Detection Games with a Fully Active Attacker

Benedetta Tondi [1], Mauro Barni [1], Neri Merhav [2]

[1] *Department of Information Engineering and Mathematics, University of Siena, Siena, ITALY*

[2] *Department of Electrical Engineering Technion - Israel Institute of Technology Technion City, Haifa, ISRAEL*

benedettatondi@gmail.com, barni@dii.unisi.it, merhav@ee.technion.ac.il

***Abstract*—We analyze a binary hypothesis testing problem in which a defender has to decide whether or not a test sequence has been drawn from a given source P_0 whereas, an attacker strives to impede the correct detection. In contrast to previous works, the adversarial setup addressed in this paper considers a fully active attacker, i.e. the attacker is active under both hypotheses. Specifically, the goal of the attacker is to distort the given sequence, no matter whether it has emerged from P_0 or not, to confuse the defender and induce a wrong decision. We formulate the defender-attacker interaction as a game and study two versions of the game, corresponding to two different setups: a Neyman-Pearson setup and a Bayesian one. By focusing on asymptotic versions of the games, we show that there exists an attacking strategy that is both dominant (i.e., optimal no matter what the defence strategy is) and universal (i.e., independent of the underlying sources) and we derive equilibrium strategies for both parties.**

I. Introduction

There are many fields in signal processing and communications where the detection problem should naturally be framed within an adversarial setting [1]: multimedia forensics (MF) [2], [3], spam filtering [4], biometric-based verification [5], one-bit watermarking [6], digital/analogue transmission under jammer attacks, just to mention a few.

In recent literature, game theory and information theory have been combined to address the problem of adversarial detection, see e.g. [7], [8]. Specifically, in [8] the general problem of binary hypothesis testing under adversarial conditions has been addressed and formulated as a game between two players, the defender and the attacker, which have conflicting goals. Given two discrete memoryless sources, P_0 and P_1, the goal of the defender is to decide whether a given test sequence has been generated by P_0 (null hypothesis, $\mathcal{H}_0$) or P_1 (alternative hypothesis, $\mathcal{H}_1$). By adopting the Neyman-Pearson (NP) approach, the set of strategies the defender can choose from is the set of decision regions for $\mathcal{H}_0$ ensuring that the false positive error probability is lower than a given threshold. On the other hand, the ultimate goal of the attacker in [8] is causing a false negative decision, so the attacker acts under $\mathcal{H}_1$ only. In other words, the attacker modifies a sequence generated by P_1, in attempt to move it into the acceptance region of $\mathcal{H}_0$. In doing so, the attacker has to respect a distortion constraint, limiting the amount of modifications that can be introduced into the sequence. In [8], such a struggle between the defender and the attacker is modeled as a competitive zero-sum game and the asymptotic equilibrium, that is, the equilibrium when the length of the observed sequence tends to infinity, is derived under the assumption that the defender bases its decision on the analysis of first order statistics only. Such an assumption, where the detector has access to a limited sets of empirical statistics of the sequence, is referred to as a *limited resources* assumption (see [9] for an introduction on this terminology). In the scenario considered in this paper, the limitation of the detection resources to the first order is motivated by the assumed memorylessness of the sources for the attack–free scenario (the empirical distribution is a sufficient statistic for the source). Another motivation is that it serves as a basis for the more general case, where the detector implements the calculation of the test statistics by means of a finite–state machine that is fed by the data in a sequential manner.[1] The analysis conducted in [8] extends the one of [10] to the adversarial scenario.

Some variants of this attack-detection game have also been studied: in [11], the setting was extended to the case where the sources are known to neither the defender nor the attacker, yet training data from both sources is available to both parities; within this framework, the case where part of the training data available to the defender is corrupted by the attacker has also been studied (see [12]).

There are many situations in which it is reasonable to assume that the attacker is active under both hypotheses with the goal of causing both false positive and false negative detection errors. For instance, in applications of fingerprint detection, an adversary might be interested to remove the fingerprint from a given image so that the generating camera would not be identified and, at the same time, to modify the specific fingerprint to blame an innocent victim, [13], [14].

With the above ideas in mind, in this paper, we extend the game–theoretic formulation of the defender-attacker interaction to the case where the attacker acts under both hypotheses. We refer to this scenario as a detection game with a *fully-active attacker*. We address both the case where the underlying hypothesis is known to the attacker and the case where it is not. We define and solve two versions of the detection game with fully active attackers, corresponding to two different detection setups: in the former, we assume that the defender bases the decision on an adversary-aware NP test; in the latter, a Bayesian approach is adopted, where the role of the two error probabilities is symmetrized, and the decision is based on the minimization of a Bayesian risk function.

As an additional contribution, we extend the analysis developed in [8] to consider randomized detection strategies. We also show that for both the classical version of the game and the game with fully active adversary, there exists an attacking strategy that is both dominant (i.e., optimal no matter what the defence strategy is) and universal (i.e., independent of the underlying sources). This marks a significant difference with respect to previous works, where the existence of a dominant strategy was proven only with reference to the defender.

The outline of the paper is the following: we introduce the notation and the main concepts in Section II; then, in Section III, we revisit the analysis developed in [8] to consider randomized detection strategies and show that there exists an attack strategy which is both universal and asymptotically dominant.

In Section IV, these findings are extended to a setting where the attacker is active under both hypotheses: the two versions of the game are studied in Section IV-A and IV-B.

