, in the adversarial example phenomena investigated in the context of deep neural networks [5]–[9], [21]–[27], the attacker observes the original image and then carefully designs perturbations on the image. Even though these perturbations are hardly perceptible to human eyes, the decision of a deep neural network can be easily manipulated.

The goal of this paper is to understand adversarial robustness of hypothesis testing rules. In the considered model, after data samples are generated by the underlying hypothesis, an adversary can observe the samples and then modify them to other values. The decision maker only observes the modified data but still needs to determine which underlying hypothesis is true. We formulate this as a minimax hypothesis testing problem, in which the adversary aims at designing attack strategies to modify the data so as to maximize the error probability while the goal of the decision maker is to design decision rules to minimize the error probability.

We first focus on the hypothesis-aware scenario, in which the adversary knows which hypothesis is used to generate the data sample. The study of this powerful adversarial model can provide performance bounds for other attack models. Under this setting, we show that the formulated minimax problem has a saddle-point solution, which reveals the structures of the optimal attack and defense strategies. In particular, the optimal defense strategy is to perform the Bayesian test on the corresponding probability mass functions (PMFs) after the attack. As a result, we can write the cost function in terms of the attack strategy only, and characterizing the optimal attack strategy is equivalent to solving a maximization problem that is non-convex over the attack strategy. In this paper, we solve this problem for a special case where the optimal Bayesian decision regions corresponding to the PMFs before attack consist of two consecutive regions. Under this assumption, we first derive an upper-bound on the prediction error, which only depends on the PMFs before attack. Afterwards, we design a specific attack scheme and show that the designed attack scheme achieves the upper-bound. This implies that the specific attack scheme is optimal. However, we also note that the attack strategy that achieves the maximum error probability is not unique.

We then study a more practical and challenging hypothesis-unaware scenario, where the attacker does not know the prior information about the underlying hypothesis. Despite the additional challenge, we show that the method developed for the hypothesis-aware case can be properly modified and extended to this scenario. In particular, following a similar saddle-point analysis, we reveal the structure of the optimal attack and defense strategy and convert the problem into a complicated non-convex optimization problem over the attack strategy. We then derive an upper-bound on the error probability and design

Y. Jin and L. Lai are with the Department of Electrical and Computer Engineering, University of California, Davis, CA. Email: {yuljin,lflai}@ucdavis.edu. The work of Y. Jin and L. Lai was supported by the National Science Foundation under Grants CCF-1717943, ECCS-1711468, CNS-1824553, CCF-1908258 and ECCS-2000415. This paper has been presented in part in the 2019 Asilomar Conference on Signals, Systems, and Computers [1].

a specific attack strategy to achieve the upper-bound.

Our work is related to several recent interesting papers [28]–[30]. Similar to our setup, in these papers, the decision maker has to make a decision using samples that might have been compromised by an attacker. Same as our paper, these papers also consider the setup with discrete random variables. On the other hand, in these papers, the decision maker is assumed to rely only on first order statistics, and uses error exponents as the performance metrics. Using tools from information theory and game theory, these papers characterize the asymptotic equilibrium of the games between the attacker and detector, as the number of samples increases. Different from these papers, we focus on the non-asymptotic case. In particular, we use the exact error probability as the performance metric and characterize the corresponding optimal attack and defense strategies.

The derived algorithms could potentially be useful for the quickest detection setup [31]–[45]. In particular, consider a system where an attacker appears at an unknown time, and we are interested in detecting the presence of attacks with minimum delay (under certain delay metric). The presence of the attacker is reflected on the change of the distribution of the data, and hence the quickest detection framework can be employed. Most of the existing works on quickest detection assume that post-change distribution is known. In the setup with an attacker, this assumption may not be practical. The algorithms developed in this paper could be used to identify which distribution is most beneficial to the attacker and hence could be the most likely post-change distribution used by the attacker.

Compared with the conference paper [1], this journal paper provides several new contributions. Firstly, [1] focuses only on the hypothesis-aware scenario. In this journal paper, we also investigate the hypothesis-unaware scenario. Since the information about the underlying hypothesis is hidden from the adversary in this case, the design of the optimal adversary is more difficult. Secondly, we improve the design of optimal attack strategy so that it is more general and can be applied to both hypothesis-aware and hypothesis-unaware adversaries. Finally, we add more comprehensive numerical examples to illustrate the analytical results obtained in this paper.

The remainder of this paper is organized as follows. In Section II, we present our problem formulation. In Section III, we depict the optimal solution for the hypothesis-aware setting. In Section IV, we focus on the hypothesis-unaware case. In Section V, we provide numerical examples to illustrate the analytical results. In Section VI, we offer concluding remarks. The main notations used in the paper are listed in Table I.

Notation	Description
$\boldsymbol{p}_0, \boldsymbol{p}_1$	Original PMF under $\mathcal{H}_0$ and $\mathcal{H}_1$
$\boldsymbol{A}, \boldsymbol{B}$	Attack matrix under $\mathcal{H}_0$ and $\mathcal{H}_1$
$\boldsymbol{q}_0, \boldsymbol{q}_1$	PMF after attack under $\mathcal{H}_0$ and $\mathcal{H}_1$
$\boldsymbol{t}$	Decision rule
$I_{t,i}, K_{t,i}$	Moved mass between regions under $\mathcal{H}_t$
$F_j(\boldsymbol{A}, \boldsymbol{B})$	Upper-bound for prediction error at step j

TABLE I
MAIN NOTATIONS

II. Problem Formulation

Suppose there is a discrete random variable X defined on a finite set $\mathcal{X} = \{x_1, x_1, ..., x_n\}$. Consider the binary hypothesis testing problem:

$$\mathcal{H}_0 : X \sim \boldsymbol{p}_0,$$
$$\mathcal{H}_1 : X \sim \boldsymbol{p}_1,$$

in which $\boldsymbol{p}_0$ is a $1 \times n$ PMF vector with $p_{0,j} = \Pr(X = x_j|\mathcal{H}_0)$. $\boldsymbol{p}_1$ is defined in a similar manner. Here, $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are assumed to be known to both the adversary and the decision maker.

In this paper, we focus on adversary hypothesis testing problem. In the considered model, after a sample is generated, an adversary can modify it to another value. The decision maker then observes the corrupted data. We consider two different adversary models with different capabilities.

A. Hypothesis-aware adversary

We first consider a powerful hypothesis-aware adversary, who knows the true underlying hypothesis with which the sample is generated. The study of this worst-case scenario will provide performance limits of other adversary models. In the considered model, the attacker can conduct randomized attacks. In particular, after observing sample $X = x_i$, the adversary can change it to an attacked sample $X' = x_j$ with a certain probability, where X' is also a random variable defined on $\mathcal{X}$. Since the adversary knows the true underlying hypothesis, different attack rules can be applied depending on whether the true hypothesis is $\mathcal{H}_0$ or $\mathcal{H}_1$. We denote the attack strategy of the attacker as $(\boldsymbol{A}, \boldsymbol{B})$, in which the components of $\boldsymbol{A}$ are $A_{i,j} = \Pr(X' = x_j|X = x_i, \mathcal{H}_0)$ and the components of $\boldsymbol{B}$ are $B_{i,j} = \Pr(X' = x_j|X = x_i, \mathcal{H}_1)$.

Motivated by adversarial example phenomena studied in deep neural networks, we assume that the change introduced by the adversary has limited amplitude. In particular, as mentioned in Section I, an adversarial example is data that has been modified by the attacker to fool the classifier. However, to avoid human eye detection, the amplitude of these modifications should be limited so that they are not perceptible to human eyes [5]–[9], [21]–[27]. Formally, we assume $A_{i,j} = B_{i,j} = 0$ when $|i-j| > \delta$, in which δ denotes the largest change allowed. We will use $\mathcal{A}, \mathcal{B}$ to denote the whole sets of all amplitude-constrained attackers under $\mathcal{H}_0, \mathcal{H}_1$ correspondingly. For any given attack rule $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$, the PMF of X' can be written as $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$ under $\mathcal{H}_0$ and $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{B}$ under $\mathcal{H}_1$, with $q_{k,j} = \Pr(X' = x_j|\mathcal{H}_k)$, with $k = 0, 1$.

Let $\mathcal{T} = [0, 1]^n$ be the set of all decision rules. Denote $\boldsymbol{t} = [t_1, \cdots, t_n] \in \mathcal{T}$ as a randomized decision rule such that if $X = x_i$, the detector selects $\mathcal{H}_1$ with probability t_i, where $0 \leq t_i \leq 1$. For decision rule $\boldsymbol{t}$, the probability of false alarm and miss detection are

$$P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_0\boldsymbol{A}\boldsymbol{t}^T, P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t}) = \boldsymbol{p}_1\boldsymbol{B}(\boldsymbol{1} - \boldsymbol{t})^T. \quad (1)$$

Assuming that the prior probability of two hypotheses are equal, i.e., $\Pr(\mathcal{H}_0) = \Pr(\mathcal{H}_1)$, the error probability P_E can be written as

$$P_E(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) = \frac{1}{2}[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) + P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t})]. \quad (2)$$

In the following, to simplify the notation, we will drop $\boldsymbol{p}_0, \boldsymbol{p}_1$ from the expression of P_E and will simply write it as $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$.

The goal of the attacker is to choose the attack rule $(\boldsymbol{A}, \boldsymbol{B})$ to maximize the error probability (2), while the goal of the defender is to choose the decision rule $\boldsymbol{t}$ to minimize the error probability (2). In this paper, we seek to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ and $\boldsymbol{t}^*$ by solving the minimax problem

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \quad (3)$$

B. Hypothesis-unaware adversary

We also consider a more practical scenario, in which the attacker does not know the true underlying hypothesis when it sees a sample. In this hypothesis-unaware adversary case, there is only one attack matrix $\mathbf{A}$, with $A_{i,j} = \Pr(X' = x_j | X = x_i)$ being the probability that the attacker will change x_i to x_j.

Correspondingly, for a decision rule $\boldsymbol{t}$, the probability of false alarm and miss detection are

$$P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_0 \boldsymbol{A} \boldsymbol{t}^T, P_M(\boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_1 \boldsymbol{A}(\mathbf{1} - \boldsymbol{t})^T. \quad (4)$$

And the error probability P_E can be written as

$$P_E(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t}) = \frac{1}{2}[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) + P_M(\boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t})]. \quad (5)$$

Similarly, we will drop $\boldsymbol{p}_0, \boldsymbol{p}_1$ from the expression of P_E and will simply write it as $P_E(\boldsymbol{A}, \boldsymbol{t})$.

Moreover, we seek to characterize the optimal $\boldsymbol{A}^*$ and $\boldsymbol{t}^*$ by solving the minimax problem

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{\boldsymbol{A}\in\mathcal{A}} P_E(\boldsymbol{A}, \boldsymbol{t}). \quad (6)$$

In the problem formulations (3) and (6) discussed above, the distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$, i.e. $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$, are known to the attacker and decision maker. These problem formulations can be generalized to the scenario where there are uncertainties about the distributions. Suppose the actual distribution $\boldsymbol{p}_t, t = 0, 1$ under $\mathcal{H}_t$ belongs to the neighborhood of a nominal distribution. The neighborhood, denoted by $\mathcal{P}_t$ can be defined by KL-divergence [19], α-divergence [15], etc. The optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ and $\boldsymbol{t}^*$ for the hypothesis-aware case can be found by solving the complex optimization problem

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} \min_{(\boldsymbol{p}_0,\boldsymbol{p}_1)\in\mathcal{P}_0\times\mathcal{P}_1} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}, \boldsymbol{p}_0, \boldsymbol{p}_1).$$

Similarly, the optimal $\boldsymbol{A}^*$ and $\boldsymbol{t}^*$ for the hypothesis-unaware case can be found by solving the optimization problem

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{\boldsymbol{A}\in\mathcal{A}} \min_{(\boldsymbol{p}_0,\boldsymbol{p}_1)\in\mathcal{P}_0\times\mathcal{P}_1} P_E(\boldsymbol{A}, \boldsymbol{t}, \boldsymbol{p}_0, \boldsymbol{p}_1).$$

These problem formulations are much more complex than (3) and (6), and are left as future work.

