The Security Margin: a Measure of Source Distinguishability under Adversarial Conditions

Mauro Barni
Dept. Information Engineering and Mathematical Sciences
University of Siena, ITALY
barni@dii.unisi.it

Benedetta Tondi
Dept. Information Engineering and Mathematical Sciences
University of Siena, ITALY
benedettatondi@gmail.com

***Abstract*—We analyze the distinguishability of two sources under adversarial conditions, when the error exponents of type I and type II error probabilities are allowed to take an arbitrarily small, yet positive, values. 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 reliably distinguished. We compute the security margin for some classes of sources and derive a general upper bound which is valid for any kind of sources assuming that the distortion is measured in terms of the mean square error between the original and the attacked sequences.**

I. Introduction

Driven by the necessity of understanding the fundamental limits of source identification in the presence of an adversary, [1] and [2] have introduced the source identification game as a game played by a Defender (D) and an Attacker (A) defined as follows: given two discrete memoryless sources $X \sim P_X$ and $Y \sim P_Y$, with the same alphabet $\mathcal{X}$, and a test sequence $x^n = (x_1, x_2 \dots x_n)$, the goal of D is to decide whether x^n has been drawn from X (hypothesis H_0) or not (hypothesis H_1). At the same time, 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 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 that can be introduced into y^n. The payoff of the game is the false negative error probability (i.e. the probability of deciding for H_0 when H_1 holds), in that D and A aim, respectively, at minimizing and maximizing it.

In [2], a version of the game in which the defender is confined to base its analysis only on first order statistics of x^n is studied, and the asymptotic equilibrium point of the game is determined when the length of the test sequence tends to infinity and the false positive error probability is required to tend to zero exponentially fast with decay rate at least equal to some λ. Given two probability mass functions (pmf) P_X and P_Y, a false positive error exponent λ, and the maximum allowed distortion D_{max}, the analysis in [2] permits to determine whether, at the equilibrium, the false negative error probability P_{fn} tends to 0 or to 1 when $n \to \infty$. This, in turn, permits to define the so-called indistinguishability region $\Gamma(P_X, \lambda, D_{max})$ as the set of the pmf's that can not be distinguished reliably from P_X when $n \to \infty$ due to the presence of the attacker. If $P_Y \in \Gamma(P_X, \lambda, D_{max})$, in fact, a strictly positive false negative error exponent can not be achieved and the attacker is going to *win* the game.

An undesirable feature of the analysis carried out in [2] is the asymmetric role of the false positive and the false negative error exponents, i.e. λ and, say, ε ($\varepsilon = \lim_{n\to\infty} -\frac{1}{n}\log P_{fn}$). In [2], in fact, the defender is required to ensure a given λ, while he is satisfied with any strictly positive ε. In this paper, we relax this constraint and allow the defender to diminish λ so to ease attaining a positive ε. More precisely, in a way that resembles Stein's lemma [3], we analyze the behavior of $\Gamma(P_X, \lambda, D_{max})$ when $\lambda \to 0$ to see whether, given a maximum allowable distortion D_{max}, it is possible for D to simultaneously attain strictly positive error exponents for the two kinds of error, hence permitting to reliably distinguish between P_X and P_Y despite the presence of the adversary. Having done so, we will adopt a slightly difference perspective and introduce the new concept of Security Margin, defined as the maximum distortion allowed to the attacker for which two sources X and Y can be reliably distinguished. 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.

The rest of this paper is organized as follows. In Sec. II, we formally introduce the source identification game and summarize the main results proven in [2]. With respect to [2], we give a novel, more insightful, interpretation of the set of strategies available to the attacker, based on the concept of transportation map. In Sec. III, we investigate the relationship between the optimal attacker's strategy and optimal transport theory, laying the basis for the subsequent analysis. In Sec. IV, we study the behavior of $\Gamma(P_X, \lambda, D_{max})$ when $\lambda \to 0$ and give a rigorous definition of Security Margin. In Sec. V, we demonstrate the powerfulness of our analysis by deriving the Security Margin for some common pmf's. The paper ends with some conclusions and directions for future work in Sec. VI.