II. Notation and definitions

Given a random variable X, we denote by $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, $x_i \in \mathcal{A}$, $i = 1, 2, \dots, n$, a sequence of n (independent) copies of X.

[1] The first–order empirical statistics considered here is then a special case, where this finite–state machine has one state only.

978-1-4673-6802-5/15/$31.00 ©2015 IEEE

Throughout this paper, we make an extensive use of the concept of typicality and the method of types (see, [15], [16], [17]). The type of a sequence $\boldsymbol{x}$ is defined as the empirical probability distribution $\hat{P}_{\boldsymbol{x}}$, that is, the vector $\{P_{\boldsymbol{x}}(x),\ x \in \mathcal{A}\}$ of the relative frequencies of the various alphabet symbols in $\boldsymbol{x}$. A type class $\mathcal{T}(\boldsymbol{x})$ is defined as the set of all the sequences having the same type of $\boldsymbol{x}$. Similarly, given a pair of sequences $(\boldsymbol{x}, \boldsymbol{y})$, the conditional type class $\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$ is the set of the sequences having empirical conditional probability distribution (or conditional type) $\hat{P}_{\boldsymbol{y}|\boldsymbol{x}}$.

We denote by $A(\boldsymbol{y}|\boldsymbol{x})$ the conditional probability distribution of a channel with input $\boldsymbol{x}$ and output $\boldsymbol{y}$. Given a permutation-invariant distortion function $d : \mathcal{A}^n \times \mathcal{A}^n \to \mathbb{R}^{+}$[2] and a maximum per-symbol distortion Δ, we define the class of admissible channels $\mathcal{C}$ as the class of channels A that assigns zero probability to output sequences $\boldsymbol{y}$ such that the distance from $\boldsymbol{x}$ is larger than the prescribed maximum value; i.e., $A(\boldsymbol{y}|\boldsymbol{x}) = 0\ \forall \boldsymbol{y}$ s.t. $d(\boldsymbol{x}, \boldsymbol{y}) > n\Delta$.

For sake of clarity, we introduce some basic definitions of game theory. A 2-player game is defined as a quadruple $(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, where $\mathcal{S}_1 = \{s_{1,1} \dots s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1} \dots s_{2,n_2}\}$ are the set of strategies the first and the second player can choose from, and $u_l(s_{1,i}, s_{2,j}), l = 1, 2$, is the payoff of the game for player l, when the first player chooses the strategy $s_{1,i}$ and the second chooses $s_{2,j}$. A pair of strategies $(s_{1,i}, s_{2,j})$ is called a profile. When $u_1(s_{s1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$, the win of a player is equal to the loss of the other and the game is said to be a zero-sum game. The sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players. In addition, we consider strategic games, i.e., games in which the players choose their strategies before starting the game without knowing the strategy chosen by the opponent player.

A common goal in game theory is to determine the existence of equilibrium points, i.e. profiles that in *some sense* represent a *satisfactory* choice for both players [18]. The most famous notion of equilibrium is due to Nash. A profile is said to be a Nash equilibrium if no player can improve its payoff by changing its strategy unilaterally.

Despite its popularity, the practical meaning of Nash equilibrium is often unclear, since there is no guarantee that the players will end up playing at the equilibrium. A particular kind of games for which stronger forms of equilibrium exist are the so called *dominance solvable* games [18]. The concept of dominance-solvability is directly related to the notion of dominant and dominated strategies. In particular, a strategy is said to be strictly dominant for one player if it is the best strategy for the player, i.e., the strategy which corresponds to the largest payoff, no matter how the other player decides to play. When one such strategy exists for one of the players, he will surely adopt it. In a similar way, we say that a strategy $s_{l,i}$ is strictly dominated by strategy $s_{l,j}$, if the payoff achieved by player l choosing $s_{l,i}$ is always lower than that obtained by playing $s_{l,j}$ regardless of the choice made by the other player. The recursive elimination of dominated strategies is one common technique for solving games. In the first step, all the dominated strategies are removed from the set of available strategies, since no rational player would ever play them. In this way, a new, smaller game is obtained. At this point, some strategies, that were not dominated before, may be dominated in the remaining game, and hence are eliminated. The process goes on until no dominated strategy exists for any player. A *rationalizable equilibrium* is any profile which survives the iterated elimination of dominated strategies [19], [20]. If at the end of the process only one profile is left, the remaining profile is said to be the *only rationalizable equilibrium* of the game, which is also the only Nash equilibrium point. Dominance solvable games are easy to analyze since, under the assumption of rational players, we can anticipate that the players will choose the strategies corresponding to the unique rationalizable equilibrium. An interesting notion of equilibrium is that of dominant equilibrium. A dominant equilibrium is a profile which corresponds to dominant strategies for both players and is the strongest kind of equilibrium that a strategic game may have.

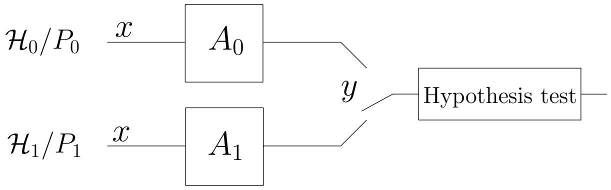

Fig. 1. Schematic representation of the adversarial setup considered in this paper. In the case of partially active attacker, channel A_0 corresponds to the identity channel.