III. Optimal hypothesis-aware adversary

In this section, we focus on the hypothesis-aware case and characterize the optimal solution to the complicated minimax optimization problem (3). To achieve this, we will first conduct a saddle-point analysis to reveal the structure of the optimal solution. Based on this, we will derive an upper-bound on the error probability. We will then develop an attack strategy to achieve this bound.

A. Saddle-point Analysis

In this subsection, we characterize the structure of the optimal decision rules by analyzing the saddle-point of the minimax problem (3).

Note that, given $\boldsymbol{t}$, $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$ is continuous and linear, and therefore is both convex and concave in $(\boldsymbol{A}, \boldsymbol{B})$. Similarly, given $(\boldsymbol{A}, \boldsymbol{B})$, $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$ is continuous and linear, and therefore is both convex and concave in $\boldsymbol{t}$. Furthermore, sets $\mathcal{A} \times \mathcal{B}$ and $\mathcal{T}$ are both compact and convex. Therefore, using Von Neumann minimax theorem [46] (which allows the swapping of the $\min$ and $\max$ operators under certain conditions), we have

$$\begin{aligned} &\min_{\boldsymbol{t}\in\mathcal{T}} \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) \\ &= \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} \min_{\boldsymbol{t}\in\mathcal{T}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \end{aligned} \quad (7)$$

This implies that the solution $(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*)$ to this minimax problem satisfies the saddle-point property

$$P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}) \geq P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*) \geq P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*). \quad (8)$$

From these two inequalities, we can characterize the structure of the optimal attack and decision strategies.

The first inequality in (8) indicates that the best decision rule must be the Bayesian test with respect to the best adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$. It is well known that, for a given arbitrary adversary attack rule $(\boldsymbol{A}, \boldsymbol{B})$, the optimal detection rule, denoted as $\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B})$, is simply a threshold rule

$$t_i^*(\boldsymbol{A}, \boldsymbol{B}) = \begin{cases} 0 & q_{0,i} > q_{1,i}, \\ \text{arbitrary} & q_{0,i} = q_{1,i}, \\ 1 & q_{0,i} < q_{1,i}, \end{cases} \quad (9)$$

where $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$, $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{B}$. For the optimal adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, the optimal decision rule is $\boldsymbol{t}^* = \boldsymbol{t}^*(\boldsymbol{A^*}, \boldsymbol{B^*})$.

With the optimal form of t^* in terms of $(\boldsymbol{A}, \boldsymbol{B})$ characterized in (9), we can then use the second inequality in (8) to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ by solving

$$\max_{\boldsymbol{A},\boldsymbol{B}} \quad \frac{1}{2}[\boldsymbol{p}_0\boldsymbol{A}(\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T + \boldsymbol{p}_1\boldsymbol{B}(1 - (\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T)], \quad (10)$$

$$\text{s.t.} \quad A_{i,j} \geq 0, B_{i,j} \geq 0, i, j = 1, .., n, \quad (11)$$

$$\sum_{j=1}^{n} A_{i,j} = 1, \sum_{j=1}^{n} B_{i,j} = 1, i = 1, .., n, \quad (12)$$

$$\mathbf{1}_{|i-j|>\delta} A_{i,j} = \mathbf{1}_{|i-j|>\delta} B_{i,j} = 0, i, j = 1, .., n, \quad (13)$$

in which $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Here, constraints (11) and (12) guarantee that each row of $\boldsymbol{A}$ and $\boldsymbol{B}$ is a conditional

PMF, while constraint (13) makes sure that the changes introduced by the attacker has a limited amplitude.

Once we solve (10) and obtain $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, the optimal $t^*(\boldsymbol{A}^*, \boldsymbol{B}^*)$ can be obtained by using (9). Due to the decision rule in (9), the objective function in (10) is a complicated function of $(\boldsymbol{A}, \boldsymbol{B})$. In the following, we will characterize the optimal solution to this challenging optimization problem under the following assumptions on $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$. Let $R_0 = \{i|p_{0,i} \geq p_{1,i}\}$ and $R_1 = \{i|p_{0,i} < p_{1,i}\}$, i.e., R_0 is the set of index where $p_{0,i}$ is larger while R_1 is the set of index where $p_{1,i}$ is larger. We will assume that R_0 (and hence R_1) is a consecutive region in $[1, n]$. Without loss of generality, we write $R_0 = \{i|1 \leq i \leq m\}$ and $R_1 = \{i|m+1 \leq i \leq n\}$.

We now compare this assumption with the assumptions used in the study of classic robust hypothesis testing [19], in which the nominal PMF is assumed to satisfy certain monotonicity and symmetry properties. Specifically, in [19], monotonicity means that $\frac{p_{1,i}}{p_{0,i}}$ is a monotonically increasing function of i and symmetry means $p_{1,n-i+1} = p_{0,i}, 1 \leq i \leq n$. It is easy to check that the monotonicity assumption implies the assumption made in this paper. Moreover, our assumption does not require the symmetry condition. Hence, our assumption is significantly weaker than the assumptions in [19].

B. Upper-bound for P_E

In this section, we develop an upper-bound on the objective function (10) that holds for any attack strategy.

We first present a lemma that simplifies $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*})$ into two equivalent forms, both of which will be used in the sequel.

Lemma 1: $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*})$ can be written as

$$P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*}) = \frac{1}{2} - \frac{1}{4}\sum_{i=1}^{n} |q_{0,i} - q_{1,i}| \quad (14)$$

$$= \frac{1}{2}\sum_{i=1}^{n} \min\{q_{0,i}, q_{1,i}\}. \quad (15)$$

Proof: Please see Appendix A. ∎

From (14), we can see that the most powerful attacker is the one that minimizes the ℓ_1 distance between $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$, which inspires us to optimize the error probability by components.

To proceed further, we denote the mass moved into $[1, i]$ as $I_{t,i}$ for $t = 0$ (i.e., under hypothesis $\mathcal{H}_0$) and $t = 1$ (i.e., under hypothesis $\mathcal{H}_1$) respectively. Similarly, define the mass moved out from $[1, i]$ as $K_{t,i}$. For example, for region $[1, m]$, we have

$$I_{1,m} = \sum_{j=m+1}^{m+\delta} p_{1,j} \left(\sum_{i=j-\delta}^{m} B_{j,i} \right),$$

$$K_{0,m} = \sum_{j=m+1-\delta}^{m} p_{0,j} \left(\sum_{i=m+1}^{j+\delta} A_{j,i} \right),$$

as shown in Fig. 1.

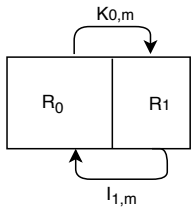

Fig. 1. Mass moved between two regions

Define

$$F_0 = F_0(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{n} q_{0,i},$$

$$F_j(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{j} \min\{q_{0,i}, q_{1,i}\} + \sum_{i=j+1}^{n} q_{0,i}.$$

Then we can see that

$$F_{j+1}(\boldsymbol{A}, \boldsymbol{B}) = F_j(\boldsymbol{A}, \boldsymbol{B}) + \min\{q_{1,j+1} - q_{0,j+1}, 0\},$$

and thus

$$2P_E(\boldsymbol{A}, \boldsymbol{B}) = F_n(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_m(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_0.$$

We are now ready to derive an upper-bound on the error probability P_E that holds for any attack strategy $(\boldsymbol{A}, \boldsymbol{B})$.

Theorem 1: For $\forall(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$,

$$F_m(\boldsymbol{A}, \boldsymbol{B}) \leq \min\left\{1, \min_{1+\delta \leq j \leq m} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}\right\}, \quad (16)$$

$$2P_E = F_n(\boldsymbol{A}, \boldsymbol{B}) \leq \min\left\{1, \min_{1+\delta \leq j \leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}\right\}, \quad (17)$$

in which

$$G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{j-\delta} p_{0,i} + \sum_{i=1}^{\min\{n, j+\delta\}} p_{1,i}. \quad (18)$$

Furthermore, for $j^* = \arg\min_{1+\delta \leq j \leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$, if $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq 1$, the equality in (17) holds when there exists $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$ such that:

(i) $q_{1,i} \leq q_{0,i}, 1 \leq i \leq j^*$;

(ii)

$$K_{0,j^*} - I_{0,j^*} = \sum_{i=j^*-\delta+1}^{j^*} p_{0,i}, \quad (19)$$

$$I_{1,j^*} - K_{1,j^*} = \sum_{i=j^*+1}^{\min\{n, j^*+\delta\}} p_{1,i}; \quad (20)$$

(iii) $F_k(\boldsymbol{A}, \boldsymbol{B}) = F_{j^*}(\boldsymbol{A}, \boldsymbol{B}), j^* < k \leq n$.

If $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > 1$, the equality is achieved when

$$F_i(\boldsymbol{A}, \boldsymbol{B}) = 1, 1 \leq i \leq n. \quad (21)$$

Proof: Please see Appendix B. ∎

We note that the bound in Theorem 1 depends only on $(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the original PMFs before attack.

C. Optimal Adversary Design

In this section, we design the attack matrix $(\boldsymbol{A}, \boldsymbol{B})$ to achieve the upper-bound in (17). As the designed attack matrix achieves the upper-bound, it is an optimal solution to (10).

The construction process is motivated by the form in (14), which shows that the component-wise absolute difference (ℓ_1 distance) between $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$ needs to be minimized. To minimize the ℓ_1 distance, we find the optimal $(\boldsymbol{A}, \boldsymbol{B})$ column by column. In particular, at the first step, we determine $\boldsymbol{A}_{:,1}, \boldsymbol{B}_{:,1}$ (based on some criteria to be detailed in the sequel). Once $\boldsymbol{A}_{:,1}, \boldsymbol{B}_{:,1}$ are determined, $q_{t,1}$ and F_1 are also determined. We denote these values as $\hat{q}_{t,1}$ and $\hat{F}_1$ respectively. We also have the constrained attack set $\mathcal{A}_1 \times \mathcal{B}_1 = \{(\boldsymbol{A}, \boldsymbol{B}) | \hat{q}_{0,1} \text{ and } \hat{q}_{1,1} \text{ are obtained}\}$. After step $j-1$, the first $j-1$ columns have been determined, and the constrained set is $\mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$. Then at step j, among all valid attack matrices in $\mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$, we determine $\boldsymbol{A}_{:,j}$, $\boldsymbol{B}_{:,j}$ (based on a process to be detailed in the sequel) and obtain $\hat{q}_{t,j}, \hat{F}_j$. The constrained set is further refined to be $\mathcal{A}_j \times \mathcal{B}_j = \{(\boldsymbol{A}, \boldsymbol{B}) | \hat{q}_{0,j} \text{ and } \hat{q}_{1,j} \text{ are obtained}\} \subset \mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$. The process ends at step n.

In the following, we describe our design of $(\boldsymbol{A}, \boldsymbol{B})$ to achieve the upper-bound. We will first focus on $1 \le j \le m$, i.e., $j \in R_0$, to obtain the equality in (16). Then focus on $m+1 \le j \le n$, i.e., $j \in R_1$, to achieve the equality in (17).

Column design for $j \in R_0$:

In R_0, we design the columns of $(\boldsymbol{A}, \boldsymbol{B})$ to satisfy

1) $\hat{q}_{1,1} = \sum_{i=1}^{1+\delta} p_{1,i}$;
2) $\hat{q}_{1,j} = p_{1,j+\delta}, 2 \le j \le m$;
3) $\hat{q}_{0,j} = \hat{q}_{1,j}, 1 \le j \le \delta$;
4) $\hat{q}_{0,j} = \max\{p_{0,j-\delta} A_{j-\delta,j}, \hat{q}_{1,j}\}, \delta+1 \le j \le m$,

which will then be shown to achieve the optimal value of $F_m(\boldsymbol{A}, \boldsymbol{B})$ in (16). These conditions are also listed in Table II.

