

# Detection Games Under Fully Active Adversaries

Benedetta Tondi, *Member, IEEE*, Neri Merhav, *Fellow, IEEE*, Mauro Barni *Fellow, IEEE*

## Abstract

We study a binary hypothesis testing problem in which a defender must decide whether or not a test sequence has been drawn from a given memoryless source  $P_0$  whereas, an attacker strives to impede the correct detection. With respect to previous works, the adversarial setup addressed in this paper considers an attacker who is active under both hypotheses, namely, a fully active attacker, as opposed to a partially active attacker who is active under one hypothesis only. In the fully active setup, the attacker distorts sequences drawn both from  $P_0$  and from an alternative memoryless source  $P_1$ , up to a certain distortion level, which is possibly different under the two hypotheses, in order to maximize the confusion in distinguishing between the two sources, i.e., to induce both false positive and false negative errors at the detector, also referred to as the defender. We model the defender-attacker interaction as a game and study two versions of this game, the Neyman-Pearson game and the Bayesian game. Our main result is in the characterization of an attack strategy that is asymptotically both dominant (i.e., optimal no matter what the defender's strategy is) and universal, i.e., independent of  $P_0$  and  $P_1$ . From the analysis of the equilibrium payoff, we also derive the best achievable performance of the defender, by relaxing the requirement on the exponential decay rate of the false positive error probability in the Neyman-Pearson setup and the tradeoff between the error exponents in the Bayesian setup. Such analysis permits to characterize the conditions for the distinguishability of the two sources given the distortion levels.

## Index Terms

Adversarial signal processing, binary hypothesis testing, statistical detection theory, game theory, the method of types.

M. Barni and B. Tondi are with the Department of Information Engineering and Mathematical Sciences, University of Siena, Siena, ITALY, e-mail: {benedettatondi@gmail.com, barni@dii.unisi.it}; N. Merhav is with the the Andrew and Erna Viterbi Faculty of Electrical Engineering - Israel Institute of Technology Technion City, Haifa, ISRAEL, email: {merhav@ee.technion.ac.il}.

## I. INTRODUCTION

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

In particular, the need for adversarial modeling has become evident in security-related applications and game theory is often harnessed as a useful tool in many research areas, such as steganalysis [7], watermarking [4], intrusion detection systems [8] and adversarial machine learning [9], [10]. In recent literature, game theory and information theory have also been combined to address the problem of adversarial detection, especially in the field of digital watermarking, see, for instance, [4], [11], [12], [13]. In all these works, the problem of designing watermarking codes that are robust to intentional attacks, is studied as a game between the information hider and the attacker.

An attempt to develop a general theory for the binary hypothesis testing problem in the presence of an adversary was made in [14]. Specifically, in [14] the general problem of binary decision 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 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 [14] is to cause 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  $\mathcal{H}_0$ . The attacker is subjected to a distortion constraint, which limits his freedom in doing so. Such a struggle between the defender and the attacker is modeled in [14] 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 his decision on the analysis of first order statistics only. In this respect, the analysis conducted in [14] extends the one of [15] to the adversarial scenario. Some variants of this attack-detection game have also been studied: in [16], 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 parties: 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 [17]).

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 camera 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 implant the fingerprint from another camera [18], [19]. Another example comes from watermarking, where an attacker can be interested in either removing or injecting the watermark from an image or a video, to redistribute the content with a fake copyright and no information (erased information) about the true ownership [20]. Attacks under both hypotheses may also be present in applications of network intrusion detection [21]. Network intrusion detection systems, in fact, can be subject to both evasion attacks [22], in which an adversary tries to avoid detection by manipulating malicious traffic, and overstimulation attacks [23], [24], in which the network is overstimulated by an adversary who sends synthetic traffic (matching the legitimate traffic) in order to cause a denial of service.

With the above ideas in mind, in this paper, we consider the game-theoretic formulation of the defender-attacker interaction when the attacker acts under both hypotheses. We refer to this scenario as a detection game with a *fully active attacker*. By contrast, when the attacker acts under hypothesis  $\mathcal{H}_1$  only (as in [14] and [16]), he is referred to as a *partially active attacker*. A distinction is made between the case where the underlying hypothesis is known to the attacker and the case where it is not. A little thought, however, immediately indicates that the latter is a special case of the former, and therefore, we focus on the former. We define and solve two versions of the *detection game with fully active attackers*, corresponding to two different formulations of the problem: the Neyman-Pearson formulation and the Bayesian formulation. In contrast to [14], here the players are allowed to adopt randomized strategies. Specifically, the defender adopts *randomized decision* strategies, while in [14] the defender's strategies were confined to deterministic decision rules. As for the attack, it consists of the application of a *channel*, whereas in [14] it was confined to the application of a deterministic function. Moreover, the partially active case of [14] can easily be obtained as a special case of the fully active case considered here. The problem of solving the game and then finding the optimum detector in the adversarial setting is not trivial and may not be possible in general. Thus, we limit the complexity of the problem and make the analysis tractable by confining the decision to depend on a given set of statistics of the observation. Such an assumption, according to which the detector has access to a limited set of empirical statistics of the sequence, is referred to as *limited resources* assumption (see [15] for an introduction on this terminology). In particular, as done in [14], [16], we limit the detection resources to first order statistics, which are, as is well known, sufficient statistics for memoryless systems [25, Section 2.9]. While the sources are indeed

assumed memoryless, one might still be concerned regarding the sufficiency of first order statistics, in our setting, since the attack channel is not assumed memoryless in the first place. Adopting, nonetheless, the limited-resources assumption to first order statistics, is motivated mainly by its simplicity, but with the understanding that the results can easily be extended to deal with arbitrarily higher order empirical statistics as well. Moreover, an important bonus of this framework is that it allows us to obtain fairly strong results concerning the game between the defender and the attacker, as will be described below.

One of the main results of this paper is the characterization of an attack strategy which is both *dominant* (i.e., optimal no matter what the defence strategy is), and *universal*, i.e., independent of the (unknown) underlying sources. Moreover, this optimal attack is the same for both the Neyman-Pearson and Bayesian games. This result continues to hold also for the partially active case, thus creating a significant difference relative to previous works, where the existence of a dominant strategy was established regarding the defender only.

Some of our results (in particular, the derivation of the equilibrium point for both the Neyman-Pearson and the Bayesian games), have already appeared mostly without proofs in [26]. Here we provide the full proofs of the main theorems, evaluate the payoff at equilibrium for both the Neyman-Pearson and Bayesian games and include the analysis of the ultimate performance of the games. Specifically, we characterize the so called indistinguishability region (to be defined formally in Section VI), namely the set of the sources for which it is not possible to attain strictly positive exponents for both false positive and false negative probabilities under the Neyman-Pearson and the Bayesian settings. Furthermore, the setup and analysis presented in [26] is extended by considering a more general case in which the maximum allowed distortion levels the attacker may introduce under the two hypotheses are different.

The paper is organized as follows. In Section II, we establish the notation and introduce the main concepts. In Section III, we formalize the problem and define the detection game with a fully active adversary for both the Neyman-Pearson and the Bayesian games, and then prove the existence of a dominant and universal attack strategy. The complete analysis of the Neyman-Pearson and Bayesian detection games, namely, the study of the equilibrium point of the game and the computation of the payoff at the equilibrium, are carried out in Sections IV and V, respectively. Finally, Section VI is devoted to the analysis of the best achievable performance of the defender and the characterization of the source distinguishability.

## II. NOTATION AND DEFINITIONS

Throughout the paper, random variables will be denoted by capital letters and specific realizations will be denoted by the corresponding lower case letters. All random variables that denote signals in the system, will be assumed to have the same finite alphabet, denoted by  $\mathcal{A}$ . Given a random variable  $X$  and a positive integer  $n$ , we denote by  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ ,  $X_i \in \mathcal{A}$ ,  $i = 1, 2, \dots, n$ , a sequence of  $n$  independent copies of  $X$ . According to the above-mentioned notation rules, a specific realization of  $\mathbf{X}$  is denoted by  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ . Sources will be denoted by the letter  $P$ . Whenever necessary, we will subscript  $P$  with the name of the relevant random variables: given a random variable  $X$ ,  $P_X$  denotes its probability mass function (PMF). Similarly,  $P_{XY}$  denotes the joint PMF of a pair of random variables,  $(X, Y)$ . For two positive sequences,  $\{a_n\}$  and  $\{b_n\}$ , the notation  $a_n \doteq b_n$  stands for exponential equivalence, i.e.,  $\lim_{n \rightarrow \infty} 1/n \ln(a_n/b_n) = 0$ , and  $a_n \dot{\leq} b_n$  designates that  $\limsup_{n \rightarrow \infty} 1/n \ln(a_n/b_n) \leq 0$ . For a given real  $s$ , we denote  $[s]_+ \triangleq \max\{s, 0\}$ . We use notation  $U(\cdot)$  for the Heaviside step function.

The type of a sequence  $\mathbf{x} \in \mathcal{A}^n$  is defined as the empirical probability distribution  $\hat{P}_{\mathbf{x}}$ , that is, the vector  $\{\hat{P}_{\mathbf{x}}(x), x \in \mathcal{A}\}$  of the relative frequencies of the various alphabet symbols in  $\mathbf{x}$ . A type class  $\mathcal{T}(\mathbf{x})$  is defined as the set of all sequences having the same type as  $\mathbf{x}$ . When we wish to emphasize the dependence of  $\mathcal{T}(\mathbf{x})$  on  $\hat{P}_{\mathbf{x}}$ , we will use the notation  $\mathcal{T}(\hat{P}_{\mathbf{x}})$ . Similarly, given a pair of sequences  $(\mathbf{x}, \mathbf{y})$ , both of length  $n$ , the joint type class  $\mathcal{T}(\mathbf{x}, \mathbf{y})$  is the set of sequence pairs  $\{(\mathbf{x}', \mathbf{y}')\}$  of length  $n$  having the same empirical joint probability distribution (or joint type) as  $(\mathbf{x}, \mathbf{y})$ ,  $\hat{P}_{\mathbf{x}\mathbf{y}}$ , and the conditional type class  $\mathcal{T}(\mathbf{y}|\mathbf{x})$  is the set of sequences  $\{\mathbf{y}'\}$  with  $\hat{P}_{\mathbf{x}\mathbf{y}'} = \hat{P}_{\mathbf{x}\mathbf{y}}$ .

Regarding information measures, the entropy associated with  $\hat{P}_{\mathbf{x}}$ , which is the empirical entropy of  $\mathbf{x}$ , is denoted by  $\hat{H}_{\mathbf{x}}(X)$ . Similarly,  $\hat{H}_{\mathbf{x}\mathbf{y}}(X, Y)$  designates the empirical joint entropy of  $\mathbf{x}$  and  $\mathbf{y}$ , and  $\hat{H}_{\mathbf{x}\mathbf{y}}(X|Y)$  is the conditional joint entropy. We denote by  $\mathcal{D}(P||Q)$  the Kullback–Leibler (K-L) divergence between two sources,  $P$  and  $Q$  with the same alphabet (see [25]).

Finally, we use letter  $A$  to denote an attack channel; accordingly,  $A(\mathbf{y}|\mathbf{x})$  is the conditional probability of the channel output  $\mathbf{y}$  given the channel input  $\mathbf{x}$ . Given a permutation-invariant distortion function<sup>1</sup>  $d : \mathcal{A}^n \times \mathcal{A}^n \rightarrow \mathbb{R}^+$  and a maximum distortion  $\Delta$ , we define the class  $\mathcal{C}_{\Delta}$  of admissible channels  $\{A(\mathbf{y}|\mathbf{x}), \mathbf{x}, \mathbf{y} \in \mathcal{A}^n\}$  as those that assign zero probability to every  $\mathbf{y}$  with  $d(\mathbf{x}, \mathbf{y}) > n\Delta$ .

<sup>1</sup>A permutation-invariant distortion function,  $d(\mathbf{x}, \mathbf{y})$ , is a distortion function that is invariant if the same permutation is applied to both  $\mathbf{x}$  and  $\mathbf{y}$ .

### A. Basics of Game Theory

For the sake of completeness, we introduce some basic definitions and concepts of game theory. A two-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 sets of strategies from which the first and the second player can choose, respectively, 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 one chooses  $s_{2,j}$ . Each player aims at maximizing its payoff function. A pair of strategies  $(s_{1,i}, s_{2,j})$  is called a *profile*. When  $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$ , the game is said to be a *zero-sum game*. For such games, the payoff of the game  $u(s_{1,i}, s_{2,j})$  is usually defined by adopting the perspective of one of the two players: that is,  $u(s_{1,i}, s_{2,j}) = u_1(s_{1,i}, s_{2,j}) = -u_2(s_{1,i}, s_{2,j})$  if the defender's perspective is adopted or vice versa. The sets  $\mathcal{S}_1, \mathcal{S}_2$  and the payoff functions are assumed known to both players. In addition, we consider *strategic games*, i.e., games in which the players choose their strategies ahead of time, without knowing the strategy chosen by the opponent.

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 [27]. The most famous notion of equilibrium is due to Nash [28]. 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 Nash equilibrium. A particular kind of games for which stronger forms of equilibrium exist are the so called *dominance solvable* games [27]. 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 this player, i.e., the strategy that maximizes the payoff, no matter what the strategy of the opponent may be. 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 strategy of the other player. Recursive elimination of dominated strategies is a 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<sup>2</sup> would ever use them. In this way, a new, smaller game is obtained. At this point, some strategies that were not dominated before, may become dominated in the new, smaller version of the game, and hence are eliminated as well. The process goes on until no dominated strategy exists for either player. A *rationalizable equilibrium* is any profile which survives the iterated elimination of dominated strategies

<sup>2</sup>In game theory, a rational player is supposed to act in a way that maximizes its payoff.

[29], [30]. 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. Another, related, 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.

