Source Distinguishability Under Distortion-Limited Attack: An Optimal Transport Perspective

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

***Abstract*— We analyze the distinguishability of two sources in a Neyman–Pearson setup when an attacker is allowed to modify the output of one of the two sources subject to an additive distortion constraint. By casting the problem in a game-theoretic framework and by exploiting the parallelism between the attacker's goal and optimal transport theory, we introduce the concept of security margin defined as the maximum average per-sample distortion introduced by the attacker for which the two sources can be distinguished ensuring arbitrarily small, yet positive, error exponents for type I and type II error probabilities. Several versions of the problem are considered according to the available knowledge about the sources. We compute the security margin for some classes of sources and derive general bounds assuming that the distortion is measured in terms of the mean square error between the original and the attacked sequence. The analysis of the game and the study of the distinguishability of the sources are extended to the case in which the distortion constraint is defined in terms of the maximum distance.**

***Index Terms*— Adversarial signal processing, hypothesis testing, source identification, multimedia forensics, cybersecurity, game theory, optimal transportation theory, earth mover distance (EMD).**

I. Introduction

ADVERSARIAL Signal Processing (Adv-SP) is an emerging research field targeting the study of signal processing techniques explicitly thought to withstand the intentional attacks of one or more adversaries aiming at system failure. Adv-SP methods can be applied to a wide variety of security-oriented applications including multimedia forensics, biometrics, digital watermarking, steganography and steganalysis, network intrusion detection, traffic monitoring, video-surveillance, just to mention a few [1].

Source identification, often modeled as a binary decision or hypothesis testing problem, is one of the most common problems in Adv-SP. In multimedia forensics [2], for instance, a forensic analyst may be asked to decide whether an image has been acquired by a given camera, notwithstanding the presence of an adversary aiming at deleting the acquisition traces left by the camera. Similarly, the analyst may be asked to decide whether a document has undergone a certain processing or not, by taking into account the possibility that someone deliberately tried to delete the traces left by the processing operator. Biometric authentication provides a further example. In this case, the authentication system must decide whether a biometric trait belongs to a certain individual, despite the presence of an attacker aiming at building a fake template that passes the authentication test [3], [4]. In 1-bit watermarking, the detector is asked to decide if a document contains a given watermark or not [5], in the presence of possible attackers aiming at undermining the detection process. Other examples include: steganalysis, in which the steganalyzer has to distinguish between cover and stego images [6] and network intrusion detection, wherein anomalous traffic conditions must be distinguished from normal ones [7]. In all these fields, taking into account the presence of the adversary in the design phase is essential to build a system which works properly also in hostile settings.

In [8], a game-theoretic framework is proposed to analyze the source identification problem under adversarial conditions. To be specific, [8] introduces the so called source identification game. The game is played by a Defender (D) and an Attacker (A) and is defined as follows: given two discrete memoryless sources X and Y with alphabet $\mathcal{X}$ and probability mass functions (pmf) P_X and P_Y, and a test sequence $x^n = (x_1, x_2 \dots x_n)$, the goal of D is to decide between hypothesis H_0 that x^n has been drawn from X and hypothesis H_1 that x^n has been generated by Y. The goal of A is to take a sequence y^n generated by Y and modify it in such a way that D classifies it as being generated by X. In doing so, D must ensure that the type I error probability (usually referred to as false positive error probability P_{fp}) of deciding for H_1 when H_0 holds stays below a given threshold, whereas A has to respect a distortion constraint, limiting the amount of modifications he can introduce into y^n. The payoff of the game is the type II error probability, or false negative error probability P_{fn}, i.e., the probability of deciding for H_0 when H_1 holds. Of course, D aims at minimizing P_{fn}, while A wishes to maximize it. The above scenario accounts for a situation in which P_X corresponds to so-to-say normal conditions and P_Y refers to an anomalous situation. It is the goal of the attacker to modify a sequence produced under anomalous conditions in such a way that the defender does not recognize that the observed system exited the normal state.

Under the assumption that the defender bases its analysis only on first order statistics of x^n, [8] derives the asymptotic equilibrium point of the game when the length of the test sequence tends to infinity and P_{fp} tends to zero exponentially fast with decay rate at least equal to λ (λ is nothing but the error exponent of the false positive error probability).

Manuscript received July 20, 2015; revised January 21, 2016 and April 13, 2016; accepted May 12, 2016. Date of publication May 19, 2016; date of current version July 8, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tanya Ignatenko.

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

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.2016.2570739

Given two pmf's P_X and P_Y, an additive distortion measure $d(\cdot,\cdot)$ and the maximum allowed distortion L_{max}, the analysis in [8] permits to determine whether, at the equilibrium, the false negative error probability P_{fn} tends to 0 or to 1 when $n \to \infty$, for a prescribed false positive error exponent λ. This, in turn, permits to define the so-called indistinguishability region $\Gamma(P_X, \lambda, L_{max})$ as the set of pmf's that can not be distinguished from P_X when $n \to \infty$ due to the presence of the attacker. If $P_Y \in \Gamma(P_X, \lambda, L_{max})$, in fact, a strictly positive false negative error exponent can not be achieved and the attacker is going to *win* the game. A similar analysis is carried out in [9] and [10] for a scenario in which P_X and P_Y are not known, and the statistics of the two sources are obtained through the observation of training sequences.

A. Contribution

A drawback with the analysis carried out in [8]–[10] is the asymmetric role of the false positive and false negative error exponents, namely λ and ε ($\varepsilon = \lim_{n\to\infty} -\frac{1}{n}\log P_{fn}$). In such works, in fact, the defender aims at ensuring a given λ, but is satisfied with any strictly positive ε. In this paper, we make a more reasonable assumption and say that the defender wins the game if - at the equilibrium - both error probabilities tend to zero exponentially fast, regardless of the particular values assumed by the error exponents. More precisely, by mimicking Stein's lemma [11], we analyze the behavior of $\Gamma(P_X, \lambda, L_{max})$ when $\lambda \to 0$ to see whether, given a maximum allowable distortion L_{max}, it is possible for D to simultaneously attain strictly positive error exponents for the two kinds of error. Having done so, we introduce a new distinguishability measure, called Security Margin ($\mathcal{SM}$), defined as the *maximum distortion allowed to the attacker, for which two sources can be distinguished reliably*. As we will see, this is a powerful concept that permits to summarize in a single quantity the distinguishability of two sources X and Y under adversarial conditions. In order to derive our main results, we parallel the optimum attacker's strategy to an *optimal transport theory* problem [12]. This allows to get an insightful interpretation of the optimum attacker's strategy and to find out that the distinguishability of two sources is ruled by a quantity (namely the security margin $\mathcal{SM}$), which corresponds to the Earth Mover Distance (*EMD*) [13]. We derive the $\mathcal{SM}$ for a wide class of pmf's in both the discrete and the continuous case and, by relying on some results in the field of optimal transport theory, we present a numerical algorithm for its efficient computation. We also extend the analysis to a version of the source identification game in which P_X and P_Y are known only through training sequences [10]. Eventually, we introduce a new version of the game in which the distortion constraint is expressed in terms of maximum absolute distancesx. This is a very interesting, yet non-trivial, scenario, that opens the way to the application of our theory to all the cases in which the distortion constraint is applied uniformly to all the elements of y^n.

It is worth stressing that this paper complements and generalizes some recent studies in the field of Multimedia security, namely [14]–[16], regarding image counterforensics, and [17], related to perfect steganography. As a matter of fact, all the solutions proposed in those papers can be seen as particular instances of the general optimal transport problem addressed and solved in Section VI.

Some of the results presented in this paper have already been stated (but not proven) in [18]. With respect to [18], the current paper contains a complete proof of all the main theorems, the extension to the case of source identification with training data, the derivation of a fast numerical methodology to compute the security margin between any two discrete sources, and the extension of the analysis to the case of L_∞ distortion.

The rest of this paper is organized as follows. In Section II, we introduce the notation used throughout the paper, give some definitions and review some basic concepts in game theory. In Section III, we give a rigorous definition of the addressed problem and summarize the main results proven in [8]. Section IV is the core of the paper: we use optimal transport to shed new light on the addressed problem and introduce the security margin concept. In Section V, we extend the analysis to cover the case of source identification with training data. In Section VI, we derive the security margin for several classes of sources, and provide an efficient algorithm to compute it when a close form solution does not exist. Section VII extends the analysis to a situation in which the allowed distortion is defined in terms of L_∞ distance. In Section VIII, we discuss the value of our analysis in practical applications, and show how the security margin concept can be applied to a well known problem in image forensics. The paper ends in Section IX, with some conclusions and highlights for future research. The most technical proofs are given in the appendices to avoid interrupting the flow of ideas.

II. Notations and Definitions

We will use capital letters (e.g. X) to indicate discrete memoryless sources. Sequences of length n drawn from a source will be indicated with the corresponding lowercase letter (e.g. x^n); accordingly, x_i will denote the $i-$th element of a sequence x^n. The alphabet of a source will be indicated by the corresponding calligraphic capital letter (e.g. $\mathcal{X}$). The probability mass function (pmf) of a discrete memoryless source X will be denoted by P_X, while the cumulative mass function will be indicated with C_X. For the sake of simplicity, the same notation will be adopted to denote the probability density function (pdf) of a continuous random variable X. The notation P_X will also 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 recoverable from the context wherein it is used. Finally, we will use the notation $P_X(A)$ to indicate the probability of an event A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$) under the probability measure P_X. The calligraphic letter $\mathcal{P}$ will be used to indicate the set of all pmf's.

Our analysis relies on the concepts of type and type class defined as follows [11], [19]. Let x^n be a sequence with elements belonging to a finite alphabet $\mathcal{X}$. The type P_{x^n} of x^n is the empirical pmf induced by the sequence x^n,

i.e. $\forall a \in \mathcal{X}$, $P_{x^n}(a) = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i, a)$, where $\delta(x_i, a) = 1$ if $x_i = a$ and zero otherwise. 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. We denote by $\mathcal{D}(P||Q)$ the Kullback-Leibler (KL) divergence between two distributions P and Q, defined on the same finite alphabet $\mathcal{X}$ [11].

A. Game Theory in a Nutshell

A 2-player 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 actions (usually called strategies) the first and the second player can choose from, and $u_l(s_{1,i}, s_{2,j}), l = 1, 2$, is the payoff of the game for player l, when the first player chooses the strategy $s_{1,i}$ and the second chooses $s_{2,j}$. A pair of strategies $(s_{1,i}, s_{2,j})$ is called a profile. When $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$, the game is said to be a competitive (or zero-sum) game. In the set-up adopted in this paper, $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to the two players. In addition, we assume that the players choose their strategies before starting the game without knowing the strategy chosen by the other player (strategic game).

A common goal in game theory is to determine the existence of equilibrium points, i.e. profiles that in *some way* represent a *satisfactory* choice for both players [20]. The most famous equilibrium notion is due to Nash. Intuitively, a profile is a Nash equilibrium if each player does not have any interest in changing his choice assuming the other does not change his strategy. Despite its popularity, the practical meaning of Nash equilibrium is doubtful, since there is no guarantee that the players will end up playing at the equilibrium. A notion with a more practical meaning is that of *dominant equilibrium*. A strategy is said to be strictly dominant for one player if it is the best strategy for the player, regardless of the strategy chosen by the other player. When a dominant strategy exists for one of the players, he will surely adopt it. The other players, in turn, will choose their strategies anticipating that the first player will play the dominant strategy. As a consequence, in a two-player game, if a dominant strategy exists the players have only one rational choice called the only rationalizable equilibrium of the game [21]. Games with the above property are called *dominance solvable* games.

