Source Distinguishability under Corrupted Training

Mauro Barni [1,2], Benedetta Tondi [1]

[1]*Department of Information Engineering and Mathematics, University of Siena, Siena, ITALY*
[2]*Consorzio Nazionale Interuniversitario per le Telecomunicazioni, CNIT*

barni@dii.unisi.it, benedettatondi@gmail.com

***Abstract*—We study a new variant of the source identification game with training data in which part of the training data is corrupted by an adversary. In such a scenario, the defender wants to decide whether a test sequence x^n has been drawn from the same source which generated a training sequence t^N, part of which has been corrupted by the adversary. By adopting a game theoretical formulation, we derive the unique rationalizable equilibrium of the game in the asymptotic setup. Moreover, by mimicking Stein's lemma, we derive the best achievable performance for the defender, permitting us to analyze the ultimate distinguishability of the two sources. We conclude the paper by comparing the performance of the test with corrupted training to the simpler case in which the adversary can not modify the training sequence, and by deriving the percentage of samples that the adversary needs to modify to make source identification impossible.**

I. Introduction

Adversarial Signal Processing (AdvSP) is an emerging discipline aiming at modeling the interplay between a defender wishing to carry out a certain processing task, and an attacker aiming at impeding it. Binary decision in an adversarial setup is one of the most recurrent problems in AdvSP, due to its importance in many application scenarios [1]. Among binary decision problems, source identification is one of the most studied subjects, since it lies at the heart of several security-oriented disciplines, like Multimedia forensics, anomaly detection, steganalysis and so on.

In [2] the source identification game is introduced to model the interplay between the defender and the attacker by resorting to concepts drawn from game and information theory. According to the model put forward in [2], the defender and the attacker have a perfect knowledge of the sources. In [3] the analysis is pushed a step forward, considering a scenario in which the to-be-distinguished sources are known only through the observation of a training sequence. Finally, [4] introduces the security margin concept, a powerful parameter characterizing the ultimate distinguishability of two sources under adversarial conditions.

In this paper, we move the analysis even further, by considering a situation in which the attacker may interfere with the learning phase by corrupting part of the training sequence. As a matter of fact, adversarial learning is a rather novel concept, which has been studied for some years from a machine learning perspective [5], [6]. Due to the natural vulnerability of machine learning systems, in fact, the attacker may take an important advantage if no countermeasures are adopted by the defender. The use of a training sequence to gather information about the statistics of the to-be-distinguished sources can be seen as a very simple learning mechanism, and the analysis of the impact that an attack, carried out in such a phase, has on the performance of a decision system, may help shedding new light on this important problem. To be specific, we extend the game-theoretic framework introduced in [3] and [4] to model a situation in which the attacker is given the possibility of corrupting part of the training sequence. We then derive the optimal strategy for the defender and the optimal corruption strategy for the attacker. Given such optimum strategies, expressed in the form of game equilibrium point, we analyze the best achievable performance in an asymptotic set up, that is when the length of the training and test sequences tend to infinity and the error probabilities of the decision tend to zero exponentially fast. Specifically, we study the distinguishability of the sources in function of the percentage α of training samples corrupted by the attacker and when the test sequence can be modified up to a certain distortion level. The results of the analysis are summarized in terms of blinding percentage α_b, defined as the percentage of corrupted samples making a reliable distinction between the two sources impossible, and security margin, defined as the maximum distortion level for which a reliable distinction is possible (see [4] and [7]).

The rest of the paper is organized as follows. In Section II we describe the scenario analyzed in the paper and give a rigorous definition of the Source Identification game with corrupted training. In Section III, we derive the equilibrium point of the game and compute the payoff at the equilibrium. In Section IV, we study the best achievable performance of the game when the defender requires only that the error probabilities of the two kinds tend to zero exponentially fast, regardless of the error exponent. We do so by introducing two summarizing parameters, namely the *security margin* under corrupted samples, and the *blind corruption percentage*. Finally, in Section V we draw some conclusions and highlight directions for future work. For lack of space, throughout the paper, we focus on the main flow of ideas without providing a complete proof of the theorems.

II. Source identification game with corrupted training

In the following we give a rigorous game-theoretic formulation of the scenario addressed in this paper. Given two discrete and memoryless sources $X \sim P_X$ and $Y \sim P_Y$ and a test sequence x^n, the goal of the defender (D) is to decide whether x^n has been drawn from X or not. On the other side, the goal of the attacker (A) is to take a sequence y^n drawn from Y and modify it in such a way that D decides that the modified sequence z^n has been generated by X. As in previous works, A must respect a distortion constraint. In the scenario considered in this paper, A and D know the statistics of X through a training sequence, however the training sequence available to D has been partly corrupted by A. More specifically, the attacker has access to a sequence τ^{m_1} drawn from X. Then he/she corrupts τ^{m_1} by adding a sequence of fake samples τ^{m_2}, then he/she reorders the sequence in a random way so to hide the position of the fake samples. Note that reordering does not alter the statistics of the training sequence since the sequence is supposed to be generated from a memoryless source. We assume that D knows that a certain percentage of samples in the training sequence may be corrupted, but he has no clue about the position of corrupted samples. According to the classification given in [6], the above scenario can be referred to as a *causative* attack with control over training data. In the following, we will denote by N the final length of the training sequence ($N = m_1 + m_2$), by α the portion of fake samples ($m_2 = \alpha N$) and by $1 - \alpha$ the portion of original samples