### III. DETECTION GAME WITH FULLY ACTIVE ATTACKER

#### A. Problem formulation

Given two discrete memoryless sources,  $P_0$  and  $P_1$ , defined over a common finite alphabet  $\mathcal{A}$ , we denote by  $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{A}^n$  a sequence emitted by one of these sources. The sequence  $\mathbf{x}$  is available to the attacker. Let  $\mathbf{y} = (y_1, y_2, \dots, y_n) \in \mathcal{A}^n$  denote the sequence observed by the defender: when an attack occurs under both  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , the observed sequence  $\mathbf{y}$  is obtained as the output of an attack channel fed by  $\mathbf{x}$ .

In principle, we must distinguish between two cases: in the first, 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 different conditional probability distributions, i.e., two different attack channels:  $A_0(\mathbf{y}|\mathbf{x})$ , applied when  $\mathcal{H}_0$  holds, and  $A_1(\mathbf{y}|\mathbf{x})$ , applied under  $\mathcal{H}_1$ . Let us denote by  $Q_i(\cdot)$  the PMF of  $\mathbf{y}$  under  $\mathcal{H}_i, i = 0, 1$ . The attack induces the following PMFs on  $\mathbf{y}$ :  $Q_0(\mathbf{y}) = \sum_{\mathbf{x}} P_0(\mathbf{x})A_0(\mathbf{y}|\mathbf{x})$  and  $Q_1(\mathbf{y}) = \sum_{\mathbf{x}} P_1(\mathbf{x})A_1(\mathbf{y}|\mathbf{x})$ .

Clearly, in the hypothesis-unaware case, the attacker will apply the same channel under  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , that is,  $A_0 = A_1$ , and we will denote the common attack channel simply by  $A$ . Throughout the paper, we focus on the hypothesis-aware case as in view of this formalism, the hypothesis-unaware case is just a special case.

Regarding the defender, we assume a randomized decision strategy, defined by  $\Phi(\mathcal{H}_i|\mathbf{y})$ , which designates the probability of deciding in favor of  $\mathcal{H}_i, i = 0, 1$ , given  $\mathbf{y}$ . Accordingly, the probability of a *false positive* (FP) decision error is given by

$$P_{\text{FP}}(\Phi, A_0) = \sum_{\mathbf{y}} Q_0(\mathbf{y})\Phi(\mathcal{H}_1|\mathbf{y}), \quad (1)$$

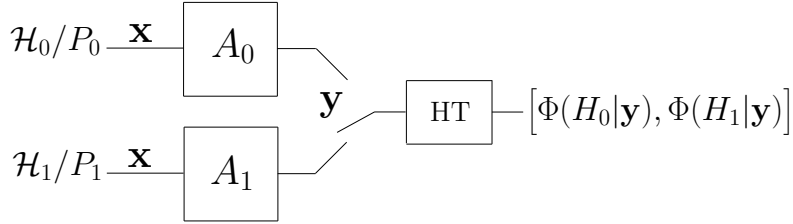


Fig. 1. Schematic representation of the adversarial setup considered in this paper. In the case of partially active attacker, channel  $A_0$  corresponds to the identity channel.

and similarly, the *false negative* (FN) probability assumes the form:

$$P_{\text{FN}}(\Phi, A_1) = \sum_{\mathbf{y}} Q_1(\mathbf{y}) \Phi(\mathcal{H}_0|\mathbf{y}). \quad (2)$$

Figure 1 provides a block diagram of the system with a fully active attacker. Obviously, the partially active case, where no attack occurs under  $\mathcal{H}_0$ , can be seen as a degenerate case of the fully active one, where  $A_0$  is the identity channel  $I$ . As in [14], due to the limited resources assumption, the defender makes a decision based on first order empirical statistics of  $\mathbf{y}$ , which implies that  $\Phi(\cdot|\mathbf{y})$  depends on  $\mathbf{y}$  only via its type class  $\mathcal{T}(\mathbf{y})$ .

Concerning the attack, in order to limit the amount of distortion, we assume a distortion constraint. In the hypothesis-aware case, we allow the attacker different distortion levels,  $\Delta_0$  and  $\Delta_1$ , under  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , respectively. Then,  $A_0 \in \mathcal{C}_{\Delta_0}$  and  $A_1 \in \mathcal{C}_{\Delta_1}$ , where, for simplicity, we assume that a common (permutation-invariant) distortion function  $d(\cdot, \cdot)$  is adopted in the two cases.

### B. Definition of the Neyman–Pearson and Bayesian Games

One of the difficulties associated with the fully active setting is that, in the presence of a fully active attacker, both the FP and FN probabilities depend on the attack channels. We therefore consider two different approaches which lead to different formulations of the detection game: in the first, the detection game is based on the Neyman-Pearson criterion, and in the second one, the Bayesian approach is adopted.

For the Neyman-Pearson setting, we define the game by assuming that the defender adopts a conservative approach and imposes an FP constraint pertaining to the worst-case attack under  $\mathcal{H}_0$ .

**Definition 1.** *The Neyman-Pearson detection game is a zero-sum, strategic game defined as follows.*

- *The set  $\mathcal{S}_D$  of strategies allowed to the defender is the class of randomized decision rules  $\{\Phi\}$  that satisfy*



- (i)  $\Phi(\mathcal{H}_0|\mathbf{y})$  depends on  $\mathbf{y}$  only via its type.
- (ii)  $\max_{A_0 \in \mathcal{C}_{\Delta_0}} P_{FP}(\Phi, A_0) \leq e^{-n\lambda}$  for a prescribed constant  $\lambda > 0$ , independent of  $n$ .
- The set  $\mathcal{S}_A$  of strategies allowed to the attacker is the class of pairs of attack channels  $(A_0, A_1)$  such that  $A_0 \in \mathcal{C}_{\Delta_0}$ ,  $A_1 \in \mathcal{C}_{\Delta_1}$ ; that is,  $\mathcal{S}_A = \mathcal{C}_{\Delta_0} \times \mathcal{C}_{\Delta_1}$ .
- The payoff of the game is  $u(\Phi, A_1) = P_{FN}(\Phi, A_1)$ ; the attacker is in the quest of maximizing  $u(\Phi, A_1)$  whereas the defender wishes to minimize it.

In the above definition, we require that the FP probability decays exponentially fast with  $n$ , with an exponential rate *at least* as large as  $\lambda$ . In the case of partially-active attack (see the formulation in [26]), the FP probability does not depend on the attack but on the defender only; accordingly, the constraint imposed by the defender in the above formulation becomes  $P_{FP}(\Phi) \leq e^{-n\lambda}$ . Regarding the attacker, we have  $\mathcal{S}_A \equiv \mathcal{C}_0 \times \mathcal{C}_{\Delta_1}$ , where  $\mathcal{C}_0$  is a singleton that contains the identity channel only.

Another version of the detection game is defined by assuming that the defender follows a less conservative approach, that is, the Bayesian approach, and tries to minimize a particular Bayes risk.

**Definition 2.** *The Bayesian detection game is a zero-sum, strategic game defined as follow.*

- The set  $\mathcal{S}_D$  of strategies allowed to the defender is the class of the randomized decision rules  $\{\Phi\}$  where  $\Phi(\mathcal{H}_0|\mathbf{y})$  depends on  $\mathbf{y}$  only via its type.
- The set  $\mathcal{S}_A$  of strategies allowed to the attacker is  $\mathcal{S}_A = \mathcal{C}_{\Delta_0} \times \mathcal{C}_{\Delta_1}$ .
- The payoff of the game is

$$u(\Phi, (A_0, A_1)) = P_{FN}(\Phi, A_1) + e^{an} P_{FP}(\Phi, A_0), \quad (3)$$

for some constant  $a$ , independent of  $n$ .

We observe that, in the definition of the payoff, the parameter  $a$  controls the tradeoff between the two terms in the exponential scale; whenever possible, the optimum defence strategy is expected to yield error exponents that differ exactly by  $a$ , so as to balance the contributions of the two terms of (3).

Notice also that, by defining the payoff as in (3), we are implicitly considering for the defender only the strategies  $\Phi(\cdot|\mathbf{y})$  such that  $P_{FP}(\Phi, A_0) \leq e^{-an}$ . In fact, any strategy that does not satisfy this inequality yields a payoff  $u > 1$ , that cannot be optimal, as it can be improved by always deciding in favor of  $\mathcal{H}_0$  regardless of  $\mathbf{y}$  ( $u = 1$ ).

As in [14], we focus on the asymptotic behavior of the game as  $n$  tends to infinity. In particular, we

are interested in the FP and FN exponents defined as:

$$\varepsilon_{\text{FP}} = -\limsup_{n \rightarrow \infty} \frac{\ln P_{\text{FP}}(\Phi, A_0)}{n}; \quad \varepsilon_{\text{FN}} = -\limsup_{n \rightarrow \infty} \frac{\ln P_{\text{FN}}(\Phi, A_1)}{n}. \quad (4)$$

We say that a strategy is *asymptotically optimum* (or *dominant*) if it is optimum (dominant) with respect to the asymptotic exponential decay rate (or the exponent, for short) of the payoff.

### C. Asymptotically Dominant and Universal Attack

In this subsection, we characterize an attack channel that, for both games, is asymptotically dominant and universal, in the sense of being independent of the unknown underlying sources. This result paves the way to the solution of the two games.

Let  $u$  denote a generic payoff function of the form

$$u = \gamma P_{\text{FN}}(\Phi, A_1) + \beta P_{\text{FP}}(\Phi, A_0), \quad (5)$$

where  $\beta$  and  $\gamma$  are given positive constants, possibly dependent on  $n$ .

We notice that the payoff of the Neyman-Pearson and Bayesian games defined in the previous section can be obtained as particular cases: specifically,  $\gamma = 1$  and  $\beta = 0$  for the Neyman-Pearson game and  $\gamma = 1$  and  $\beta = e^{an}$  for the Bayesian one.

**Theorem 1.** *Let  $c_n(\mathbf{x})$  denote the reciprocal of the total number of conditional type classes  $\{\mathcal{T}(\mathbf{y}|\mathbf{x})\}$  that satisfy the constraint  $d(\mathbf{x}, \mathbf{y}) \leq n\Delta$  for a given  $\Delta > 0$ , namely, admissible conditional type classes<sup>3</sup>.*

*Define:*

$$A_{\Delta}^*(\mathbf{y}|\mathbf{x}) = \begin{cases} \frac{c_n(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} & d(\mathbf{x}, \mathbf{y}) \leq n\Delta \\ 0 & \text{elsewhere} \end{cases}. \quad (6)$$

*Among all pairs of channels  $(A_0, A_1) \in \mathcal{S}_A$ , the pair  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  minimizes the asymptotic exponent of  $u$  for every  $P_0, P_1$ , every  $\gamma, \beta \geq 0$  and every permutation-invariant  $\Phi(\mathcal{H}_0|\cdot)$ .*

*Proof:* We first focus on the attack under  $\mathcal{H}_1$  and therefore on the FN probability.

Consider an arbitrary channel  $A_1 \in \mathcal{C}_{\Delta_1}$ . Let  $\Pi : \mathcal{A}^n \rightarrow \mathcal{A}^n$  denote a permutation operator that permutes any member of  $\mathcal{A}^n$  according to a given permutation matrix and let

$$A_{\Pi}(\mathbf{y}|\mathbf{x}) \triangleq A_1(\Pi\mathbf{y}|\Pi\mathbf{x}). \quad (7)$$

<sup>3</sup>From the method of the types it is known that  $1 \geq c_n(\mathbf{x}) \geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)}$  for any  $\mathbf{x}$  [25].

Since the distortion function is assumed permutation–invariant, the channel  $A_{\Pi}(\mathbf{y}|\mathbf{x})$  introduces the same distortion as  $A_1$  and hence satisfies the distortion constraint. Due to the memorylessness of  $P_1$  and the assumption that  $\Phi(\mathcal{H}_0|\mathbf{y})$  belongs to  $\mathcal{S}_D$ , we have:

$$\begin{aligned}
 P_{\text{FN}}(\Phi, A_{\Pi}) &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\mathbf{x}) A_{\Pi}(\mathbf{y}|\mathbf{x}) \Phi(\mathcal{H}_0|\mathbf{y}) \\
 &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\mathbf{x}) A_1(\Pi\mathbf{y}|\Pi\mathbf{x}) \Phi(\mathcal{H}_0|\mathbf{y}) \\
 &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\Pi\mathbf{x}) A_1(\Pi\mathbf{y}|\Pi\mathbf{x}) \Phi(\mathcal{H}_0|\Pi\mathbf{y}) \\
 &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\mathbf{x}) A_1(\mathbf{y}|\mathbf{x}) \Phi(\mathcal{H}_0|\mathbf{y}) \\
 &= P_{\text{FN}}(\Phi, A_1),
 \end{aligned} \tag{8}$$

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

$$\bar{A}(\mathbf{y}|\mathbf{x}) = \frac{1}{n!} \sum_{\Pi} A_{\Pi}(\mathbf{y}|\mathbf{x}) = \frac{1}{n!} \sum_{\Pi} A_1(\Pi\mathbf{y}|\Pi\mathbf{x}), \tag{9}$$

which also introduces the same distortion as  $A_1$ . Now, notice that this channel assigns the same conditional probability to all sequences in the same conditional type class  $\mathcal{T}(\mathbf{y}|\mathbf{x})$ . To see why this is true, we observe that any sequence  $\mathbf{y}' \in \mathcal{T}(\mathbf{y}|\mathbf{x})$  can be seen as being obtained from  $\mathbf{y}$  through the application of a permutation  $\Pi'$  which leaves  $\mathbf{x}$  unaltered. Then, we have:

$$\begin{aligned}
 \bar{A}(\mathbf{y}'|\mathbf{x}) &= \bar{A}(\Pi'\mathbf{y}|\Pi'\mathbf{x}) = \frac{1}{n!} \sum_{\Pi} A_1(\Pi(\Pi'\mathbf{y})|\Pi(\Pi'\mathbf{x})) \\
 &= \frac{1}{n!} \sum_{\Pi} A_1(\Pi\mathbf{y}|\Pi\mathbf{x}) = \bar{A}(\mathbf{y}|\mathbf{x}).
 \end{aligned} \tag{10}$$

Therefore, since  $\bar{A}(\mathcal{T}(\mathbf{y}|\mathbf{x})|\mathbf{x}) \leq 1$ , we argue that

$$\begin{aligned}
 \bar{A}(\mathbf{y}|\mathbf{x}) &\leq \begin{cases} \frac{1}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} & d(\mathbf{x}, \mathbf{y}) \leq n\Delta \\ 0 & \text{elsewhere} \end{cases} \\
 &= \frac{A_{\Delta_1}^*(\mathbf{y}|\mathbf{x})}{c_n(\mathbf{x})} \\
 &\leq (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} A_{\Delta_1}^*(\mathbf{y}|\mathbf{x}),
 \end{aligned} \tag{11}$$

which implies that, for every permutation–invariant defence strategy  $\Phi$ ,

$$P_{\text{FN}}(\Phi, A_1) \leq (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} P_{\text{FN}}(A_{\Delta_1}^*, \Phi) \tag{12}$$

or equivalently

$$P_{\text{FN}}(\Phi, A_{\Delta_1}^*) \geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} P_{\text{FN}}(A_1, \Phi). \tag{13}$$

We conclude that  $A_{\Delta_1}^*$  minimizes the error exponent of  $P_{\text{FN}}(\Phi, A_1)$  across all channels in  $\mathcal{C}_{\Delta_1}$  and for every  $\Phi \in \mathcal{S}_D$ , regardless of  $P_1$ .