III. The SI_{ks} Game

In this section, we formally define the source identification game and summarize the main results proven in [8].

A. Definition of the SI_{ks} Game and Equilibrium Point

We start with the definition of the source identification game with known sources (SI_{ks}). Given a test sequence x^n, we indicate with H_0 the hypothesis that x^n has been generated by P_X and with H_1 the alternative hypothesis that x^n has been generated by P_Y.

Defender's strategies. The set of strategies of the Defender ($\mathcal{S}_D$) consists of all possible acceptance regions for H_0. More precisely, by following [8], we require that D bases its analysis only on the first order statistics of x^n. This is equivalent to ask that the acceptance region for H_0, hereafter referred to as Λ^n, is a union of type classes. Since a type class is univocally defined by the empirical pmf of the sequences it contains, Λ^n can be seen as a union of types $P \in \mathcal{P}_n$. We consider an asymptotic version of the game and require that the false positive error probability P_{fp} decreases exponentially with decay rate at least equal to λ. Under the above assumptions, the space of strategies of D is given by:

$$\mathcal{S}_D = \{\Lambda^n \in 2^{\mathcal{P}_n} : P_{fp} \leq 2^{-\lambda n}\}, \tag{1}$$

where $2^{\mathcal{P}_n}$ indicates the power set of $\mathcal{P}_n$.

Attacker's strategies. Given a sequence y^n drawn from Y, the goal of A is to transform it into a sequence z^n belonging to Λ^n. Let us indicate by $n(i, j)$ the number of times that the i-th symbol of the alphabet is transformed into the j-th one as a consequence of the attack. Similarly, we indicate by $S^n_{YZ}(i, j) = n(i, j)/n$ the relative frequency with which the i-th symbol is transformed into the j-th one. In the following, we refer to S^n_{YZ} as *transportation map*. For any additive distortion measure, the overall distortion introduced by the attack can be expressed as $d(y^n, z^n) = \sum_{i,j} n(i, j)d(i, j)$, where $d(i, j)$ is the distortion introduced when symbol i is transformed into symbol j. Similarly, the average per-sample distortion depends only on S^n_{YZ}; in fact, $d(y^n, z^n)/n = \sum_{i,j} S^n_{YZ}(i, j)d(i, j)$. The map S^n_{YZ} determines also the type of the attacked sequence. In fact, by indicating with $P_{z^n}(j)$ the relative frequency of symbol j in z^n, we have $P_{z^n}(j) = \sum_i S^n_{YZ}(i, j) \triangleq S^n_Z(j)$. Finally, we observe that the attacker can not change more symbols than there are in the sequence y^n; as a consequence a map S^n_{YZ} can be applied to a sequence y^n only if $S^n_Y(i) \triangleq \sum_j S^n_{YZ}(i, j) = P_{y^n}(i)$. The above reasoning suggests an interesting interpretation of S^n_{YZ}, which can be seen as the joint empirical pmf between the sequences y^n and z^n. In the same way, S^n_Y and S^n_Z correspond, respectively, to the type of y^n and z^n.

By remembering that Λ^n depends only on the type of the test sequence, and given that the type of the attacked sequence depends on S^n_Z only through S^n_{YZ}, we can define the action of the attacker as the choice of a transportation map in the set of *admissible* maps defined as:

$$\mathcal{A}^n(L_{max}, P_{y^n}) = \begin{cases} S^n_Y = P_{y^n} \\ \sum_{i,j} S^n_{YZ}(i, j)d(i, j) \leq L_{max}, \end{cases} \tag{2}$$

where the second condition expresses the per-letter distortion constraint the attacker is subject to, and L_{max} is the maximum allowed (average) per-letter distortion. With the above definitions, the space of strategies of the attacker is the set of all the possible ways of associating an admissible transformation map to the to-be-attacked sequence. In the following, we will refer to the result of such an association as $S^n_{YZ}(y^n)$, or $S^n_{YZ}(i, j; y^n)$. In the same way, $S^n_Z(j; y^n)$ indicates the output marginal of $S^n_{YZ}(i, j; y^n)$. By adopting the above symbolism, the space of strategies for the attacker can be defined as:

$$\mathcal{S}_A = \{S^n_{YZ}(i, j; y^n) : S^n_{YZ}(i, j) \in \mathcal{A}^n(L_{max}, P_{y^n})\}. \tag{3}$$

The payoff. Having fixed the maximum false positive error probability, we adopt a Neyman-Pearson approach and define the payoff as the false negative error probability:

$$u_D = -u_A = - \sum_{y^n : S_Z^n(j;y^n) \in \Lambda^n} P_Y(y^n). \quad (4)$$

The main result of [8] is given by the following theorem.

Theorem 1: Let

$$\Lambda^{n,*} = \left\{ P \in \mathcal{P}_n : \mathcal{D}(P||P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}, \quad (5)$$

and

$$S_{YZ}^{n,*}(i, j; y^n) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L_{max}, P_{y^n})} \mathcal{D}(S_Z^n||P_X). \quad (6)$$

Then $\Lambda^{n,*}$ is a dominant equilibrium for D and the profile $(\Lambda^{n,*}, S_{YZ}^{n,*}(i, j; y^n))$ is the only rationalizable equilibrium of the SI_{ks} game, which, then, is a dominance solvable game.

B. Payoff of the SI_{ks} Game at the Equilibrium

Given the optimal acceptance region $\Lambda^{n,*}$ and the optimum attacking strategy $S_{YZ}^{n,*}(y^n)$, we can introduce the finite-length indistinguishability region $\Gamma^n(P_X, \lambda, L_{max})$ as follows:

$$\Gamma^n(P_X, \lambda, L_{max}) = \{P \in \mathcal{P}_n : \exists\ S_{YZ}^n \in \mathcal{A}^n(L_{max}, P) \text{ s.t. } S_Z^n \in \Lambda^{n,*}\}. \quad (7)$$

The indistinguishability region defines all the type classes (with denominator n) whose sequences can be moved within $\Lambda^{n,*}$ by the attacker. When n tends to infinity, we can define the asymptotic counterpart of Γ^n [8]:

$$\Gamma(P_X, \lambda, L_{max}) = \{P \in \mathcal{P} :\ \exists\ S_{YZ} \in \mathcal{A}(L_{max}, P) \text{ s.t. } S_Z \in \Lambda^*(P_X, \lambda)\}, \quad (8)$$

where

$$\Lambda^*(P_X, \lambda) = \{P \in \mathcal{P} : \mathcal{D}(P||P_X) \leq \lambda\}, \quad (9)$$

and where the definitions of $S_{YZ}(i, j)$, $S_Z(j)$ and $\mathcal{A}(L_{max}, P)$ derive from those of $S_{YZ}^n(i, j)$, $S_Z^n(j)$ and $\mathcal{A}^n(L_{max}, P)$, by relaxing the requirement that $S_{YZ}(i, j)$, $S_Z(j)$ and $P(i)$ are rational numbers with denominator n. More precisely, we can state the following theorem:

Theorem 2: For the SI_{ks} game, the error exponent of the false negative error probability at the equilibrium is given by[1]*:*

$$\varepsilon = \min_{P \in \Gamma(P_X, \lambda, L_{max})} \mathcal{D}(P||P_Y), \quad (10)$$

leading to the following cases:

1) $\varepsilon = 0$, *if* $P_Y \in \Gamma(P_X, \lambda, L_{max})$;
2) $\varepsilon \neq 0$, *if* $P_Y \notin \Gamma(P_X, \lambda, L_{max})$.

According to Theorem 2, $\Gamma(P_X, \lambda, L_{max})$ can be interpreted as the region with the sources that cannot be distinguished from P_X guaranteeing a false positive error exponent at least equal to λ in the presence of an adversary with allowed distortion L_{max}. A geometric interpretation of Theorem 2 is given in Figure 1.

[1]Here and in the rest of the paper the use of the minimum instead of the infimum is justified by the compactness of $\Gamma(P_X, \lambda, L_{max})$.

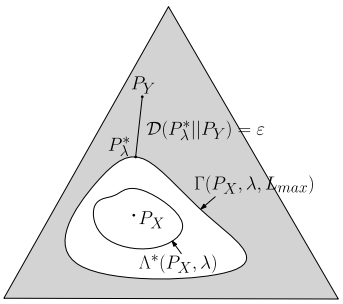


Fig. 1. Geometric interpretation of Theorem 2.

IV. The Security Margin

In this section, we use the optimal transport interpretation of the attacker's strategy to introduce a measure of source distinguishability in the set-up defined by the SI_{ks} game.

A. Characterization of the Indistinguishability Region Using Optimal Transport

To start with, we find it convenient to rephrase the results described in the previous section as an optimal transport problem [12]. Let P and Q be two pmf's defined over the same finite alphabet, and let $c(i, j)$ be the cost of transporting the i-th symbol into the j-th one. In one of its instances, optimal transport theory looks for the transportation map that transforms P into Q by minimizing the average cost of the transport. By using the notation introduced in the previous section, this corresponds to solving the following problem:

$$\min_{S_{YZ} : S_Y = P, S_Z = Q} \sum_{i,j} S_{YZ}(i, j) c(i, j). \quad (11)$$

A nice interpretation of the problem defined by equation (11) is obtained by interpreting the pmf's P and Q as two different ways of piling up a certain amount of earth, and $c(i, j)$ as the cost necessary to move a unitary amount of earth from position i to position j. In this case, the minimum cost achieved in (11) can be seen as the minimum effort required to turn one pile into the other. Due to such a viewpoint, in computer vision applications, the minimum in equation (11) is usually known as Earth Mover Distance (*EMD*) between P and Q [13]. When P and Q are probability mass functions and $c(i, j) = d(i, j)^p$ for some distance measure d (with $p \geq 1$), the *EMD* has a more general statistical meaning. Given two random variables with probability distributions P_X and P_Y, the *EMD* between P_X and P_Y corresponds to the minimum expected p-th power distance between the random variables X and Y taken over all joint probability distributions P_{XY} with marginal distributions respectively equal to P_X and P_Y:

$$EMD_{d^p}(P_X, P_Y) = \min_{P_{XY} : \sum_y P_{XY} = P_X, \sum_x P_{XY} = P_Y} E_{XY}[d(X, Y)^p]. \quad (12)$$

In transport theory terminology, expression (12) is the p-th power of the Wasserstein distance [12], [22] (or the Monge-Kantorovich metric of order p [23], [24]). In particular, when $c(i, j) = |i - j|^2$ (i.e. $d(i, j) = |i - j|$ and $p = 2$) the earth mover distance $EMD_{L_2^2}(P_X, P_Y)$ is equivalent to the

squared Mallows distance between P_X and P_Y [25]. In the following, we will continue to refer to (11) as *EMD(P,Q)*. We also observe that even if we introduced the *EMD* by considering finite-alphabet sources, there is no need to restrict the definition in (12) to discrete random variables. In fact, in the second part of the paper, we will extend our analysis and use the *EMD* to measure the distinguishability of continuous sources.