978-1-4799-8882-2/14/$31.00 ©2014 IEEE

$(m_1 = (1-\alpha)N)$. The training sequence made available to D will be indicated by t^N. Finally, we hypothesize a linear relationship between the length of the test and the corrupted training sequence, i.e. $N = cn$. By adopting a Neyman-Pearson perspective, D is interested in accepting or rejecting the hypothesis H_0 that the sequence has been generated by X, ensuring that the false positive error probability (P_{fp}) of rejecting H_0 when H_0 holds (type I error) is lower than a given threshold. On the other side, the attacker aims at inducing a type II error, i.e. at making the system accept a sequence generated by Y (alternative hypothesis H_1) as if it were drawn from X. By referring again to the taxonomy introduced in [6], this is a typical *integrity* attack to the source identification system. Similarly to the previous versions of the game studied in [2] and [3], we assume that D relies only on the first order statistics of x^n and t^N to make a decision. For mathematical tractability, likewise the earlier versions (e.g. [2]), we study the asymptotic version of the game when $n \to \infty$, by requiring that P_{fp} decays exponentially fast with error exponent at least equal to λ.

We start by observing that source identification with training data is equivalent to deciding whether the training and test sequences have been generated by the same, unknown, source. The additional difficulty we have to face with in the case of a corrupted training is that the defender must take into account the fact that only part of the training sequence has been generated by the correct source. Let, then, $\mathcal{I}$ denote a subset of $(1-\alpha)N$ indexes taken in $\{1, 2, ..., N\}$, and let $\bar{\mathcal{I}}$ be the indexes in $\{1, 2, ..., N\}$ which are not contained in $\mathcal{I}$. We indicate by $t_{\mathcal{I}}^{(1-\alpha)N}$ the subsequence of t^N obtained by removing the elements indexed by $\bar{\mathcal{I}}$ (or, equivalently, keeping only those indexed by $\mathcal{I}$). Since D does not know P_X and he/she can not make any assumptions on the position of corrupted samples in t^N, in order to guarantee that the false positive error probability is lower than $2^{-n\lambda}$, he/she needs to impose that:

$$\max_{\mathcal{I}} \max_{P_X} P_X\{(x^n, t_{\mathcal{I}}^{(1-\alpha)N}) \notin \Lambda^n\} \leq 2^{-n\lambda}, \qquad (1)$$

where Λ^n is the acceptance region of the test, i.e. the set with the pairs of sequences for which D accepts the hypothesis (H_0) that x^n and t^N have been generated by the same source. The assumption behind (1) is that, since D does not have any information about the way the fake samples have been generated by A, the only reasonable choice for him is to ignore such samples.

With regard to A, the attack now consists of two parts. Given a sequence y^n drawn according to P_Y, and the original training sequence $\tau^{(1-\alpha)N}$, the attacker generates the sequence of fake samples $\tau^{\alpha N}$, which are randomly mixed up with those in $\tau^{(1-\alpha)N}$, and transforms y^n into z^n, trying to generate a pair (t^N, z^n)[1] belonging to Λ^n. In doing so, he has to ensure that $d(y^n, z^n) \leq nL$ for some proper distortion function d.

With the above ideas in mind, we define the source identification game with corrupted training as follows.

Definition 1. *The $SI_{c\text{-}tr}(\mathcal{S}_D, \mathcal{S}_A, u)$ is a zero-sum, strategic, game played by D and A, defined by the following strategies and payoff.*

- *The strategies available to D are all the acceptance regions for which the false positive error probability is guaranteed to tend to zero exponentially fast:*

$$\mathcal{S}_D = \left\{ \Lambda^n : \max_{\mathcal{I}, P_X} P_X\{(x^n, t_{\mathcal{I}}^{(1-\alpha)N}) \notin \Lambda^n\} \leq 2^{-n\lambda} \right\}, \qquad (2)$$

where we exploited the independence, under H_0, of x^n (test) and t^N (corrupted training).

- *The set of strategies of A consists of all the possible ways he can choose two functions, g and f. Function $g(\cdot)$ rules the generation of the fake training samples, while $f(\cdot)$ is a mapping function, which maps a sequence y^n generated by Y into a new sequence z^n subject to a distortion constraint. In formulas:*

$$\mathcal{S}_A = \Big\{ g : \tau^{\alpha N} = g(\tau^{(1-\alpha)N}, y^n) \qquad (3)$$
$$f : d(y^n, f(y^n, \tau^{(1-\alpha)N})) \leq nL \Big\},$$

where L denotes the maximum allowed average per-letter distortion.

- *The payoff function is defined in terms of the false negative error probability, namely:*

$$u(\Lambda^n, (g, f)) = -P_{fn}, \qquad (4)$$

where P_{fn} is the false negative probability (that is the probability of accepting H_0 when H_1 holds).

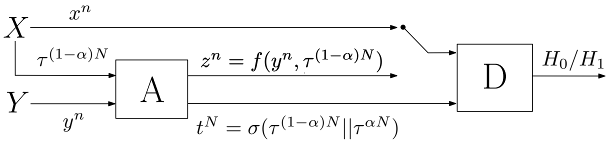
X
x^n
$\tau^{(1-\alpha)N}$
A
$z^n = f(y^n, \tau^{(1-\alpha)N})$
D
H_0/H_1
Y
y^n
$t^N = \sigma(\tau^{(1-\alpha)N}||\tau^{\alpha N})$

Fig. 1. Block diagram of the adversarial setup considered in the paper. Symbol $||$ denotes the concatenation between sequences and $\sigma()$ a possible reordering of the samples.

A critical observation to be made regards the dependence of $g(\cdot)$ on y^n. This means that the fake samples used to corrupt the training sequence may depend on the sequence that the attacker wants to pass off as generated by X. This might seem too strong an assumption, since it is not reasonable that the attacker generates a new corrupted training set for each new sequence generated by Y. As we will see later on, however, due to the asymptotic nature of our analysis, the dependence on y^n will be transformed into a dependence on P_Y, which is a much more reasonable assumption.