	$j=1$	$2 \le j \le \delta$	$\delta+1 \le j \le m$
$\mathcal{H}_0 : \hat{q}_{0,j}$	$\sum_{i=1}^{1+\delta} p_{1,i}$	$p_{1,j+\delta}$	$\max\{p_{0,j-\delta} A_{j-\delta,j}, p_{1,j+\delta}\}$
$\mathcal{H}_1 : \hat{q}_{1,j}$	$\sum_{i=1}^{1+\delta} p_{1,i}$	$p_{1,j+\delta}$	$p_{1,j+\delta}$

TABLE II
PMF DESIGN IN R_0

First, we specify how to design each element of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ so that $\hat{q}_{0,j}$s and $\hat{q}_{1,j}$s are set to be these values.

For the first step, by 1), 3), we have $\hat{q}_{0,1} = \hat{q}_{1,1} = \sum_{i=1}^{1+\delta} p_{1,i}$, and thus $\hat{F}_1 = F_0 = 1$. Moreover, we can achieve this by setting the first column of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ as

$$\hat{A}_{1,1} = \min\left\{1, \frac{\hat{q}_{0,1}}{p_{0,1}}\right\},$$

$$\hat{A}_{i,1} = \min\left\{1, \frac{\max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}}{p_{0,i}}\right\}, 2 \le i \le n,$$

$$\hat{B}_{i,1} = 1, 1 \le i \le 1+\delta, B_{i,1} = 0, 2+\delta \le i \le n.$$

We continue to next columns. For columns $2 \le j \le \delta$, from 2) and 3), we have $\hat{q}_{0,j} = \hat{q}_{1,j} = p_{1,j+\delta}$, then $\hat{F}_j = \hat{F}_{j-1}$. We can achieve this by setting the j-th columns of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ as

$$\forall 1 \le i \le n, \hat{A}_{i,j} = \min\left\{1 - \sum_{k=1}^{j-1} \hat{A}_{i,k}, \frac{\max\left\{\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right), 0\right\}}{p_{0,i}}\right\}, \tag{22}$$

$$\hat{B}_{j+\delta,j} = 1, \hat{B}_{i,j} = 0, \forall i \ne j+\delta. \tag{23}$$

For columns $\delta+1 \le j \le m$, $\hat{\boldsymbol{A}}_{:,j}$ and $\hat{\boldsymbol{B}}_{:,j}$ are also designed by (22) and (23).

In Appendix C, we show that, with this design of $\hat{\boldsymbol{A}}_{:,j}$ and $\hat{\boldsymbol{B}}_{:,j}$, the requirements in 1), 2), 3), 4) are satisfied.

Remark 1: The column design for $\boldsymbol{A}$ in (22) indicates that

$$\forall 1 \le i \le n, \hat{A}_{i,j} = \min\left\{A_{i,j}^{(1)}, \max\left\{A_{i,j}^{(2)}, A_{i,j}^{(3)}\right\}\right\},$$

in which

$$A_{i,j}^{(1)} = 1 - \sum_{k=1}^{j-1} \hat{A}_{i,k} = \max_{\boldsymbol{A} \in \mathcal{A}_{j-1}} A_{i,j},$$

$$A_{i,j}^{(2)} = \frac{\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right)}{p_{0,i}},$$

$$A_{i,j}^{(3)} = 0 = \min_{\boldsymbol{A} \in \mathcal{A}_{j-1}} A_{i,j}.$$

For a given component j, looking at i which starts from 1 and goes to n, we notice that the value of $\hat{A}_{i,j}$ will goes from $A_{i,j}^{(1)}$, to $A_{i,j}^{(2)}$ and then $A_{i,j}^{(3)}$.

Second, we show that $\hat{F}_m$ achieves the equality in (16) by checking the values of $\hat{F}_j$ one by one from $j = \delta+1$ to $j = m$. We have three cases that will occur in order as j increases.

Case 1: $\hat{q}_{0,j} = \hat{q}_{1,j}$, then $\hat{F}_j = \hat{F}_{j-1}$.

Case 2: j is the first component such that $\hat{q}_{0,j} > \hat{q}_{1,j}$, or equivalently, j is the smallest component satisfying

$$\sum_{i=1}^{j} \hat{q}_{0,i} > \sum_{i=1}^{j} \hat{q}_{1,i} = \sum_{i=1}^{j+\delta} p_{1,i}.$$

This means that we have $\hat{F}_{j-1} = \hat{F}_{j-2} = ... = 1$. As for $\hat{q}_{0,j}$, if $\hat{q}_{0,j} \ne \hat{q}_{1,j}$, then $\hat{q}_{0,j} = p_{0,j-\delta} \hat{A}_{j-\delta,j} > 0$ and thus

$$\sum_{i=1}^{j} \hat{q}_{0,i} \overset{(a)}{=} \sum_{i=1}^{j-\delta} p_{0,i} > \sum_{i=1}^{j} \hat{q}_{1,i} = \sum_{i=1}^{j+\delta} p_{1,i}. \tag{24}$$

To derive (a), as discussed above, we have $\hat{A}_{j-\delta,j} > 0$, which indicates $\hat{A}_{j-\delta,j-1} \ne A_{j-\delta,j-1}^{(1)}$. Then $K_{0,j} = \sum_{i=j-\delta+1}^{j} p_{0,i}$ and (a) is true. Therefore,

$$\begin{aligned} \hat{F}_j &= 1 + \hat{q}_{1,j} - \hat{q}_{0,j} = 1 + \sum_{i=1}^{j} (\hat{q}_{1,j} - \hat{q}_{0,j}) \\ &= 1 + \sum_{i=1}^{j+\delta} p_{1,i} - \sum_{i=1}^{j-\delta} p_{0,i} = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1). \end{aligned} \tag{25}$$

Case 3: Suppose k is the largest component with

$$\hat{F}_k = G_k(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 + \sum_{i=1}^{k+\delta} p_{1,i} - \sum_{i=1}^{k-\delta} p_{0,i}. \quad (26)$$

Similar to Case 2, we have

$$\sum_{i=k+1}^{j} \hat{q}_{0,i} = \sum_{i=k-\delta+1}^{j-\delta} p_{0,i} > \sum_{i=k+1}^{j} \hat{q}_{1,i} = \sum_{i=k+\delta+1}^{j+\delta} p_{1,i}. \quad (27)$$

Therefore,

$$\begin{aligned}
\hat{F}_i &= \hat{F}_k, k+1 \le i \le j-1, \\
\hat{F}_j &= \hat{F}_k + \sum_{i=k+1}^{j} (\hat{q}_{1,i} - \hat{q}_{0,i}) \\
&\overset{(b)}{=} \hat{F}_k + \sum_{i=k+\delta+1}^{j+\delta} p_{1,i} - \sum_{i=k-\delta+1}^{j-\delta} p_{0,i} \\
&\overset{(c)}{=} 1 + \sum_{i=1}^{j+\delta} p_{1,i} - \sum_{i=1}^{j-\delta} p_{0,i} = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1), \quad (28)
\end{aligned}$$

where (b) is from (27) and (c) is true due to (26).

Taking all three cases into consideration, we have

$$\hat{F}_j = \min\left\{\hat{F}_{j-1}, G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}, \quad (29)$$

and thus $\hat{F}_m = \min\left\{1, \min_{1\le j\le m} G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}$, which achieves the equality in (16).

Column design for $j \in R_1$:

In R_1, we design the columns of $(\boldsymbol{A}, \boldsymbol{B})$ to satisfy

$$\begin{aligned}
1) \quad \hat{q}_{0,j} &= \max\left\{\min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j}, \right. \\
&\qquad \left. \min\left\{\max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}, \max_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j}\right\}\right\}, \quad (30) \\
2) \quad \hat{q}_{1,j} &= \max\left\{\min_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}, \right. \\
&\qquad \left. \min\left\{\hat{q}_{0,j}, \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}\right\}\right\}, \quad (31)
\end{aligned}$$

where

$$\begin{aligned}
&\min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} = p_{0,j-\delta} A_{j-\delta,j}, \\
&\max_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} = K_{0,j-1} - I_{0,j-1} + \sum_{i=j}^{j+\delta} p_{0,i}, \\
&\max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = \sum_{i=j}^{j+\delta} p_{1,i} - I_{1,j-1} + K_{1,j-1}, \\
&\min_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = p_{1,j-\delta} B_{j-\delta,j}.
\end{aligned}$$

First, we describe the construction of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$. Note that, the first m columns of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ have already been selected in R_0. For columns from $m+1$ to n, $\hat{\boldsymbol{A}}$ is constructed by (22) and $\hat{\boldsymbol{B}}$ is constructed by

$$\hat{B}_{i,j} = \min\left\{1 - \sum_{k=1}^{j-1} \hat{B}_{i,k}, \frac{\hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j}}{p_{1,i}}\right\}. \quad (32)$$

In Appendix C, we show that such $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ design satisfies the conditions on $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$ in $1), 2)$.

Second, we verify that $\hat{F}_n$ achieves the equality in (17) if the conditions on $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$ in $1), 2)$ are satisfied. The main idea is to derive the value of $\hat{F}_j$ based on the value of $\hat{F}_{j-1}$ by calculating $\hat{q}_{0,j} - \hat{q}_{1,j}$. According to the previously designed columns, the relationship between $\hat{q}_{0,j}$ and $\hat{q}_{1,j}$ has three different cases.

Case 1: $\hat{q}_{0,j} \le \hat{q}_{1,j}$, then $\hat{F}_j = \hat{F}_{j-1}$. Moreover, we have $\hat{q}_{0,j} \ne p_{0,j-\delta} A_{j-\delta,j}$ in this case.

Case 2: Assume that j is the smallest component in R_1 with $\hat{q}_{0,j} > \hat{q}_{1,j}$. Specifically, by $1), 2)$, for this component, we have

$$\begin{aligned}
\hat{q}_{0,j} &= \min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} \\
&= p_{0,j-\delta} A_{j-\delta,j} = K_{0,j-1} - I_{0,j-1} - \sum_{i=j-\delta+1}^{j-1} p_{0,i}, \\
\hat{q}_{1,j} &= \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = \sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} - I_{1,j-1} + K_{1,j-1}.
\end{aligned}$$

Then

$$\begin{aligned}
&\hat{q}_{0,j} - \hat{q}_{1,j} = p_{0,j-\delta} A_{j-\delta,j} - \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} \\
&= (K_{0,j-1} - I_{0,j-1} - \sum_{i=j-\delta+1}^{j-1} p_{0,i}) \\
&\quad - \left(\sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} - I_{1,j-1} + K_{1,j-1}\right) \\
&\overset{(a)}{=} \hat{F}_{j-1} - 1 + \sum_{i=1}^{j-1}(p_{0,i} - p_{1,i}) - \sum_{i=j-\delta+1}^{j-1} p_{0,i} \\
&\quad - \sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} \\
&= \hat{F}_{j-1} - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1),
\end{aligned}$$

in which (a) comes from the following fact,

$$\begin{aligned}
&F_j(\boldsymbol{A}, \boldsymbol{B}) = 1 + \sum_{i=1}^{j} \min\{(q_{1,i} - q_{0,i}), 0\} \\
&\overset{(b)}{\le} 1 + \sum_{i=1}^{j}(q_{1,i} - q_{0,i}) = 1 + \sum_{i=1}^{j} q_{1,i} - \sum_{i=1}^{j} q_{0,j} \\
&= 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j},
\end{aligned}$$

where the equality in (b) is attained when $q_{1,i} \le q_{0,i}, 1 \le i \le j$. Recall that for $i \in R_0$, we have $\hat{q}_{0,i} \ge \hat{q}_{1,i}$. Then based on

the assumption, we have $\hat{q}_{1,i} \leq \hat{q}_{0,i}, 1 \leq i \leq j$ and hence (a) is true.