A similar argument applies to the FP probability to derive the optimum channel under  $\mathcal{H}_0$ ; that is, from the memorylessness of  $P_0$  and the permutation-invariance of  $\Phi(\mathcal{H}_1|\cdot)$ , we have:

$$P_{\text{FP}}(\Phi, A_{\Delta_0}^*) \geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} P_{\text{FP}}(A_0, \Phi), \quad (14)$$

for every  $A_0 \in \mathcal{C}_{\Delta_0}$ . Accordingly,  $A_{\Delta_0}^*$  minimizes the error exponent of  $P_{\text{FP}}(\Phi, A_0)$ .

We then have:

$$\begin{aligned} & \gamma P_{\text{FN}}(\Phi, A_1) + \beta P_{\text{FP}}(\Phi, A_0) \\ & \leq (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} (\gamma P_{\text{FN}}(\Phi, A_{\Delta_1}^*) + \beta P_{\text{FP}}(\Phi, A_{\Delta_0}^*)) \\ & \doteq \gamma P_{\text{FN}}(\Phi, A_{\Delta_1}^*) + \beta P_{\text{FP}}(\Phi, A_{\Delta_0}^*), \end{aligned} \quad (15)$$

for every  $A_0 \in \mathcal{C}_{\Delta_0}$  and  $A_1 \in \mathcal{C}_{\Delta_1}$ . Notice that, since the asymptotic equality is defined in the logarithmic scale, eq. (15) holds no matter what the values of  $\beta$  and  $\gamma$  are, including values that depend on  $n$ . Hence, the pair of channels  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  minimizes the asymptotic exponent of  $u$  for any permutation-invariant decision rule  $\Phi(\mathcal{H}_0|\cdot)$  and for any  $\gamma, \beta \geq 0$ . ■

According to Theorem 1, for every zero-sum game with payoff function of the form in (5), if  $\Phi$  is permutation-invariant, the pair of attack channels which is the most favorable to the attacker is  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$ , which does not depend on  $\Phi$ . Then, the optimum attack strategy  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  is *dominant*. Specifically, given  $\mathbf{x}$ , in order to generate  $\mathbf{y}$  which causes a detection error with the prescribed maximum allowed distortion, the attacker cannot do any better than randomly selecting an admissible conditional type class according to the uniform distribution and then choose at random  $\mathbf{y}$  within this conditional type class. Figure 2 illustrates the intuition behind the definition of the attack channel in (6): since the number of conditional type classes is only polynomial in  $n$ , the random choice of the conditional type class does not affect the exponent of the error probabilities; besides, since the decision is the same for all sequences within the same conditional type class, the choice of  $\mathbf{y}$  within that conditional type class is immaterial.

As an additional result, Theorem 1 states that, whenever an adversary aims at maximizing a payoff function of the form (5), and as long as the defence strategy is confined to the analysis of the first order statistics, the (asymptotically) optimum attack strategy is *universal* w.r.t. the sources  $P_0$  and  $P_1$ , i.e., it depends neither on  $P_0$  nor on  $P_1$ .

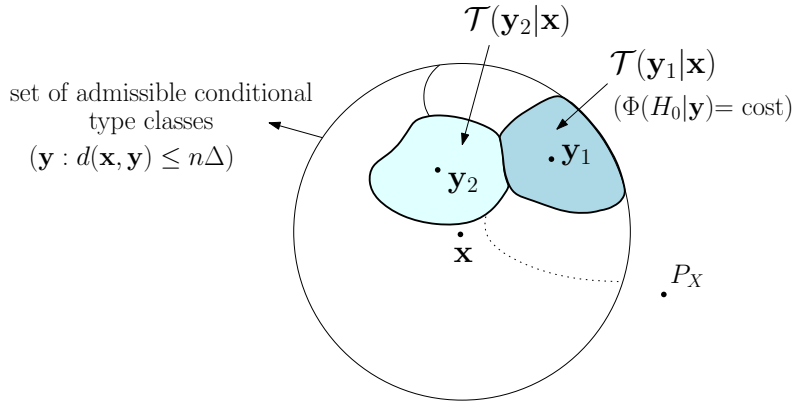


Fig. 2. Graphical interpretation of the behavior of the attack channel  $A_\Delta^*$ .

Finally, if  $\Delta_0 = \Delta_1 = \Delta$ , the optimum attack consists of applying the same channel  $A_\Delta^*$  regardless of the underlying hypothesis and then the optimum attack strategy is *fully-universal*: the attacker needs to know neither the sources ( $P_0$  and  $P_1$ ), nor the underlying hypothesis. In this case, it becomes immaterial whether the attacker is aware or unaware of the true hypothesis. As a consequence of this property, in the hypothesis-unaware case, when the attacker applies the same channel under both hypotheses, subject to a fixed maximum distortion  $\Delta$ , the optimum channel remains  $A_\Delta^*$ .

As a final remark, according to Theorem 1, for the partially active case, there exists an (asymptotically) dominant and universal attack channel. This result marks a considerable difference relative to the results of [14], where the optimum deterministic attack function is found using the rationalizability argument, that is, by exploiting the existence of a dominant defence strategy, and hence it is neither dominant nor universal.

#### IV. THE NEYMAN-PEARSON DETECTION GAME

In this section, we study the detection game with a fully active attacker in the Neyman-Pearson setup as defined in Definition 1. From the analysis of Section III-C, we already know that there exists a dominant attack strategy. Regarding the defender, we will determine the asymptotically optimum strategy regardless of the dominant pair of attack channels; in particular, as will be seen in Lemma 1 below, an asymptotically dominant defense strategy can be derived from a detailed analysis of the FP constraint. As a consequence, the Neyman-Pearson detection game has a dominant equilibrium.

##### A. Optimal Detection and Game Equilibrium

The following lemma characterizes the optimal detection strategy in the Neyman-Pearson setting.

**Lemma 1.** *For the Neyman-Pearson game of Definition 1, the defence strategy*

$$\Phi^*(\mathcal{H}_1|\mathbf{y}) \triangleq \exp \left\{ -n \left[ \lambda - \min_{\mathbf{x}:d(\mathbf{x},\mathbf{y})\leq n\Delta_0} \mathcal{D}(\hat{P}_{\mathbf{x}}\|P_0) \right]_+ \right\}, \quad (16)$$

*is asymptotically dominant for the defender.*

The proof appears in Appendix I-A.

We point out that when the attacker is partially-active, it is known from [26] that the optimum defence strategy is

$$\Phi^*(\mathcal{H}_1|\mathbf{y}) \triangleq \exp \left\{ -n \left[ \lambda - \mathcal{D}(\hat{P}_{\mathbf{y}}\|P_0) \right]_+ \right\}. \quad (17)$$

From (17), it is easy to argue that there exists a deterministic strategy, corresponding to the Hoeffding test [31], which is asymptotically equivalent to  $\Phi^*(\mathcal{H}_1|\mathbf{y})$ . This result is in line with the one in [14] (Lemma 1), where the class of defence strategies is confined to deterministic decision rules.

Intuitively, the extension from (17) to (16) is explained as follows. In the case of fully active attacker, the defender is subject to a constraint on the maximum FP probability over  $\mathcal{S}_A$ , that is, the set of the admissible channels  $A \in \mathcal{C}_{\Delta_0}$  (see Definition 1). From the analysis of Section III-C, channel  $A_{\Delta_0}^*$  minimizes the FP exponent over this set. In order to satisfy the constraint for a given sequence  $\mathbf{y}$ , the defender must handle the worst-case value (i.e., the minimum) of  $\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_0)$  over all the type classes  $\mathcal{T}(\mathbf{x}|\mathbf{y})$  which satisfy the distortion constraint, or equivalently, all the sequences  $\mathbf{x}$  such that  $d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0$ .

According to Lemma 1, the best defence strategy is asymptotically dominant. Also, since  $\Phi^*$  depends on  $P_0$  only, and not on  $P_1$ , it is referred to as *semi-universal*.

Concerning the attacker, since the payoff is a special case of (5) with  $\gamma = 1$  and  $\beta = 0$ , the optimum pair of attack channels is given by Theorem 1 and corresponds to  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$ .

The following comment is in order. Since the payoff of the game is defined in terms of the FN probability only, it is independent of  $A_0 \in \mathcal{C}_{\Delta_0}$ . Furthermore, since the defender adopts a conservative approach to guarantee the FP constraint for every  $A_0$ , the constraint is satisfied for every  $A_0$  and therefore all channel pairs of the form  $(A_0, A_{\Delta_1}^*)$ ,  $A_0 \in \mathcal{S}_A$ , are equivalent in terms of the payoff. Accordingly, in the hypothesis-aware case, the attacker can employ any admissible channel under  $\mathcal{H}_0$ . In the Neyman-Pearson setting, the sole fact that the attacker is active under  $\mathcal{H}_0$  forces the defender to take countermeasures that make the choice of  $A_0$  immaterial.

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

**Theorem 2.** Consider the Neyman-Pearson detection game of Definition 1. Let  $\Phi^*$  and  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  be the strategies defined in Lemma 1 and Theorem 1, respectively. The profile  $(\Phi^*, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  is an asymptotically dominant equilibrium of the game.

### B. Payoff at the Equilibrium

In this section, we derive the payoff of the Neyman-Pearson game at the equilibrium of Theorem 2. To do this, we will assume an additive distortion function, i.e.,  $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n d(x_i, y_i)$ . In this case,  $d(\mathbf{x}, \mathbf{y})$  can be expressed as  $\sum_{ij} n_{\mathbf{x}\mathbf{y}}(i, j)d(i, j)$ , where  $n_{\mathbf{x}\mathbf{y}}(i, j) = n\hat{P}_{\mathbf{x}\mathbf{y}}(i, j)$  denotes the number of occurrences of the pair  $(i, j) \in \mathcal{A}^2$  in  $(\mathbf{x}, \mathbf{y})$ . Therefore, the distortion constraint regarding  $A_0$  can be rewritten as  $\sum_{(i,j) \in \mathcal{A}^2} \hat{P}_{\mathbf{x}\mathbf{y}}(i, j)d(i, j) \leq \Delta_0$ . A similar formulation holds for  $A_1$ .

Let us define

$$\tilde{\mathcal{D}}_{\Delta}^n(\hat{P}_{\mathbf{y}}, P) \triangleq \min_{\{\hat{P}_{\mathbf{x}} | E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta\}} \mathcal{D}(\hat{P}_{\mathbf{x}} \| P), \quad (18)$$

where  $E_{\mathbf{x}\mathbf{y}}$  denotes the empirical expectation, defined as

$$E_{\mathbf{x}\mathbf{y}}d(X, Y) = \sum_{(i,j) \in \mathcal{A}^2} \hat{P}_{\mathbf{x}\mathbf{y}}(i, j)d(i, j) \quad (19)$$

and the minimization is carried out for a given  $\hat{P}_{\mathbf{y}}$ . Accordingly, the strategy in (16) can be rewritten as

$$\Phi^*(\mathcal{H}_1 | \mathbf{y}) \triangleq \exp \left\{ -n \left[ \lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}} \| P_0) \right]_+ \right\}. \quad (20)$$

When  $n \rightarrow \infty$ ,  $\tilde{\mathcal{D}}_{\Delta}^n$  becomes<sup>4</sup>

$$\tilde{\mathcal{D}}_{\Delta}(P_Y, P) \triangleq \min_{\{P_{X|Y} : E_{X|Y}d(X, Y) \leq \Delta\}} \mathcal{D}(P_X \| P), \quad (21)$$

where  $E_{X|Y}$  denotes expectation w.r.t.  $P_{X|Y}$ .

Definition (21) can be stated for any PMF  $P_Y$  in the probability simplex in  $\mathbb{R}^{|\mathcal{A}|}$ . Note that the minimization problem in (21) has a unique solution as it is a convex program.

The function  $\tilde{\mathcal{D}}_{\Delta}$  will have an important role in the remaining part of the paper, especially in the characterization of the asymptotic behavior of the games. To draw a parallelism,  $\tilde{\mathcal{D}}_{\Delta}$  plays a role similar to that of the Kullback–Leibler divergence  $\mathcal{D}$  in classical detection theory for the non-adversarial case.

