Adversarial Source Identification Game with Corrupted Training

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

***Abstract*—We study a variant of the source identification game with training data in which part of the training data is corrupted by an attacker. In the addressed scenario, the defender aims at deciding whether a test sequence has been drawn according to a discrete memoryless source $X \sim P_X$, whose statistics are known to him through the observation of a training sequence generated by X. In order to undermine the correct decision under the alternative hypothesis that the test sequence has not been drawn from X, the attacker can modify a sequence produced by a source $Y \sim P_Y$ up to a certain distortion and corrupt the training sequence either by adding some fake samples or by replacing some samples with fake ones. We derive the unique rationalizable equilibrium of the two versions of the game in the asymptotic regime and by assuming that the defender makes his decision by relying only on the first order statistics of the test and the training sequences. By mimicking Stein's lemma, we derive the best achievable performance for the defender when the first type error probability is required to tend to zero exponentially fast with an arbitrarily small, yet positive, error exponent. We then use such a result to analyze the ultimate distinguishability of any two sources as a function of the allowed distortion and the fraction of corrupted samples injected into the training sequence.**

***Index Terms*—Hypothesis testing, adversarial signal processing, cybersecurity, game theory, source identification, optimal transportation theory, earth mover distance, adversarial learning, Sanov's theorem.**

I. Introduction

Adversarial Signal Processing (AdvSP) is an emerging discipline aiming at modelling the interplay between a defender wishing to carry out a certain processing task, and an attacker aiming at impeding it [1]. Binary decision in an adversarial setup is one of the most recurrent problems in AdvSP, due to its importance in many application scenarios. Among binary decision problems, source identification is one of the most studied subjects, since it lies at the heart of several security-oriented disciplines. In network monitoring, for instance, the analyst is asked to decide if the observed traffic has been generated under normal network conditions or it provides enough evidence that an attack is ongoing. In multimedia forensics, we may want to distinguish which between two sources (e.g. two photo cameras) generated a given image, in other situations, the security of a system relies on the capability of distinguishing the behaviour of malevolent and fair users.

The source identification game has been introduced in [2] to allow the study of source identification under adversarial conditions on a rigorous basis. According to the setup considered in [2], the analyst (hereafter referred to as the defender) is asked to decide whether an observed sequence v^n has been generated by a discrete memoryless source X with known probability mass function P_X or not. The attacker, on his hand, takes a sequence generated by another source Y and modifies it in the attempt to make the defender decide that the modified sequence was generated by X. The attack must satisfy a distortion constraint, setting the maximum amount of distortion that can be introduced in the attacked sequence. The structure of the game is defined by assuming that the decision strategy adopted by the defender satisfies a constraint on the false positive error probability (that is, the probability of deciding that v^n was not generated by X, when in fact it was), while the payoff of the game is defined in terms of the false negative error probability (that is, the probability that a sequence generated by Y and modified by the attacker is said to be generated by X). In [2], the asymptotic equilibrium point of the game is derived by assuming that the length n of the observed sequence tends to infinity.

According to the model put forward in [2], the defender and the attacker have a perfect knowledge of the source X, i.e. they know P_X. In [3], the analysis is pushed a step forward by considering a scenario in which the source X is known only through the observation of a training sequence t^n. This is an interesting extension of the setup studied in [2]. The use of a training sequence to gather information about the statistics of the sources, in fact, can be seen as a simple learning mechanism, linking the results derived in [3] to machine learning and adversarial machine learning [4] in particular. The analysis carried out in [2] and [3] has been further revised and extended in [5], with the introduction of the notion of security margin, a synthetic parameter characterising the ultimate distinguishability of two sources under adversarial conditions.

In this paper, we extend the analysis further, by considering a situation in which the attacker interferes with the learning phase by corrupting part of the training sequence. From a theoretical point of view, this represents a major deviation from the analysis carried out in [2], [3], [5]. The first and most important consequence of the possibility that the training sequence has been corrupted by the attacker, is that, under the new setup, the attack influences also the accuracy of the decision under the hypothesis that the sequence has been generated by X. In other words, the action of the attacker has an impact on both the false positive and false negative error probability. This was not the case in the setup originally introduced in [2] and inherited in [3], [5], where the false positive error probability was independent of the strategy chosen by the

M. Barni is with the Department of Information Engineering and Mathematics, University of Siena, Via Roma 56, 53100 - Siena, ITALY, phone: +39 0577 234850 (int. 1005), e-mail: barni@dii.unisi.it; B. Tondi is with the Department of Information Engineering and Mathematics, University of Siena, Via Roma 56, 53100 - Siena, ITALY, e-mail: benedettatondi@gmail.com.

attacker. As a result, the fulfilment of the constraint on the false positive error probability requires that the possible actions of the attackers are taken into account, by adopting a worst case approach. Such a fundamental modification of the structure of the game influences all the rest of the analysis, thus calling for the adoption of more powerful tools, and leading to new results that incorporate those derived in [2], [3], [5] as limit cases, but substantially departs form them. From a practical point of view, encompassing the case of a corrupted training sequence permits to extend the applicability of the analysis to situations in which the collection of the training data is not under the full control of the analyst. This is the case in many modern applications of machine learning wherein the data used in the training phase is collected in a non-controlled environment, e.g. by resorting to crowdsourcing or on-line learning with the risk that part of the data is altered with the aim of facilitating a subsequent attack [6].

A. Links with adversarial machine learning and sample applications

The analysis carried out in this paper is strongly related to adversarial machine learning [7]. Adversarial learning is a rather novel concept, which is receiving a growing attention due to the ubiquitous use of machine learning techniques in an ever increasing number of applications [4], [6], [8], [9]. 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. Attacks against a machine learning system can be classified according to different perspectives. To start with, and by following the taxonomy introduced in [7], we can distinguish the attacks according to the moment when the attack is applied. According to such a perspective, we can distinguish between *causative* and *exploratory* attacks. In the former case, the attacker corrupts the training process to cause a subsequent classification error. In the latter situation, the attack is carried during the classification phase, trying to build a test sample that causes a classification error. Another possible taxonomy considers the kind of error the attacker aims at: the aim of an in *integrity violation* is to cause a false positive error, e.g. to avoid that an anomalous situation is detected, or to allow the access to a system or service to a non-allowed user. On the contrary, attacks aiming at an *availability violation* try to induce a false negative error, e.g., to deny the access to a service to a legitimate user. According to the above taxonomy, the scenario considered in this paper corresponds to a *causative* attack, while prior works, noticeably [3], were focusing on *exploratory* attacks only. With regard to the kind of errors the attacker aims at, we are considering an *integrity violation* attack (an example of a work considering also availability violation is given in [10]).

Despite the mostly theoretical nature of our study, due to the difficulties to model all the subtleties typical of real world applications, the results derived in this work may guide a first rough analysis in a wide range of application scenarios. As an example, we may consider a network traffic monitoring system aiming at discriminating normal traffic conditions from anomalous situations possibly indicating the presence of a denial of service, or any other kind of attack. Let us assume that the system relies on inter-packet arrival time to make its decision about the status of the network; due to the lack of a good theoretical model describing the statistics of inter-arrival times, the system observes the network under normal conditions to learn a model of the observations in the absence of attacks. On his hand, the hacker may shape the characteristics of the attack so to evade the detection, however, if he can also corrupt at least part of the data gathered during the learning phase, e.g., by injecting within it a certain percentage of false samples, he will surely be able to mount a more powerful attack. A similar situation may occur in spam-filtering applications. If the filtering service is built by relying on a training phase in which the system learns the statistics of legitimate e-mails, the spammer may ease the subsequent construction of spam samples capable of evading the filter control, by corrupting the learning process, e.g., by introducing within the legitimate e-mails observed during the learning phase some words that he is going to use afterwards to pass the anti-spam check.

B. Contributions and main results

Following [2] and [3], we model the interplay between the attacker and the defender by using a game theoretic approach according to which each player knows only the possible strategies available to his opponent, but does not know which strategy he is actually going to use. The set of strategies available to the defender corresponds to the possible detection rules he can adopt, while the attacker must decide how to corrupt the training data (up to his maximum capacity) and the test data so to induce a decision error. As to the payoff, we assume a zero-sum competitive game, where the attacker aims at increasing the false negative error probability while the defender aims at minimising it.

Given the above general framework, the main contributions of this work can be summarised as follows:

1) We give a rigorous definition of the game, and we derive the optimum choices for the defender and the attacker in the form of equilibrium points of the game, when the length of the training sequence and the observed sequence tends to infinity. In doing so, we prove that the game is a dominance solvable game, since a dominant strategy exists for the defender which is optimum regardless o the choice made by the attacker (Theorem 1 and following discussion in Section III).
2) Given the equilibrium point, we analyse the payoff (namely the false negative error probability) at the equilibrium. Given a source X, such an analysis permits to determine the region (called indistinguishability region) of the sources that can not be reliably distinguished from X due to the attack (Theorems 2 and 3, Section III).
3) By mimicking, and considerably extending, the analysis made in [5], for any two sources X and Y, we derive the security margin and the blinding corruption level, defined as the maximum distortion of the test sequence and the maximum fraction of fake samples introduced into the training set, still allowing the distinction of X and Y while ensuring positive error exponents for the two kinds of errors of the test (Section V).

Throughout the paper, we consider two different scenarios wherein the attacker is allowed respectively to *add* a certain amount of fake samples to the training sequence and to selectively *replace* a fraction of the samples of the training sequences with fake samples. As we will see, the second case is more favourable to the attacker, since a lower distortion and a lower number of corrupted training samples are enough to prevent a correct decision.

A further, methodological, contribution regards the techniques used to prove the main results of the paper. As opposed to previous works, such proofs rely on a generalised version of Sanov's theorem [11], [12], which is proven in Appendix A. The use of such a generalised version of Sanov's theorem, in fact, permits to simplify considerably some of the proofs.

This work considerably extends the analysis presented in [13], by providing a formal proof of the results anticipated in [13][1] and by studying the more complex corruption scenario in which the attacker has the freedom to replace a given percentage of the training samples rather than simply adding some fake samples to the original training sequence (which was the only case considered in [13]).

The rest of this paper is organised as follows. Section II summarises the notation used throughout the paper, gives some definitions and introduces some basic concepts of Game theory that will be used in the sequel. Section III gives a rigorous definition of the source identification game with corrupted training. In Section IV, we prove the main theorems of the paper regarding the asymptotic equilibrium point of the game and the payoff at the equilibrium. Section V leverages on the results proven in Section IV to introduce the concepts of blind corruption level and security margin. Section VI, introduces and solves a version of the game in which the attacker can selectively replace a percentage of training samples, by paying attention to compare the results of the analysis with the results proven in the previous sections. The paper ends in Section VII, with a summary of the main results proven in the paper and the description of possible directions for future work. In order to avoid burdening the main body of the paper, the most technical details of the proofs are gathered in the Appendix.

II. Notation and definitions

In this section, we introduce the notation and definitions used throughout the paper. We will use capital letters to indicate discrete memoryless sources (e.g. X). Sequences of length n drawn from a source will be indicated with the corresponding lowercase letters (e.g. x^n); accordingly, x_i will denote the i-th element of a sequence x^n. The alphabet of an information 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. The calligraphic letter $\mathcal{P}$ will be used to indicate the class of all the probability mass functions, namely, the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$. The notation P_X will be also used to indicate the probability measure ruling the emission of sequences from a source X, so we will use the expressions $P_X(a)$ and $P_X(x^n)$ to indicate, respectively, the probability of symbol $a{\in}\mathcal{X}$ and the probability that the source X emits the sequence x^n, the exact meaning of P_X being always clearly recoverable from the context wherein it is used. We will use the notation $P_X(A)$ to indicate the probability of A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$) under the probability measure P_X. Finally, the probability of a generic will be denoted by $Pr\{\}$.

Our analysis relies extensively on the concepts of type and type class defined as follows (see [11] and [14] for more details). 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]:

$$\mathcal{D}(P||Q){=}\sum_{a\in\mathcal{X}}P(a)\log_2\frac{P(a)}{Q(a)}. \qquad (1)$$

Most of our results are expressed in terms of the generalised log-likelihood ratio function h (see [3], [15], [16]), which for any two given sequences x^n and t^m is defined as:

$$h(P_{x^n},P_{t^m}){=}\mathcal{D}(P_{x^n}||P_{r^{n+m}}){+}\frac{m}{n}\mathcal{D}(P_{t^m}||P_{r^{n+m}}), \qquad (2)$$

where $P_{r^{n+m}}$ denotes the type of the sequence r^{n+m}, obtained by concatenating x^n and t^m, i.e. $r^{n+m}{=}x^n||t^m$. The intuitive meaning behind the above definition is that $P_{r^{n+m}}$ is the pmf which maximises the probability that a memoryless source generates two independent sequences belonging to $T(P_{x^n})$ and $T(P_{t^m})$, and that such a probability is equal to $2^{-nh(P_{x^n},P_{t^m})}$ at the first order in the exponent (see [16] or Lemma 1 in [3]).

Throughout the paper, we will need to compute *limits* and *distances* in $\mathcal{P}$. We can do so by choosing one of the many available distances defined over $\mathbb{R}^{|\mathcal{X}|}$ and for which $\mathcal{P}$ is a bounded set, for instance the L_p distance for which we have:

$$d_{L_p}(P,Q){=}\left(\sum_{a\in\mathcal{X}}|P(a){-}Q(a)|^p\right)^{1/p}. \qquad (3)$$

Without loss of generality, we will prove all our results by adopting the L_1 distance, the generalisation to different L_p metrics being straightforward. In the sequel, distances between pmf's in $\mathcal{P}$ will be simply indicated as $d(\cdot,\cdot)$ as a shorthand for $d_{L_1}(\cdot,\cdot)$[2].

We also need to introduce the *Hausdorff distance* as a way to measure distances between subsets of a metric space [17].

[1] We also give a more precise formulation of the problem, by correcting some inaccuracies present in [13]. In particular, we reformulated the definition of the space of strategies of the defender in a more general form, without constraining the defender to base his decision on subsequences of the training sequence (which, in principle, might not be the optimum strategy for the defender and then should not be assumed a-priori). We also corrected a formal inaccuracy in the definition of the game given in [13], regarding the space wherein the acceptance region is defined.

[2] Throughout the paper, we will use the symbol $d(\cdot,\cdot)$ to indicate both the distortion between two sequences in $\mathcal{X}^n$ and the L_1 distance between two pmf's in $\mathcal{P}$, the exact meaning being always clear from the context,

Let S be a generic space and d a distance measure defined over S. For any point $x{\in}S$ and any non-empty subset $A{\subseteq}S$, the distance of x from the subset A is defined as:

$$d(x,A) = \inf_{a \in A} d(a,x). \tag{4}$$

Given the above definition, the Hausdorff distance between any two subsets of S is defined as follows.

Definition 1. *For any two subsets A and B of S, let us define $\delta_B(A){=}\sup_{b \in B} d(b,A)$. The Hausdorff distance $\delta_H(A,B)$ between A and B is given by:*

$$\delta_H(A,B) = \max\{\delta_A(B), \delta_B(A)\}. \tag{5}$$

If the sets A and B are bounded with respect to d, then the Hausdorff distance always takes a finite value. The Hausdorff distance does not define a true metric, but only a pseudometric, since $\delta_H(A,B){=}0$ implies that the closures of the sets A and B coincide, namely $cl(A){=}cl(B)$, but not necessarily that $A{=}B$. For this reason, in order for δ_H to be a metric, we need to restrict its definition to closed subsets[3]. Let then $\mathcal{L}(S)$ denote the space of non-empty closed and limited subsets of S and let $\delta_H{:}\mathcal{L}(S){\times}\mathcal{L}(S){\rightarrow}[0,\infty)$. Then, the space $\mathcal{L}(S)$ endowed with the Hausdorff distance is a metric space [18] and we can give the following definition:

Definition 2. *Let $\{K_n\}$ be a sequence of closed and limited subsets of S, i.e., $K_n{\in}\mathcal{L}(S)\ \forall n$. We use the notation $K_n \overset{H}{\rightarrow} K$ to indicate that the sequence has limit in $(\mathcal{L}(S),\delta_H)$ and the limiting set is K.*

A. Basic notions of Game Theory

In this section, we introduce some basic notions and definitions of Game Theory.

A 2-player game is defined as a quadruple $(\mathcal{S}_1,\mathcal{S}_2,u_1,u_2)$, where $\mathcal{S}_1{=}\{s_{1,1} \ldots s_{1,n_1}\}$ and $\mathcal{S}_2{=}\{s_{2,1} \ldots s_{2,n_2}\}$ 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 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 $u_1(s_{s1,i},s_{2,j}){=}{-}u_2(s_{1,i},s_{2,j})$, the win of a player is equal to the loss of the other and the game is said to be a zero-sum game. For such games, the payoff of the game $u(s_{s1,i},s_{2,j})$ is defined by adopting the perspective of one of the two players (that is, $u(s_{s1,i},s_{2,j}){=}u_1(s_{s1,i},s_{2,j}){=}{-}u_2(s_{1,i},s_{2,j})$ if the defender's perspective is adopted or viceversa). The sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players. Throughout the paper we consider strategic games, i.e., games in which the players choose their strategies beforehand without knowing the strategy chosen by the opponent player.