II. The source identification game (SI^{lr}_{ks})

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 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. In a zero-sum competitive game we have $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$, so that the win of a player is equal to the loss of the other. In our set up, the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players. In addition, we assume that the players choose their strategies before starting the game so that they have no hints about the strategy actually chosen by the other player (strategic game). Given the above definitions, the source identification game with known sources and limited resources, hereafter referred to as SI^{lr}_{ks}, is defined as follows.

Defender's strategies. To start with, we consider the strategies of the Defender ($\mathcal{S}_D$). As outlined in the Introduction, we simplify the problem by requiring that D bases its analysis only on first order statistics of x^n, that is we require that the acceptance region for hypothesis H_0 (referred to as Λ_0^n) is a union of type classes[1]. Since

[1]A type class is defined as the set of all the sequences having the same empirical distribution [3], [4].

a type class is univocally defined by the empirical pmf (type) of the sequences contained in it, this is equivalent to define Λ_0^n as a union of types $P \in \mathcal{P}_n$, where $\mathcal{P}_n$ is the set of all the possible types for sequences of length n. In addition, we consider the asymptotic version of the game and require that P_{fp} decreases exponentially fast with decay rate at least equal to λ. Under the above assumptions, the set of strategies of D is defined as follows:

$$\mathcal{S}_D = \{\Lambda_0^n \in 2^{\mathcal{P}_n} : P_{fp} \leq 2^{-\lambda n}\}, \tag{1}$$

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

Attacker's strategies. Given a sequence y^n drawn from Y, the goal of the attacker is to transform it into a sequence z^n belonging to the acceptance region chosen by D. 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 $S_{YZ}(i,j) = n(i,j)/n$ the relative frequency with which the i-th symbol is transformed into the j-th one. In the following, we refer to S_{YZ} as *transportation map*. An interesting property of S_{YZ} is that, for any additive distortion measure, the per-sample distortion $d(y^n, z^n)/n$ between y^n and z^n depends only on S_{YZ}, since we have $d(y^n, z^n) = \sum_{i,j} n(i,j)d(i,j)$, where $d(i,j)$ is the distortion introduced when the symbol i is transformed into the symbol j. Equivalently, the average per-sample distortion between y^n and z^n is $\sum_{i,j} S_{YZ}(i,j)d(i,j)$. Even more interestingly, the empirical pmf (i.e. the type) of the attacked sequence is univocally determined by S_{YZ}. In fact, by indicating with $P_{z^n}(j)$ the relative frequency of symbol j into z^n, we have:

$$P_{z^n}(j) = \sum_i S_{YZ}(i,j) \triangleq S_Z(j). \tag{2}$$

Finally, we observe that the attacker can not change more symbols than there are in the sequence y^n, as a result a transportation map S_{YZ} can be applied to a sequence y^n only if:

$$S_Y(i) \triangleq \sum_j S_{YZ}(i,j) = P_{y^n}(i). \tag{3}$$

Equation (3), together with (2) suggests an interesting interpretation of S_{YZ}, S_Y and S_Z, which can be seen, respectively, as the joint empirical pmf between the sequences y^n and z^n, the empirical pmf of y^n and the empirical pmf of z^n.

Since due to the limited resources assumption, Λ_0^n depends only on the empirical pmf of the test sequence, and given that the empirical pmf of the attacked sequence depends only on S_{YZ} through S_Z, we can restrict the action of the attacker to the choice of a transportation map among all the *admissible* maps, a map being admissible if:

$$\begin{aligned} &S_Y = P_{y^n} \\ &\sum_{i,j} S_{YZ}(i,j)d(i,j) \leq D_{max}. \end{aligned} \tag{4}$$

In the following, we will refer to the set of admissible maps as $\mathcal{A}(D_{max}, P_{y^n})$. Given the above, the set of strategies available to 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 $S_{YZ}(i,j;y^n)$.