Recall that $j^* = \arg\min_{1\leq j\leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$ and we prove that $j^* \in R_1$ by contradiction. Suppose $j^* \in R_0$. Then note that $\hat{F}_m = G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq G_j(\boldsymbol{p}_0, \boldsymbol{p}_1), \forall j \in R_1$ and thus

$$\begin{aligned} \hat{q}_{0,j} - \hat{q}_{1,j} &= \hat{F}_{j-1} - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) \\ &\leq \hat{F}_m - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq 0, \end{aligned}$$

which contradicts with the assumption that $\hat{q}_{0,j} > \hat{q}_{1,j}$. Hence, $j^* \in R_1$ and we have $\hat{F}_j = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)$.

Case 3: For $k > j$ such that $\hat{q}_{0,j} = p_{0,j-\delta} A_{j-\delta,j} > \hat{q}_{1,j}$, by the similar idea of proving (28) in R_0, we will also have $\hat{F}_j = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)$.

Taking all three cases into consideration, in R_1, (29) also holds, which indicates that the equality in Theorem 1 is obtained for the designed adversary.

IV. Optimal hypothesis-unaware adversary

In Section III, we have considered a powerful hypothesis-aware adversary who knows the true underlying hypothesis before attack. In this section, we consider a more practical scenario with a hypothesis-unaware adversary who does not know the true underlying hypothesis that generates the observed data. In this section, we will investigate the optimal solution to the minimax problem characterized in (6).

Under the hypothesis-unaware setting, as the adversary has less information, the attack is more difficult to carry out. However, the approach in Section III can be modified and applied here.

First, the saddle point analysis in Section III-A can be easily extended to the hypothesis-unaware case to simplify (6). In particular, following a similar saddle-point analysis, for any given attack matrix $\boldsymbol{A}$, we have that the optimal form of the decision rule is

$$t_i^*(\boldsymbol{A}) = \begin{cases} 0 & q_{0,i} > q_{1,i}, \\ \text{arbitrary} & q_{0,i} = q_{1,i}, \\ 1 & q_{0,i} < q_{1,i}, \end{cases} \tag{33}$$

where $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$, $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{A}$. The optimal attack matrix $\boldsymbol{A}^*$ is the solution to

$$\max_{\boldsymbol{A}} \quad \frac{1}{2}[\boldsymbol{p}_0\boldsymbol{A}(\boldsymbol{t}^*(\boldsymbol{A})^T + \boldsymbol{p}_1\boldsymbol{A}(1 - \boldsymbol{t}^*(\boldsymbol{A})].$$

This can be further rewritten as

$$\begin{aligned} \max_{\boldsymbol{A}} \quad & \frac{1}{2}[1 + (\boldsymbol{p}_0 - \boldsymbol{p}_1)\boldsymbol{A}\boldsymbol{t}^*(\boldsymbol{A})^T], \\ \text{subject to} \quad & A_{i,j} \geq 0, i, j = 1, .., n, \\ & \sum_{j=1}^{n} A_{i,j} = 1, \\ & 1_{|i-j|>\delta} A_{i,j} = 0, i, j = 1, .., n. \end{aligned} \tag{34}$$

In the following, we will generalize the approach in Section III to characterize the optimal solution to (34).

A. Upper-bound for P_E

Let $F_{m-\delta}(\boldsymbol{A}) = \sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{n} q_{0,i}$ and

$$f(j, \boldsymbol{A}) = \sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{j} \min\{q_{0,i}, q_{1,i}\} + \sum_{i=j+1}^{n} q_{0,i}.$$

Define

$$F_j(\boldsymbol{A}) = \begin{cases} F_{m-\delta}(\boldsymbol{A}) & 1 \leq j \leq m-\delta, \\ f(j, \boldsymbol{A}) & m-\delta+1 \leq j \leq m+\delta, \\ f(m+\delta, \boldsymbol{A}) & m+\delta+1 \leq j \leq n. \end{cases}$$

Then from the definition, we have

$$F_{j+1}(\boldsymbol{A}, \boldsymbol{B}) = F_j(\boldsymbol{A}, \boldsymbol{B}) + \min\{q_{1,j+1} - q_{0,j+1}, 0\},$$

and thus

$$\begin{aligned} &2P_E(\boldsymbol{A}, \boldsymbol{B}) \overset{(a)}{=} F_n(\boldsymbol{A}, \boldsymbol{B}) = ... = F_{m+\delta}(\boldsymbol{A}, \boldsymbol{B}) \leq ... \\ &\leq F_m(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_{m-\delta} = F_{m-\delta-1} = ... = F_0, \end{aligned}$$

where (a) is due to the fact that

$$\begin{aligned} &\sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{m+\delta} \min\{q_{0,i}, q_{1,i}\} + \sum_{i=m+\delta+1}^{n} q_{0,i} \\ &= \sum_{i=1}^{n} \min\{q_{0,i}, q_{1,i}\}. \end{aligned}$$

Similar to Theorem 1, we have the following bound.

Theorem 2: For $\forall \boldsymbol{A} \in \mathcal{A}$,

$$F_m(\boldsymbol{A}) \leq \min_{m-\delta\leq j\leq m} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}, \tag{35}$$

$$\begin{aligned} 2P_E(\boldsymbol{A}) = F_{m+\delta}(\boldsymbol{A}) &\leq \min_{m-\delta\leq j\leq m+\delta} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\} \\ &:= E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1), \end{aligned} \tag{36}$$

in which

$$\begin{aligned} &E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{m-2\delta} (p_{0,i} - p_{1,i}), \\ &E_j(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{j-\delta} (p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{\min\{n, j+\delta\}} (p_{1,i} - p_{0,i}), \\ &j^* = \arg \min_{m-\delta\leq j\leq m+\delta} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}. \end{aligned} \tag{37}$$

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality in (36) holds when there exists $\boldsymbol{A} \in \mathcal{A}$ such that:

(i) $K_{0,m-\delta} - K_{1,m-\delta} = \sum_{i=m-2\delta+1}^{m-\delta} (p_{0,i} - p_{1,i})$;
(ii) $q_{1,i} \leq q_{0,i}, m-\delta+1 \leq i \leq j^*$;
(iii) $K_{0,j^*} - K_{1,j^*} = \sum_{i=j^*-\delta+1}^{\min\{j^*,m\}} (p_{0,i} - p_{1,i})$,
$I_{1,j^*} - I_{0,j^*} = \sum_{i=\max\{m+1,j^*+1\}}^{\min\{n,j^*+\delta\}} (p_{1,i} - p_{0,i})$;
(iv) $F_k(\boldsymbol{A}) = F_{j^*}(\boldsymbol{A}), j^* < k \leq m+\delta$.

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality is achieved when

$$F_i(\boldsymbol{A}) = E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1), m-\delta \leq i \leq m+\delta.$$

Proof: Please see Appendix D. ■

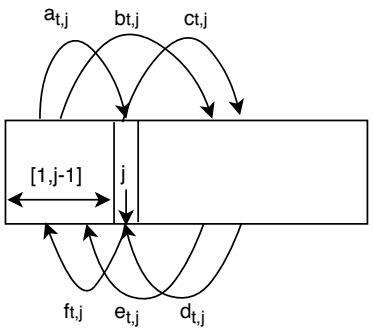

Fig. 2. Moved mass between different regions at component j

B. Attack strategy design

In this section, we design an attack matrix $\hat{\boldsymbol{A}}$ to achieve the upper-bound in (36). As the designed matrix achieves the upper-bound, it is an optimal solution to (34). Similar to the design of hypothesis-aware attack matrix, we construct the optimal $\boldsymbol{A}$ column by column.

Before proceeding further, we need to define quantities related to mass moving between different regions. In particular, for $t = 0, 1$, define

- $a_{t,j}$: $[1, j-1] \to j$,
- $b_{t,j}$: $[1, j-1] \to [j+1, n]$,
- $c_{t,j}$: $j \to [j+1, n]$,
- $d_{t,j}$: $[j+1, n] \to j$,
- $e_{t,j}$: $[j+1, n] \to [1, j-1]$,
- $f_{t,j}$: $j \to [1, j-1]$.

These quantities are illustrated in Fig. 2.

Moreover, we will use $\hat{a}, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{f}$ to denote the value of a, b, c, d, e, f determined by $\hat{\boldsymbol{A}}$ while using $\hat{F}_j$ to denote the value of F_j achieved by $\hat{\boldsymbol{A}}$.

Column design for $j \in R_0$:

In R_0, for $t = 0, 1$, we design columns of attack matrix $\hat{\boldsymbol{A}}$ to achieve

1) $\hat{q}_{t,j} = p_{t,j}, 1 \le j \le m - 2\delta$;
2) $\hat{q}_{t,j} = 0, m - 2\delta + 1 \le j \le m - \delta$;
3) $\hat{q}_{t,j} = p_{t,j-\delta} + \hat{d}_{t,j}, m - \delta + 1 \le j \le m$, where $\hat{d}_{t,j}$ is selected to satisfy

$$\hat{d}_{1,j} - \hat{d}_{0,j} = \min\{p_{0,j-\delta} - p_{1,j-\delta}, \hat{F}_{m-\delta} - \hat{F}_{j-1} + \sum_{i=m+1}^{\min\{n,j+\delta\}} (p_{1,i} - p_{0,i}) + \sum_{i=m-2\delta+1}^{j-\delta-1} (p_{1,i} - p_{0,i})\}.$$

To summarize, these conditions listed in Table III.

	$\mathcal{H}_t : \hat{q}_{t,j}$
$1 \le j \le m - 2\delta$	$p_{t,j}$
$m - 2\delta + 1 \le j \le m - \delta$	0
$m - \delta + 1 \le j \le m$	$p_{t,j-\delta} + \hat{d}_{t,j}$

TABLE III
PMF DESIGN IN R_0 FOR THE HYPOTHESIS-UNAWARE ADVERSARY

Here, again, we will first describe how to design $\hat{\boldsymbol{A}}$ so that 1), 2) and 3) are satisfied. We will then show that, once these conditions are satisfied, the equality in (35) is achieved. Hence, the designed $\hat{\boldsymbol{A}}$ is optimal.

In particular, we set columns 1 to m of $\hat{\boldsymbol{A}}$ to be

a) $1 \le j \le m - 2\delta, \hat{A}_{j,j} = 1, \hat{A}_{i,j} = 0, i \ne j$;
b) $m - \delta + 1 \le j \le m, \hat{A}_{j-\delta,j} = 1$,
$\hat{A}_{i,j} = \min\left\{1, \max\left\{\frac{\hat{d}_{1,j} - \hat{d}_{0,j} - \sum_{k=m+1}^{i-1}(p_{1,k} - p_{0,k})}{p_{1,i} - p_{0,i}}, 0\right\}\right\}$,
$m + 1 \le i \le n$.

Following the same proof in Appendix C, we can show that using design specified in a), b), the equalities in $1), 2), 3)$ are satisfied for $1 \le j \le m$. Details of the proofs are omitted for brevity.