The final goal of game theory is to determine the existence of equilibrium points, i.e. profiles that in *some sense* represent the *best* choice for both players [19]. The most famous notion of equilibrium is due to Nash. A profile is said to be a Nash equilibrium if no player can improve its payoff by changing its strategy unilaterally. Despite its popularity, the practical meaning of Nash equilibrium is often unclear, since there is no guarantee that the players will end up playing at the equilibrium. A particular kind of games for which stronger forms of equilibrium exist are the so called *dominance solvable* games [19]. To be specific, a strategy is said to be strictly dominant for one player if it is the best strategy for the player, i.e., the strategy which corresponds to the largest payoff, no matter how the other player decides to play. When one such strategy exists for one of the players, he will surely adopt it. In a similar way, we say that a strategy $s_{l,i}$ is strictly dominated by strategy $s_{l,j}$, if the payoff achieved by player l choosing $s_{l,i}$ is always lower than that obtained by playing $s_{l,j}$ regardless of the choice made by the other player. The recursive elimination of dominated strategies is a common technique for solving games. In the first step, all the dominated strategies are removed from the set of available strategies, since no rational player would ever play them. In this way, a new, smaller game is obtained. At this point, some strategies, that were not dominated before, may be dominated in the remaining game, and hence are eliminated. The process goes on until no dominated strategy exists for any player. A *rationalizable equilibrium* is any profile which survives the iterated elimination of dominated strategies [20], [21]. If at the end of the process only one profile is left, the remaining profile is said to be the *only rationalizable equilibrium* of the game. The corresponding strategies are the only rational choice for the two players and the game is said *dominance solvable*.

III. Source identification game with addition of corrupted training samples ($SI^a_{c\text{-}tr}$)

In this section, we give a rigorous definition of the Source Identification game with addition of corrupted training samples.

Given a discrete and memoryless source $X{\sim}P_X$ and a test sequence v^n, the goal of the defender ($\mathscr{D}$) is to decide whether v^n has been drawn from X (hypothesis H_0) or not (alternative hypothesis H_1). By adopting a Neyman-Pearson perspective, we assume that $\mathscr{D}$ must ensure that the false positive error probability (P_{fp}), i.e., the probability of rejecting H_0 when H_0 holds (type I error) is lower than a given threshold. Similarly to the previous versions of the game studied in [2] and [3], we assume that $\mathscr{D}$ relies only on first order statistics to make a decision. For mathematical tractability, like earlier papers, we study the asymptotic version of the game when $n{\rightarrow}\infty$, by requiring that P_{fp} decays exponentially fast when n increases, with an error exponent at least equal to λ, i.e. $P_{fp}{\leq}2^{-n\lambda}$. On his side, the attacker ($\mathscr{A}$) aims at increasing the false negative error probability (P_{fn}), i.e., the probability of accepting H_0 when H_1 holds (type II error). Specifically, A takes a sequence y^n drawn from a source $Y{\sim}P_Y$ and modifies it in such a way that $\mathscr{D}$ decides that the modified sequence z^n has been generated by X. In doing so, A must respect a distortion constraint requiring that the average per-letter distortion between y^n and z^n is lower than L.

Players $\mathscr{A}$ and $\mathscr{D}$ know the statistics of X through a training sequence, however the training sequence can be partly corrupted by $\mathscr{A}$. Depending on how the training sequence

[3] Note that in this case the inf and sup operations involved in the definition of the Hausdorff distance can be replaced with min and max, respectively.

is modified by the attacker, we can define different versions of the game. In this paper, we focus on two possible cases: in the first case, hereafter referred to as source identification game with addition of corrupted samples $SI^a_{c\text{-}tr}$, the attacker can add some fake samples to the original training sequence. In the second case, analysed in Section VI, the attacker can replace some of the training samples with fake values (source identification game with replacement of training samples - $SI^r_{c\text{-}tr}$). It is worth stressing that, even if the goal of the attacker is to increase the false negative error probability, the training sequence is corrupted regardless of whether H_0 or H_1 holds, hence, in general, this part of the attack also affects the false positive error probability. As it will be clear later on, this forces the defender to adopt a worst case perspective to ensure that P_{fp} is surely lower than $2^{-\lambda n}$.

As to Y, we assume that the attacker knows P_Y exactly. For a proper definition of the payoff of the game, we also assume that $\mathscr{D}$ knows P_Y. This may seem a too strong assumption. However, we will show later on that the optimum strategy of $\mathscr{D}$ does not depend on P_Y, thus allowing us to relax the assumption that $\mathscr{D}$ knows P_Y.

With the above ideas in mind, we are now ready to give a formal definition of the $SI^a_{c\text{-}tr}$ game.

A. Structure of the $SI^a_{c\text{-}tr}$ game

A schematic representation of the $SI^a_{c\text{-}tr}$ game is given in Figure 1.

Let τ^{m_1} be a sequence drawn from X. We assume that τ^{m_1} is accessible to $\mathscr{A}$, who corrupts it by concatenating to it a sequence of fake samples τ^{m_2}. Then $\mathscr{A}$ reorders the overall 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. In the following, we denote by m the final length of the training sequence ($m{=}m_1{+}m_2$), and by $\alpha{=}\frac{m_2}{m_1+m_2}$ the portion of fake samples within it. The corrupted training sequence observed by $\mathscr{D}$ is indicated by t^m. Eventually, we hypothesize a linear relationship between the lengths of the test and the corrupted training sequence, i.e. $m{=}cn$, for some constant value c[4].

The goal of $\mathscr{D}$ is to decide if an observed sequence v^n has been drawn from the same source that generated t^m (H_0) or not (H_1). We assume that $\mathscr{D}$ knows that a certain percentage of samples in the training sequence are corrupted, but he has no clue about the position of the corrupted samples. The attacker can also modify the sequence generated by Y so to induce a decision error. The corrupted sequence is indicated by z^n. With regard to the two phases of the attack, we assume that $\mathscr{A}$ first corrupts the training sequence, then he modifies the sequence y^n. This means that, in general, z^n will depend both on y^n and t^m, while t^m (noticeably τ^{m_2}) does not depend on y^n. Stated in another way, the corruption of the training sequence can be seen as a preparatory part of the attack, whose goal is to ease the subsequent camouflage of y^n.

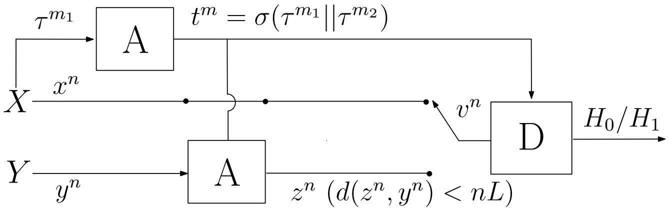
τ^{m_1}
A
$t^m = \sigma(\tau^{m_1}||\tau^{m_2})$
X
x^n
v^n
D
H_0/H_1
Y
y^n
A
$z^n \ (d(z^n, y^n) < nL)$

Fig. 1. Schematic representation of the $SI^a_{c\text{-}tr}$ game. Both training and test sequences are corrupted by the attacker. Symbol $||$ denotes concatenation of sequences and $\sigma()$ is a random permutation of sequence samples.

For a formal definition of the $SI^a_{c\text{-}tr}$ game, we must define the set of strategies available to $\mathscr{D}$ and $\mathscr{A}$ (respectively $\mathcal{S}_{\mathscr{D}}$ and $\mathcal{S}_{\mathscr{A}}$) and the corresponding payoffs.

B. Defender's strategies

The basic assumption behind the definition of the space of strategies available to $\mathscr{D}$ is that to make his decision $\mathscr{D}$ relies only on the first order statistics of v^n and t^m. This assumption is equivalent to requiring that the acceptance region for hypothesis H_0, hereafter referred to as $\Lambda^{n\times m}$, is a union of pairs of type classes[5], or equivalently, pairs of types (P,R), where $P{\in}\mathcal{P}^n$ and $R{\in}\mathcal{P}^m$. To define $\Lambda^{n\times m}$, $\mathscr{D}$ follows a Neyman-Pearson approach, requiring that the false positive error probability is lower than a certain threshold. Specifically, we require that the false positive error probability tends to zero exponentially fast with a decay rate at least equal to λ. Given that the pmf P_X ruling the emission of sequences under H_0 is not known and given that the corruption of the training sequence is going to impair $\mathscr{D}$'s decision under H_0, we adopt a worst case approach and require that the constraint on the false positive error probability holds for all possible P_X and for all the possible strategies available to the attacker. Given the above setting, the space of strategies available to $\mathscr{D}$ is defined as follows:

$$\mathcal{S}_{\mathscr{D}}{=}\{\Lambda^{n\times m}{\subset}\mathcal{P}^n{\times}\mathcal{P}^m\!: \max_{P_X\in\mathcal{P}} \max_{s\in\mathcal{S}_{\mathscr{A}}} P_{fp} \leq 2^{-\lambda n}\}, \tag{6}$$

where the inner maximization is performed over all the strategies available to the attacker. We will refine this definition at the end of the next section, after the exact definition of the space of strategies of the attacker.

C. Attacker's strategies

With regard to $\mathscr{A}$, the attack consists of two parts. Given a sequence y^n drawn from P_Y, and the original training sequence τ^{m_1}, the attacker first generates a sequence of fake samples τ^{m_2} and mixes them up with those in τ^{m_1} producing the training sequence t^m observed by $\mathscr{D}$. Then he transforms y^n into z^n, eventually trying to generate a pair of sequences

[4] In this paper, we are interested in studying the equilibrium point of the source identification game when the length of the test and training sequences tend to infinity. Strictly speaking, we should ensure that when n grows, all the quantities m, m_1 and m_2 are integer numbers for the given c and α. In practice, we will neglect such an issue, since when n grows the ratios m/n and $m_1/(m_1{+}m_2)$ can approximate any real values c and α. More rigorously, we could consider only rational values of c and α, and focus on subsequences of n including only those values for which $m/n{=}c$ and $m_1/(m_1{+}m_2){=}\alpha$.

[5] We use the superscript $n{\times}m$ to indicate explicitly that $\Lambda^{n\times m}$ refers to n-long test sequences and $(m{=}cn)$-long training sequences.

(z^n,t^m)[6] whose types belong to $\Lambda^{n\times m}$. In doing so, he must ensure that $d(y^n,z^n)\leq nL$ for some distortion function d.

Let us consider the corruption of the training sequence first. Given that the defender bases his decision only on the type of t^m, we are only interested in the effect that the addition of the fake samples has on P_{t^m}. By considering the different length of τ^{m_1} and τ^{m_2}, we have:

$$P_{t^m}=\alpha P_{\tau^{m_2}}+(1-\alpha)P_{\tau^{m_1}}, \tag{7}$$

where $P_{t^m}\in\mathcal{P}^m$, $P_{\tau^{m_1}}\in\mathcal{P}^{m_1}$ and $P_{\tau^{m_2}}\in\mathcal{P}^{m_2}$. The first part of the attack, then, is equivalent to choosing a pmf in $\mathcal{P}^{m_2}$ and mixing it up with $P_{\tau^{m_1}}$. By the same token, it is reasonable to assume that the choice of the attacker depends only on $P_{\tau^{m_1}}$ rather than on the single sequence τ^{m_1}. Arguably, the best choice of the pmf in $\mathcal{P}^{m_2}$ will depend on P_Y, since the corruption of the training sequence is instrumental in letting the defender think that a sequence generated by Y has been drawn by the same source that generated t^m.

To describe the part of the attack applied to the test sequence, we follow the approach used in [5] based on transportation theory [22]. 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, let $S^n_{YZ}(i,j)=n(i,j)/n$ be the relative frequency with which such a transformation occurs. In the following, we refer to S^n_{YZ} as *transportation map*. For any additive distortion measure, the distortion introduced by the attack can be expressed in terms of $n(i,j)$ and S^n_{YZ}. In fact, we have:

$$d(y^n,z^n) = \sum_{i,j} n(i,j)d(i,j), \tag{8}$$

$$\frac{d(y^n,z^n)}{n} = \sum_{i,j} S^n_{YZ}(i,j)d(i,j). \tag{9}$$

where $d(i,j)$ is the distortion introduced when symbol i is transformed into symbol j.

The map S^n_{YZ} also determines the type of the attacked sequence. In fact, by indicating with $P_{z^n}(j)$ the relative frequency of symbol j into z^n, we have:

$$P_{z^n}(j) = \sum_i S^n_{YZ}(i,j) \triangleq S^n_Z(j). \tag{10}$$

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 $S^n_Y(i)\triangleq\sum_j S^n_{YZ}(i,j)=P_{y^n}(i)$. Sometimes, we find convenient to explicitly denote the dependence of the map chosen by the attacker on the type of t^m and y^n, and hence we will also adopt the notation $S^n_{YZ}(P_{t^m},P_{y^n})$.

By remembering that $\Lambda^{n\times m}$ depends on v^n only through its type, and given that the type of the attacked sequence depends on S^n_Y only through S^n_{YZ}, we can define the second phase of the attack as the choice of a transportation map among all *admissible* maps, a map being admissible if:

$$\begin{aligned} S^n_Y &= P_{y^n} \\ \sum_{i,j} S^n_{YZ}(i,j)d(i,j) &\leq L. \end{aligned} \tag{11}$$

Hereafter, we will refer to the set of admissible maps as $\mathcal{A}^n(L,P_{y^n})$.

With the above ideas in mind, the set of strategies of the attacker can be defined as follows:

$$\mathcal{S}_{\mathscr{A}} = \mathcal{S}_{\mathscr{A},T}\times\mathcal{S}_{\mathscr{A},O}, \tag{12}$$

where $\mathcal{S}_{\mathscr{A},T}$ and $\mathcal{S}_{\mathscr{A},O}$ indicate, respectively, the part of the attack affecting the training sequence and the observed sequence, and are defined as:

$$\mathcal{S}_{\mathscr{A},T} = \left\{ Q(P_{\tau^{m_1}})\colon \mathcal{P}^{m_1}\to\mathcal{P}^{m_2} \right\}, \tag{13}$$

$$\mathcal{S}_{\mathscr{A},O} = \left\{ S^n_{YZ}(P_{y^n},P_{t^m})\colon \mathcal{P}^n\times\mathcal{P}^m\to\mathcal{A}^n(L,P_{y^n}) \right\}. \tag{14}$$

Note that the first part of the attack ($\mathcal{S}_{\mathscr{A},T}$) is applied regardless of whether H_0 or H_1 holds, while the second part ($\mathcal{S}_{\mathscr{A},O}$) is applied only under H_1. We also stress that the choice of $Q(P_{\tau^{m_1}})$ depends only on the training sequence τ^{m_1}, while the transportation map used in the second phase of the attack is a function of both on y^n and τ^{m_1} (through t^m). Finally, we observe that with these definitions, the set of strategies of the defender can be redefined by explicitly indicating that the constraint on the false positive error probability must be verified for all possible choices of $Q(\cdot)\in\mathcal{S}_{\mathscr{A},T}$, since this is the only part of the attack affecting P_{fp}. Specifically, we can rewrite (6) as

$$\mathcal{S}_{\mathscr{D}}=\{\Lambda^{n\times m}\subset\mathcal{P}^n\times\mathcal{P}^m\colon \max_{P_X} \max_{Q(\cdot)\in\mathcal{S}_{\mathscr{A},T}} P_{fp} \leq 2^{-\lambda n}\}. \tag{15}$$

D. Payoff

The payoff of the game is defined in terms of the false negative error probability, namely:

$$u(\Lambda^{n\times m},(Q(\cdot),\, S^n_{YZ}(\cdot,\cdot))) = -P_{fn}, \tag{16}$$

where the defender's perspective is adopted; then, $\mathscr{D}$ aims at maximising u while $\mathscr{A}$ wants to minimise it.

In Section IV-B, we will show that the game defined above is dominance solvable and we will derive the *rationalizable equilibrium* (see discussion in Section II-A), indicated by $(\Lambda^{n\times m,*},(Q^*(\cdot),\, S^{n,*}_{YZ}(\cdot,\cdot)))$.

E. The $SI^a_{c\text{-}tr}$ game with targeted corruption ($SI^{a,t}_{c\text{-}tr}$ game)

The $SI^a_{c\text{-}tr}$ game is difficult to solve directly, because of the 2-step attacking strategy. We will work around this difficulty by tackling first with a slightly different version of the game, namely the source identification game with targeted corruption of the training sequence, $SI^{a,t}_{c\text{-}tr}$, depicted in Fig. 2 (this corresponds to a targeted attack according to the taxonomy introduced in [7]).

[6] While reordering is essential to hide the position of fake samples to $\mathscr{D}$, it does not have any impact on the position of (z^n,t^m) with respect to $\Lambda^{n\times m}$, since we assumed that the defender bases his decision only on the first order statistic of the observed sequences. For this reason, we omit to indicate the reordering operator σ in the attacking procedure.

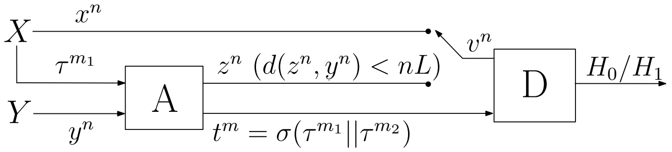
X
x^n
τ^{m_1}
Y
y^n
A
$z^n\ (d(z^n, y^n) < nL)$
$t^m = \sigma(\tau^{m_1}||\tau^{m_2})$
v^n
D
H_0/H_1

Fig. 2. $SI^a_{c\text{-}tr}$ game with targeted corruption of the training sequence ($SI^{a,t}_{c\text{-}tr}$ game).