Regarding the notation, for two positive sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means that $\lim_{n\to\infty} 1/n \log(a_n/b_n) = 0$, and $a_n \dot{\leq} b_n$ designates that $\limsup_{n\to\infty} 1/n \log(a_n/b_n) \leq 0$. Throughout the paper, for a given quantity s, we adopt the following notation: $[s]_+ \triangleq \max\{s, 0\}$.

Given two random variables X and Y, we use notation $\hat{H}_{\boldsymbol{x}}(X)$ ($\hat{H}_{\boldsymbol{y}}(Y)$) for the empirical entropy of a sequence $\boldsymbol{x}$ ($\boldsymbol{y}$) and notation $\hat{H}_{\boldsymbol{xy}}(X, Y)$ ($\hat{H}_{\boldsymbol{xy}}(X|Y)$) for the joint (conditional) entropy, see [15]. Finally, we denote with $\mathcal{D}(\cdot\|\cdot)$ the Kullback–Leibler (K-L) divergence, see again [15].

III. Detection Game with Partially Active Attacker

In this section, we extend the analysis of the binary hypothesis testing game developed in [8], where the attacker is active under $\mathcal{H}_1$ only. The analysis not only introduces new results with respect to [8], but it also represents the basis for studying the version of the game with a fully active attacker.

Given two discrete memoryless sources, P_0 and P_1, defined over a same source alphabet $\mathcal{A}$, we denote by $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{A}^n$ a sequence emitted by one of these sources. The sequence $\boldsymbol{x}$ is available to the attacker. Let $\boldsymbol{y} = (y_1, y_2, ..., y_n) \in \mathcal{A}^n$ denote the sequence observed by the defender: in the scenario considered in this section, we have $\boldsymbol{y} = \boldsymbol{x}$ under $\mathcal{H}_0$ (no attack occurs), whereas under $\mathcal{H}_1$, $\boldsymbol{y}$ is obtained as the output of an attack channel defined by a conditional probability distribution $A_1(\boldsymbol{y}|\boldsymbol{x})$. Figure 1 illustrated the general framework.

Let us denote by $Q_i(\cdot)$ the probability distribution of $\boldsymbol{y}$ under hypothesis $\mathcal{H}_i$; then, we have $Q_0(\boldsymbol{y}) = P_0(\boldsymbol{y})$ and $Q_1(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P_1(\boldsymbol{x}) A_1(\boldsymbol{y}|\boldsymbol{x})$.

With regard to the defender, we assume a possibly randomized decision strategy, where $D(\mathcal{H}_i|\boldsymbol{y})$ designates the probability of deciding in favor of $\mathcal{H}_i$, $i = 0, 1$, given the observed sequence $\boldsymbol{y}$. Accordingly, the probability of a false positive (FP) decision error is given by

$$P_{\text{FP}}(D) = \sum_{\boldsymbol{y}} P_0(\boldsymbol{y}) D(\mathcal{H}_1|\boldsymbol{y}), \tag{1}$$

[2] For a permutation-invariant distance, the distance does not change by applying the same permutation to $\boldsymbol{x}$ and $\boldsymbol{y}$.

whereas the false negative (FN) error probability assumes the form:

$$P_{\text{FN}}(D, A_1) = \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{x}) A_1(\boldsymbol{y}|\boldsymbol{x}) D(\mathcal{H}_0|\boldsymbol{y}). \tag{2}$$

As in [8], due to the limited resources assumption, the defender makes a decision based on first order empirical statistics of $\boldsymbol{y}$, which implies that $D(\cdot|\boldsymbol{y})$ depends on $\boldsymbol{y}$ only via its type class $\mathcal{T}(\boldsymbol{y})$. The set $\mathcal{T}(\boldsymbol{y})$ can be interpreted as the set of all the sequences which can be obtained by permuting $\boldsymbol{y}$. In other words, $D(\cdot|\boldsymbol{y})$ is invariant to permutations of $\boldsymbol{y}$. Concerning the attack, in order to limit the amount of distortion, we will assume a distortion constraint: for some chosen permutation-invariant distortion function $d(\cdot,\cdot)$ and per-letter distortion Δ, the attack channel A_1 belongs to $\mathcal{C}$.

We define the generalized detection game with a *partially active attacker* (i.e., an attacker active under $\mathcal{H}_1$ only) as follows.

Definition 1. *The* DG-PA*($\mathcal{S}_D$, $\mathcal{S}_A$,u) game is a zero-sum, strategic game played by a defender and an attacker, defined as follows:*

- *The set of strategies the defender is the class $\mathcal{S}_D$ of randomized decision rules that satisfy the following properties:*
  - (i) $D(\mathcal{H}_0|\boldsymbol{y}) = D(\mathcal{H}_0|\boldsymbol{y}')$ *whenever $\boldsymbol{y}'$ is a permutation of $\boldsymbol{y}$.*
  - (ii) $P_{FP}(D) \le e^{-\lambda n}$ *for a given prescribed $\lambda > 0$.*
- *The set of strategies for the attacker is the class $\mathcal{S}_A$ of attack channels A_1 with the property that $d(\boldsymbol{x},\boldsymbol{y}) > n\Delta$ implies $A_1(\boldsymbol{y}|\boldsymbol{x}) = 0$; that is $\mathcal{S}_A \equiv \mathcal{C}$.*
- *The payoff function: $u(D, A) = P_{FN}(D, A)$, where the attacker's perspective is adopted (the attacker is in the quest for maximizing $u(D, A)$ and the defender wishes to minimize $u(D, A)$).*

We point out that the *DG-PA* game is an extension of the source identification (*SI*) game defined in [8], since in the *DG-PA* both players of the game are allowed to employ randomized strategies, while in the *SI*, only deterministic strategies were considered. Specifically, in [8], the defence strategies are confined to deterministic rules (decision regions) and the attack is confined to the application of deterministic functions to the to-be-attacked sequence.

