The Source Identification Game: An Information-Theoretic Perspective

Mauro Barni, *Fellow, IEEE*, and Benedetta Tondi

***Abstract*—We introduce a theoretical framework in which to cast the source identification problem. Thanks to the adoption of a game-theoretic approach, the proposed framework permits us to derive the ultimate achievable performance of the forensic analysis in the presence of an adversary aiming at deceiving it. The asymptotic Nash equilibrium of the source identification game is derived under an assumption on the resources on which the forensic analyst may rely. The payoff at the equilibrium is analyzed, deriving the conditions under which a successful forensic analysis is possible and the error exponent of the false-negative error probability in such a case. The difficulty of deriving a closed-form solution for general instances of the game is alleviated by the introduction of an efficient numerical procedure for the derivation of the optimum attacking strategy. The numerical analysis is applied to a case study to show the kind of information it can provide.**

***Index Terms*—Multimedia forensics, source identification, counter-forensics, game theory, hypothesis testing, adversarial signal processing.**

I. Introduction

MULTIMEDIA forensics [1] is a new discipline aiming at collecting evidence about the past history of multimedia documents, including the identification of the source of the document [2], the distinction between computer generated and *real world* documents (e.g., distinction between real and synthetic images [3]), the detection of traces left by the application of certain processing tools like resampling [4] or JPEG compression [5], the modification of the semantic content of the document through cut and paste [6] or copy-move operations [7] and so on. Early works in this field did not consider the presence of an adversary aiming at impeding the forensic analysis; as a result most multimedia forensic techniques do not work properly if some simple countermeasures (collectively referred to as counter-forensic techniques) are taken in order to delete the traces left by the acquisition device or the processing tool that has been used to create the forgery [8]–[12]. In an attempt to re-establish the validity of forensic analysis, researchers have started building new tools to detect the traces left by counter-forensic algorithms, as, for example, in [13]–[15].

It is evident that any attempt to improve the forensic analysis will be accompanied by a dual effort to devise more powerful counter-forensic techniques that leave less and less evidence in the forged documents. While this is an unavoidable and possibly virtuous loop that will finally lead to more powerful forensic and counter-forensic tools, the need to investigate the ultimate limits of forensic (and counter-forensic) analysis clearly exists.

A. Previous Works

Few attempts have been made so far to cast the multimedia forensic problem into a rigorous framework and systematically study the relationship between forensics, counter-forensics and counter counter-forensics. In [12], Böhme and Kirchner cast the forensic problem in a hypothesis testing framework. Several versions of the problem are defined according to the particular hypothesis (or hypotheses) being tested, including distinction between natural and computer generated images, manipulation detection, source identification. Still in [12], counter-forensics is defined as a way to degrade the performance of the hypothesis test envisaged by the forensic analyst. By relying on arguments similar to those used in steganography and steganalysis [16], [17], Böhme and Kirchner argue that the divergence between the probability density functions of the observed signals after the application of the counter-forensic attack is a proper measure to evaluate the reliability of the attack, thus introducing the concept of ε-reliable and perfectly reliable counter-forensics. Noticeably, such measures do not depend on the particular investigation technique adopted by the forensic analyst. Even if Böhme and Kirchner do not explicitly refer to a game-theoretic framework, their attempt to decouple the counterattack from a specific forensic strategy can be seen as a first-implicit-step towards the definition of the equilibrium point of a general multimedia forensics game. The analysis presented in this paper provides a formal game-theoretic framework to cast the above concepts in, clarifying their exact meaning and the conditions under which they hold.

Another work loosely related to the present paper is [18], where the authors introduce a game-theoretic framework to evaluate the effectiveness of a given attacking strategy and derive the optimum countermeasures. As opposed to our analysis, in [18] the attacker's strategy is fixed and the game-theoretic framework is used only to determine the optimum parameters of the forensic analysis and the attack, thus failing to provide a complete characterization of the game between the attacker and the forensic analyst.

Manuscript received October 17, 2012; revised December 17, 2012; accepted December 22, 2012. Date of publication January 01, 2013; date of current version February 05, 2013. This work was supported in part by the REWIND Project, funded by the Future and Emerging Technologies (FET) programme within the 7FP of the EC, under Grant 268478. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

The authors are with the Department of Information Engineering and Mathematical Sciences, University of Siena, 53100 Siena, Italy (e-mail: barni@dii.unisi.it; benedettatondi@libero.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2012.2237397

1556-6013/$31.00 © 2013 IEEE

Finally, we would like to observe that source identification in the presence of an adversary has already been studied, from a more practical point of view and without resorting to game theory, in [13]. The scenario depicted in [13], however, does not fit into the framework introduced in this paper, since the underlying statistical model based on PRNU (Photo Response Nonuniformity) can not be described as a stationary source. On the positive side, the framework introduced in this paper is more general, since it can be applied also to manipulation detection and many other forensic problems. In manipulation detection, for instance, the analyst wants to distinguish between original images following some distribution and manipulated images following a different distribution, while the attacker's goal is to impede manipulation detection by concealing the changes introduced as a consequence of the manipulation under some distortion constraints (see also [12]).

B. Contribution

In this paper, we consider the following problem: let $X \simeq P_X$ be a source of information known to both the Forensic Analyst (FA) and the Adversary (AD). The goal of the FA is to distinguish sequences generated by X from those generated by other sources. Let $Y \simeq P_Y$ be a second source known to the AD (and possibly to the FA), and let $y^n = (y_1, y_2 \dots y_n)$ be a sequence drawn from Y. It is the aim of the AD to transform y^n into a new sequence z^n as close as possible to y^n in such a way that FA believes that z^n has been generated by X.

Given the above scenario, the first contribution of this paper is to propose a rigorous game-theoretic framework that can be used to analyze the source identification problem, and derive the equilibrium point of the game for some simple, yet meaningful cases. Specifically, we show that under certain assumptions on the set of strategies available to the FA, the game admits an asymptotic Nash equilibrium, and derive the optimum strategies for the AD and the FA at the equilibrium. With respect to [12], the use of a game-theoretic framework has the advantage of clarifying the exact conditions under which the divergence can be used as a measure of the effectiveness of the forensic and counter-forensic strategies.

As a second contribution, we analyze the asymptotic behavior of the payoff at the equilibrium. In this way we are able to distinguish the cases in which the FA will succeed from those in which the AD will eventually win the game. This analysis significantly extends the concept of *vulnerability* of a forensic strategy [12], since we prove that, at least in the scenario we are considering, the vulnerability of *any* forensic strategy ultimately depends only on the relationship between the sources X and Y, the allowed attack distortion and the target false alarm error probability.

Finding a closed-form expression for the Nash equilibrium point is possible only in very simple cases, hence the third contribution of the paper is the description of an efficient numerical procedure whereby the asymptotically optimum strategies can be identified and the payoff at the equilibrium evaluated. We then use the numerical analysis to get some insights into the best achievable performance for a close-to-reality case study.

Some of the ideas presented in this paper were already introduced in [19]. With respect to such a work, though, several novelties are introduced. From a theoretical point of view, the most relevant differences are the way the resource limited assumption (see Section III-A) is introduced and the details of the proofs, that correct some imprecisions present in [19]. The part dedicated to multivalued sources and the numerical analysis are also a complete novelty with respect to [19], where the discussion focused on the binary case only. The mathematical machinery used to prove our main results relies heavily on the methods of types [20], [21] and is somewhat similar to the techniques used in [22], with reference to watermarking and [23], with regard to hypothesis testing. Despite the similarities, several differences exist between our work and the analysis carried out in [22]. First of all, the watermarking scenario described in [22] does not refer directly to a game theoretic formulation: as a matter of fact, the analysis in [22] is carried under the assumption that no attack is present or that the attack channel is fixed, and the resort to a min–max optimization is due to the necessity of finding the jointly optimum watermark embedding and detection strategies. Secondly, in [22] no attempt is made to derive the error exponents for the jointly optimum watermarking scheme. In our case, instead, such an analysis plays a major role since it determines the winner of the game under asymptotic conditions.

Finally, we observe that even if the paper focuses on the source identification problem, the same set up can be used to model a much wider category of problems. As a matter of fact, any situation in which an analyst is interested in distinguishing between two hypotheses characterized by different probability distributions, despite the presence of an adversary, can be profitably analyzed by using the framework proposed in this work.

The rest of this paper is organized as follows. In Section II, the notation used throughout the paper is introduced, together with some basic notions of game theory. Section III contains the main results of the paper. First it introduces a rigorous definition of the source identification game, then the equilibrium point of the game is looked for and the behavior of the payoff at the equilibrium is analyzed. Section IV is devoted to the numerical analysis and its application to a simple case-study. The paper ends with Section V, where some conclusions are drawn and directions for future research highlighted.

II. Basic Concepts, Notation, and Definitions

In this section we summarize the notation and definitions used throughout the paper. We also introduce some basic concepts of game theory that will be used to model the source identification problem.

For the rest of this work we will use capital letters to indicate scalar random variables (RVs), whose specific realizations will be represented by the corresponding lower case letters. Random sequences, whose length will be denoted by n, are indicated by X^n. Instantiations of random sequences are indicated by the corresponding lowercase letters, so x^n indicates a specific realization of the random sequence X^n, and X_i, x_i, $i = 1, n$ indicate the i-th element of X^n and x^n respectively. Information sources will also be defined by capital letters. The alphabet of an information source will be indicated by the corresponding calligraphic capital letter (e.g., $\mathcal{X}$). Calligraphic letters will also be used to indicate classes of information sources ($\mathcal{C}$) and classes of probability density functions ($\mathcal{P}$). The probability density function (pdf) of a random variable X will be denoted by P_X. The same notation will be used to indicate the probability measure ruling the emission of sequences from a source X, so we will use the expressions $P_X(a)$ and $P_X(x^n)$ to indicate, respectively, the probability of symbol $a \in \mathcal{X}$ and the probability that the source X emits the sequence x^n, the exact meaning of P_X being always clearly recoverable from the context wherein it is used. Given an event A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$), we will use the notation $P_X(A)$ to indicate the probability of the event A under the probability measure P_X. Given two sequences x^n and y^n, their Hamming distance is defined as the number of locations for which $x_i \neq y_i$, i.e.,

$$d_H(x^n, y^n) = n - \sum_{i=1}^{n} \delta(x_i, y_i), \tag{1}$$

with $\delta(x_i, y_i) = 1$ if $x_i = y_i$ and 0 otherwise.