III. Payoff at the equilibrium

Having defined the $SI_{c\text{-}tr}$ game, we must study the existence of an equilibrium point. The analysis goes along the same lines followed in [3] for the source identification game with non-corrupted training. For this reason, and for lack of space, we limit ourselves to stating the main results without proving them.

A. Equilibrium point

To start with, we observe that since the defender bases its analysis on the first order statistics of t^N and x^n, the acceptance region Λ^n can be expressed as a union of pairs of types or type classes[2]. Then we need to define the generalized log-likelihood ratio function $h(P_{x^n}, P_{t^N})$ (see [10], [11], [3]):

$$h(P_{x^n}, P_{t^N}) = \mathcal{D}(P_{x^n}||P_{r^{n+N}}) + \mathcal{D}(P_{t^N}||P_{r^{n+N}}), \qquad (5)$$

where $\mathcal{D}$ indicates the Kullback-Leibler distance (or divergence function) [8] and $P_{r^{n+N}}$ denotes the empirical probability mass

[1] While reordering is essential to hide the position of fake samples to D, it does not have any impact on the position of (t^N, z^n) with respect to Λ^n, since we assumed that the defender bases its decision only on the first order statistic of the observed sequences.

[2] The type (P_{x^n}) of a sequence x^n is the empirical probability distribution induced by the sequence. A type class is defined as the set of all the sequences having the same type [8], [9]. Throughout the paper we indicate by $\mathcal{P}_n$ the set of types induced by sequences of length n.

function (pmf) of the sequence r^{n+N}, obtained by concatenating x^n and t^N, i.e. $r^{n+N} = x^n \| t^N$. The optimum strategy for the defender stems from the following lemma.

Lemma 1. $\Lambda^{n,*}$ *being defined as follows:*

$$\Lambda^{n,*}=\Big\{(P_{x^n}, P_{t^N})\colon \min_{\mathcal{I}} h(P_{x^n}, P_{t_{\mathcal{I}}^{(1-\alpha)N}}) \leq \lambda - \delta_{n,c}\Big\} \quad (6)$$

with

$$\delta_{n,c} = |\mathcal{X}| \frac{\log(n+1)(nc+1)}{n}, \quad (7)$$

where $|\mathcal{X}|$ *is the cardinality of the source alphabet and where* $P_{\tau_{\mathcal{I}}^{(1-\alpha)N}}$ *denotes the type of the subsequence of* τ^N *obtained by removing the samples indexed by* $\bar{\mathcal{I}}$*. Then:*

1) $\max_{\mathcal{I}} \max_{P_X} P_X\{(x^n, t_{\mathcal{I}}^{(1-\alpha)N}) \notin \Lambda^{n,*}\} \leq 2^{-n(\lambda-\nu_n)}$, *with* $\nu_n \to 0$, *for* $n \to \infty$,
2) $\forall \Lambda^n \in \mathcal{S}_D$, *we have* $\bar{\Lambda^n} \subseteq \bar{\Lambda}^{n,*}$.

Proof: The lemma follows immediately from Lemma 2 in [3]. ■

The above lemma shows that strategy $\Lambda^{n,*}$ is admissible (point 1.) and optimal (point 2.) for D, regardless of the attack. From a game-theoretic perspective, this means that such a strategy is a dominant strategy for D and implies that the game is dominance solvable [12]. In such a situation, the defender and the attacker will end up playing, respectively, the dominant strategy and the strategy which results from the resolution of the decision problem (i.e. the problem obtained by assuming that D plays the dominant strategy). Then, given the original training sequence $\tau^{(1-\alpha)N}$, the optimum attacking strategy is given by the following double minimization:

$$\begin{aligned}(g^*(\tau^{(1-\alpha)N}, y^n), f^*(y^n, \tau^{(1-\alpha)N})) = \quad (8)\\ \arg\min_{\substack{(\tau^{\alpha N}, z^n):\\ d(z^n, y^n) \leq nL}} \min_{\mathcal{I}} h(P_{y^n}, P_{t_{\mathcal{I}}^{(1-\alpha)N}}),\end{aligned}$$

Given the optimum strategies for both players, it is immediate to state the following:

Theorem 1. *The* $SI_{c\text{-}tr}$ *game is a dominance solvable game, whose only rationalizable equilibrium corresponds to profile* $(\Lambda^{n,*}, (g^*, f^*))$.

Proof: the theorem is a direct consequence of the fact that $\Lambda^{n,*}$ is a dominant strategy for D. ■

It is worth observing that the concept of rationalizable equilibrium is much stronger than the usual notion of Nash equilibrium, since the strategies corresponding to such an equilibrium are the only ones that two rational players may adopt [12].

Before going on with the analysis of the game at the equilibrium, we observe that, from (8), it is possible to reformulate the optimum strategy for A entirely as a function of types, instead of sequences. Indeed, by adopting an optimal transport perspective [13] (likewise in [4]), we can rewrite the distortion constraint between y^n and z^n in terms of admissibility of the transportation map which moves the distribution P_{y^n} into the attacked distribution P_{z^n}. Formally, let S_{PQ}^n be a transportation map moving the pmf P into Q (when present, the superscript n indicates that the map is applied to empirical pmf's in $\mathcal{P}_n$)[3]. Given a map and a distortion function $d(i,j)$ measuring the *cost* of moving symbol i into j, the average per-letter distortion associated to the map can be written as $\sum_{i,j} S_{PQ}^n(i,j) d(i,j)$. For a certain source pmf P and a maximum distortion L, we define $\mathcal{A}^n(L, P)$ as the set of admissible maps that can be applied to P and introduce an average per-letter distortion lower than L.