Whereas the strategies available to the defender remain the same, for the attacker, the choice of $Q(\cdot)$ is targeted to the counterfeiting of a given sequence y^n. In other words, we will assume that the attacker corrupts the training sequence τ^{m_1} to ease the counterfeiting of a specific sequence y^n rather than to increase the probability that the second part of the attack succeeds. This means that the part of the attack aiming at corrupting the training sequence also depend on y^n, that is:

$$\mathcal{S}_{\mathscr{A},T} = \left\{ Q(P_{\tau^{m_1}},P_{y^n})\colon \mathcal{P}^{m_1}\times\mathcal{P}^{n}\to\mathcal{P}^{m_2} \right\}. \tag{17}$$

Even if this setup is not very realistic and is more favourable to the attacker, who can exploit the exact knowledge of y^n (rather than its statistical properties) also for the corruption of the training sequence, in the next section we will show that, for large n, the $SI^{a,t}_{c\text{-}tr}$ game is equivalent to the non-targeted version of the game we are interested in.

With the above ideas in mind, the $SI^{a,t}_{c\text{-}tr}$ game is formally defined as follows.

1) Defender's strategies:

$$\mathcal{S}_{\mathscr{D}}=\{\Lambda^{n\times m}\subset\mathcal{P}^{n}\times\mathcal{P}^{m}\colon \max_{P_X}\ \max_{Q(\cdot,\cdot)\in\mathcal{S}_{\mathscr{A},T}} P_{fp}\leq 2^{-\lambda n}\}. \tag{18}$$

2) Attacker's strategies:

$$\mathcal{S}_{\mathscr{A}} = \mathcal{S}_{\mathscr{A},T}\times\mathcal{S}_{\mathscr{A},O} \tag{19}$$

with $\mathcal{S}_{\mathscr{A},T}$ and $\mathcal{S}_{\mathscr{A},O}$ defined as in (17) and (14) respectively. It is worth observing that $\mathcal{S}_{\mathscr{D}}$ is always non empty, since it contains at least the degenerate strategy that always accepts H_0. For such a strategy, P_{fp} is identically equal to zero, hence it satisfies the constraint on the false positive error probability regardless of the values of λ and α. In fact, it is easy to realise that when $\alpha=1$ this is the only possible choice available to the defender. Such a strategy obviously results in a false negative error probability equal to 1, hence determining the *win* of the adversary.

Another observation regards the assumption that $\mathscr{D}$ knows α, that is the (maximum) percentage of training samples that $\mathscr{A}$ may corrupt. This is an implicit and necessary assumption in the definition of the game, since for a proper definition it is necessary that the players know the space of strategies of the other players. Assuming that the value of α is not known to the defender would require that we redefine the game as a game with incomplete information, namely a Bayesian game, possibly looking for Bayesian equilibria [23]. As a matter of fact, coping with attacks often implies making decisions under uncertainty; hence, the interaction between the defender and the attacker has already been modelled as a Bayesian game in other security-oriented works, e.g. [24], [25], [26], [27], for intrusion detection applications, or [28], for image forensics. In our problem, the Bayesian formulation of the game would dramatically complicate the analysis of the problem, so we decided to stick to a classical definition and interpret the value of α as a kind of worst case estimate that the defender has on the capability of $\mathscr{A}$ to corrupt the training data. As a matter of fact, in the Neyman-Pearson setup adopted in this paper (and in prior works), some estimate on the maximum percentage of samples corrupted by the attacker is necessary, since in the absence of such an estimate the constraint on the false positive error probability could not be satisfied, given that the possibility that all the training samples have been corrupted could not be ruled out.

3) Payoff: The payoff of the game is still equal to the false negative error probability:

$$u(\Lambda^{n\times m},(Q(\cdot,\cdot),\ S^n_{YZ}(\cdot,\cdot))) = -P_{fn}. \tag{20}$$

IV. Asymptotic equilibrium and payoff of the $SI^{a,t}_{c\text{-}tr}$ and $SI^{a}_{c\text{-}tr}$ games

In this section, we focus on the behavior of the game when the length of the test and training sequences tends to infinity; we first derive the equilibrium of the $SI^{a,t}_{c\text{-}tr}$ and the $SI^{a}_{c\text{-}tr}$ games and then evaluate the payoff at the equilibrium.

A. Optimum defender's strategy

We start by deriving the asymptotically optimum strategy for $\mathscr{D}$. As we will see, a dominant and universal strategy with respect to P_Y exists for $\mathscr{D}$. In other words, the optimum choice of $\mathscr{D}$ depends on neither the strategy chosen by the attacker nor P_Y. In addition, since the constraint on the false positive probability must be satisfied for all attackers' strategy, the optimum strategy for the defender is the same for both the targeted and non-targeted versions of the game.

As a first thing, we look for an explicit expression of the false positive error probability. Such a probability depends on P_X and on the strategy used by $\mathscr{A}$ to corrupt the training sequence. In fact, the mapping of y^n into z^n does not have any impact on $\mathscr{D}$'s decision under H_0. We carry out our derivations by focusing on the game with targeted corruption. It will be clear from our analysis that the dependence on y^n has no impact on P_{fp}, and hence the same results hold for the game with non-targeted corruption.

For a given P_X and $Q(\cdot,\cdot)$, P_{fp} is equal to the probability that Y generates a sequence y^n and X generates two sequences x^n and τ^{m_1}, such that the pair of type classes $(P_{x^n},\alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}})$ falls outside $\Lambda^{n\times m}$. Such a probability can be expressed as:

$$\begin{aligned} P_{fp} &= Pr\{(P_{x^n},\alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}})\in\bar{\Lambda}^{n\times m}\} \\ &= \sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot \\ &\qquad \sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} P_X(T(P_{x^n}))\cdot \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}: \\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}})), \end{aligned} \tag{21}$$

where $\bar{\Lambda}^{n\times m}$ is the complement of $\Lambda^{n\times m}$, and where we have exploited the fact that under H_0 the training sequence τ^{m_1} and the test sequence x^n are generated independently by X. Given the above formulation, the set of strategies available to $\mathscr{D}$ can be rewritten as:

$$\mathcal{S}_{\mathscr{D}}=\Bigg\{\Lambda^{n\times m}\colon \max_{P_X}\max_{Q(\cdot,\cdot)}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot \tag{22}$$

$$\sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} P_X(T(P_{x^n}))\cdot \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}}))\le 2^{-\lambda n}\Bigg\}.$$

We are now ready to prove the following lemma, which describes the asymptotically optimum strategy for the defender for both versions of the game.

Lemma 1. *Let $\Lambda^{n\times m,*}$ be defined as follows:*

$$\Lambda^{n\times m,*}=\left\{(P_{v^n},P_{t^m})\colon \min_{Q\in\mathcal{P}^{m_2}} h\left(P_{v^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\le\lambda-\delta_n\right\} \tag{23}$$

with

$$\delta_n=|\mathcal{X}|\frac{\log(n+1)((1-\alpha)nc+1)}{n}, \tag{24}$$

where $|\mathcal{X}|$ is the cardinality of the source alphabet, $c=\frac{m}{n}$, and where the minimisation over Q is limited to all the Q's such that $P_{t^m}-\alpha Q$ is nonnegative for all the symbols in $\mathcal{X}$. Then:

1) $\max_{P_X}\max_{s\in\mathcal{S}_{\mathscr{A}}} P_{fp}\le 2^{-n(\lambda-\nu_n)}$, *with* $\lim_{n\to\infty}\nu_n=0$,
2) $\forall\Lambda^{n\times m}\in\mathcal{S}_{\mathscr{D}}$, *we have* $\bar{\Lambda}^{n\times m}\subseteq\bar{\Lambda}^{n\times m,*}$.

where ν^n is an arbitrary sequence approaching 0 when n tends to infinity.

Proof. To prove the first part of the lemma, we see that from the expression of the false positive error probability given by (21), we can write:

$$\max_{P_X}\max_{Q(\cdot,\cdot)} P_{fp}\le \tag{25}$$

$$\max_{P_X}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot \sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))\cdot$$

$$\max_{Q(\cdot,\cdot)}\sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}})). \tag{26}$$

Let us consider the term within the inner summation. For each $P_{\tau^{m_1}}$ such that $\alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}$, we have[7]:

$$P_X(T(P_{\tau^{m_1}}))\le\max_{Q\in\mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right), \tag{27}$$

with the understanding that the maximisation is carried out only over the Q's such that $P_{t^m}-\alpha Q$ is nonnegative for all the symbols in $\mathcal{X}$.

Thanks to the above observation, we can upper bound the false positive error probability as follows:

$$\max_{P_X}\max_{Q(\cdot,\cdot)} P_{fp}\le \tag{28}$$

$$\max_{P_X}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot$$

$$\sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))\cdot|\mathcal{P}^{m_1}|\cdot\max_{Q\in\mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)$$

$$\overset{(a)}{=}\max_{P_X}\sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))|\mathcal{P}^{m_1}|\max_{Q\in\mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)$$

$$\le|\mathcal{P}^{m_1}|\sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}}\max_{Q\in\mathcal{P}^{m_2}}\max_{P_X} P_X(T(P_{x^n}))P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)$$

where in (a) we exploited the fact that the rest of the expression no longer depends on P_{y^n}. From this point, the proof goes along the same line of the proof of Lemma 2 in [3], by observing that $\max_{P_X} P_X(T(P_{x^n}))P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)$is upper bounded by $2^{-nh(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha})}$, and that for each pair of types in $\bar{\Lambda}^{n\times m,*}$, $h(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha})$ is larger than $\lambda-\delta_n$ for every Q by the very definition of $\Lambda^{n\times m,*}$.

We now pass to the second part of the lemma. Let $\Lambda^{n\times m}$ be a strategy in $\mathcal{S}_{\mathscr{D}}$, and let (P_{x^n},P_{t^m}) be a pair of types contained in $\bar{\Lambda}^{n\times m}$. Given that $\Lambda^{n\times m}$ is an admissible decision region (see (18)), the probability that X emits a test sequence belonging to $T(P_{x^n})$ and a training sequence τ^{m_1} such that after the attack $(\tau^{m_1}||\tau^{m_2})\in T(P_{t^m})$ must be lower than $2^{-\lambda n}$ for all P_X and all possible attacking strategies, that is:

$$2^{-\lambda n}>\max_{P_X}\max_{Q(\cdot,\cdot)}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot \tag{29}$$

$$\Big[P_X(T(P_{x^n}))\cdot\sum_{\substack{P_{\tau^{m_1}}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}}))\Big]$$

$$\overset{(a)}{=}\max_{P_X}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot$$

$$\Big[P_X(T(P_{x^n}))\cdot\max_{Q(\cdot,P_{y^n})}\sum_{\substack{P_{\tau^{m_1}}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}}))\Big]$$

$$\overset{(b)}{\ge}\max_{P_X}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(T(P_{y^n}))\cdot\Big[P_X(T(P_{x^n}))\cdot$$

$$\max_{Q(P_{\tau^{m_1}},P_{y^n})} P_X\left(T\left(\frac{P_{t^m}-\alpha Q(P_{\tau^{m_1}},P_{y^n})}{1-\alpha}\right)\right)\Big]$$

$$\overset{(c)}{=}\max_{P_X} P_X(T(P_{x^n}))\max_{Q\in\mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right),$$

where (a) is obtained by replacing the maximisation over all possible strategies $Q(\cdot,\cdot)$, with a maximisation over $Q(\cdot,P_{y^n})$ for each specific P_{y^n}, and (b) is obtained by considering only one term $P_{\tau^{m_1}}$ of the inner summation and optimising

[7] It is easy to see that the bound (27) holds also for the non-targeted game, when Q depends on the training sequence only ($Q(P_{\tau^{m_1}})$).

$Q(P_{\tau^{m_1}},P_{y^n})$ for that term. Finally, (c) follows by observing that the optimum $Q(\cdot,P_{y^n})$ is the same for any P_{y^n}. As usual, the maximization over Q in the last expression is restricted to the Q's for which $P_{t^m}-\alpha Q \geq 0$ for all the symbols in $\mathscr{X}$ [8]

By lower bounding the probability that a memoryless source X generates a sequence belonging to a certain type class (see [11], chapter 12), we can continue the above chain of inequalities as follows

$$2^{-\lambda n} > \frac{\max\limits_{P_X}\max\limits_{Q\in\mathcal{P}^{m_2}} 2^{-n\left[\mathcal{D}(P_{x^n}||P_X)+\frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}||P_X\right)\right]}}{(n{+}1)^{|\mathcal{X}|}(m_1{+}1)^{|\mathcal{X}|}} \tag{30}$$

$$\geq \frac{2^{-n\min\limits_{Q\in\mathcal{P}^{m_2}}\min\limits_{P_X}\left[\mathcal{D}(P_{x^n}||P_X)+\frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}||P_X\right)\right]}}{(n{+}1)^{|\mathcal{X}|}(m_1{+}1)^{|\mathcal{X}|}}$$

$$\overset{(a)}{=} \frac{2^{-n\min\limits_{Q\in\mathcal{P}^{m_2}} h\left(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)}}{(n{+}1)^{|\mathcal{X}|}(m_1{+}1)^{|\mathcal{X}|}},$$

where (a) derives from the minimisation properties of the generalised log-likelihood ratio function $h()$ (see Lemma 1, in [3]). By taking the log of both terms we have:

$$\min_{Q\in\mathcal{P}^{m_2}} h\left(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right) > \lambda{-}\delta_n, \tag{31}$$

thus completing the proof of the lemma. □

Lemma 1 shows that the strategy $\Lambda^{n\times m,*}$ is asymptotically admissible (point 1) and optimal (point 2), regardless of the attack. From a game-theoretic perspective, this means that such a strategy is a dominant strategy for $\mathscr{D}$ and implies that the game is dominance solvable [20]. Similarly, the optimum strategy is a semi-universal one, since it depends on P_X but it does not depend on P_Y. At first sight, the minimisation required by the optimum defender's strategy seems to be computationally prohibitive, however this is not the case since the minimisation can be carried out efficiently by exploiting the convexity of the h function. More specifically, since the minimisation is limited to the Q's such that $P_{t^m}-\alpha Q$ is nonnegative for all the symbols in $\mathscr{X}$, the log-sum inequality [11] can be applied to show that $h\left(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)$ is a convex function with respect to those Q, that is within the set $\{Q{\in}\mathcal{P}^{m_2}{:}\frac{P_{t^m}-\alpha Q}{1-\alpha}{\in}\mathcal{P}^{m_1}\}$. Being this set linear in Q and limited (corresponding to a subset of the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$), the optimisation problem in (23) is a convex mixed integer nonlinear problem, namely, *convex* MINLP [29], for which a global optimum solution exists. For this kind of problems, there are several efficient solvers yielding the optimum solution [30]. The number of optimisation variables, which determines the computational complexity, corresponds to the cardinality of the alphabet, i.e $|\mathcal{X}|$, and hence the minimisation is viable in many practical scenarios.

It is clear from the proof of Lemma 1 that the same optimum strategy holds for the targeted and non-targeted versions of the game. The situation is rather different with regard to the optimum strategy for the attacker. Despite the existence of a dominant strategy for the defender, in fact, the identification of the optimum attacker's strategy for the $SI^{a}_{c\text{-}tr}$ game is not easy due to the 2-step nature of the attack. For this reason, in the following sections, we will focus on the targeted version of the game, which is easier to study. We will then use the results obtained for the $SI^{a,t}_{c\text{-}tr}$ game to derive the best achievable performance for the case of non-targeted attack.

B. The $SI^{a,t}_{c\text{-}tr}$ game: optimum attacker's strategy and equilibrium point

Given the dominant strategy of $\mathscr{D}$, for any given τ^{m_1} and y^n, the optimum attacker's strategy for the $SI^{a,t}_{c\text{-}tr}$ game boils down to the following double minimisation:

$$(Q^*(P_{\tau^{m_1}},P_{y^n}),\ S^{n,*}_{YZ}(P_{y_n},P_{t^m})) = \tag{32}$$
$$\arg\min_{\substack{Q\in\mathcal{P}^{m_2}\\ S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})}} \left(\min_{Q'} h\left(P_{z^n},\frac{(1{-}\alpha)P_{\tau^{m_1}}{+}\alpha Q{-}\alpha Q'}{1{-}\alpha}\right)\right),$$

where P_{z^n} is obtained by applying the transformation map S^n_{YZ} to P_{y^n}, and where $P_{t^m}{=}(1{-}\alpha)P_{\tau^{m_1}}{+}\alpha Q$. As usual, the minimisation over Q' is limited to the Q' such that all the entries of the resulting pmf are nonnegative.