The basic properties of the functional  $\tilde{\mathcal{D}}_{\Delta}(P_Y, P)$  are the following: (i) it is continuous in  $P_Y$ ; (ii) it has convex level sets, i.e., the set  $\{P_Y : \tilde{\mathcal{D}}_{\Delta}(P_Y, P) \leq t\}$  is convex for every  $t \geq 0$ . Point (ii) is a

<sup>4</sup>Due to the density of rational numbers on the real line, the admissibility set in (18) is dense in that of (21); since the divergence functional is continuous, the sequence  $\{\tilde{\mathcal{D}}_{\Delta}^n(\hat{P}_{\mathbf{y}}, P)\}_{n \geq 1}$  tends to  $\tilde{\mathcal{D}}_{\Delta}(P_Y, P)$  as  $n \rightarrow \infty$ .

consequence of the following property, which will turn out to be useful for proving some of the results in the sequel (in particular, Theorem 3, 7 and also 8).

**Property 1.** *The function  $\tilde{\mathcal{D}}_{\Delta}(P_Y, P)$  is convex in  $P_Y$  for every fixed  $P$ .*

The proof follows from the convexity of the divergence functional (see Appendix I-B).

Using the above definitions, the equilibrium payoff is given by the following theorem:

**Theorem 3.** *Let the Neyman-Pearson detection game be as in Definition 1. Let  $(\Phi^*, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  be the equilibrium profile of Theorem 2. Then,<sup>5</sup>*

$$\begin{aligned} \varepsilon_{FN}(\lambda) &= - \lim_{n \rightarrow \infty} \frac{1}{n} \ln P_{FN}(\Phi^*, A_{\Delta_1}^*) \\ &= \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \end{aligned} \quad (22)$$

The proof, which appears in Appendix I-C, is based on Sanov's theorem [32], [33], by exploiting the compactness of the set  $\{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$ .

From Theorem 3 it follows that  $\varepsilon_{FN}(\lambda) = 0$  whenever there exists a PMF  $P_Y$  inside the set  $\{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$  with  $\Delta_1$ -limited expected distortion from  $P_1$ . When this condition does not hold,  $P_{FN}(\Phi^*, A_{\Delta_1}^*) \rightarrow 0$  exponentially rapidly.

For a partially-active attacker, the error exponent in (22) becomes

$$\varepsilon_{FN}(\lambda) = \min_{P_Y: \mathcal{D}(P_Y, P_0) \leq \lambda} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \quad (23)$$

It can be shown that the error exponent in (23) is the same as the error exponent of Theorem 2 in [14] (and Theorem 2 in [34]), where deterministic strategies are considered for both the defender and the attacker. Such equivalence can be explained as follows. As already pointed, the optimum defence strategy in (17) and the deterministic rule found in [14] are asymptotically equivalent (see the discussion immediately after Lemma 1). Concerning the attacker, even in the more general setup (with randomized strategies) considered here, an asymptotically optimum attack could be derived as in [14], that is, by considering the best response to the dominant defence strategy in [14]. Such attack consists of minimizing the divergence w.r.t.  $P_0$ , namely  $\mathcal{D}(\hat{P}_Y || P_0)$ , over all the admissible sequences  $\mathbf{y}$ , and then is deterministic. Therefore, concerning the partially active case, the asymptotic behavior of the game is equivalent to the one in [14]. The main difference between the setup in [14] and the more general one addressed in this paper relies

<sup>5</sup>We make explicit the dependence on the parameter  $\lambda$  in the notation of the error exponent, since this will turn to be useful in the sequel.



on the *kind* of game equilibrium, which is stronger here (namely, a *dominant* equilibrium) due to the existence of dominant strategies for both the defender and the attacker, rather than for the defender only.

When the distortion function  $d$  is a metric, we can state the following result, whose proof appears in Appendix I-D.

**Theorem 4.** *When the distortion function  $d$  is a metric, eq. (22) can be rephrased as*

$$\varepsilon_{FN}(\lambda) = \min_{P_Y: \mathcal{D}(P_Y \| P_0) \leq \lambda} \tilde{\mathcal{D}}_{\Delta_0 + \Delta_1}(P_Y, P_1). \quad (24)$$

Comparing eq. (24) with (23) is insightful for understanding the difference between the fully active and partially active cases. Specifically, the FN error exponents of both cases are the same when the distortion under  $\mathcal{H}_1$  in the partially-active case is  $\Delta_0 + \Delta_1$  (instead of  $\Delta_1$ ).

When  $d$  is not a metric, (24) is only an upper bound on  $\varepsilon_{FN}(\lambda)$ , as can be seen from the proof of Theorem 4. Accordingly, in the general case ( $d$  is not a metric), applying distortion levels  $\Delta_0$  and  $\Delta_1$  to sequences from, respectively,  $\mathcal{H}_0$  and  $\mathcal{H}_1$  (in the fully active setup) is more favorable to the attacker with respect to applying a distortion  $\Delta_0 + \Delta_1$  to sequences from  $\mathcal{H}_0$  only (in the partially active setup).

## V. THE BAYESIAN DETECTION GAME

In this section, we study the Bayesian game (Definition 2). In contrast to the Neyman–Pearson game, in the Bayesian game, the optimal defence strategy is found by assuming that the strategy played by the attacker, namely the optimum pair of channels  $(A_0^*, A_1^*)$  of Theorem 1, is known to the defender, that is, by exploiting the rationalizability argument (see Section II-A). Accordingly, the resulting optimum strategy is not dominant, and so, the associated equilibrium is weaker compared to that of the Neyman–Pearson game.

### A. Optimum Defence and Game Equilibrium

Since the payoff in (3) is a special case of (5) with  $\gamma = 1$  and  $\beta = e^{an}$ , for any defence strategy  $\Phi \in \mathcal{S}_D$ , the asymptotically optimum attack channels under  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are given by Theorem 1, and correspond to the pair  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$ . Then, we can determine the best defence strategy by assuming that the attacker will play  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  and evaluating the best response of the defender to this pair of channels.

Our solution for the Bayesian detection game is given in the following theorem, whose proof appears in Appendix II-A.

**Theorem 5.** Consider the Bayesian detection game of Definition 2. Let  $Q_0^*(\mathbf{y})$  and  $Q_1^*(\mathbf{y})$  be the probability distributions induced by channels  $A_{\Delta_0}^*$  and  $A_{\Delta_1}^*$ , respectively.

Then,<sup>6</sup>

$$\Phi^\#(\mathcal{H}_1|\mathbf{y}) = U\left(\frac{1}{n} \log \frac{Q_1^*(\mathbf{y})}{Q_0^*(\mathbf{y})} - a\right) \quad (25)$$

is the optimum defence strategy.

If, in addition, the distortion measure is additive, the defence strategy

$$\Phi^\dagger(\mathcal{H}_1|\mathbf{y}) = U\left(\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}\mathbf{y}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}\mathbf{y}, P_1) - a\right) \quad (26)$$

is asymptotically optimum.

It is useful to provide the asymptotically optimum strategy,  $\Phi^\dagger$ , in addition to the optimal one,  $\Phi^\#$ , for the following reason: while  $\Phi^\#$  requires the non-trivial computation of the two probabilities  $Q_1(\mathbf{y})$  and  $Q_0(\mathbf{y})$ , the strategy  $\Phi^\dagger$ , which leads to the same payoff asymptotically, is easier to implement because of its single-letter form.

We now state the following theorem.

**Theorem 6.** Consider the Bayesian game of Definition 2. Let  $(A_{\Delta_0}^*, A_{\Delta_1}^*)$  be the attack strategy of Theorem 1 and let  $\Phi^\#$  and  $\Phi^\dagger$  be the defence strategies defined, respectively, in (25) and (26). The profiles  $(\Phi^\#, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  and  $(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  are asymptotic rationalizable equilibria of the game.

The analysis in this section can be easily generalized to any payoff function defined as in (5), i.e., for any  $\gamma, \beta \geq 0$ .

Finally, we observe that, the fact that the equilibrium found in the Bayesian case (namely, a rationalizable equilibrium) is weaker with respect to the equilibrium derived for the Neyman–Pearson game (namely, a dominant equilibrium) is a consequence of the fact that the Bayesian game is defined in a less restrictive manner than the Neyman–Pearson game. This is due to the conservative approach adopted in the latter: while in the Bayesian game the defender cares about both FP and FN probabilities and their tradeoff, in the Neyman–Pearson game the defender does not care about the value of the FP probability provided that its exponent is larger than  $\lambda$ , which is automatically guaranteed by restricting the set of strategies. This restriction simplifies the game so that a dominant strategy can be found for the restricted game.

<sup>6</sup>We remind that  $U(\cdot)$  denotes the Heaviside step function.

### B. Equilibrium Payoff

We now derive the equilibrium payoff of the Bayesian game. As in the Neyman–Pearson game, we assume an additive distortion measure. For simplicity, we focus on the asymptotically optimum defence strategy  $\Phi^\dagger$ . We have the following theorem.

**Theorem 7.** *Let the Bayesian detection game be as in Definition 2. Let  $(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  be the equilibrium profile of Theorem 6. The asymptotic exponential rate of the equilibrium Bayes payoff  $u$  is given by*

$$-\lim_{n \rightarrow \infty} \frac{1}{n} \ln \left( u(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*)) \right) = \min_{P_Y} \left( \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), (\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - a) \right\} \right). \quad (27)$$

The proof appears in Appendix II-B.

According to Theorem 7, the asymptotic exponent of  $u$  is zero if there exists a PMF  $P_Y^*$  with  $\Delta_1$ -limited expected distortion from  $P_1$  such that  $\tilde{\mathcal{D}}_{\Delta_0}(P_Y^*, P_0) \leq a$ . Therefore, when we focus on the case of zero asymptotic exponent of the payoff, the parameter  $a$  plays a role similar to  $\lambda$  in the Neyman–Pearson game. By further inspecting the exponent expressions of Theorems 7 and 3, we observe that, when  $a = \lambda$ , the exponent in (27) is smaller than or equal to the one in (22), where equality holds only when both (27) and (22) vanish. However, comparing these two cases in the general case is difficult because of the different definition of the payoff functions and, in particular, the different role taken by the parameters  $\lambda$  and  $a$ . In the Neyman–Pearson game, in fact, the payoff corresponds to the FN probability and is not affected by the value of the FP probability, provided that its exponent is larger than  $\lambda$ ; in this way, the ratio between FP and FN error exponent at the equilibrium is generally smaller than  $\lambda$  (a part for the case in which the asymptotic exponent of the payoff is zero). In the Bayesian case, the payoff is a weighted combination of the two types of errors and then the term with the largest exponent is the dominating term, namely, the one which determines the asymptotic behavior; in this case, the parameter  $a$  determines the exact tradeoff between the FP and FN exponent in the equilibrium payoff.

## VI. SOURCE DISTINGUISHABILITY

In this section, we investigate the performance of the Neyman–Pearson and Bayesian games as functions of  $\lambda$  and  $a$  respectively. From the expressions of the equilibrium payoff exponents, it is clear that the Neyman–Pearson and the Bayesian payoffs increase as  $\lambda$  and  $a$  decrease, respectively. In particular, by setting  $\lambda = 0$  and  $a = 0$ , we obtain the largest achievable payoffs of both games which correspond to

the best achievable performance for the defender. Therefore, we say that two sources are *distinguishable* under the Neyman–Pearson (resp. Bayesian) setting, if there exists a value of  $\lambda$  (resp.  $\alpha$ ) such that the FP and FN exponents at the equilibrium of the game are simultaneously strictly positive. When such a condition does not hold, we say that the sources are *indistinguishable*. Specifically, in this section, we characterize, under both the Neyman–Pearson and the Bayesian settings, the *indistinguishability region*, defined as the set of the alternative sources that cannot be distinguished from a given source  $P_0$ , given the attack distortion levels  $\Delta_0$  and  $\Delta_1$ . Although each game has a different asymptotic behavior, we will see that the indistinguishability regions in the Neyman–Pearson and the Bayesian settings are the same. The study of the distinguishability between the sources under adversarial conditions, performed in this section, in a way extends the Chernoff-Stein lemma [25] to the adversarial setup (see [34]).

We start by proving the following result for the Neyman–Pearson game.

**Theorem 8.** *Given two memoryless sources  $P_0$  and  $P_1$  and distortion levels  $\Delta_0$  and  $\Delta_1$ , the maximum achievable FN exponent for the Neyman–Pearson game is:*

$$\lim_{\lambda \rightarrow 0} \varepsilon_{FN}(\lambda) = \varepsilon_{FN}(0) = \min_{\{P_{Y|X}: E_{XY}d(X,Y) \leq \Delta_0, (P_{XY})_X = P_0\}} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \quad (28)$$

where  $\varepsilon_{FN}(\lambda)$  is as in Theorem 3.

The theorem is an immediate consequence of the continuity of  $\varepsilon_{FN}(\lambda)$  as  $\lambda \rightarrow 0^+$ , which follows by the continuity of  $\tilde{\mathcal{D}}_{\Delta}$  with respect to  $P_Y$  and the density of the set  $\{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$  in  $\{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) = 0\}$  as  $\lambda \rightarrow 0^+$ <sup>7</sup>.

We notice that, if  $\Delta_0 = \Delta_1 = 0$ , there is only an admissible point in the set in (28), for which  $P_Y = P_0$ ; then,  $\varepsilon_{FN}(0) = \mathcal{D}(P_0||P_1)$ , which corresponds to the best achievable FN exponent known from the classical literature for the non-adversarial case (Stein lemma [25], Theorem 11.8.3).

Regarding the Bayesian setting, we have the following theorem, the proof of which appears in Appendix III-A.