Optimal transport theory permits to rewrite the indistinguishability region in a more compact and easier-to-interpret way. In fact, it is immediate to see that equation (8) can be rewritten as:

$$\Gamma(P_X, \lambda, L_{max}) = \{P \in \mathcal{P} : \exists\ Q \in \Lambda^*(P_X, \lambda) \text{ s.t } EMD(P, Q) \leq L_{max}\}, \tag{13}$$

where now $c(i, j)$ corresponds to the distortion metric used to constraint the strategies available to the attacker.

B. Security Margin Definition

We now study the behavior of $\Gamma(P_X, \lambda, L_{max})$ when $\lambda \to 0$. Doing so will allow us to investigate whether two sources X and Y are ultimately distinguishable in the setting defined by the SI_{ks} game. The rationale behind our analysis derives directly from equations (8) and (9). In fact, it is easy to see that decreasing λ in the definition of $\mathcal{S}_D$ leads to a more favorable game for the defender, since he can adopt a smaller acceptance region and obtain a larger payoff. Stated in another way, from D's perspective, evaluating the behavior of the game for $\lambda \to 0$ corresponds to exploring the best achievable false negative error exponent, when P_{fp} tends to 0 exponentially fast.

More formally, we start by proving the following property.

Property 1: For any two values λ_1 and λ_2 such that $\lambda_2 < \lambda_1$, $\Gamma(P_X, \lambda_2, L_{max}) \subseteq \Gamma(P_X, \lambda_1, L_{max})$.

Proof: The property follows immediately from equation (13) by observing that $\Gamma(P_X, \lambda, L_{max})$ depends on λ only through the acceptance region $\Lambda^*(P_X, \lambda)$, for which we obviously have $\Lambda^*(P_X, \lambda_2) \subseteq \Lambda^*(P_X, \lambda_1)$ whenever $\lambda_2 < \lambda_1$. ∎

Thanks to Property 1, we can compute the limit of the false negative error exponent when λ tends to zero, as summarized in the following theorem (somewhat resembling Stein's Lemma [11]).

Theorem 3: Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum average per-letter distortion L_{max} (defined according to an additive distortion measure), let us adopt the following definition:

$$\Gamma(P_X, L_{max}) = \{P \in \mathcal{P} : \mathrm{EMD}(P, P_X) \leq L_{max}\}; \tag{14}$$

then the maximum achievable false negative error exponent ε for the SI_{ks} game is

$$\lim_{\lambda \to 0} \lim_{n \to \infty} -\frac{1}{n} \log P_{fn} = \min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y). \tag{15}$$

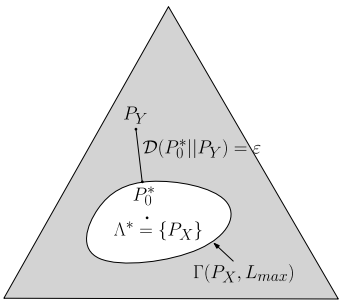


Fig. 2. Geometric interpretation of Theorem 3.

Proof: The innermost limit in (15) defines the error exponent for a fixed λ, say it $\varepsilon(\lambda)$. Thanks to (10), we know that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_{fn} = \varepsilon(\lambda) = \min_{P \in \Gamma(P_X, \lambda, L_{max})} \mathcal{D}(P||P_Y). \tag{16}$$

Then, according to Property 1, the sequence $\varepsilon(\lambda)$ is monotonically non decreasing as λ decreases. In addition, since $\Gamma(P_X, L_{max}) \subseteq \Gamma(P_X, \lambda, L_{max})\ \forall \lambda$, we have:

$$\varepsilon(\lambda) \leq \min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y). \tag{17}$$

Being $\varepsilon(\lambda)$ bounded from above and non-decreasing, the limit for $\lambda \to 0$ exists and is finite. We must now prove that the limit is indeed equal to $\min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y)$. Let P_0^* be the point achieving the minimum in (15) and P_λ^* the point achieving the minimum in the set $\Gamma(P_X, \lambda, L_{max})$, i.e. the point achieving the minimum in equation (10) (see Figure 1 for a pictorial representation of P_λ^*). Due to Lemma 1 (Appendix A), for any arbitrarily small τ, we can choose a small enough λ such that, for any P in $\Gamma(P_X, \lambda, L_{max})$, a pmf P' in $\Gamma(P_X, L_{max})$ exists whose distance from P is lower than τ. By taking $P = P_\lambda^*$ and exploiting the continuity of the $\mathcal{D}$ function, we have

$$\mathcal{D}(P'||P_Y) \leq \min_{P \in \Gamma(P_X, \lambda, L_{max})} \mathcal{D}(P||P_Y) + \delta(\tau), \tag{18}$$

for some $P' \in \Gamma(P_X, L_{max})$ and some value $\delta(\tau)$ such that $\delta(\tau) \to 0$ as $\tau \to 0$. A fortiori, relation (18) holds for $P' = P_0^*$ and then we can write

$$\varepsilon(\lambda) \geq \min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y) - \delta(\tau), \tag{19}$$

where $\delta(\tau)$ can be made arbitrarily small by decreasing λ. Equation (19), together with equation (17), shows that we can get arbitrarily close to $\min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y)$, by decreasing λ, hence proving that $\min_{P \in \Gamma(P_X, L_{max})} \mathcal{D}(P||P_Y)$ is the limit of the sequence $\varepsilon(\lambda)$ as $\lambda \to 0$. ∎

Figure 2 gives a geometric interpretation of Theorem 3. The figure is obtained from Figure 1 by observing that when $\lambda \to 0$ the optimum acceptance region collapses into the single pmf P_X, i.e., $\Lambda^* = \{P_X\}$.

By the light of Theorem 3, $\Gamma(P_X, L_{max})$ is the smallest indistinguishability region for the SI_{ks} game. Moreover, from equation (14), we see that the distinguishability of two pmf's (in the SI_{ks} setting) ultimately depends on their *EMD*. In fact, if $EMD(P_Y, P_X) > L_{max}$, the defender is able to distinguish

X from Y by adopting a sufficiently small λ. On the contrary, if $EMD(P_Y, P_X) \leq L_{max}$, there is no positive value of λ for which the sequences emitted by the two sources can be asymptotically distinguished.

By adopting a different perspective, given two sources X and Y, one may ask which is the maximum attacking distortion for which D can tell X and Y apart. The answer to this question follows immediately from Theorem 3 and leads to the following definition.

Definition 1 (Security Margin): Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum average per-letter distortion for which X and Y can be distinguished in the SI_{ks} setting is called Security Margin and is given by

$$\mathcal{SM}(P_Y, P_X) = \mathrm{EMD}(P_Y, P_X). \tag{20}$$

Interestingly, the *EMD* is a symmetric function of P_X and P_Y [13], and hence the security margin does not depend on the role of X and Y in the test, i.e. $\mathcal{SM}(P_X, P_Y) = \mathcal{SM}(P_Y, P_X)$. The security margin is a powerful measure summarizing in a single quantity how securely two sources can be distinguished.

It is worth remarking that the security margin between two sources pertains to the *security* of the hypothesis test behind the source identification problem and not to its *robustness*, since it is measured at the equilibrium of the game, i.e. by assuming that both the players of the game make optimal choices. To better exemplify the above concept, let us consider the simple case of two binary sources. Specifically, let X and Y be two Bernoulli sources with parameters $p = P_X(1)$ and $q = P_Y(1)$ respectively. Let also assume that the distortion constraint is expressed in terms of the Hamming distance between the sequences, that is $d(i, j) = 0$ when $i = j$ and 1 otherwise. Without loss of generality let $p > q$. The distortion associated to a transportation map S_{XY} can be written as:

$$\sum_{i,j} S_{YX}(i, j)d(i, j) = S_{YX}(0, 1) + S_{YX}(1, 0). \tag{21}$$

Since $p > q$, it is easy to conclude that the minimum of the above expression is obtained when $S_{YX}(1, 0) = 0$ (intuitively, if the source X outputs more 1's than Y, it does not make any sense to turn the 1's emitted by Y into 0's). As a consequence, to satisfy the constraint $S_X(1) = p$ we must let $S_{YX}(0, 1) = p - q$, yielding $\mathcal{SM}(P_Y, P_X) = p - q$, or more generally $|p-q|$. We can conclude that if the attacker is allowed to introduce an average Hamming distortion larger or equal than $|p-q|$, then there is no way for the defender to distinguish the two sources. This is not the case if the output of the source Y passes through a binary symmetric channel with crossover probability equal to $|p - q|$, since the output of the channel will still be distinguishable from the sequences emitted by X. Consider, for example, a simple case in which $q = 1/2$ and $p > 1/2$. Regardless of the crossover probability, the output of the channel will still be a binary source with equiprobable symbols, which is distinguishable from X given that $p > 1/2$. In other words, the two sources can not be distinguished in the presence of an attacker introducing a distortion equal to $|p - q|$, while they can be distinguished if the output of Y passes through a channel introducing the same average distortion introduced by the attacker.

V. Source Identification With Training Data

In this section, we extend the analysis to the case of source identification with training data (SI_{tr}). In such a scenario, the sources X and Y are not completely known to D and A, so they must base their actions on the knowledge of a training sequence drawn from X. In [10], it is proven that the source identification game with training data is more favorable to the attacker than the SI_{ks} game. Then one could argue that in the SI_{tr} setup the security margin between the two sources is smaller, implying that a lower distortion is sufficient to the attacker to make the sources undistinguishable. The remarkable result that we will prove in this section is that this is not the case, hence showing that the ultimate distinguishability of two sources is the same for the two games.

A. The Source Identification Game With Training Data (SI_{tr})

In the source identification game with training data, the defender must decide whether a test sequence x^n has been generated by a source X whose statistics are known through an N-sample training sequence t_D^N drawn from X. This is equivalent to deciding whether to accept or not the hypothesis H_0 that the test and the training sequences have been generated by the same source. Consequently, the acceptance region Λ is defined as the set with all the pairs of sequences (x^n, t_D^N) that D classifies as being generated by the same source. Once again, we require that Λ is a union of pairs of type classes, or equivalently, pairs of types (P, Q), where $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}_N$. As for the SI_{ks} case, the defender must ensure that P_{fp} tends to zero exponentially fast with a certain decay rate. Since P_X is not known, the constraint must be satisfied in a worst case sense, i.e. for all $P_X \in \mathcal{P}$.

Given a sequence y^n drawn from a source $Y \neq X$, the goal of the attacker is to transform y^n into a sequence z^n belonging to the acceptance region chosen by D while respecting a distortion constraint. Likewise the defender, all the information that the attacker has about X is a K-long training sequence t_A^K. By adopting the same formalism used in the previous section, the set of strategies of the attacker consists of all the possible ways of choosing an admissible transportation map to transform y^n into z^n. We consider a simple version of the game for which $K = N$ and $t_A^N = t_D^N \triangleq t^N$. We also assume N to be a linear function of n, i.e. $N = cn$.

The payoff still corresponds to the false negative error probability.