As a remark, for $L{=}0$ (corruption of the training sequence only), we get:

$$Q^*(P_{\tau^{m_1}},P_{y^n}) =$$
$$\arg\min_{Q\in\mathcal{P}^{m_2}}\left[\min_{Q'} h\left(P_{y^n},\ P_{\tau^{m_1}}{+}\frac{\alpha}{1{-}\alpha}(Q{-}Q')\right)\right], \tag{33}$$

while, for $\alpha{=}0$ (classical setup, without corruption of the training sequence) we have:

$$S^{n,*}_{YZ}(P_{y^n},P_{t^m}){=}\operatorname*{argmin}_{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})} h(P_{z^n},P_{t^m}), \tag{34}$$

falling back to the known case of source identification with uncorrupted training, already studied in [3]. Having determined the optimum strategies of both players, it is immediate to state the following:

Theorem 1. *The $SI^{a,t}_{c\text{-}tr}$ game is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile* $(\Lambda^{n\times m,*}{,}(Q^*(\cdot{,}\cdot),\ S^{n,*}_{YZ}(\cdot{,}\cdot))$ *given by equations* (23) *and* (32).

Proof. The theorem is a direct consequence of the fact that $\Lambda^{n\times m,*}$ is a dominant strategy for $\mathscr{D}$. □

C. The $SI^{a,t}_{c\text{-}tr}$ game: payoff at the equilibrium

In this section, we derive the asymptotic value of the payoff at the equilibrium, to see who and under which conditions is going to *win* the game.

To start with, we identify the set of pairs $(P_{y^n},P_{\tau^{m_1}})$ for which, as a consequence of $\mathscr{A}$'s action, $\mathscr{D}$ accepts H_0:

$$\begin{aligned}\Gamma^n(\lambda{,}\alpha{,}L) = \{(P_{y^n}{,}P_{\tau^{m_1}}){:}\ \exists\ (P_{z^n}{,}P_{t^m}){\in}\Lambda^{n\times m,*} \\ \text{s.t. } P_{t^m}{=}(1{-}\alpha)P_{\tau^{m_1}}{+}\alpha Q \text{ and } P_{z^n}{=}S^n_Z \\ \text{for some } Q{\in}\mathcal{P}^{m_2} \text{ and } S^n_{YZ}{\in}\mathcal{A}(L{,}P_{y^n})\}.\end{aligned} \tag{35}$$

[8] It is easy to see that the same lower bound can be derived also for the non targeted case, as the optimum Q in the second to last expression does not depend on P_{y^n}.

If we fix the type of the non-corrupted training sequence ($P_{\tau^{m_1}}$), we obtain:

$$\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L)=\{P_{y^n}\colon \exists\ P_{z^n}\in\Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}}+\alpha Q) \tag{36}$$

$$\text{s.t. } P_{z^n}=S^n_Z$$

$$\text{for some } Q\in\mathcal{P}^{m_2} \text{ and } S^n_{YZ}\in\mathcal{A}(L,P_{y^n})\},$$

where $\Lambda^{n,*}(P)$ denotes the acceptance region for a fixed type of the training sequence in $\mathcal{P}^m$. It is interesting to notice that, since in the current setting $\mathscr{A}$ has two degrees of freedom, the attack has a twofold effect: the sequence y^n is modified in order to bring it inside the acceptance region $\Lambda^{n,*}(P_{t^m})$ and the acceptance region itself is modified so to facilitate the former action.

To go on, we find it convenient to rewrite the set $\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L)$ as follows:

$$\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L) = \tag{37}$$

$$\{P_{y^n}\colon \exists S^n_{PV} \in \mathcal{A}(L,P_{y^n}) \text{ s.t. } S^n_V \in \Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha)\},$$

where

$$\Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha)= \tag{38}$$

$$\{P_{y^n}\colon \exists Q \in \mathcal{P}^{m_2} \text{ s.t. } P_{y^n}\in\Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}}+\alpha Q)\},$$

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 subscript 0 suggests, this set corresponds to the set in (36) when $\mathscr{A}$ cannot modify the sequence drawn from Y (i.e. L=0) and then tries to hamper the decision by corrupting the training sequence only.

By considering the expression of the acceptance region, the set $\Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha)$ can be expressed in a more explicit form as follows:

$$\Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha) = \big\{P_{y^n}\colon \exists Q,Q' \in \mathcal{P}^{m_2} \text{ s.t.} \tag{39}$$

$$h\bigg(P_{y^n},P_{\tau^{m_1}}+\frac{\alpha}{(1-\alpha)}(Q-Q')\bigg) \leq \lambda-\delta_n\big\},$$

where the second argument of $h()$ denotes a type in $\mathcal{P}^{m_1}$ obtained from the original training sequence τ^{m_1} by first adding m_2 samples and later removing (in a possibly different way) the same number of samples. Note that in this formulation Q accounts for the fake samples introduced by the attacker and Q' for the worst case *guess* made by the defender of the position of the corrupted samples. We also observe that since we are treating the $SI^{a,t}_{c\text{-}tr}$ game, in general Q will depend on P_{y^n}. As usual, we implicitly assume that Q and Q' are chosen in such a way that $P_{\tau^{m_1}}+\frac{\alpha}{(1-\alpha)}(Q-Q')$ is nonnegative and smaller than or equal to 1 for all the alphabet symbols.

We are now ready to derive the asymptotic payoff of the game by following a path similar to that used in [2], [3]. First of all we generalise the definition of the sets $\Lambda^{n\times m,*}$, Γ^n and Γ^n_0 so that they can be evaluated for a generic pmf in $\mathcal{P}$ (that is, without requiring that the pmf's are induced by sequences of finite length). This step passes through the generalization of the h function. Specifically, given any pair of pmf's $(P,P')\in\mathcal{P}\times\mathcal{P}$, we define:

$$h_c(P,P') = \mathcal{D}(P||U) + c\mathcal{D}(P'||U); \tag{40}$$

$$U = \frac{1}{1+c}P + \frac{c}{1+c}P',$$

where $c\in[0,1]$. Note that when $(P,P')\in\mathcal{P}^n\times\mathcal{P}^n$, $h_c(P,P')= h(P,P')$. The asymptotic version of $\Lambda^{n\times m,*}$ is:

$$\Lambda^*=\bigg\{(P,R) : \min_Q h_c\bigg(P,\ \frac{R-\alpha Q}{1-\alpha}\bigg) \leq \lambda\bigg\}. \tag{41}$$

In a similar way, we can derive the asymptotic versions of Γ^n and Γ^n_0 in (37) and (38)-(39). To do so, we first observe that, the transportation map S^n_{YZ} depends on the sources only through the pmfs. By denoting with S^n_{PV} a transportation map from a pmf $P\in\mathcal{P}^n$ to another pmf $V\in\mathcal{P}^n$ and rewriting the set Γ^n accordingly, we can easily derive the asymptotic version of the set as follows:

$$\Gamma(R,\lambda,\alpha,L) = \{P\in\mathcal{P}\colon \exists S_{PV}\in\mathcal{A}(L,P) \text{ s.t. } V\in\Gamma_0(R,\lambda,\alpha)\}, \tag{42}$$

with

$$\Gamma_0(R,\lambda,\alpha) = \tag{43}$$

$$\{P\in\mathcal{P}\colon \exists Q\in\mathcal{P} \text{ s.t. } P\in\Lambda^*((1-\alpha)R+\alpha Q)\} =$$

$$\bigg\{P\in\mathcal{P}\colon \exists Q,Q'\in\mathcal{P} \text{ s.t. } h_c\bigg(P,\ R+\frac{\alpha}{(1-\alpha)}(Q-Q')\bigg) \leq \lambda\bigg\},$$

where the definitions of S_{PV} and $\mathcal{A}(L,P)$ derive from those of S^n_{PV} and $\mathcal{A}^n(L,P)$ by relaxing the requirement that the terms $S_{PV}(i,j)$ and $P(i)$ are rational numbers with denominator n. We now have all the necessary tools to prove the following theorem.

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

$$\varepsilon = \min_R[(1-\alpha)c\mathcal{D}(R||P_X)+\min_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)]. \tag{44}$$

Accordingly,

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

Proof. The theorem could be proven going along the same lines of the proof of Theorem 4 in [3]. We instead provide a proof based on the extension of Sanov's theorem provided in the Appendix (see Theorem 6). In fact, Theorem 2, as well as Theorem 4 in [3], can be seen as an application of such a generalized version of Sanov's theorem.

Let us consider

$$\begin{aligned} P_{fn} &= \sum_{(P_{y^n},P_{\tau^{m_1}})\in\Gamma^n(\lambda,\alpha,L)} P_X(T(P_{\tau^{m_1}}))P_Y(T(P_{y^n})) \\ &= \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R)) \sum_{P\in\Gamma^n(R,\lambda,\alpha,L)} P_Y(T(P)) \\ &= \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R))P_Y(\Gamma^n(R,\lambda,\alpha,L)). \end{aligned} \tag{45}$$

We start by deriving an upper-bound of the false negative error probability. We can write:

$$\begin{aligned} P_{fn} &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R)) \sum_{P\in\Gamma^n(R,\lambda,\alpha,L)} 2^{-n\mathcal{D}(P||P_Y)} \\ &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R))(n{+}1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma^n(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)} \\ &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R))(n{+}1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)} \\ &\leq (n{+}1)^{|\mathcal{X}|}(m_1{+}1)^{|\mathcal{X}|} \\ &\quad \cdot 2^{-n \min\limits_{R\in\mathcal{P}^{m_1}} [\frac{m_1}{n}\mathcal{D}(R||P_X) + \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)]} \\ &\leq (n{+}1)^{|\mathcal{X}|}(m_1{+}1)^{|\mathcal{X}|} \\ &\quad \cdot 2^{-n \min\limits_{R\in\mathcal{P}} [(1-\alpha)c\mathcal{D}(R||P_X) + \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)]}, \end{aligned} \tag{46}$$

where the use of the minimum instead of the infimum is justified by the fact that $\Gamma^n(R,\lambda,\alpha,L)$ and $\Gamma(R,\lambda,\alpha,L)$ are compact sets. By taking the log and dividing by n we find:

$$\begin{aligned} &-\frac{\log P_{fn}}{n} \geq \\ &\quad \min_{R\in\mathcal{P}} \big[(1-\alpha)c\mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)\big] - \beta_n, \end{aligned} \tag{47}$$

where $\beta_n = |\mathcal{X}| \frac{\log(n+1)((1-\alpha)nc+1)}{n}$ tends to 0 when n tends to infinity.

We now turn to the analysis of a lower bound for P_{fn}. Let R^* be the pmf achieving the minimum in the outer minimisation of (44). Due to the density of rational numbers within real numbers, we can find a sequence of pmfs' $R_{m_1} \in \mathcal{P}^{m_1}$ ($m_1 = (1-\alpha)nc$) that tends to R^* when n (and hence m_1) tends to infinity. We can write:

$$\begin{aligned} P_{fn} &= \sum_{R\in\mathcal{P}^{m_1}} P_X(T(R)) P_Y(\Gamma^n(R,\lambda,\alpha,L)) \\ &\geq P_X(T(R_{m_1})) P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)), \\ &\geq \frac{2^{-m_1\mathcal{D}(R_{m_1}||P_X)}}{(m_1{+}1)^{|\mathcal{X}|}} P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)), \end{aligned} \tag{48}$$

where in the first inequality we have replaced the sum with the single element of the subsequence R_{m_1} defined previously, and where the second inequality derives from the well known lower bound on the probability of a type class [11]. From (48), by taking the log and dividing by n, we obtain:

$$\begin{aligned} &-\frac{\log P_{fn}}{n} \leq \\ &\quad (1-\alpha)c\mathcal{D}(R_{m_1}||P_X) - \frac{1}{n}\log P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)) + \beta'_n, \end{aligned} \tag{49}$$

where $\beta'_n = |\mathcal{X}| \frac{\log(m_1+1)}{n}$ tends to 0 when n tends to infinity. In order to compute the probability $P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L))$, we resort to Corollary 1 of the the generalised version of Sanov's Theorem given in Appendix A.

To apply the corollary, we must show that $\Gamma^n(R_{m_1},\lambda,\alpha,L) \xrightarrow{H} \Gamma(R^*,\lambda,\alpha,L)$.

First of all, we observe that by exploiting the continuity of the h_c function and the density of rational numbers into the real ones, it is easy to prove that $\Gamma_0^n(R_{m_1},\lambda,\alpha) \xrightarrow{H} \Gamma_0(R^*,\lambda,\alpha)$. Then the Hausdorff convergence of $\Gamma^n(R_{m_1},\lambda,\alpha,L)$ to $\Gamma(R^*,\lambda,\alpha,L)$ follows from the regularity properties of the set of transportation maps stated in Appendix B. To see how, we observe that any transformation $S_{PV} \in \mathcal{A}(L,P)$ mapping P into V can be applied in inverse order through the transformation $S_{VP}(i,j) = S_{PV}(j,i)$. It is also immediate to see that S_{VP} introduces the same distortion introduced by S_{PV}, that is $S_{VP} \in \mathcal{A}(L,V)$. Let now P be a point in $\Gamma(R^*,\lambda,\alpha,L)$. By definition we can find a map $S_{PV} \in \mathcal{A}(L,P)$ such that $V \in \Gamma_0(R^*,\lambda,\alpha)$. Since $\Gamma_0^n(R_{m_1},\lambda,\alpha) \xrightarrow{H} \Gamma_0(R^*,\lambda,\alpha)$, for large enough n, we can find a point $V' \in \Gamma_0^n(R_{m_1},\lambda,\alpha)$ which is arbitrarily close to V. Thanks to the second part of Theorem 7 in Appendix B, we know that a map $S_{V'P'} \in \mathcal{A}^n(L,V')$ exists such that P' is arbitrarily close to P and $P' \in \mathcal{P}^n$. By applying the inverse map $S_{P'V'}$ to P', we see that $P' \in \Gamma^n(R_{m_1},\lambda,\alpha,L)$, thus permitting us to conclude that, when n increases, $\delta_{\Gamma(R^*,\lambda,\alpha,L)}(\Gamma^n(R_{m_1},\lambda,\alpha,L)) \rightarrow 0$. In a similar way, we can prove that $\delta_{\Gamma^n(R_{m_1},\lambda,\alpha,L)}(\Gamma(R^*,\lambda,\alpha,L)) \rightarrow 0$, hence permitting us to conclude that $\Gamma^n(R_{m_1},\lambda,\alpha,L) \xrightarrow{H} \Gamma(R^*,\lambda,\alpha,L)$.

We can now apply the generalised version of Sanov Theorem as expressed in Corollary 1 of Appendix A to conclude that:

$$-\lim_{n\to\infty} \frac{1}{n} \log P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)) = \min_{P\in\Gamma(R^*,\lambda,\alpha,L)} \mathcal{D}(P||P_Y). \tag{50}$$

Going back to (49), and by exploiting the continuity of the divergence function, we can say that for large n we have:

$$-\frac{\log P_{fn}}{n} \leq (1-\alpha)c\mathcal{D}(R^*||P_X) + \min_{P\in\Gamma(R^*,\lambda,\alpha,L)} \mathcal{D}(P||P_Y) + \nu_n, \tag{51}$$

where the sequence ν_n tends to zero when n tends to infinity. By coupling equations (47) and (51) and by letting $n \rightarrow \infty$, we eventually obtain:

$$\begin{aligned} &-\lim_{n\to\infty} \frac{\log P_{fn}}{n} = \\ &\quad \min_{R} [(1-\alpha)c \cdot \mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)], \end{aligned} \tag{52}$$

thus proving the theorem.

□

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 $\mathscr{A}$ induces $\mathscr{D}$ to decide in favour of H_0 even if H_1 holds.

D. Analysis of the $SI^a_{c\text{-}tr}$ game

We now focus on the $SI^a_{c\text{-}tr}$ game. For a given choice of $Q(P_{\tau^{m_1}}) \in \mathcal{S}_{\mathscr{A},T}$ (and hence t^m), given a sequence y^n, the optimum choice of the second part of the attack derives quite easily from the definition of $\Lambda^{n\times m,*}$, namely

$$\begin{aligned} &S^{n,*}_{YZ}(P_{y^n},P_{t^m}) = \\ &\arg \min_{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})} \left(\min_{Q\in\mathcal{P}^{m_2}} h\left(P_{z^n}, \frac{P_{t^m} - \alpha Q}{1-\alpha}\right)\right). \end{aligned} \tag{53}$$

Now the point is to determine the strategy $Q(P_{\tau^{m_1}})$ which maximises the probability that the attack in (53) succeeds. To this purpose, of course, the attacker must exploit the knowledge of P_Y. Since solving such a maximisation problem is not an easy task, we will proceed in a different way. We first introduce a simple (and possibly suboptimum) strategy, then we argue that such a strategy is asymptotically optimum, in that the set of the sources that cannot be distinguished from X with this choice is the same set that we have obtained for the $SI^{a,t}_{c\text{-}tr}$ setup, which is known to be more favourable to the attacker. More specifically, we consider the following two-part attack. In the first part, $\mathscr{A}$ does not know y^n, hence he trusts the law of large numbers and optimises $Q(P_{\tau^{m_1}})$ by using P_Y as a proxy for P_{y^n}. To do so, he applies (32), by replacing P_{y^n} with P_Y. Specifically, by indicating with $Q^\dagger$, the resulting strategy for the first part of the attack, we have

$$Q^\dagger(P_{\tau^{m_1}}) = \arg\min_{Q\in\mathcal{P}^{m_2}} \tag{54}$$

$$\min_{\substack{Q'\in\mathcal{P}^{m_2}\\ S_{YZ}\in\mathcal{A}(L,P_Y)}} h_c\left(P_Z,P_{\tau^{m_1}}+\frac{\alpha}{1-\alpha}(Q-Q')\right). \tag{55}$$

As a by-product of the above minimisation, the attacker also finds the map $S^{n,\dagger}_{YZ}$ representing the optimum attack when $P_{y^n}{=}P_Y$. Let us indicate the result of the application of such a map to P_Y by $P^\dagger_Z$.