As in [8], we focus on the asymptotic behavior of the *DG-PA* game, that is, the behavior when n tends to infinity. We say that a strategy is asymptotically optimum (or dominant) strategy if the strategy is optimum (dominant) with respect to the exponent of the payoff.

We start by asserting the following lemma:

Lemma 1. *The defence strategy*

$$D^*(\mathcal{H}_1|\boldsymbol{y}) \triangleq \exp\{-n[\lambda - \mathcal{D}(\hat{P}_{\boldsymbol{y}}\|P_0)]_+\}, \tag{3}$$

is an asymptotically dominant strategy for the defender.

Proof: The asymptotic optimality of $D^*(\cdot|\boldsymbol{y})$ follows directly from the false positive (FP) constraint:

$$\begin{aligned} e^{-\lambda n} &\ge \sum_{\boldsymbol{y}'} P_0(\boldsymbol{y}') D(\mathcal{H}_1|\boldsymbol{y}') \ge |\mathcal{T}(\boldsymbol{y})| \cdot P_0(\boldsymbol{y}) D(\mathcal{H}_1|\boldsymbol{y}) \\ &\dot{\ge} e^{-nD(\hat{P}_{\boldsymbol{y}}\|P_0)} D(\mathcal{H}_1|\boldsymbol{y}), \quad \forall \boldsymbol{y}, \end{aligned} \tag{4}$$

where we have exploited the permutation–invariance of $D(\mathcal{H}_1|\boldsymbol{y})$ and the memoryless of nature P_0, which implies $P_0(\boldsymbol{y}) = P_0(\boldsymbol{y}')$ whenever $\boldsymbol{y}'$ is a permuted version of $\boldsymbol{y}$. It follows that

$$D(\mathcal{H}_1|\boldsymbol{y}) \dot{\le} \min\{1, e^{-n[\lambda - D(\hat{P}_{\boldsymbol{y}}\|P_0)]}\} = D^*(\mathcal{H}_1|\boldsymbol{y}).$$

By using the method of types [16], it is easy to see that D^* satisfies the false positive constraint within a polynomial factor. Since $D^*(\mathcal{H}_1|\boldsymbol{y}) \dot{\ge} D(\mathcal{H}_1|\boldsymbol{y})$, obviously, $D^*(\mathcal{H}_0|\boldsymbol{y}) \dot{\le} D(\mathcal{H}_0|\boldsymbol{y})$, and so, $P_{\text{FN}}(D^*, A_1) \dot{\le} P_{\text{FN}}(D, A_1)$ for every attack channel A_1. ■

According to Lemma 1, the best defending strategy is dominant, and then it is the optimum strategy regardless of the attacking channel. Furthermore, we argue that the optimum decision function asymptotically tends to a deterministic function which essentially corresponds to the Hoeffding test [21]. Note that Lemma 1 is in line with the results obtained in [8] where the analysis is confined to deterministic decision rules. As the optimum strategy D^* depends only on P_0, but not on P_1, it is said to be semi–universal.

We now move on to the analysis of the attack. One of the main results of the paper is stated by the following theorem.

Theorem 1. *Let $c_n(\boldsymbol{x})$ denote the reciprocal of the total number of conditional type classes $\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$ that satisfy the constraint $d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta$, namely, admissible conditional type classes*[3]*. The attack channel*

$$A^*(\boldsymbol{y}|\boldsymbol{x}) = \begin{cases} \frac{c_n(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} & d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta \\ 0 & \textit{elsewhere} \end{cases}, \tag{5}$$

is an asymptotically dominant strategy for the attacker.

Proof: Consider an arbitrary channel $A_1 \in \mathcal{S}_A$. Let $\Pi : \mathcal{A}^n \to \mathcal{A}^n$ denote a permutation operator that permutes any member of $\mathcal{A}^n$ according to a given permutation matrix and let

$$A_\Pi(\boldsymbol{y}|\boldsymbol{x}) \triangleq A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}), \tag{6}$$

Since the distortion is invariant to permutations, channel $A_\Pi(\boldsymbol{y}|\boldsymbol{x})$ introduces the same distortion as A_1 and hence satisfies the distortion constraint. Thanks to the memorylessness of P_1 and the assumed permutation–invariance of $D(\mathcal{H}_0|\boldsymbol{y})$, we have

$$\begin{aligned} P_{\text{FN}}(D, A_\Pi) &= \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{y}) A_\Pi(\boldsymbol{y}|\boldsymbol{x}) D(\mathcal{H}_0|\boldsymbol{y}) \\ &= \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{y}) A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}) D(\mathcal{H}_0|\boldsymbol{y}) \\ &= \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\Pi\boldsymbol{y}) A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}) D(\mathcal{H}_0|\Pi\boldsymbol{y}) \\ &= \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{y}) A_1(\boldsymbol{y}|\boldsymbol{x}) D(\mathcal{H}_0|\boldsymbol{y}) \\ &= P_{\text{FN}}(D, A_1), \end{aligned} \tag{7}$$

and so, $P_{\text{FN}}(D, A_1) = P_{\text{FN}}(D, \bar{A})$ where we have defined