Throughout the paper we make extensive use of the concepts of type and type class defined as follows (for more insights into the use of type classes in information theory and statistics we refer to [20]). Let x^n be a sequence with elements belonging to an alphabet $\mathcal{X}$. The type P_{x^n} of x^n is the empirical probability distribution induced by the sequence x^n, i.e., $\forall a \in \mathcal{X}, P_{x^n}(a) = \sum_{i=1}^{n} \delta(x_i, a)/n$. In the following we indicate with $\mathcal{P}_n$ the set of types with denominator n, i.e., the set of types induced by sequences of length n. Given $P \in \mathcal{P}_n$, we indicate with $T(P)$ the type class of P, i.e., the set of all the sequences in $\mathcal{X}^n$ having type P.

The Kullback–Leibler (KL) divergence between two distributions P and Q on the same finite alphabet $\mathcal{X}$ is defined as:

$$\mathcal{D}(P\|Q) = \sum_{a \in \mathcal{X}} P(a) \log \frac{P(a)}{Q(a)}, \tag{2}$$

where, according to usual conventions, $0 \log 0 = 0$ and $p \log p/0 = \infty$ if $p > 0$. Empirical distributions can be used to calculate empirical information theoretic quantities, thus the empirical entropy of a sequence will be denoted by:

$$H(P_{x^n}) = -\sum_{a \in \mathcal{X}} P_{x^n}(a) \log P_{x^n}(a). \tag{3}$$

Similar definitions hold for other information theoretic quantities (e.g., KL-divergence and conditional entropy) governed by empirical distributions.

A. Game Theory

Game theory is a branch of mathematics devoted to the analysis of strategic situations, referred to as games, in which the success of one player depends on the choices made by the other players. Traditionally, analysis in game theory aims at finding the equilibrium points of the game, i.e., a set of strategies for the various players of the game such that each player cannot improve his outcome, given the others' strategies. Game theory encompasses a great variety of situations depending, among other things, on the number of players, the way the degree of success of each player is defined, the knowledge that a player has on the strategies adopted by the others, the deterministic or probabilistic nature of the game and so on. In this paper, we are concerned with a rather simple class of games, i.e., the class of strategic, two-player, zero-sum games. In this setup, a game is defined as a 4-uple $G(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, where $\mathcal{S}_1 = \{s_{1,1} \dots s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1} \dots s_{2,n_2}\}$ are the set of strategies (actions) the first and the second player can choose from, and $u_l(s_{1,i}, s_{2,j}), l = 1, 2$ is the payoff of the game for player l, when the first player chooses the strategy $s_{1,i}$ and the second chooses $s_{2,j}$. A pair of strategies $s_{1,i}$ and $s_{2,j}$ is called a profile. In a zero-sum competitive game the two payoff functions are strictly related to each other since for any profile we have $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$. In other words, the win of a player is equal to the loss of the other. In the particular case of a zero-sum game, then, only one payoff function needs to be defined. Without loss of generality we can specify the payoff of the first player (generally indicated by u), with the understanding that the payoff of the second player u_2 is equal to $-u$. In the most common formulation, the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players. In addition, it is assumed that the players choose their strategies before starting the game so that they have no hints about the strategy actually chosen by the other player (strategic game).

Given a game, the determination of the best strategy that each player should follow to maximize its payoff is not an easy task, all the more that a profile that is optimum for both the players may not exist. As we said, 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. While there are many definitions of equilibrium, the most famous and commonly adopted is the one due by Nash [24], [25]. For the particular case of a two-player game, a profile (s_{1,i^*}, s_{2,j^*}) is a Nash equilibrium if:

$$\begin{aligned} u_1((s_{1,i^*}, s_{2,j^*})) &\geq u_1((s_{1,i}, s_{2,j^*})) \; \forall s_{1,i} \in \mathcal{S}_1 \\ u_2((s_{1,i^*}, s_{2,j^*})) &\geq u_2((s_{1,i^*}, s_{2,j})) \; \forall s_{2,j} \in \mathcal{S}_2, \end{aligned} \tag{4}$$

where for a zero-sum game $u_2 = -u_1$. In practice, a profile is a Nash equilibrium if each player does not have any interest in changing its choice assuming the other does not change its strategy. In the rest of this paper we will formulate the source identification game as a zero-sum, competitive game, between the FA and the AD, and we will derive the Nash equilibrium profile for some particular versions of the game.

III. Source Identification With Known Sources

Our definition of the Source Identification (SI) game starts from the observation that the task of the FA is the definition of a test to accept or reject the hypothesis that the sequence under analysis was produced by a certain source X. On the other side, the goal of the AD is to take a sequence generated by a different source and modify it in such a way that the FA accepts the hypothesis that the modified sequence has been generated by X. In doing so the AD may want to minimize the amount of modifications it has to introduce to deceive the FA. In the following, we cast the above ideas into a rigorous framework. To start with, we assume that both the FA and the AD have a full knowledge of the source X. We assume the source Y is also known to both the FA and the AD. This may seem a questionable choice, since in practice it may be difficult for the FA to have full access to the source Y. We will see, however, that, at least asymptotically, the assumption that the FA knows Y can be removed, thus leading to a more realistic model. One may also argue that perfect knowledge of sources X and Y can never be reached in practice, yet we believe that the analysis of even this simplified version of the game can be extremely insightful and open the way to the analysis of more realistic and complex scenarios.

Let, then, $\mathcal{C}$ be a class of information sources with finite alphabet $\mathcal{X}$ (e.g., the class of memoryless sources, or the class of k-order Markov source) and let X and Y be two sources belonging to $\mathcal{C}$. As we said, we assume that the probability measures P_X and P_Y ruling the emission of sequences by X and Y are known to both the FA and the AD.

Let y^n be a sequence drawn from Y and let z^n be a modified version of y^n produced by the AD in the attempt to deceive the FA. Let H_0 be the hypothesis that the test sequence has been generated by X, and let H_1 be the opposite hypothesis that the sequence has been generated by Y. We define the source identification game under the known source assumption (SI_{ks}) as follows.

Definition 1: The $SI_{ks}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a zero-sum, strategic, game played by the FA and the AD, defined by the following strategies and payoff.

- The set of strategies the FA can choose from is the set of acceptance regions for H_0 for which the false positive probability (i.e., the probability of rejecting H_0 when H_0 is true) is below a certain threshold:
$$\mathcal{S}_{FA} = \{\Lambda_0 : P_X(x^n \notin \Lambda_0) \leq P_{fp}\}, \tag{5}$$
where Λ_0 is the acceptance region for H_0 (similarly we indicate with $\Lambda_1 = \Lambda_0^c$ the rejection region for H_0), P_{fp} is a prescribed maximum false positive probability, and where $P_X(x^n \notin \Lambda_0)$ indicates the probability that a sequence generated by X does not belong to Λ_0.
- The set of strategies the AD can choose from is formed by all the functions that map a sequence $y^n \in \mathcal{X}^n$ into a new sequence $z^n \in \mathcal{X}^n$ subject to a distortion constraint:
$$\mathcal{S}_{AD} = \{f(y^n) : d(y^n, f(y^n)) \leq nD\}, \tag{6}$$
where $d(\cdot,\cdot)$ is a proper distance function and D is the maximum allowed average per-letter distortion[1].
- The payoff function is defined as the false negative error probability (P_{fn}), namely:
$$u(\Lambda_0, f) = -P_{fn} = -\sum_{y^n : f(y^n) \in \Lambda_0} P_Y(y^n). \tag{7}$$

A few comments are in order to clarify some of the choices we made to formulate the SI_{ks} game. First of all we decided to limit the strategies available to the AD to deterministic functions of y^n. This may seem a limiting choice, however we will see in Section III-A that, at least asymptotically, the optimum strategy of the FA depends neither on the strategy chosen by the AD nor on P_Y, then, it does not make sense for the AD to adopt a randomized strategy to confuse the FA.

The second comment regards the assumption that the FA knows P_Y. As it is evident from (7), this is a necessary assumption, since for a proper definition of the game it is required that both players have a full knowledge of the payoff for all possible profiles. An alternative possibility could be to define the payoff under a worst case assumption on P_Y, however such a choice has two major drawbacks. First of all, if X and Y belong to the same class of sources $\mathcal{C}$, the worst case for the FA would always be $P_X = P_Y$, a condition under which no meaningful analysis can be made to distinguish sequences drawn from X and Y. One could require that X and Y belong to different source classes, however such classes should have to be known to the FA for a proper definition of the game, thus raising the same concerns raised by the assumption that the FA knows Y. Secondly, adopting a worst case analysis leads to the necessity of differentiating the payoffs of the FA and the AD, since for the FA the worst case corresponds to the highest false negative error probability across all $Y \in \mathcal{C}$, while the AD knows P_Y and hence can compute the actual error probability. This observation would lead to the definition of a noncompetitive version of the game in which two different payoffs are specified for the FA and the AD. While this is an interesting direction to look into, we leave it for future work, all the more that in the sequel of the paper we will focus on the asymptotic solution of the game, for which the optimum strategy of the FA does not depend on P_Y, thus making the assumption that the FA knows P_Y irrelevant.

A. Asymptotically Optimum Strategies for FA With Limited Resources

Solving the SI_{ks} game as stated in Definition 1 is a cumbersome task, hence in this section we focus on the asymptotic optimum strategies that are obtained when the length n of the observed sequence tends to infinity. In order to make the problem tractable, we also limit the kind of acceptance regions the FA can choose from. We will do so by using an approach similar to that used in [22] to derive the optimum embedding and detection strategies for a general watermarking problem. Specifically, we limit the complexity of the analysis carried out by the FA by confining it to depend on a limited set of statistics computed on the test sequence. To fix the ideas, we assume that the sources X and Y belong to the class of discrete memoryless sources, however our analysis can be extended to other source classes as outlined in Section III-D. Given the memoryless nature of X and Y, it makes sense to require that the FA bases its decision by relying only on P_{x^n}, i.e., on the empirical probability density function induced by the test sequence[2]. Note that, strictly speaking, P_{x^n} is not a sufficient statistics for the FA; in fact, even if Y is a memoryless source, the AD could introduce some memory within the sequence as a result of the application of f. This is the reason why we need to introduce explicitly the requirement that the FA bases its decision only on P_{x^n}.