We also observe that type of the corrupted training sequence t^N has the following general form:

$$P_{t^N} = (1-\alpha) P_{\tau^{(1-\alpha)N}} + \alpha C', \text{for some } C' \in \mathcal{P}_{\alpha N}, \quad (9)$$

which implies that, given an attacked test sequence z^n (for the moment we do not care how z^n is obtained), the optimum strategy for corrupting the training set is equivalent to find a C^* s.t.

$$C^* = \arg \min_{C' \in \mathcal{P}_{\alpha N}} \min_{\mathcal{I}} h(P_{z^n}, P_{t_{\mathcal{I}}^{(1-\alpha)N}}), \quad (10)$$

where P_{t^N} is written as in (9). Consequently, the optimum strategy of A in (8) can be rewritten as:

$$\arg\min_{\substack{(P_{t^N}, P_{z^n}):\\ P_{t^N} = (1-\alpha)P_{\tau^{(1-\alpha)N}} + \alpha C', C' \in \mathcal{P}_{\alpha N},\\ S_{yz}^n \in \mathcal{A}(L, P_{y^n})}} \min_{\mathcal{I}} h(P_{z^n}, P_{t_{\mathcal{I}}^{(1-\alpha)N}}). \quad (11)$$

With the attacking strategy rewritten as in (11), it is straightforward to define the set of the pairs $(P_{y^n}, P_{\tau^{(1-\alpha)N}})$ for which, because of A's action, D is forced to accept H_0:

$$\begin{aligned}\Gamma^n(\lambda, \alpha, L) = \{(P_{y^n}, P_{\tau^{(1-\alpha)N}}) : \exists (P_{z^n}, P_{t^N}) \in \Lambda^{n,*} \text{ s.t.} \quad (12)\\ P_{t^N} = (1-\alpha) P_{\tau^{(1-\alpha)N}} + \alpha C',\\ \text{for some } C' \in \mathcal{P}_{\alpha N}, \text{ and } S_{yz}^n \in \mathcal{A}(L, P_{y^n})\},\end{aligned}$$

which, by fixing the type of the original training sequence ($P_{\tau^{(1-\alpha)N}}$) becomes:

$$\begin{aligned}\Gamma^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha, L) = \quad (13)\\ \big\{P_{y^n} \in \mathcal{P}_n :\\ \exists P_{z^n} \in \Lambda^{n,*}((1-\alpha)P_{\tau^{(1-\alpha)N}} + \alpha C'),\\ \text{for some } C' \in \mathcal{P}_{\alpha N}, \text{ and s.t. } S_{yz}^n \in \mathcal{A}(L, P_{y^n})\big\},\end{aligned}$$

where, similarly, we referred to the acceptance region for a fixed training type in $\mathcal{P}_N$. It is interesting to notice that, since in the current setting A has two degrees of freedom (he/she can both modify the test sequence and include fake samples in the training sequence), the attack has a double effect: the sequence y^n is modified in order to bring it inside the acceptance region $\Lambda^{n,*}(P_{t^N})$ and the acceptance region itself $\Lambda^{n,*}(P_{t^N})$ is modified so to make the former action easier.

B. Payoff of the game at the equilibrium

In this section we study the payoff of the game at the equilibrium, thus trying to understand who and under which conditions is going to *win* game. To do so, we first reformulate the set in (13) in a more convenient way. First, we rewrite the region $\Gamma^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha, L)$ as follows:

$$\begin{aligned}\Gamma^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha, L) = \{P_{y^n} \in \mathcal{P}_n : \exists S_{yz}^n \in \mathcal{A}(L, P_{y^n}) \quad (14)\\ \text{s.t. } P_{z^n} \in \Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha)\},\end{aligned}$$

where

$$\begin{aligned}\Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha) = \big\{P \in \mathcal{P}_n : \exists C' \in \mathcal{P}_{\alpha N} \text{ s.t.} \quad (15)\\ P \in \Lambda^{n,*}((1-\alpha)P_{\tau^{(1-\alpha)N}} + \alpha C')\big\}.\end{aligned}$$

is the set containing all the test sequences (or, equivalently, test types) for which it is possible to corrupt the training set in such a way that they fall within the acceptance region. As the notation suggests, this set corresponds to the set in (14) when A cannot modify the sequence

[3] Throughout the paper we will adopt the lighter notation S_{yz}^n when we refer to a map which moves the empirical pmf of a sequence y^n (i.e. P_{y^n}) into the empirical pmf of another sequence z^n (i.e. P_{z^n}).

drawn from Y (i.e. $L = 0$) and then tries to hamper the decision by corrupting the training sequence only.