$$\bar{A}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{n!}\sum_{\Pi} A_\Pi(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{n!}\sum_{\Pi} A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}), \tag{8}$$

which also introduces the same distortion as A_1. This channel assigns the same probability to all the sequences in the same conditional type class $\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$. To prove it, we observe that any sequence $\boldsymbol{y}' \in \mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$ can be seen as being obtained from $\boldsymbol{y}$ through the application of a permutation Π' which leaves $\boldsymbol{x}$ unaltered. Then, we have:

$$\begin{aligned} \bar{A}(\boldsymbol{y}'|\boldsymbol{x}) = \bar{A}(\Pi'\boldsymbol{y}|\Pi'\boldsymbol{x}) &= \frac{1}{n!}\sum_{\Pi} A_1(\Pi(\Pi'\boldsymbol{y})|\Pi(\Pi'\boldsymbol{x})) \\ &= \frac{1}{n!}\sum_{\Pi} A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}) = \bar{A}(\boldsymbol{y}|\boldsymbol{x}). \end{aligned} \tag{9}$$

[3] From the method of the types it is known that $1 \ge c_n(\boldsymbol{x}) \ge (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)}$ for any $\boldsymbol{x}$ [15].

Therefore, we argue that

$$\begin{aligned}\bar{A}(\boldsymbol{y}|\boldsymbol{x}) \overset{\cdot}{\leq} &\begin{cases} \frac{1}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} & d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta \\ 0 & \text{elsewhere} \end{cases} \\ =&\frac{A^*(\boldsymbol{y}|\boldsymbol{x})}{c_n(\boldsymbol{x})} \\ \leq&(n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} A^*(\boldsymbol{y}|\boldsymbol{x}), \end{aligned} \tag{10}$$

which implies that, for every permutation–invariant defence strategy D, $P_{\text{FN}}(D, A_1) \leq (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D, A^*)$, or equivalently

$$P_{\text{FN}}(D, A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D, A_1). \tag{11}$$

We conclude that A^* minimizes the error exponent of $P_{\text{FN}}(D, A_1)$ for every given $A_1 \in \mathcal{S}_A$ and $D \in \mathcal{S}_D$. ■

Theorem 1 states that strategy A^* is *dominant* for the attacker, and so the optimum attacking channel does not depend on the decision strategy $D(\cdot|\boldsymbol{y})$. Then, given a sequence $\boldsymbol{x}$, in order to generate an attacked sequence $\boldsymbol{y}$ which undermines the detection (with the prescribed maximum allowed distortion), the best way is to choose an admissible conditional type class according to the uniform distribution (at random) and then select at random a sequence $\boldsymbol{y}$ within this conditional type class. As a further result, Theorem 1 states that the optimum attacking strategy is *universal*, i.e., it depends neither on P_0 nor on P_1. The existence of dominant strategies for both players directly leads to the following result.

Theorem 2. *The profile (D^*, A^*) is an asymptotically dominant equilibrium for the* DG-PA *game.*

IV. Detection Games with fully active attacker

We now consider the detection game when the attacker is active under both hypotheses. In principle, we must distinguish between two cases: in the first one, the attacker is aware of the underlying hypothesis (hypothesis-aware attacker), whereas in the second case, it is not (hypothesis-unaware attacker).

In the hypothesis-aware case, the attack strategy is defined by two attack channels: A_0 (carried out when $\mathcal{H}_0$ holds) and A_1 (carried out under $\mathcal{H}_1$). This attack induces the following distributions on the observed sequence $\boldsymbol{y}$: $Q_0(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) A_0(\boldsymbol{y}|\boldsymbol{x})$ and $Q_1(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P_1(\boldsymbol{x}) A_1(\boldsymbol{y}|\boldsymbol{x})$. The FP probability becomes:

$$P_{\text{FP}}(D, A_0) = \sum_{\boldsymbol{x},\boldsymbol{y}} P_0(\boldsymbol{x}) A_0(\boldsymbol{y}|\boldsymbol{x}) D(\mathcal{H}_1|\boldsymbol{y}), \tag{12}$$

while for the FN probability, equation (2) continues to hold.

The schematic representation of the fully-active case is given in Figure 1. It is easy to argue that the partially active case is a degenerate case of the fully active one (where A_0 is the identity channel).

By the same reasoning as in the proof of Theorem 1, we now show that the (asymptotically) optimum attacking strategy is independent on the underlying hypothesis. As a consequence, the best attack under the fully active regime is to apply the same A^* regardless of which hypothesis holds. Due to this property, it becomes immaterial whether the attacker is aware or unaware of the true hypothesis.

To be more specific, let u denote a payoff function of the form

$$u = \gamma P_{\text{FN}}(D, A_1) + \beta P_{\text{FP}}(D, A_0), \tag{13}$$

where β and γ are given positive constants, possibly dependent on n. The following Theorem asserts the asymptotic dominance of the channel A^* w.r.t. the payoff function u for every choice of β and γ.

Theorem 3. *Let A^* denote the attack channel in* (5)*. Among all pairs of channels in $\mathcal{C}$, the pair (A_0^*, A_1^*) with $A_0^* = A_1^* = A^*$ minimizes the asymptotic exponent of u for any $\gamma, \beta \geq 0$ and any permutation–invariant decision rule $D(\mathcal{H}_0|\cdot)$.*

Proof: Due to the memorylessness of P_1 and the permutation-invariance of $D(\mathcal{H}_0|\cdot)$, and by reasoning as we did in Theorem 1, we know that, for every $A_1 \in \mathcal{C}$, we have:

$$P_{\text{FN}}(D, A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D, A_1), \tag{14}$$

and then A^* minimizes the error exponent of $P_{\text{FN}}(D, A_1)$.