In the second part of the attack, $\mathscr{A}$ tries to move P_{y^n} as close as possible to $P^\dagger_Z$, that is:

$$S^{n,\dagger}_{YZ}(P_{y^n},P^\dagger_{t^m}) = \arg\min_{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})} d(S^n_Z,P^\dagger_Z), \tag{56}$$

where $S^{n,\dagger}_{YZ}(P_{y^n},P^\dagger_{t^m})$ depends upon the corrupted training sequence obtained after the application of the first part of the attack, namely $P^\dagger_{t^m}{=}(1{-}\alpha)P_{\tau^{m_1}}{+}\alpha Q^\dagger(P_{\tau^{m_1}})$, through $P^\dagger_Z$.

The asymptotic optimality of the strategy $(Q^\dagger(P_{\tau^{m_1}})$, $S^{n,\dagger}_{YZ}(P_{y^n},P^\dagger_{t^m}))$ derives from the following theorem

Theorem 3 (Indistinguishability region of the $SI^a_{c\text{-}tr}$ game)**.** *The indistinguishability region of $SI^a_{c\text{-}tr}$ game is equal to that of the $SI^{a,t}_{c\text{-}tr}$ game (see* (42)*) and is asymptotically achieved by the attacking strategy $(Q^\dagger(P_{\tau^{m_1}})$, $S^{n,\dagger}_{YZ}(P_{y^n},P^\dagger_{t^m}))$.*

Proof (sketch). The theorem derives from the observation that due to the law of large numbers, when n grows, P_{y^n} tends to P_Y; hence, for large enough n, optimising the first part of the attack by replacing P_{y^n} with P_Y does not introduce a significant performance loss. The rigorous proof goes along similar lines to those used to prove Theorem 2 and ultimately relies on the continuity of the h_c function and the regularity properties of the set $\mathcal{A}^n(L,P_{y^n})$. The details of the proof are omitted for sake of brevity. □

Given the asymptotic equivalence of the $SI^a_{c\text{-}tr}$ and the $SI^{a,t}_{c\text{-}tr}$ games, in the rest of the paper, we will generally refer to the $SI^a_{c\text{-}tr}$ game without specifying if we are considering the targeted or non-targeted case.

V. Source distinguishability for the $SI^a_{c\text{-}tr}$ game

In this section, we study the behaviour of the $SI^a_{c\text{-}tr}$ game when we vary the decay rate of the false positive error probability λ. It is clear, in fact, that D can improve his payoff at the equilibrium (linked to the false negative error probability), by relaxing the constraint on the false positive error exponent, i.e. by decreasing λ. By letting λ tend to zero, then, we can derive the best achievable performance of the defender when we require only that P_{fp} tends to zero exponentially fast with an arbitrarily low - yet strictly positive - error exponent. This corresponds to extending the Chernoff-Stein lemma [11] to the adversarial setup considered in this paper. Eventually, we use such a result to derive the conditions under which the reliable distinction between two sources is possible in terms of the number of corrupted training samples α and maximum allowed distortion L.

A. Ultimate achievable performance of the game

As we said, the goal of this section is to study the limit of the indistinguishability region when $\lambda{\to}0$. This limit, in fact, determines 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).

We start by exploiting optimal transport theory to rewrite the indistinguishability region as:

$$\Gamma(P_X,\lambda,\alpha,L) = \{P\colon \exists V{\in}\Gamma_0(P_X,\lambda,\alpha) \text{ s.t. } \mathit{EMD}(P,V){\le}L\}, \tag{57}$$

where *EMD* (Earth Mover Distance) is the term used in computer vision to denote the minimum transportation cost [22], [31], that is

$$\mathit{EMD}(P,V) = \min_{S_{PV}:S_P=P,S_V=V} \sum_{i,j} S_{PV}(i,j)d(i,j). \tag{58}$$

With this definition, the main result of this section is stated by the following theorem.

Theorem 4. *Given two sources X and 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 $SI^a_{c\text{-}tr}$ game is:*

$$\lim_{\lambda\to 0}\lim_{n\to\infty} -\frac{1}{n}\log P_{fn} = \min_R[(1{-}\alpha)c\mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\alpha,L)}\mathcal{D}(P||P_Y)], \tag{59}$$

where $\Gamma(R,\alpha,L){=}\Gamma(R,\lambda{=}0,\alpha,L)$. Accordingly, the ultimate indistinguishability region is given by:

$$\Gamma(P_X,\alpha,L){=}\{P\colon \exists V{\in}\Gamma_0(P_X,\alpha) \textit{ s.t. } \mathrm{EMD}(P,V){\le}L\}, \tag{60}$$

where $\Gamma_0(P_X,\alpha){=}\Gamma_0(P_X,\lambda{=}0,\alpha)$. Moreover, $\Gamma(P_X,\alpha,L)$ can be rewritten as:

$$\begin{aligned}\Gamma(P_X,\alpha,L){=}&\left\{P\colon \min_{V:\mathrm{EMD}(P,V)\le L}\sum_i[V(i){-}P_X(i)]^+{\le}\frac{\alpha}{(1{-}\alpha)}\right\}\\ {=}&\left\{P\colon \min_{V:\mathrm{EMD}(P,V)\le L} d_{L_1}(V,P_X) \le \frac{2\alpha}{(1{-}\alpha)}\right\}.\end{aligned} \tag{61}$$

with $[a]^+{=}\max\{a,0\}$.

Proof. The proof of the first part goes along the same steps used in the proof of Theorems 3 and 4 in [5] and is not repeated here. We show, instead, that $\Gamma(P_X,\alpha,L)$ can be rewritten as in (61).

By observing that $h_c(P,Q)=0$ if and only if $P=Q$, it is immediate to see that the set $\Gamma_0(P_X,\lambda=0,\alpha)$ takes the following expression:

$$\Gamma_0(P_X,\alpha)=\{P\colon \exists Q,Q'\in\mathcal{P} \text{ s.t. } P = P_X+\frac{\alpha}{(1-\alpha)}(Q-Q')\}. \tag{62}$$

Expression (62) can be rewritten by avoiding the introduction of the auxiliary pmf's Q and Q'. To do so, we observe that $Q(i)$ must be larger than $Q'(i)$ for all the bins i for which $P(i)>P_X(i)$ (and viceversa). In addition, Q and Q' must be valid pmf's, hence we have $\sum_i[Q(i)-Q'(i)]^+=\sum_i[Q'(i)-Q(i)]^+\leq 1$. Then, it is easy to see that (62) is equivalent to the following definition:

$$\begin{aligned}\Gamma_0(P_X,\alpha)&=\left\{P\colon \sum_i[P(i)-P_X(i)]^+\leq\frac{\alpha}{(1-\alpha)}\right\} \qquad (63)\\ &=\left\{P\colon d_{L_1}(P,P_X)\leq\frac{2\alpha}{(1-\alpha)}\right\},\end{aligned}$$

where the second equality follows by observing that $d_{L_1}(P,P_X)=\sum_i[P(i)-P_X(i)]^+ +\sum_i[P_X(i)-P(i)]^+$. Eventually, (61) derives immediately from the expression of $\Gamma_0(P_X,\alpha)$ given in (63). □

According to Theorem 4, $\Gamma(P_X,\alpha,L)$ provides the *ultimate indistinguishability region* of the test, that is the set of all the pmf's for which $\mathscr{A}$ wins the game.

Before going on, we pose to discuss the geometrical meaning of the set $\Gamma_0(P_X,\alpha)$ in (62). To do so, we introduce the set Λ_0^*, obtained from Λ^* by letting $\lambda\to\infty$:

$$\Lambda_0^*=\left\{(P,P')\colon \exists Q \text{ s.t. } P' = \frac{P-\alpha Q}{(1-\alpha)}\right\}. \tag{64}$$

As usual, we can fix the pmf P and define:

$$\Lambda_0^*(P)=\left\{P'\colon \exists Q \text{ s.t. } P' = \frac{P-\alpha Q}{(1-\alpha)}\right\}. \tag{65}$$

By referring to Figure 3 (left part), we can geometrically interpret $\Lambda_0^*(P)$ as the set of the pmf's P' such that P is a convex combination (with coefficient α) of P' with a point Q of the probability simplex. Starting from (43), we can then rewrite $\Gamma_0(P_X,\alpha)$ as follows:

$$\Gamma_0(P_X,\alpha)=\{P\colon \exists Q\in\mathcal{P} \text{ s.t. } P\in\Lambda_0^*((1-\alpha)P_X+\alpha Q)\}. \tag{66}$$

Accordingly, $\Gamma_0(P_X,\alpha)$ is geometrically obtained as the union of the acceptance regions built from the points which can be written as a convex combination of P_X with some point Q in the simplex. As shown in the right part of Figure 3, such a region corresponds to a hexagon centred in P_X, which, in the probability simplex, is equivalent to the set of points whose L_1 distance from P_X is smaller than or equal to $2\alpha/(1-\alpha)$ (as stated in (63)). Of course, only the points of the hexagon that lie inside the simplex are valid pmf's and then must be accounted for.

A pictorial representation of the set $\Gamma(P_X,\alpha,L)$ is given in Figure 4.

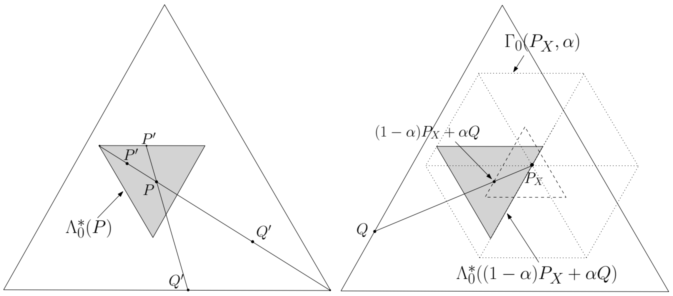

Fig. 3. Geometrical interpretation of $\Lambda_0^*(P)$ (left) and geometrical construction of $\Gamma_0(P_X,\alpha)$ (right). The size of the sets are exaggerated for graphical purposes.

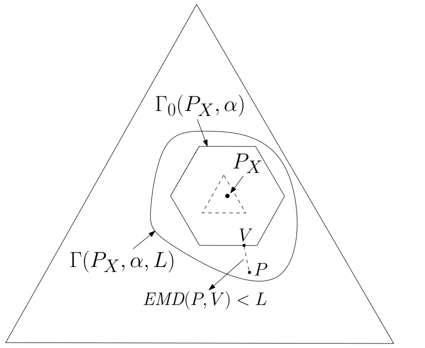

Fig. 4. Geometrical interpretation of $\Gamma(P_X,\alpha,L)$ as stated in Theorem 4.

B. Security margin and blinding corruption level (α_b)

By a closer inspection of the *ultimate indistinguishability region* $\Gamma(P_X,\alpha,L)$, we can derive some interesting parameters characterising the distinguishability of two sources in adversarial setting. 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 $\Gamma_0(P_X,\alpha)$. Then we conclude that $\mathscr{D}$ can tell the two sources apart if $d_{L_1}(P_Y,P_X)>\frac{2\alpha}{(1-\alpha)}$. On the contrary, if $d_{L_1}(P_Y,P_X)\leq\frac{2\alpha}{(1-\alpha)}$, $\mathscr{A}$ is able to make the sources indistinguishable by corrupting the training sequence. Clearly, the larger the α the easier is for $\mathscr{A}$to win the game. We can define the *blinding corruption level* α_b, as the minimum value of α for which two sources X and Y can not be distinguished. Specifically, we have:

$$\alpha_b(P_X,P_Y) = \frac{d_{L_1}(P_Y,P_X)}{2+d_{L_1}(P_Y,P_X)} = \frac{\sum_i[P_Y(i)-P_X(i)]^+}{1+\sum_i[P_Y(i)-P_X(i)]^+}. \tag{67}$$

From (67) it is easy to see that α_b is always lower than $1/2$, with the limit case $\alpha_b=1/2$ corresponding to a situation in which P_X and P_Y have completely disjoint supports[9]. It is interesting to notice that α_b is symmetric with respect to the two sources. Since the attacker is allowed only to add samples to the training sequence without removing existing samples, this might seem a counterintuitive result. Actually, the symmetry of α_b is a consequence of the worst case approach adopted by the defender. In fact, $\mathscr{D}$ himself discards

[9] We remind that for any pair of pmf's (P,Q), $d_{L_1}(P,Q)\leq 2$.

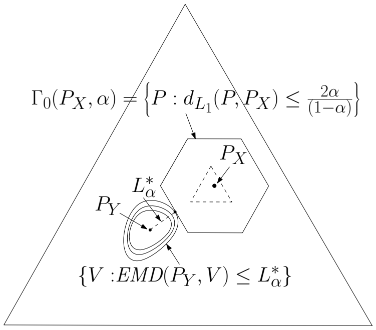
$\Gamma_0(P_X,\alpha) = \left\{P : d_{L_1}(P,P_X) \leq \frac{2\alpha}{(1-\alpha)}\right\}$
P_X
L^*_α
P_Y
$\{V : EMD(P_Y,V) \leq L^*_\alpha\}$

Fig. 5. Geometrical interpretation of the Security Margin between two sources X and Y.

a subset of samples from the training sequence in such a way to maximise the probability that the remaining part of the training sequence and the test sequence have been drawn from the same source.

Let us now consider the more general case in which $L \neq 0$. For a given $\alpha < \alpha_b$, we look for the maximum distortion allowed to $\mathscr{A}$ for which it is possible to reliably distinguish between the two sources. From equation (61), we see that the attack does not succeed if:

$$\min_{V:EMD(P_Y,V)\leq L} d_{L_1}(V,P_X) > \frac{2\alpha}{(1-\alpha)}. \tag{68}$$

This leads to the following definition, which extends the concept of security margin, introduced in [5], to the more general setup considered in this paper.

Definition 3 (Security Margin in the $SI^a_{c\text{-}tr}$ setup)**.** *Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum distortion allowed to the attacker for which the two sources can be reliably distinguished in the $SI^a_{c\text{-}tr}$ setup with a fraction α of possibly corrupted samples, is called Security Margin and is given by*

$$\mathcal{SM}_\alpha(P_X,P_Y) = L^*_\alpha, \tag{69}$$

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

$$\min_{V:\mathrm{EMD}(P_Y,V)\leq L^*_\alpha} d_{L_1}(V,P_X) = \frac{2\alpha}{(1-\alpha)}. \tag{70}$$

A geometric interpretation of L^*_α is given in Figure 5. By focusing on the case $P_Y \notin \Gamma_0(P_X,\alpha)$, and by observing that

$$\min_{V:EMD(P_Y,V)\leq L} d_{L_1}(V,P_X) \tag{71}$$

is a monotonic non-increasing function of L, the security margin can be expressed in explicit form as

$$\mathcal{SM}_\alpha(P_X,P_Y) = \operatorname*{argmin}_{L'} \min_{V:EMD(P_Y,V)\leq L'} \left| d_{L_1}(V,P_X) - \frac{2\alpha}{(1-\alpha)} \right|. \tag{72}$$

When $L > \mathcal{SM}_\alpha(P_X,P_Y)$, it is not possible for $\mathscr{D}$ 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, whenever the fraction of corrupted samples reaches the critical value, the sources can not be distinguished even if the attacker does not introduce any distortion. On the contrary, setting $\alpha = 0$ corresponds to studying the distinguishability of the sources with uncorrupted training; in this case we have $\mathcal{SM}_0(P_X,P_Y) = \mathit{EMD}(P_X,P_Y)$, in agreement with [5]. With reference to Figure 5, it is easy to see that when $\alpha = 0$ the hexagon representing $\Gamma_0(P_X,\alpha)$ collapses into the single point P_X and the security margin corresponds to the Earth Mover Distance between Y and X. Eventually, we notice that, for $\alpha > 0$, the value of the security margin in (72) is less than $\mathit{EMD}(P_X,P_Y)$. This is also an expected behaviour since the general setting considered in this paper is more favourable to the attacker than the setting in [5].

By looking at (72), we can argue that the Security Margin is symmetric with respect to the two sources X and Y, that is, $\mathcal{SM}_\alpha(P_Y,P_X) = \mathcal{SM}_\alpha(P_X,P_Y)$.

To show that this is the case, we observe that the pmf V' associated with the minimum L, for which we have $\mathit{EMD}(P_Y,V') = \mathcal{SM}_\alpha(P_X,P_Y)$, can be obtained through the application of a map $S_{P_Y V}$ that works as follows: it does not modify a portion $\alpha/(1-\alpha)$ of P_Y and moves the remaining mass into an equal amount of P_X in a convenient way (i.e., in such a way to minimise the overall distance between the masses). The inverse map can be applied to bring the same quantity of mass from P_X to P_Y, while leaving as is the remaining mass, thus obtaining a V'' which satisfies $\mathit{EMD}(P_X,V'') = \mathit{EMD}(P_Y,V')$ (because of the symmetry of the per-symbol distortion d) and $d_{L_1}(V'',P_Y) = d_{L_1}(V',P_X) = 2\alpha/(1-\alpha)$. Arguably, V'' is the pmf for which $\mathit{EMD}(P_X,V'') = \mathcal{SM}_\alpha(P_Y,P_X)$; hence, $\mathcal{SM}_\alpha(P_Y,P_X) = \mathcal{SM}_\alpha(P_X,P_Y)$.