A fundamental consequence of the limited resources assumption is that it forces Λ_0 to be a union of type classes, i.e., if x^n belongs to Λ_0, then the whole type class of x^n, namely $T(P_{x^n})$, will be contained in Λ_0. Since a type class is univocally defined

[1] While D can be interpreted as the average per-letter distortion, the AD is not obliged to introduce a distortion that is lower than D for each sample of the sequence, since (6) defines only a global constraint.

[2] In order to keep the notation as light as possible, we use the symbol x^n to indicate the test sequence even if, in principle, it is not known whether x^n originates from X or not.

by the empirical probability density function of the sequences contained in it, we can redefine the acceptance region Λ_0 as a union of types $P \in \mathcal{P}_n$, where $\mathcal{P}_n$ is the set of all possible types with denominator n.

With the above ideas in mind we can define the (asymptotic) SI_{ks}^{lr} game as follows.

Definition 2: The $SI_{ks}^{lr}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a game between the FA and the AD defined by the following strategies and payoff:

$$\begin{aligned} \mathcal{S}_{FA} &= \{\Lambda_0 \in 2^{\mathcal{P}_n} : P_{fp} \leq 2^{-\lambda n}\}, &(8)\\ \mathcal{S}_{AD} &= \{f(y^n) : d(y^n, f(y^n)) \leq nD\}, &(9)\\ u(\Lambda_0, f) &= -P_{fn}, &(10) \end{aligned}$$

where in the definition of $\mathcal{S}_{FA}$, $2^{\mathcal{P}_n}$ indicates the power set of $\mathcal{P}_n$, i.e., all the possible unions of types[3]. Note also that we now ask that the false positive error probability decay exponentially fast with n, thus opening the way to the asymptotic solution of the game.

We start our derivation by proving the following lemma.

Lemma 1: Let Λ_1^* be defined as follows:

$$\Lambda_1^* = \left\{ P \in \mathcal{P}_n : \mathcal{D}(P \| P_X) \geq \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}, \quad (11)$$

and let Λ_0^* be the corresponding acceptance region. Then we have:

1) $P_{fp} \leq 2^{-n(\lambda - e_n)}$, with $e_n \to 0$ for $n \to \infty$,
2) for every $\Lambda_0 \in \mathcal{S}_{FA}$ (with $\mathcal{S}_{FA}$ defined as in (8)) we have $\Lambda_1 \subseteq \Lambda_1^*$.

Proof: Since Λ_1^* and Λ_0^* are unions of type classes, $P_{fp}(\Lambda_0^*)$ can be rewritten as

$$P_{fp}(\Lambda_0^*) = \sum_{P \in \Lambda_1^*} P_X(T(P)), \quad (12)$$

where $P_X(T(P))$ indicates the collective probability (under P_X) of all the sequences in $T(P)$. For the class of discrete memoryless sources, the number of types is upper bounded by $(n+1)^{|\mathcal{X}|}$ and the probability of a type class $T(P)$ by $2^{-n\mathcal{D}(P\|P_X)}$ (see [20] chapter 12), hence we have:

$$\begin{aligned} P_{fp}(\Lambda_0^*) &\leq (n+1)^{|\mathcal{X}|} \max_{P \in \Lambda_1^*} P_X(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Lambda_1^*} \mathcal{D}(P\|P_X)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n(\lambda - |\mathcal{X}| \frac{\log(n+1)}{n})} \\ &= 2^{-n(\lambda - 2|\mathcal{X}| \frac{\log(n+1)}{n})}, \quad (13) \end{aligned}$$

proving the first part of the lemma with $e_n = 2|\mathcal{X}|(\log(n+1))/n$ and where the last inequality derives from (11).

We now pass to the second part of the lemma. Let Λ_0 be in $\mathcal{S}_{FA}$ and let P be in Λ_1. Then we have (see [20] chapter 12 for a justification of the last inequality):

$$\begin{aligned} 2^{-\lambda n} &\geq P_X(\Lambda_1) \\ &\geq P_X(T(P)) \\ &\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P\|P_X)}, \quad (14) \end{aligned}$$

that, by taking the logarithm of both sides, proves that indeed $P \in \Lambda_1^*$. ■

The first relation proved in lemma 1 says that, asymptotically, Λ_0^* defines a valid strategy for the FA, while the second one implies the optimality of Λ_1^*. In fact, if for a certain strategy of the AD we have that $P \notin \Lambda_1^*$, *a fortiori* we have that $P \notin \Lambda_1$ for any other choice of Λ_1 hence resulting in a higher false negative error probability.

An interesting consequence of lemma 1 is that the optimum strategy for the FA does not depend on: i) the strategy chosen by the AD, and ii) P_Y, i.e., the optimum strategy is universally optimum across all the probability density functions in $\mathcal{C}$. As we anticipated, this result makes the assumption that the FA knows P_Y unnecessary. In the same way, it is not necessary for the FA to know the probability density function of the attacked sequences.

We also observe that the strategy expressed by (11) has a simple heuristic interpretation: the FA will accept only the sequences whose empirical pdf is close enough (in divergence terms) to the known pdf of the source X.

We now pass to the determination of the optimum strategy for the AD. Since the acceptance region is fixed, the AD can optimize its strategy by assuming that $\Lambda_0 = \Lambda_0^*$. We start by observing that the goal of the AD is to maximize P_{fn}. Such a goal is obtained by trying to bring the sequences produced by Y within Λ_0^*, i.e., by trying to reach the condition:

$$\mathcal{D}(P_{f(y^n)} \| P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n}. \quad (15)$$

In doing so, the AD must only respect the constraint that $d(y^n, f(y^n)) \leq nD$. The optimum strategy for the AD can then be expressed as follows:

$$f^*(y^n) = \arg \min_{z^n : d(z^n, y^n) \leq nD} \mathcal{D}(P_{z^n} \| P_X). \quad (16)$$

Together with lemma 1, the above observation permits to state the first fundamental result of the paper, summarized in the following theorem.

Theorem 1: The profile (Λ_0^*, f^*) defined by lemma 1 and (16) defines an asymptotic Nash equilibrium for the SI_{ks}^{lr} game.

Proof: Adapting (4) to the case at hand yields:

$$\begin{aligned} u(\Lambda_0^*, f^*) &\geq u(\Lambda_0, f^*) \;\forall \Lambda_0 \in \mathcal{S}_{FA} &(17)\\ -u(\Lambda_0^*, f^*) &\geq -u(\Lambda_0^*, f) \;\forall f \in \mathcal{S}_{AD}, &(18) \end{aligned}$$

where the minus sign in the second inequality is due to the fact that in a zero-sum game we have $u_2 = -u_1$. By remembering that for the SI_{ks}^{lr} game $-u$ is the false negative error probability, the inequality (17) derives immediately from lemma 1.

[3] In the rest of the paper we will refer at Λ_0 as a union of sequences or a union of types interchangeably, the two perspectives being equivalent and clearly understandable from the context.

In the same way, since f^* maximizes the false negative error probability given Λ_0^*, the inequality (18) is always verified thus proving the theorem. ■

We conclude this section by observing that even if in the definition of the SI game the payoff is the average false negative error probability, the strategy defined by (16) represents the optimum attack the AD can use for each single sequence. In fact, if the minimization in (16) fails to bring y^n into Λ_0^*, any other attack will also fail.

B. Payoff at the Equilibrium and Error Exponents

The next step is the computation of the payoff at the equilibrium. Given the asymptotic nature of the solution we found, it makes sense to compute the asymptotic behavior of P_{fn} at the equilibrium. From the foregoing discussion it is easy to argue that P_{fn} will either tend to 0 or to 1 for $n \to \infty$ depending on the relationship between the maximum allowed distortion and the KL-divergence between P_X and P_Y. Then, for a more accurate analysis, we will also evaluate the error exponent of the false negative error probability defined as[4]

$$\varepsilon_{fn} = \lim_{n\to\infty} -\frac{\log P_{fn}}{n}. \tag{19}$$

In this framework we are interested in understanding the conditions under which P_{fn} tends to 0, and the value of ε_{fn} in this case[5].

Let Γ_{fn} be the set of sequences generated by Y that can be moved into Λ_0^*. We can write:

$$\Gamma_{fn} = \{y^n : \exists z^n \in \Lambda_0^* \text{ s.t. } d(y^n, z^n) \le nD\}. \tag{20}$$

The false negative error probability is clearly equal to the probability that $y^n \in \Gamma_{fn}$. We start by observing that, under some very general assumptions, Γ_{fn} is still a union of type classes.

Property 1: The set Γ_{fn} defined in (20) is a union of type classes for any permutation-invariant distance-measure.

Proof: Let $y^n \in \Gamma_{fn}$. Then there exists a sequence $z^n \in \Lambda_0^*$ such that $d(y^n, z^n) \le nD$. Let $\tilde{y}^n$ be a generic sequence belonging to the type class of y^n, i.e., $\tilde{y}^n \in T(P_{y^n})$. Then there exists a permutation σ of the elements of the sequence y^n such that $\tilde{y}^n = \sigma(y^n)$. If we apply the same permutation to z^n we obtain a sequence $\tilde{z}^n = \sigma(z^n)$ belonging to $T(P_{z^n})$. Given that Λ_0^* is a union of type classes, $\tilde{z}^n \in \Lambda_0^*$. We now assume that the distance measure used to define the distortion constraint is invariant to a permutation of the sequence elements. This is the case, for instance, of all additive distance measures for which:

$$d(y^n, z^n) = \sum_{i=1}^{n} g(y_i, z_i), \tag{21}$$

for any function $g(\cdot)$. Under this assumption $d(\tilde{y}^n, \tilde{z}^n) = d(y^n, z^n) \le nD$, with $\tilde{z}^n \in \Lambda_0^*$, hence proving that $\tilde{y}^n \in \Gamma_{fn}$. ■

The above property shows that Γ_{fn} is a union of type classes for all the most common distance measures, including L_p distances, the Hamming distance and the max (infinity) distance for which $d(y^n, z^n) = \max_i |y_i - z_i|$. Thanks to property 1, Γ_{fn} can be redefined in terms of types instead than sequences:

$$\Gamma_{fn}^n = \{P \in \mathcal{P}_n : \forall y^n \in T(P), \exists z^n \in \Lambda_0^* \text{ s.t. } d(y^n, z^n) \le nD\} \tag{22}$$

where we have explicitly indicated that we refer to sequences of length n.