The payoff. The payoff of the game is the false negative error probability, that is:

$$u_D = -u_A = -\sum_{y^n : S_Z(j;y^n) \in \Lambda_0^n} P_Y(y^n), \tag{5}$$

where $P_Y(y^n)$ is the probability that the source Y outputs the sequence y^n.

Equilibrium point. The main result of [2] is summarized by the following theorem.

Theorem 1. *The profile* $(\Lambda_0^{*,n}, S_{YZ}^*(i,j;y^n))$ *with*

$$\Lambda_0^{*,n} = \left\{ P \in \mathcal{P}_n : \mathcal{D}(P||P_X) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n} \right\}, \tag{6}$$

and

$$S_{YZ}^*(i,j;y^n) = \arg \min_{S_{YZ} \in \mathcal{A}(D_{max}, P_{y^n})} \mathcal{D}(S_Z||P_X), \tag{7}$$

is the only rationalizable equilibrium of the SI_{ks}^{lr} *game, which, then, is a dominance solvable game [5].*

In the above theorem, $\mathcal{D}(P||Q)$ indicates the divergence (or Kullback-Leibler distance) between two pmf's P and Q [3]. Given the optimal acceptance region Λ_0^*, we can introduce the indistinguishability region $\Gamma^n(P_X, \lambda, D_{max})$ as follows:

$$\begin{aligned} &\Gamma^n(P_X, \lambda, D_{max}) = \\ &\qquad \{P \in \mathcal{P}_n : \exists\ S_{YZ} \in \mathcal{A}(D_{max}, P) \text{ s.t. } S_Z \in \Lambda_0^{*,n}\}. \end{aligned} \tag{8}$$

Note that the above analysis applies only to sequences of length n (as explicitly indicated by the apex n in $\Lambda_0^{*,n}$ and Γ^n), so the values assumed by the empirical pmf's and the transportation map belong to $\mathbb{Q}_n$, i.e. the set of rational number with denominator n.

Due to the density of rational numbers in $\mathbb{R}$, by letting n tend to infinity we obtain the asymptotic counterpart of Γ^n, namely Γ, specifying whether two sources can eventually be distinguished for increasing values of n. Region Γ can be expressed as follows

$$\begin{aligned} &\Gamma(P_X, \lambda, D_{max}) = \\ &\qquad \{P \in \mathcal{P} : \exists\ S_{YZ} \in \mathcal{A}(D_{max}, P) \text{ s.t. } S_Z \in \Lambda_0^*(P_X, \lambda)\}, \end{aligned} \tag{9}$$

where Λ_0^* is the asymptotic version of the set $\Lambda_0^{*,n}$:

$$\Lambda_0^*(P_X, \lambda) = \{P \in \mathcal{P} : \mathcal{D}(P||P_X) \leq \lambda\}. \tag{10}$$

Note that now the values $P(i)$ and $S_{YZ}(i,j)$ are no longer required to belong to $\mathbb{Q}_n$. More precisely, by using a slightly different formulation with respect to [2], we can state the following theorem

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

$$\varepsilon = \min_{P \in \Gamma(P_X, \lambda, D_{max})} \mathcal{D}(P||P_Y), \tag{11}$$

leading to the following cases:

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

III. Relationship with optimal transport theory

Before going on with our analysis, we find it convenient to rephrase the results described in Section II as an optimal transport problem [6]. Let P and Q be two pmf's, and let $c(i,j)$ be a measure specifying the cost of transporting the i-th symbol into the j-th one. The optimal transport problem looks for the transportation map S_{PQ}^* 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 minimization problem:

$$S_{PQ}^* = \arg \min_{S_{YZ} : S_Y = P, S_Z = Q} \sum_{i,j} S_{YZ}(i,j)c(i,j). \tag{12}$$

By interpreting the pmf's P and Q as two different ways of piling up a certain amount of earth, the minimum cost achieved in (12) can be seen as the minimum effort required to turn one pile into the other,

where $c(i,j)$ is the cost necessary to move a unitary amount of earth from position i to position j. Due to such a viewpoint, the minimum in (12) is usually called the Earth Mover Distance (*EMD*) between P and Q (see [7]).