1) Bernoulli sources: In order to get some insights on the practical meaning of 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 fraction of samples that $\mathscr{A}$ must add to induce a decision error is, according to (67), $\alpha_b = \frac{|p-q|}{1+|p-q|}$. For instance, and rather obviously, when $|p-q| = 1$, to win the game $\mathscr{A}$ must introduce a number of fake samples equal to the number of samples of the correct training sequence, i.e. $\alpha = 0.5$. With regard to $\mathcal{SM}$, we have:

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p| - \frac{\alpha}{1-\alpha} & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b \end{cases}. \tag{73}$$

Figure 6 illustrates the behavior of $\mathcal{SM}_\alpha(p,q)$ as a function of α when $p = 0.3$ and $q = 0.7$. The blinding corruption value is $\alpha_b = 0.286$.

VI. Source identification game with replacement of training samples

In this section, we study a variant of the game with corrupted training, in which $\mathscr{A}$ observes the training sequence and can replace a selected fraction of samples. Let τ^m indicate the original m-sample long training sequence drawn from X

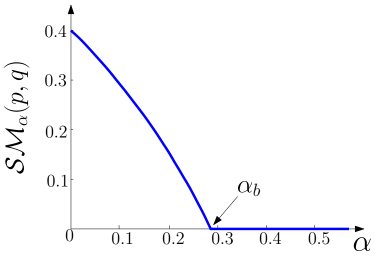

Fig. 6. Security margin as a function of α for Bernoulli sources with parameters p=0.3 and q=0.7 (α_b=0.286).

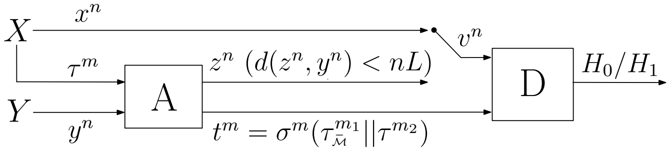
X
x^n
τ^m
Y
y^n
A
z^n $(d(z^n, y^n) < nL)$
$t^m = \sigma^m(\tau_{\bar{\mathcal{M}}}^{m_1}||\tau^{m_2})$
v^n
D
H_0/H_1

Fig. 7. Block diagram of the $SI^r_{c\text{-}tr}$ game (targeted corruption). Given the original training sequence τ^m, the adversary has the possibility to replace a selected subset of m_2 training samples with fake ones.

and let $\mathcal{M}$ be a subset of $m_2 = \alpha m$ indexes in $[1,2...m]$. The attacker can choose the index set $\mathcal{M}$ and replace the corresponding samples with m_2 fake samples. More formally, given the original training sequence τ^m, the training sequence *seen* by the defender is t^m=$\sigma(\tau_{\bar{\mathcal{M}}}^{m_1}||\tau^{m_2})$, where $\bar{\mathcal{M}}$ is the complement of $\mathcal{M}$ in $[1,2...m]$, $\tau_{\bar{\mathcal{M}}}^{m_1}$ is the set of original (non-attacked) samples, and τ^{m_2} is the sequence with the fake samples introduced by the attacker.

Figure 7 illustrates the adversarial setup considered in this section for the case of a targeted attack. Arguably, this scenario is more favourable to the attacker with respect to the $SI^a_{c\text{-}tr}$ game.

A. Formal definition of the $SI^r_{c\text{-}tr}$ game

In the sequel, we formally define the source identification game with replacement of selected samples, namely the $SI^r_{c\text{-}tr}$ game. As anticipated, we focus on a version of the game in which the corruption of the training samples depends on the to-be-attacked sequence y^n (targeted attack), the extension to the case of non-target attack, in fact, can be easily obtained by following the same approach used in Section IV-D.

1) Defender's strategies: As in the $SI^a_{c\text{-}tr}$ game, in order to be sure that the false positive error probability is lower than $2^{-n\lambda}$, the defender adopts a worst case strategy and considers the maximum of the false positive error probability over all the possible P_X and over all the possible attacks that the training sequence may have undergone, yielding:

$$\mathcal{S}_D=\{\Lambda^{n\times m}\subset\mathcal{P}^n\times\mathcal{P}^m\colon \max_{P_X\in\mathcal{P}} \max_{s\in\mathcal{S}_{\mathscr{A},T}} P_{fp} \leq 2^{-\lambda n}\}. \quad (74)$$

While the above expression is formally equal to that of the $SI^a_{c\text{-}tr}$ game (see (15)), the maximisation over $\mathcal{S}_{\mathscr{A},T}$ is now more cumbersome, due to the additional degree of freedom available to the attacker, who can selectively remove the samples of the original training sequence. In fact, even if $\mathscr{D}$ knew the position of the corrupted samples, simply throwing them away would not guarantee that the remaining part of the sequence would follow the same statistics of X, since the attacker might have deliberately altered them by selectively choosing the samples to replace.

2) Attacker's strategies: With regard to the attacker, the part of the attack working on the test sequence y^n is the same as for the $SI^a_{c\text{-}tr}$ case, while the part regarding the corruption of the training sequence must be redefined. To this purpose, we observe that the corrupted training sequence may be any sequence t^m for which $d_H(t^m,\tau^m)\leq\alpha m$, where d_H denotes the Hamming distance. Given that the defender bases his decision on the type of t^m, it is convenient to rewrite the constraint on the Hamming distance between sequences as a constraint on the L_1 distance between the corresponding types. In fact, by looking at the empirical distribution of the corrupted sequence, searching for a sequence t^m s.t. $d_H(t^m,\tau^m)\leq\alpha m$ is equivalent to searching for a pmf $P_{t^m}\in\mathcal{P}^m$ for which $d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha$ (see the proof of Lemma 2 in [2]). Therefore, the set of strategies of the attacker is defined by $\mathcal{S}_{\mathscr{A}}$=$\mathcal{S}_{\mathscr{A},T}\times\mathcal{S}_{\mathscr{A},O}$, where

$$\begin{aligned}\mathcal{S}_{\mathscr{A},T} &= \{Q(P_{\tau^m},P_{y^n})\colon \mathcal{P}^m\times\mathcal{P}^n\to\mathcal{P}^m \\ &\qquad \text{such that } d_{L_1}(Q(P_{\tau^m},P_{y^n}),P_{\tau^m}) \leq 2\alpha\}, \quad (75)\\ \mathcal{S}_{\mathscr{A},O} &= \{S^n_{YZ}(P_{y^n},P_{t^m})\colon \mathcal{P}^n\times\mathcal{P}^m\to\mathcal{A}^n(L,P_{y^n})\}. \quad (76)\end{aligned}$$

Note that, in this case, the function $Q(\cdot,\cdot)$ gives the type of the whole training sequence observed by $\mathscr{D}$ (not only the fake subpart, as it was in the $SI^a_{c\text{-}tr}$ game), that is, P_{t^m}=$Q(P_{\tau^m},P_{y^n})$.

In the following, we will find convenient to express the attacking strategies in $\mathcal{S}_{\mathscr{A},T}$ in an alternative way. Since the attacker *replaces* the samples of a subpart of the training sequence, the corruption strategy is equivalent to first removing a subpart of the training sequence and then adding a fake subsequence of the same length. Then, the sequence is reordered to hide the position of the fake samples. By focusing on the type of the observed training sequence, we can write:

$$P_{t^m} = P_{\tau^m}-\alpha Q_R(P_{\tau^m},P_{y^n})+\alpha Q_A(P_{\tau^m},P_{y^n}). \quad (77)$$

where $Q_R(P_{\tau^m},P_{y^n})$ and $Q_A(P_{\tau^m},P_{y^n})$ (both belonging to $\mathcal{P}^{m_2}$) are the types of the removed and injected subsequences respectively. In order to simplify the notation, in the following we will avoid to indicate explicitly the dependence of $Q_R(P_{\tau^m},P_{y^n})$ and $Q_A(P_{\tau^m},P_{y^n})$ on P_{τ^m}, P_{y^n}, and will indicate them as $Q_R()$ and $Q_A()$. Furthermore, we will use notation Q_R and Q_A whenever the dependence from the arguments is not relevant. By varying Q_R and Q_A, we obtain all the pmf's that can be produced from P_{τ^m} by first removing and later adding m_2 samples. Of course not all pairs (Q_R, Q_A) are admissible since the P_{t^m} resulting from (77) must be a valid pmf, i.e. it must be nonnegative for all the symbols of the alphabet $\mathcal{X}$.

3) Payoff: As usual, the payoff function is defined as

$$u(\Lambda^{n\times m},(Q(\cdot,\cdot),S^n_{YZ}(\cdot,\cdot))) = -P_{fn}. \quad (78)$$

In the following section, we will show that a *rationalizable equilibrium* $(\Lambda^{n\times m,*},(Q^*(\cdot,\cdot),\ S^{n,*}_{YZ}(\cdot,\cdot)))$ exists also for the $SI^r_{c\text{-}tr}$ game.

B. Equilibrium point and payoff at the equilibrium

In order to ensure that P_{fp} is always lower than $2^{-\lambda n}$, it is convenient to use the attack formulation given in (77). For a given P_X, Q_R and Q_A, P_{fp} is the probability that X generates two sequences x^n and τ^m, such that the pair of type classes $(P_{x^n},P_{\tau^m}-\alpha(Q_R()-Q_A()))$ falls outside $\Lambda^{n\times m}$. Accordingly, the set of strategies available to $\mathscr{D}$ can be rewritten as:

$$\mathcal{S}_D=\Bigg\{\Lambda^{n\times m}\colon \max_{P_X\in\mathcal{P}}\ \max_{Q_R(),Q_A()}\sum_{P_{y^n}\in\mathcal{P}^n}P_Y(T(P_{y^n}))\cdot \tag{79}$$

$$\sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}}P_X(T(P_{x^n}))\cdot\sum_{\substack{P_{\tau^m}\in\mathcal{P}^m:\\ P_{\tau^m}-\alpha(Q_R()-Q_A())=P_{t^m}}}P_X(T(P_{\tau^m}))\leq 2^{-\lambda n}\Bigg\}.$$

By proceeding as in the proof of Lemma 1, it is easy to prove that the asymptotically optimum strategy for the defender corresponds to the following:

$$\begin{aligned}\Lambda^{n\times m,*}=\big\{&(P_{x^n},P_{t^m})\colon\\ &\min_{Q_R,Q_A\in\mathcal{P}^{m_2}}h(P_{x^n},P_{t^m}+\alpha(Q_R-Q_A))\leq\lambda-\delta_n\big\},\end{aligned}\tag{80}$$

where δ_n tends to 0 as $n\to\infty$ and the minimisation is limited to Q_R and Q_A in $\mathcal{P}^{m_2}$ such that $P_{t^m}+\alpha(Q_R-Q_A)$ is a valid pmf. Consequently, the optimum attacking strategy is given by:

$$\begin{aligned}&(Q^*(P_{\tau^m},P_{y^n}),\ S_{YZ}^{n,*}(P_{y^n},P_{t^m}))=\\ &\operatorname*{argmin}_{\substack{P_{t^m}\ \text{s.t.}\ d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha\\ S_{YZ}^n\in\mathcal{A}^n(L,P_{y^n})}}\left[\min_{Q_R,Q_A}h(P_{z^n},P_{t^m}+\alpha(Q_R-Q_A))\right],\end{aligned}\tag{81}$$

hence resulting in the following theorem.

Theorem 5. *The $SI^r_{c\text{-}tr}$ game with targeted corruption is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile $(\Lambda^{n\times m,*},(Q^*(),\ S_{YZ}^{n,*}()))$ given by equations* (80) *and* (81).

In order to study the asymptotic payoff of the $SI^r_{c\text{-}tr}$ game at the equilibrium, we parallel the analysis carried out in Sec. IV-C. By considering the case L=0, the set of pairs of types for which $\mathscr{D}$ will accept H_0 as a consequence of the attack to the training sequence is given by

$$\begin{aligned}\Gamma_0^n(\lambda,\alpha)=\{&(P_{y^n},P_{\tau^m})\colon\\ &\exists P_{t^m}\ \text{s.t.}\ d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha\\ &\text{and}\ (P_{y^n},P_{t^m})\in\Lambda^{n\times m,*}\}.\end{aligned}\tag{82}$$

If we fix the type of the original training sequence, we get:

$$\begin{aligned}\Gamma_0^n(P_{\tau^m},\lambda,\alpha)&=\{P_{y^n}\colon\exists P_{t^m}\ \text{s.t.}\ d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha\\ &\qquad\text{and}\ P_{y^n}\in\Lambda^{n,*}(P_{t^m})\}\\ &=\{P_{y^n}\colon\exists P_{t^m},\ \exists Q,Q'\in\mathcal{P}^{m_2},\ \text{s.t.}\\ &\qquad d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha\\ &\qquad\text{and}\ h(P_{x^n},P_{t^m}-\alpha Q'+\alpha Q)\leq\lambda-\delta_n\}.\end{aligned}\tag{83}$$

By letting n go to infinity, we obtain the asymptotic counterpart of the above set, which, for a generic $R\in\mathcal{P}$, takes the following expression:

$$\begin{aligned}\Gamma_0(R,\lambda,\alpha)=\big\{&P\colon\exists P',Q,Q',\ \text{s.t.}\ d_{L_1}(P',R)\leq 2\alpha\\ &\text{and}\ h_c(P,P'-\alpha Q'+\alpha Q)\leq\lambda\big\}.\end{aligned}\tag{84}$$

When $L\neq 0$, we obtain:

$$\Gamma(R,\lambda,\alpha,L)=\{P\colon\exists V\in\Gamma_0(R,\lambda,\alpha)\ \text{s.t.}\ \textit{EMD}(P,V)\leq L\}.\tag{85}$$

With the above definitions, it is straightforward to extend Theorem 2 to the $SI^r_{c\text{-}tr}$ case, thus proving that the set in (85) evaluated in $R=P_X$ represents the indistinguishability region of the $SI^r_{c\text{-}tr}$ game.

C. Security margin and blinding corruption level

As a last contribution, we are interested in studying the ultimate distinguishability of two sources X and Y in the $SI^r_{c\text{-}tr}$ setting and compare it with the result we have obtained for the $SI^a_{c\text{-}tr}$ case. To do so, we consider the behaviour of the indistinguishability region when λ tends to 0. We have:

$$\Gamma(P_X,\alpha,L)=\{P\colon\exists V\in\Gamma_0(P_X,\alpha)\ \text{s.t.}\ \textit{EMD}(P,V)\leq L\},\tag{86}$$

where

$$\begin{aligned}\Gamma_0(P_X,\alpha)&=\big\{P\colon\exists P',Q,Q'\ \text{s.t.}\ d_{L_1}(P',P_X)\leq 2\alpha\\ &\qquad\text{and}\ P=P'+\alpha(Q-Q')\big\}\\ &=\big\{P\colon\exists P'\ \text{s.t.}\ d_{L_1}(P',P_X)\leq 2\alpha\\ &\qquad\text{and}\ d_{L_1}(P,P')\leq 2\alpha\big\}.\end{aligned}\tag{87}$$

The set in (87) can be equivalently rewritten as

$$\Gamma_0(P_X,\alpha)=\big\{P\colon d_{L_1}(P,P_X)\leq 4\alpha\big\}.\tag{88}$$

To see why, we first notice that set (87) is contained in (88). Indeed, from the triangular inequality we have that, for any P', $d(P,P_X)\leq d_{L_1}(P,P')+d_{L_1}(P',P_X)$. Then, if P belongs to $\Gamma_0(P_X,\alpha)$ in (87), it also belongs to the set in (88). To see that the two sets are indeed equivalent, it is sufficient to show that the reverse implication also holds. To this purpose, we observe that, whenever $d_{L_1}(P,P_X)\leq 4\alpha$, a type P^* can be found such that its distance from both P and P_X is less or at most equal to 2α. In fact, by letting $P^*=\frac{P+P_X}{2}$, we have

$$\begin{aligned}d_{L_1}(P,P^*)&=d_{L_1}(P^*,P_X)=\sum_i\left|\frac{P(i)-P_X(i)}{2}\right|\\ d_{L_1}(P,P_X)&=\sum_i\Big|P_X(i)-P(i)\Big|=2d_{L_1}(P,P^*).\end{aligned}\tag{89}$$

If $d_{L_1}(P,P_X)\leq 4\alpha$, then, $d_{L_1}(P,P^*)=d_{L_1}(P^*,P_X)=d_{L_1}(P,P_X)/2\leq 2\alpha$, permitting us to conclude that the sets in (87) and (88) are equivalent.