To go on, we need to find a more explicit expression for $\Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha)$. To this purpose, we reformulate the acceptance region, defined in (6), in an easier-to-handle manner. Let us observe that in order for $P \in \mathcal{P}_{(1-\alpha)N}$ to be the type of a sequence $t_{\mathcal{I}}^{(1-\alpha)N}$, obtained from t^N by removing the αN samples indexed by $\bar{\mathcal{I}}$, the sequence t^N should have at least $(1-\alpha)N \cdot P(i)$ symbols i for each i. Equivalently, any type $P \in \mathcal{P}_{(1-\alpha)N}$ must be such that $P_{t^N} = (1-\alpha)P + \alpha C$ for some $C \in \mathcal{P}_{\alpha N}$. More explicitly, $P = (P_{t^N} - \alpha C)/(1-\alpha)$ for some $C \in \mathcal{P}_{\alpha N}$. Accordingly, by varying C in $\mathcal{P}_{\alpha N}$ we span all the possible types of the sequence $t_{\mathcal{I}}^{(1-\alpha)N}$. The acceptance region in (6) can then be rewritten as follows:

$$\begin{aligned}\Lambda^{n,*} = \{&(P_{x^n}, P_{t^N}) : \exists C \in \mathcal{P}_{\alpha N} \\ &\text{s.t. } h\left(P_{x^n}, \frac{P_{t^N} - \alpha C}{(1-\alpha)}\right) \leq \lambda - \delta_{n,c}\bigg\},\end{aligned} \tag{16}$$

which for a fixed corrupted training sequence corresponds to:

$$\begin{aligned}\Lambda^{n,*}(P_{t^N}) = \{&P_{x^n} : \exists C \in \mathcal{P}_{\alpha N} \\ &\text{s.t. } h\left(P_{x^n}, \frac{P_{t^N} - \alpha C}{(1-\alpha)}\right) \leq \lambda - \delta_{n,c}\bigg\}.\end{aligned} \tag{17}$$

Accordingly, the set $\Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha)$ takes the form:

$$\begin{aligned}&\Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha) = \\ &\quad \{P \in \mathcal{P}_n : \exists C', C \in \mathcal{P}_{\alpha N} \text{ s.t.} \\ &\quad h\bigg(P, P_{\tau^{(1-\alpha)N}} + \frac{\alpha}{(1-\alpha)}(C' - C)\bigg) \leq \lambda - \delta_{n,c}\},\end{aligned} \tag{18}$$

where the second argument of h denotes the generic type in $\mathcal{P}_{(1-\alpha)N}$ obtained from the original training sequence $\tau^{(1-\alpha)N}$ by first adding αN samples and later removing (in a possibly different way) the same number of samples. Note that in this formulation C' accounts for the fake samples introduced by the attacker and C for the part of the samples removed by the defender.

We are now ready to derive the asymptotic payoff of the game by following the same path used in [2], [3]. Such a path consists in the following steps: i) the sets $\Gamma^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha)$ and $\Gamma_0^n(P_{\tau^{(1-\alpha)N}}, \lambda, \alpha)$ are generalized so that they can be applied to a generic pmf $Q \in \mathcal{P}$ (that is, without requiring that the pmf is induced by a sequence of length n); ii) the asymptotic counterparts of Γ^n and Γ_0^n are obtained by letting n tend to infinity. Point i) passes through the generalization of the h function so that it can be applied to two generic pmf's. Specifically we define:

$$\begin{aligned}h_c(P,Q) &= \mathcal{D}(P||U) + c\mathcal{D}(Q||U); \\ U &= \frac{1}{1+c}P + \frac{c}{1+c}Q.\end{aligned} \tag{19}$$

Then we redefine the sets Γ^n and Γ_0^n for a generic pmf Q:

$$\begin{aligned}&\Gamma^n(Q,\lambda,\alpha,L) = \\ &\quad \{P \in \mathcal{P}_n : \exists S_{PR} \in \mathcal{A}(L,P) \text{ s.t. } R \in \Gamma_0^n(Q,\lambda,\alpha)\},\end{aligned} \tag{20}$$

$$\begin{aligned}\Gamma_0^n(Q,\lambda,\alpha) = &\{P \in \mathcal{P}_n : \exists C' \in \mathcal{P}_{\alpha N} \text{ s.t.} \\ &\quad P \in \Lambda^{n,*}((1-\alpha)Q + \alpha C')\} \\ = &\{P \in \mathcal{P}_n : \exists C', C \in \mathcal{P}_{\alpha N} \text{ s.t.} \\ &\quad h_c\left(P, Q + \frac{\alpha}{(1-\alpha)}(C' - C)\right) \leq \lambda - \delta_{n,c}\}.\end{aligned} \tag{21}$$

where the set $\Lambda^{n,*}(Q)$ is generalized in the same way. Finally, the asymptotic extensions of $\Gamma^n(Q,\lambda,\alpha,L)$, $\Gamma_0^n(Q,\lambda,\alpha)$ and $\Lambda^{n,*}(Q)$ are obtained by letting $n \to \infty$ and removing the constraint that P belong to $\mathcal{P}_n$ (that is the element of P no longer need to be rational number with denominator n). In the following we will refer to such sets as $\Gamma(Q,\lambda,\alpha,L)$, $\Gamma_0(Q,\lambda,\alpha)$ and $\Lambda^*(Q)$. We now have all the necessary tools to state the following theorem.

Theorem 2 (Asymptotic payoff of the $SI_{c\text{-}tr}$ game)**.** *For the $SI_{c\text{-}tr}$ game, the false negative error exponent at the equilibrium is given by*

$$\varepsilon = \min_Q [\mathcal{D}(Q||P_X) + \min_{P \in \Gamma(Q,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)]. \tag{22}$$

Accordingly,

1) $P_Y \in \Gamma(P_X,\lambda,\alpha,L)$ *then* $\varepsilon = 0$*;*
2) $P_Y \notin \Gamma(P_X,\lambda,\alpha,L)$ *then* $\varepsilon > 0$.

