entropy

MDPI

Article

Detection Games under Fully Active Adversaries

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

[1] Department of Information Engineering and Mathematical Sciences, University of Siena, 53100 Siena, Italy; barni@dii.unisi.it

[2] The Andrew and Erna Viterbi Faculty of Electrical Engineering—Israel Institute of Technology Technion City, Haifa 3200003, Israel; merhav@ee.technion.ac.il

* Correspondence: benedettatondi@gmail.com; Tel.: +39-0577-23-3752

Received: 23 November 2018; Accepted: 25 December 2018; Published: 29 December 2018

Abstract: We study a binary hypothesis testing problem in which a defender must decide whether a test sequence has been drawn from a given memoryless source P_0, while 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, 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 characterizing the conditions for the distinguishability of the two sources given the distortion levels.

Keywords: adversarial signal processing; binary hypothesis testing; statistical detection theory; game theory; the method of types

1. Introduction

Signal processing techniques are routinely applied in the great majority of security-oriented applications. In particular, the detection problem plays a fundamental role in many security-related scenarios, including secure detection and classification, target detection in radar-systems, intrusion detection, biometric-based verification, one-bit watermarking, steganalysis, spam-filtering, and multimedia forensics (see [1]). A unifying characteristic of all these fields is the presence of an adversary explicitly aiming at hindering the detection task, with the consequent necessity of adopting proper models that take into account the interplay between the system designer and the adversary. Specifically, the adopted models should be able to deal, on the one hand, with the uncertainty that the designer has about the attack the system is subject to, and, on the other hand, with the uncertainty of the adversary about the target system (i.e., the system it aims at defeating). Game theory has recently been proposed as a way to study the strategic interaction between the choices made by the system designer and the adversary. Among the fields wherein such an approach has been used, we mention steganalysis [2,3], watermarking [4], intrusion detection [5] and adversarial machine learning [6,7]. With regard to binary detection, game theory and information theory have

been combined to address the problem of adversarial detection, especially in the field of digital watermarking (see, for instance, [4,8–10]). 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 is made in [11]. Specifically, in [11], the general problem of binary decision under adversarial conditions is 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 [11] 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 [11] as a competitive zero-sum game; the asymptotic equilibrium, i.e., 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 [11] extends the one of [12] to the adversarial scenario. Some variants of this attack-detection game have also been studied; for instance, in [13], the setting is extended to the case where the sources are known to neither the defender nor the attacker, while the training data from both sources are 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 [14]).

The assumption that the attacker is active only under H_1 stems from the observation that in many cases H_0 corresponds to a kind of *normal* or *safe* situation and its rejection corresponds to revealing the presence of a threat or that something anomalous is happening. This is the case, for instance, in intrusion detection systems, tampering detection, target detection in radar systems, steganalysis and so on. In these cases, the goal of the attacker is to avoid the possibility that the monitoring system raises an alarm by rejecting the hypothesis H_0, while having no incentive to act under H_0. Moreover, in many cases, the attack is not even present under H_0 or it does not have the possibility of manipulating the observations the detector relies on when H_0 holds.

In many other cases, however, 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. As an example, we may consider the case of a radar target detection system, where the defender wishes to distinguish between the presence and the absence of a target, by considering the presence of a hostile jammer. To maximize the damage caused by his actions, the jammer may decide to work under both the hypotheses: when H_1 holds, to avoid that the defender detects the presence of the target, and in the H_0 case, to increase the number of false alarms inducing a waste of resources deriving from the adoption of possibly expensive countermeasures even when they are not needed. In a completely different scenario, we may consider an image forensic system aiming at deciding whether a certain image has been shot by a given camera, for instance because the image is associated to a crime such as child pornography or terrorism. Even in this case, the attacker may be interested in causing a missed detection event to avoid that he is accused of the crime associated to the picture, or to induce a false alarm to accuse an innocent victim. Other examples come from digital watermarking, where an attacker can be interested in either removing or injecting the watermark from an image or a video, to redistribute the content without any ownership information or in such a way to support a false copyright statement [15], and network intrusion detection systems [16], wherein the adversary may try to both avoid the detection of the intrusion by manipulating a malicious traffic, and to implement an overstimulation attack [17,18] to cause a denial of service failure.

According to the scenario at hand, the attacker may be aware of the hypothesis under which it is operating (hypothesis-aware attacker) or not (hypothesis-unaware attacker). While the setup considered in this paper can be used to address both cases, we focus mainly on the case of an hypothesis-aware attacker, since this represents a worst-case assumption for the defender. In addition, the case of an hypothesis-unaware attacker can be handled as a special case of the more general case of an aware attacker subject to identical constraints under the two hypotheses.

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 [11,13]), it is referred to as a *partially active attacker*. As we show, the game in the partially active case turns out to be a special case of the game with fully active, hypothesis-aware, attacker. Accordingly, the hypothesis-aware fully active scenario forms a unified framework that includes the hypothesis-unaware case and the partially active scenario as special cases.

We define and solve two versions of the *detection game with fully active attackers*, corresponding to two different formulations of the problem: a Neyman–Pearson formulation and a Bayesian formulation. Another difference with respect to [11] is that here the players are allowed to adopt randomized strategies. Specifically, the defender can adopt a *randomized decision* strategy, while in [11] the defender's strategies are confined to deterministic decision rules. As for the attack, it consists of the application of a *channel*, whereas in [11] it is confined to the application of a deterministic function. Moreover, the partially active case of [11] 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 optimal 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 [12] for an introduction on this terminology). In particular, as already done in related literature [11,13,14], we limit the detection resources to first-order statistics, which are known to be a sufficient statistic for memoryless systems (Section 2.9, [19]). In the setup studied in this paper, the sources are indeed assumed to be memoryless, however 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. Forcing, nonetheless, the defender to rely on first-order statistics is motivated mainly by its simplicity. In addition, the use of first-order statistics is common in a number of application scenarios even if the source under analysis is not memoryless. In image forensics, for instance, several techniques have been proposed which rely on the analysis of the image histogram or a subset of statistics derived from it, e.g., for the detection of contrast enhancement [20] or cut-and-paste [21] processing. As another example, the analysis of statistics derived from the histograms of block-DCT coefficients is often adopted for detecting both single and multiple JPEG compression [22]. More generally, we observe that the assumption of limited resources is reasonable in application scenarios where the detector has a small computational power. Having said that, it should be also emphasized that the analysis in this work can be extended to richer sets of empirical statistics, e.g., higher-order statistics (see the Conclusions Section for a more elaborated discussion on this point).

As a last note, we observe that, although we limit ourselves to memoryless sources, our results can be easily extended to more general models (e.g., Markov sources), as long as a suitable extension of the method of types is available.

One of the main results of this paper is the characterization of an attack strategy that is both *dominant* (i.e., optimal no matter what the defense strategy is), and *universal*, i.e., independent of the (unknown) underlying sources. Moreover, this optimal attack strategy turns out to be the same under both hypotheses, thus rendering the distinction between the hypothesis-aware and the hypothesis-unaware scenarios completely inconsequential. In other words, the optimal attack strategy is universal, not only with respect to uncertainty in the source statistics under either hypothesis, but also

with respect to the unknown hypothesis in the hypothesis-unaware case. Moreover, the 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 marking a significant difference with respect to previous works [11,13], where the existence of a dominant strategy wasestablished with regard to 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 [23]. 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 6), 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 [23] 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 2, we establish the notation and introduce the main concepts. In Section 3, 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 4 and 5, respectively. Finally, Section 6 is devoted to the analysis of the best achievable performance of the defender and the characterization of the source distinguishability.

2. Notation and Definitions

Throughout the paper, random variables are denoted by capital letters and specific realizations are denoted by the corresponding lower case letters. All random variables that denote signals in the system are assumed to have the same finite alphabet, denoted by $\mathcal{A}$. Given a random variable X and a positive integer n, we denote by $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, $X_i \in \mathcal{A}$, $i = 1, 2, \ldots, n$, a sequence of n independent copies of X. According to the above-mentioned notation rules, a specific realization of $\boldsymbol{X}$ is denoted by $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Sources are denoted by the letter P. Whenever necessary, we 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., $\lim_{n\to\infty} 1/n \ln(a_n/b_n) = 0$, and $a_n \dot{\leq} b_n$ designates that $\limsup_{n\to\infty} 1/n \ln(a_n/b_n) \leq 0$.

For a given real s, we denote $[s]_+ \stackrel{\triangle}{=} \max\{s, 0\}$. We use notation $U(\cdot)$ for the Heaviside step function.

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

Regarding information measures, the entropy associated with $\hat{P}_{\boldsymbol{x}}$, which is the empirical entropy of $\boldsymbol{x}$, is denoted by $\hat{H}_{\boldsymbol{x}}(X)$. Similarly, $\hat{H}_{\boldsymbol{xy}}(X, Y)$ designates the empirical joint entropy of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\hat{H}_{\boldsymbol{xy}}(X|Y)$ 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 [19]).