A similar argument can be applied to the FP probability; that is, from the memorylessness of P_0 and the permutation–invariance of $D(\mathcal{H}_1|\cdot)$, we have:

$$P_{\text{FP}}(D, A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FP}}(D, A_0), \tag{15}$$

for every $A_0 \in \mathcal{C}$. Accordingly, A^* minimizes the asymptotic exponent of $P_{\text{FP}}(D, A_0)$ as well. We then have:

$$\begin{aligned}&\gamma P_{\text{FN}}(D, A_1) + \beta P_{\text{FP}}(D, A_0) \\ &\quad \leq (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} (\gamma P_{\text{FN}}(D, A^*) + \beta P_{\text{FP}}(D, A^*)) \\ &\quad \doteq \gamma P_{\text{FN}}(D, A^*) + \beta P_{\text{FP}}(D, A^*), \end{aligned} \tag{16}$$

for every $A_0 \in \mathcal{C}$ and $A_1 \in \mathcal{C}$. Notice that, since the asymptotic equality is defined in logarithmical scale, relation (16) holds whichever is the (eventual) dependence of γ and β on n. Hence, $A_0 = A_1 = A^*$ minimizes the asymptotic exponent of u for any permutation–invariant decision rule $D(\mathcal{H}_0|\cdot)$ and for any $\gamma, \beta > 0$. ■

We point out that, whenever γ (res. β) is equal to 0, all the attacking strategies A_1 (res. A_0) are equivalent, in the sense that all the pairs (A^*, A_1) for every A_1 (res. (A_0, A^*), for every A_0) lead to the same asymptotic payoff. From Theorem 3 we deduce that, whenever an adversary aims at maximizing a payoff function of the form (13), and as long as the defence strategy is confined to the analysis of the first order statistics, the asymptotically optimal attack is A^* under either hypothesis.

We now turn the attention to the defender. The main difficulty relies in the fact that in the presence of a fully active attacker P_{FP} also depends on the attack, thus forcing us to reconsider the constraint on P_{FP}.

In the sequel, we consider two different approaches which lead to different formulations of the detection game with fully active attacker (DG-FA).

A. The DG-FA game: the Neyman Pearson approach

We consider the detection based on the NP test. To define a DG-FA game in this setup, we assume that the defender adopts a conservative approach by imposing an FP constraint pertaining to the worst–case attack under $\mathcal{H}_0$ Specifically, we define the game as follows.

Definition 2. *The* DG-FA1*($\mathcal{S}_D$, $\mathcal{S}_A$,u) game is a zero-sum, strategic game defined by*

- *The set of strategies for the defender is the the class $\mathcal{S}_D$ of randomized decision rules that satisfy*
  - (i) $D(\mathcal{H}_0|\boldsymbol{y}) = D(\mathcal{H}_0|\boldsymbol{y}')$ *whenever $\boldsymbol{y}'$ is a permutation of $\boldsymbol{y}$.*
  - (ii) $\max_{A_0 \in \mathcal{C}} P_{FP}(D, A_0) \leq e^{-n\lambda}$ *for a prescribed $\lambda > 0$.*
- *The set of strategies for the attacker is the class $\mathcal{S}_A$ of the pairs of attack channels (A_0, A_1) such that $A_0, A_1 \in \mathcal{C}$.*
- *The payoff function: $u(D, A) = P_{FN}(D, A_1)$.*

Having already determined the best attacking strategy, we focus on the best defender's strategy. We start with the following lemma:

Lemma 2. *The strategy*

$$D^*(\mathcal{H}_1|\boldsymbol{y}) \triangleq \exp\left\{-n\left[\lambda - \min_{\boldsymbol{x}:d(\boldsymbol{x},\boldsymbol{y})\le n\Delta} \mathcal{D}(\hat{P}_{\boldsymbol{x}}||P_0) - |\mathcal{A}|^2\frac{\log(n+1)}{n}\right]_+\right\} \quad (17)$$

is asymptotically dominant for the defender.

Proof: Due to space limitations, we only explain the intuition behind and provide the sketch of the proof. We know from Lemma 1 that for the case of no attack under $\mathcal{H}_0$, the asymptotically optimal detection rule is based on $\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)$. In the setup of the *DG-FA1* game, where the attacker is active also under $\mathcal{H}_0$, the defender is subject to a constraint on the maximum FP probability over $\mathcal{S}_A$. We know from Theorem 3 that, in the asymptotic exponent sense, this maximum value is achieved when $A_0 = A^*$. We observe that A^* assigns a probability which is the reciprocal of a polynomial term at each conditional type class that satisfies the distortion constraint (admissible conditional type class). Then, in order to be compliant with the constraint, for a given sequence $\boldsymbol{y}$, the defender has to consider the minimum of $\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)$ over all the type classes $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ which satisfy the distortion constraint, or equivalently, all the sequences $\boldsymbol{x}$ such that $d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta$.