Proof: The proof can be seen as a particular application of Sanov's theorem [8] which exploits the density of rational numbers in the real line to show that $\Gamma^n(P_X,\lambda,\alpha,L)$ approaches $\Gamma(P_X,\lambda,\alpha,L)$ when $n \to \infty$. Being the proof very similar to that of Theorem 4 in [3], we omitt it for sake of brevity. ■

We observe that the expression of the error exponent given in (22) has the same form of the error exponent of the SI_{tr} game studied in [3], the only difference being the shape of the region over which the inner minimization is performed. As an immediate consequence of Theorem 2, the set $\Gamma(P_X,\lambda,\alpha,L)$ defines the *indistinguishability region* of the test, that is the set of all the sources for which A, by directly modifying the sequences emitted by the source and by properly corrupting the training set, is able to induce D to decide in favor of H_0.

We conclude this section by observing that the asymptotic version of the optimum attacker's strategy does not depend anymore on the to-be-attacked sequence y^n. In fact, the attacker needs only to find a C' which modifies the acceptance region in such a way that it is possible to find an admissible transportation map moving P_Y within it. Then, the optimum corruption strategy depends on P_Y rather than P_{y^n}. In hindsight, the reason for such a result is that, due to the law of large numbers, the type of the sequences generated by Y will tend to P_Y in probability hence making it possible to the attacker to rely only on the knowledge of P_Y.

IV. Security margin and blinding corruption level (α_b)

As a final step, we are interested in studying the behavior of the game for $\lambda \to 0$ in order to derive the best achievable performance for D. Stated in another way, our goal is to study the limit of the indistinguishability region when $\lambda \to 0$. This limit, in fact, provides all the pmf's P_Y that can not be distinguished from P_X ensuring that the two types of error probabilities tend to zero exponentially fast (with vanishingly small, yet positive, error exponents). Such an analysis corresponds to extend the Stein lemma to the adversarial setup considered here, in a way that resembles [4]. First of all, we observe that optimal transport theory permits us to rewrite the indistinguishability region $\Gamma(P_X,\lambda,\alpha,L)$ as:

$$\Gamma(P_X,\lambda,\alpha,L) = \{P : \exists R \in \Gamma_0(P_X,\lambda,\alpha) \text{ s.t. } \mathit{EMD}(P,R) \leq L\}, \tag{23}$$

where *EMD* (Earth Mover Distance) is the term used in signal and image processing applications to denote the minimum cost of the transportation [14], [15], that is

$$\mathit{EMD}(P,R) = \min_{S_{PR}: S_P = P, S_R = R} \sum_{i,j} S_{PR}(i,j)d(i,j). \tag{24}$$

To get a more insightful interpretation of the set (23), we investigate the behavior of the game when no distortion is allowed to the attacker (i.e. $L = 0$), in which case the indistinguishability region of the test is given by $\Gamma_0(P_X, \lambda, \alpha)$. By observing that $h_c(P, Q) = 0$ if and only if $P = Q$, when λ tends to 0, the defender asymptotically accepts H_0 only if $P_Y \in \Gamma_0(P_X, \alpha)$, with $\Gamma_0(P_X, \alpha)$ defined as:

$$\Gamma_0(P_X, \alpha) = \{P : \exists C, C' \in \mathcal{P} \text{ s.t. } P = P_X + \frac{\alpha}{(1-\alpha)}(C' - C)\}. \tag{25}$$

It is possible to rewrite the set (25) in a way which avoids the reference to the auxiliary pmf's C and C'. To do so, we observe that $C'(i)$ must be larger than $C(i)$ for all the bins i for which $P(i) > P_X(i)$ (and viceversa). Since C' and C must be valid pmf's, we argue that $\sum_i [C'(i) - C(i)]^+ = \sum_i [C(i) - C'(i)]^+ \leq 1$ (where $[a]^+ = a$ if $a \geq 0$ and zero otherwise). Then, it is easy to see that (25) is equivalent to the following definition:

$$\begin{aligned} \Gamma_0(P_X, \alpha) &= \left\{ P : \sum_i [P(i) - P_X(i)]^+ \leq \frac{\alpha}{(1-\alpha)} \right\} \\ &= \left\{ P : d_{L_1}(P, P_X) \leq \frac{2\alpha}{(1-\alpha)} \right\}, \end{aligned} \tag{26}$$

where d_{L_1} denotes the L_1 distance. With $\Gamma_0(P_X, \alpha)$ defined as in (26), we can prove the following theorem.

Theorem 3. *Given two sources $X \sim P_X$ and $Y \sim P_Y$, a maximum allowed average per-letter distortion L and a fraction α of training samples provided by the attacker, the maximum achievable false negative error exponent ε for the $SI_{c\text{-}tr}$ game is:*

$$\lim_{\lambda \to 0} \lim_{n \to \infty} -\frac{1}{n} \log P_{fn} = \min_Q [\mathcal{D}(Q||P_X) + \min_{P \in \Gamma(P_X, \alpha, L)} \mathcal{D}(P||P_Y)], \tag{27}$$

where,

$$\begin{aligned} \Gamma(P_X, \alpha, L) &= \{P : \exists R \in \Gamma_0(P_X, \alpha) \textit{ s.t. } \text{EMD}(R, P) \leq L\} \\ &= \left\{ P : \min_{R:\text{EMD}(P,R) \leq L} \sum_i [R(i) - P_X(i)]^+ \leq \frac{\alpha}{(1-\alpha)} \right\}. \end{aligned} \tag{28}$$

Proof: The proof goes along the same steps used in the sketched-proof of Theorem 3 in [4] and is skipped for lack of space. ■