Finally, we use the letter A to denote an attack channel; accordingly, $A(\boldsymbol{y}|\boldsymbol{x})$ is the conditional probability of the channel output $\boldsymbol{y}$ given the channel input $\boldsymbol{x}$. Given a permutation-invariant distortion

function $d\colon \mathcal{A}^n \times \mathcal{A}^n \to \mathbb{R}^+$ (a permutation-invariant distortion function $d(\boldsymbol{x},\boldsymbol{y})$ is a distortion function that is invariant if the same permutation is applied to both $\boldsymbol{x}$ and $\boldsymbol{y}$) and a maximum distortion Δ, we define the class $\mathcal{C}_\Delta$ of admissible channels $\{A(\boldsymbol{y}|\boldsymbol{x}),\ \boldsymbol{x},\boldsymbol{y} \in \mathcal{A}^n\}$ as those that assign zero probability to every $\boldsymbol{y}$ with $d(\boldsymbol{x},\boldsymbol{y}) > n\Delta$.

2.1. 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 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 $\mathcal{S}_1$ and $\mathcal{S}_2$ 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 [24]. The most famous notion of equilibrium is due to Nash [25]. 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 [24]. The concept of dominance-solvability is directly related to the notion of strict dominance 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. Similarly, 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 (in game theory, a rational player is supposed to act in a way that maximizes its payoff) 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 [26,27]. 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 that corresponds to dominant strategies for both players and is the strongest kind of equilibrium that a strategic game may have.

3. Detection Game with Fully Active Attacker

3.1. Problem Formulation

Given two discrete memoryless sources, P_0 and P_1, defined over a common finite alphabet $\mathcal{A}$, we denote by $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{A}^n$ a sequence emitted by one of these sources. The sequence $\boldsymbol{x}$ is available to the attacker. Let $\boldsymbol{y} = (y_1, y_2, ..., y_n) \in \mathcal{A}^n$ denote the sequence observed by the defender:

when an attack occurs under both $\mathcal{H}_0$ and $\mathcal{H}_1$, the observed sequence $\boldsymbol{y}$ is obtained as the output of an attack channel fed by $\boldsymbol{x}$.

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: $A_0(\boldsymbol{y}|\boldsymbol{x})$, applied when $\mathcal{H}_0$ holds, and $A_1(\boldsymbol{y}|\boldsymbol{x})$, applied under $\mathcal{H}_1$. Let us denote by $Q_i(\cdot)$ the PMF of $\boldsymbol{y}$ under $\mathcal{H}_i$, $i = 0, 1$. The attack induces the following PMFs on $\boldsymbol{y}$: $Q_0(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P_0(\boldsymbol{x})A_0(\boldsymbol{y}|\boldsymbol{x})$ and $Q_1(\boldsymbol{y}) = \sum_{\boldsymbol{x}} P_1(\boldsymbol{x})A_1(\boldsymbol{y}|\boldsymbol{x})$.

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 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. Obviously, the partially active case, where no attack occurs under $\mathcal{H}_0$ can be seen as a degenerate case of the hypothesis-aware fully active one, where A_0 is the identity channel I.

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

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

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

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

As in [11], due to the limited resources assumption, the defender makes a decision based on first-order empirical statistics of $\boldsymbol{y}$, which implies that $\Phi(\cdot|\boldsymbol{y})$ depends on $\boldsymbol{y}$ only via its type class $\mathcal{T}(\boldsymbol{y})$.

Concerning the attack, to limit the amount of distortion, we assume a distortion constraint. In the hypothesis-aware case, we allow the attacker different distortion levels, Δ_0 and Δ_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 both cases.

Figure 1 provides a block diagram of the system with a fully active attacker studied in this paper.

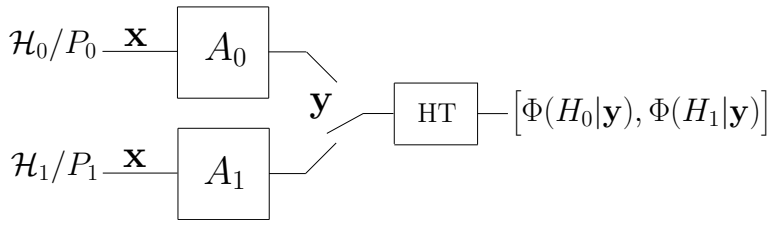

Figure 1. Schematic representation of the adversarial setup considered in this paper. In the case of partially active attacker, channel A_0 corresponds to the identity channel.

3.2. 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)** *$\Phi(\mathcal{H}_0|\boldsymbol{y})$ depends on $\boldsymbol{y}$ 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 λ. Such a requirement is relatively strong, its main consequence being that the strategy used by the attacker under H_0 is irrelevant, in that the payoff is the same whichever is the channel $A_0 \in \mathcal{C}_{\Delta_0}$ played by the attacker. It is worth observing that, to be more general, we could have defined the problem as a non-zero sum game, where the defender has payoff $u_D(\phi, A_1) = -P_{\text{FN}}(\Phi, A_1)$, whereas for the attacker we consider a payoff function of the form $u_A(\phi, (A_0, A_1)) = \beta P_{\text{FP}}(\Phi, A_0) + \gamma P_{\text{FN}}(\Phi, A_1)$, for some positive constant β and γ. As is made clear in the following section, this non-zero sum version of the game has the same equilibrium strategies of the zero-sum game defined above.

In the case of partially active attack (see the formulation in [23]), 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 $P_{\text{FP}}(\Phi) \leq e^{-n\lambda}$. 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 he 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 $\Phi(\mathcal{H}_0|\boldsymbol{y})$ depends on $\boldsymbol{y}$ 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*

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

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 optimal defense strategy is expected to yield error exponents that differ exactly by a, so as to balance the contributions of the two terms of Equation (3).

Notice also that, by defining the payoff as in Equation (3), we are implicitly considering for the defender only the strategies $\Phi(\cdot|\boldsymbol{y})$ such that $P_{\text{FP}}(\Phi, A_0) \dot{\leq} e^{-an}$. 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 $\boldsymbol{y}$ ($u = 1$).

As in [11], we focus on the asymptotic behavior of the game as n tends to infinity. In particular, we are interested in the exponents of the FP and FN probabilities, defined as follows:

$$\varepsilon_{\text{FP}} = -\limsup_{n\to\infty} \frac{1}{n} \ln P_{\text{FP}}(\Phi, A_0); \quad \varepsilon_{\text{FN}} = -\limsup_{n\to\infty} \frac{1}{n} \ln P_{\text{FN}}(\Phi, A_1). \tag{4}$$

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

3.3. 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

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

where β and γ 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 $c_n(\boldsymbol{x})$ denote the reciprocal of the total number of conditional type classes $\{\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})\}$ that satisfy the constraint $d(\boldsymbol{x}, \boldsymbol{y}) \leq n\Delta$ for a given $\Delta > 0$, namely, admissible conditional type classes (from the method of the types, it is known that $1 \geq c_n(\boldsymbol{x}) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)}$ for any $\boldsymbol{x}$ [19]).*

Define:

$$A^*_\Delta(\boldsymbol{y}|\boldsymbol{x}) = \begin{cases} \frac{c_n(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} & d(\boldsymbol{x}, \boldsymbol{y}) \leq n\Delta \\ 0 & \textit{elsewhere} \end{cases}. \tag{6}$$

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 and 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 $\Pi : \mathcal{A}^n \to \mathcal{A}^n$ denote a permutation operator that permutes any member of $\mathcal{A}^n$ according to a given permutation matrix and let

$$A_\Pi(\boldsymbol{y}|\boldsymbol{x}) \stackrel{\triangle}{=} A_1(\Pi\boldsymbol{y}|\Pi\boldsymbol{x}). \tag{7}$$

Since the distortion function is assumed to be permutation-invariant, the channel $A_\Pi(\boldsymbol{y}|\boldsymbol{x})$ introduces the same distortion as A_1, and hence it satisfies the distortion constraint. Due to the memorylessness of P_1 and the assumption that $\Phi(\mathcal{H}_0|\cdot)$ belongs to $\mathcal{S}_D$ (i.e., that $\Phi(\mathcal{H}_0|\cdot)$ depends on the observed sequence via its type), both $P_1(\boldsymbol{y})$ and $\Phi(\mathcal{H}_0|\boldsymbol{y})$ are invariant to permutations on $\boldsymbol{y}$. Then, we have:

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

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

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

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 $\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$. To see why this is true, we observe that any sequence $\boldsymbol{y}' \in \mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$ can be seen as being obtained from $\boldsymbol{y}$ through the application of a permutation Π', which leaves $\boldsymbol{x}$ unaltered. Then, we have:

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

Therefore, since the probability assigned by $\bar{A}$ to the sequences in $\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$ is surely less than or equal to 1, we argue that

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

which implies that, $P_{\text{FN}}(\Phi, \bar{A}) \dot{\leq} (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(A^*_{\Delta_1}, \Phi)$.

Then,

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

or equivalently

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

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 for the derivation of the optimal 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:

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

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:

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

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 exponential scale, Equation (15) holds no matter what the values of β and γ 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 Equation (5), if Φ 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 Φ. Then, the optimal attack strategy $(A^*_{\Delta_0}, A^*_{\Delta_1})$ is *dominant*. The intuition behind the attack channel in Equation (6) is illustrated in Figure 2 and explained below. Given $\boldsymbol{x}$, generated by a source P_X, the set $\{\boldsymbol{y} : d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta\}$ corresponds to a set of conditional type classes (for a permutation-invariant distortion function, $d(\boldsymbol{x},\boldsymbol{y})$ is the same for every $\boldsymbol{y} \in \mathcal{T}(\boldsymbol{y}|\boldsymbol{x})$). We say that a conditional type class is *admissible* if it belongs to this set. Then, to generate $\boldsymbol{y}$ 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 choosing at random $\boldsymbol{y}$ within this conditional type class. 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 $\boldsymbol{y}$ within that conditional type class is immaterial.