Upon inspection of (88), we can conclude that, as expected, the indistinguishability region for L=0 (and hence, also for the case $L\neq 0$) is larger than that of the $SI^a_{c\text{-}tr}$ game (see (63)), thus confirming that the game with sample replacement is more favourable to the attacker (a graphical comparison

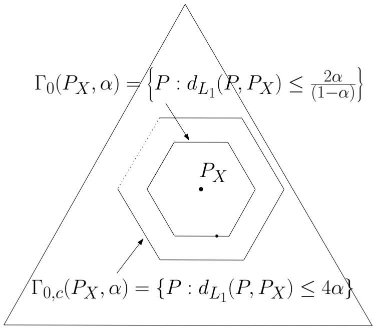
$\Gamma_0(P_X,\alpha) = \left\{P : d_{L_1}(P,P_X) \leq \frac{2\alpha}{(1-\alpha)}\right\}$
P_X
$\Gamma_{0,c}(P_X,\alpha) = \{P : d_{L_1}(P,P_X) \leq 4\alpha\}$

Fig. 8. Comparison of the indistinguishability regions for the $SI^a_{c\text{-}tr}$ and $SI^r_{c\text{-}tr}$ games with L=0.

between the indistinguishability regions for the two setups is shown in Figure 8). As a matter of fact, for the attacker, the advantage of the $SI^r_{c\text{-}tr}$ game with respect to the $SI^a_{c\text{-}tr}$ game depends on α. For small α and for α close to $1/2$, the indistinguishability regions of the two games are very similar, while for intermediate values of α the indistinguishability region of the $SI^r_{c\text{-}tr}$ game is considerably larger than that of the $SI^a_{c\text{-}tr}$ game (the maximum difference between the two regions is obtained for $\alpha \approx 0.3$). When $\alpha = 1/2$ the attacker always wins, since he is able to bring any pmf inside the acceptance region regardless of the game version, while for $\alpha = 0$, we fall back into the source identification game without corruption of the training sequence, thus making the two versions of the game equivalent.

Given two sources X and Y, the blinding corruption level value takes the expression:

$$\alpha_b = \frac{d_{L_1}(P_Y,P_X)}{4}. \qquad (90)$$

Since $d_{L_1}(P_Y,P_X) \leq 2$ for any couple (P_Y,P_X) (the maximum value 2 is taken when the two distribution have disjoint support), the blinding value for the $SI^r_{c\text{-}tr}$ game is lower than the blinding value of $SI^a_{c\text{-}tr}$ game. The two expressions are identical when the two sources have disjoint support, in which case $\alpha_b = 1/2$.

When the attacker can also corrupt the test sequence, the *ultimate indistinguishability region* of the $SI^r_{c\text{-}tr}$game is:

$$\Gamma(P_X,\alpha,L) = \{P: \min_{V:EMD(P,V)\leq L} d_{L_1}(V,P_X) \leq 4\alpha\}. \qquad (91)$$

Starting from (91) we can define the security margin in the $SI^r_{c\text{-}tr}$ setup.

Definition 4 (Security Margin in the $SI^r_{c\text{-}tr}$ 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 $SI^r_{c\text{-}tr}$ setup is called Security Margin and is given by*

$$\mathcal{SM}_\alpha(P_X,P_Y) = L^*_\alpha, \qquad (92)$$

where L^*_α is the quantity which satisfies the following relation

$$\min_{V:\text{EMD}(P_Y,V)\leq L^*_\alpha} d_{L_1}(V,P_X) = 4\alpha, \qquad (93)$$

if $P_Y \notin \Gamma_0(P_X,\alpha)$, and $L^*_\alpha = 0$ otherwise.

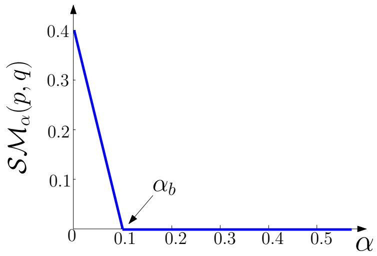

Fig. 9. Security margin as a function of α for Bernoulli sources with parameters p=0.3 and q=0.7 (α_b=0.1).

Considering again the case of two Bernoulli sources and by adopting the same notation of Section V-B1, we have that $\alpha_b = |p-q|/4$, while the security margin is

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p|-2\alpha & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b \end{cases}. \qquad (94)$$

Figure 6 plots $\mathcal{SM}_\alpha$ as a function of α when p=0.3 and q=0.7. The blinding value is α_b=0.1 which, as expected, is lower than the value we found for the $SI^a_{c\text{-}tr}$ setup.

VII. Conclusions and final remarks

We studied the distinguishability of two sources in an adversarial setup when the sources are known through training data, part of which can be corrupted by the attacker himself. We considered two different scenarios. In the first one, the attacker simply adds fake samples to the original training sequence, while in the second one, the attacker replaces a selected subset of training samples with fake ones. We formalised both cases in a game-theoretic setup, then we derived the equilibrium point of the games and analysed the (asymptotic) payoff at the equilibrium. The result of the game can be summarised in a compact and elegant way by introducing two parameters, namely the Security Margin under corruption of the training sequence, and the blinding corruption level α_b, defined as the portion of fake samples the attacker must introduce to make impossible any reliable distinction between the sources. Based on these two parameters, the performance of the two games with corruption of the training data can be easily compared.

Though rather theoretical, our findings can guide more practical researches in several fields belonging to the emerging areas of adversarial signal processing [1] and secure machine learning [8]. In many cases, in fact, the defender must take into account the possibility that the data he is using to tune the system he is working at, or during the learning phase, is corrupted by the attacker. Of course, the extent to which the results proven in this paper can be applied to practical scenarios depends much on each specific application and the degree to which the assumptions under the theory are satisfied. In several image forensics applications, for instance, the forensic analysis relies on image histogram [32], [33] or on the histogram of DCT coefficients [34], [35], thus justifying the assumption that the defender relies only on the first order statistics of the observed sequence. In other cases, the restrictions imposed by the theory can be interpreted as a worst case assumption of the defender. This is again the case of the first order statistics assumption. If the performance achieved in this setup are good enough, the defender can be sure that in practical applications, when the analysis is not limited to first order statistics, the results will be even more favourable.

The dependence of the optimum strategy of the defender on α (see (23)) is another critical point for the application to the real scenarios. When it is not reasonable to assume that the percentage of corrupted training samples is know to the defender, we can interpret the value of α used by $\mathscr{D}$ as a worst case estimate of the capabilities of the attacker. If the actual value of α used by the attacker is larger than the estimate, then $\mathscr{D}$ fails to even satisfy the false positive constraint hence resulting in a complete failure of the Neyman-Pearson setup. If instead the value used by the attacker is smaller, then the constraint on the false positive is surely satisfied, and the payoff of the game will be better than that predicted by theory (from the point of view of the defender), hence justifying the interpretation on α as a worst case estimate made by $\mathscr{D}$.

From the point of view of the attacker, the optimum strategies derived in Sections IV and VI can be the starting point for the development of theoretically-sound attacks capable of defeating any detector relying on a certain class of statistics (namely first order statistics). Such a path has already been followed in the case of source identification with uncorrupted training data, whose results have been applied to devise a universal attack against image forensics techniques based on image histogram [36]. We expect that a similar exploitation of the theoretical analysis be possible also in the case of training with corrupted samples.

The analysis carried out in this paper can be extended in several ways, for instance by considering continuous sources, or by assuming that the sources X and Y are not memoryless, but still amenable to be studied by using the method of types [37]. Following the analysis in [38], we could also consider a more general setup in which the attacker is active under both H_0 and H_1. An interesting generalisation consists in studying a symmetric setup in which the training and the test sequences can be corrupted by applying the same kinds of processing. For instance, the attacker could be allowed to replace samples in both the training and the set sequences, or he could be allowed to modify the training sequence up to a certain distortion. Other kinds of attacks to the training data could also be considered, like sample removal with no addition of fake samples. As a matter of fact, the kind of attack strongly depends on the application scenario, and it is arguable that the availability of a large variety of theoretical models would help bridging the gap between theory and practice.

Acknowledgment

This work has been partially supported by a research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-16-2-0173. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

Appendix

A. Generalized Sanov's theorem

Let us consider a sequence of n i.i.d. discrete random variables taking values in a finite alphabet $\mathcal{X}$ and distributed according to a pmf P. We denote with P_n the empirical pmf of the sequence. Let $E \subseteq \mathcal{P}$ be a set of pmf's. Sanov's theorem [11], [39], [12] states that

$$\begin{aligned}\inf_{Q\in E}\mathcal{D}(Q||P) &\leq -\limsup_{n\to\infty}\frac{1}{n}\log P(P_n\in E)\\ &\leq -\liminf_{n\to\infty}\frac{1}{n}\log P(P_n\in E)\\ &\leq \inf_{Q\in \mathit{int}\ E}\mathcal{D}(Q||P),\end{aligned} \tag{A1}$$

where *int* S denote the interior part of the set S.

When $\mathit{cl}(E)$ = $\mathit{cl}(\mathit{int}(E))$[10], or, $E \subseteq \mathit{cl}(\mathit{int}(E))$, the left and right-hand side of (A1) coincide and we get the exact rate:

$$-\lim_{n\to\infty}\frac{1}{n}\log P(P_n\in E) = \inf_{Q\in E}\mathcal{D}(Q||P). \tag{A2}$$

If we define the set E_n=$E\cap\mathcal{P}^n$, we have: $P(P_n\in E)$=$P(P_n\in E_n)$ and we can rewrite Sanov's theorem as:

$$\begin{aligned}\inf_{Q\in E}\mathcal{D}(Q||P) &\leq -\limsup_{n\to\infty}\frac{1}{n}\log P(P_n\in E_n)\\ &\leq -\liminf_{n\to\infty}\frac{1}{n}\log P(P_n\in E_n)\\ &\leq \inf_{Q\in \mathit{int}\ E}\mathcal{D}(Q||P),\end{aligned} \tag{A3}$$

Note that, by construction, we have $\mathit{cl}(E)$ = $\mathit{cl}(\cup_n E_n)$.

In the following, we extend the formulation of Sanov's theorem given in (A3) to more general sequences of sets E_n for which it does not necessary hold that E_n=$E\cap\mathcal{P}^n$ for some set E.

We start by introducing the notion of convergence for sequences of subsets due to Kuratowsky, which is a more general notion of convergence with respect to the one based on Hausdorff distance. Let (S,d) be a metric space. We first provide the definition of *lower closed limit* or Kuratowski limit inferior [40].

Definition 5. *A point p belongs to the lower limit* $\underset{n\to\infty}{Li}\ K_n$ *(or simply* LiK_n*) of a sequence of sets* K_n*, if every neighborhood of* p *intersects all the* K_n*'s from a sufficiently great index* n *onward.*

Given the above definition, the expression $p \in \underset{n\to\infty}{Li}\ K_n$ *is equivalent to the existence of a sequence of points* $\{p_n\}$ *such that:*

$$p = \lim_{n\to\infty} p_n, \quad p_n\in K_n. \tag{A4}$$

Stated in another way, LiK_n *is the set of the accumulation points of sequences in* K_n*. As an alternative, equivalent, definition we can let:*

$$\underset{n\to\infty}{Li}\ K_n = \{p \in X \ \textit{s.t.}\ \limsup_{n\to\infty} d(x,K_n) = 0\}. \tag{A5}$$

[10] $\mathit{cl}(E)$ denotes the closure of E. Clearly, $\mathit{cl}(E) \equiv E$ if E is a closed set.

Similarly, we have the following definition of *upper closed limit* or Kuratowski limit superior [40].

Definition 6. *A point p belongs to the upper limit $\underset{n\to\infty}{Ls}\, K_n$ (or simply LsK_n) of a sequence of sets K_n, if every neighborhood of p intersects an infinite number of terms in K_n.*

The expression $p \in Ls_{n\to\infty}K_n$ is equivalent to the existence of a subsequence of points $\{p_{k_n}\}$ such that

$$k_1{<}k_2{<}..., \quad p = \lim_{n\to\infty} p_{k_n}, \quad p_{k_n}{\in}\ K_{k_n}.$$

As an alternative, equivalent, definition we can let:

$$\underset{n\to\infty}{Ls}\, K_n = \{p \in X \ s.t. \liminf_{n\to\infty} d(x,K_n) = 0\}. \qquad \text{(A6)}$$

It can be proven that the Kuratowski limit inferior and superior are always closed set (see [40]).

Given the above, we can state the following:

Definition 7. *The sequence of sets $\{K_n\}$ is said to be convergent to K in the sense of Kuratowski, that is $K_n \overset{K}{\to} K$, if $LiK_n{=}K{=}LsK_n$, in which case we write $K{=}LimK_n$.*

We observe that Kuratowski convergence is weaker than convergence in Hausdorff metric; in fact, given a sequence of closed sets $\{K_n\}$, $K_n \overset{H}{\to} K$ implies $K_n \overset{K}{\to} K$ [41]. For compact metric spaces, the reverse implication also holds and the two kinds of convergence coincide.

In this work, we are interested in the space $\mathcal{P}$ of probability mass functions defined over a finite alphabet $\mathcal{X}$, i.e., the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$, equipped with the L_1 metric. Being $\mathcal{P}$ a closed subset of $\mathbb{R}^{|\mathcal{X}|}$, $\mathcal{P}$ is a complete set. In addition, with the L_1 metric, $\mathcal{P}{\in}\mathcal{L}(\mathbb{R}^{|\mathcal{X}|})$, that is, $\mathcal{P}$ is bounded. The space $(\mathcal{P}, d_{L_1})$, then, is a compact metric space and then, for our purposes, Kuratowski and Hausdorff convergence are equivalent.

We are now ready to prove the following generalisation of Sanov's theorem:

Theorem 6 (Generalized Sanov's theorem)**.** *Let $\{E_{(n)}\}$ be a sequence of sets in $\mathcal{P}$, such that $Li(E_{(n)}{\cap}\mathcal{P}^n){\neq}\ \emptyset$. Then:*

$$\begin{aligned}\min_{Q\in\ LsE_{(n)}} \mathcal{D}(Q||P) &\le -\limsup_{n\to\infty} \frac{1}{n}\log P(P_n{\in}E_{(n)})\\ &\le -\liminf_{n\to\infty} \frac{1}{n}\log P(P_n{\in}E_{(n)})\\ &\le \min_{Q\in\ Li\ (E_{(n)}\cap\mathcal{P}^n)} \mathcal{D}(Q||P), \qquad \text{(A7)}\end{aligned}$$

If, in addition, $LsE_{(n)}{=}Li(E_{(n)}{\cap}\mathcal{P}^n)$, the generalized Sanov's limit exists as follows:

$$-\lim_{n\to\infty} \frac{1}{n}\log P(P_n{\in}E_{(n)}) = \min_{Q\in LimE_{(n)}} \mathcal{D}(Q||P). \qquad \text{(A8)}$$

Proof. We first prove the expression for the lower bound. Let $E_n = E_{(n)}{\cap}\mathcal{P}^n$. We have:

$$\begin{aligned}P(E_{(n)}) &= \sum_{Q\in E_n} P_X(T(Q))\\ &\le (n{+}1)^{|\mathcal{X}|}2^{-n\min_{Q\in E_n}\mathcal{D}(Q||P)}\\ &\le (n{+}1)^{|\mathcal{X}|}2^{-n\inf_{Q\in E_{(n)}}\mathcal{D}(Q||P)}\\ &= (n{+}1)^{|\mathcal{X}|}2^{-n\min_{Q\in cl(E_{(n)})}\mathcal{D}(Q||P)}. \qquad \text{(A9)}\end{aligned}$$

In the last inequality we exploited the fact that, being each $E_{(n)}$ a bounded set of $\mathcal{P}$, and $\mathcal{D}$ lower bounded in $\mathcal{P}$, the infimum over $E_{(n)}$ corresponds to the minimum over its closure. By taking the logarithm of each side and dividing by n, we get:

$$\frac{1}{n}\log P(E_{(n)}) \le -\min_{Q\in cl(E_{(n)})} \mathcal{D}(Q||P) + \frac{\log(n{+}1)^{|\mathcal{X}|}}{n}, \quad \text{(A10)}$$

We now prove that, for any δ and for sufficiently large n, we have

$$\min_{Q\in cl(E_{(n)})} \mathcal{D}(Q||P) \ge \min_{Q\in LsE_{(n)}} \mathcal{D}(Q||P) - \delta. \qquad \text{(A11)}$$

First, according to the properties of the limit superior, $LsE_{(n)}{=}Ls(cl(E_{(n)}))$ [40], hence proving (A11) is equivalent to showing that:

$$\min_{Q\in cl(E_{(n)})} \mathcal{D}(Q||P) \ge \min_{Q\in Ls(cl(E_{(n)}))} \mathcal{D}(Q||P) - \delta. \qquad \text{(A12)}$$

Let Q_n be the sequence of points achieving the minimum of the left-hand side of (A12) (for simplicity we assume that the minimum is unique, the extension to a more general case being straightforward). Let $Q_{n(j)}$ be a subsequence of Q_n formed only by the elements of Q_n that do not belong to $Ls(cl(E_{(n)}))$[11]. If the number of elements in $Q_{n(j)}$ is finite, then for n large enough $Q_n \in Ls(cl(E_{(n)}))$ and eq. (A12) is verified with $\delta{=}0$. If the number of elements in $Q_{n(j)}$ is infinite, then, due to the boundedness of $\mathcal{P}$, the elements of $Q_{n(j)}$ must have at least one accumulation point (Bolzano-Weierstrass theorem). Let A_i's be the accumulation points of $Q_{n(j)}$. By definition of Ls, all A_i's belong to $Ls(cl(E_{(n)}))$. In addition, for any radius ρ, from a certain j on, all the points in $Q_{n(j)}$ belong to $\mathcal{R}{=}\bigcup_i\mathcal{B}(A_i,\rho)$[12]. For large enough n, then we have:

$$\begin{aligned}\min_{Q\in cl(E_{(n)})} \mathcal{D}(Q||P) &\ge \min_{Q\in Ls(cl(E_{(n)}))\cup\mathcal{R}} \mathcal{D}(Q||P) \qquad \text{(A13)}\\ &\ge \min_{Q\in Ls(cl(E_{(n)}))} \mathcal{D}(Q||P) - \delta,\end{aligned}$$

where the second inequality derives from the continuity of the $\mathcal{D}$ function and the arbitrariness of ρ.