According to Theorem 3, $\Gamma(P_X, \alpha, L)$ provides the *ultimate indistinguishability region* of the test, that is the set of all the pmf for which D will be defeated. Before going on, it is interesting to clarify the geometrical meaning of set $\Gamma_0(P_X, \alpha)$ in (25), by rewriting it as follows

$$\Gamma_0(P_X, \alpha) = \{P : \exists C' \in \mathcal{P} \text{ s.t. } P \in \Lambda^*_{\lambda \to 0}((1-\alpha)P_X + \alpha C')\}, \tag{29}$$

where $\Lambda^*_{\lambda \to 0}(P)$ plays the role of the *ultimate* acceptance region of the test and derives from $\Lambda^*(P)$ by letting λ go to 0:

$$\Lambda^*_{\lambda \to 0}(P) = \left\{ P' : \exists C \in \mathcal{P} \text{ s.t. } P' = \frac{P - \alpha C}{(1-\alpha)} \right\}. \tag{30}$$

With reference to Figure 2, left, we can geometrically interpret $\Lambda^*_{\lambda \to 0}(P)$ as the set of the points P' such that P is convex combination (with coefficient α) of P' with a point C of the probability simplex. Then, according to (29), $\Gamma_0(P_X, \alpha)$ is geometrically obtained as the union of the acceptance regions built over the points which are convex combination of P_X with some point C' in the simplex; this corresponds to an hexagonal space around P_X which, in the probability simplex, is equivalent to the set of the points whose L_1 distance from P_X is constrained to $2\alpha/(1-\alpha)$ (as stated in (26)). Obviously, only the points of this space which lie inside the simplex are valid pmf's and then must be accounted for. A pictorial representation of set $\Gamma(P_X, \alpha, L)$ is given in Figure 2, right, for a smaller value of α.

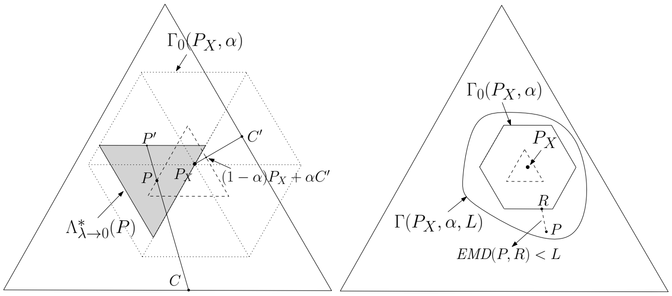

Fig. 2. Geometrical construction of $\Gamma_0(P_X, \alpha)$ (left) and geometrical interpretation of Theorem 3 (right).

By a closer inspection of the *ultimate indistinguishability region* $\Gamma(P_X, \alpha, L)$, we can derive some interesting parameters characterizing the distinguishability of two sources in adversarial setting (both with or without corrupted training, the latter case corresponding to $\alpha = 0$.). Let $X \sim P_X$ and $Y \sim P_Y$ be two sources. Let us focus first on the case in which the attacker can not modify the test sequence ($L = 0$). In this situation, the ultimate indistinguishability region boils down to $\Gamma_0(P_X, \alpha)$. We conclude that D can tell the two sources apart if $d_{L_1}(P_Y, P_X) > \frac{2\alpha}{(1-\alpha)}$. On the contrary, if $d_{L_1}(P_Y, P_X) \leq \frac{2\alpha}{(1-\alpha)}$, A is able to make the sources indistinguishable by corrupting the training sequence. As expected, the larger the α the easier is for A to win the game. By adopting a different perspective, we can defined the blinding corruption level α_b, for which two sources can not be distinguished. Specifically, we have:

$$\alpha_b(P_X, P_Y) = \frac{\sum_i [P_Y(i) - P_X(i)]^+}{1 + \sum_i [P_Y(i) - P_X(i)]^+} = \frac{d_{L_1}(P_Y, P_X)}{2 + d_{L_1}(P_Y, P_X)}. \tag{31}$$

From (31) it is easy to see that α_b is always lower that $1/2$. indeed, for $\alpha \geq 1/2$, there is always a choice of the set $\mathcal{I}$ for which no original sample remains in the training subsequence analyzed by D, hence making a reliable decision impossible. The limit situation $\alpha_b = 1/2$ corresponds to a case in which the P_X and P_Y have completely disjoint supports.

Let us now consider the more general case in which $L \neq 0$. For a given $\alpha < \alpha_b$, we look for the maximum attacking distortion allowing D to reliably distinguish between the two sources. From equation (28), it is easy to argue that the defender is able to distinguish X and Y despite the attack if $\min_{R:EMD(P_Y,R) \leq L} d_{L_1}(R, P_X) > \frac{2\alpha}{(1-\alpha)}$. This leads to the following definition, which extends the concept of security margin, introduced in [4], to the more general setup considered in his paper.

Definition 2 (Security Margin in the $SI_{c\text{-}tr}$ setup)**.** *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 $SI_{c\text{-}tr}$ setup is called Security Margin and is given by*

$$\mathcal{SM}_\alpha(P_X, P_Y) = L^*_\alpha, \tag{32}$$

where $L^*_\alpha = 0$ if $P_Y \in \Gamma_0(P_X, \alpha)$, whereas, if $P_Y \notin \Gamma_0(P_X, \alpha)$, L^*_α is the quantity which satisfies

$$\min_{R:\text{EMD}(P_Y,R) \leq L^*_\alpha} d_{L_1}(R, P_X) = \frac{2\alpha}{(1-\alpha)}. \tag{33}$$

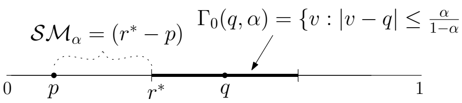
$\mathcal{SM}_\alpha = (r^* - p)$
$\Gamma_0(q,\alpha) = \{v : |v-q| \leq \frac{\alpha}{1-\alpha}\}$
0
p
r^*
q
1

Fig. 3. Geometrical interpretation of the security margin between X and Y. When $\alpha = 0$, $\Gamma_0(q,\alpha)$ boils down to point p and $\mathcal{SM} = (q - p)$ (see [4]).