Optimal transport theory permits us to rewrite the indistinguishability region in a more compact and easier-to-interpret way. In fact, it is immediate to see that (9) can be rewritten as:

$$\Gamma(P_X,\lambda,D_{max}) = \{P\in\mathcal{P}:\exists\ Q\in\Lambda_0^*(P_X,\lambda)\ \text{s.t}\ \textit{EMD}(P,Q)\le D_{max}\}, \tag{13}$$

where in the definition of the *EMD* we let $c(i,j)=d(i,j)$.

We point out that when the L_2 distance is used to measure the distance between symbols, that is when $d(i,j)=(i-j)^2$, the *EMD* between two probability distributions P_Y and P_X corresponds to the squared Mallows distance [8]. Given two sources $X\sim P_X$ and $Y\sim P_Y$, the squared Mallows distance between P_X and P_Y is defined as the minimum mean square error between X and Y taken over all joint probability distributions P_{XY} such that the marginal distribution are respectively P_X and P_Y:

$$M_2^2(P_X,P_Y)=\min_{\substack{P_{XY}:\sum_x P_{XY}=P_Y\\ \sum_y P_{XY}=P_X}} E_{XY}[(X-Y)^2]. \tag{14}$$

Note that even if we introduced the *EMD* by considering finite-alphabet sources, there is no need to restrict the definition of the Mallows distance to discrete random variables; in fact at the end of this paper we will extend our analysis and use the *EMD* or Mallows distance to measure the distinguishability of continuous sources.

IV. The security margin

We now study the behavior of $\Gamma(P_X,\lambda,D_{max})$ 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}^{lr} game. The rationale behind our analysis derives directly from (9) and (10). 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 u_D. 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 under adversarial conditions. In a more rigorous way, we can state the following theorem (closely reminding Stein's lemma for attack-free hypothesis testing [3]):

Theorem 3. *Given two sources $X\sim P_X$ and $Y\sim P_Y$ and a maximum allowable average per-letter distortion D_{max}, the maximum achievable false negative error exponent ε for the SI_{ks}^{lr} game is*

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

where

$$\Gamma(P_X,D_{max})=\{P\in\mathcal{P}:\textit{EMD}(P,P_X)\le D_{max}\}. \tag{16}$$

Proof: Due to space limitations, we only provides a sketch of the proof. When $\lambda\to 0$, we see from (10) that the set $\Lambda_0^*(P_X,\lambda)$ collapses into the single point P_X. From (13), it is easy to argue that for $\lambda\to 0$, $\Gamma(P_X,\lambda,D_{max})$ takes the form expressed in (16). Reasoning like in Stein's Lemma we can prove that $\Gamma(P_X,D_{max})$ is exactly the smallest indistinguishability region that can be achieved by D and that the expression of the false negative error exponent in (15) holds. Figure 1 gives a geometrical interpretation of Theorems 2 and 3 and the indistinguishability regions. Point P^* represents the minimum-achieving pmf in the two cases, while the shaded area

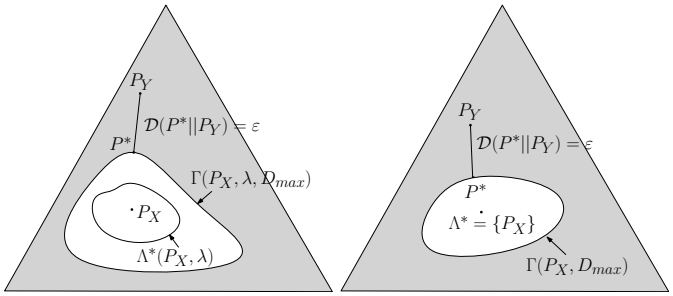

Fig. 1. Geometric interpretation of Theorems 2 and 3.

represents the set of points (pmf's) P for which $\textit{EMD}(P,Q)>D_{max}$ for any $Q\in\Lambda^*$, and hence can be distinguished from P_X. ∎

Theorem 3 permits to interpret $\Gamma(P_X,D_{max})$ as the smallest indistinguishability region for the SI_{ks}^{lr} game. In particular, definition (16) suggests that the distinguishability of two pmf's under adversarial conditions ultimately depends on their *EMD*. In fact, if $\textit{EMD}(P_Y,P_X)>D_{max}$, the defender will be able to distinguish X from Y by adopting a sufficiently small λ. On the contrary, if $\textit{EMD}(P_Y,P_X)\le D_{max}$, then there is no positive value of λ for which the two sources can be asymptotically distinguished.