The results in [10] which are most relevant for the present analysis can be summarized as follows. Let Λ_{tr}^n be the acceptance region of the test. Let h_c denote the generalized log-likelihood ratio function defined as in [26] and [27] extended to general pmf's in $\mathcal{P}$; explicitly, given $P, Q \in \mathcal{P}$:

$$h_c(P, Q) = \mathcal{D}(P||U) + c\mathcal{D}(Q||U);$$
$$U = \frac{1}{1+c}P + \frac{c}{1+c}Q. \tag{22}$$

Let

$$\Gamma_{tr}(Q, \lambda, L_{max}) = \{P \in \mathcal{P} : \exists R \in \Lambda^*_{tr}(Q, \lambda) \text{ s.t. } EMD(P, R) \leq L_{max}\}, \quad (23)$$

where

$$\Lambda^*_{tr}(Q, \lambda) = \{P \in \mathcal{P} : h_c(P, Q) \leq \lambda\}, \quad (24)$$

For the SI_{tr} game with equal training sequences, the error exponent of the false negative error probability at the equilibrium is given by:

$$\varepsilon_{tr}(\lambda) = \min_R \left[c \cdot \mathcal{D}(R||P_X) + \min_{P \in \Gamma_{tr}(R, \lambda, L_{max})} \mathcal{D}(P||P_Y) \right]. \quad (25)$$

It follows that $\varepsilon_{tr}(\lambda) = 0$ if $P_Y \in \Gamma_{tr}(P_X, \lambda, L_{max})$, $\varepsilon_{tr}(\lambda) \neq 0$ otherwise; then, the sources that cannot be distinguished from X are those inside $\Gamma_{tr}(P_X, \lambda, L_{max})$. The only difference with respect to the case of known sources consists in the asymptotic acceptance region $\Lambda^*_{tr}(P_X, \lambda)$, which is proven to be strictly larger than $\Lambda^*(P_X, \lambda)$, given that the h_c function is always lower than $\mathcal{D}$. As a consequence, it is straightforward to argue that $\Gamma_{tr}(P_X, \lambda, L_{max}) \supset \Gamma(P_X, \lambda, L_{max})$.

B. Security Margin for the SI_{tr} Game

To study the behavior of the SI_{tr} game when $\lambda \to 0$, we observe that both $\mathcal{D}(P||Q)$ and $h_c(P, Q)$ are convex functions and are equal to zero if and only if $P = Q$. This permits to extend Property 1 to Γ_{tr} yielding:

Property 2: For any R and any two values λ_1 and λ_2 such that $\lambda_2 < \lambda_1$, $\Gamma_{tr}(R, \lambda_2, L_{max}) \subseteq \Gamma_{tr}(R, \lambda_1, L_{max})$.

In a similar way, Lemma 1 can be extended to the set $\Gamma_{tr}(R, \lambda, L_{max})$ (Appendix A).

We are now ready to prove the counterpart of Theorem 3 for the SI_{tr} game.

Theorem 4: Given two sources X and Y and a maximum allowable average per-letter distortion L_{max} (defined according to an additive distortion measure), the maximum achievable false negative error exponent for the SI_{tr} game is

$$\lim_{\lambda \to 0} \varepsilon_{tr}(\lambda) = \min_R \left[c \cdot \mathcal{D}(R||P_X) + \min_{P \in \Gamma(R, L_{max})} \mathcal{D}(P||P_Y) \right], \quad (26)$$

where $\Gamma(R, L_{max})$ is by replacing P_X with R in (14).[2]

Proof: The proof goes along the same line of the proof of Theorem 3. From Property 2, we see immediately that $\varepsilon(\lambda)$ is non-increasing when λ decreases, since the innermost minimization in equation (25) is taken over a smaller set when λ decreases. By the same token, we have:

$$\varepsilon_{tr}(\lambda) \leq \min_R \left(c\mathcal{D}(R||P_X) + \min_{P \in \Gamma(R, L_{max})} \mathcal{D}(P||P_Y) \right). \quad (27)$$

This implies that $\lim_{\lambda \to 0} \varepsilon(\lambda)$ exists and is finite. Given that Lemma 1 still holds for the set $\Gamma_{tr}(R, \lambda, L_{max})$ $\forall R$, we can reason as in the proof of Theorem 3 to conclude that:

$$\min_{P \in \Gamma_{tr}(R, \lambda, L_{max})} \mathcal{D}(P||P_Y) \geq \min_{P \in \Gamma(R, L_{max})} \mathcal{D}(P||P_Y) - \delta(\tau), \quad (28)$$

where $\delta(\tau)$ can be made arbitrarily small by decreasing λ. By adding the term $c\mathcal{D}(R||P_X)$ to both sides of (28) and considering that the relation holds for any $R \in \mathcal{P}$, we have:

$$\begin{aligned} \varepsilon_{tr}(\lambda) &= \min_R \left[c\mathcal{D}(R||P_X) + \min_{P \in \Gamma_{tr}(R, \lambda, L_{max})} \mathcal{D}(P||P_Y) \right] \\ &\geq \min_R \left[c\mathcal{D}(R||P_X) + \min_{P \in \Gamma(R, L_{max})} \mathcal{D}(P||P_Y) \right] - \delta(\tau), \end{aligned} \quad (29)$$

which concludes the proof due to the arbitrariness of $\delta(\tau)$. ■

A consequence of Theorem 4 is that $\lim_{\lambda \to 0} \varepsilon(\lambda) = 0$ if and only if $P_Y \in \Gamma(P_X, L_{max})$, which then can be seen as the smallest indistinguishability region for the SI_{tr} game. We conclude that the smallest indistinguishability regions for the two cases are equal implying that the security margin for the SI_{tr} setting, say $\mathcal{SM}_{tr}$, is the same of the SI_{ks} game, that is $\mathcal{SM}_{tr}(P_X, P_Y) = EMD(P_X, P_Y)$.

We remark that, for any allowed distortion $L_{max} < EMD(P_X, P_Y)$, the minimum value of λ which allows D to make a reliable decision in the SI_{tr} setting is lower than that in the SI_{ks} setting. However, the difference between the two settings regards the decay rate of the error probabilities, not the ultimate distinguishability of the sources.

We conclude this section by briefly discussing the SI_{tr} game with different training sequences ($t_D^N \neq t_A^K$). It is known from [10] that, as long as the length of both sequences grows linearly with n, the indistinguishability region is equal to that of the game with equal training sequences. By relying on this result, it is not difficult to prove that the security margin remains the same even for such a version of the game.

VI. Security Margin Computation

In this section we focus on the actual computation of the security margin. We first consider the case of discrete sources, then we extend the analysis to continuous sources.

Given two discrete sources $X \sim P_X$ and $Y \sim P_Y$, the computation of the security margin requires the evaluation of $EMD(P_X, P_Y)$. A closed form solution can be found only in some simple cases. More generally, the *EMD* between two sources can be computed by resorting to numerical analysis, and in fact, due to its wide use as a similarity measure in computer vision applications, several efficient algorithms have been proposed (see [28] for example). In the following, we describe a fast iterative algorithm for the computation of the *EMD* between any two sources assuming that the distortion (or cost) function has the general form $d(i, j) = |i - j|^p$, with $p \geq 1$. A case of great interest is $p = 1$ and $p = 2$, according to which the distortion between y^n and the attacked sequence z^n corresponds, respectively, to the L_1 and L_2^2 distance.

A. Hoffman's Greedy Algorithm for Computing $\mathcal{SM}$

The transportation problem we have to solve for computing $\mathcal{SM}(P_Y, P_X)$, i.e. $EMD(P_Y, P_X)$, is known in modern literature as *Hitchcock transportation problem* [29], which, in turn, can be formulated as a linear programming problem in the

[2]Note that when λ tends to 0, we do not need anymore to differentiate between the SI_{ks} and SI_{tr} games in the definition of $\Gamma(R, L_{max})$.

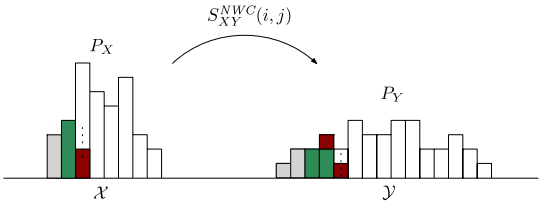


Fig. 3. Graphical representation of the NWC rule. P_X and P_Y are two generic earth piles (source and sink) $\mathcal{X}$ and $\mathcal{Y}$, while $S_{XY}^{NWC}(i,j)$ denotes the amount of earth moved from location i to j.

following way:

$$EMD(P_X, P_Y) = \min_{S_{XY}} \sum_{i,j} d(i,j) S_{XY}(i,j), \tag{30}$$

where S_{XY} must satisfy the linear constraints:

$$\begin{aligned} \sum_j S_{XY}(i,j) &= P_X(i) \quad \forall i \in \mathcal{X} \\ \sum_i S_{XY}(i,j) &= P_Y(j) \quad \forall j \in \mathcal{Y} \\ S_{XY}(i,j) &\geq 0 \quad \forall i,j, \end{aligned} \tag{31}$$

and where, by referring to the original Monge formulation [30], $S_{XY}(i,j)$ denotes the quantity of soil shipped from location (source) i to location (sink) j and $d(i,j)$ is the cost for shipping a unitary amount of soil from i to j.

A Transportation Problem (TP) like the one defined by equations (30) and (31) is a particular minimum cost flow problem [31] which, being linear, can be solved through the simplex method [32]. In general, the solution of TP depends on the cost function $d(\cdot,\cdot)$, however there are some classes of cost functions for which the solution can be found through a simple greedy algorithm. Specifically, the algorithm proposed by A.J. Hoffman in 1963 [33], allows to solve the transportation problem whenever $d(\cdot,\cdot)$ satisfies the so called Monge property [34], that is when:

$$d(i,j) + d(r,s) \leq d(i,s) + d(r,j), \tag{32}$$

$\forall (i,j,r,s)$ such that $1 \leq i < r \leq |\mathcal{X}|$ and $1 \leq j < s \leq |\mathcal{Y}|$.

It is easy to verify that Monge property is satisfied by any cost function of the form $d(i,j) = |i-j|^p$, and, more in general, by any convex function of the quantity $|i-j|$. The iterative procedure proposed by Hoffman to solve the optimal transport problem is known as *North-West Corner (*NWC*) rule* [33] and works as follows: take the bin of $\mathcal{X}$ with the smallest value and start moving its elements into the bin with the smallest value in $\mathcal{Y}$. When the smallest bin of $\mathcal{Y}$ is filled, go on with the second smallest bin in $\mathcal{Y}$. Similarly, when the smallest bin in $\mathcal{X}$ is emptied, go on with the second smallest bin in $\mathcal{X}$. The procedure is iterated until all the bins in $\mathcal{X}$ have been moved into those of $\mathcal{Y}$. The above procedure is described graphically in Figure 3. In the figure, we chose two distributions with disjoint supports for sake of clarity, however the procedure is valid regardless of how the two distributions are spread along the real line. Interestingly, the *NWC* rule does not depend explicitly on the cost matrix, so the transportation map obtained through it is the same regardless of the Monge cost. According to Hoffman's algorithm, when the cost function satisfies Monge's property, the *EMD* can be computed in linear running time: the number of elementary operations, in fact, is at most equal to $|\mathcal{X}| + |\mathcal{Y}|$. This represents a dramatic simplification with respect to the complexity required to solve a general Hitchcock transportation problem [35].