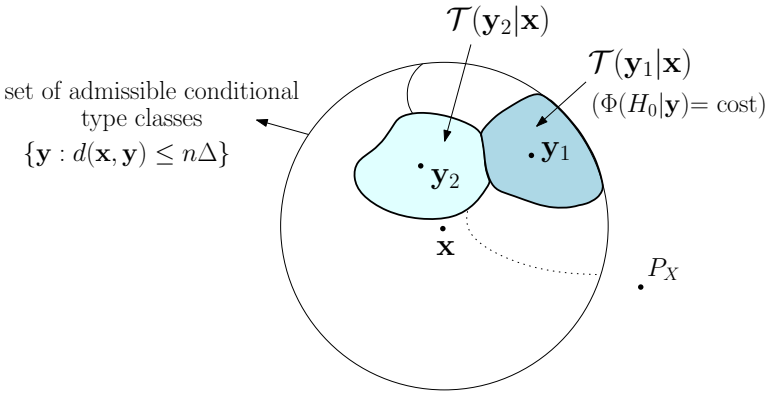

Figure 2. Graphical interpretation of the behavior of the attack channel A^*_Δ. The number of admissible conditional type classes is polynomial in n, that is $\{\boldsymbol{y} : d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta\} = \bigcup_{i \in p(n)} \mathcal{T}(\boldsymbol{y}_i|\boldsymbol{x})$ where $p(n)$ is a polynomial of n.

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

We observe that, if $\Delta_0 = \Delta_1 = \Delta$, the optimal attack consists of applying the same channel A^*_Δ regardless of the underlying hypothesis and then the optimal 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 Δ, the optimal channel remains A^*_Δ.

According to Theorem 1, for the partially active case, there exists an (asymptotically) dominant and universal attack channel. This result marks a considerable difference with respect to the results of [11], where the optimal deterministic attack function is found by using the rationalizability argument, that is, by exploiting the existence of a dominant defense strategy, and it is hence neither dominant nor universal.

Finally, we point out that, similar to [11], although the sources are memoryless and only first-order statistics are used by the defender, the output probability distributions induced by the attack channel A^*_Δ, namely Q_0 and Q_1, are not necessarily memoryless.

4. 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 3.3, we already know that there exists a dominant attack strategy. Regarding the defender, we determine the asymptotically optimal strategy regardless of the dominant pair of attack channels; in particular, as shown in Lemma 1, an asymptotically dominant defense strategy can be derived from a detailed analysis of the FP constraint. Consequently, the Neyman–Pearson detection game has a dominant equilibrium.

4.1. 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 defense strategy*

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

is asymptotically dominant for the defender.

In the special case $\Delta_0 = 0$, $\Phi^*(\mathcal{H}_1|\boldsymbol{y})$ in Equation (16) is (asymptotically) equivalent to the Hoeffding test for non-adversarial hypothesis testing [28].

The proof of Lemma 1 appears in Appendix A.1.

We point out that, when the attacker is partially active, it is known from [23] that the optimal defense strategy is

$$\Phi^*(\mathcal{H}_1|\boldsymbol{y}) \stackrel{\triangle}{=} \exp\left\{-n\left[\lambda - \mathcal{D}(\hat{P}_{\boldsymbol{y}}\|P_0)\right]_+\right\}, \tag{17}$$

which is in line with the one in [11] (Lemma 1), where the class of defense strategies is confined to deterministic decision rules.

Intuitively, the extension from Equations (16) and (17) 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 3.3, channel $A^*_{\Delta_0}$ minimizes the FP exponent over this set. To satisfy the constraint for a given sequence $\boldsymbol{y}$, the defender must handle the worst-case value (i.e., the minimum) of $\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)$ over all the type classes $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ which satisfy the distortion constraint, or equivalently, all the sequences $\boldsymbol{x}$ such that $d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta_0$.

According to Lemma 1, the best defense strategy is asymptotically dominant. In addition, since Φ^* 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 Equation (5) with $\gamma = 1$ and $\beta = 0$, the optimal 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 Φ^* 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.*

4.2. 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 assume an additive distortion function, i.e., $d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n d(x_i, y_i)$. In this case, $d(\boldsymbol{x},\boldsymbol{y})$ can be expressed as $\sum_{ij} n_{\boldsymbol{xy}}(i,j)d(i,j)$, where $n_{\boldsymbol{xy}}(i,j) = n\hat{P}_{\boldsymbol{xy}}(i,j)$ denotes the number of occurrences

of the pair $(i,j) \in \mathcal{A}^2$ in $(\boldsymbol{x},\boldsymbol{y})$. Therefore, the distortion constraint regarding A_0 can be rewritten as $\sum_{(i,j)\in\mathcal{A}^2} \hat{P}_{\boldsymbol{xy}}(i,j)d(i,j) \leq \Delta_0$. A similar formulation holds for A_1.

Let us define

$$\tilde{\mathcal{D}}^n_\Delta(\hat{P}_{\boldsymbol{y}}, P) \stackrel{\triangle}{=} \min_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{xy}} d(X,Y) \leq \Delta\}} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P), \tag{18}$$

where $E_{\boldsymbol{xy}}$ denotes the *empirical expectation*, defined as

$$E_{\boldsymbol{xy}} d(X,Y) = \sum_{(i,j)\in\mathcal{A}^2} \hat{P}_{\boldsymbol{xy}}(i,j)d(i,j) \tag{19}$$

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

$$\Phi^*(\mathcal{H}_1|\boldsymbol{y}) \stackrel{\triangle}{=} \exp\left\{-n\left[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}} \| P_0)\right]_+\right\}. \tag{20}$$

When $n \to \infty$, $\tilde{\mathcal{D}}^n_\Delta$ becomes (due to the density of rational numbers on the real line, the admissibility set in Equation (18) is dense in that of Equation (21); since the divergence functional is continuous, the sequence $\{\tilde{\mathcal{D}}^n_\Delta(\hat{P}_{\boldsymbol{y}}, P)\}_{n\geq 1}$ tends to $\tilde{\mathcal{D}}_\Delta(P_Y, P)$ as $n \to \infty$)

$$\tilde{\mathcal{D}}_\Delta(P_Y, P) \stackrel{\triangle}{=} \min_{\{P_{X|Y}: E_{XY} d(X,Y) \leq \Delta\}} \mathcal{D}(P_X \| P), \tag{21}$$

where E_{XY} denotes expectation with respect to P_{XY}.

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

The function $\tilde{\mathcal{D}}_\Delta$ has 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; and (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 consequence of the following property, which is useful for proving some of the results in the sequel (in particular, Theorems 3, 7 and 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 A.2).

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,*

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

In Equation (22), we made explicit the dependence on the parameter λ in the notation of the error exponent, since this is useful in the sequel.

The proof of Theorem 3, which appears in Appendix A.3, is based on Sanov's theorem [29,30], by exploiting the compactness of the set $\{P_Y : \tilde{D}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$.

From Theorem 3 it follows that $\varepsilon_{\text{FN}}(\lambda) = 0$ whenever there exists a PMF P_Y inside the set $\{P_Y : \tilde{D}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$ with Δ_1-limited expected distortion from P_1. When this condition does not hold, $P_{\text{FN}}(\Phi^*, A^*_{\Delta_1}) \to 0$ exponentially rapidly.

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

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

It can be shown that the error exponent in Equation (23) is the same as the error exponent of Theorem 2 in [11] (and Theorem 2 in [31]), where deterministic strategies are considered for both the defender and the attacker. Such equivalence can be explained as follows. As already pointed, the optimal defense strategy in Equation (17) and the deterministic rule found in [11] 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 optimal attack could be derived as in [11], that is, by considering the best response to the dominant defense strategy in [11]. Such attack consists of minimizing the divergence with respect to P_0, namely $\mathcal{D}(\hat{P}_{\mathbf{y}} || P_0)$, 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 [11]. The main difference between the setup in [11] and the more general one addressed in this paper relies 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 A.4.

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

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

Comparing Equations (23) and (24) 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 Δ_1).

When d is not a metric, Equation (24) is only an upper bound on $\varepsilon_{\text{FN}}(\lambda)$, as can be seen from the proof of Theorem 4. Accordingly, in the general case (d is not a metric), applying distortion levels Δ_0 and Δ_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).

5. 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 defense strategy is found by assuming that the strategy played by the attacker, namely the optimal pair of channels (A^*_0, A^*_1) of Theorem 1, is known to the defender, that is, by exploiting the rationalizability argument (see Section 2.1). Accordingly, the resulting optimal strategy is not dominant, thus the associated equilibrium is weaker compared to that of the Neyman–Pearson game.