We are now ready to investigate the asymptotic behavior of P_{fn}. To this aim, we need to introduce the asymptotic version of Γ_{fn}, defined as:

$$\Gamma_{fn}^\infty = cl\left(\bigcup_n \Gamma_{fn}^n\right), \tag{23}$$

where $cl(S)$ indicates the *closure* of a set S.

At this point we can distinguish two cases: P_Y may either belong to Γ_{fn}^∞ or not. In the former case $\varepsilon_{fn} = 0$, and the FA will not be able to correctly distinguish between original and fake sequences. In the latter case ε_{fn} is strictly positive and the probability that the FA will not distinguish original and fake sequences tends to zero exponentially fast when n increases. In both cases, the error exponent of the false negative error probability is given by the following theorem.

Theorem 2: For the SI_{ks}^{lr} game, the error exponent of the false negative error probability at the equilibrium is given by:

$$\varepsilon_{fn} = \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P \| P_Y), \tag{24}$$

leading to the following cases:

1) $\varepsilon_{fn} = 0$, if $P_Y \in \Gamma_{fn}^\infty$;
2) $\varepsilon_{fn} = \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P \| P_Y)$, if $P_Y \notin \Gamma_{fn}^\infty$.

Proof: The theorem can be seen as a special case of Sanov's theorem (see [20], chapter 12, and [23]), however proving that the hypothesis of Sanov's theorem are satisfied is not a straightforward task due to the complicated expression used to define Γ_{fn}^∞. For this reason in the following we give a complete proof of the theorem. We start by showing that ε_{fn} is lower bounded by the expression in (24). Later we demonstrate that the same expression is also an upper bound for ε_{fn} thus proving the theorem. For any value of n we have

$$\begin{aligned} P_{fn} &= \sum_{P \in \Gamma_{fn}^n} P_Y(T(P)) \\ &\overset{(a)}{\le} \sum_{P \in \Gamma_{fn}^n} 2^{-n\mathcal{D}(P\|P_Y)} \\ &\overset{(b)}{\le} (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Gamma_{fn}^n} \mathcal{D}(P\|P_Y)} \\ &\overset{(c)}{\le} (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P\|P_Y)}, \end{aligned} \tag{25}$$

where inequalities (a) and (b) follow from the properties of types [20], and (c) follows from the fact that $\Gamma_{fn}^n \subseteq \Gamma_{fn}^\infty$. Passing to the error exponent yields:

$$\begin{aligned} \varepsilon_{fn} &\ge -\lim_{n\to\infty} \frac{1}{n} \log(n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P\|P_Y)} \\ &= -\lim_{n\to\infty} |\mathcal{X}| \frac{\log(n+1)}{n} + \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P\|P_Y) \\ &= \min_{P \in \Gamma_{fn}^\infty} \mathcal{D}(P\|P_Y). \end{aligned} \tag{26}$$

[4] Due to (8), the false positive error exponent ε_{fp} is always larger than or equal to λ.

[5] In the same way we could investigate how fast the probability of a correct decision tends to zero when $\varepsilon_{fn} = 0$. Such an analysis goes along the same lines we will use for the computation of ε_{fn} and will not be detailed.

We now prove that the same expression is also an upper bound for ε_{fn}. Let P^* the probability distribution which satisfies (24). Even if P^* does not need to belong to Γ^n_{fn} for any n, in the Appendix we show that if n is large enough we can find a type in Γ^n_{fn} that is arbitrarily close to P^*. Due to the continuity of $\mathcal{D}$, we can then find a sequence of types $Q_n \in \Gamma^n_{fn}$ such that $\mathcal{D}(Q_n\|P_Y) \to \mathcal{D}(P^*\|P_Y)$ when $n \to \infty$. So, for large n, we have:

$$\begin{aligned} P_{fn} &= \sum_{P\in\Gamma^n_{fn}} P_Y(T(P)) \\ &\geq P_Y(T(Q_n)) \\ &\geq \frac{2^{-n\mathcal{D}(Q_n\|P_Y)}}{(n+1)^{|\mathcal{X}|}}, \end{aligned} \tag{27}$$

where the last inequality follows from the properties of types [20]. By passing to the error exponents, we have:

$$\begin{aligned} \varepsilon_{fn} &\leq -\lim_{n\to\infty}\frac{1}{n}\log\frac{2^{-n\mathcal{D}(Q_n\|P_Y)}}{(n+1)^{|\mathcal{X}|}} \\ &= \lim_{n\to\infty}|\mathcal{X}|\frac{\log(n+1)}{n} + \lim_{n\to\infty}\mathcal{D}(Q_n\|P_Y) \\ &= \lim_{n\to\infty}\mathcal{D}(Q_n\|P_Y) \\ &= \mathcal{D}(P^*\|P_Y) = \min_{P\in\Gamma^\infty_{fn}}\mathcal{D}(P\|P_Y). \end{aligned} \tag{28}$$

Combining the two bounds the theorem is proved. ■

The main consequence of Theorem 2 is that, given P_X, D and λ, the set of sources P_Y can be split into two distinct regions, the subset of sources for which the false negative probability tends to zero exponentially fast ($P_Y \in \Gamma^{\infty,c}_{fn}$) and the sources for which, as a consequence of the attack, the false negative probability tends to 1. Stated in another way, Theorem 2 permits to say whether two sources P_X and P_Y are asymptotically distinguishable with a false positive error exponent equal to λ, in the presence of an attack subject to a distortion constraint nD.

A problem with Theorem 2 is that the expression of Γ^∞_{fn}, does not allow an easy computation of the pdf's P_Y for which $P_{fn} \to 0$ and the corresponding error exponents. In Section IV, we show how such a problem can be solved numerically. Here we specialize the expression of Γ^∞_{fn} to the case in which the distortion constraint is expressed in terms of the Hamming distance between y^n and z^n. In this case, in fact, a closed-form expression can be found for Γ^∞_{fn} thus greatly simplifying the analysis. The simplification relies on the following lemma.

Lemma 2: If $d(y^n, z^n) = d_H(y^n, z^n)$, the set Γ^n_{fn} can be expressed as:

$$\begin{aligned} \Gamma^n_{fn} = \Gamma^* = \{P \in \mathcal{P}_n : \exists P' \in \Lambda^*_0 \\ \text{s.t. } \|P - P'\|_{L_1} \leq 2D_H\} \end{aligned} \tag{29}$$

where the L_1 distance between P and P' (sometimes called variational distance) is defined as:

$$d_{L_1}(P, P') = \|P - P'\|_{L_1} = \sum_{a\in\mathcal{X}} |P(a) - P'(a)|. \tag{30}$$

Proof: We start by proving that a sequence whose type has a distance larger than $2D_H$ from all the types in Λ^*_0 can not belong to Γ^n_{fn}. Let y^n and z^n be two sequences, and let P_{y^n} and P_{z^n} be their types. The distance between P_{y^n} and P_{z^n} can be rewritten as follows:

$$\begin{aligned} \|P_{y^n} - P_{z^n}\|_{L_1} &= \sum_{a\in\mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)] \\ &\quad + \sum_{a\in\mathcal{X}^-} [P_{z^n}(a) - P_{y^n}(a)] \\ &= 2\sum_{a\in\mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)], \end{aligned} \tag{31}$$

where $\mathcal{X}^+$ (respectively $\mathcal{X}^-$, $\mathcal{X}^=$) indicates the set of a's for which $P_{y^n}(a) > P_{z^n}(a)$ (respectively $P_{y^n}(a) < P_{z^n}(a)$, $P_{y^n}(a) = P_{z^n}(a)$), and where the last equality follows from the observation that:

$$\sum_{a\in\mathcal{X}^-} P_{y^n}(a) = 1 - \sum_{a\in\mathcal{X}^+} P_{y^n}(a) - \sum_{a\in\mathcal{X}^=} P_{y^n}(a). \tag{32}$$

Let us consider now the Hamming distance between the sequences y^n and z^n. By considering $\mathcal{X}^+$, we see that $d_H(y^n, z^n)$ is larger or equal to $\sum_{a\in\mathcal{X}^+} n[P_{y^n}(a) - P_{z^n}(a)]$. In fact, for each $a \in \mathcal{X}^+$, there must be at least $n[P_{y^n}(a) - P_{z^n}(a)]$ positions in which the sequences y^n and z^n differ, so to justify the presence of $n[P_{y^n}(a) - P_{z^n}(a)]$ more a's in y^n than in z^n, thus yielding:

$$\|P_{y^n} - P_{z^n}\|_{L_1} \leq \frac{2d_H(y^n, z^n)}{n}. \tag{33}$$

For the sequences y^n whose type does not satisfy (29), we have $\|P_{y^n} - P_{z^n}\|_{L_1} > 2D_H \; \forall z^n \in \Lambda^*_0$, yielding

$$2D_H < \|P_{y^n} - P_{z^n}\|_{L_1} \leq \frac{2d_H(y^n, z^n)}{n}, \tag{34}$$

showing that $\Gamma^n_{fn} \subseteq \Gamma^*$.

We now show that $\Gamma^* \subseteq \Gamma^n_{fn}$. Let P be a type in Γ^*. Then there exists a type $P' \in \Lambda^*_0$ whose distance from P is lower than or equal to $2D_H$. Let y^n be a sequence belonging to $T(P)$, the type class of P. Starting form y^n we can easily build a new sequence z^n whose type is equal to P' by proceeding as follows. Let $\mathcal{X}^+$ be the set of a's for which $P_{y^n}(a) > P'(a)$. For each $a \in \mathcal{X}^+$ we take $n[P_{y^n}(a) - P'(a)]$ positions where $y_i = a$, and replace a with a value $b \in \mathcal{X}^-$, in such a way that at the end we have $P_{z^n}(a) = P'(a) \; \forall a \in \mathcal{X}$. Note that this is always possible as we have

$$\sum_{a\in\mathcal{X}^+} [P_{y^n}(a) - P'(a)] = \sum_{b\in\mathcal{X}^-} [P'(b) - P_{y^n}(b)]. \tag{35}$$

Since to pass from y^n to z^n we modified only $\sum_{a\in\mathcal{X}^+} n[P_{y^n}(a) - P'(a)]$ positions of y^n we have:

$$\begin{aligned} d_H(y^n, z^n) &= \sum_{a\in\mathcal{X}^+} n[P_{y^n}(a) - P'(a)] \\ &= \frac{n\|P_{y^n} - P'\|_{L_1}}{2} \\ &\leq nD_H, \end{aligned} \tag{36}$$

showing that $y^n \in \Gamma^n_{fn}$, and hence $\Gamma^* \subseteq \Gamma^n_{fn}$, thus concluding the proof of the lemma. ■

Lemma 2 permits to use a simpler expression for Γ^{∞}_{fn}, that passes from the definition of the asymptotic version of Λ^*_0, as specified in the following:

$$\begin{aligned}\Lambda^{*,\infty}_0 &= \{P \in \mathcal{P} : \mathcal{D}(P\|P_X) < \lambda\}\\ \Lambda^{*,\infty}_1 &= \{P \in \mathcal{P} : \mathcal{D}(P\|P_X) \geq \lambda\}\\ \Gamma^{\infty}_{fn} &= cl\{P \in \mathcal{P} : \exists P' \in \Lambda^{*,\infty}_0\\ &\qquad \text{s.t. } \|P - P'\|_{L_1} \leq 2D_H\}. \qquad (37)\end{aligned}$$

Given the above definition it is easy to restate Theorem 2 by adopting the more convenient expression of Γ^{∞}_{fn}. The proof goes along the same lines followed to prove Theorem 2, and is omitted.