By inserting (A11) in (A10), we have that, for large n,

$$\frac{1}{n}\log P(E_{(n)}){\le}{-}\min_{Q\in LsE_{(n)}} \mathcal{D}(Q||P) + \frac{\log(n{+}1)^{|\mathcal{X}|}}{n} + \delta, \quad \text{(A14)}$$

and hence, by the arbitrariness of δ,

$$-\limsup_{n\to\infty} \frac{1}{n}\log P(E_{(n)}) \ge \min_{Q\in LsE_{(n)}} \mathcal{D}(Q||P). \qquad \text{(A15)}$$

We now pass to the upper bound. Let Q^* be a point achieving the minimum of the divergence over the set LiE_n. By definition of limit inferior, there exists a sequence of points $\{Q_n\}$, $Q_n{\in}E_n$ such that $Q_n{\to}Q^*$ as $n{\to}\infty$. Then, by exploiting the continuity of $\mathcal{D}$, it follows that:

$$\mathcal{D}(Q_n||P) \le D(Q^*||P) + \gamma, \qquad \text{(A16)}$$

[11] $n(i) > n(j), \forall i > j$

[12] $\mathcal{B}(A_i,\rho)$ is a ball with radius ρ centred in A_i.

where γ can be made arbitrarily small for large n. We can then write:

$$\begin{aligned} P(E_{(n)}) &= \sum_{Q\in E_n} P(T(Q)) \\ &\geq P(T(Q_n)) \geq \frac{2^{-n\mathcal{D}(Q_n||P)}}{(n+1)^{|\mathcal{X}|}}. \end{aligned} \tag{A17}$$

Hence, we get

$$\begin{aligned} \frac{1}{n}\log P(E_{(n)}) &\geq -\mathcal{D}(Q_n||P) - |\mathcal{X}|\frac{\log(n+1)}{n}, \\ &\geq -\mathcal{D}(Q^*||P) - \gamma - |\mathcal{X}|\frac{\log(n+1)}{n}, \\ &\geq -\min_{Q\in LiE_n}\mathcal{D}(Q||P) - \gamma - |\mathcal{X}|\frac{\log(n+1)}{n}, \end{aligned} \tag{A18}$$

and then, by the arbitrariness of γ,

$$-\liminf_{n\to\infty}\frac{1}{n}\log P(E_{(n)}) \leq \min_{Q\in LiE_n}\mathcal{D}(Q||P), \tag{A19}$$

which concludes the proof of the first part (relation (A7)).

For the proof of the second part, we observe that, when $LsE_{(n)}{=}Li(E_{(n)}{\cap}\mathcal{P}^n)$, the two bounds in (A7) coincides. Moreover, the following chain of inclusions holds, $LiE_{(n)} \subseteq LsE_{(n)} = Li(E_{(n)}{\cap}\mathcal{P}^n) \subseteq LiE_{(n)}$, and then $LiE_{(n)} = LsE_{(n)} = LimE_{(n)}$, yielding (A8). □

We observe that, in general, the Kuratowski convergence of $E_{(n)}$ is a *necessary* condition for the existence of the generalized Sanov limit in (A8), but it is not sufficient. In fact, we could have $LiE_{(n)} \supseteq Li(E_{(n)}{\cap}\mathcal{P}^n)$, in which case the lower and upper bound in (A7) do not coincide. It is also interesting to notice that when $E_{(n)} \in \mathcal{P}^n$ is a sequence of sets in $\mathcal{P}^n$, then Sanov's limit holds whenever $E_{(n)}\overset{K}{\to}E$ for some set E, or, by exploiting the compactness of $\mathcal{P}$, $E_{(n)}\overset{H}{\to}E$. Based on the above observation, we can state the following corollary:

Corollary 1. *Let $E_{(n)}$ be a sequence of sets in $\mathcal{P}^n$, such that $E_{(n)}\overset{H}{\to}E$. Then:*

$$-\lim_{n\to\infty}\frac{1}{n}\log P(P_n{\in}E_{(n)}) = \min_{Q\in E}\mathcal{D}(Q||P). \tag{A20}$$

B. Regularity properties of the set of admissible maps

To prove the theorems on the asymptotic behaviour of the payoff in the two versions of the source identification game studied in this paper, we need to prove some regularity theorems on the set of admissible maps.

To start with, we need to define a distance between transportation maps, that is a function d_s: $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}{\times}\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}{\to}\mathbb{R}^{+}$. In accordance with the rest of the paper, let us choose the L_1 distance, that is, given two maps (S_{PV},S_{QR}), we define $d_s(S_{PV},S_{QR}){=}\sum_{i,j}|S_{PV}(i,j){-}S_{QR}(i,j)|$.

Our first result regards the regularity of $\mathcal{A}(L,P)$ as a function of P.

Lemma 2. *Let $P{\in}\mathcal{P}$ and let P' be any pmf in the neighbourhood of P of radius τ, i.e., $P'{\in}\mathcal{B}(P,\tau)$. Then*

$$\delta_H(\mathcal{A}(L,P),\, \mathcal{A}(L,P')) \leq \tau$$

and hence $\lim_{\tau\to 0}\delta_H(\mathcal{A}(L,P),\mathcal{A}(L,P')) = 0$, uniformly in $\mathcal{P}$. Moreover, if we insist that $P'{\in}\mathcal{P}^n$, the following result holds: $\forall\varepsilon{>}0$, $\exists\tau^*$ and n^* such that $\forall\tau{<}\tau^*$ and $n{>}n^*$,

$$\delta_H(\mathcal{A}(L,P),\, \mathcal{A}^n(L,P')) \leq \varepsilon \quad \forall P'{\in}\mathcal{B}(P,\tau){\cap}\mathcal{P}^n,\, \forall P{\in}\mathcal{P}.$$

Proof. From a general perspective, the lemma follows from the fact that $\mathcal{A}^n(L,P_{y^n})$ (and $\mathcal{A}(L,P)$) is built by imposing a number of linear constraints on the admissible transportation maps (see (11)), i.e. $\mathcal{A}(L,P)$ is a convex polytope [42], [43]. By considering a P' close to P, we are perturbing the vector of the known terms of the linear constraints which defines the admissibility set. Instead of invoking the above general principle, in the following we give an explicit proof of the lemma.

Given $P{\in}\mathcal{P}$ and $P'{\in}\mathcal{B}(P,\tau)$, let $\tau(i){=}P(i){-}P'(i)$ be the excess (or defect) of mass of P with respect to P' in bin i. For any map in $\mathcal{A}(L,P)$, we can choose a map $S_{P'V'}$ that works as follows: for the bins i such that $\tau(i){\leq}0$, let $S_{P'V'}(i,j){=}S_{PV}(i,j)$ for $j{\neq}i$, while for $j{=}i$, we let $S_{P'V'}(i,j){=}S_{PV}(i,j){+}|\tau(i)|$. For the bins i for which $\tau(i){>}0$, we first sort the index set $\{j{:}S_{PV}(i,j){\neq}0\}$ in decreasing order with respect to the amount of distortion introduced per unit of mass delivered from i to j ($d(i,j)$). Then, starting from the first index in the ordered list, we let $S_{P'V'}(i,j){=}\max(0,\, S_{PV}(i,j){-}\tau(i))$. If $S_{P'V'}(i,j){=}0$, we update $\tau(i)$ to a new value $\tau'(i){=}\tau(i){-}S_{PV}(i,j)$, and iterate the previous procedure by subtracting the updated value of $\tau'(i)$ from the second $S_{PV}(i,j)$ in the list. This procedure goes on until the subtraction gives $S_{P'V'}(i,j){\neq}0$, that is when we have removed all the excess mass from the i-th row of $S_{PV}(i,j)$.

It is easy to see that the map built in this way satisfies the distortion constraint, in fact, by construction the distortion associated to $S_{P'V'}$ is less than that introduced by S_{PV}. Then, $S_{P'V'}{\in}\mathcal{A}(L,P')$. In addition, by construction, $\sum_j|S_{P'V'}(i,j){-}S_{PV}(i,j)|{\leq}|\tau(i)|$, and hence $\sum_{ij}|S_{P'V'}(i,j){-}S_{PV}(i,j)|{\leq}\tau$. Accordingly, we have:

$$\begin{aligned} &\delta_{\mathcal{A}(L,P)}(\mathcal{A}(L,P'))= \\ &\max_{S_{PV}\in\mathcal{A}(L,P)}\ \min_{S_{P'V'}\in\mathcal{A}(L,P')} d_s(S_{PV},S_{P'V'}) \leq \tau \end{aligned} \tag{A21}$$

since, as we have shown with the preceding construction, the inner minimum is always lower or equal than τ. By repeating the same argument exchanging the role of $\mathcal{A}(L,P)$ and $\mathcal{A}(L,P')$, we find that $\delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)){\leq}\tau$, thus concluding the first part of the proof.

In the second part of the lemma, we require that $P'{\in}\mathcal{P}^n$ and that the map produces a sequence in $\mathcal{P}^n$. The proof is easily achieved by exploiting the first part of the lemma according to which for any map S_{PV} in $\mathcal{A}(L,P)$, we can find a map $S_{P'V'}$ in $\mathcal{A}(L,P')$ which is arbitrarily close to S_{PV}, and then approximating $S_{P'V'}$ with a map $S^n_{P'V'}{\in}\mathcal{A}^n(L,P')$. Due to the density of rational numbers in real numbers, such an approximation can be made arbitrarily accurate by increasing n, thus completing the proof. □

Given a transformation S_{PV} mapping P into V, Lemma 2 states that, for any pmf P' close to P, we can find a map

$S_{P'V'}$ close to S_{PV}. The following theorem extends such a result to the pmf resulting from the application of the mapping.

Theorem 7. *Let $P\in\mathcal{P}$, and let P' be any pmf in the neighbourhood of P of radius τ, i.e., $P'\in\mathcal{B}(P,\tau)$. Let $S_{PV}\in\mathcal{A}(L,P)$. Then, we can always find a map $S_{P'V'}\in\mathcal{A}(L,P')$ such that $V'\in\mathcal{B}(V,\tau)$.*

Similarly, for any $\varepsilon>0$, there exist τ^* and n^* such that $\forall$ $\tau<\tau^*$ and $n>n^*$, given a $P\in\mathcal{P}$, a map $S_{PV}\in\mathcal{A}(L,P)$ and $P'\in\mathcal{P}^n\cap\mathcal{B}(P,\tau)$, we can find a map $S^n_{P'V'}$ in $\mathcal{A}^n(L,P')$ such that $V'_n\in\mathcal{B}(V,\varepsilon)\cap\mathcal{P}^n$.

Proof. For any two maps S_{PV} and $S_{P'V'}$, we have:

$$\begin{aligned} V'(j) &= \sum_i S_{P'V'}(i,j) \\ &= \sum_i (S_{PV}(i,j)+(S_{P'V'}(i,j)-S_{PV}(i,j))) \\ &\le V(j)+\sum_i |S_{P'V'}(i,j)-S_{PV}(i,j)|, \end{aligned} \tag{A22}$$

and

$$\begin{aligned} V'(j) &= \sum_i S_{P'V'}(i,j) \\ &= \sum_i (S_{PV}(i,j)+(S_{P'V'}(i,j)-S_{PV}(i,j))) \\ &\ge V(j)-\sum_i |S_{P'V'}(i,j)-S_{PV}(i,j)|, \end{aligned} \tag{A23}$$

yielding:

$$|V'(j)-V(j)| \le \sum_i |S_{P'V'}(i,j)-S_{PV}(i,j)|. \tag{A24}$$

By summing over j and exploiting Lemma 2, we can choose $S_{P'V'}$ so that:

$$\begin{aligned} \sum_j |V'(j)-V(j)| &\le \sum_{i,j} |S_{P'V'}(i,j)-S_{PV}(i,j)| \\ &\le \delta_H(\mathcal{A}(L,P'),\, \mathcal{A}(L,P)) \le \tau, \end{aligned} \tag{A25}$$

and hence $V'\in\mathcal{B}(V,|\tau)$.

Similarly to the second part of Lemma 2, the second part of the theorem follows immediately from the density of rational numbers in the real line. □

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, August 2014, doi:10.1109/TIT.2014.2325571.

[4] 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

[5] M. Barni and B. Tondi, "Source distinguishability under distortion-limited attack: an optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, October 2016, doi:10.1109/TIFS.2016.2570739.

[6] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition : A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, Sept 2015.

[7] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.

[8] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[9] B. Biggio, I. Corona, B. Nelson, B. I. P. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, *Security Evaluation of Support Vector Machines in Adversarial Environments*. Cham: Springer International Publishing, 2014, pp. 105–153.

[10] B. Tondi, M. Barni, and N. Neri Merhav, "Detection games with a fully active attacker," in *Proc. of WIFS 2015. IEEE International Workshop on Information Forensics and Security*. IEEE, 2015, pp. 1–6.

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

[12] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer Science & Business Media, 2009.

[13] M. Barni and B. Tondi, "Source distinguishability under corrupted training," in *Proc. of Wifs 2014, IEEE International Workshop on Information Forensics and Security*, Atlanta, Georgia, 3-5 December 2014.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems. 2nd edition*. Cambridge University Press, 2011.

[15] 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.

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

[17] J. Munkres, *Topology*, ser. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. [Online]. Available: https://books.google.it/books?id=XjoZAQAAIAAJ

[18] J. Henrikson, "Completeness and total boundedness of the Hausdorff metric," *MIT Undergraduate Journal of Mathematics*, vol. 1, pp. 69–80, 1999.

[19] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[20] 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.

[21] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.

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

[23] J. C. Harsanyi, "Games with incomplete information," *The American Economic Review*, vol. 85, no. 3, pp. 291–303, 1995.

[24] Y. Liu, C. Comaniciu, and H. Man, "A bayesian game approach for intrusion detection in wireless ad hoc networks," in *Proceeding from the 2006 workshop on Game theory for communications and networks*. ACM, 2006, p. 4.

[25] A. Garnaev and W. Trappe, "A bandwidth monitoring strategy under uncertainty of the adversary's activity," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 4, pp. 837–849, 2016.

[26] Y. E. Sagduyu, R. A. Berry, and A. Ephremides, "Jamming games in wireless networks with incomplete information," *IEEE Communications Magazine*, vol. 49, no. 8, 2011.

[27] A. Garnaev, M. Baykal-Gursoy, and H. V. Poor, "Incorporating attack-type uncertainty into network protection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1278–1287, 2014.

[28] H. Zeng, J. Liu, J. Yu, X. Kang, Y. Q. Shi, and Z. J. Wang, "A framework of camera source identification bayesian game," *IEEE transactions on cybernetics*, 2017.

[29] M. R. Bussieck and A. Pruessner, "Mixed-integer nonlinear programming," *SIAG/OPT Newsletter: Views & News*, vol. 14, no. 1, pp. 19–22, 2003.

[30] P. Bonami, M. Kilinç, and J. Linderoth, "Algorithms and software for convex mixed integer nonlinear programs," *Mixed integer nonlinear programming*, pp. 1–39, 2012.

[31] 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.

[32] G. Cao, Y. Zhao, R. Ni, and X. Li, “Contrast enhancement-based forensics in digital images,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 515–525, March 2014.

[33] M. Stamm and K. J. R. Liu, “Blind forensics of contrast enhancement in digital images,” in *Proc. of ICIP 2008, 15th IEEE International Conference on Image Processing, year = 2008, pages = 3112–3115*.

[34] T. Pevny and J. Fridrich, “Detection of double-compression in jpeg images for applications in steganography,” *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, June 2008.

[35] S. Milani, M. Tagliasacchi, and S. Tubaro, “Discriminating multiple jpeg compression using first digit features,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2253–2256.

[36] M. Barni, M.Fontani, and B. Tondi, “A universal attack against histogram-based image forensics,” *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 5, no. 3, 2013.

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

[38] B. Tondi, M. Barni, and N. Merhav, “Detection games with a fully active attacker,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2015, pp. 1–6.

[39] S. I.N., “On the probability of large deviations of random variables,” *Math. Sbornik*, vol. 42, pp. 11–44, 1957.

[40] K. Kuratowski, *Topology*, ser. Topology. Academic Press, 1968, vol. 1.

[41] G. Salinetti and J. B. Wets, “On the convergence of sequences of convex sets in finite dimensions,” *Siam review*, vol. 21, no. 1, pp. 18–33, 1979.

[42] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Belmont, MA: Athena Scientific, 1997, vol. 6.

[43] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.