5.1. Optimal Defense and Game Equilibrium

Since the payoff in Equation (3) is a special case of Equation (5) with $\gamma = 1$ and $\beta = e^{an}$, for any defense strategy $\Phi \in \mathcal{S}_D$, the asymptotically optimal 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 defense 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 B.1.

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

Then,

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

is the optimal defense strategy.

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

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

is asymptotically optimal.

It is useful to provide the asymptotically optimal 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 $Q_1(\boldsymbol{y})$ and $Q_0(\boldsymbol{y})$, the strategy $\Phi^{\dagger}$, which leads to the same payoff asymptotically, is easier to implement because of its single-letter form.

According to the asymptotically optimal defense strategy defined by the above theorem, given the observed sequence $\boldsymbol{y}$, the decision is made by comparing the value of the difference $\tilde{\mathcal{D}}_{\Delta_0}^n(\hat{P}_{\boldsymbol{y}}|P_0) - \tilde{\mathcal{D}}_{\Delta_1}^n(\hat{P}_{\boldsymbol{y}}|P_1)$ against the threshold. Then, the functions $\tilde{\mathcal{D}}_{\Delta_0}^n$ and $\tilde{\mathcal{D}}_{\Delta_1}^n$ take the role of the K-L divergence functions when the likelihood ratio test is performed in the non-adversarial case [19].

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 defense strategies defined, respectively, in Equations (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 Equation (5), i.e., for any $\gamma, \beta \geq 0$.

Finally, we observe 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 Neymam–Pearson game the defender does not care about the value of the FP probability provided that its exponent is larger than λ, 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.

5.2. 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 optimal defense 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*

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

The proof appears in Appendix B.2.

According to Theorem 7, the asymptotic exponent of u is zero if there exists a PMF P^*_Y with Δ_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 λ 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 Equation (27) is smaller than or equal to the one in Equation (22), where equality holds only when both Equations (22) and (27) 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 λ 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 λ; in this way, the ratio between FP and FN error exponent at the equilibrium is generally smaller than λ (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.

6. Source Distinguishability

In this section, we investigate the performance of the Neyman–Pearson and Bayesian games as a function of λ and a, respectively. From the expressions of the equilibrium payoff exponents, it is clear that increase (and then the equilibrium payoffs of the games decrease) as λ and a decrease, respectively. In particular, by letting $\lambda=0$ and $a=0$, we obtain the largest achievable exponents 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 (respectively, Bayesian) setting, if there exists a value of λ (respectively, a) 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 Δ_0 and Δ_1. Although each game has a different asymptotic behavior, we show 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 [19] to the adversarial setup (see [31]).

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* Δ_0 *and* Δ_1*, the maximum achievable FN exponent for the Neyman–Pearson game is:*

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

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

The theorem is an immediate consequence of the continuity of $\varepsilon_{\text{FN}}(\lambda)$ as $\lambda \to 0^+$, 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 $\lambda \to 0^+$ (It holds true from Property 1).

We notice that, if $\Delta_0 = \Delta_1 = 0$, there is only an admissible point in the set in Equation (28), for which $P_Y = P_0$; then, $\varepsilon_{\text{FN}}(0) = \mathcal{D}(P_0||P_1)$, which corresponds to the best achievable FN exponent known from the classical literature for the non-adversarial case (Stein lemma [19] (Theorem 11.8.3)).

Regarding the Bayesian setting, we have the following theorem, the proof of which appears in Appendix C.1.

Theorem 9. *Given two memoryless sources P_0 and P_1 and distortion levels Δ_0 and Δ_1, the maximum achievable exponent of the equilibrium Bayes payoff is*

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

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, respectively P_0 (the fact that $\tilde{\mathcal{D}}_{\Delta_0}$ ($\tilde{\mathcal{D}}_{\Delta_1}$) is 0 in a Δ_0-limited (Δ_1-limited) neighborhood of P_0 (P_1), and not just in P_0 (P_1), does not affect the argument), the minimum over P_Y of the maximum between these quantities (right-hand side of Equation (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 [19] (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 [19]).

From Theorems 8 and 9, it follows that there is no positive λ, respectively 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:

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

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:

$$\min_{Q_{XY}:\substack{(Q_{XY})_X=P_X\\(Q_{XY})_Y=P_Y}} E_{XY}d(X,Y) \leq \Delta, \tag{31}$$

where the expectation E_{XY} is with respect to Q_{XY}. For ease of notation, in Equation (31), we used $(Q_{XY})_Y = P_Y$ (respectively $(Q_{XY})_X = P_X$) as short for $\sum_x Q_{XY}(x,y) = P_Y(y)$, $\forall y \in \mathcal{A}$ (respectively $\sum_y Q_{XY}(x,y) = P_X(x)$, $\forall x \in \mathcal{A}$), where Q_{XY} is a joint PMF and P_X and P_Y are its marginal PMFs.

In computer vision applications, the left-hand side of Equation (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, $EMD_d(P_Y, P_X)$) [32]. It is also known as the ρ-bar distortion measure [33].

A brief comment concerning the analogy between the minimization in Equation (31) and *optimal transport theory* is worth expounding. The minimization problem in Equation (31) is known in the Operations Research literature as *Hitchcock Transportation Problem* (TP) [34]. Referring to the original Monge formulation of this problem [35], 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 parallel can be found in [31].

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 Δ_0 and Δ_1, the set of the PMFs that cannot be distinguished from P_0 in both the Neyman–Pearson and Bayesian settings is given by*

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

Set Γ is the indistinguishability region. By definition (see the beginning of this section), the PMFs inside Γ 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 optimal 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 C.2.

Corollary 2 (Corollary to Theorems 8 and 9)**.** *When d is a metric, given the source P_0 and distortion levels Δ_0 and Δ_1, for any fixed P, the minimum in Equation (32) is achieved when*

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

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

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

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 Equation (34) is sufficient in order for P_0 and P be indistinguishable, that is $\Gamma \supseteq \{P : EMD_d(P_0, P) \leq \Delta_0 + \Delta_1\}$ (see discussion in Appendix C.2, 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., $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} |x_i - y_i|^p$, 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 Γ can be bounded as follows*

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

Corollary 3 can be proven by exploiting the Hölder inequality [36] (see Appendix C.3).

7. Concluding Remarks

We consider 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 [11], 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 optimal 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 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, such as Markov sources. As mentioned in the Introduction, it would be also relevant to extend the analysis to higher order statistics. In fact, the techniques introduced in this paper are very general and lend themselves to extensions that allow richer sets of sufficient statistics (as currently pursued in an ongoing work of the first two coauthors). This generalization, however, comes at the price of an increased complexity of the analysis with regard to both the expressions of the equilibrium strategies and the performance achieved by the two parties at the equilibrium. More specifically, given the set of statistics used by the defender, it turns out that the optimal attacker also uses the very same statistics (plus the empirical correlation with the source sequence, due to the distortion constraint) and the optimal attack channel is given in terms of uniform distributions across conditional types that depend on these set of statistics. Therefore, in a sequence of games, where the defender exploits more and more statistics, so would the attacker, and it would be interesting to explore the limiting behaviour of this process.

We also mention the case of unknown sources, where the source statistics 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 [13,14]. The extension of such analyses to the fully active scenario represents a further interesting direction for future research.

Finally, we stress that, in this paper, as well as in virtually all prior works using game theory to model the interplay between the attacker and the defender, it is assumed that the attacker is always present. This is a pessimistic or worst-case assumption, leading to the adoption of an overly conservative defense strategy (when the attacker is not present, in fact, better performance could be in principle achieved by adopting a non-adversarial detection strategy). A more far-reaching extension would require that the ubiquitous presence of the attacker is reconsidered, for instance by resorting to a two-steps analysis, where the presence or absence of the attacker is established in the first step, or by defining and studying more complex game models, e.g., games with incomplete information [37].

Author Contributions: All the authors have contributed equally to this research work and the writing of this paper.

Funding: This research received no external funding.

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

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Neyman–Pearson Detection Game

This appendix contains the proofs of the results in Section 4.