C. Bernoulli Sources

The exact computation of Γ^{∞}_{fn} and the payoff at the equilibrium for the SI^{lr}_{ks} game in a general case is very cumbersome, and depends heavily on the particular relationship between P_X and P_Y. In order to exemplify the general concepts described in the previous section, we now apply them to the case of two Bernoulli sources. For this kind of sequences the Hamming distance is a natural choice to define the distortion constraint, thus permitting to adopt the simplified definition of Γ^{∞}_{fn} given in (37).

Let X and Y be Bernoulli sources with parameters p and q respectively. In this case the acceptance region for H_0 assumes a very simple form. In fact, the KL-divergence between P_{x^n} and P_X depends only on the number of 1's in x^n, the divergence being a monotonic increasing[6] function of $|\nu_x(1) - p|$, where we indicated with $\nu_x(1)$ the relative frequency of 1's in x^n. When seen as an union of types, the acceptance region may be defined in terms of $P(1)$ (the probability of 1 under P) only:

$$\Lambda^*_0 = \{P \in \mathcal{P}_n : P(1) \in (\nu_{inf}(\lambda), \nu_{sup}(\lambda))\}, \qquad (38)$$

with $\nu_{inf}(\lambda)$ and $\nu_{sup}(\lambda)$ derived from the equality

$$\mathcal{D}(P\|P_X) = \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}, \qquad (39)$$

and where we have explicitly indicated the dependence of ν_{inf} and ν_{sup} on λ. Note that in some cases we may have $\nu_{inf} = 0$ and/or $\nu_{sup} = 1$, since (39) may admit a solution only for $P(1) > p$, $P(1) < p$, or no solution at all.

The optimum strategy of the AD is also easy to define. Given the monotonic nature of the KL-divergence, the AD will increase (decrease) the number of 1's in y^n to make the relative frequency of 1's in z^n as close as possible to p. The AD will succeed in inducing a decision error if the relative frequency of ones in z^n belongs to the interval (ν_{inf}, ν_{sup}). Since the distortion constraint states that $d(y^n, z^n) \leq nD_H$, we clearly have:

$$\Gamma^n_{fn} = \{P \in \mathcal{P}_n : P(1) \in (\nu_{inf}(\lambda) - D_H, \nu_{sup}(\lambda) + D_H)\}, \qquad (40)$$

with the boundaries of the interval truncated to 0 or 1 when needed. For the computation of the error exponent of P_{fn} at the equilibrium we first consider the asymptotic version of Λ^*_0 and Γ_{fn}:

$$\Lambda^{*,\infty}_0 = \{P \in \mathcal{P} : P(1) \in (\nu^{\infty}_{inf}(\lambda), \nu^{\infty}_{sup}(\lambda))\} \qquad (41)$$

where ν^{∞}_{inf} and ν^{∞}_{sup} are now derived from the equality $\mathcal{D}(P\|P_X) = \lambda$ and

$$\Gamma^{\infty}_{fn} = \{P \in \mathcal{P} : P(1) \in [\nu^{\infty}_{inf}(\lambda) - D_H, \nu^{\infty}_{sup}(\lambda) + D_H]\}. \qquad (42)$$

As stated by theorem 2, we can distinguish two cases:

$$\begin{aligned} q = P_Y(1) &\in [\nu^{\infty}_{inf}(\lambda) - D_H, \nu^{\infty}_{sup}(\lambda) + D_H]\\ q = P_Y(1) &\notin [\nu^{\infty}_{inf}(\lambda) - D_H, \nu^{\infty}_{sup}(\lambda) + D_H]. \qquad (43)\end{aligned}$$

In the first case $\varepsilon_{fn} = 0$. In the second case P_{fn} tends to 0 for $n \to \infty$ and the error exponent can be computed by resorting to (24) and (37). Let us suppose for instance that $q > \nu^{\infty}_{sup} + D_H$. The type in Γ^{∞}_{fn} closest to P_Y in divergence is a Bernoulli source with parameter $p^* = \nu^{\infty}_{sup} + D_H$, and hence the error exponent will be $\varepsilon_{fn} = \mathcal{D}(\mathcal{B}(p^*)\|\mathcal{B}(q))$.

D. Sources With Memory

The existence of a Nash equilibrium for the SI^{lr}_{ks} game has been proven by assuming that the FA bases its analysis on the empirical pdf of the test sequence. This assumption makes sense for the class of DMS sources whose characteristics are completely described by first order statistics, but is not reasonable for sources with memory. A closer inspection to the methods used in Sections III-A and III-B, however, reveals that the analysis carried out therein can be extended to sources with memory, as long as the concepts of type and type classes can still be used. As a matter of fact, even if the method of types was initially developed to work with memoryless sources [21], it can be extended to more complex models as well. Given a class $\mathcal{C}$ of sources with alphabet $\mathcal{X}$, we say that a partition of $\mathcal{X}^n$ into N_n disjoint sets $T_1, \dots T_{N_n}$, is a partition into type classes if all the sequences in the same T_i are equiprobable for all the sources in $\mathcal{C}$. If the number N_n of type classes grows subexponentially with n, then the method of types can be applied to sources in $\mathcal{C}$, and the analysis we carried out in Sections III-A and III-B can be extended to such sources, if we insist with the limited resources assumption, i.e., if we continue to assume that the FA is restricted to define the acceptance region as a union of type classes. Now it turns out that the concept of types can be applied to some of the most commonly used source models, including Markov sources with finite order and renewal processes.

For Markov sources with finite order, a model that is commonly used to describe a wide variety of sources with memory, it is known that the number of type classes grows polynomially with n [21], hence making the extension of our analysis straightforward. For instance, in this case, the limited resources assumption is equivalent to ask that the FA bases its decision on the empirical transition probabilities induced by x^n plus the pdf of x_1. While the final form of the optimum acceptance region and the minimization problem to be solved by the AD will be

[6] Actually the KL-divergence may have an asymmetric behavior for $n_x(1) < np$ and $n_x(1) > np$ however this asymmetry does not have any impact on our analysis.

much more complicated, their general form will remain essentially the same.

Renewal processes are another class of sources amenable to be analyzed by relying on the concept of types. Given a binary source, let us indicate by $\tau_0, \tau_0+\tau_1, \tau_0+\tau_1+\tau_2\dots$ the positions of the 1's in the sequences produced by the source. The τ_i $(i \geq 1)$ are called interarrival times, and τ_0 initial waiting time. If the τ_i are independent and identically distributed random variables, the output of the source is called a renewal process. In the same way, if the t_i sequence forms a $k-$order Markov chain, the output of the source is called a Markov renewal process of order k. Renewal processes can be used, for instance, to model run length sequences and hence could be of interest in forensic problems dealing with compressed streams adopting run-length coding (e.g., the JPEG coding standard). In [26], it is shown that the number of type classes of renewal processes and Markov renewal processes (of finite order) grows subexponentially with n, thus opening the way to the extension of our analysis to this class of sources.

IV. Numerical Analysis

Finding a closed-form solution for the case of multivalued sources seems a prohibitive task. While the formula defining the optimum acceptance region does not change and can be easily implemented by the FA, the task of the AD is more complex due to the necessity of solving the minimization problem in (16). Such a minimization resembles some instances of the optimal transport problem [27], however here we are interested in minimizing the divergence subject to a distortion constraint, whereas, classically, optimal transport faces with the somewhat-dual problem of minimizing the distortion needed to make a source pdf equal to a target one. In the following we introduce an efficient numerical procedure to determine the optimum strategy of the AD. At first sight, numerical analysis could also seem a difficult problem, since the number of variables involved in the optimization grows with the sequence length n, which in turn needs to be very large due to the asymptotic nature of the analysis. After a closer look, however, the minimization problem can be greatly simplified. In order to show how, let us reformulate it in a more convenient way. Let $n_z(i)$ and $n_y(i)$ be the number of times the symbol i appears, respectively, in z^n and y^n, and let $\nu_z(i)$ and $\nu_y(i)$ be the corresponding relative frequencies ($\nu_z(i) = n_z(i)/n$, $\nu_y(i) = n_y(i)/n$). Let $n(i \to j)$ be the number of times that a symbol i is transformed into a symbol j as a result of the application of $f(\cdot)$[7]. The first constraint that the AD must satisfy when defining $f(\cdot)$ is a structural one, since $f(\cdot)$ can not modify more samples than there are in the sequence. Specifically we have:

$$n(i \to i) + \sum_{k \neq i} n(i \to k) = \sum_k n(i \to k) = n_y(i). \tag{44}$$

The second constraint is the distortion constraint, that can be reformulated in terms of $n(i \to j)$ as follows

$$d(z^n, y^n) = \sum_{i,j} n(i \to j) g(i,j) \leq nD, \tag{45}$$

where we have assumed that an additive metric is used to compute the distance between z^n and y^n, e.g., if the popular L_p norm is used, we would have $g(i,j) = |i-j|^p$. We now must express the objective function of the minimization in (16) as a function of $n(i \to j)$. To do so we observe that:

$$n_z(i) = n(i \to i) + \sum_{k \neq i} n(k \to i) = \sum_k n(k \to i). \tag{46}$$

By observing that

$$\mathcal{D}(P_{z^n} \| P_X) = \sum_{j=1}^{|\mathcal{X}|} \nu_z(j) \log \frac{\nu_z(j)}{P_X(j)}, \tag{47}$$

we can rephrase the optimization problem in (16) as follows

$$\min_{n(i\to j)} \sum_{j=1}^{|\mathcal{X}|} \frac{\sum_k n(k \to j)}{n} \cdot \log \left(\frac{\sum_k n(k \to j)}{n P_X(j)} \right) \tag{48}$$

subject to the constraints (let us call them $\mathcal{K}$):

$$\begin{cases} n(i \to j) \geq 0 \\ \sum_j n(i \to j) = n_y(i) \forall i \\ \sum_{i,j} n(i \to j) g(i,j) \leq nD. \end{cases} \tag{49}$$

Given the optimum values of the $n(i \to j)$'s, the AD can determine $f(\cdot)$ easily. For instance, for any pair (i,j) he can pick at random $n(i \to j)$ samples of y^n equal to i and transform them into j.