The full proof goes along the following lines: first, it is shown that $P_{\text{FN}}(D^*, A_1) \stackrel{\cdot}{\le} P_{\text{FN}}(D, A_1)$ for every $D \in \mathcal{S}_D$; then, by proving that $\max_A P_{\text{FP}}(D^*, A)$ fulfills the FP constraint, it follows that $D^*(\cdot|\boldsymbol{y})$ is the optimum defence strategy (asymptotically). ■

Lemma 2 asserts the dominance and the semi-universality of the defence strategy, which depends only on the source P_0.

With regard to the attack, since the payoff of the game is a special case of (13) with $\gamma = 1$ and $\beta = 0$, the optimum pair of attacking channels is given by Theorem 3 and is (A^*, A^*). We point out that, as a consequence of Theorem 3, the optimum attacking strategy is *fully universal*: the attacker does not need to know either sources (P_0 and P_1) or the underlying hypothesis.

We observe that, since the defender adopted a conservative approach to ensure the constraint on FP, the pairs (A_0, A^*), for every $A_0 \in \mathcal{S}_A$, are all equivalent, that is, they lead to the same payoff, and then the attacker does not even need to perform the attack under the null hypothesis. Therefore, if the attacker is aware of the true hypothesis, then she could play any channel under $\mathcal{H}_0$. In the NP decision setup, the sole fact that the attacker is allowed to attack under $\mathcal{H}_0$ forces the defender to take countermeasures that render the attack under $\mathcal{H}_0$ useless.

Due to the existence of dominant strategies for both players, we can immediately state the following theorem:

Theorem 4. *Profile* $(D^*, (A^*, A^*))$ *is an asymptotically dominant equilibrium for the* DG-FA1 *game.*

B. The DG-FA: the Bayesian approach

In this section, we define another version of the DG-FA game. Specifically, we assume that the defender follows a less conservative Bayesian approach and tries to minimize a particular Bayes risk. The resulting game is defined as follows:

Definition 3. *The* DG-FA2*($\mathcal{S}_D, \mathcal{S}_A$,u) game is a zero-sum, strategic game defined by*

- *The set of strategies for the defender is the class $\mathcal{S}_D$ of randomized decision rules that satisfy $D(\mathcal{H}_0|\boldsymbol{y}) = D(\mathcal{H}_0|\boldsymbol{y}')$ whenever $\boldsymbol{y}'$ is a permutation of $\boldsymbol{y}$.*
- *The set of strategies for the attacker is the same set as before;*
- *The payoff function:*

$$u = P_{\text{\tiny FN}}(D, A_1) + e^{an}P_{\text{\tiny FP}}(D, A_0), \quad (18)$$

for some positive a.

We observe that, in the definition of the payoff, the parameter a controls the tradeoff between the two error exponents; we anticipate that the optimum strategy D will be the one making the difference between the two error exponents exactly equal to a. Notice also that, with definition (18), we are implicitly considering for the defender only the strategies $D(\cdot|\boldsymbol{y})$ such that $P_{\text{FP}}(D, A_0) \stackrel{\cdot}{\le} e^{-an}$. Indeed, any $D(\cdot|\boldsymbol{y})$ which does not satisfy this constraint cannot be the optimum strategy, yielding a payoff $u > 1$ which can be improved by always deciding in favor of $\mathcal{H}_0$ ($u = 1$).

Let us define:

$$\tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_0) \triangleq \min_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}:E_{\boldsymbol{x}\boldsymbol{y}}(d(X,Y))\le\Delta\}} \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0), \quad (19)$$

where $E_{\boldsymbol{x}\boldsymbol{y}}(\cdot)$ defines the empirical expectation and the minimization is carried out for a given empirical distribution of $\boldsymbol{y}$, $\hat{P}_{\boldsymbol{y}}$. A similar definition can be given for $\tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_1)$.

Our solution for the *DG-FA2* game is given by the following theorem.

Theorem 5. *Let*

$$D^{\#,1}(\mathcal{H}_1|\boldsymbol{y}) = U\left(\frac{1}{n}\log\frac{Q_1(\boldsymbol{y})}{Q_0(\boldsymbol{y})} - a\right), \quad (20)$$

where $U(\cdot)$ denotes the Heaviside step function, and let A^* be defined as usual. denote the attacking channel in (5)*. Strategy $D^{\#,1}$ is an optimum strategy for the defender.*

If, in addiction, the distortion measure is additive, i.e., $d(\boldsymbol{x},\boldsymbol{y}) = \sum_i d(x_i, y_i)$ for some single-letter distortion function, strategy

$$D^{\#,2}(\mathcal{H}_1|\boldsymbol{y}) = U\left(\tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_1) - a\right) \quad (21)$$

is asymptotically optimum for the defender.

The reason why it is meaningful to provide also the asymptotical optimum strategy, is the following: although strategy $D^{\#,1}$ is preferable for the defender in a game theoretical sense (being optimal for finite n), it requires the non trivial computation of the two probabilities $Q_1(\boldsymbol{y})$ and $Q_0(\boldsymbol{y})$. Strategy $D^{\#,2}$, instead, it is easier to implement because of its single-letter form, and leads to the same payoff asymptotically.