Appendix A.1. Proof of Lemma 1

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

$$\min_{\Phi \in \mathcal{S}_D} P_{\text{FN}}(\Phi, A_1), \tag{A1}$$

for any attack channel A_1. Below, we first show that $P_{\text{FN}}(\Phi^*, A_1) \dot{\leq} P_{\text{FN}}(\Phi, A_1)$ for every $\Phi \in \mathcal{S}_D$ and for every A_1, that is, Φ^* 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 Φ^* is also admissible. Therefore, we can conclude that $\Phi^*(\cdot|\boldsymbol{y})$ asymptotically solves Equation (A1). Exploiting the memorylessness of P_0 and the permutation invariance of $\Phi(\mathcal{H}_1|\boldsymbol{y})$ and $d(\boldsymbol{x}, \boldsymbol{y})$, for every $\boldsymbol{y}' \in \mathcal{A}^n$ we have,

$$
\begin{aligned}
e^{-\lambda n} \geq & \max_{A} \sum_{\boldsymbol{x},\boldsymbol{y}} P_0(\boldsymbol{x}) A(\boldsymbol{y}|\boldsymbol{x}) \Phi(\mathcal{H}_1|\boldsymbol{y}) \\
\geq & \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) A^*_{\Delta_0}(\boldsymbol{y}|\boldsymbol{x}) \right) \Phi(\mathcal{H}_1|\boldsymbol{y}) \\
= & \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta_0} P_0(\boldsymbol{x}) \cdot \frac{c_n(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \right) \Phi(\mathcal{H}_1|\boldsymbol{y}) \\
\geq & (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta_0} \cdot \frac{P_0(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \right) \Phi(\mathcal{H}_1|\boldsymbol{y}) \qquad \text{(A2)} \\
\overset{(a)}{\geq} & (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} |\mathcal{T}(\boldsymbol{y}')| \left(\max_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}') \leq n\Delta_0} |\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}')| \cdot \frac{P_0(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}'|\boldsymbol{x})|} \right) \Phi(\mathcal{H}_1|\boldsymbol{y}') \\
\overset{(b)}{=} & (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \Phi(\mathcal{H}_1|\boldsymbol{y}') \max_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}') \leq n\Delta_0} P_0(\boldsymbol{x}) \cdot |\mathcal{T}(\boldsymbol{x})| \\
\geq & \Phi(\mathcal{H}_1|\boldsymbol{y}') \max_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}') \leq n\Delta_0} \frac{e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0)}}{(n+1)^{|\mathcal{A}|^2 \cdot (|\mathcal{A}|-1)}} \\
= & \Phi(\mathcal{H}_1|\boldsymbol{y}') \frac{\exp\left\{ -n \min_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}') \leq n\Delta_0} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0) \right\}}{(n+1)^{|\mathcal{A}|^2 \cdot (|\mathcal{A}|-1)}},
\end{aligned}
$$

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

It now follows that

$$
\Phi(\mathcal{H}_1|\boldsymbol{y}) \dot{\leq} \exp\left\{ -n \left[\lambda - \min_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0) \right] \right\}. \qquad \text{(A3)}
$$

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

$$
\begin{aligned}
\Phi(\mathcal{H}_1|\boldsymbol{y}) \dot{\leq} & \min\left\{ 1, \exp\left[-n \left(\lambda - \min_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \leq n\Delta_0} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_0) \right) \right] \right\} \\
= & \Phi^*(\mathcal{H}_1|\boldsymbol{y}). \qquad \text{(A4)}
\end{aligned}
$$

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

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

so that $\Phi^*(\mathcal{H}_1|\boldsymbol{y}) = \min\{1, e^{-n \cdot k_n(\boldsymbol{y})}\}$. We now show that $\Phi^*(\mathcal{H}_1|\boldsymbol{y})$ 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) &\le (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FP}}(\Phi^*, A^*) \\
&= (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \sum_{\boldsymbol{x},\boldsymbol{y}} P_0(\boldsymbol{x}) A^*_{\Delta_0}(\boldsymbol{y}|\boldsymbol{x}) \Phi^*(\mathcal{H}_1|\boldsymbol{y}) \\
&= (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \sum_{(\boldsymbol{x},\boldsymbol{y}):d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} P_0(\boldsymbol{x}) \cdot \frac{c_n(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \cdot \Phi^*(\mathcal{H}_1|\boldsymbol{y}) \\
&\le (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \sum_{(\boldsymbol{x},\boldsymbol{y}):d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} \frac{P_0(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \cdot \Phi^*(\mathcal{H}_1|\boldsymbol{y}) \\
&\le (n+1)^{2|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \sum_{\boldsymbol{y}} \left(\max_{\boldsymbol{x}:d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} |\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})| \cdot \frac{P_0(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \right) \Phi^*(\mathcal{H}_1|\boldsymbol{y}) \qquad \text{(A5)} \\
&= (n+1)^{2|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \left(\sum_{\hat{P}_{\boldsymbol{y}}:k_n(\boldsymbol{y})\ge 0} e^{-nk_n(\boldsymbol{y})} \left(\max_{\boldsymbol{x}:d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} |\mathcal{T}(\boldsymbol{x})| \cdot P_0(\boldsymbol{x}) \right) + \right. \\
&\qquad \left. + \sum_{\hat{P}_{\boldsymbol{y}}:k_n(\boldsymbol{y})<0} \left(\max_{\boldsymbol{x}:d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} |\mathcal{T}(\boldsymbol{x})| \cdot P_0(\boldsymbol{x}) \right) \right) \\
&\le (n+1)^{2|\mathcal{A}|\cdot(|\mathcal{A}|-1)} \left(\sum_{\hat{P}_{\boldsymbol{y}}:k_n(\boldsymbol{y})\ge 0} e^{-n\lambda} + \right. \\
&\qquad \left. + \sum_{\hat{P}_{\boldsymbol{y}}:k_n(\boldsymbol{y})<0} \exp\left\{ -n \min_{\boldsymbol{x}:d(\boldsymbol{x},\boldsymbol{y})\le n\Delta_0} \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0) \right\} \right) \\
&\le (n+1)^{|\mathcal{A}|(|\mathcal{A}|^2+|\mathcal{A}|-1)} e^{-n\lambda}.
\end{aligned}$$

Appendix A.2. Proof of Property 1

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

$$\tilde{\mathcal{D}}_\Delta(\lambda P_{Y_1} + (1-\lambda)P_{Y_2}, P) \le \lambda \tilde{\mathcal{D}}_\Delta(P_{Y_1}, P) + (1-\lambda)\tilde{\mathcal{D}}_\Delta(P_{Y_2}, P). \qquad \text{(A6)}$$

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

$$\tilde{\mathcal{D}}_\Delta(P_Y, P) \stackrel{\triangle}{=} \min_{\{Q_{XY}: E_{XY}d(X,Y)\le\Delta, (Q_{XY})_Y = P_Y\}} \mathcal{D}\big((Q_{XY})_X \| P\big), \qquad \text{(A7)}$$

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:

$$\tilde{\mathcal{D}}_\Delta(\lambda P_{Y_1} + (1-\lambda)P_{Y_2}, P) = \min_{\{Q_{XY}: E_{XY}d(X,Y)\le\Delta,\ (Q_{XY})_Y = \lambda P_{Y_1} + (1-\lambda)P_{Y_2}\}} \mathcal{D}\big((Q_{XY})_X \| P\big). \qquad \text{(A8)}$$

We find it convenient to rewrite the right-hand side of Equation (A8) by minimizing over pairs of PMFs (Q'_{XY}, Q''_{XY}) and considering the convex combination of these PMFs with weights λ and $(1-\lambda)$, in place of Q_{XY}; hence,

$$\tilde{\mathcal{D}}_\Delta(\lambda P_{Y_1} + (1-\lambda)P_{Y_2}, P) = \min_{(Q'_{XY}, Q''_{XY})\in\mathcal{H}} \mathcal{D}\big(\lambda (Q'_{XY})_X + (1-\lambda)(Q''_{XY})_X \| P\big), \qquad \text{(A9)}$$

where

$$\mathcal{H} = \big\{(Q'_{XY}, Q''_{XY}) : \lambda(Q'_{XY})_Y + (1-\lambda)(Q''_{XY})_Y = \lambda P_{Y_1} + (1-\lambda)P_{Y_2}, \\ \lambda E'_{XY} d(X,Y) + (1-\lambda) E''_{XY} d(X,Y) \leq \Delta\big\}. \quad \text{(A10)}$$

Let

$$\mathcal{H}' = \big\{Q'_{XY} : E''_{XY} d(X,Y) \leq \Delta, (Q'_{XY})_Y = P_{Y_1}\big\} \times \big\{Q''_{XY} : E''_{XY} d(X,Y) \leq \Delta, (Q''_{XY})_Y = P_{Y_2}\big\}; \quad \text{(A11)}$$

then, $\mathcal{H}' \subset \mathcal{H}$, where the set $\mathcal{H}'$ is separable in Q'_{XY} and Q'_{XY}. Accordingly, Equations (A9) and (A10) can be upper bounded by

$$\min_{Q'_{XY}: E''_{XY} d(X,Y) \leq \Delta, (Q'_{XY})_Y = P_{Y_1}} \; \min_{Q''_{XY}: E''_{XY} d(X,Y) \leq \Delta, (Q''_{XY})_Y = P_{Y_2}} \mathcal{D}\big(\lambda(Q'_{XY})_X + (1-\lambda)(Q''_{XY})_X \| P\big). \quad \text{(A12)}$$

By the convexity of $\mathcal{D}\big((Q_{XY})_X \| P\big)$ with respect to Q_{XY} (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, [19])), it follows that

$$\mathcal{D}\big(\lambda(Q'_{XY})_X + (1-\lambda)(Q''_{XY})_X \| P\big) \leq \lambda \mathcal{D}\big((Q'_{XY})_X \| P\big) + (1-\lambda)\mathcal{D}\big((Q''_{XY})_X \| P\big). \quad \text{(A13)}$$

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

$$\min_{Q'_{XY}: \sum_x Q'_{XY} = P_{Y_1}, E''_{XY} d(X,Y) \leq \Delta} \lambda \mathcal{D}\big((Q'_{XY})_X \| P\big) + \min_{Q''_{XY}: (Q''_{XY})_Y = P_{Y_2}, E''_{XY} d(X,Y) \leq \Delta} (1-\lambda)\mathcal{D}\big((Q''_{XY})_X \| P\big), \quad \text{(A14)}$$

thus proving Equation (A6).