B. Security Margin for Continuous Sources

The analysis carried out in the previous sections is limited to discrete sources. When continuous sources are considered, we can quantize the probability density functions of the sources and apply the analysis for discrete sources. By letting the quantization step tend to zero, the *EMD* between P_X and P_Y can still be regarded as the security margin between the two sources. In this case, a general expression for the $\mathcal{SM}$ can be derived by considering the *continuous transportation problem* (CTP), known as Monge-Kantorovic formulation of the mass transportation problem, [22]. If the continuous cost function $c(x,y)$, $c: X \times Y \to \mathbb{R}$, satisfies the continuous Monge property [34], that is if $c(x,y)+c(x',y') \leq c(x',y)+c(x,y')$, $\forall x \leq x'$, $y \leq y'$, the optimum cumulative transportation map corresponds to the Hoeffding distribution [24], defined as:

$$C_{XY}^*(x,y) = \min\{C_X(x), C_Y(y)\}, \quad \forall (x,y) \in \mathbb{R}^2, \tag{33}$$

where $C_X(x)$ and $C_Y(y)$ are the cumulative distributions of X and Y respectively. The continuous map in (33) generalizes the *NWC* rule. Therefore, one can compute $\mathcal{SM}(P_Y, P_X)$ by evaluating the mean value $E_{XY}[c(x,y)]$ over the continuous distribution $C_{XY}^*(x,y)$. In general, however, finding a closed form expression is not an easy task.

A particularly simple and insightful formula can be obtained when the cost function corresponds to the squared Euclidean distance. Let us assume, then, that $c(x,y) = (x-y)^2$, and let X and Y be two continuous sources with means μ_X and μ_Y, variances σ_X and σ_Y and covariance $covXY$. As shown in [36] (decomposition theorem), the expectation in (12) can be rewritten as follows:

$$\begin{aligned} E_{XY}[(X-Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \\ + 2[\sigma_X \sigma_Y - covXY], \end{aligned} \tag{34}$$

where the three terms express, respectively, the difference in location, spread and shape between X and Y. Interestingly, $covXY$ is the only term depending on the joint pdf of X and Y. Then, in order to compute $\mathcal{SM}$, we only have to compute the maximum covariance over all the possible joint pdf's:

$$\begin{aligned} \mathcal{SM}_{L_2^2}(P_X, P_Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \\ + 2[\sigma_X \sigma_Y - \max_{P_{XY}: \frac{\sum_y P_{XY} = P_X}{\sum_x P_{XY} = P_Y}} covXY]. \end{aligned} \tag{35}$$

Since $0 \leq covXY \leq \sigma_X \sigma_Y$, the security margin can be bounded as follows:

$$\begin{aligned} (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \leq \mathcal{SM}_{L_2^2}(P_X, P_Y) \\ \leq (\mu_X - \mu_Y)^2 + \sigma_X^2 + \sigma_Y^2. \end{aligned} \tag{36}$$

Accordingly, whenever the distortion introduced by the attacker is less than the quantity on the left-hand side of (36), X and Y are asymptotically distinguishable, regardless of their

specific distribution. Instead, if the distortion is above this value, the distinguishability of the two sources depends on their specific probability distributions. Finally, for distortions greater than the quantity on the right-hand side of (36), there is no way to distinguish X and Y. When P_X and P_Y have the same form, for instance when the random variables X and Y are both distributed according to a Gaussian distribution, the security margin takes the minimum value and the lower bound in (36) holds with equality. In this case, in fact, it is possible to turn P_X into P_Y by imposing a deterministic relationship between X and Y, namely $Y = \frac{\sigma_Y}{\sigma_X}X + (\mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X)$; in this way, the covariance term is equal to $\sigma_X\sigma_Y$, and hence the contribution of the shape term in the security margin vanishes. This is a remarkable result stating that the distinguishability of two sources belonging to the same class depends only on their means and variances.

VII. The Security Margin With L_∞ Distance

In this section, we extend the definition of $\mathcal{SM}$ to the case in which the distortion measure constraining the attacker is expressed in terms of the maximum absolute distance between the samples of y^n and z^n, that is to the case in which the distortion is measured by relying on the L_∞ norm. In many cases, in fact, the distortion constraint must be satisfied locally, thus requiring that the maximum absolute distance between the elements of y^n and z^n is limited rather than its average. This is the case, for instance, in biomedical and remote sensing image compression, for which the maximum error introduced at each pixel location must be strictly controlled, thus calling for the adoption of near-lossless image coding schemes [37]. Another example in which the use of the L_∞ distance is recommended, is when it must be ensured that two versions of the same image, an original and a processed one, are visually indistinguishable. In such a case, it is necessary that the absolute difference between the two images is lower than the just noticeable distortion at each pixel location.

In our analysis, we will refer to the case of known sources, the extension to the SI_{tr} game being immediate.

A. The SI_{ks} Game With L_∞ Distance

We start by observing that the adoption of the L_∞ distance requires that the SI_{ks} game is, partly, redefined due to the non-additive nature of the distortion constraint. In this case, in fact, it does not make any sense to define the distortion constraint in terms of average per-letter distortion and let the overall allowed distortion to increase with n.

Similarly to the previous cases, it is possible to express the distortion constraint by limiting the set of transportation maps the attacker can choose from. Specifically, the maximum distance between y^n and z^n can be rewritten as follows:

$$d_{L_\infty}(y^n, z^n) = \max_j |z_j - y_j| = \max_{(i,j):S^n_{YZ}(i,j)\neq 0} |i-j|. \quad (37)$$

By using the above formula in the definition of the set of admissible maps (i.e. in the second line of (2)), we can still define the set of strategies of the attacker as the set of rules associating an admissible map to the to-be-attacked sequence, as in (3). In the following, we will refer to the set of admissible maps resulting from the use of the L_∞ distance as $\mathcal{A}^n_{L_\infty}(L_{max}, P_{y^n})$.

Passing to the analysis of the indistinguishability region, it is easy to see that relation (7) continues to hold by replacing $\mathcal{A}^n(L_{max}, P_{y^n})$ with $\mathcal{A}^n_{L_\infty}(L_{max}, P_{y^n})$. In fact, the dominant strategy for the defender does not depend on the set of strategies available to the attacker. The asymptotic version of $\Gamma^n_{L_\infty}(P_X, \lambda, L_{max})$ can also be defined as in (8), namely:

$$\Gamma_{L_\infty}(P_X, \lambda, L_{max}) = \{P \in \mathcal{P} : \exists\ S_{YZ} \in \mathcal{A}_{L_\infty}(L_{max}, P) \text{ s.t. } S_Z \in \Lambda^*(P_X, \lambda)\}, \quad (38)$$

where $\mathcal{A}_{L_\infty}(L_{max}, P)$ is the asymptotic counterpart of $\mathcal{A}^n_{L_\infty}(L_{max}, P)$. The next step requires the extension of Theorem 2 to the SI_{ks} game with L_∞ distance, that is we need to prove that the set in (38) contains all the sources that can not be distinguished from X because of the attack, even when the length of the observed sequence tends to infinity. This is a critical step since such theorem was proved in [11] by assuming an additive distortion measure, which is not the case when the L_∞ distance is adopted. Roughly speaking, we need to prove that when $n \to \infty$ the elements of $\Gamma^n_{L_\infty}(P_X, \lambda, L_{max})$ are dense in $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$ (in which case Theorem 2 can be proven in a way similar to Sanov's Theorem [11]). More formally, we need to prove that for any $P_Y \in \Gamma_{L_\infty}(P_X, \lambda, L_{max})$ and any $\delta > 0$, a pmf $Q^n \in \Gamma^n_{L_\infty}(P_X, \lambda, L_{max})$ exists such that the distance between P_Y and Q^n is smaller than δ. The proof requires only some minor modifications with respect to the proof of [8, Th. 2] and is omitted for sake of brevity.

B. Security Margin for the SI_{ks} Game With L_∞ Distance

As a next step, we must study the behavior of the indistinguishability region when $\lambda \to 0$. As we will see, even if the adoption of a distance based on the L_∞ norm prevents a direct formulation of the problem in terms of *EMD*, the distinguishability of X and Y is still closely related to the optimal transportation map between P_X and P_Y. Such a connection is rooted in the following property.

Property 3: Given two distributions P and Q, the transportation map S^{NWC}_{PQ} obtained by applying the NWC *rule to P and Q is a solution of the problem*

$$\min_{S_{YZ}:S_Y=P,S_Z=Q} \left(\max_{(i,j)\in S_{YZ}(i,j)\neq 0} |i-j| \right). \quad (39)$$

Proof: Let $S^* \neq S^{NWC}_{PQ}$ be a generic transformation mapping P into Q. Given that $S^* \neq S^{NWC}_{PQ}$ there exists at least one quadruple of bins (t, r, v, s), with $t < r$ and $v < s$, for which, $S^*(t, s) > 0$ and $S^*(r, v) > 0$. Let us assume, without loss of generality, that $S^*(t, s) \leq S^*(r, v)$. We now define a new map S' which is obtained from S^* by letting:

$$\begin{aligned} S'(t, v) &= S^*(t, v) + S^*(t, s) \\ S'(t, s) &= 0 \\ S'(r, v) &= S^*(r, v) - S^*(t, s) \\ S'(r, s) &= S^*(r, s) + S^*(t, s). \end{aligned} \quad (40)$$

Since $\max\{|t-s|, |r-v|\} > \max\{|t-v|, |r-s|\}$, the maximum distortion introduced by S' is lower than or equal to that introduced by S^*, that is:

$$\max_{(i,j)\in S^*(i,j)\neq 0} |i-j| \geq \max_{(i,j)\in S'(i,j)\neq 0} |i-j|. \quad (41)$$

We now inspect S', if there is another quadruple of bins (t', r', v', s') satisfying the same properties of (t, r, v, s), we let $S^* = S'$ and iterate the above procedure. The process ends when no quadruple of bins with the required properties exists and hence when $S' = S_{PQ}^{NWC}$. Since at each step the distortion introduced by the new map does not increase, the above procedure proves that S_{PQ}^{NWC} introduces a distortion lower than or equal to that introduced by any other S^* mapping P into Q, thus proving that S_{PQ}^{NWC} achieves the minimum in (39). ∎

Thanks to Property 3, the set $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$ in (38) can be rewritten as follows:

$$\Gamma_{L_\infty}(P_X, \lambda, L_{max}) = \{P \in \mathcal{P} : \exists\, Q \in \Lambda^*(P_X, \lambda) \text{ s.t } \max_{(i,j):S_{PQ}^{NWC}(i,j)\neq 0} |i-j| \leq L_{max}\}. \quad (42)$$

By letting λ tend to 0, we obtain the smallest indistinguishability region, thus extending Theorem 3 to the SI_{ks} game with L_∞ distance.

Theorem 5: Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum allowable per-letter distortion L_{max}, and given:

$$\Gamma(P_X, L_{max}) = \{P \in \mathcal{P} : \max_{(i,j)\in S_{PP_X}^{\text{NWC}}} |i-j| \leq L_{max}\}, \quad (43)$$

the maximum achievable false negative error exponent ε for the SI_{ks} game with L_∞ distance is

$$\lim_{\lambda\to 0}\lim_{n\to\infty} -\frac{1}{n}\log P_{fn} = \min_{P\in\Gamma_{L_\infty}(P_X, L_{max})} \mathcal{D}(P||P_Y). \quad (44)$$

Proof: The proof relies on the extension of Property 1 and Lemma 1 to the L_∞ case. The extension of Property 1 is immediate since, once again, the indistinguishability region depends on λ only through $\Lambda^*(P_X, \lambda)$, whose form does not depend on the particular norm adopted to express the distortion constraint. The extension of Lemma 1 requires some more care and is proven in Appendix B. For the rest, the theorem can be proven by reasoning as in the proof of Theorem 3. ∎

As a consequence of Theorem 5, the distinguishability of two sources depends again on the optimum transportation map between the pmf's of the sources. Specifically, the defender is able to distinguish between X and Y if and only if

$$\max_{(i,j)\in S_{P_Y P_X}^{NWC}} |i-j| > L_{max}. \quad (45)$$

Condition (45) can be used to determine the maximum attacking distortion for which D is able to distinguish X and Y, i.e. $\mathcal{SM}(P_X, P_Y)$.