We now verify that we can achieve the equality in the upper-bound (35) once conditions 1), 2) and 3) are satisfied.

$$\begin{aligned}\hat{F}_{m-\delta} &= \sum_{i=1}^{m-\delta}(p_{1,i} - p_{0,i}) + I_{1,m-\delta} - I_{0,m-\delta} - K_{1,m-\delta} \\ &\quad + K_{0,m-\delta} + 1 \\ &= 1 - \sum_{i=1}^{m-2\delta}(p_{0,i} - p_{1,i}) \coloneqq E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1).\end{aligned}$$

For $\forall m - \delta \le j \le m$,

$$\begin{aligned}\hat{F}_j &= \hat{F}_{j-1} + \min\{0, \hat{q}_{1,j} - \hat{q}_{0,i}\} \\ &= \min\{\hat{F}_{j-1}, \hat{F}_{j-1} + \hat{q}_{1,j} - \hat{q}_{0,i}\} \\ &= \min\{\hat{F}_{j-1}, \hat{F}_{j-1} + p_{1,j-\delta} - p_{0,j-\delta} + \hat{d}_{1,j} - \hat{d}_{0,i}\} \\ &= \min\left\{\hat{F}_{j-1}, 1 - \sum_{i=1}^{j-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{\min\{n,j+\delta\}}(p_{1,i} - p_{0,i})\right\} \\ &\coloneqq \min\left\{\hat{F}_{j-1}, E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\},\end{aligned}$$

and thus

$$\hat{F}_m = \min_{m-\delta \le j \le m}\left\{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}, \tag{38}$$

which reaches the equality in (35).

Column design for $j \in R_1$:

In R_1, the first m columns of $\hat{\boldsymbol{A}}$ have been determined in R_0 and for $j \in R_1$, we further design $\boldsymbol{A}_{:,m+1:n}$ to achieve

$$\begin{aligned}\hat{q}_{1,j} - \hat{q}_{0,j} &= \max\{\min_{\boldsymbol{A} \in \mathcal{A}_{j-1}}(q_{1,j} - q_{0,j}), \\ &\qquad \min\{0, \max_{\boldsymbol{A} \in \mathcal{A}_{j-1}}(q_{1,j} - q_{0,j})\}\}.\end{aligned}$$

We will design $\boldsymbol{A}_{:,m+1:n}$ in two cases. For the first case, we always have $j^* \in R_0$ and $\boldsymbol{A}_{:,m+1:n}$ can be designed in a simple way. For the second case, similar procedure in Section III-C **Case 2** can be applied. In the following part, we will provide the assumptions of two cases and analyze the first scenario in detail while skip the details for the second scenario.

Case 1:

$$\min_{m-\delta \le j \le m-1}\{\sum_{i=j-\delta+1}^{m}(p_{0,i} - p_{1,i}) - \sum_{i=j+\delta+1}^{m+\delta}(p_{1,i} - p_{0,i})\} \le 0.$$

By applying (38), this condition is equivalent to

$$\hat{F}_m \leq 1 + \sum_{i=1}^{m+\delta}(p_{1,i} - p_{0,1}),$$

and thus $\forall j \in [m+1, m+\delta]$,

$$E_j(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{j-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{j+\delta}(p_{1,i} - p_{0,i})$$
$$\geq 1 - \sum_{i=1}^{m}(p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{m+\delta}(p_{1,i} - p_{0,1}) \geq \hat{F}_m.$$

Therefore, we will be able to find an $\hat{\boldsymbol{A}} \in \mathcal{A}_m$, such that $\hat{F}_{m+\delta} = \hat{F}_m$.

The desired $\boldsymbol{A}_{:,m+1:n}$ is designed by

1) $\forall m - \delta + 1 \leq j \leq m, \hat{A}_{j,m+1} = 1$;
2) $\forall m + 1 \leq j \leq m + \delta, \hat{A}_{j,m+1} = 1 - \sum_{i=1}^{m} \hat{A}_{j,i}$;
3) $\forall m + \delta + 1 \leq j \leq n, \hat{A}_{j,j} = 1$.

Then we have

$$\begin{aligned}
&\hat{q}_{1,m+1} - \hat{q}_{0,m+1} = \sum_{k=m-\delta+1}^{m+\delta+1}(p_{1,k} - p_{0,k})\hat{A}_{k,m+1} \\
= \;& K_{1,m} - K_{0,m} + \sum_{k=m+1}^{m+\delta+1}(p_{1,k} - p_{0,k})(1 - \sum_{i=1}^{m}\hat{A}_{k,i}) \\
= \;& K_{1,m} - K_{0,m} + \sum_{k=m+1}^{m+\delta+1}(p_{1,k} - p_{0,k}) - I_{1,m} + K_{0,m} \\
\overset{(a)}{=} \;& 1 + \sum_{i=1}^{m+\delta}(p_{1,i} - p_{0,1}) - \hat{F}_m \geq 0,
\end{aligned}$$

where (a) is because $\forall m - \delta + 1 \leq j \leq m + \delta, \forall \boldsymbol{A} \in \mathcal{A}$,

$$\begin{aligned}
F_j(\boldsymbol{A}) &= F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j}\min\{(q_{1,i} - q_{0,i}), 0\} \\
&\overset{(b)}{\leq} F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j}(q_{1,i} - q_{0,i}) \\
&= \sum_{i=1}^{j} q_{1,i} + \sum_{i=j+1}^{n} q_{0,i} \\
&= 1 + \sum_{i=1}^{j}(q_{1,i} - q_{0,i}) \\
&= 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) \\
&\quad + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j},
\end{aligned}$$

and the equality in (b) holds when $q_{1,i} - q_{0,i} \leq 0, m - \delta + 1 \leq i \leq j$. For here, $j = m$ and we have $q_{1,i} - q_{0,i} \leq 0$ in R_0 and (a) is true.

Furthermore, we have $\hat{q}_{1,j} = \hat{q}_{0,j} = 0, m + 2 \leq j \leq m + \delta$. Therefore, for the designed $\hat{\boldsymbol{A}}$, we have

$$\begin{aligned}
&\hat{F}_{m+\delta}(\hat{\boldsymbol{A}}) = \hat{F}_{m+1}(\hat{\boldsymbol{A}}) \\
= \;& \hat{F}_m(\hat{\boldsymbol{A}}) + \min\{\hat{q}_{1,m+1} - \hat{q}_{0,m+1}, 0\} = \hat{F}_m(\hat{\boldsymbol{A}}),
\end{aligned}$$

and thus the equality in Theorem 2 is achieved.

Case 2:

$$\min_{m-\delta \leq j \leq m-1}\{\sum_{i=j-\delta+1}^{m}(p_{0,i} - p_{1,i}) - \sum_{i=j+\delta+1}^{m+\delta}(p_{1,i} - p_{0,i})\} > 0.$$

Under this condition, by the same idea in III-C **Case 2**, we will have $\hat{F}_j = \min\{\hat{F}_{j-1}, E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$. Therefore, the equality in Theorem 2 is attained.

V. Numerical Results

In this section, we provide numerical examples to illustrate results obtained in this paper.

In the first example, we give two specific PMFs with a few components and perform hypothesis-aware and hypothesis-unaware attacks to show how the adversary works. In this example, the PMF before attack is provided in (39) and Fig. 3. It is easy to calculate that for this PMF, if there is no adversary, the error probability corresponding to the optimal Bayesian detection rule is $P_E = \frac{11}{32}$. Assume that the attack amplitude is $\delta = 1$. Following the design process in Section III-C and IV-B, the optimal hypothesis-aware attack strategy $\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a$ and the optimal hypothesis-unaware attack strategy $\hat{\boldsymbol{A}}_u$ are

$$\hat{\boldsymbol{A}}_a = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{7} & \frac{4}{7} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{B}}_a = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{A}}_u = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Thus, the PMFs after attack can be calculated and are provided in (40) and Fig. 4 for the hypothesis-aware model and the PMFs of hypothesis-unaware model are provided in (41) and Fig. 5. It is easy to check that, for the constructed adversary, the error probabilities are $P_E(\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a, t^*(\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a)) = \frac{1}{2}$ and $P_E(\hat{\boldsymbol{A}}_u, t^*(\hat{\boldsymbol{A}}_u)) = \frac{7}{16}$ correspondingly. Since the error probability is 1/2 (the largest possible value) for the hypothesis-aware attack, the designed attack matrix is clearly optimal. For the hypothesis-unaware attack, the error probability under $\hat{\boldsymbol{A}}_u$ is less than $\frac{1}{2}$. This already achieves the maximal value of $P_E(\boldsymbol{A}_u, t^*(\boldsymbol{A}_u))$ by Theorem 2, $2P_E(\boldsymbol{A}_u, t^*(\boldsymbol{A}_u)) \leq E_4(\boldsymbol{p}_0, \boldsymbol{p}_1) = \frac{7}{8}$. Therefore, for this particular example,

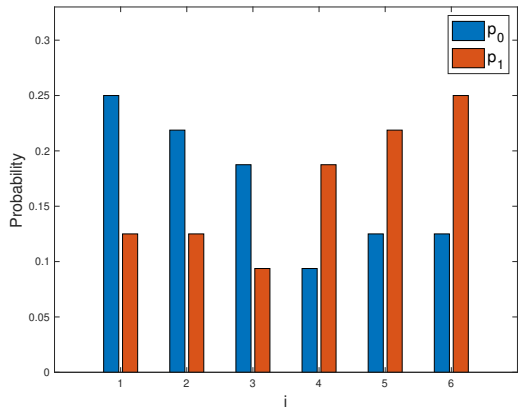

Fig. 3. PMFs $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$

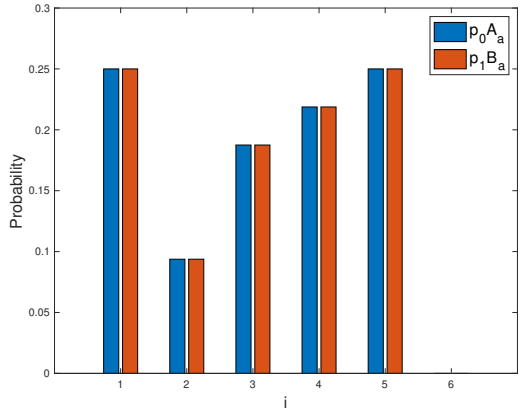

Fig. 4. PMFs $\boldsymbol{p}_0\hat{\boldsymbol{A}}_a$ and $\boldsymbol{p}_1\hat{\boldsymbol{B}}_a$ for the hypothesis-aware case

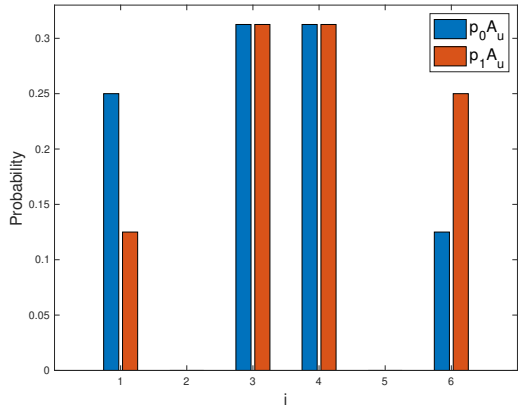

Fig. 5. PMFs $\boldsymbol{p}_0\hat{\boldsymbol{A}}_u$ and $\boldsymbol{p}_1\hat{\boldsymbol{A}}_u$ for the hypothesis-unaware case

the hypothesis-unaware attacker is not as powerful as the hypothesis-aware attacker.

$$\begin{array}{cccccccc} \boldsymbol{p}_0 & 8/32 & 7/32 & 6/32 & 3/32 & 4/32 & 4/32 \\ \boldsymbol{p}_1 & 4/32 & 4/32 & 3/32 & 6/32 & 7/32 & 8/32 \end{array} \tag{39}$$

$$\begin{array}{cccccccc} \boldsymbol{p}_0\hat{\boldsymbol{A}}_a & 8/32 & 3/32 & 6/32 & 7/32 & 8/32 & 0 \\ \boldsymbol{p}_1\hat{\boldsymbol{B}}_a & 8/32 & 3/32 & 6/32 & 7/32 & 8/32 & 0 \end{array} \tag{40}$$

$$\begin{array}{cccccccc} \boldsymbol{p}_0\hat{\boldsymbol{A}}_u & 8/32 & 0 & 10/32 & 10/32 & 0 & 4/32 \\ \boldsymbol{p}_1\hat{\boldsymbol{A}}_u & 4/32 & 0 & 10/32 & 10/32 & 0 & 8/32 \end{array} \tag{41}$$

In the second example, we explore how δ affects the prediction error for a randomly selected $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ under two attack models. In our simulation, we generate $2n$ random numbers in $[0, 1]$ by uniform distribution, divide them into two sequences and normalize each sequence to make it a PMF while maintaining two consecutive regions to meet the assumption made in Section III-A. After $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are generated, they are fixed throughout the experiment. We then apply the proposed attack schemes to find one of the optimal attackers and calculate its prediction error under the Bayesian test. The results are shown in Fig. 6, where both the upper-bounds for the error probability and the error probability under constructed optimal attackers are presented. There are only two lines in Fig. 6 since the upper-bounds are achieved by the designed adversary and they overlap each other, which verifies the correctness of the construction process. From Fig. 6, we can see that, for each adversary, the attacker becomes more powerful as δ increases. In particular, for the hypothesis-aware case, when δ is large enough, the prediction error will reach $\frac{1}{2}$, the largest possible value.