**Theorem 9.** *Given two memoryless sources  $P_0$  and  $P_1$  and distortion levels  $\Delta_0$  and  $\Delta_1$ , the maximum achievable exponent of the equilibrium Bayes payoff is*

$$-\lim_{a \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left( u(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*)) \right) = \min_{P_Y} \left( \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \right\} \right), \quad (29)$$

<sup>7</sup>It holds true from Property 1.

where the inner limit at the left hand side is as defined in Theorem 7.

Since  $\tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1)$ , and similarly  $\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0)$ , are convex functions of  $P_Y$ , and reach their minimum in  $P_1$ , resp.  $P_0$ ,<sup>8</sup> the minimum over  $P_Y$  of the maximum between these quantities (right-hand side of (29)) is attained when  $\tilde{\mathcal{D}}_{\Delta_1}(P_Y^*, P_1) = \tilde{\mathcal{D}}_{\Delta_0}(P_Y^*, P_0)$ , for some PMF  $P_Y^*$ . This resembles the best achievable exponent in the Bayesian probability of error for the non-adversarial case, which is attained when  $\mathcal{D}(P_Y^* \| P_0) = \mathcal{D}(P_Y^* \| P_1)$  for some  $P_Y^*$  (see [25], Theorem 11.9.1). In that case, from the expression of the divergence function, such  $P_Y^*$  is found in a closed form and the resulting exponent is equivalent to the Chernoff information (see Section 11.9 in [25]).

From Theorem 8 and 9, it follows that there is no positive  $\lambda$ , res.  $a$ , for which the asymptotic exponent of the equilibrium payoff is strictly positive, if there exists a PMF  $P_Y$  such that the following conditions are both satisfied:

$$\begin{cases} \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) = 0 \\ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) = 0. \end{cases} \quad (30)$$

In this case, then,  $P_0$  and  $P_1$  are indistinguishable under both the Neyman–Pearson and the Bayesian settings. We observe that the condition  $\tilde{\mathcal{D}}_{\Delta}(P_Y, P_X) = 0$  is equivalent to the following:<sup>9</sup>

$$\min_{Q_{XY}: \begin{matrix} (Q_{XY})_X = P_X \\ (Q_{XY})_Y = P_Y \end{matrix}} E_{XY} d(X, Y) \leq \Delta, \quad (31)$$

where the expectation  $E_{XY}$  is w.r.t  $Q_{XY}$ . In computer vision applications, the left-hand side of (31) is known as the *Earth Mover Distance* (EMD) between  $P_X$  and  $P_Y$ , which is denoted by  $EMD_d(P_X, P_Y)$  (or, by symmetry,  $EMD_d(P_Y, P_X)$ ) [35]. It is also known as the  $\rho$ -bar distortion measure [36].

A brief comment concerning the analogy between the minimization in (31) and *optimal transport theory* is worth. The minimization problem in (31) is known in the Operations Research literature as *Hitchcock Transportation Problem* (TP) [37]. Referring to the original Monge formulation of this problem [38],  $P_X$  and  $P_Y$  can be interpreted as two different ways of piling up a certain amount of soil; then,  $P_{XY}(x, y)$  denotes the quantity of soil shipped from location (source)  $x$  in  $P_X$  to location (sink)  $y$  in  $P_Y$  and  $d(x, y)$  is the cost for shipping a unitary amount of soil from  $x$  to  $y$ . In transport theory terminology,  $P_{XY}$  is referred to as *transportation map*. According to this perspective, evaluating the *EMD* corresponds to finding the minimal transportation cost of moving a pile of soil into the other. Further insights on this

<sup>8</sup>The fact that  $\tilde{\mathcal{D}}_{\Delta_0}$  ( $\tilde{\mathcal{D}}_{\Delta_1}$ ) is 0 in a  $\Delta_0$ -limited ( $\Delta_1$ -limited) neighborhood of  $P_0$  ( $P_1$ ), and not just in  $P_0$  ( $P_1$ ), does not affect the argument.

<sup>9</sup>For ease of notation, given a joint PMF  $Q_{XY}$  with marginal PMFs  $P_X$  and  $P_Y$ , we use notation  $(Q_{XY})_Y = P_Y$  (res.  $(Q_{XY})_X = P_X$ ) as short for  $\sum_x Q_{XY}(x, y) = P_Y(y)$ ,  $\forall y \in \mathcal{A}$  (res.  $\sum_y Q_{XY}(x, y) = P_X(x)$ ,  $\forall x \in \mathcal{A}$ ).

parallel can be found in [34].

We summarize our findings in the following corollary, which characterizes the conditions for distinguishability under both the Neyman–Pearson and the Bayesian setting.

**Corollary 1** (Corollary to Theorems 8 and 9). *Given a memoryless source  $P_0$  and distortion levels  $\Delta_0$  and  $\Delta_1$ , the set of the PMFs that cannot be distinguished from  $P_0$  in both the Neyman–Pearson and Bayesian settings is given by*

$$\Gamma = \left\{ P : \min_{P_Y: \text{EMD}_d(P_Y, P_0) \leq \Delta_0} \text{EMD}_d(P_Y, P) \leq \Delta_1 \right\}. \quad (32)$$

Set  $\Gamma$  is the indistinguishability region. By definition (see the beginning of this section), the PMFs inside  $\Gamma$  are those for which, as a consequence of the attack, the FP and FN probabilities cannot go to zero simultaneously with strictly positive exponents. Clearly, if  $\Delta_0 = \Delta_1 = 0$ , that is, in the non-adversarial case,  $\Gamma = \{P_0\}$ , as any two distinct sources are always distinguishable.

When  $d$  is a metric, for a given  $P \in \Gamma$ , the computation of the optimum  $P_Y$  can be traced back to the computation of the *EMD* between  $P_0$  and  $P$ , as stated by the following corollary, whose proof appears in Appendix III-B.

**Corollary 2** (Corollary to Theorems 8 and 9). *When  $d$  is a metric, given the source  $P_0$  and distortion levels  $\Delta_0$  and  $\Delta_1$ , for any fixed  $P$ , the minimum in (32) is achieved when*

$$P_Y = \alpha P_0 + (1 - \alpha)P, \quad \alpha = 1 - \frac{\Delta_0}{\text{EMD}(P_0, P)}. \quad (33)$$

*Then, the set of PMFs that cannot be distinguished from  $P_0$  in the Neyman–Pearson and Bayesian setting is given by*

$$\Gamma = \{P : \text{EMD}_d(P_0, P) \leq \Delta_0 + \Delta_1\}. \quad (34)$$

According to Corollary 2, when  $d$  is a metric, the performance of the game depends only on the sum of distortions,  $\Delta_0 + \Delta_1$ , and it is immaterial how this amount is distributed between the two hypotheses.

In the general case ( $d$  not a metric), the condition on the *EMD* stated in (34) is sufficient in order for  $P_0$  and  $P$  be indistinguishable, that is  $\Gamma \supseteq \{P : \text{EMD}_d(P_0, P) \leq \Delta_0 + \Delta_1\}$  (see discussion in Appendix III-B, at the end of the proof of Corollary 2). Furthermore, in the case of an  $L_p^p$  distortion function ( $p \geq 1$ ), i.e.,  $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|^p$ , we have the following corollary.

**Corollary 3** (Corollary to Theorems 8 and 9). *When  $d$  is the  $L_p^p$  distortion function, for some  $p \geq 1$ , the set  $\Gamma$  can be bounded as follows*

$$\Gamma \subseteq \{P : \text{EMD}_{L_p^p}(P_0, P) \leq (\Delta_0^{1/p} + \Delta_1^{1/p})^p\}. \quad (35)$$

Corollary 3 can be proven by exploiting the Hölder inequality [39] (see Appendix III-C).

## VII. CONCLUSIONS

We considered the problem of binary hypothesis testing when an attacker is active under both hypotheses, and then an attack is carried out aiming at both false negative and false positive errors. By modeling the defender-attacker interaction as a game, we defined and solved two different detection games: the Neyman–Pearson and the Bayesian game. This paper extends the analysis in [14] [14], where the attacker is active under the alternative hypothesis only. Another aspect of greater generality is that here both players are allowed to use randomized strategies. By relying on the method of types, the main result of this paper is the existence of an attack strategy which is both *dominant* and *universal*, that is, optimal regardless of the statistics of the sources. The optimum attack strategy is also independent of the underlying hypothesis, namely *fully-universal*, when the distortion introduced by the attacker in the two cases is the same. From the analysis of the asymptotic behavior of the equilibrium payoff we are able to establish conditions under which the sources can be reliably distinguished in the fully-active adversarial setup. The theory developed permits to assess the security of the detection in adversarial setting and give insights on how the detector should be designed in such a way to make the attack hard.

Among the possible directions for future work, we mention the extension to multiple hypothesis testing. Another interesting direction is the extension to continuous alphabets, which calls for an extension of the method of types to this case, or to more realistic models of finite alphabet sources, still amenable to analysis, like Markov sources. As mentioned in the introduction, it would be also relevant to overcome the limitation to first order statistics, by extending the analysis to higher order statistics and getting equilibria in a similar fashion. Finally, we mention the case of unknown sources, where the sources are estimated from training data, possibly corrupted by the attacker. In this scenario, the detection game has been studied for a partially active case, with both uncorrupted and corrupted training data [16], [17]. The extension of such analyses to the fully active scenario considered in this paper is a further interesting direction for future research.

## ACKNOWLEDGMENT

We thank Alessandro Agnetis of the University of Siena, for the useful discussions on optimization concepts underlying the computation of the *EMD*.

## APPENDIX I

## NEYMAN–PEARSON DETECTION GAME

This appendix contains the proofs of the results in Section IV.

## A. Proof of Lemma 1

Whenever existent, the dominant defence strategy can be obtained by solving:

$$\min_{\Phi \in \mathcal{S}_D} P_{\text{FN}}(\Phi, A_1), \quad (\text{I.1})$$

for any attack channel  $A_1$ . Below, we first show that  $P_{\text{FN}}(\Phi^*, A_1) \leq P_{\text{FN}}(\Phi, A_1)$  for every  $\Phi \in \mathcal{S}_D$  and for every  $A_1$ , that is,  $\Phi^*$  is asymptotically dominant. Then, by proving that  $\max_{A \in \mathcal{C}_{\Delta_0}} P_{\text{FP}}(\Phi^*, A)$  fulfills the FP constraint, we show that  $\Phi^*$  is also admissible. Therefore, we can conclude that  $\Phi^*(\cdot|\mathbf{y})$  asymptotically solves (I.1). Exploiting the memorylessness of  $P_0$  and the permutation invariance of  $\Phi(\mathcal{H}_1|\mathbf{y})$  and  $d(\mathbf{x}, \mathbf{y})$ , for every  $\mathbf{y}' \in \mathcal{A}^n$  we have,

$$\begin{aligned} e^{-\lambda n} &\geq \max_A \sum_{\mathbf{x}, \mathbf{y}} P_0(\mathbf{x}) A(\mathbf{y}|\mathbf{x}) \Phi(\mathcal{H}_1|\mathbf{y}) \\ &\geq \sum_{\mathbf{y}} \left( \sum_{\mathbf{x}} P_0(\mathbf{x}) A_{\Delta_0}^*(\mathbf{y}|\mathbf{x}) \right) \Phi(\mathcal{H}_1|\mathbf{y}) \\ &= \sum_{\mathbf{y}} \left( \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} P_0(\mathbf{x}) \cdot \frac{c_n(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \right) \Phi(\mathcal{H}_1|\mathbf{y}) \\ &\geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\mathbf{y}} \left( \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \frac{P_0(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \right) \Phi(\mathcal{H}_1|\mathbf{y}) \\ &\stackrel{(a)}{\geq} (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} |\mathcal{T}(\mathbf{y}')| \left( \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}') \leq n\Delta_0} |\mathcal{T}(\mathbf{x}|\mathbf{y}')| \cdot \frac{P_0(\mathbf{x})}{|\mathcal{T}(\mathbf{y}'|\mathbf{x})|} \right) \Phi(\mathcal{H}_1|\mathbf{y}') \\ &\stackrel{(b)}{=} (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \Phi(\mathcal{H}_1|\mathbf{y}') \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}') \leq n\Delta_0} P_0(\mathbf{x}) \cdot |\mathcal{T}(\mathbf{x})| \\ &\geq \Phi(\mathcal{H}_1|\mathbf{y}') \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}') \leq n\Delta_0} \frac{e^{-n\mathcal{D}(\hat{P}\mathbf{x} \| P_0)}}{(n+1)^{|\mathcal{A}|^2 \cdot (|\mathcal{A}|-1)}} \\ &= \Phi(\mathcal{H}_1|\mathbf{y}') \frac{\exp \left\{ -n \min_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}') \leq n\Delta_0} \mathcal{D}(\hat{P}\mathbf{x} \| P_0) \right\}}{(n+1)^{|\mathcal{A}|^2 \cdot (|\mathcal{A}|-1)}}, \quad (\text{I.2}) \end{aligned}$$



where (a) is due to the permutation invariance of the distortion function  $d$  and (b) is due to the identity  $|\mathcal{T}(\mathbf{x})| \cdot |\mathcal{T}(\mathbf{y}|\mathbf{x})| \equiv |\mathcal{T}(\mathbf{y})| \cdot |\mathcal{T}(\mathbf{x}|\mathbf{y})| \equiv |\mathcal{T}(\mathbf{x}, \mathbf{y})|$ .