Appendix A.3. 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_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{x}) A^*_{\Delta_1}(\boldsymbol{y}|\boldsymbol{x}) \Phi^*(\mathcal{H}_0|\boldsymbol{y}) \\
&= \sum_{\boldsymbol{y}} \sum_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta_1} P_1(\boldsymbol{x}) \frac{c_n(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \\
&\le \sum_{\boldsymbol{y}} \sum_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta_1} \frac{P_1(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \\
&= \sum_{\boldsymbol{y}} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} |\mathcal{T}(\hat{P}_{\boldsymbol{x}|\boldsymbol{y}})| \frac{e^{-n[\hat{H}_{\boldsymbol{x}}(X) + \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)]}}{|\mathcal{T}(\hat{P}_{\boldsymbol{y}|\boldsymbol{x}})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \\
&= \sum_{\hat{P}_{\boldsymbol{y}}} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} |\mathcal{T}(\hat{P}_{\boldsymbol{x}})| e^{-n[\hat{H}_{\boldsymbol{x}}(X) + \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)]} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \\
&= \sum_{\hat{P}_{\boldsymbol{y}}} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \qquad \text{(A15)} \\
&= \sum_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) < \lambda} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0))}\right) \\
&\le \sum_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) < \lambda} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} \\
&\le (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp\left\{ -n \min_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) < \lambda} \left[\min_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1) \right] \right\} \\
&\le (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp\left\{ -n \inf_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) < \lambda} \left[\min_{P_{X|Y}: E_{XY} d(X,Y) \le \Delta_1} \mathcal{D}(P_X \| P_1) \right] \right\} \\
&\le (n+1)^{2|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \exp\left\{ -n \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \le \lambda} \left[\min_{P_{X|Y}: E_{XY} d(X,Y) \le \Delta_1} \mathcal{D}(P_X \| P_1) \right] \right\}.
\end{aligned}$$

Then:

$$\limsup_{n \to \infty} \frac{1}{n} \ln P_{\text{FN}}(\Phi^*, A^*_{\Delta_1}) \le - \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \le \lambda} \left[\min_{P_{X|Y}: E_{XY} d(X,Y) \le \Delta_1} \mathcal{D}(P_X \| P_1) \right]. \qquad \text{(A16)}$$

We now move on to the lower bound.

$$\begin{aligned}
P_{\text{FN}}(\Phi^*, A^*_{\Delta_1}) &= \sum_{\boldsymbol{x},\boldsymbol{y}} P_1(\boldsymbol{x}) A^*_{\Delta_1}(\boldsymbol{y}|\boldsymbol{x}) \Phi^*(\mathcal{H}_1|\boldsymbol{y}) \\
&\ge (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\boldsymbol{y}} \sum_{\boldsymbol{x}: d(\boldsymbol{x},\boldsymbol{y}) \le n\Delta_1} \frac{P_1(\boldsymbol{x})}{|\mathcal{T}(\boldsymbol{y}|\boldsymbol{x})|} \left(1 - e^{-n[\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]_+}\right) \qquad \text{(A17)} \\
&= (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} \sum_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) < \lambda} \sum_{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}: E_{\boldsymbol{x}\boldsymbol{y}} d(X,Y) \le \Delta_1} e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0))}\right) \\
&\ge (n+1)^{-|\mathcal{A}| \cdot (|\mathcal{A}|-1)} e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} (1 - e^{-n(\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0))}),
\end{aligned}$$

where, for a fixed n, $\hat{P}_{\boldsymbol{y}}$ is a PMF that satisfies $\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) \le \lambda - (\ln n)/n$ and $\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}$ 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 $(\hat{P}_{\boldsymbol{y}}, \hat{P}_{\boldsymbol{x}|\boldsymbol{y}}) \to (P^*_Y, P^*_{X|Y})$ (we are implicitly

exploiting the fact that set $\{\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) < \lambda\}$ is dense in $\{\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda\}$, for every $\lambda > 0$, which holds true from Property 1), where

$$(P_Y^*, P_{X|Y}^*) = \arg\min_{(P_Y, P_{X|Y})} \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda} \left[\min_{P_{X|Y} : E_{XY} d(X,Y) \leq \Delta_1} \mathcal{D}(P_X \| P_1) \right]. \tag{A18}$$

Therefore, we can assert that:

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \ln P_{\text{FN}}(\Phi^*, A^*_{\Delta_1}) &\geq \lim_{n\to\infty} \frac{1}{n} \ln \left[e^{-n\mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1)} \left(1 - e^{-n(\lambda - \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0))} \right) \right] \\ &= - \lim_{n\to\infty} \mathcal{D}(\hat{P}_{\boldsymbol{x}} \| P_1) \\ &= - \mathcal{D}(P_X^* \| P_1) \\ &= - \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y \| P_0) \leq \lambda} \left[\min_{P_{X|Y} : E_{XY} d(X,Y) \leq \Delta_1} \mathcal{D}(P_X \| P_1) \right]. \end{aligned} \tag{A19}$$

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

Appendix A.4. Proof of Theorem 4

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

$$\begin{aligned} \varepsilon_{\text{FN}} &= \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) \leq \lambda} \left(\min_{P_{X|Y} : E_{XY} d(X,Y) \leq \Delta_1} \mathcal{D}(P_X \| P_1) \right) \\ &= \min_{P_Z : \mathcal{D}(P_Z \| P_0) \leq \lambda} \min_{P_{Y|Z} : E_{YZ} d(Y,Z) \leq \Delta_0} \left(\min_{P_{X|Y} : E_{XY} d(X,Y) \leq \Delta_1} \mathcal{D}(P_X \| P_1) \right). \end{aligned} \tag{A20}$$

To prove the theorem, we now show that Equation (A20) can be simplified as follows:

$$\varepsilon_{\text{FN}} = \min_{P_Z : \mathcal{D}(P_Z \| P_0) \leq \lambda} \left(\min_{P_{X|Z} : E_{XZ} d(X,Z) \leq \Delta_0 + \Delta_1} \mathcal{D}(P_X \| P_1) \right), \tag{A21}$$

which is equivalent to Equation (24) (see Section 3). The equivalence of the expressions in Equations (A20) and (A21) follows from the equivalence of the two feasible sets for the PMF P_X. We first show that any feasible P_X in Equation (A20) is also feasible in Equation (A21). Let P_X be a feasible PMF in Equation (A20). 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 Equation (A20),

$$E_{XZ} d(X,Z) \leq E_{XYZ}[d(X,Y) + d(Y,Z)] = E_{XY} d(X,Y) + E_{YZ} d(Y,Z) \leq \Delta_0 + \Delta_1, \tag{A22}$$

and then P_X is a feasible PMF in Equation (A21). 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

$$\begin{aligned} P_{XY}(i,j) &= (1-\alpha) P_{XZ}(i,j) \quad \forall i, \forall j \neq i, \\ P_{XY}(i,i) &= (1-\alpha) P_{XZ}(i,i) + \alpha P_X(i) \quad \forall i. \end{aligned} \tag{A23}$$

By adopting the transportation theory perspective introduced towards the end of Section 6, 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 Equation (A23) is the one which leaves in place a percentage α of the mass and moves the remaining $(1-\alpha)$ percentage to fill the pile $(1-\alpha)P_Z$ according to map $(1-\alpha)P_{XZ}$.

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

$$\begin{aligned} P_{YZ}(i,j) &= \alpha P_{XZ}(i,j) \quad \forall i, \forall j \neq i, \\ P_{YZ}(i,i) &= \alpha P_{XZ}(i,i) + (1-\alpha)P_Z(i) \quad \forall i. \end{aligned} \tag{A24}$$

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 Equation (A21) also belongs to the one in Equation (A20).

Appendix B. Bayesian Detection Game

This appendix contains the proofs for Section 5.

Appendix B.1. Proof of Theorem 5

Given the probability distributions $Q_0(\boldsymbol{y})$ and $Q_1(\boldsymbol{y})$ induced by $A^*_{\Delta_0}$ and $A^*_{\Delta_1}$, respectively, the optimal decision rule is deterministic and is given by the likelihood ratio test (LRT) [38]:

$$\frac{1}{n}\ln\frac{Q_1(\boldsymbol{y})}{Q_0(\boldsymbol{y})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} a, \tag{A25}$$

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

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

$$\begin{aligned} Q_0(\boldsymbol{y}) &= \sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) A^*_{\Delta_0}(\boldsymbol{y}|\boldsymbol{x}) \\ &\doteq \sum_{\boldsymbol{x}:\, d(\boldsymbol{x},\boldsymbol{y})\leq n\Delta_0} e^{-n[\hat{H}_{\boldsymbol{x}}(X)+\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)]} \cdot e^{-n\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(Y|X)} \\ &\doteq \max_{\boldsymbol{x}:\, d(\boldsymbol{x},\boldsymbol{y})\leq n\Delta_0} e^{n\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(X|Y)} \cdot \Big(e^{-n[\hat{H}_{\boldsymbol{x}}(X)+\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)]} \\ &\qquad\qquad \cdot e^{-n\hat{H}_{\boldsymbol{x}\boldsymbol{y}}(Y|X)} \Big) \\ &= \max_{\boldsymbol{x}:\, d(\boldsymbol{x},\boldsymbol{y})\leq n\Delta_0} e^{-n[\hat{H}_{\boldsymbol{y}}(Y)+\mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)]} \\ &\overset{(a)}{=} \exp\Big\{-n\Big[\hat{H}_{\boldsymbol{y}}(Y)+ \\ &\qquad\qquad + \min_{\{\hat{P}_{\boldsymbol{x}|\boldsymbol{y}}:E_{\boldsymbol{x}\boldsymbol{y}}d(X,Y)\leq\Delta_0\}} \mathcal{D}(\hat{P}_{\boldsymbol{x}}\|P_0)\Big]\Big\} \\ &= \exp\Big\{-n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0)]\Big\}, \end{aligned} \tag{A26}$$

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

$$Q_1(\boldsymbol{y}) \doteq \exp\Big\{-n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)]\Big\}. \tag{A27}$$

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

$$\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} a, \tag{A28}$$

which proves the second part of the theorem.

