Smart Detection of Line-Search Oracle Attacks

Benedetta Tondi *Student Member, IEEE*, Pedro Comesaña-Alfaro *Senior Member, IEEE*, Fernando Pérez-González *Fellow, IEEE*, Mauro Barni *Fellow, IEEE*

***Abstract*—We discuss how a binary detector can learn whether it is being subject to an oracle attack by resorting to a higher level of detection (*metadetection*). On a second step, assuming that the attacker is aware of the fact that the detector takes countermeasures, we investigate a possible way for him to react. Then, we study the interplay between the defender and the attacker when both of them try to do their best for pursuing their opposite goals. We focus our analysis on the metadetection of oracle attacks based on line search algorithms, as they are prevalent in the literature. In such scenario, we propose metadetectors which work under very general conditions, that is, when the oracle is not exclusively fed with line search attacking queries, but only some of the malicious queries are made along the lines, whereas the others are done by mimicking the behavior of honest users. We theoretically evaluate the final achievable performance of these metadetectors, deriving conditions under which asymptotic powerful testing is possible. Experimental results show the power of metadetection for countering the line search attacks in both synthetic and real application scenarios.**

***Index Terms*—Adversarial signal processing, composite hypothesis testing, intentional attacks, sample covariance matrix (high-dimensionality regime), singular value decomposition, zero-bit watermarking.**

I. Introduction

Adversarial Binary Detection, that is, the study of binary detection under intentional attacks, is a prominent problem in security-related applications, like forensic detection, biometric authentication, fingerprint detection, network intrusion detectors, spam filtering, reputation systems, etc. In these applications, in fact, detection is often hampered by adversaries that actively modify their behavior and the observations the detection relies on to cause misclassification. As a consequence, classical detection theory and design methods must be revised to account for the existence of adversaries aimed at misleading the system [1]. Among intentional attacks, those based on the information gathered by repeatedly querying the detector, often referred to in the literature by the name of *oracle attacks*, have been shown to be very powerful (see Section II).

Oracle attacks are a serious threat for watermarking systems [2], [3]; such type of attacks can also be found in many other multimedia security applications: for instance, hill-climbing attacks in biometric recognition systems belong to this category [4]–[6], as well as similar kind of attacks in spam filtering and intrusion detection [7]–[9]. The effectiveness and generality of oracle attacks, then, calls for the development of proper countermeasures.

A. Introduction to Smart detection

Arming against malicious attackers that query the classifier to learn information and then construct their attack is becoming a common need in security-oriented applications (e.g., [10], [11], or [12], [13] for a generic adversarial strategy). Specifically, a novel direction for counteracting oracle attacks, relying on the use of *smart detectors*, has been recently explored in [10]. A smart detector is defined as a detector that is able to *learn from* and *react to* repeated query attacks. Notice that detectors producing a random output close to the boundary are not smart according to the previous definition, because they are not able to determine whether they are being attacked. To learn whether the system is being subject to an oracle attack, a *metadetector* is proposed that works at a higher level than the primary detector. While the operation of the latter is not modified, the former is specifically devoted to detect malicious queries and its definition is not affected by the specific purpose of the primary detector. Once the smart detector decides that an oracle attack is ongoing, effective countermeasures can be enforced, including the prevention of further accesses to the detector (banning), or the conservative switch to a more convoluted detection function.

In [10], two different metadetectors are proposed; one of them, named Closeness-To-the-Boundary (CTB) metadetector, works under general attack models, and simply exploits the fact that oracle attacks generally produce a large number of queries close to the detection boundary; the other one, named Line Search (LS) metadetector, targets the line searches typically performed by those attacks (cf. [2], [14], [15]). Both strategies are shown to successfully detect oracle attacks with very few queries. The analysis of CTB-based metadetection is generalized in [16] by removing the assumption that the detector is exclusively fed with either malicious or honest queries. It is worth stressing that, although [10] and [16] are focused on watermark detection, the proposed metadetectors are higher-level detectors that can be applied to any binary decision problem where oracle attacks can be a threat, i.e., regardless of the underlying detection problem.

B. Contribution

In this paper, we propose a generalized LS metadetection which works in a very general attacking scenario, i.e., under very general attacking conditions. Starting from the metadetector in [10], we make a step forward by considering the possible reaction of the attacker to the countermeasures taken by the defender in an attempt to restore the effectiveness of

M. Barni and B. Tondi are with the Dept. of Information Engineering, University of Siena, Via Roma 56, 53100 - Siena, ITALY, e-mail: barni@dii.unisi.it; benedettatondi@gmail.com.