It now follows that

$$\Phi(\mathcal{H}_1|\mathbf{y}) \leq \exp \left\{ -n \left[ \lambda - \min_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\mathbf{x}} \| P_0) \right] \right\}. \quad (\text{I.3})$$

Since  $\Phi(\mathcal{H}_1|\mathbf{y})$  is a probability,

$$\begin{aligned} \Phi(\mathcal{H}_1|\mathbf{y}) &\leq \min \left\{ 1, \exp \left[ -n \left( \lambda - \min_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\mathbf{x}} \| P_0) \right) \right] \right\} \\ &= \Phi^*(\mathcal{H}_1|\mathbf{y}). \end{aligned} \quad (\text{I.4})$$

Consequently,  $\Phi^*(\mathcal{H}_0|\mathbf{y}) \leq \Phi(\mathcal{H}_0|\mathbf{y})$  for every  $\mathbf{y}$ , and so,  $P_{\text{FN}}(\Phi^*, A_1) \leq P_{\text{FN}}(\Phi, A_1)$  for every  $A_1$ . For convenience, let us denote

$$k_n(\mathbf{y}) = \lambda - \min_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\mathbf{x}} \| P_0),$$

so that  $\Phi^*(\mathcal{H}_1|\mathbf{y}) = \min\{1, e^{-n \cdot k_n(\mathbf{y})}\}$ . We now show that  $\Phi^*(\mathcal{H}_1|\mathbf{y})$  satisfies the FP constraint, up to a

polynomial term in  $n$ , i.e., it satisfies the constraint asymptotically.

$$\begin{aligned}
 \max_{A \in \mathcal{C}_{\Delta_0}} P_{\text{FP}}(\Phi^*, A) &\leq (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} P_{\text{FP}}(\Phi^*, A^*) \\
 &= (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\mathbf{x}, \mathbf{y}} P_0(\mathbf{x}) A_{\Delta_0}^*(\mathbf{y}|\mathbf{x}) \Phi^*(\mathcal{H}_1|\mathbf{y}) \\
 &= (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{(\mathbf{x}, \mathbf{y}): d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} P_0(\mathbf{x}) \cdot \frac{c_n(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \cdot \Phi^*(\mathcal{H}_1|\mathbf{y}) \\
 &\leq (n+1)^{|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{(\mathbf{x}, \mathbf{y}): d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \frac{P_0(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \cdot \Phi^*(\mathcal{H}_1|\mathbf{y}) \\
 &\leq (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\mathbf{y}} \left( \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} |\mathcal{T}(\mathbf{x}|\mathbf{y})| \cdot \frac{P_0(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \right) \Phi^*(\mathcal{H}_1|\mathbf{y}) \\
 &= (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \left( \sum_{\hat{P}_{\mathbf{y}}: k_n(\mathbf{y}) \geq 0} e^{-nk_n(\mathbf{y})} \left( \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} |\mathcal{T}(\mathbf{x})| \cdot P_0(\mathbf{x}) \right) + \right. \\
 &\quad \left. + \sum_{\hat{P}_{\mathbf{y}}: k_n(\mathbf{y}) < 0} \left( \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} |\mathcal{T}(\mathbf{x})| \cdot P_0(\mathbf{x}) \right) \right) \\
 &\leq (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \left( \sum_{\hat{P}_{\mathbf{y}}: k_n(\mathbf{y}) \geq 0} e^{-n\lambda} + \right. \\
 &\quad \left. + \sum_{\hat{P}_{\mathbf{y}}: k_n(\mathbf{y}) < 0} \exp \left\{ -n \min_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\mathbf{x}} \| P_0) \right\} \right) \\
 &\leq (n+1)^{(|\mathcal{A}|^2 + 2|\mathcal{A}|) \cdot (|\mathcal{A}|-1) + |\mathcal{A}|} e^{-n\lambda}. \tag{I.5}
 \end{aligned}$$

### B. Proof of Property 1

We next prove that for any two PMFs  $P_{Y_1}$  and  $P_{Y_2}$  and any  $\lambda \in (0, 1)$ ,

$$\tilde{\mathcal{D}}_{\Delta}(\lambda P_{Y_1} + (1-\lambda)P_{Y_2}, P) \leq \lambda \tilde{\mathcal{D}}_{\Delta}(P_{Y_1}, P) + (1-\lambda) \tilde{\mathcal{D}}_{\Delta}(P_{Y_2}, P). \tag{I.6}$$

Let us rewrite  $\tilde{\mathcal{D}}_{\Delta}$  in (21) by expressing the minimization in terms of the joint PMF  $P_{XY}$ :

$$\tilde{\mathcal{D}}_{\Delta}(P_Y, P) \triangleq \min_{\{Q_{XY}: E_{XY} d(X, Y) \leq \Delta, (Q_{XY})_Y = P_Y\}} \mathcal{D}((Q_{XY})_X \| P), \tag{I.7}$$

where we used  $(Q_{XY})_Y = P_Y$  as short for  $\sum_x Q_{XY}(x, y) = P_Y(y)$ ,  $\forall y$ , and we made explicit the dependence of  $\mathcal{D}(P_X \| P)$  on  $Q_{XY}$ . Accordingly:

$$\tilde{\mathcal{D}}_{\Delta}(\lambda P_{Y_1} + (1-\lambda)P_{Y_2}, P) = \min_{\{Q_{XY}: E_{XY} d(X, Y) \leq \Delta, (Q_{XY})_Y = \lambda P_{Y_1} + (1-\lambda)P_{Y_2}\}} \mathcal{D}((Q_{XY})_X \| P). \tag{I.8}$$

We find convenient to rewrite the right-hand side of (I.8) by minimizing over pairs of PMFs  $(Q'_{XY}, Q''_{XY})$  and considering the convex combination of these PMFs with weights  $\lambda$  and  $(1 - \lambda)$ , in place of  $Q_{XY}$ ; hence

$$\tilde{\mathcal{D}}_{\Delta}(\lambda P_{Y_1} + (1 - \lambda)P_{Y_2}, P) = \min_{(Q'_{XY}, Q''_{XY}) \in \mathcal{H}} \mathcal{D}(\lambda(Q'_{XY})_X + (1 - \lambda)(Q''_{XY})_X \| P), \quad (\text{I.9})$$

where

$$\mathcal{H} = \{(Q'_{XY}, Q''_{XY}) : \lambda(Q'_{XY})_Y + (1 - \lambda)(Q''_{XY})_Y = \lambda P_{Y_1} + (1 - \lambda)P_{Y_2}, \\ \lambda E'_{XY} d(X, Y) + (1 - \lambda)E''_{XY} d(X, Y) \leq \Delta\}. \quad (\text{I.10})$$

Let

$$\mathcal{H}' = \{Q'_{XY} : E'_{XY} d(X, Y) \leq \Delta, (Q'_{XY})_Y = P_{Y_1}\} \times \{Q''_{XY} : E''_{XY} d(X, Y) \leq \Delta, (Q''_{XY})_Y = P_{Y_2}\}; \quad (\text{I.11})$$

then,  $\mathcal{H}' \subset \mathcal{H}$ , where the set  $\mathcal{H}'$  is separable in  $Q'_{XY}$  and  $Q''_{XY}$ . Accordingly, (I.9)-(I.10) can be upper bounded by

$$\min_{Q'_{XY} : E'_{XY} d(X, Y) \leq \Delta, (Q'_{XY})_Y = P_{Y_1}} \min_{Q''_{XY} : E''_{XY} d(X, Y) \leq \Delta, (Q''_{XY})_Y = P_{Y_2}} \mathcal{D}(\lambda(Q'_{XY})_X + (1 - \lambda)(Q''_{XY})_X \| P). \quad (\text{I.12})$$

By the convexity of  $\mathcal{D}((Q_{XY})_X \| P)$  with respect to  $Q_{XY}$ <sup>10</sup>, it follows that

$$\mathcal{D}(\lambda(Q'_{XY})_X + (1 - \lambda)(Q''_{XY})_X \| P) \leq \lambda \mathcal{D}((Q'_{XY})_X \| P) + (1 - \lambda) \mathcal{D}((Q''_{XY})_X \| P). \quad (\text{I.13})$$

Note that the above relation is not strict since it might be that  $(Q'_{XY})_X = (Q''_{XY})_X = P$ . Then, an upper bound for  $\tilde{\mathcal{D}}_{\Delta}(\lambda P_{Y_1} + (1 - \lambda)P_{Y_2}, P)$  is given by

$$\min_{Q'_{XY} : \sum_x Q'_{XY} = P_{Y_1}, E'_{XY} d(X, Y) \leq \Delta} \lambda \mathcal{D}((Q'_{XY})_X \| P) + \min_{Q''_{XY} : \sum_x Q''_{XY} = P_{Y_2}, E''_{XY} d(X, Y) \leq \Delta} (1 - \lambda) \mathcal{D}((Q''_{XY})_X \| P), \quad (\text{I.14})$$

thus proving (I.6).

<sup>10</sup>This is a consequence of the fact that the divergence function is convex in its arguments and the operation  $(\cdot)_X$  is linear (see Theorem 2.7.2 in [25]).

### C. Proof of Theorem 3

We start by proving the upper bound for the FN probability:

$$\begin{aligned}
P_{\text{FN}}(\Phi^*, A_{\Delta_1}^*) &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\mathbf{x}) A_{\Delta_1}^*(\mathbf{y}|\mathbf{x}) \Phi^*(\mathcal{H}_0|\mathbf{y}) \\
&= \sum_{\mathbf{y}} \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_1} P_1(\mathbf{x}) \frac{c_n(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
&\leq \sum_{\mathbf{y}} \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_1} \frac{P_1(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
&= \sum_{\mathbf{y}} \sum_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} |\mathcal{T}(\hat{P}_{\mathbf{x}}|\mathbf{y})| \frac{e^{-n[\hat{H}\mathbf{x}(X) + \mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)]}}{|\mathcal{T}(\hat{P}_{\mathbf{y}}|\mathbf{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
&= \sum_{\hat{P}_{\mathbf{y}}} \sum_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} |\mathcal{T}(\hat{P}_{\mathbf{x}})| e^{-n[\hat{H}\mathbf{x}(X) + \mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)]} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
&= \sum_{\hat{P}_{\mathbf{y}}} \sum_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
&= \sum_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) < \lambda} \sum_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0))}\right) \\
&\leq \sum_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) < \lambda} \sum_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \\
&\leq (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp \left\{ -n \min_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) < \lambda} \left[ \min_{\hat{P}_{\mathbf{x}}|\mathbf{y}: E_{\mathbf{x}\mathbf{y}}d(X, Y) \leq \Delta_1} \mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1) \right] \right\} \\
&\leq (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp \left\{ -n \inf_{P_{\mathbf{Y}}: \tilde{\mathcal{D}}_{\Delta_0}(P_{\mathbf{Y}}, P_0) < \lambda} \left[ \min_{P_{\mathbf{X}|\mathbf{Y}}: E_{\mathbf{X}\mathbf{Y}}d(X, Y) \leq \Delta_1} \mathcal{D}(P_{\mathbf{X}}\|P_1) \right] \right\} \\
&\leq (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp \left\{ -n \min_{P_{\mathbf{Y}}: \tilde{\mathcal{D}}_{\Delta_0}(P_{\mathbf{Y}}, P_0) \leq \lambda} \left[ \min_{P_{\mathbf{X}|\mathbf{Y}}: E_{\mathbf{X}\mathbf{Y}}d(X, Y) \leq \Delta_1} \mathcal{D}(P_{\mathbf{X}}\|P_1) \right] \right\}. \quad (\text{I.15})
\end{aligned}$$

Then:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \ln P_{\text{FN}}(\Phi^*, A_{\Delta_1}^*) \leq - \min_{P_{\mathbf{Y}}: \tilde{\mathcal{D}}_{\Delta_0}(P_{\mathbf{Y}}, P_0) \leq \lambda} \left[ \min_{P_{\mathbf{X}|\mathbf{Y}}: E_{\mathbf{X}\mathbf{Y}}d(X, Y) \leq \Delta_1} \mathcal{D}(P_{\mathbf{X}}\|P_1) \right]. \quad (\text{I.16})$$

We now move on to the lower bound.

$$\begin{aligned}
 P_{\text{FN}}(\Phi^*, A_{\Delta_1}^*) &= \sum_{\mathbf{x}, \mathbf{y}} P_1(\mathbf{x}) A_{\Delta_1}^*(\mathbf{y}|\mathbf{x}) \Phi^*(\mathcal{H}_1|\mathbf{y}) \\
 &\geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\mathbf{y}} \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_1} \frac{P_1(\mathbf{x})}{|\mathcal{T}(\mathbf{y}|\mathbf{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0)]_+}\right) \\
 &= (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) < \lambda} \sum_{\hat{P}_{\mathbf{x}|\mathbf{y}}: E_{\mathbf{x}\mathbf{y}} d(X, Y) \leq \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0))}\right) \\
 &\geq (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0))}\right), \tag{I.17}
 \end{aligned}$$

where, for a fixed  $n$ ,  $\hat{P}_{\mathbf{y}}$  is a PMF that satisfies  $\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) \leq \lambda - (\ln n)/n$  and  $\hat{P}_{\mathbf{x}|\mathbf{y}}$  is such that the distortion constraint is satisfied. Since the set of rational PMFs is dense in the probability simplex, two such sequences can be chosen in such a way that  $(\hat{P}_{\mathbf{y}}, \hat{P}_{\mathbf{x}|\mathbf{y}}) \rightarrow (P_Y^*, P_{X|Y}^*)$ ,<sup>11</sup> where

$$(P_Y^*, P_{X|Y}^*) = \arg \min_{(P_Y, P_{X|Y})} \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda} \left[ \min_{P_{X|Y}: E_{XY} d(X, Y) \leq \Delta_1} \mathcal{D}(P_X\|P_1) \right]. \tag{I.18}$$

Therefore, we can assert that:

$$\begin{aligned}
 \liminf_{n \rightarrow \infty} \frac{1}{n} \ln P_{\text{FN}}(\Phi^*, A_{\Delta_1}^*) &\geq \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left[ e^{-n\mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0))}\right) \right] \\
 &= - \lim_{n \rightarrow \infty} \mathcal{D}(\hat{P}_{\mathbf{x}}\|P_1) \\
 &= - \mathcal{D}(P_X^*\|P_1) \\
 &= - \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y\|P_0) \leq \lambda} \left[ \min_{P_{X|Y}: E_{XY} d(X, Y) \leq \Delta_1} \mathcal{D}(P_X\|P_1) \right]. \tag{I.19}
 \end{aligned}$$

By combining the upper and lower bounds, we conclude that  $\limsup$  and  $\liminf$  coincide. Therefore the limit of the sequence  $1/n \ln P_{\text{FN}}$  exists and the theorem is proven.