In the third example, we investigate the impact of the alphabet size n on the prediction error. The PMFs before the attack are generated in the same manner as the second example. From Fig. 7, we have that, for a fixed attack amplitude $\delta = 50$, the prediction error decreases as the alphabet size n increases. The reason is that, as n increases, the relative attack strength δ/n decreases, and hence the impact of the attack on the error probability also decreases. However, if the ratio between δ and n is fixed (for example, $\delta/n = 0.03, 0.06, 0.1$ as shown in Fig. 8 and $\delta/n = 0.1, 0.15, 0.2$ as shown in Fig. 9), there is no significant change in the prediction error as the alphabet size increases. In particular, from the hypothesis-aware result given in Fig. 8, we see that the prediction error reaches 0.5 when $\delta = 0.1n$ for n varies from 400 to 1000, indicating that even a relatively small perturbation could have a big impact on the prediction accuracy. On the other hand, for the hypothesis-unaware model, from Fig. 9, we see that it is harder for the prediction error to reach $\frac{1}{2}$, indicating that the strength of attack has been highly restricted if the hypothesis information is hidden from the adversary.

In the fourth example, we illustrate the characteristic of PMFs before and after attack. First, we generate the PMFs by truncating a Poisson distribution with parameter $\lambda_t, t = 0, 1$, since the normal Poisson distribution is defined on an infinite set. To normalize the distribution, we then move the mass on

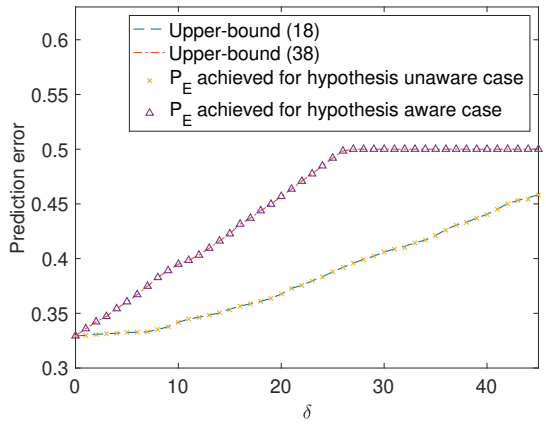

Fig. 6. Prediction error v.s. $\delta, n = 200, m = 97$

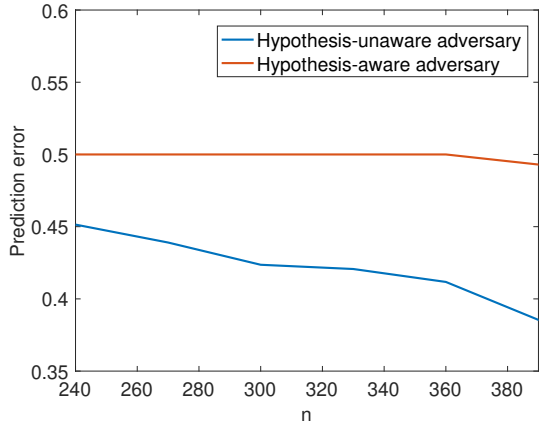

Fig. 7. Prediction error v.s. alphabet size n for $\delta = 50$

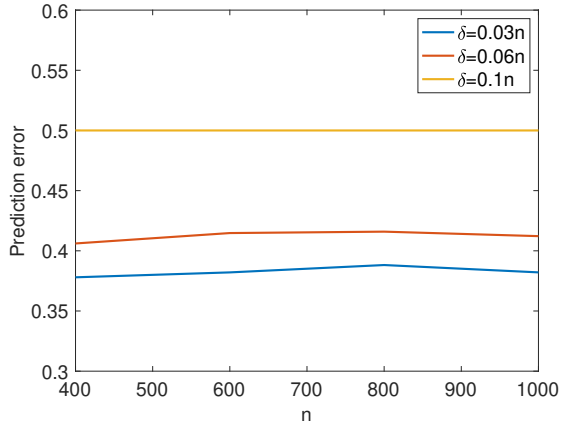

Fig. 8. Prediction error v.s. alphabet size n (hypothesis-aware)

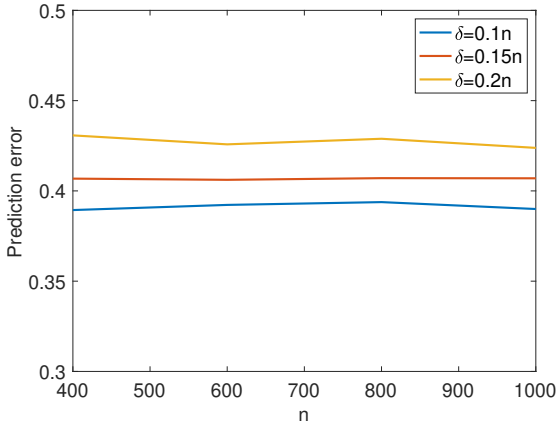

Fig. 9. Prediction error v.s. alphabet size n (hypothesis-unaware)

the tails to the finite alphabet equally and name the distribution as truncated Poisson distribution. Thus, the PMFs can be written as $\boldsymbol{p}_{t,i} = \boldsymbol{p}^0_{t,i} + d, t = 0, 1, 1 \leq i \leq n$, where $\boldsymbol{p}^0_{t,i} = \frac{(\lambda_t^i e^{-\lambda_t})}{i!}$ and $d = \frac{1-\sum_{i=1}^n \boldsymbol{p}^0_{t,i}}{n}$. By setting $n = 110$ and $\lambda_0 = 35, \lambda_1 = 75$ for $\mathcal{H}_0, \mathcal{H}_1$ respectively, the PMFs before attack are shown in Fig. 10. Under this setup, the PMFs after attack are shown in Fig. 10 and Fig. 11 for the hypothesis-aware and hypothesis-unaware attackers respectively. In these figures, we set $\delta = 24$. The results show that, for both hypothesis-aware and hypothesis-unaware adversary, the PMFs after attack can be made the same under two hypotheses. As the result, for both adversary models, $P_E = \frac{1}{2}$ after the attack.

Fig. 12 illustrates the PMFs before and after attack for the hypothesis-unaware case when $\delta = 20$. From this figure, we can see that, $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$, the PMFs after attack for different hypotheses, are not the same under the optimal hypothesis-unaware adversary. On the other hand, for the hypothesis-aware attacker, the error probability is equal to 1/2.

Fig. 13 illustrates how P_E increases as the attack amplitude δ increases. From this figure, we can see that, for both attack models, P_E increases with δ. Furthermore, the prediction error in the hypothesis-aware case is always larger than hypothesis-unaware case and reaches 1/2 earlier than the hypothesis-unaware case. This is consistent with the simulation result in the previous random distribution scenario.

VI. Conclusion

In this paper, we have investigated the adversarial robustness of hypothesis testing rules. We have formulated this as a minimax hypothesis testing problem. We have characterized the optimal attack and the corresponding optimal decision rules for both hypothesis-aware and hypothesis-unaware adversary models. We have also provided numerical examples to illustrate the analytical results obtained in this paper.

Building on the problem formulation and analysis in this paper, there are several interesting future research directions. Firstly, as discussed in Section I, it is important to extend the

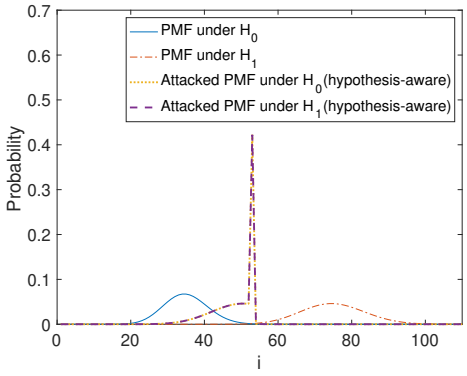

Fig. 10. The PMF before and after attack (hypothesis-aware)

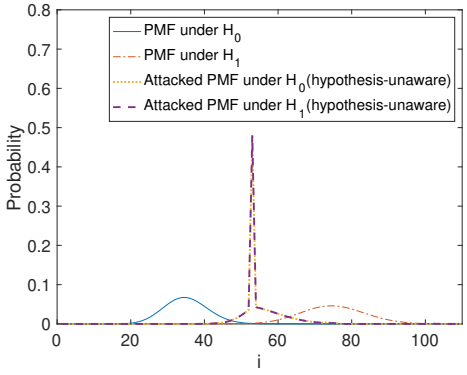

Fig. 11. The PMF before and after attack (hypothesis-unaware)

analysis to the scenario where the true underlying distributions are unknown to both the attacker and decision-maker. Secondly, the application to steganography and steganalysis [47], in which steganography aims to hide secret messages in the cover media while steganalysis tries to detect hidden secret information embedded in the cover media, is another interesting research direction. Thirdly, our work can be applied to the decentralized detection setup [48]–[50], with a fusion center and distributed nodes, some of which might be compromised. The compromised nodes may send fake messages to the fusion center, and the goal of the fusion center is to make correct decisions in spite of the presence of misbehaving nodes. Finally, other than the amplitude constraint considered in this paper, it is important to investigate other types of constraints on the adversary.