By adopting a slightly different perspective, given two sources X and Y, one may wonder which is the maximum attacking distortion for which D can distinguish X and Y despite the presence of the adversary, naturally leading 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 distortion for which the two sources can be reliably distinguished in the SI_{ks}^{lr} setting is called Security Margin and is given by*

$$\mathcal{SM}(P_Y,P_X)=\textit{EMD}(P_Y,P_X). \tag{17}$$

We observe that the *EMD* is a symmetric function of P_X and P_Y [7], and hence the security margin does not depend on the role of X and Y in the test, i.e. $\mathcal{SM}(P_X,P_Y)=\mathcal{SM}(P_Y,P_X)$.

V. Notable examples

A. Discrete sources

In general the *EMD* between two sources can be computed by resorting to numerical analysis, and in fact, due to its wide use as similarity measure in computer vision applications, several efficient algorithms have been developed to compute the *EMD* between discrete sources (see [9] for example). In some cases, however, a closed form expression can be found, as shown in the following.

Bernoulli sources. 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 $p>q$. Clearly, we have:

$$\sum_{i,j} S_{YX}(i,j)d(i,j)=S_{YX}(0,1)+S_{YX}(1,0). \tag{18}$$

Since $p>q$, it is easy to conclude that the minimum of the above expression is obtained when $S_{YX}(1,0)=0$ (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 $\mathcal{SM}(P_X,P_Y)=|p-q|$.

Uniform sources with multiple cardinalities. Let X and Y be two uniform pmf's with alphabets $\mathcal{X}$ and $\mathcal{Y}$. It is possible to prove that the

optimum transportation map is obtained by taking the bin of $\mathcal{Y}$ with the smallest value and start moving it into the bin with the smallest value in $\mathcal{X}$. When the smallest bin of $\mathcal{X}$ is full, we go on with the second smallest bin in $\mathcal{X}$. When the smallest bin in $\mathcal{Y}$ has been emptied, we go on with the second smallest bin in $\mathcal{Y}$. The procedure is iterated until all the bins in $\mathcal{Y}$ have been moved into those of $\mathcal{X}$. When the cardinality of $\mathcal{Y}$ is a multiple of the cardinality of $\mathcal{X}$ such a procedure permits to express the $\mathcal{SM}$ in closed form as follows:

$$\mathcal{SM}(P_X, P_Y) = \frac{1}{|\mathcal{Y}|} \sum_{i=0}^{|\mathcal{X}|-1} \sum_{j=0}^{\alpha-1} (|X_l - Y_l| + j + (\alpha-1)i)^2, \quad (19)$$

where we have assumed that $|\mathcal{Y}| = \alpha|\mathcal{X}|$, with $\alpha \in \mathbb{N}$, and where X_l and Y_l denote the lower non-empty bins of $\mathcal{X}$ and $\mathcal{Y}$ respectively. The formula implicitly assumes that $Y_l > X_l$, the extension to the case in which such a relationship does not hold being immediate.