Proof: Since the payoff in (18) a special case of (13) (with $\gamma = 1$ and $\beta = e^{\alpha n}$), Since (18) is a special case of (13) (with $\gamma = 1$ and $\beta = e^{\alpha n}$), for any defence strategy $D \in \mathcal{S}_D$, the asymptotically optimum attacking channel under both hypotheses is A^*, the same and corresponds to the channel A^* defined in (5), see Theorem 3. Then, we can determine the best defence strategy by assuming that the attacker will play (A^*, A^*) and evaluating the best response of the defender. Given the probability distributions $Q_0(\boldsymbol{y})$ and $Q_1(\boldsymbol{y})$ induced by A^*, the optimum decision rule is deterministic and is given by the likelihood ratio test (LRT) [22]:

$$\frac{1}{n}\log\frac{Q_1(\boldsymbol{y})}{Q_0(\boldsymbol{y})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} a, \quad (22)$$

which proves the optimality of the decision rule in (20).

To prove the asymptotic optimality of the decision rule in (21), let us approximate $Q_0(\boldsymbol{y})$ and $Q_1(\boldsymbol{y})$ using the method of types as follows:

$$\begin{aligned}
Q_0(\boldsymbol{y}) &= \sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) A^*(\boldsymbol{y}|\boldsymbol{x}) \\
&\doteq \sum_{\boldsymbol{x}:\ d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta} e^{-n[\hat{H}_{\boldsymbol{x}}(X) + \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0)]} \cdot e^{-n\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(Y|X)} \\
&\doteq \max_{\boldsymbol{x}:\ d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta} e^{nH_{\boldsymbol{x}\boldsymbol{y}}(X|Y)} \cdot \Big(e^{-n[\hat{H}_{\boldsymbol{x}}(X) + \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0)]} \\
&\qquad \cdot e^{-n\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(Y|X)} \Big) \\
&= \max_{\boldsymbol{x}:\ d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta} e^{-n[\hat{H}_{\boldsymbol{y}}(Y) + \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0)]} \\
&\overset{(a)}{\doteq} \exp\Big\{ -n \Big[\hat{H}_{\boldsymbol{y}}(Y) + \\
&\qquad + \min_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}} : E_{\boldsymbol{x}\boldsymbol{y}}(d(X,Y)) \leq \Delta\}} \mathcal{D}(P_X \| P_0) \Big] \Big\} \\
&= \exp\Big\{ -n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_0)] \Big\},
\end{aligned} \tag{23}$$

where in (a) we exploited the additivity of the distortion function d. Similarly,

$$Q_1(\boldsymbol{y}) \doteq \exp\Big\{ -n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_1)] \Big\}. \tag{24}$$

Thus, we have the following asymptotic approximation to the LRT:

$$\tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}(\hat{P}_{\boldsymbol{y}}, P_1) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} a, \tag{25}$$

which proves the second part of the theorem. ■

Given the above, we can assert the following:

Theorem 6. *The profile* $(D^{\#,1}, (A^*, A^*))$ *and* $(D^{\#,2}, (A^*, A^*))$ *are asymptotic rationalizable equilibria for the* DG-FA2 *game.*

As final remark, we observe that the analysis in this section can be easily generalized to any payoff function defined as in (13), i.e., for any $\gamma, \beta \geq 0$.

V. Conclusions

We considered the problem of adversarial binary hypothesis testing when the attack is carried out under both hypotheses, aiming at causing both false negative and false positive errors. By modeling the defender-attacker interaction as a game, we first extended the results in [8] to the case of an attacker which is active under the alternative hypothesis only, then we defined and solved two different versions of the detection game with fully active attacker corresponding to different decision setups: the case of decision based on NP approach and Bayesian approach. Among the possible directions for future work, we mention the extension to the case of multiple hypothesis testing, or classification. Another interesting direction is the extension of the results to more realistic models (of wider applicability), like for instance Markov sources.

References

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, 2013, pp. 8682–8686.

[2] E. Delp, N. Memon, and M. Wu, "Special issue on digital forensics," *IEEE Signal Processing Magazine*, vol. 26, no. 2, March 2009.

[3] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Forensics vs anti-forensics: a decision and game theoretic framework," in *ICASSP 2012, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[4] N. Dalvi, P. Domingos, P. Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 99–108.

[5] S. Marcel, M. S. Nixon, and S. Z. Li, Eds., *Handbook of Biometric Anti-Spoofing - Trusted Biometrics under Spoofing Attacks*, ser. Advances in Computer Vision and Pattern Recognition. Springer, 2014. [Online]. Available: http://dx.doi.org/10.1007/978-1-4471-6524-8

[6] I. Cox, M. Miller, and J. Bloom, *Digital watermarking*. Morgan Kaufmann, 2002.

[7] A. Somekh-Baruch and N. Merhav, "On the capacity game of public watermarking systems," *IEEE Transactions on Information Theory*, vol. 50, no. 3, pp. 511–524, March 2004.

[8] M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, March 2013.

[9] N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 255–274, Jan 2008.

[10] ——, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 255–274, January 2008.

[11] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. PP, no. 99, pp. 1–1, 2014.

[12] ——, "Source distinguishability under corrupted training," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec 2014, pp. 197–202.

[13] M.Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, March 2008.

[14] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 227–236, March 2011.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

[16] I. Csiszar, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, October 1998.

[17] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems. 2nd edition*. Cambridge University Press, 2011.

[18] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[19] Y. C. Chen, N. Van Long, and X. Luo, "Iterated strict dominance in general games," *Games and Economic Behavior*, vol. 61, no. 2, pp. 299–315, November 2007.

[20] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.

[21] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, 1965.

[22] H. L. Van Trees, *Detection, estimation and modulation theory. vol. 2. , nonlinear modulation theory*. New York: J. Wiley and sons, 1971. [Online]. Available: http://opac.inria.fr/record=b1108665