Definition 2 (Security Margin for the L_∞ Case): Let X and Y be two discrete memoryless sources. The maximum distortion for which the two sources can be reliably distinguished in the SI_{ks} setting with L_∞ distance is given by

$$\mathcal{SM}_{L_\infty}(P_Y, P_X) = \max_{(i,j):S_{P_Y P_X}^{\text{NWC}}(i,j)\neq 0} |i-j|, \quad (46)$$

where $S_{P_Y P_X}^{\text{NWC}}$ is obtained by applying the NWC *rule to map P_Y into P_X.*

Even if we proved Theorem 5 for the case of known sources, it is possible to extend it to the SI_{tr} game. The proof goes along the same lines of the SI_{ks} case and is omitted.

VIII. Use of $\mathcal{SM}$ in Practical Applications

The Security Margin is a powerful concept which permits to summarize into a single quantity the asymptotic behavior of the game between the attacker and the defender. Its practical application, however, poses a number of problems due to the assumptions behind the definition of $\mathcal{SM}$. In this section we first discuss the impact of these assumptions in real applications, then we present the possible use of $\mathcal{SM}$ within a multimedia forensics scenario.

A. Impacts of Theoretical Assumptions in Practical Setups

The two main assumptions behind our analysis are that the sources X and Y are memoryless, and that the defender relies only on first order statistics to make his decision. We stress that by first order statistics we mean all the statistics that can be derived from the analysis of the relative occurrences of the symbols within the observed sequence, including high order moments like, for instance, the empirical skewness and the kurtosis of the sequence. On the other hand, they do not include joint statistics among samples, like transition probabilities and co-occurrence matrices [38].

As as a first observation, we note that while the use of first order statistics may seem to be fully justified by the DMS assumption for X and Y, this is not necessarily the case, and we must explicitly set it as a working condition. The use of first order statistics to distinguish between two discrete memoryless sources, in fact, is optimum only when no attack is present [11]. In general, the attacker could introduce memory within z^n, thus making the use of first order statistics suboptimum. As an alternative path, we could have imposed that the attack corresponds to a memoryless channel, since in this case first order statistics would represent a set of sufficient statistics for the source identification test. By proceeding in this way, however, we would simply move the first order constraint from the defender to the attacker.[3]

From a practical point of view, the main problem with the memoryless assumption, is that it may not be met in real-world applications. Real signals, like images, for instance, can not be assimilated to memoryless sources and consequently, the defender could decide to go beyond first order statistics to make his decision. In some cases, the memoryless assumption can be justified because the defender operates in a transformed domain, e.g the DCT domain, or in a random projections domain [39]. In any event, the use of first order detectors is quite common in real applications even when dealing with correlated sources. In the case of image forensics, for instance, several techniques rely only on the analysis of the image histogram or a subset of features derived from it. Since even in

[3]By adopting the defender's point of view, avoiding to impose any additional constraint on the action of the attacker may be interpreted as a worst case assumption.

the case of sources with memory by the law of large numbers the sources will end up generating sequences with a type arbitrarily close to the marginal pmf, we conjecture that the definition and the meaning of the security margin remains the same, as long as the defender decides to rely only on the empirical marginal distributions for his analysis.

Another assumption underlying the theoretical analysis which may not be valid in practice is that X and Y are stationary sources. Time varying sources are encountered in many practical applications. In PRNU-based camera identification [40], for instance, images produced by a specific camera are detected due to the presence of a distinctive time varying signature, the Photo Response Non-Uniformity noise, introduced by the camera during image acquisition. Another example is given by the camera model identification scheme presented in [41], where the time varying nature of the images prevent the use of a stationary noise model. Other examples can be drawn from biometric recognition, where the biometric templates used for identity verification can not be assimilated to stationary signals [3], and steganalysis, where the cover image is sometimes modelled as a sequence of independent Gaussian variable with different variance [42]. Yet, even when dealing with time varying signals, the use of first order statistics obtained by a global analysis of the analyzed signal is common practice. This is the case, for instance, of the detection of histogram-based image enhancement which, due to the time varying nature of the underlying image, can not be described by a memoryless model, but is usually faced with by resorting to first order detectors [43]. Even in biometrics, first order statistics are sometimes used instead of more powerful joint statistics, like in [44], where the adoption of the arbitrarily varying sources (AVS) model [45] permits to account for a (slightly) time-varying behavior of the sources and justifies the resort to a memoryless formulation of the problem. Even in the case of time-varying sources, then, the use of the security margin concept is not totally unrealistic.

B. $\mathcal{SM}$ in Data-Driven Image Forensics

Source identification is one of the most common problems in image and multimedia forensics. In fact, gathering information about the device that was used to produce a certain image plays a crucial role in many investigations. In a similar way, the analyst may be interested to decide if a certain processing operator has been applied to a given image, that is to distinguish between the class of images that underwent a certain processing and those which did not. When a statistical model for the two classes of images is available, the $\mathcal{SM}$ between the two classes can be calculated as detailed in the previous sections. In most cases, though, such model does not exist. In these cases, the forensic analyst may adopt a data-driven approach, usually based on machine learning techniques, wherein the characteristics of the image classes are derived from a number of examples. Let, then, $\mathcal{K}_1$ and $\mathcal{K}_2$ be two classes of images, for instance images acquired by a scanner and images produced by a camera. Given a test image I, the goal of the defender is to accept or reject the hypothesis that I belongs to $\mathcal{K}_1$. To make his decision, the defender can rely on two sets of sample images (often referred to as training sets) belonging to $\mathcal{K}_1$ and $\mathcal{K}_2$, let us call such sets $\mathcal{T}_1$ and $\mathcal{T}_2$. Moreover, let us assume that the defender relies only on the first order statistics of I, that is the image histogram h_I. The goal of the attacker is to take an image J belonging to $\mathcal{K}_2$ and modify it in such a way that the defender classifies it as belonging to $\mathcal{K}_1$. Even if the theoretical formulation leading to the definition of $\mathcal{SM}$ can not be directly applied, we can argue that *in some sense* the security margin between J and $\mathcal{T}_1$ (which is the only available representation of $\mathcal{K}_1$) is the minimum *EMD* between h_J and the histograms of the images in $\mathcal{T}_1$, namely

$$\mathcal{SM}(J, \mathcal{T}_1) = \min_{I \in \mathcal{T}_1} EMD(h_J, h_I). \tag{47}$$

In fact, if the distortion allowed to the attacker is larger than $\mathcal{SM}(J, \mathcal{T}_1)$, A can modify J in such a way that its histogram is equal to the histogram of one of the images in $\mathcal{T}_1$, thus making a reliable distinction impossible. In the same way, we could define the $\mathcal{SM}$ between two classes of images as the average minimum *EMD* between the histograms of the images in one class and those of the images in the other class:

$$\mathcal{SM}(\mathcal{T}_2, \mathcal{T}_1) = \frac{1}{|\mathcal{T}_2|} \sum_{J \in \mathcal{T}_2} \min_{I \in \mathcal{T}_1} EMD(h_J, h_I). \tag{48}$$

A similar analysis can be applied when the distinction between the classes $\mathcal{K}_1$ and $\mathcal{K}_2$ is carried out in a transformed domain, e.g. the block DCT domain.

In the next paragraph, we exemplify the above ideas by applying them to a well-known problem in image forensics.

1) Histogram-Based Detection of Contrast Enhancement: Detection of the traces left within an image by contrast-enhancement operators is an active research topic in image forensics. In fact, knowing that an image, or part of it, has been subject to a contrast enhancement operator may help understanding the history of the image, and whether some parts of the image have been cut-and-pasted from another image which underwent a different processing history. In some cases, the analysis must be carried out by taking into account the possibility that an adversary has modified the contrast-enhanced image so to hinder the analysis. Given that most contrast enhancement operators work directly on the image histogram, forensic tools for contrast-enhancement detection usually rely on the analysis of the image histogram and hence fit well the theoretical setup adopted in this paper [43], [46]. In this framework, estimating the $\mathcal{SM}$ between the classes of original and contrast-enhanced images as specified in equations (47) and (48), may help to understand how difficult is for the adversary to completely delete the traces left by the enhancement operator.

To exemplify the above ideas we considered the images contained in the MIRFLICKR dataset [47]. These are 25,000 original, never processed images of size 333×500. We randomly split the images in two sets $\mathcal{T}_1$ containing 24,000 images and $\mathcal{T}_2$ with 1,000 images. Then we contrast-enhanced the images in $\mathcal{T}_2$ by applying a gamma correction operator with various γ [48]. Eventually, we used equation (47) to compute the $\mathcal{SM}$ between the images in $\mathcal{T}_2$ and $\mathcal{T}_1$. The results we

TABLE I

AVERAGE $\mathcal{SM}$ BETWEEN $\mathcal{T}_1$ AND $\mathcal{T}_2$ FOR VARIOUS VALUES OF γ

	γ					
	0.3	0.8	1.3	1.8	2.3	2.8
$\mathcal{SM}_{L^\infty}$	27.3	13.8	13.8	14.9	16.1	17.4
$\mathcal{SM}_{L_2^2}$	48.8	27.4	25.8	26.1	26.3	26.8

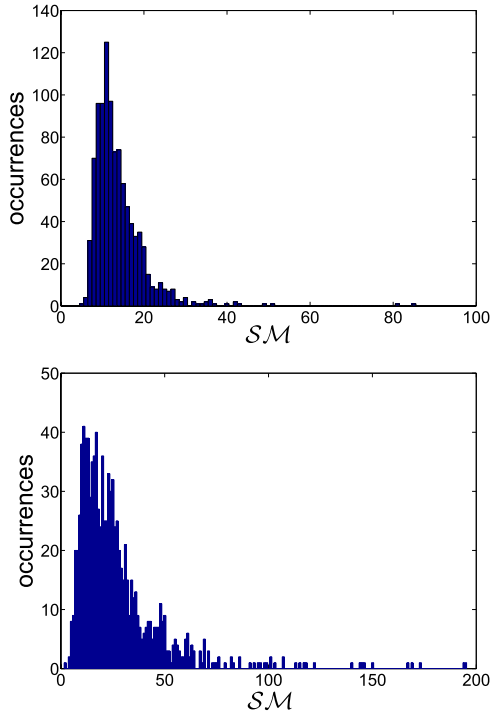


Fig. 4. Distribution of the $\mathcal{SM}$ across the images in $\mathcal{T}_2$ for the case of L_∞ (above) and L_2^2 (below) distance. The strength of the enhancement operator is $\gamma = 0.8$.

obtained are reported in Fig. 4 where we show the distribution of the $\mathcal{SM}$ across all the images in $\mathcal{T}_2$ for both the cases of squared Euclidean distance and maximum distance. The $\mathcal{SM}$ ranges from a minimum of 1.6 to a maximum of 195.3 for the square Euclidean distance, and from 5 to 85 for the L_∞ case. In Table I, we show the average $\mathcal{SM}$, computed as stated in (48), for different values of γ. The values in the table suggest that for a perfect concealment of the traces left by the γ correction operator, the attacker must introduce an average square distortion in the order of 25 and a maximum (non-squared) distortion in the order of 13-15 grey levels.