#### D. Proof of Theorem 4

First, observe that, by exploiting the definition of  $\tilde{\mathcal{D}}_{\Delta}$ , (22) can be rewritten as

$$\begin{aligned}
 \varepsilon_{\text{FN}} &= \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda} \left( \min_{P_{X|Y}: E_{XY} d(X, Y) \leq \Delta_1} \mathcal{D}(P_X\|P_1) \right) \\
 &= \min_{P_Z: \mathcal{D}(P_Z\|P_0) \leq \lambda} \min_{P_{Y|Z}: E_{YZ} d(Y, Z) \leq \Delta_0} \left( \min_{P_{X|Y}: E_{XY} d(X, Y) \leq \Delta_1} \mathcal{D}(P_X\|P_1) \right). \tag{I.20}
 \end{aligned}$$

<sup>11</sup>We are implicitly exploiting the fact that set  $\{\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) < \lambda\}$  is dense in  $\{\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$ , for every  $\lambda > 0$ , which holds true from Property 1.

To prove the theorem, we now show that (I.20) can be simplified as follows:

$$\varepsilon_{\text{FN}} = \min_{P_Z: \mathcal{D}(P_Z \| P_0) \leq \lambda} \left( \min_{P_{X|Z}: E_{XZ}d(X, Z) \leq \Delta_0 + \Delta_1} \mathcal{D}(P_X \| P_1) \right), \quad (\text{I.21})$$

which is equivalent to (24) (see Section III). The equivalence of the expressions in (I.20) and (I.21) follows from the equivalence of the two feasible sets for the PMF  $P_X$ . We first show that any feasible  $P_X$  in (I.20) is also feasible in (I.21). Let then  $P_X$  be a feasible PMF in (I.20). By exploiting the properties of the triangular inequality property of the distance, we have that, regardless of the specific choice of the distributions  $P_{Y|Z}$  and  $P_{X|Y}$  in (I.20),

$$E_{XZ}d(X, Z) \leq E_{XYZ}[d(X, Y) + d(Y, Z)] = E_{XY}d(X, Y) + E_{YZ}d(Y, Z) \leq \Delta_0 + \Delta_1, \quad (\text{I.22})$$

and then  $P_X$  is a feasible PMF in (I.21). To prove the opposite inclusion, we observe that, for any  $P_Z$  and  $P_{X|Z}$  such that  $\mathcal{D}(P_Z \| P_0) \leq \lambda$  and  $E_{XZ}d(X, Z) \leq \Delta_0 + \Delta_1$ , it is possible to define a variable  $Y$ , and then two conditional PMFs  $P_{Y|Z}$  and  $P_{X|Y}$ , such that  $E_{XY}d(X, Y) \leq \Delta_1$  and  $E_{YZ}d(Y, Z) \leq \Delta_0$ . To do so, it is sufficient to let  $P_Y$  be the convex combination of  $P_X$  and  $P_Z$ , that is  $P_Y = \alpha P_X + (1 - \alpha)P_Z$  where  $\alpha = \Delta_0 / (\Delta_0 + \Delta_1)$ . With this choice for the marginal, we can define  $P_{X|Y}$  so that  $P_{XY}$  satisfies<sup>12</sup>

$$\begin{aligned} P_{XY}(i, j) &= (1 - \alpha)P_{XZ}(i, j) \quad \forall i, \forall j \neq i, \\ P_{XY}(i, i) &= (1 - \alpha)P_{XZ}(i, i) + \alpha P_X(i) \quad \forall i; \end{aligned} \quad (\text{I.23})$$

similarly,  $P_{Y|Z}$  can be chosen such that  $P_{YZ}$  satisfies

$$\begin{aligned} P_{YZ}(i, j) &= \alpha P_{XZ}(i, j) \quad \forall i, \forall j \neq i, \\ P_{YZ}(i, i) &= \alpha P_{XZ}(i, i) + (1 - \alpha)P_Z(i) \quad \forall i. \end{aligned} \quad (\text{I.24})$$

It is easy to see that, with the above choices,  $E_{XY}d(X, Y) = (1 - \alpha)E_{XZ}d(X, Z)$  and  $E_{YZ}d(Y, Z) = \alpha E_{XZ}d(X, Z)$ . Then,  $E_{XY}d(X, Y) \leq (1 - \alpha)(\Delta_0 + \Delta_1) \leq \Delta_1$  and  $E_{YZ}d(Y, Z) \leq \Delta_0$ . Consequently, any  $P_X$  belonging to the set in (I.21) also belongs to the one in (I.20).

## APPENDIX II

### BAYESIAN DETECTION GAME

This appendix contains the proofs for Section V.

<sup>12</sup>By adopting the transportation theory perspective introduced towards the end of Section VI, we can look at  $P_X$  and  $P_Y$  as two ways of piling up a certain amount of soil; then  $P_{XY}$  can be interpreted as a map which moves  $P_X$  into  $P_Y$  ( $P_{XY}(i, j)$  corresponds to the amount of soil moved from position  $i$  to  $j$ ). The map in (I.23) is the one which leaves in place a percentage  $\alpha$  of the mass and moves the remaining  $(1 - \alpha)$  percentage to fill the pile  $(1 - \alpha)P_Z$  according to map  $(1 - \alpha)P_{XZ}$ .

### A. Proof of Theorem 5

Given the probability distributions  $Q_0(\mathbf{y})$  and  $Q_1(\mathbf{y})$  induced by  $A_{\Delta_0}^*$  and  $A_{\Delta_1}^*$  respectively, the optimum decision rule is deterministic and is given by the likelihood ratio test (LRT) [40]:

$$\frac{1}{n} \ln \frac{Q_1(\mathbf{y})}{Q_0(\mathbf{y})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} a, \quad (\text{II.1})$$

which proves the optimality of the decision rule in (25).

To prove the asymptotic optimality of the decision rule in (26), let us approximate  $Q_0(\mathbf{y})$  and  $Q_1(\mathbf{y})$  using the method of types as follows:

$$\begin{aligned} Q_0(\mathbf{y}) &= \sum_{\mathbf{x}} P_0(\mathbf{x}) A_{\Delta_0}^*(\mathbf{y}|\mathbf{x}) \\ &\doteq \sum_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} e^{-n[\hat{H}\mathbf{x}(X) + \mathcal{D}(\hat{P}\mathbf{x} \| P_0)]} \cdot e^{-n\hat{H}\mathbf{x}\mathbf{y}(Y|X)} \\ &\doteq \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} e^{n\hat{H}\mathbf{x}\mathbf{y}(X|Y)} \cdot \left( e^{-n[\hat{H}\mathbf{x}(X) + \mathcal{D}(\hat{P}\mathbf{x} \| P_0)]} \right. \\ &\quad \left. \cdot e^{-n\hat{H}\mathbf{x}\mathbf{y}(Y|X)} \right) \\ &= \max_{\mathbf{x}: d(\mathbf{x}, \mathbf{y}) \leq n\Delta_0} e^{-n[\hat{H}\mathbf{y}(Y) + \mathcal{D}(\hat{P}\mathbf{x} \| P_0)]} \\ &\stackrel{(a)}{=} \exp \left\{ -n \left[ \hat{H}\mathbf{y}(Y) + \right. \right. \\ &\quad \left. \left. + \min_{\{\hat{P}\mathbf{x}, \mathbf{y}: E\mathbf{x}\mathbf{y}d(X,Y) \leq \Delta_0\}} \mathcal{D}(\hat{P}\mathbf{x} \| P_0) \right] \right\} \\ &= \exp \left\{ -n[\hat{H}\mathbf{y}(Y) + \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}\mathbf{y}, P_0)] \right\}, \end{aligned} \quad (\text{II.2})$$

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

$$Q_1(\mathbf{y}) \doteq \exp \left\{ -n[\hat{H}\mathbf{y}(Y) + \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}\mathbf{y}, P_1)] \right\}. \quad (\text{II.3})$$

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

$$\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}\mathbf{y}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}\mathbf{y}, P_1) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} a, \quad (\text{II.4})$$

which proves the second part of the theorem.

### B. Proof of Theorem 7

In order to make the expression of  $u(P_{\text{FN}}(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*)))$  explicit, let us first evaluate the two error probabilities at equilibrium. Below, we derive the lower and upper bound on the probability of  $\mathbf{y}$  under  $\mathcal{H}_1$ , when the attack channel is  $A_{\Delta_1}^*$ :

$$(n+1)^{-|\mathcal{A}||\mathcal{A}-1|} e^{-n[\hat{H}\mathbf{y}(Y) + \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}\mathbf{y}, P_1)]} \leq Q_1^*(\mathbf{y}) < (n+1)^{|\mathcal{A}|^2} e^{-n[\hat{H}\mathbf{y}(Y) + \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}\mathbf{y}, P_1)]}. \quad (\text{II.5})$$

The same bounds hold for  $Q_0^*(\mathbf{y})$ , with  $\tilde{\mathcal{D}}_{\Delta_0}$  replacing  $\tilde{\mathcal{D}}_{\Delta_1}$ . For the FN probability, the upper bound is

$$\begin{aligned}
P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*) &= \sum_{\mathbf{y}} Q_1^*(\mathbf{y}) \cdot \Phi^\dagger(\mathcal{H}_0|\mathbf{y}) \\
&= \sum_{\mathbf{y}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} Q_1^*(\mathbf{y}) \\
&\leq (n+1)^{|\mathcal{A}|^2} \sum_{\mathbf{y}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} e^{-n[\hat{H}\mathbf{y} + \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1)]} \\
&\leq (n+1)^{|\mathcal{A}|^2 + |\mathcal{A}|} \max_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} e^{-n\tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1)} \\
&= (n+1)^{|\mathcal{A}|^2 + |\mathcal{A}|} \exp \left\{ -n \left( \min_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) \right) \right\}. \quad (\text{II.6})
\end{aligned}$$

Then,

$$-\limsup_{n \rightarrow \infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*)) \leq \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \quad (\text{II.7})$$

For the lower bound,

$$\begin{aligned}
P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*) &\geq (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \sum_{\mathbf{y}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} e^{-n[\hat{H}\mathbf{y} + \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1)]} \\
&\geq (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \max_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} e^{-n\tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1)} \\
&= (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \exp \left\{ -n \left( \min_{\hat{P}_{\mathbf{y}}: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a} \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) \right) \right\}. \quad (\text{II.8})
\end{aligned}$$

Then

$$\begin{aligned}
-\liminf_{n \rightarrow \infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*)) &\geq \lim_{n \rightarrow \infty} \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) \\
&= \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \quad (\text{II.9})
\end{aligned}$$

where  $\hat{P}_{\mathbf{y}}$  is a properly chosen PMF, belonging to the set  $\{\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a\}$  for every  $n$ , and such that  $\hat{P}_{\mathbf{y}} \rightarrow P_Y^*$  where<sup>13</sup>

$$P_Y^* = \arg \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \quad (\text{II.10})$$

<sup>13</sup>By Property 1, set  $\{\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\mathbf{y}}, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\mathbf{y}}, P_1) < a\}$  is dense in  $\{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a\}$  and then such a sequence of PMFs can always be found.



By combining (II.7) and (II.9), we get

$$\varepsilon_{\text{FN}} = - \lim_{n \rightarrow \infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*)) = \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \quad (\text{II.11})$$

Therefore, from (II.6) and (II.8) we have

$$P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*) \doteq \exp \left\{ -n \left( \min_{\hat{P}_Y: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1) < a} \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1) \right) \right\}, \quad (\text{II.12})$$

and the limit of  $\frac{1}{n} \ln P_{\text{FN}}$  exists and is finite.

Similar bounds can be derived for the FP probability, resulting in

$$P_{\text{FP}}(\Phi^*, A_{\Delta_0}^*) \doteq \exp \left\{ -n \left( \min_{\hat{P}_Y: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1) \geq a} \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) \right) \right\}, \quad (\text{II.13})$$

and in particular

$$\varepsilon_{\text{FP}} = - \lim_{n \rightarrow \infty} \frac{1}{n} \ln(P_{\text{FP}}(\Phi^*, A_{\Delta_0}^*)) = \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \geq a} \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0). \quad (\text{II.14})$$

From (II.14), we see that, as argued, the profile  $(\Phi^\dagger, (A_{\Delta_0}^*, A_{\Delta_1}^*))$  leads to a FP exponent always at least as large as  $a$ .

We are now ready to evaluate the asymptotic behavior of the payoff of the Bayesian detection game:

$$\begin{aligned} u &= P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*) + e^{an} P_{\text{FP}}(\Phi^\dagger, A_{\Delta_0}^*) \\ &\doteq \max\{P_{\text{FN}}(\Phi^\dagger, A_{\Delta_1}^*), e^{an} P_{\text{FP}}(\Phi^\dagger, A_{\Delta_0}^*)\} \\ &\doteq \exp \left\{ -n \min \left( \min_{\hat{P}_Y: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1) < a} \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1), \min_{\hat{P}_Y: \tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_Y, P_1) \geq a} (\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_Y, P_0) - a) \right) \right\} \\ &= \exp \left\{ -n \min_{P_Y} \left( \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}^n(P_Y, P_1), (\tilde{\mathcal{D}}_{\Delta_0}^n(P_Y, P_0) - a) \right\} \right) \right\} \\ &\doteq \exp \left\{ -n \min_{P_Y} \left( \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), (\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - a) \right\} \right) \right\}, \end{aligned} \quad (\text{II.15})$$

where the asymptotic equality in the last line follows from the density of the set of empirical probability distributions of  $n$ -length sequences in the probability simplex and from the continuity of the to-be-minimized expression in round brackets as a function of  $P_Y$ .