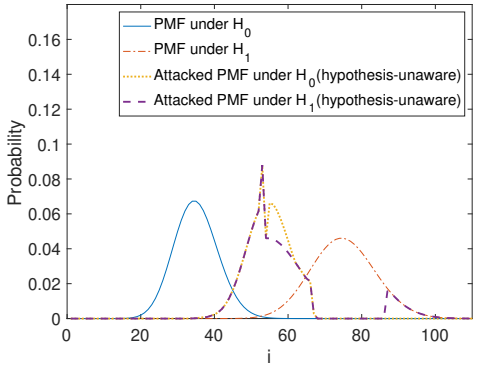

Fig. 12. The PMF before and after attack (hypothesis-unaware) when $\delta = 20$

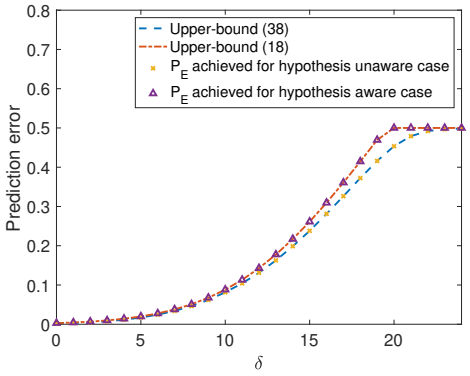

Fig. 13. Prediction error v.s. δ for truncated Poisson distribution

Appendix A
Proof of Lemma 1

We first prove (14):

$$\begin{aligned} P_E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{t}^*) &= \frac{1}{2}\left[P_F(\boldsymbol{p}_0,\boldsymbol{A},\boldsymbol{t}^*)+P_M(\boldsymbol{p}_1,\boldsymbol{B},\boldsymbol{t}^*)\right] \\ &= \frac{1}{2}\left[\sum_{i=1}^{n} q_{0,i}t_i^* + \sum_{i=1}^{n} q_{1,i}(1-t_i^*)\right] \\ &\overset{(a)}{=} \frac{1}{2}+\frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}}(q_{0,i}-q_{1,i}) \\ &= \frac{1}{2}-\frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}}|q_{0,i}-q_{1,i}| \qquad (42) \\ &\overset{(b)}{=} \frac{1}{2}-\frac{1}{4}\sum_{i=1}^{n}|q_{0,i}-q_{1,i}|\,. \end{aligned}$$

Here, (a) is true to due to the form of $\boldsymbol{t}^*$ specified in (9). We now show (b) is true:

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(q_{0,i}-q_{1,i}) \\ &= \sum_{i:q_{0,i}\geq q_{1,i}}(q_{0,i}-q_{1,i}) + \sum_{i:q_{0,i}<q_{1,i}}(q_{0,i}-q_{1,i}) \\ &= \sum_{i:q_{0,i}\geq q_{1,i}}|q_{0,i}-q_{1,i}| - \sum_{i:q_{0,i}<q_{1,i}}|q_{0,i}-q_{1,i}|\,, \end{aligned}$$

which implies

$$\begin{aligned} \sum_{i:q_{0,i}\geq q_{1,i}}|q_{0,i}-q_{1,i}| &= \sum_{i:q_{0,i}<q_{1,i}}|q_{0,i}-q_{1,i}| \\ &= \frac{1}{2}\sum_{i=1}^{n}|q_{0,i}-q_{1,i}|\,. \end{aligned}$$

We now prove (15). Using step (a) of (42), we have

$$
\begin{aligned}
P_E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{t^*}) &= \frac{1}{2}-\frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}}(q_{1,i}-q_{0,i})\\
&=\frac{1}{2}\left[\sum_{i:q_{0,i}\geq q_{1,i}} q_{1,i}+\sum_{i:q_{0,i}<q_{1,i}} q_{1,i}-\sum_{i:q_{0,i}<q_{1,i}}(q_{1,i}-q_{0,i})\right]\\
&=\frac{1}{2}\left[\sum_{i:q_{0,i}>q_{1,i}} q_{1,i}+\sum_{i:q_{0,i}=q_{1,i}} q_{1,i}+\sum_{i:q_{0,i}<q_{1,i}} q_{0,i}\right]\\
&=\frac{1}{2}\sum_{i=1}^{n}\min\{q_{0,i},q_{1,i}\}.
\end{aligned}
$$

Appendix B
Proof of Theorem 1

For $\forall(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}$, we have

$$
\begin{aligned}
F_j(\boldsymbol{A},\boldsymbol{B}) &= 1+\sum_{i=1}^{j}\min\{(q_{1,i}-q_{0,i}),0\}\\
&\overset{(a)}{\leq} 1+\sum_{i=1}^{j}(q_{1,i}-q_{0,i})\\
&= 1+\sum_{i=1}^{j}q_{1,i}-\sum_{i=1}^{j}q_{0,j}\\
&\quad -\left(\sum_{i=1}^{j}p_{0,i}-K_{0,j}+I_{0,j}\right)\\
&= 1-\sum_{i=1}^{j}(p_{0,i}-p_{1,i})+K_{0,j}-K_{1,j}+I_{1,j}-I_{0,j},\\
&\overset{(b)}{\leq} 1-\sum_{i=1}^{j}(p_{0,i}-p_{1,i})+\sum_{i=\max\{1,j-\delta+1\}}^{j}p_{0,i}\\
&\quad +\sum_{i=j+1}^{\min\{j+\delta,n\}}p_{1,i}\\
&= 1-\sum_{i=1}^{j-\delta}p_{0,i}+\sum_{i=1}^{j+\delta}p_{1,i}=G_j(\boldsymbol{p}_0,\boldsymbol{p}_1),
\end{aligned}
\tag{43}
$$

Here, the equality in (a) holds when $q_{1,i}-q_{0,i}\leq 0, 1\leq i\leq j$, inequality (b) comes from the natural restrictions on I,K, in which the equality holds when $K_{0,j}-I_{0,j}=\sum_{i=\max\{1,j-\delta+1\}}^{j}p_{0,i}$, and $I_{1,j}-K_{1,j}=\sum_{i=j+1}^{\min\{j+\delta,n\}}p_{1,i}$.

Note that $2P_E(\boldsymbol{A},\boldsymbol{B})=F_n(\boldsymbol{A},\boldsymbol{B})\leq ...\leq F_m(\boldsymbol{A},\boldsymbol{B})\leq ...\leq F_0=1$. As (43) holds for $\forall\boldsymbol{A},\boldsymbol{B}\in\Omega$, we have $F_n(\boldsymbol{A},\boldsymbol{B})\leq G_j(\boldsymbol{p}_0,\boldsymbol{p}_1),\forall 1\leq j\leq n$. Therefore,

$$
\begin{aligned}
F_m(\boldsymbol{A},\boldsymbol{B}) &\leq \min_{1\leq j\leq m}\{1,G_j(\boldsymbol{p}_0,\boldsymbol{p}_1)\},\\
F_n(\boldsymbol{A},\boldsymbol{B}) &\leq \min_{1\leq j\leq n}\{1,G_j(\boldsymbol{p}_0,\boldsymbol{p}_1)\}.
\end{aligned}
$$

Let $j^*=\arg\min_{1\leq j\leq n}\{G_j(\boldsymbol{p}_0,\boldsymbol{p}_1)\}$. If $G_{j^*}(\boldsymbol{p}_0,\boldsymbol{p}_1)\leq 1$, we have $F_n(\boldsymbol{A},\boldsymbol{B})\leq G_{j^*}(\boldsymbol{p}_0,\boldsymbol{p}_1)$ and the equality is achieved when

- $F_{j^*}(\boldsymbol{A},\boldsymbol{B})=G_{j^*}(\boldsymbol{p}_0,\boldsymbol{p}_1)$, which is equivalent to

$$
\begin{aligned}
&q_{1,i}-q_{0,i}\leq 0, 1\leq i\leq j^*,\\
&K_{0,j^*}-I_{0,j^*}=\sum_{i=\max\{1,j^*-\delta+1\}}^{j^*}p_{0,i},\\
&I_{1,j^*}-K_{1,j^*}=\sum_{i=j^*+1}^{\min\{j^*+\delta,n\}}p_{1,i}.
\end{aligned}
$$

- $F_k(\boldsymbol{A},\boldsymbol{B})=F_{j^*}(\boldsymbol{A},\boldsymbol{B}), j^*<k\leq n$.

If $G_{j^*}(\boldsymbol{p}_0,\boldsymbol{p}_1)>1$, we have $F_n(\boldsymbol{A},\boldsymbol{B})\leq 1$ and the equality is achieved if

- $F_i(\boldsymbol{A},\boldsymbol{B})=1, 1\leq i\leq n$.

Appendix C
Proof of the designed attack matrices achieving $\hat{\boldsymbol{q}}_0,\hat{\boldsymbol{q}}_1$

We will calculate the PMF $\hat{\boldsymbol{q}}_0,\hat{\boldsymbol{q}}_1$ achieved by the attack matrices $(\hat{\boldsymbol{A}},\hat{\boldsymbol{B}})$ designed according to (22) and (23) for columns $2,\cdots,m$, and according to (22) and (32) for columns $m+1,\cdots,n$. We will show that these satisfy the desired conditions specified in Section III-C.

For $j=1$,

$$
\begin{aligned}
q_{0,1} &= \sum_{i=1}^{1+\delta}p_{0,i}\hat{A}_{i,1}\\
&= \min\{p_{0,1},\hat{q}_{0,1}\}\\
&\quad +\sum_{i=2}^{1+\delta}\min\{p_{0,i},\max\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\}\}
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&\overset{(a)}{=} \hat{q}_{0,1},\\
q_{0,1} &= \sum_{i=1}^{1+\delta}p_{1,i}\hat{B}_{i,1}=\sum_{i=1}^{1+\delta}p_{1,i}.
\end{aligned}
$$

Here, (a) is true because 1) if $p_{0,1}\geq\hat{q}_{0,1}$, then $\min\{p_{0,1},\hat{q}_{0,1}\}=\hat{q}_{0,1}$ and $\max\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\}=0$, which indicates $q_{0,1}=\hat{q}_{0,1}$; 2) if $p_{0,1}<\hat{q}_{0,1}$, then

$$
\begin{aligned}
&\min\{p_{0,1},\hat{q}_{0,1}\}+\sum_{i=2}^{1+\delta}\min\left\{p_{0,i},\max\left\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\right\}\right\}\\
&= p_{0,1}+\sum_{i=2}^{1+\delta}\min\left\{p_{0,i},\max\left\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\right\}\right\}\\
&= \min\{p_{0,2},\max\{0,\hat{q}_{0,1}\}\}\\
&\quad +\sum_{i=3}^{1+\delta}\min\left\{p_{0,i},\max\left\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\right\}\right\}\\
&= \min\{p_{0,2},\hat{q}_{0,1}\}\\
&\quad +\sum_{i=3}^{1+\delta}\min\left\{p_{0,i},\max\left\{0,\hat{q}_{0,1}-\sum_{k=1}^{i-1}p_{0,k}\right\}\right\}.
\end{aligned}
\tag{45}
$$

Note that (44) and (45) are in the same form. Then by continuing this process, we will have $q_{0,1}=\hat{q}_{0,1}$.

For $2 \le j \le m$, under $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$, we have

$$\begin{aligned} q_{1,j} &= \sum_{i=1}^{j+\delta} p_{1,i}\hat{B}_{i,j} = p_{1,j+\delta}, \\ q_{0,j} &= \sum_{i=1}^{j+\delta} p_{0,i}\hat{A}_{i,j} = \sum_{i=1}^{j+\delta} \min\left\{ p_{0,i}(1-\sum_{k=1}^{j-1}\hat{A}_{i,k}), \right. \\ &\qquad \left. \max\left\{ \hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1-\sum_{t=1}^{j-1}\hat{A}_{k,t}\right), 0 \right\} \right\} \quad (46) \\ &\overset{(b)}{=} \hat{q}_{0,j}, \end{aligned}$$

in which (b) can be derived using the similar steps discussed in $j = 1$. Specifically, for $i = 1$, the index term in (46) is $\min\left\{p_{0,1}(1-\sum_{k=1}^{j-1}\hat{A}_{1,k}), \max\{\hat{q}_{0,j}, 0\}\right\}$. If $p_{0,1}(1-\sum_{k=1}^{j-1}\hat{A}_{1,k}) \ge \hat{q}_{0,j}$, we have $q_{0,j} = \hat{q}_{0,j}$ directly. On the other hand, if $p_{0,1}(1-\sum_{k=1}^{j-1}\hat{A}_{1,k}) < \hat{q}_{0,j}$, the index term for $i = 1$ is $p_{0,1}(1-\sum_{k=1}^{j-1}\hat{A}_{1,k})$, which will cancel out with a term in $\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1-\sum_{t=1}^{j-1}\hat{A}_{k,t}\right)$ and we will have $q_{0,j} = \hat{q}_{0,j}$ after a series of cancellations.