Some observations are in order regarding (48). First of all the number of optimization variables is quadratic in $|\mathcal{X}|$. This is a great improvement with respect to (16) where the number of variables involved in the optimization was n. Secondly, it obviously makes sense to consider only solutions for which one between $n(i \to j)$ and $n(j \to i)$ is equal to 0, nevertheless it is not necessary to explicitly express this constraint in $\mathcal{K}$, since the solutions for which this condition does not hold can be easily pruned after the optimization problem is solved. Last, but not least, as it is shown in the Appendix, the objective function is convex in the optimization variables $n(i \to j)$, hence several efficient solutions exist to solve it [28]. Among them we mention the possibility of solving a relaxed version of the problem in which the $n(i \to j)$'s do not need to be integer. Given the convexity of the objective function the relaxed problem can be solved efficiently by resorting to steepest gradient methods. Once the relaxed solution has been obtained, the optimum integer solution can be found by searching in the surrounding of the relaxed minimum. A further significant simplification can be achieved by observing that, actually, the AD does not need to minimize the divergence between P_{z^n} and P_X, all

[7] According to this notation $n(i \to j)$ is always a positive number and $n(i \to i)$ indicates the number of times that symbol i is left as is by $f(\cdot)$.

he needs is to find a sequence within Λ_0^* that satisfies the distortion constraint. Two cases are possible: let $\mathcal{D}^*$ be the minimum divergence achieved by solving the relaxed problem. If $\mathcal{D}^* \geq \lambda - |\mathcal{X}|(\log(n+1))/n$, then there's no way for the AD to move y^n within Λ_0^* and the FA will win the game. On the contrary, if $\mathcal{D}^* < \lambda - |\mathcal{X}|(\log(n+1))/n$, the AD can proceed by quantizing the values $n(i \to j)$ until he finds an integer solution that satisfies the distortion constraint and for which $f(y^n) \in \Lambda_0^*$. If n is large the impact of the quantization on the objective function in (48) will be minimal and the search very fast.

A. A Case Study

In order to show the potentiality of the numerical analysis outlined above, we apply it to a case study somewhat related to a class of practical source identification problems. Let us assume that two signal sources X and Y differ for the *noisiness* of the signals they produce. In order to test the hypothesis that a signal has been generated by X, the FA applies a wavelet decomposition to the signal and considers the statistics of the DWT (Discrete Wavelet Transform) coefficients at a certain decomposition level. The FA knows that the DWT coefficients are independent and follow a Laplacian distribution $P_X(x) = 0.5\alpha e^{-\alpha|x|}$. The DWT coefficients of the signal produced by the source Y also follow a Laplacian distribution but with a different decay parameter β. Alternatively, we could consider images produced by cameras characterized by a different noise variance [29], [30]. In order to distinguish between images acquired by the two cameras, the FA identifies a flat region of the image and analyzes how the pixel grey levels are distributed around the mean value of the area. The FA knows that if the image has been produced by the first source the pixels follow a Laplacian distribution with decay parameter α, while for images acquired by the second camera, the decay parameter is equal to β. Note that we used a Laplacian distribution only as an example, the whole derivation remaining valid for any other distribution. Given a sequence y^n of DWT coefficients (or pixel gray levels) produced by Y and a distortion constraint nD, we would like to derive the optimum attacking strategy for the AD. We would also like to investigate whether the FA can distinguish between sequences (images) generated by X and Y by ensuring that the false positive error probability tends to zero exponentially fast with error exponent at least equal to λ.

A first problem we have to solve is that the analysis described in the previous sections does not apply to continuous sources. The simplest way to get around this problem is to quantize the continuous pdf's. If the quantization step is small enough, the analysis of the discrete case will provide useful indications about the continuous problem. Without loss of generality, in the following we quantize the Laplacian pdf's onto the set of integers by restricting the pdf to values that have a nonnegligible probability of appearing in a sequence of a certain length. Specifically, the probability $P_X(i)$ is computed as:

$$P_X(i) = \int_{i-1/2}^{i+1/2} \frac{\alpha}{2} e^{-\alpha|x|} dx. \tag{50}$$

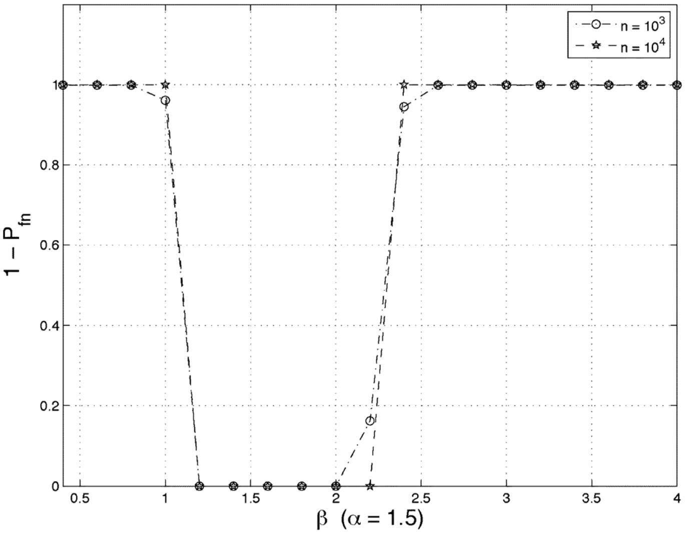

Fig. 1. False negative error probability obtained through Monte Carlo simulations (1000 random sequences). The plots have been obtained by letting $D = 0.01, \lambda = 0.06$.

For the values of n, α and β used in our simulations, it is enough to consider values until $i = \pm 20$ since the probability that a value outside the interval $[-20, 20]$ shows up is significantly lower than one[8]. For instance, with $n = 10^6$ and $\alpha = 1$, such a probability is about 0.002. A similar procedure is adopted to discretize P_Y. Let us call $\hat{P}_X$ and $\hat{P}_Y$ the discretized versions of P_X and P_Y. A first possibility to use the numerical analysis is through Monte Carlo simulations. We generate a great number of sequences according to $\hat{P}_Y$ and use the numerical optimization to move them within Λ_0^*. Measuring the success rate will provide an estimate of the false negative error probability. Fig. 1 shows the results obtained with the above procedure for $\alpha = 1.5$, various values of β, $D = 1/100$, $\lambda = 0.06$ and 2 different values of n ($n = 10^3$ and $n = 10^4$). Each point has been obtained by generating 10^3 sequences according to $\hat{P}_Y$.

In Fig. 2, the results of the optimum attack are exemplified, by reporting the target pdf $\hat{P}_X$, the type of a sequence generated according to $\hat{P}_Y$, and the type of the attacked sequence for two different values of D. As it can be seen, the type of the attacked sequence gets closer to the target pdf thus reducing the divergence between P_{z^n} and $\hat{P}_X$, possibly entering the detection region Λ_0^*.

The behavior of P_{fn} agrees with the insights provided by Theorem 2: the values of β can be split into 2 main classes, those for which the false negative error probability approaches zero and those for which the false negative probability tends to 1. Of course, the former class corresponds to the cases for which β is further from α thus easing the job of the FA. Such a dual behavior is more evident for large values of n since for such values the numerical analysis gets closer to the asymptotic conditions underlying the theoretical analysis. The numerical analysis, then permits, for each value of λ (and for a fixed D), to compute the minimum and maximum values of β for which the FA is going to fail. An example of this analysis is given in Fig. 3, where the range of β for which the FA fails is plotted as a function of λ.

[8] For $i = 20$ we let $P_X(i) = \int_{i-1/2}^{\infty} (\alpha)/(2) e^{-\alpha|x|} dx$. Similarly for $i = -20$.

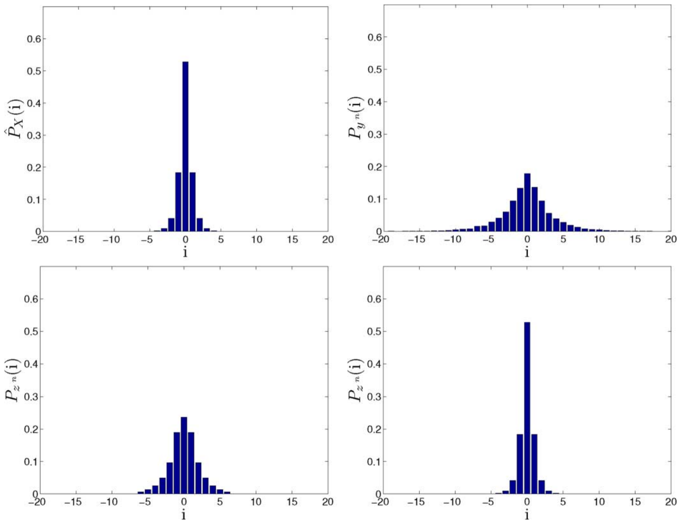

Fig. 2. Result of the optimal attack (from top left to bottom right): target pdf $\hat{P}_X$ $(\alpha = 1.5)$, P_{y^n} (drawn from a source with $\beta = 0.4$), P_{z^n} for $D = 1$ and $D = 2$. The divergence $D(P_{z^n} \| \hat{P}_X)$ after the attack is, respectively, 0.453 $(D = 1)$ and 7×10^{-5} $(D = 2)$.