### APPENDIX III

#### SOURCE DISTINGUISHABILITY

This appendix contains the proofs for Section VI.

### A. Proof of Theorem 9

The theorem directly follows from Theorem 7. In fact, by letting

$$e_a(P_Y) = \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - a \right\}, \quad (\text{III.1})$$

$a \geq 0$ , the limit in (29) can be derived as follows:

$$\begin{aligned} \lim_{a \rightarrow 0} \min_{P_Y} e_a(P_Y) &= \min_{P_Y} \lim_{a \rightarrow 0} e_a(P_Y) \\ &= \min_{P_Y} \left( \max \left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \right\} \right), \end{aligned} \quad (\text{III.2})$$

where the order of limit and minimum can be exchanged because of the uniform convergence of  $e_a(P_Y)$  to  $e_0(P_Y)$  as  $a$  tends to 0.

### B. Proof of Corollary 2

The corollary can be proven by exploiting the fact that, when  $d$  is a metric, the *EMD* is a metric and then  $EMD_d(P_0, P)$  satisfies the triangular inequality. In this case, it is easy to argue that the  $P_Y$  achieving the minimum in (32) is the one for which the triangular relation holds at the equality, which corresponds to the convex combination of  $P_0$  and  $P$  (i.e., the PMF lying on the straight line between  $P_0$  and  $P$ ) with combination coefficient  $\alpha$  such that  $EMD_d(P_0, P_Y)$  (or equivalently, by symmetry,  $EMD_d(P_Y, P_0)$ ) is exactly equal to  $\Delta_0$ .

Formally, let  $X \sim P_0$  and  $Z \sim P$ . We want to find the PMF  $P_Y$  which solves

$$\min_{P_Y: EMD_d(P_Y, P_0) \leq \Delta_0} EMD_d(P_Y, P). \quad (\text{III.3})$$

For any  $Y \sim P_Y$  and any choice of  $P_{XY}$  and  $P_{YZ}$  (that is,  $P_{Y|X}$  and  $P_{Z|Y}$ ), by exploiting the triangular inequality property of the distance, we can write

$$E_{XZ}d(X, Z) \leq E_{XY}d(X, Y) + E_{YZ}d(Y, Z), \quad (\text{III.4})$$

where  $P_{XZ}$  can be any joint distribution with marginals  $P_0$  and  $P$ . Then,

$$EMD(P_0, P) \leq E_{XY}d(X, Y) + E_{YZ}d(Y, Z). \quad (\text{III.5})$$

From the arbitrariness of the choice of  $P_{XY}$  and  $P_{YZ}$ , if we let  $P_{XY}^*$  and  $P_{YZ}^*$  be the joint distributions achieving the *EMD* between  $X$  and  $Y$ , and  $Y$  and  $Z$ , we get

$$EMD(P_0, P) \leq EMD(P_0, P_Y) + EMD(P_Y, P). \quad (\text{III.6})$$

From the above relation, we can derive the following lower bound for the to-be-minimized quantity in (III.3):

$$EMD(P_Y, P) \geq EMD(P_0, P) - EMD(P_0, P_Y) \quad (\text{III.7})$$

$$\geq EMD(P_0, P) - \Delta_0. \quad (\text{III.8})$$

We now show that  $P_Y$  defined as in (33) achieves the above lower bound while satisfying the constraint  $EMD(P_0, P_Y) \leq \Delta_0$ , and then gets the minimum value in (III.3).

Let  $P_{XZ}^*$  be the joint distribution achieving the  $EMD$  between  $X$  and  $Z$ . Then,  $E_{XZ}^* d(X, Z) = EMD(P_0, P)$  (where the star on the apex indicates that the expectation is taken under  $P_{XZ}^*$ ). Given the marginal  $P_Y = \alpha P_0 + (1 - \alpha)P$ , we can define  $P_{XY}$  and  $P_{YZ}$ , starting from  $P_{XZ}^*$ , as in the proof of Theorem 4 ((I.23) and (I.24)). With this choice,  $E_{XY} d(X, Y) = (1 - \alpha)EMD(P_0, P)$  and  $E_{YZ} d(Y, Z) = \alpha EMD(P_0, P)$ . Then, for the value of  $\alpha$  in (33) we have that  $E_{XY} d(X, Y) = \Delta_0$  and

$$E_{YZ} d(Y, Z) = EMD(P_0, P) - \Delta_0. \quad (\text{III.9})$$

By combining (III.9) and (III.8), we argue that  $EMD(P_Y, P) = EMD(P_0, P) - \Delta_0$ <sup>14</sup>. Therefore,  $P_Y$  in (33) solves (III.3).

To prove the second part of the corollary, we just need to observe that a PMF  $P$  belongs to the indistinguishability set in (32) if and only if

$$EMD(P_Y, P) = EMD(P_0, P) - \Delta_0 \leq \Delta_1, \quad (\text{III.10})$$

that is  $EMD(P_0, P) \leq \Delta_0 + \Delta_1$ .

From the above proof, we notice that, for any  $P$  in the set in (34), i.e., such that  $EMD_d(P_0, P) \leq \Delta_0 + \Delta_1$ , the PMF  $P_Y = \alpha P_0 + (1 - \alpha)P$  with  $\alpha$  as in (33) satisfies  $EMD(P_Y, P_0) = \Delta_0$  and  $EMD(P_Y, P_1) = \Delta_1$  for any choice of  $d$ . Then, when  $d$  is not a metric, the region in (34) is contained in the indistinguishability region.

<sup>14</sup>We also argue that the choice made for  $P_{YZ}$  minimizes the expected distortion between  $Y$  and  $Z$ , i.e., it yields  $E_{YZ} d(Y, Z) = EMD(P_Y, P)$ . Furthermore, being  $E_{XY} d(X, Y) = \Delta_0$ , it holds  $EMD(P_Y, P) = EMD(P_0, P) - E_{XY} d(X, Y)$  and then, from the triangular inequality in (III.6), it follows that  $EMD(P_0, P_Y) = E_{XY} d(X, Y) = \Delta_0$ .

### C. Proof of Corollary 3

By inspecting the minimization in (32), we see that for any source  $P$  that cannot be distinguished from  $P_0$ , it is possible to find a source  $P_Y$  such that  $EMD_d(P_Y, P) \leq \Delta_1$  and  $EMD_d(P_Y, P_0) \leq \Delta_0$ . In order to prove the corollary, we need to show that such  $P$  lies inside the set defined in (35).

We give the following definition. Given two random variables  $X$  and  $Y$ , the Hölder inequality applied to the expectation function ([39]) reads:

$$E_{XY}|XY| \leq (E_X[|X|^r])^{1/r} (E_Y[|Y|^q])^{1/q}, \quad (\text{III.11})$$

where  $r \geq 1$  and  $q = r/(r-1)$ , namely, the Hölder conjugate of  $r$ .

We use the notation  $E_{XY}^*$  for the expectation of the pair  $(X, Y)$  when the probability map is the one achieving the  $EMD(P_X, P_Y)$ , namely  $P_{XZ}^*$ . Then, we can write:

$$\begin{aligned} EMD_{L_p^p}(P_0, P) &= E_{XZ}^*[||X - Z||^p] \\ &\stackrel{(a)}{\leq} E_{XYZ}^*[(||X - Y|| + ||Y - Z||)^p] \\ &\stackrel{(b)}{\leq} E_{XYZ} [ ||X - Y||^p + ||Y - Z||^p + p \cdot ||X - Y||^{p-1} ||Y - Z|| + \\ &\quad + p(p-1)/2 \cdot ||X - Y||^{p-2} ||Y - Z||^2 + \dots + p \cdot ||X - Y|| ||Y - Z||^{p-1} ] \\ &= E_{XYZ}[||X - Y||^p] + E_{XYZ}[||Y - Z||^p] + p \cdot E_{XYZ}[||X - Y||^{p-1} ||Y - Z||] + \\ &\quad + p(p-1)/2 \cdot E_{XYZ}[||X - Y||^{p-2} ||Y - Z||^2] + \dots + p \cdot E_{XYZ}[||X - Y|| ||Y - Z||^{p-1}] \\ &\stackrel{(c)}{\leq} E_{XYZ}[||X - Y||^p] + E_{XYZ}[||Y - Z||^p] + p \cdot (E_{XYZ}[||X - Y||^p])^{\frac{p-1}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{1}{p}} \\ &\quad + p(p-1)/2 \cdot (E_{XYZ}[||X - Y||^p])^{\frac{p-2}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{2}{p}} + \dots \\ &\quad \dots + p \cdot (E_{XYZ}[||X - Y||^p])^{\frac{1}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{p-1}{p}} \\ &= E_{XY}[||X - Y||^p] + E_{YZ}[||Y - Z||^p] + p \cdot (E_{XY}[||X - Y||^p])^{\frac{p-1}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{1}{p}} \\ &\quad + p(p-1)/2 \cdot (E_{XY}[||X - Y||^p])^{\frac{p-2}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{2}{p}} + \dots \\ &\quad \dots + p \cdot (E_{XY}[||X - Y||^p])^{\frac{1}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{p-1}{p}} \\ &= \left( (E_{XYZ}[||X - Y||^p])^{1/p} + (E_{XYZ}[||Y - Z||^p])^{1/p} \right)^p \\ &\leq \left( \Delta_0^{1/p} + \Delta_1^{1/p} \right)^p, \end{aligned} \quad (\text{III.12})$$

where in (a) we considered the joint distribution  $P_{XYZ}$  such that  $\sum_Z P_{XYZ} = P_{XY}^*$ ,  $\sum_X P_{XYZ} = P_{YZ}^*$  (and, consequently,  $\sum_Y P_{XYZ} = P_{XZ}^*$ ) and in (b) we developed the  $p$ -power of the binomial (binomial

theorem). Finally, in (c), we applied the Hölder's inequality to the various terms of Newton's binomial: specifically, for each term  $E_{XYZ}[||X - Y||^{p-t} ||Y - Z||^t]$ , with  $t = 1, \dots, p - 1$ , the Hölder inequality is applied with  $r = p/(p - t)$  (and  $q = r/(r - 1)$ ).

## REFERENCES

- [1] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds. Springer Berlin / Heidelberg, 2012.
- [2] D. Lowd and C. Meek, "Adversarial learning," in *Proc. of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 641–647.
- [3] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *Proc. of EUSIPCO'05, European Signal Processing Conference*. IEEE, 2005, pp. 1–4.
- [4] A. Cohen and A. Lapidot, "The Gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1639–1667, Jun 2002.
- [5] K. Tan, K. Killourhy, and R. Maxion, "Undermining an anomaly-based intrusion detection system using common exploits," in *Recent Advances in Intrusion Detection*. Springer, 2002, pp. 54–73.
- [6] 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.
- [7] A. D. Ker, "Batch steganography and the threshold game," in *Security, Steganography, and Watermarking of Multimedia Contents*, 2007, p. 650504.
- [8] Y. Wu, B. Wang, K. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 1, pp. 4–15, 2012.
- [9] N. Dalvi, P. Domingos, P. Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 99–108.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [11] J. O'Sullivan, P. Moulin, and J. Ettinger, "Information theoretic analysis of steganography," in *Proc. of IEEE International Symposium on Information Theory, 1998*, Aug 1998, pp. 297–.
- [12] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 563–593, Mar 2003.
- [13] 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.
- [14] 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.
- [15] 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, January 2008.
- [16] 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.
- [17] —, "Adversarial source identification game with corrupted training," *submitted to IEEE Transactions on Information Theory (on second review round)*, arXiv preprint arXiv:1703.09244, 2017.

- [18] 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.
- [19] 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.
- [20] P. Moulin and A. Ivanovic, "Game-theoretic analysis of watermark detection," in *Proc. of International Conference on Image Processing, 2001*, vol. 3. IEEE, 2001, pp. 975–978.
- [21] I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *Information Sciences*, vol. 239, pp. 201–225, 2013.
- [22] G. Vigna, W. Robertson, and D. Balzarotti, "Testing network-based intrusion detection signatures using mutant exploits," in *Proceedings of the 11th ACM conference on Computer and communications security*. ACM, 2004, pp. 21–30.
- [23] S. Patton, W. Yurcik, and D. Doss, "An achilles' heel in signature-based ids: Squealing false positives in snort," in *Proceedings of RAID*, vol. 2001, 2001.
- [24] D. Mutz, C. Kruegel, W. Robertson, G. Vigna, and R. A. Kemmerer, "Reverse engineering of network signatures," in *In proceedings of Auscert Asia Pacific Information Technology Security Conference, Gold*. Citeseer, 2005.
- [25] T. M. Cover and J. A. Thomas, "Elements of information theory 2nd edition," 2006.
- [26] B. Tondi, M. Barni, and N. Merhav, "Detection games with a fully active attacker," in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, Nov 2015, pp. 1–6.
- [27] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.
- [28] J. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.
- [29] 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.
- [30] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.
- [31] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, pp. 369–401, 1965.
- [32] S. I.N., "On the probability of large deviations of random variables," *Math. Sbornik*, vol. 42, pp. 11–44, 1957.
- [33] I. Csiszár and P. Shields, *Information Theory and Statistics: a Tutorial*. Now Publishers Inc., 2004.
- [34] M. Barni and B. Tondi, "Source distinguishability under distortion-limited attack: An optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, Oct 2016.
- [35] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, November 2000.
- [36] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.
- [37] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematical Physics*, vol. 20, pp. 224–230.
- [38] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.
- [39] A. F. Karr, *Probability*. Springer-Verlag New York, 1993.
- [40] 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>