For $j \in R_1$, since the formula of $\hat{A}_{i,j}$ stays the same, $q_{0,j} = \hat{q}_{0,j}$ still holds. Under $\hat{B}_{i,j}$, suppose l is the smallest component with $\hat{B}_{l,j} > 0$, then we have

$$\begin{aligned} q_{1,j} &= \sum_{i=1}^{j+\delta} p_{1,i}\hat{B}_{i,j} \\ &= \sum_{i=1}^{j+\delta} \min\left\{ p_{1,i}(1-\sum_{k=1}^{j-1}\hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j} \right\} \\ &= \sum_{i=l}^{j+\delta} \min\left\{ p_{1,i}(1-\sum_{k=1}^{j-1}\hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j} \right\} \\ &= \min\left\{ p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}), \hat{q}_{1,j} - \sum_{k=1}^{l-1} p_{1,k}\hat{B}_{k,j} \right\} \\ &\quad + \sum_{i=l+1}^{j+\delta} \min\left\{ p_{1,i}(1-\sum_{k=1}^{j-1}\hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j} \right\} \\ &= \min\left\{ p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}), \hat{q}_{1,j} \right\} \quad (47) \\ &\quad + \sum_{i=l+1}^{j+\delta} \min\left\{ p_{1,i}(1-\sum_{k=1}^{j-1}\hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j} \right\} \\ &\overset{(c)}{=} \hat{q}_{1,j}, \end{aligned}$$

in which (c) can be proved by considering two different cases. First, if $p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}) \ge \hat{q}_{1,j}$, in (47), we have $\min\left\{p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}), \hat{q}_{1,j}\right\} = \hat{q}_{1,j} = p_{1,l}\hat{B}_{l,j}$ and $\hat{q}_{1,j} - p_{1,l}\hat{B}_{l,j} = 0$, which implies $\hat{B}_{l+1,j} = 0$ and thus $\hat{B}_{i,j} = 0, n \ge i \ge l+1$. Therefore, (c) holds. Second, if $p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}) \le \hat{q}_{1,j}$, in (47), we have $\min\left\{p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}), \hat{q}_{1,j}\right\} = p_{1,l}(1-\sum_{k=1}^{j-1}\hat{B}_{l,k}) = p_{1,l}\hat{B}_{l,j}$, which will cancel out with a term in $\hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j}$ and thus after a series of cancellation, we have $q_{0,j} = \hat{q}_{0,j}$.

Appendix D
Proof of Theorem 2

For $j = m - \delta$, we have

$$\begin{aligned} & F_{m-\delta}(\boldsymbol{A}) \\ &= \sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{n} q_{0,i} \\ &= 1 - \sum_{i=1}^{m-\delta}(p_{0,i} - p_{1,i}) + K_{0,m-\delta} - K_{1,m-\delta} \\ &\quad + I_{1,m-\delta} - I_{0,m-\delta} \\ &\overset{(a)}{\le} 1 - \sum_{i=1}^{m-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m-2\delta+1}^{m-\delta}(p_{0,i} - p_{1,i}) + 0 \\ &= 1 - \sum_{i=1}^{m-2\delta}(p_{0,i} - p_{1,i}) \\ &= E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1). \end{aligned}$$

For $\forall m - \delta + 1 \le j \le m + \delta, \forall \boldsymbol{A} \in \mathcal{A}$, we have

$$\begin{aligned} F_j(\boldsymbol{A}) &= F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j} \min\{(q_{1,i} - q_{0,i}), 0\} \\ &\overset{(b)}{\le} F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j}(q_{1,i} - q_{0,i}) \\ &= \sum_{i=1}^{j} q_{1,i} + \sum_{i=j+1}^{n} q_{0,i} \\ &= 1 + \sum_{i=1}^{j}(q_{1,i} - q_{0,i}) \\ &= 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + K_{0,j} - K_{1,j} \\ &\quad + I_{1,j} - I_{0,j} \\ &\overset{(c)}{\le} 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + \sum_{i=j-\delta+1}^{\min\{j,m\}}(p_{0,i} - p_{1,i}) \\ &\quad + \sum_{i=\max\{m+1,j+1\}}^{\min\{n,j+\delta\}}(p_{1,i} - p_{0,i}) \\ &= 1 - \sum_{i=1}^{j-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{\min\{n,j+\delta\}}(p_{1,i} - p_{0,i}) \\ &= E_j(\boldsymbol{p}_0, \boldsymbol{p}_1), \end{aligned}$$

in which the inequalities in $(a), (c)$ follow from the observation about I, K and the equality in (b) holds when $q_{1,i} \le q_{0,i}, m - \delta + 1 \le i \le j$.

Since the above inequality holds for $\forall \boldsymbol{A} \in \mathcal{A}$ and we have shown that $F_{m+\delta}(\boldsymbol{A}) \leq F_{m+\delta-1}(\boldsymbol{A}) \leq ... \leq F_{m-\delta}(\boldsymbol{A})$, then

$$\begin{aligned} F_m(\boldsymbol{A}) &\leq \min_{m-\delta \leq j \leq m} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}, \\ F_{m+\delta}(\boldsymbol{A}) &\leq \min_{m-\delta \leq j \leq m+\delta} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}. \end{aligned}$$

Furthermore, if $j^* > m-\delta$, $F_{m+\delta}(\boldsymbol{A}) \leq E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1)$ and the equality is achieved when

(i)

$$K_{0,m-\delta} - K_{1,m-\delta} = \sum_{i=m-2\delta+1}^{m-\delta} (p_{0,i} - p_{1,i});$$

(ii) $q_{1,i} \leq q_{0,i}, m-\delta+1 \leq i \leq j^*$;

(iii)

$$\begin{aligned} K_{0,j^*} - K_{1,j^*} &= \sum_{i=j^*-\delta+1}^{\min\{j^*,m\}} (p_{0,i} - p_{1,i}), \\ I_{1,j^*} - I_{0,j^*} &= \sum_{i=\max\{m+1,j^*+1\}}^{\min\{n,j^*+\delta\}} (p_{1,i} - p_{0,i}); \end{aligned}$$

(iv) $F_k(\boldsymbol{A}) = F_{j^*}(\boldsymbol{A}), j^* < k \leq m+\delta$.

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality is achieved when

$$F_i(\boldsymbol{A}) = E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1), m-\delta \leq i \leq m+\delta.$$

REFERENCES

[1] Y. Jin and L. Lai, "Adversarially robust hypothesis testing," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, (Pacific Grove, CA), Nov. 2019.

[2] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, pp. 1287–1289, Mar. 2019.

[3] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 2017, pp. 70–76, Jan. 2017.

[4] B. Biggio, G. fumera, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition : A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, Aug. 2015.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv pp.1312.6199*, Dec. 2013.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conference on Learning Representations*, (San Diego, CA), May. 2015.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symposium on Security and Privacy*, (San Jose, CA), pp. 39–57, May. 2017.

[8] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," *arXiv pp.1710.11342*, Oct. 2017.

[9] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, Apr. 2020.

[10] S. Sihag and A. Tajer, "Secure estimation under causative attacks," *IEEE Transactions on Information Theory*, 2020. To appear.

[11] M. Bande and V. V. Veeravalli, "Asversarial multi-user bandits for uncoordinated spectrum access," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, (Brighton, UK), pp. 1–5, May 2019.

[12] Y. Yang, "Robust estimation for dependent observations," *Manuscripta geodaetica*, vol. 19, no. 1, pp. 10–17, Oct. 1994.

[13] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, vol. 196. San Francisco, CA: John Wiley & Sons, 2011.

[14] P. J. Huber, *Robust statistics*. New York, NY: Springer, 2011.

[15] G. Gl and A. M. Zoubir, "Robust hypothesis testing with α -divergence," *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4737–4750, May. 2016.

[16] S. Sarta, S. Gezici, and S. Yksel, "Hypothesis testing under subjective priors and costs as a signaling game," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 5169–5183, Aug. 2019.

[17] N. Halay, K. Todros, and A. O. Hero, "Binary hypothesis testing via measure transformed quasi-likelihood ratio test," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6381–6396, Sep. 2017.

[18] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, Apr. 2017.

[19] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, Dec. 2008.

[20] L. Lai and E. Bayraktar, "On the adversarial robustness of robust estimators," *IEEE Transactions on Information Theory*, 2020. To appear.

[21] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi, "Measuring neural net robustness with constraints," in *Proc. Advances in neural information processing systems*, (Quebec, Canada), pp. 2613–2621, Dec. 2016.

[22] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symposium on Security and Privacy*, (San Jose, CA), pp. 582–597, May. 2016.

[23] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, Sep. 2018.

[24] A. Fawzi, O. Fawzi, and P. Frossard, "Fundamental limits on adversarial robustness," in *Proc. Int. Conference on Machine Learning, Workshop on Deep Learning*, (Lille, France), Jul. 2015.

[25] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proc. Advances in Neural Information Processing Systems*, (Quebec, Canada), pp. 1178–1187, Dec. 2018.

[26] Y. Chen, S. Kar, and J. M. F. Moura, "Resilient distributed estimation through adversary detection," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2455–2469, Mar. 2018.

[27] F. Li, L. Lai, and S. Cui, "On the adversarial robustness of subspace learning," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1470–1483, Mar. 2020.

[28] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, Jun. 2014.

[29] M. Barni and B. Tondi, "Multiple-observation hypothesis testing under adversarial conditions," in *2013 IEEE International Workshop on Information Forensics and Security (WIFS)*, (Guangzhou, China), pp. 91–96, Nov. 2013.

[30] M. Barni and B. Tondi, "Adversarial source identification game with corrupted training," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3894–3915, Feb. 2018.

[31] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, May. 2017.

[32] T. Banerjee, H. Firouzi, and A. O. Hero, "Quickest detection for changes in maximal KNN coherence of random matrices," *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4490–4503, Jul. 2018.

[33] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," in *Proc. Information Theory and Applications Workshop*, pp. 1–20, San Diego, CA, May. 2013.

[34] S. Zou, G. Fellouris, and V. V. Veeravalli, "Quickest change detection under transient dynamics: Theory and asymptotic analysis," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1397–1412, Oct. 2018.

[35] H. V. Poor and O. Hadjiliadis, *Quickest detection*. Cambridge University Press, 2008.

[36] Y. Wang and Y. Mei, "Large-scale multi-stream quickest change detection via shrinkage post-change estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, Dec. 2015.

[37] O. Hadjiliadis, H. Zhang, and H. V. Poor, "One shot schemes for decentralized quickest change detection," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3346–3359, Jun. 2009.

[38] R. Jana and S. Dey, "Change detection in teletraffic models," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 846–853, Mar. 2000.
[39] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a Markov process across a sensor array," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1961–1981, Mar. 2010.
[40] G. Fellouris, E. Bayraktar, and L. Lai, "Efficient Byzantine sequential change detection," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3346–3360, Sept. 2017.
[41] T. Banerjee and V. V. Veeravalli, "Data-efficient quickest change detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3727–3735, May. 2015.
[42] A. G. Tartakovsky and V. V. Veeravalli, "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Analysis*, vol. 27, no. 4, pp. 441–475, Oct. 2008.
[43] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, Nov. 1986.
[44] S. Li, Y. Yılmaz, and X. Wang, "Quickest detection of false data injection attack in wide-area smart grids," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2725–2735, Dec. 2014.
[45] D. Li, L. Lai, and S. Cui, "Quickest change detection and identification across a sensor array," in *Proc. IEEE Global Conference on Signal and Information Processing*, pp. 145–148, Austin, TX, Dec. 2013.
[46] J.-P. Aubin and I. Ekeland, *Applied nonlinear analysis*. North Chelmsford, MA: Courier Corporation, 2006.
[47] R. Cogranne, T. Denemark, and J. Fridrich, "Theoretical model of the fld ensemble classifier based on hypothesis testing theory," in *Proc. IEEE International Workshop on Information Forensics and Security*, (Atlanta, GA), pp. 167–172, IEEE, Dec. 2014.
[48] D. Ciuonzo, P. S. Rossi, and P. Willett, "Generalized rao test for decentralized detection of an uncooperative target," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 678–682, May 2017.
[49] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, "Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 205–215, Nov. 2012.
[50] D. Ciuonzo, A. De Maio, and P. S. Rossi, "A systematic framework for composite hypothesis testing of independent bernoulli trials," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1249–1253, Jan. 2015.