B. 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 (pdf's) of the sources and apply the analysis for discrete sources. By letting the quantization step tend to zero, the *EMD* (which in this case can be interpreted as the Mallows distance) between P_X and P_Y can still be regarded as the security margin between the two sources. In the following we assume that the squared Euclidean norm (L_2^2) is used as distance metric so to use the expression given in (14). Let then X and Y be two continuous sources with means μ_X and μ_Y and variances σ_X and σ_Y. As shown in [10] (decomposition theorem), the expectation in (14) can be rewritten as follows

$$E_{XY}[(X-Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2[\sigma_X\sigma_Y - covXY], \quad (20)$$

where the three terms express, respectively, the difference in location, spread and shape between the variables X and Y [11]. Interestingly, the covariance term $covXY$ is the only term in (20) which depends on the joint pdf of X and Y. Then, in order to find the security margin, we only have to compute:

$$\max_{P_{XY}: X\sim P_X, Y\sim P_Y} covXY. \quad (21)$$

By assuming $P_{X,Y} = P_X P_Y$, i.e. by assuming that X and Y are independent, we have $covXY = 0$, hence permitting us to derive a general upper bound of $\mathcal{SM}$:

$$\mathcal{SM}_{L_2^2}(P_X, P_Y) \leq (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2\sigma_X\sigma_Y. \quad (22)$$

When the pdf's of the two sources have the same shape, for instance when they are both distributed according to a Gaussian or a Laplacian distribution, the security margin assumes a particularly simple expression. 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 case the covariance term is maximum and equal to $\sigma_X\sigma_Y$, and hence the contribution of the shape term in the security margin vanishes, yielding:

$$\mathcal{SM}_{L_2^2}(P_X, P_Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2. \quad (23)$$

This is a remarkable, and in principle unexpected, result, stating that the distinguishability of two sources belonging to the same class depends only on their mean values and variances, regardless of their particular pdf.

VI. Conclusions

In the attempt to analyze the distinguishability of two sources in an adversarial setting, we studied the behavior of the SI_{ks}^{lr} game when the defender is allowed to arbitrarily reduce the false positive error exponent. This allowed us to study the ultimate distinguishability of two sources X and Y. It turns out that when an adversary is present, the source distinguishability can be summarized into a single parameter called Security Margin. If the attacker introduces a distortion lower than the security margin, in fact, the defender will always be able to distinguish the two sources assuming that the length of the sequence he observes tends to infinity. By exploiting the parallelism between the optimum attacker's strategy and optimal transport theory, we have shown that the security margin corresponds to the Earth Mover Distance between the two sources and computed it for some common class of sources. In practice, just to mention a possible real application, the concept of security margin allows to give a measure of distinguishability between untouched images and processed ones in the image forensic scenario under adversarial conditions considered in [12]. As a future work, we plan to compute the security margin for a wider class of sources, and extend the concept of security margin to different versions of the source identification game, like, for instance, the game with training sequences [13].

References

[1] M. Barni, "A game theoretic approach to source identification with known statistics," in *ICASSP 2012, IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

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

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

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

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

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

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

[8] E. Levina and P. Bickel, "The Earth Mover's distance is the Mallows distance: some insights from statistics," in *Proc. of ICCV 2001, Eight IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 251–256 vol.2.

[9] O. Pele and M. Werman, "Fast and robust Earth Mover's distances," in *Proc. ICCV'09, 12th IEEE International Conference on Computer Vision*, 2009, pp. 460–467.

[10] A. Irpino and E. Romano, "Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation." in *EGC*, ser. Revue des Nouvelles Technologies de l'Information, M. Noirhomme-Fraiture and G. Venturini, Eds., vol. RNTI-E-9. Cepadues-Editions, 2007, pp. 99–110.

[11] K. Košmelj and L. Billard, "Mallows' L_2 distance in some multivariate methods and its application to histogram-type data," *Metodoloski Zvezki*, vol. 9, no. 2, pp. 107–118, 2012.

[12] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. of MM&Sec 2012, 14th ACM Workshop on Multimedia & Security*. New York, NY, USA: ACM, 2012, pp. 97–104.

[13] M. Barni and B. Tondi, "Optimum forensic and counter-forensic strategies for source identification with training data," in *Proc. of WIFS'12, IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, 2-5 December 2012, pp. 199–204.

Acknowledgement

This work was partially supported by the REWIND Project funded by the Future and Emerging Technologies (FET) program of the European Commission, under FET-Open grant number: 268478.