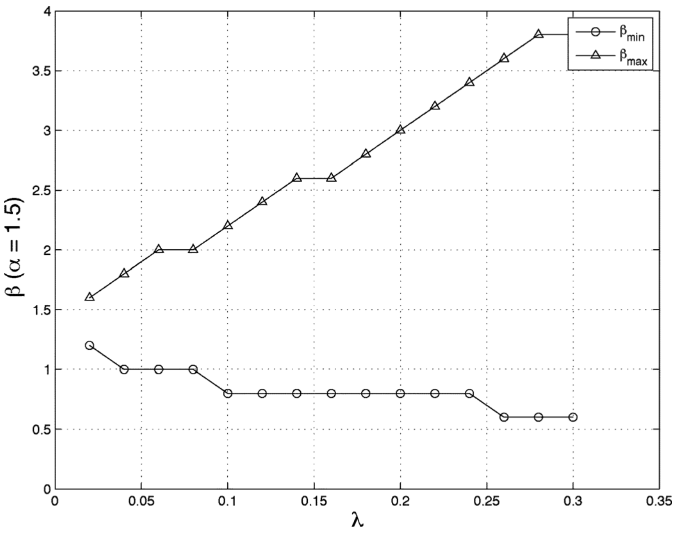

Fig. 3. Range of β for which increasing n results in a false-negative error probability tending to 1. The plots have been obtained by letting $\alpha = 1.5$, $D = 1/100$, and $n = 10^4$.

The above observation suggests an alternative way to use the numerical analysis to decide who between the AD and the FA is going to *win* the game. Since due to the law of large numbers, for large n all the sequences will eventually belong to the same type, and that such a type will approach $\hat{P}_Y$, we can investigate whether the AD will succeed in bringing such sequences within Λ_0^* (of course this analysis does not allow to estimate ε_{fn}). Doing so requires only that the numerical optimization is applied directly to the type obtained by multiplying the probabilities in $\hat{P}_Y$ by (a large enough) n, and see whether the value of the divergence obtained in this way is lower than $\lambda - |\mathcal{X}|(\log(n+1))/n$. An example of the results that can be obtained in this way is given in Fig. 4, where the maximum value of λ for which the AD is not able to fool the forensic analysis is given. Upon inspection of the figure, for any value of D, we can determine the values of β for which no reliable forensic analysis is possible $(\lambda_{\max} = 0)$. For the other values, the FA can distinguish between sequences produced by X and Y assuming that he chooses a value of λ lower than $\lambda_{\max}$.

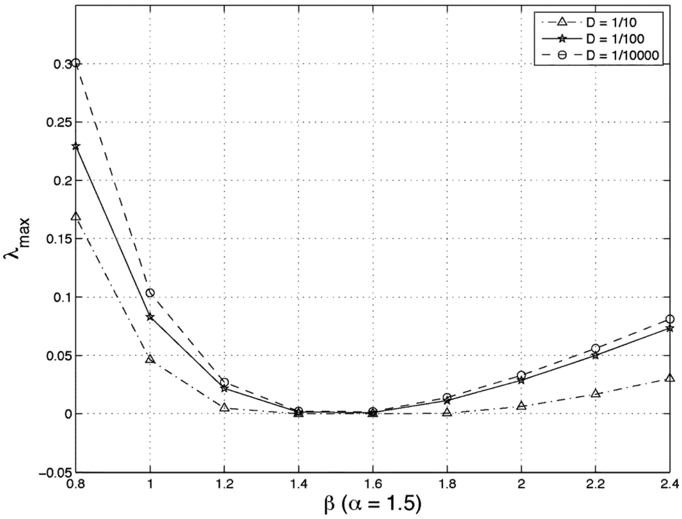

Fig. 4. Maximum value of λ achievable as function of β. The plots refer to different values of D and have been obtained by letting $\alpha = 1.5$, $n = 10^6$.

V. Conclusion

The definition of the SI_{ks} and SI_{ks}^{lr} games, and the derivation of the Nash equilibrium for the SI_{ks}^{lr} game, represent a first step towards the construction of a theoretical framework to cast multimedia forensics and counter-forensics in. While we recognize that the proposed framework does not account for all the subtleties involved in real forensic analysis, e.g., the necessity of preserving the perceptual plausibility of the forged documents, and that the statistical models under which some instances of the problem are solved do not grasp the complexity of real signals, we believe our analysis to be an important step towards the construction of a rigorous theoretical background for multimedia forensics research that can be used to guide the search for practical algorithms. The identification of an efficient numerical procedure to determine the optimum AD strategy is also important for the application of the theoretical framework to real scenarios, since it can contribute to fill the gap between the simplicity of theoretical models and the complexity of real life applications. In [31], for instance, such numerical procedure is used to derive a universal attacking strategy to conceal manipulation traces left in the image histogram.

A lesson that can be learned from our analysis is that constraining the FA to rely only on the empirical pdf of the test sequence for taking a decision may result in the impossibility for him to successfully distinguish between the sources X and Y. The only way out in this case is to resort to higher order analysis based on more accurate image models (as usually done with camera identification methods based on PRNU-Photo Response Nonuniformity [2]). In fact, this is a path similar to the one followed by steganalysis researchers as a way to cope with steganographic schemes designed according to the perfect steganography approach [16].

Several directions for future research can be outlined, among them we mention the extension of the results presented in this paper to more realistic models, e.g., Markov sources, or continuous sources, and the adoption of a source model capable of describing nonstationary acquisition artifacts like PRNU and other forms of sensor noise. The definition of the source identification game when the sources are known only through training data is another promising research direction that we plan to investigate. A further research direction to look into is the definition of a proper game for a multisource scenario in which the binary hypothesis test is replaced by a multiple hypotheses test. Such an extension could follow the approach adopted in [32], where the optimum asymptotic decision strategy - in the absence of attacks—is derived under a constraint on the error exponents involved in the test.

APPENDIX

A. Proof of Theorem 2

In order to complete the proof of the theorem given in the main body of the paper, we have to show that, given P^*, if n is large enough we can find a type in Γ^n_{fn} that is arbitrarily close to P^*. We can distinguish two cases, P^* may either belong to $\bigcup_n \Gamma^n_{fn}$ or not. In the sequel we provide the proof for the former case. In the latter case, in fact, for the very definition of Γ^∞_{fn} we can find a type P' in $\bigcup_n \Gamma^n_{fn}$ which is arbitrarily close to P^*, and then a type in Γ^n_{fn} (with n large enough) that is arbitrarily close to P' and hence to P^*. Let then P^* be such that $P^* \in \Gamma^m_{fn}$ for some m. Due to the definition of Γ^m_{fn} this means that taken a sequence $y^m \in T(P^*)$ a mapping $f_m(y^m)$ exists such that

$$\begin{aligned} &\mathcal{D}(P_{z^m}\|P_X) = \lambda - \frac{\log(m+1)}{m} - \varepsilon < \lambda - \frac{\log(m+1)}{m} \\ &d(z^m, y^m) \le mD, \end{aligned} \tag{A1}$$

where, as usual, $z^m = f_m(y^m)$ and where ε is a strictly positive quantity. In order to characterize the mapping $f_m(\,\cdot\,)$ we use the same notation used in Section IV, i.e., we consider the quantity $m(i \to j)$[9] indicating how many times a value i is changed into j by the application of $f_m(\,\cdot\,)$, and the quantity $m_y(i)$ indicating the number of occurrences of symbol i in the sequence y^m. Due to the density of rational number on the real line, for large n we can find a type $Q_n \in \mathcal{P}_n$ that is arbitrarily close to P^*. More precisely, by indicating with v^n a sequence in $T(Q_n)$ and with $n_v(i)$ the number of times that the symbol i appears in v^n, we can say that for any positive value δ and for large n we have

$$\left|\frac{m_y(i)}{m} - \frac{n_v(i)}{n}\right| < \delta\ \forall i. \tag{A2}$$

We now prove that if δ is small enough (and hence n large enough) $Q_n \in \Gamma^n_{fn}$. To do so, we have to prove that it is possible to map any sequence in $T(Q_n)$ in a sequence whose type belongs to $\Lambda_0^{*,n}$. To do so we consider a mapping $f_n(\,\cdot\,)$ defined as follows:

$$\begin{aligned} n(i \to j) &= \left\lfloor m(i \to j)\frac{n}{m} \right\rfloor = m(i \to j)\frac{n}{m} + \nu_{i,j}\ i \ne j \\ n(i \to i) &= n_v(i) - \sum_{j \ne i} n(i \to j), \end{aligned} \tag{A3}$$

where $\nu_{i,j}$, indicating the truncation error, is a negative quantity strictly larger than -1. We now prove that the sequence w^n resulting from the application of f_n to v^n satisfies the distortion constraint and belongs to $\Lambda_0^{*,n}$. For any additive distortion measure $d(\cdot,\cdot)$, the distortion constraint can be rewritten as:

$$\begin{aligned} d(w^n, v^n) &= \sum_{i \ne j} n(i \to j) g(i,j) \\ &= \sum_{i \ne j} \left(m(i \to j)\frac{n}{m} + \nu_{i,j} \right) g(i,j) \\ &\le \frac{n}{m} \sum_{i,j} m(i \to j) g(i,j) \le \frac{n}{m} mD = nD, \end{aligned} \tag{A4}$$

where the last inequality derives from the fact that the application of the mapping $f_m(\,\cdot\,)$ to y^m satisfies the distortion constraint for sequences of length m. We only have to show that $w^n \in \Lambda_0^{*,n}$. We can do that by means of the following chain of equalities:

$$\begin{aligned} n_w(j) &= n(j \to j) + \sum_{i \ne j} n(i \to j) \\ &= n(j \to j) + \sum_{i \ne j} m(i \to j)\frac{n}{m} + \sum_{i \ne j} \nu_{i,j} \\ &= n_v(j) - \sum_{i \ne j} n(j \to i) + \sum_{i \ne j} m(i \to j)\frac{n}{m} + \sum_{i \ne j} \nu_{i,j} \\ &= n_v(j) - \sum_{i \ne j} m(j \to i)\frac{n}{m} + \sum_{i \ne j} \nu_{j,i} \\ &\quad + \sum_{i \ne j} m(i \to j)\frac{n}{m} + \sum_{i \ne j} \nu_{i,j}. \end{aligned} \tag{A5}$$

By passing to relative frequency we can write:

$$\frac{n_w(j)}{n} = \frac{n_v(j)}{n} - \sum_{i \ne j} \frac{m(j \to i)}{m} + \sum_{i \ne j} \frac{m(i \to j)}{m} + \delta_n$$

where δ_n is a sequence that tends to 0 for $n \to \infty$. By exploiting the relation (A2), we can say that:

$$\begin{aligned} \frac{n_w(j)}{n} &= \frac{m_y(j)}{m} - \sum_{i \ne j} \frac{m(j \to i)}{m} \\ &\quad + \sum_{i \ne j} \frac{m(i \to j)}{m} + \delta_n + \delta' \\ &= \frac{m(j \to j)}{m} + \sum_{i \ne j} \frac{m(i \to j)}{m} + \delta_n + \delta' \\ &= \frac{m_z(j)}{m} + \delta_n + \delta', \end{aligned} \tag{A6}$$

with the absolute value of δ' smaller than δ in (A2) and hence arbitrarily small. In other words, for large n the type of the sequence w^n can be made arbitrarily close to the type of z^m. Due to the continuity of $\mathcal{D}(P_{w^n}\|P_X)$ as a function of the elements of P_{w^n}, it is possible to choose a large enough n such that

[9] We use the symbol m instead of n to indicate explicitly that we are considering sequences of length m.