We conclude this section by observing that the values given Fig. 4 and Table I must be interpreted with care. First of all, they ensure the success of the attack asymptotically and in the presence of an optimum detector. Deceiving practical forensic operators may be significantly easier, and hence may require a considerably lower distortion. Secondly, the visual impact of the attack can not be measured only in terms of L_2^2 or even L_∞ distance, since it also depends on how the attack is implemented in the pixel domain, that is on which specific pixels are chosen to realise the mapping defined by the NWC rule (readers may refer to [49] for an example of the visual impact that a practical implementation of histogram remapping has in the pixel domain).

IX. CONCLUSIONS

By interpreting the attacker's optimum strategy in the SI_{ks} (and SI_{tr}) game as the solution of an optimum transport problem, we have introduced the concept of security margin, a single measure summarising the distinguishability of two sources under adversarial conditions. We also described an efficient algorithm to compute the security margin between several classes of sources. By relying on the security margin concept, we can understand who between the attacker and the defender is going to asymptotically win the source identification game.

The analysis carried out in this paper can be extended in several directions, with different difficulty levels. As a first extension, we mention a scenario in which the system under analysis is observed through a noisy (memoryless) channel. If the attacker acts after the channel, then $\mathcal{SM}$ can be calculated both at the input and the output of the channel, to measure the security loss caused by the channel. In case the attacker acts before the channel, the situation is slightly more involved, since the attacker must take into account the presence of the channel when devising the optimum attack. To calculate the security margin, then, we must consider the backward channel having at the input the sequence observed by the defender and at the output the attacked sequence (a similar approach is used in [50] for biometric identification). As an alternative setup we could consider the effect that the maximum transmission rate allowed by the channel has on source distinguishability, linking the security margin to the degradation introduced by the channel in a typical rate distortion setup.

Another possible extension (already mentioned in Section VIII), regards the generalization of the security margin concept to the case of Markov sources of finite order. The method of types, in fact, still holds in this case [51], hence making it possible to reformulate our main theorems for such sources.

A more far reaching extension regards the case of multiple source identification, or classification. Though non-trivial, such an extension could be applied to a large number of practical applications, including biometric identification, multiple-camera identification, multiple JPEG compression and so on.

APPENDIX

A. Behavior of the Sets Γ and Γ_{tr} for $\lambda \to 0$

We start by showing that for small values of λ, $\Gamma(P_X, \lambda, L_{max})$ approaches $\Gamma(P_X, L_{max})$ smoothly. As a first step, we prove the following property.

Property 4: $\mathrm{EMD}(P, Q)$ *is a continuous and convex function of* P *and* Q.

Proof: Property 4 follows immediately if we look at the *EMD* as the solution of a Linear Programming (LP) problem (see Section VI-A), wherein P and Q are the known terms of the linear constraints. In fact, it is a known result in operations research that the minimum of the objective function of an LP problem is a continuous and convex function of the known terms of the linear constraints [52]. ■

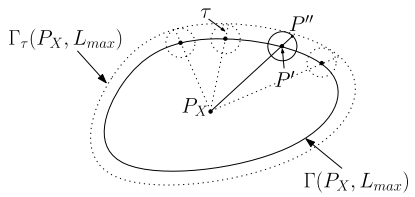


Fig. 5. Graphical representation of the set $\Gamma_\tau(P_X, L_{max})$.

By exploiting the continuity of the divergence and the continuity and convexity of the *EMD*, we now show that when λ tends to 0, $\Gamma(P_X, \lambda, L_{max})$ tends to $\Gamma(P_X, L_{max})$ regularly.

Lemma 1: Let $X \sim P_X$ be an information source and L_{max} the maximum allowable average per-letter distortion in the SI_{ks} game. The set $\Gamma(P_X, \lambda, L_{max})$, defined in (13)*, satisfies the following property:*

$$\forall \tau > 0, \quad \exists \lambda > 0 \ \ s.t.\ \forall P \in \Gamma(P_X, \lambda, L_{max}) \\ \exists P' \in \Gamma(P_X, L_{max}) \ \ s.t.\ P \in B(P', \tau), \tag{A1}$$

where $\Gamma(P_X, L_{max})$ is defined as in (14) *and $B(P', \tau)$ is a ball centered in P' with radius τ.*

Proof: Throughout the proof we will refer to Figure 5 where all the sets and quantities involved in the proof are sketched. For any $\tau > 0$, we consider the set:

$$\Gamma_\tau(P_X, L_{max}) \\ = \{P : \exists P' \in \Gamma(P_X, L_{max}) \ \ \text{s.t. } P \in B(P', \tau)\}. \tag{A2}$$

With such a definition, we can rephrase (A1) as follows:

$$\forall \tau > 0, \exists \lambda > 0 \text{ s.t. } \Gamma(P_X, \lambda, L_{max}) \subseteq \Gamma_\tau(P_X, L_{max}). \tag{A3}$$

For sake of simplicity, we will prove a slightly stronger version of the lemma by means of the following two-step proof. First, we will show that a subset of $\Gamma_\tau(P_X, L_{max})$ exists having the following form:

$$\Gamma_\tau^{sub}(P_X, L_{max}) = \{P : EMD(P, P_X) \le L_{max} + \delta(\tau)\}, \tag{A4}$$

for some $\delta(\tau) > 0$. Then, we will prove that for small enough λ, any $P \in \Gamma(P_X, \lambda, L_{max})$ belongs to $\Gamma_\tau^{sub}(P_X, L_{max})$.

To start with, let P' be any point in $\mathcal{B}(\Gamma(P_X, L_{max}))$, the boundary of $\Gamma(P_X, L_{max})$. Among all the the points belonging to the boundary of the ball of radius τ and centered in P', consider the one, name it P'', lying along the direction given by the line joining P_X and P' and falling outside $\Gamma(P_X, L_{max})$ (see Figure 5). By the convexity of the *EMD* (Property 4) and since $EMD = 0$ if and only if $P = P_X$, we conclude that $EMD(P'', P_X) > EMD(P', P_X)$. Since P' lies on the boundary of $\Gamma(P_X, L_{max})$, we know that $EMD(P'', P_X) = L_{max} + \mu$, where $\mu = \mu(P', \tau)$ is a strictly positive quantity. We now show that the first part the proof holds by letting $\delta(\tau) = \min_{P' \in \mathcal{B}(\Gamma(P_X, L_{max}))} \mu(P', \tau)$. To this purpose, let P be any point in $\Gamma_\tau^{sub}(P_X, L_{max})$ for the above choice of $\delta(\tau)$. If $P \in \Gamma(P_X, L_{max})$, then, by definition, P also belongs to $\Gamma_\tau(P_X, L_{max})$. On the other hand, if P lies outside $\Gamma(P_X, L_{max})$, let us denote by P^* the point lying on the boundary of the set $\Gamma(P_X, L_{max})$ along the line joining P and P_X, and let P^{**} be the point where the same line crosses the ball $B(P^*, \tau)$ outside $\Gamma(P_X, L_{max})$. Now, $EMD(P, P_X) \le L_{max} + \delta(\tau) \le EMD(P^{**}, P_X)$ by construction. Because of the convexity of *EMD*, then $P \in B(P^*, \tau)$ as required.

Let us now pass to the second part of the proof. First, we observe that the set $\Gamma(P_X, \lambda, L_{max})$ depends on λ only through the acceptance region $\Lambda^*(P_X, \lambda)$. If λ is small, due to the continuity of the divergence, for any $Q \in \Lambda^*(P_X, \lambda)$ we have $Q \in B(P_X, \kappa(\lambda))$ for some $\kappa(\lambda)$ such that $\kappa(\lambda) \to 0$ when $\lambda \to 0$. Let, then, P be a pmf in $\Gamma(P_X, \lambda, L_{max})$. By definition, a $Q \in \Lambda^*(P_X, \lambda)$ exists s.t. $EMD(P, Q) \le L_{max}$. If λ is small, due to the proximity of Q to P_X and the continuity of the *EMD* we have that $EMD(P, P_X) < EMD(P, Q) + \eta(\lambda) \le L_{max} + \eta(\lambda)$ with $\eta(\lambda)$ approaching 0 when $\lambda \to 0$. In particular, if λ is small enough $\eta(\lambda) < \delta(\tau)$ and hence $P \in \Gamma_\tau^{sub}(P_X, L_{max})$ which in turn is entirely contained in $\Gamma_\tau(P_X, L_{max})$ thus completing the proof. ■

In the same way, we can prove that Lemma 1 holds also when $\Gamma(P_X, \lambda, L_{max})$ is replaced by $\Gamma_{tr}(R, \lambda, L_{max})$ and $\Gamma(P_X, L_{max})$ by $\Gamma(R, L_{max})$ with a generic R instead of P_X. To be convinced about that, it is sufficient to note that the only difference between Γ and Γ_{tr} is the test function which defines the acceptance region, respectively the divergence and the h_c function. Since h_c is still a continuous and convex function and, likewise $\mathcal{D}$, is equal to zero if and only if its arguments are identical, the proof that we used for Lemma 1 still holds.

B. Behavior of $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$ for $\lambda \to 0$

We prove that when $\lambda \to 0$, $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$ approaches $\Gamma_{L_\infty}(P_X, L_{max})$ regularly, as stated by the following lemma.

Lemma 2 (Extension of Lemma 1 to the L_∞ Case): Let $X \sim P_X$ be an information source and L_{max} the maximum per-sample distortion allowed to the attacker. The set $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$, defined in Section VII, satisfies the following property:

$$\forall \tau > 0, \quad \exists \lambda > 0 \ s.t., \ \forall P \in \Gamma_{L_\infty}(P_X, \lambda, L_{max}) \\ \exists P' \in \Gamma_{L_\infty}(P_X, L_{max}) \ s.t.\ P \in B(P', \tau), \tag{A5}$$

where $B(P', \tau)$ is a ball centered in P' with radius τ.

Proof: We will prove the lemma by assuming that the distance defining the ball $B(P', \tau)$ is the L_1 distance, extending the proof to other distances being straightforward.

For a fixed $\tau > 0$, let P be a pmf in $\Gamma_{L_\infty}(P_X, \lambda, L_{max})$ for some λ. This means that at least one pmf $Q \in \Lambda^*(P_X, \lambda)$ exists, such that P can be mapped into Q with maximum shipment distance lower than or equal to L_{max}. From equation (9) and by exploiting the continuity of the divergence function, we argue that $Q \in \mathcal{B}(P_X, \gamma(\lambda))$ for some positive $\gamma(\lambda)$, and where $\gamma(\lambda) \to 0$ as $\lambda \to 0$. Accordingly, P_X can be written as $P_X(j) = Q(j) + \gamma(j)$, $\forall j$, where $\sum_{j \in \mathcal{X}} |\gamma(j)| < \gamma(\lambda)$. Note that, by construction, $\sum_j \gamma(j) = 0$ and $\gamma(j) \to 0$ when $\lambda \to 0$. Let S_{PQ} be an admissible map bringing P into Q (such a map surely exists by construction). We prove the lemma by explicitly building a pmf P' and a new admissible transportation map S', such that, P' is arbitrarily close to P (for a small enough λ) and S' maps P' into P_X.