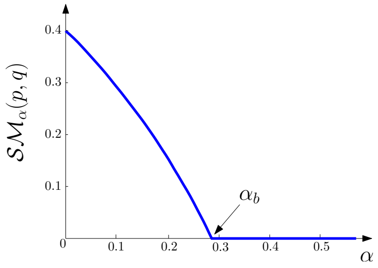

Fig. 4. Security margin as a function of α for Bernoulli sources with parameters $p = 0.3$ and $q = 0.7$ ($\alpha_b = 0.286$).

By focusing on the case $P_Y \notin \Gamma_0(P_X,\alpha)$, since the left-hand side of (33) is a monotonic non increasing function of L_α, the security margin $\mathcal{SM}_\alpha(P_X,P_Y)$ can be expressed in explicit form as

$$\arg\min_{L_\alpha} \min_{R:EMD(P_Y,R)\leq L_\alpha} \left| d_{L_1}(R,P_X) - \frac{2\alpha}{(1-\alpha)} \right|. \qquad (34)$$

When $L > \mathcal{SM}_\alpha(P_X,P_Y)$, it is not possible for D to distinguish between the two sources with positive error exponents of the two kinds. By looking at the behavior of the security margin as a function of α, we see that $\mathcal{SM}_{\alpha_b}(P_X,P_Y) = 0$, meaning that the sources can not be distinguished even if the attacker can not introduce any distortion. On the contrary, setting $\alpha = 0$ corresponds to study the distinguishability of the sources with uncorrupted training, in which case we have $\mathcal{SM}_0 = EMD(P_X,P_Y)$ (in agreement with the results derived in [4]). Moreover, for any $\alpha > 0$, value security margin in (32) is less than $EMD(P_X,P_Y)$. This is also an expected behavior since the general setting considered in this paper is more favorable to the attacker, with respect to the setting in [4].

A. Bernoulli sources

In order to get some insights about the practical meaning of the analysis carried out in the previous sections and the parameters α_b and $\mathcal{SM}_\alpha$, we consider the simple case of two Bernoulli sources with parameter $q = P_X(1)$ and $p = P_Y(1)$. Assuming that no distortion is allowed to the attacker, the (minimum) percentage of samples that A has to modify for inducing a decision error is, according to (31), $\alpha_b = \frac{|p-q|}{1+|p-q|}$. As suggested by intuition, when $|p - q| = 1$, in order for A to win the game, the number of fake samples should be equal to the number of samples of the correct training sequence (i.e. $\alpha = 0.5$). When some distortion is allowed ($L \neq 0$), we have

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p| - \frac{\alpha}{1-\alpha} & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b \end{cases}. \qquad (35)$$

The geometrical meaning of (35) is illustrated in Figure 3 for two generic Bernoulli sources with $p > q$ (w.l.o.g.). Figure 4 depicts the behavior of the $\mathcal{SM}_\alpha(p,q)$ as a function of α when $p = 0.3$ and $q = 0.7$.

V. Conclusions

In this paper we analyzed the distinguishability of two sources in an adversarial setting when the sources are known through training data, part of which can be corrupted by the attacker. We did so by introducing the Source Identification game with corrupted training, then we derived the equilibrium point of the game and analyzed the (asymptotic) payoff at the equilibrium. To summarize all our findings in a compact way, we introduced two parameters, namely the Security Margin under corruption (extending the analysis in [4]), and the blinding corruption percentage α_b, defined as the portion of fake training samples the attacker must introduce to make source distinction impossible. All together, the results we got provide a general framework to cast the source identification problem with training data in, and derive the ultimate performance achievable by the defender for different settings. The goal of our future work will be to extend the analysis so to cover different corrupting scenarios, e.g. a case in which the attacker can also remove some of the correct training samples before making the system available to D. It would also be interesting to further investigate the link between our analysis and secure machine learning as outlined in [6].

Acknowledgment

This work was partially supported by the European Office of Aerospace Research and Development under Grant FA8655-12-1-2138: AMULET - A multi-clue approach to image forensics - and by the REWIND Project, funded by the Future and Emerging Technologies (FET) programme within the 7FP of the EC, under grant 268478.

References

[1] M. Barni and F. Pérez-González, “Coping with the enemy: advances in adversary-aware signal processing,” in *ICASSP 2013, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vancouver, Canada, 26-31 May 2013, pp. 8682–8686.

[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] ——, “Binary hypothesis testing game with training data,” *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, August 2014.

[4] ——, “The security margin: a measure of source distinguishability under adversarial conditions,” in *Proc. of GlobalSip’13, IEEE Global Conference on Signal and Information Processing*, Austin, Texas, 3-5 December 2013.

[5] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?” in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS ’06. New York, NY, USA: ACM, 2006, pp. 16–25. [Online]. Available: http://doi.acm.org/10.1145/1128817.1128824

[6] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, “The security of machine learning,” *Machine Learning*, vol. 81, pp. 121–148, 2010.

[7] M. Barni and B. Tondi, “Source distinguishability under distortion-limited attack: an optimal transport perspective,” *submitted on IEEE Transactions on Information Theory*, July 2014.

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

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

[10] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, March 1989.

[11] M. Kendall and S. Stuart, *The Advanced Theory of Statistics, vol. 2, 4th edition*. New York: MacMillan, 1979.

[12] 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.

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

[14] S. T. Rachev, *Mass Transportation Problems: Volume I: Theory*. Springer, 1998, vol. 1.

[15] 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.