$|\mathcal{D}(P_{w^n}\|P_X) - \mathcal{D}(P_{z^m}\|P_X)| < \varepsilon'$ with $\varepsilon' < \varepsilon$. Then we can write:

$$\begin{aligned}\mathcal{D}(P_{w^n}\|P_X) &\le D(P_{z^m}\|P_X) + \varepsilon' \\ &= \lambda - \frac{\log(m+1)}{m} + \varepsilon' - \varepsilon \\ &< \lambda - \frac{\log(n+1)}{n}, \qquad \text{(A7)}\end{aligned}$$

which completes our proof.

B. Convexity of $\mathcal{D}(P_{z^n}\|P_X)$ as a Function of $n(k \to j)$

Let N be a matrix with $N(k,j) = n(k \to j)$. We have to prove that the objective function, let us indicate it by $g(N)$, of the optimization problem expressed in (48) is convex in N, i.e., that for any two matrices N_1 and N_2 and any two values $\alpha \in [0,1]$ and $\beta = 1 - \alpha$, we have

$$g(\alpha N_1 + \beta N_2) \le \alpha g(N_1) + \beta g(N_2). \qquad \text{(A8)}$$

Let N^j be the j-th column of N and let $g_j(N^j)$ be defined as:

$$g_j(N^j) = \left(\sum_k n(k \to j)\right) \cdot \log \frac{(\sum_k n(k \to j))}{nP_X(j)}. \qquad \text{(A9)}$$

We clearly have:

$$\begin{aligned}g(N) &= \sum_{j=1}^{|\mathcal{X}|} \frac{(\sum_k n(k \to j))}{n} \cdot \log \frac{(\sum_k n(k \to j))}{nP_X(j)} \\ &= \frac{1}{n} \sum_{j=1}^{|\mathcal{X}|} g_j(N^j). \qquad \text{(A10)}\end{aligned}$$

By definition g_j does not depend on $n(k \to i)\ \forall i \neq j$, hence, if relation (A8) holds for each g_j, then it also holds for the overall function $g(N)$. We have

$$\begin{aligned}g_j\left(\alpha N_1^j + \beta N_2^j\right) &= \sum_k (\alpha n_1(k \to j) + \beta n_2(k \to j)) \\ &\cdot \log \frac{\sum_k (\alpha n_1(k \to j) + \beta n_2(k \to j))}{nP_X(j)}, \qquad \text{(A11)}\end{aligned}$$

that we conveniently rewrite as:

$$\begin{aligned}g_j\left(\alpha N_1^j + \beta N_2^j\right) &= \left(\alpha \sum_k n_1(k \to j) + \beta \sum_k n_2(k \to j)\right) \\ &\cdot \log \frac{\alpha \sum_k n_1(k \to j) + \beta \sum_k n_2(k \to j)}{\alpha n P_X(j) + \beta n P_X(j)}. \qquad \text{(A12)}\end{aligned}$$

Being $n(k \to j)$ nonnegative, we can apply the log-sum inequality [20] to (A12), obtaining:

$$\begin{aligned}g_j\left(\alpha N_1^j + \beta N_2^j\right) &\le \alpha \sum_k n_1(k \to j) \cdot \log \frac{\alpha(\sum_k n_1(k \to j))}{\alpha n P_X(j)} \\ &+ \beta \sum_k n_2(k \to j) \cdot \log \frac{\beta(\sum_k n_2(k \to j))}{\beta n P_X(j)} \\ &= \alpha g_j(N_1^j) + \beta g_j(N_2^j), \qquad \text{(A13)}\end{aligned}$$

which completes the proof.

References

[1] E. Delp, N. Memon, and M. Wu, "Special issue on digital forensics," *IEEE Signal Process. Mag.*, vol. 26, no. 2, p. , Mar. 2009.

[2] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 74–90, Mar. 2008.

[3] S. Lyu and H. Farid, "How realistic is photorealistic?," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 845–850, Feb. 2005.

[4] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, Feb. 2005.

[5] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 154–160, Mar. 2009.

[6] Y.-F. Hsu and S.-F. Chang, "Camera response functions for image forensics: An automatic algorithm for splicing detection," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 816–825, Dec. 2010.

[7] S. Bayram, H. T. Sencar, and N. Memon, "An efficient and robust method for detecting copy-move forgery," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2009)*, Apr. 2009, pp. 1053–1056.

[8] T. Gloe, M. Kirchner, A. Winkler, and R. Böhme, "Can we trust digital image forensics?," in *Proc. ACM Multimedia 2007*, Augsburg, Germany, Sep. 2007, pp. 78–86.

[9] M. Kirchner and R. Böhme, "Hiding traces of resampling in digital images," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 582–292, Dec. 2008.

[10] G. Cao, Y. Zhao, R. Ni, and H. Tian, "Anti-forensics of contrast enhancement in digital images," in *Proc. 12th ACM Multimedia and Security Workshop 2010*, Sep. 2010, pp. 24–34.

[11] M. C. Stamm and K. J. R. Liu, "Anti-forensics of digital image compression," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 50–65, Sep. 2011.

[12] R. Böhme and M. Kirchner, , H. T. Sencar and N. Memon, Eds., "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*. Berlin/Heidelberg, Germany: Springer, 2012.

[13] M. Goljan, J. Fridrich, and M. Chen, "Sensor noise camera identification: Countering counter forensics," in *Proc. SPIE Conf. Media Forensics and Security*, San Jose, CA, USA, 2010.

[14] S. Lai and R. Böhme, "Countering counter-forensics: The case of JPEG compression," in *Information Hiding*, ser. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2011, vol. 6958, pp. 285–298.

[15] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro, "Countering JPEG anti-forensics," in *IEEE Int. Conf. Image Processing (ICIP'11)*, Brussels, Belgium, Sep. 2011, pp. 1949–1952.

[16] C. Cachin, "An information-theoretic model for steganography," in *IH98, Second International Workshop on Information Hiding*, ser. ser. Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 1998, vol. 6958, pp. 306–318.

[17] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

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

[19] M. Barni, "A game theoretic approach to source identification with known statistics," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2012)*, Kyoto, Japan, Mar. 25–30, 2012.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[21] I. Csiszar, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

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

[23] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Annals Math. Statist.*, vol. 36, no. 2, pp. 369–401, Apr. 1965.

[24] J. Nash, "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci.*, vol. 36, no. 1, pp. 48–49, 1950.

[25] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA, USA: MIT Press, 1994.

[26] I. Csiszar and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2065–2072, Nov. 1996.

[27] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer-Verlag, 2009.

[28] P. Bonami, M. Kilinc, and J. Linderoth, , J. Lee and S. Leyffer, Eds. , "Algorithms and software for convex mixed integer nonlinear programs," in *Mixed Integer Nonlinear Programming*. New York, NY, USA: Springer, 2012, pp. 1–39.

[29] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image Vis. Comput.*, vol. 27, pp. 1497–1503, 2009.

[30] X. Pan, X. Zhang, and S. Lyu, "Exposing image forgery with blind noise estimation," in *Proc. ACM Multimedia and Security Workshop 2011*, Buffalo, NY, USA, Sep. 2011, pp. 15–20.

[31] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. ACM Multimedia and Security Workshop*, Coventry, U.K., Sep. 6–7, 2012, pp. 97–104.

[32] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

Mauro Barni (M'92–SM'06–F'12) graduated in electronic engineering from the University of Florence in 1991, where he received the Ph.D. degree in informatics and telecommunications in 1995.

He is currently working as associate professor at the University of Siena. During the last decade, his activity has focused on digital image processing and information security, with particular reference to the application of image processing techniques to copyright protection (digital watermarking) and multimedia forensics. Lately he has been studying the possibility of processing signals that have been previously encrypted without decrypting them. He led several national and international research projects on these subjects. He is author of about 270 papers, and holds four patents in the field of digital watermarking and document protection. He is coauthor of the book *Watermarking Systems Engineering* (Dekker, 2004).

In 2008, Dr. Barni was the recipient of the *IEEE Signal Processing Magazine* best column award. In 2011, he was the recipient of the IEEE Geoscience and Remote Sensing Society Transactions Prize Paper Award. He has been the chairman of the IEEE Multimedia Signal Processing Workshop (Siena, 2004), and the chairman of the International Workshop on Digital Watermarking (Siena, 2005). He has been the founding editor in chief of the *EURASIP Journal on Information Security*. He currently serves as associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. From 2010 to 2011, he was the chairman of the IEEE Information Forensic and Security Technical Committee of the Signal Processing Society. He has been a member of the IEEE Multimedia Signal Processing Technical Committee and the conference board of the IEEE Signal Processing Society. He has been appointed distinguished lecturer of the IEEE Signal Processing Society for the years 2012–2013.

Benedetta Tondi graduated (*cum laude*) in electronics and communications engineering from the University of Siena, Siena, Italy, in 2012, with a thesis on *Adversary-aware source identification* in the area of multimedia forensics. She is currently a member of the Visual Information Processing and Protection (VIPP) group in the Department of Information Engineering and Mathematical Sciences, University of Siena, where she is working toward the Ph.D. degree.

Her research interest focuses on the application of information theory and game theory concepts to forensics and counter-forensics analysis.