Appendix B.2. Proof of Theorem 7

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 $\boldsymbol{y}$ under $\mathcal{H}_1$, when the attack channel is $A^*_{\Delta_1}$:

$$(n+1)^{-|\mathcal{A}||\mathcal{A}-1|} e^{-n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)]} \leq Q^*_1(\boldsymbol{y}) < (n+1)^{|\mathcal{A}|^2} e^{-n[\hat{H}_{\boldsymbol{y}}(Y) + \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)]}. \tag{A29}$$

The same bounds hold for $Q^*_0(\boldsymbol{y})$, with $\tilde{\mathcal{D}}_{\Delta_0}$ replacing $\tilde{\mathcal{D}}_{\Delta_0}$. For the FN probability, the upper bound is

$$\begin{aligned}
P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1}) =& \sum_{\boldsymbol{y}} Q^*_1(\boldsymbol{y}) \cdot \Phi^\dagger(\mathcal{H}_0|\boldsymbol{y}) \\
=& \sum_{\boldsymbol{y}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} Q^*_1(\boldsymbol{y}) \\
\leq& (n+1)^{|\mathcal{A}|^2} \sum_{\boldsymbol{y}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} e^{-n[\hat{H}_{\boldsymbol{y}} + \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)]} \\
\leq& (n+1)^{|\mathcal{A}|^2 + |\mathcal{A}|} \max_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} e^{-n\tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)} \\
=& (n+1)^{|\mathcal{A}|^2 + |\mathcal{A}|} \exp\left\{ -n \left(\min_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \right) \right\}.
\end{aligned} \tag{A30}$$

Then,

$$-\limsup_{n\to\infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1})) \leq \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \tag{A31}$$

For the lower bound,

$$\begin{aligned}
P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1}) \geq& (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \sum_{\boldsymbol{y}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} e^{-n[\hat{H}_{\boldsymbol{y}} + \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)]} \\
\geq& (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \max_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} e^{-n\tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1)} \\
=& (n+1)^{-|\mathcal{A}||\mathcal{A}-1|} \exp\left\{ -n \left(\min_{\hat{P}_{\boldsymbol{y}}: \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \right) \right\}.
\end{aligned} \tag{A32}$$

Then,

$$\begin{aligned}
-\liminf_{n\to\infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1})) \geq& \lim_{n\to\infty} \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \\
=& \min_{P_Y: \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \leq a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1),
\end{aligned} \tag{A33}$$

where $\hat{P}_{\boldsymbol{y}}$ is a properly chosen PMF, belonging to the set $\{\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a\}$ for every n, and such that $\hat{P}_{\boldsymbol{y}} \to P^*_Y$ where

$$P^*_Y = \arg \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \le a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \tag{A34}$$

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

By combining Equations (A31) and (A33), we get

$$\varepsilon_{\text{FN}} = -\lim_{n\to\infty} \frac{1}{n} \ln(P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1})) = \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \le a} \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1). \tag{A35}$$

Therefore, from Equations (A30) and (A32) we have

$$P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1}) \doteq \exp\left\{-n \left(\min_{\hat{P}_{\boldsymbol{y}} : \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \right) \right\}, \tag{A36}$$

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

$$P_{\text{FP}}(\Phi^*, A^*_{\Delta_0}) \doteq \exp\left\{-n \left(\min_{\hat{P}_{\boldsymbol{y}} : \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \ge a} \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) \right) \right\}, \tag{A37}$$

and in particular

$$\varepsilon_{\text{FP}} = -\lim_{n\to\infty} \frac{1}{n} \ln(P_{\text{FP}}(\Phi^*, A^*_{\Delta_0})) = \min_{P_Y : \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1) \ge a} \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0). \tag{A38}$$

From Equation (A38), 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:

$$\begin{aligned}
u =& P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1}) + e^{an} P_{\text{FP}}(\Phi^\dagger, A^*_{\Delta_0}) \\
\doteq& \max\{P_{\text{FN}}(\Phi^\dagger, A^*_{\Delta_1}), e^{an} P_{\text{FP}}(\Phi^\dagger, A^*_{\Delta_0})\} \\
\doteq& \exp\left\{-n \min\left(\min_{\hat{P}_{\boldsymbol{y}} : \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) < a} \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1), \min_{\hat{P}_{\boldsymbol{y}} : \tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1) \ge a} (\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - a) \right) \right\} \\
=& \exp\left\{-n \min_{P_{\boldsymbol{y}}} \left(\max\left\{ \tilde{\mathcal{D}}^n_{\Delta_1}(\hat{P}_{\boldsymbol{y}}, P_1), (\tilde{\mathcal{D}}^n_{\Delta_0}(\hat{P}_{\boldsymbol{y}}, P_0) - a) \right\} \right) \right\} \\
\doteq& \exp\left\{-n \min_{P_Y} \left(\max\left\{ \tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), (\tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - a) \right\} \right) \right\},
\end{aligned} \tag{A39}$$

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 C. Source Distinguishability

This appendix contains the proofs for Section 6.

Appendix C.1. Proof of Theorem 9

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

$$e_a(P_Y) = \max\left\{\tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0) - a\right\}, \tag{A40}$$

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

$$\begin{aligned}\lim_{a\to 0} \min_{P_Y} e_a(P_Y) &= \min_{P_Y} \lim_{a\to 0} e_a(P_Y) \\ &= \min_{P_Y} \left(\max\left\{\tilde{\mathcal{D}}_{\Delta_1}(P_Y, P_1), \tilde{\mathcal{D}}_{\Delta_0}(P_Y, P_0)\right\}\right),\end{aligned} \tag{A41}$$

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.

Appendix C.2. 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 Equation (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 α such that $EMD_d(P_0, P_Y)$ (or equivalently, by symmetry, $EMD_d(P_Y, P_0)$) is exactly equal to Δ_0.

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

$$\min_{P_Y : EMD_d(P_Y, P_0) \leq \Delta_0} EMD_d(P_Y, P). \tag{A42}$$

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

$$E_{XZ} d(X, Z) \leq E_{XY} d(X, Y) + E_{YZ} d(Y, Z), \tag{A43}$$

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

$$EMD(P_0, P) \leq E_{XY} d(X, Y) + E_{YZ} d(Y, Z). \tag{A44}$$

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

$$EMD(P_0, P) \leq EMD(P_0, P_Y) + EMD(P_Y, P). \tag{A45}$$

From the above relation, we can derive the following lower bound for the to-be-minimized quantity in Equation (A42):

$$\begin{aligned} EMD(P_Y, P) &\geq EMD(P_0, P) - EMD(P_0, P_Y) && \text{(A46)} \\ &\geq EMD(P_0, P) - \Delta_0. && \text{(A47)}\end{aligned}$$

We now show that P_Y defined as in Equation (33) achieves the above lower bound while satisfying the constraint $EMD(P_0, P_Y) \leq \Delta_0$, and then get the minimum value in Equation (A42).

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 (Equations (A23) and (A24)). 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 α in Equation (33) we have that $E_{XY}d(X,Y) = \Delta_0$ and

$$E_{YZ}d(Y,Z) = EMD(P_0,P) - \Delta_0. \tag{A48}$$

By combining Equations (A47) and (A48), we argue that $EMD(P_Y,P) = EMD(P_0,P) - \Delta_0$ (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 Equation (A45), it follows that $EMD(P_0,P_Y) = E_{XY}d(X,Y) = \Delta_0)$. Therefore, P_Y in Equation (33) solves Equation (A42).

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

$$EMD(P_Y,P) = EMD(P_0,P) - \Delta_0 \leq \Delta_1, \tag{A49}$$

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 Equation (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 α as in Equation (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 Equation (34) is contained in the indistinguishability region.

Appendix C.3. Proof of Corollary 3

By inspecting the minimization in Equation (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 $EMD_d(P_Y,P_0) \leq \Delta_0$. To prove the corollary, we need to show that such P lies inside the set defined in Equation (35).