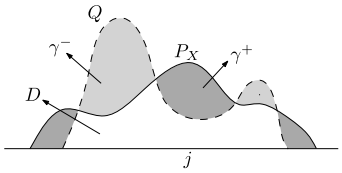


Fig. 6. Geometric interpretation of γ^+, γ^- and $D(j)$.

We first introduce two new quantities, namely $\gamma^+(j)$, defined as follows:

$$\gamma^+(j) = \gamma(j) \quad \text{if } P_X(j) - Q(j) \geq 0$$
$$\gamma^+(j) = 0 \quad \text{if } P_X(j) - Q(j) < 0, \tag{A6}$$

and $\gamma^-(j)$ defined as

$$\gamma^-(j) = -\gamma(j) \quad \text{if } P_X(j) - Q(j) < 0$$
$$\gamma^-(j) = 0 \quad \text{if } P_X(j) - Q(j) \geq 0. \tag{A7}$$

A graphical interpretation of γ^+ and γ^- is given in Figure 6. Clearly, $\sum_j \gamma^-(j) = \sum_j \gamma^+(j)$. With the above definitions, we can look at the demand distribution Q as consisting of two amounts: the mass distribution D, with $D(j) = \min\{P_X(j), Q(j)\}$, and γ^-. According to the superposition principle, the map S_{PQ} can then be split into two sub-maps: one which satisfies the demand of D (let us call it S^D_{PQ}), and one that satisfies the demand of γ^- (let us call it S^γ_{PQ}). The same distinction can be made in the source distribution:

$$P(i) = \sum_j S^D_{PQ}(i,j) + \sum_j S^\gamma_{PQ}(i,j) = P_D(i) + P_\gamma(i), \tag{A8}$$

where P_D and P_γ are the masses in the source distribution which are used to satisfy the mass demand pertaining to D and γ^- according to the mapping S_{PQ}. Then, $\sum_i P_D(i) = D$ and $\sum_i P_\gamma(i) = \gamma^-$. In order to construct the pmf P' we are looking for, we simply remove from P the amount of mass P_γ used to fill γ^- and redistribute it according to γ^+. Specifically, we have

$$P'(i) = P_D(i) + \gamma^+(i) \tag{A9}$$
$$S'(i,j) = S^D_{PQ}(i,j) + \gamma^+(j)\delta(i,j), \tag{A10}$$

where $\delta(i,j)$ is equal to 1 if $i = j$ and 0 otherwise. It is easy to see that applying the transportation map $S'(i,j)$ to P' yields P_X. Besides, from the procedure adopted to build S', it is evident that

$$\max_{(i,j):S'(i,j)\neq 0} |i-j| \leq \max_{(i,j):S_{PQ}(i,j)\neq 0} |i-j| \leq L_{max}, \tag{A11}$$

(the only new shipments introduced are from a bin to itself). In addition, the distance between P' and P is, by construction, lower than $\gamma(\lambda)$, which can be made arbitrarily small by decreasing λ, thus completing the proof of the lemma. ∎

REFERENCES

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 8682–8686.

[2] E. Delp, N. Memon, and M. Wu, "Digital forensics [From the Guest Editors]," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 14–15, Mar. 2009.

[3] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2005, pp. 1–4.

[4] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: Joint operation with a verification system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2013, pp. 98–104.

[5] I. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. San Mateo, CA, USA: Morgan Kaufmann, 2002.

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

[7] K. M. C. Tan, K. S. Killourhy, and R. A. Maxion, "Undermining an anomaly-based intrusion detection system using common exploits," in *Proc. Recent Adv. Intrusion Detect. (RAID)*, Zürich, Switzerland, Oct. 2002, pp. 54–73.

[8] M. Barni and B. Tondi, "The source identification game: An information-theoretic perspective," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 450–463, Mar. 2013.

[9] M. Barni and B. Tondi, "Optimum forensic and counter-forensic strategies for source identification with training data," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Tenerife, Spain, Dec. 2012, pp. 199–204.

[10] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4848–4866, Aug. 2014.

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

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

[13] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[14] P. Comesaña-Alfaro and F. Pérez-González, "Optimal counterforensics for histogram-based forensics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3048–3052.

[15] F. Balado, "The role of permutation coding in minimum-distortion perfect counterforensics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 6240–6244.

[16] P. Comesana and F. Pérez-González, "The optimal attack to histogram-based forensic detectors is simple(x)," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2014, pp. 137–142.

[17] F. Balado and D. Haughton, "Permutation codes and steganography," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 2954–2958.

[18] M. Barni and B. Tondi, "The security margin: A measure of source distinguishability under adversarial conditions," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Austin, TX, USA, Dec. 2013, pp. 225–228.

[19] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

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

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

[22] S. T. Rachev, *Mass Transportation Problems: Theory*, vol. 1. New York, NY, USA: Springer, 1998.

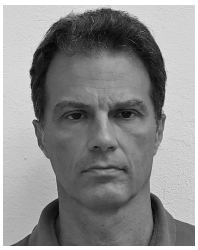
[23] C. Villani, *Topics in Optimal Transportation* (Graduate Studies in Mathematics), vol. 58. Providence, RI, USA: AMS, 2003.

[24] S. T. Rachev, "The Monge–Kantorovich mass transference problem and its stochastic applications," *Theory Probab. Appl.*, vol. 29, no. 4, pp. 647–676, 1985.

[25] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Jul. 2001, pp. 251–256.

[26] 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.

[27] M. Kendall and S. Stuart, *The Advanced Theory of Statistics*, vol. 2, 4th ed. New York, NY, USA: MacMillan, 1979.
[28] O. Pele and M. Werman, "Fast and robust Earth mover's distances," in *Proc. 12th IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 460–467.
[29] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *J. Math. Phys.*, vol. 20, nos. 1–4, pp. 224–230, Apr. 1941.
[30] G. Monge, *Mémoire sur la théorie des Déblais et des Remblais*. De l'Imprimerie Royale, Paris: Histoire de l'Academie Royale des Sciences, 1781, pp. 666–704.
[31] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[32] V. Chvatal, *Linear Programming* (Series of Books in the Mathematical Sciences), vol. 1. New York, NY, USA: Freeman, 1983, 1983.
[33] A. Hoffman, "On simple linear programming problems," in *Proc. Symp. Pure Math.*, vol. 7. 1963, pp. 317–327.
[34] R. E. Burkard, B. Klinz, and R. Rudolf, "Perspectives of Monge properties in optimization," *Discrete Appl. Math.*, vol. 70, no. 2, pp. 95–161, Sep. 1996.
[35] J. B. Orlin, "A faster strongly polynomial minimum cost flow algorithm," *Oper. Res.*, vol. 41, no. 2, pp. 338–350, 1993.
[36] A. Irpino and E. Romano, "Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation," in *Proc. EGC*, vol. RNTI-E-9, 2007, pp. 99–110.
[37] R. Ansari, N. Memon, and E. Ceran, "Near-lossless image compression techniques," *J. Electron. Imag.*, vol. 7, no. 3, pp. 486–494, 1998. [Online]. Available: http://dx.doi.org/10.1117/1.482591
[38] S. W. Zucker and D. Terzopoulos, "Finding structure in Co-occurrence matrices for texture analysis," *Comput. Graph. Image Process.*, vol. 12, no. 3, pp. 286–308, Mar. 1980.
[39] J. E. Fowler and Q. Du, "Anomaly detection and reconstruction from random projections," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 184–195, Jan. 2012.
[40] M. Chen, J. Fridrich, M. Goljan, and J. Lukás, "Determining image origin and integrity using sensor noise," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 74–90, Mar. 2008.
[41] T. H. Thai, R. Cogranne, and F. Retraint, "Camera model identification based on the heteroscedastic noise model," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 250–263, Jan. 2014.
[42] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 221–234, Feb. 2016.
[43] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 515–525, Mar. 2014.
[44] A. N. Harutyunyan, N. Grigoryan, S. Voloshynovskiy, and O. J. Koval, "A new biometric identification model and the multiple hypothesis testing for arbitrarily varying objects," in *Proc. CAST Workshop-Biometrie (BIOSIG)*, 2011, pp. 305–312.
[45] F.-W. Fu and S.-Y. Shen, "Hypothesis testing for arbitrarily varying source with exponential-type constraint," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 892–895, Mar. 1998.
[46] M. Stamm and K. J. R. Liu, "Blind forensics of contrast enhancement in digital images," in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 3112–3115.
[47] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[48] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley, 1992.
[49] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. ACM Multimedia Secur. Workshop*, Coventry, U.K., Sep. 2012, pp. 97–104.
[50] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2003, p. 82.
[51] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1393–1402, Jul. 2004.
[52] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, 1st ed. Belmont, MA, USA: Athena Scientific, 1997.



Mauro Barni (M'92–SM'06–F'12) received the degree in electronics engineering from the University of Florence, in 1991, and the Ph.D. degree in informatics and telecommunications in 1995. He has carried out his research activity for over 20 years, first with the Department of Electronics and Telecommunication, University of Florence, and then with the Department of Information Engineering and Mathematics, University of Siena, where he works as an Associate Professor. During the last decade, he has been studying the application of image processing techniques to copyright protection and authentication of multimedia, and the possibility of processing signals that has been previously encrypted without decrypting them (digital watermarking, multimedia forensics, and signal processing in the encrypted domain). Lately, he has been involved in theoretical and practical aspects of adversarial signal processing. He has authored/coauthored about 300 papers in international journals and conference proceedings, and holds four patents in the field of digital watermarking and image authentication. He has coauthored the book *Watermarking Systems Engineering: Enabling Digital Assets Security and other Applications* (Dekker Inc., 2004). He participated in several National and European research projects on diverse topics, including computer vision, multimedia signal processing, remote sensing, digital watermarking, and IPR protection. He was the Funding Editor of the *EURASIP Journal on Information Security*. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He was the Chairman of the IEEE Information Forensics and Security Technical Committee from 2010 to 2011, and the Technical Program Co-Chair of ICASSP 2014. He was appointed as an DL of the IEEE SPS from 2013 to 2014. He is a member of EURASIP. He was a recipient of the 2016 Individual Technical Achievement of EURASIP.

Benedetta Tondi (S'13) received the master's (*cum laude*) degree in electronics and communications engineering from the University of Siena, Siena, Italy, in 2012, with a thesis on the adversary-aware source identification in the area of multimedia forensics. She is a Research Associate at the Department of Information Engineering and Mathematics, University of Siena. She is currently a member of the Visual Information Processing and Protection Group in the DIISM. She is Assistant for the course of Information Theory and Coding and Multimedia Security, led by M. Barni. She is a member of the National Inter-University Consortium for Telecommunications. She is a Student Member of the IEEE Young Professionals and the IEEE Signal Processing Society. Her research interests focus on the application of information theory and game theory concepts to forensics and counter-forensics analysis and more in general on the adversarial signal processing. From 2014 to 2015, she was a Visiting Student at the Signal Processing in Communications Group, University of Vigo. She has been a Designated Reviewer on the Technical Program Committee for the IEEE GlobalSIP14- Workshop on Information Forensics and Security (WIFS) 2014, the IEEE International Conference on Multimedia and Expo 2015 and 2016, and the IEEE International Conference on Image Processing 2015 and 2016. She is a winner of the Best Student Paper Award at the IEEE WIFS 2014, and the Best Paper Award at the IEEE WIFS 2015.