We give the following definition. Given two random variables X and Y, the Hölder inequality applied to the expectation function [36] reads:

$$E_{XY}|XY| \leq \left(E_X[|X|^r]\right)^{1/r}\left(E_Y[|Y|^q]\right)^{1/q}, \tag{A50}$$

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:

$$
\begin{aligned}
EMD_{L_p^p}(P_0, P) &= E^*_{XZ}[||X - Z||^p] \\
&\overset{(a)}{\leq} E^*_{XYZ}[(||X - Y|| + ||Y - Z||)^p] \\
&\overset{(b)}{\leq} E_{XYZ}\Big[||X - Y||^p + ||Y - Z||^p + p \cdot ||X - Y||^{p-1}\ ||Y - Z|| + \\
&\qquad + p(p-1)/2 \cdot ||X - Y||^{p-2}\ ||Y - Z||^2 + ... + p \cdot ||X - Y||\ ||Y - Z||^{p-1}\Big] \\
&= E_{XYZ}[||X - Y||^p] + E_{XYZ}[||Y - Z||^p] + p \cdot E_{XYZ}[||X - Y||^{p-1}\ ||Y - Z||] + \\
&\quad + p(p-1)/2 \cdot E_{XYZ}[||X - Y||^{p-2}\ ||Y - Z||^2] + ... \\
&\qquad ..... + p \cdot E_{XYZ}[||X - Y||\ ||Y - Z||^{p-1}] \\
&\overset{(c)}{\leq} E_{XYZ}[||X - Y||^p] + E_{XYZ}[||Y - Z||^p] + p \cdot (E_{XYZ}[||X - Y||^p])^{\frac{p-1}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{1}{p}} \\
&\quad + p(p-1)/2 \cdot (E_{XYZ}[||X - Y||^p)^{\frac{p-2}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{2}{p}} + ... \\
&\qquad ... + p \cdot (E_{XYZ}[||X - Y||^p])^{\frac{1}{p}} (E_{XYZ}[||Y - Z||^p])^{\frac{p-1}{p}} \\
&= E_{XY}[||X - Y||^p] + E_{YZ}[||Y - Z||^p] + p \cdot (E_{XY}[||X - Y||^p])^{\frac{p-1}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{1}{p}} \\
&\quad + p(p-1)/2 \cdot (E_{XY}[||X - Y||^p])^{\frac{p-2}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{2}{p}} + ... \\
&\qquad ... + p \cdot (E_{XY}[||X - Y||^p])^{\frac{1}{p}} (E_{YZ}[||Y - Z||^p])^{\frac{p-1}{p}} \\
&= \Big((E_{XYZ}[||X - Y||^p])^{1/p} + (E_{XYZ}[||Y - Z||^p])^{1/p}\Big)^p \\
&\leq \Big(\Delta_0^{1/p} + \Delta_1^{1/p}\Big)^p ,
\end{aligned}
\tag{A51}
$$

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 $t = 1, .., p-1$, the Hölder inequality is applied with $r = p/(p-t)$ (and $q = r/(r-1)$).

References

1. Barni, M.; Pérez-González, F. Coping with the enemy: Advances in adversary-aware signal processing. In Proceesings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 8682–8686.
2. Ker, A.D. Batch steganography and the threshold game. In *Security, Steganography, and Watermarking of Multimedia Contents*; SPIE: Bellingham, WA, USA, 2007; p. 650504.
3. Schöttle, P.; Böhme, R. Game Theory and Adaptive Steganography. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 760–773, doi:10.1109/TIFS.2015.2509941. [CrossRef]
4. Cohen, A.; Lapidoth, A. The Gaussian watermarking game. *IEEE Trans. Inf. Theory* **2002**, *48*, 1639–1667, doi:10.1109/TIT.2002.1003844. [CrossRef]
5. Wu, Y.; Wang, B.; Liu, K.R.; Clancy, T.C. Anti-jamming games in multi-channel cognitive radio networks. *IEEE J. Sel. Areas Commun.* **2012**, *30*, 4–15. [CrossRef]
6. Dalvi, N.; Domingos, P.; Mausam, P.; Sanghai, S.; Verma, D. Adversarial classification. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 99–108.
7. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2014; pp. 2672–2680.

8. O'Sullivan, J.; Moulin, P.; Ettinger, J. Information theoretic analysis of steganography. In Proceedings of the IEEE International Symposium on Information Theory, Cambridge, MA, USA, 16–21 August 1998; p. 297, doi:10.1109/ISIT.1998.708902. [CrossRef]
9. Moulin, P.; O'Sullivan, J. Information-theoretic analysis of information hiding. *IEEE Trans. Inf. Theory* **2003**, *49*, 563–593, doi:10.1109/TIT.2002.808134. [CrossRef]
10. Somekh-Baruch, A.; Merhav, N. On the capacity game of public watermarking systems. *IEEE Trans. Inf. Theory* **2004**, *50*, 511–524. [CrossRef]
11. Barni, M.; Tondi, B. The source identification game: An information-theoretic perspective. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 450–463. [CrossRef]
12. Merhav, N.; Sabbag, E. Optimal watermark embedding and detection strategies under limited detection resources. *IEEE Trans. Inf. Theory* **2008**, *54*, 255–274. [CrossRef]
13. Barni, M.; Tondi, B. Binary Hypothesis Testing Game With Training Data. *IEEE Trans. Inf. Theory* **2014**, *60*, 4848–4866, doi:10.1109/TIT.2014.2325571. [CrossRef]
14. Barni, M.; Tondi, B. Adversarial Source Identification Game With Corrupted Training. *IEEE Trans. Inf. Theory* **2018**, *64*, 3894–3915, doi:10.1109/TIT.2018.2806742. [CrossRef]
15. Moulin, P.; Ivanovic, A. Game-theoretic analysis of watermark detection. In Proceedings of the International Conference on Image Processing, Thessaloniki, Greece, 7–10 October 2001; Volume 3, pp. 975–978.
16. Vigna, G.; Robertson, W.; Balzarotti, D. Testing network-based intrusion detection signatures using mutant exploits. In Proceedings of the 11th ACM Conference on Computer and Communications Security, Washington, DC, USA, 25–29 October 2004; pp. 21–30.
17. Patton, S.; Yurcik, W.; Doss, D. An Achilles' heel in signature-based IDS: Squealing false positives in SNORT. In Proceedings of the RAID, Davis, CA, USA, 10–12 October 2001; Volume 2001.
18. Mutz, D.; Kruegel, C.; Robertson, W.; Vigna, G.; Kemmerer, R.A. Reverse engineering of network signatures. In Prooceedings of the Auscert Asia Pacific Information Technology Security Conference, Gold Coast, Australia, 22–26 May 2005.
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
20. Cao, G.; Zhao, Y.; Ni, R.; Li, X. Contrast enhancement-based forensics in digital images. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 515–525. [CrossRef]
21. Pan, X.; Zhang, X.; Lyu, S. Exposing image splicing with inconsistent local noise variances. In Proceedigns of the 2012 IEEE International Conference on Computational Photography (ICCP), Seattle, WA, USA, 28–29 April 2012; pp. 1–10.
22. Popescu, A.C.; Farid, H. Statistical Tools for Digital Forensics. In Proceedings of the 6th International Conference on Information Hiding, IH'04, Toronto, ON, Canada, 23–25 May 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 128–147, doi:10.1007/978-3-540-30114-1_10. [CrossRef]
23. Tondi, B.; Barni, M.; Merhav, N. Detection games with a fully active attacker. In Proceedings of the 2015 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2015; pp. 1–6, doi:10.1109/WIFS.2015.7368575. [CrossRef]
24. Osborne, M.J.; Rubinstein, A. *A Course in Game Theory*; MIT Press: Cambridge, UK, 1994.
25. Nash, J. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef] [PubMed]
26. Chen, Y.C.; Van Long, N.; Luo, X. Iterated strict dominance in general games. *Games Econ. Behav.* **2007**, *61*, 299–315. [CrossRef]
27. Bernheim, D. Rationalizable Strategic Behavior. *Econometrica* **1984**, *52*, 1007–1028. [CrossRef]
28. Hoeffding, W. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **1965**, *36*, 369–401. [CrossRef]
29. Sanov, I.N. On the probability of large deviations of random variables. *Math. Sb.* **1957**, *42*, 11–44.
30. Csiszár, I.; Shields, P.C. Information theory and statistics: A tutorial. *Found. Trends Commun. Inf. Theory* **2004**, *1*, 417–528. [CrossRef]
31. Barni, M.; Tondi, B. Source Distinguishability Under Distortion-Limited Attack: An Optimal Transport Perspective. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 2145–2159, doi:10.1109/TIFS.2016.2570739. [CrossRef]
32. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
33. Gray, R.M. *Entropy and Information Theory*; Springer Science & Business Media: Berlin, Germany, 2011.

34. Hitchcock, F.L. The distribution of a product from several sources to numerous localities. *J. Math. Phys.* **1941**, *20*, 224–230. [CrossRef]
35. Monge, G. *Mémoire sur la Théorie des Déblais et des Remblais*; De l'Imprimerie Royale: Paris, France, 1781.
36. Karr, A.F. *Probability*; Springer: New York, NY, USA, 1993.
37. Osborne, M.J. *An Introduction to Game Theory*; Oxford University Press: New York, NY, USA, 2004; Volume 3.
38. Van Trees, H.L. *Detection, Estimation and Modulation Theory. vol. 2., Nonlinear Modulation Theory*; John Wiley and Sons: New York, NY, USA, 1971.