P. Comesaña-Alfaro and F. Pérez-González are with the Dept. Teoría de la Señal y Comunicaciones. EE Telecom., Universidad de Vigo, 36310 Vigo, SPAIN, e-mail: pcomesan@gts.uvigo.es; fperez@gts.uvigo.es

the attack. The analysis of the counterattack allows us, in turn, to refine the metadetection. Specifically, the new metadetector takes into account the possibility that 'useful' attacking queries (i.e., the aligned queries of the line search attack) are mixed with queries explicitly generated by the attacker to mislead the metadetection. Some practical sub-optimal metadetectors are proposed, and their asymptotic performance is theoretically analyzed, including the derivation of those conditions under which asymptotically powerful testing is possible. Furthermore, such analysis permits us to determine critical values of the system parameters for a correct detection of the oracle attack. Experiments confirm the power of smart detection, by showing that the metadetector is able to reliably detect oracle attacks, even when only very few attacking queries are hidden in a batch of misleading queries.

It is worth remarking that the refined metadetector presented in this paper also works in the practical scenario in which several users, potentially including both honest and malicious, query the system during an observation period. In such situation, the metadetector has to face the slightly different problem of discovering if there is an attacker among the users querying the system. It is easy to see that this problem has the same formalization as the one addressed in this paper, and then our analysis can be also applied in this situation.

The remainder of the paper is organized as follows: in Section II we provide a general introduction to multimedia watermarking and to related literature on the problem of oracle attacks. Then, in Section III we enter the heart of the paper: we first present the metatest and then formally define the LS metadetector. In Section IV, we discuss about possible counterattacks (higher-level oracle attacks) and we propose some LS metadetector implementations which are able to cope with them. An analysis of the performance of those LS metadetectors with respect to the design parameters is reported in Section V. Section VI is devoted to the experimental validation, with both synthetic signals and real images. The paper ends with Section VII, where some conclusions are drawn and directions for future research are highlighted.

II. Prior art (on watermarking)

An important application scenario of the techniques discussed, which is used as a leading example throughout the paper, is digital multimedia watermarking. In a multimedia watermarking system, a watermark is imperceptibly embedded in a multimedia content in order to, for example, protect the copyright of its owner, hide data, or help in the authentication of that content [17], [18]. Indeed, two general problems are typically discerned in watermarking: watermark detection (aka, zero-bit watermarking), and watermark decoding (aka, multi-bit watermarking). In watermark detection one wants to determine whether a certain watermark has been embedded in the considered content or there is no watermark; therefore, it is formalized as a binary hypothesis test. On the other hand, in watermark decoding the content is assumed to be watermarked using one among multiple possible watermarks, each encoding a different message, and the decoder must decide which watermark has been embedded. Consequently, watermark decoding is a multiple hypothesis test. Both problems have security requirements, meaning that misleading the detector or the decoder must be a hard task for those users who do not have access to a *secret key* used to generate the watermarks, and consequently defines the watermark detection/decoding regions. A particular instance of watermark embedding method, which has been extensively considered in the literature and in practical applications, is the so-called Additive Spread Spectrum (Add-SS) [19], where a pseudorandom signal, independent of the multimedia host, is generated from the secret key and added to the host. The detector in this case is typically based on the correlation between the received content and the watermark, and the comparison of that value with a decision threshold; therefore, the resulting detection region is a hyperplane.

The application scenario addressed in this work is related to watermark detection. In particular, we consider the case where the watermarking system is used for copyright enforcement purposes (e.g., so that only those users in possession of the proper rights can play a film or audio), and users have access to the watermark detector as a black box (i.e., they only see the binary output of the detector: watermarked vs. non-watermarked); an example of this scenario is the DVD copyright system [20]. In that framework, the watermark detection system may be challenged by attackers, who want to reverse the detector decision, i.e., they may want to convert a watermarked content into non-watermarked, or vice versa. In order to do so, the attackers must gain knowledge about the watermark detection region. Although the information reported by a binary black box detector might seem to be rather limited, a number of attacks have been proposed in the literature that repeatedly query the detector in order to produce an illegitimate (i.e., illegally watermarked or non-watermarked) content. They are globally named as oracle attacks.

The most popular oracle attack is the so-called *sensitivity attack*, originally proposed in [2] for attacking the correlation-based detector of Add-SS. The sensitivity attack works by changing one component of the signal at a time and observing the output of the decision function to learn the normal vector that represents the detection region boundary. For more complicated decision boundaries, more sophisticated approaches were later proposed in [21]–[23]. In [3], [15], a powerful variant of the sensitivity attack which implements Newton's descent algorithm to iteratively find a point close to the decision boundary was proposed. The algorithm is completely blind, in the sense that no *a priori* knowledge of the decision function is needed; the information on the first and second order local derivatives required by the iterative algorithm is estimated by querying the detector. The algorithm, termed Blind Newton Sensitivity Attack (BNSA), has been proven to be very effective in removing the watermark and creating forgeries for a number of existing schemes. Blind algorithms have succeeded in removing the watermark for a variety of watermarking algorithms, including those used in the BOWS (Break Our Watermarking System) and BOWS-2 contests [24], [25].

Many solutions have been proposed to counteract oracle attacks. For example, efforts have been made to complicate the

shape of the decision boundary, e.g., by means of *fractalization* [22] and *randomization* [21], [26]. These countermeasures can be easily overcome by an attacker using the 'envelope' of the fractal boundary in one case, or averaging out the boundary randomness in the other. Moreover, the use of intricate decision boundaries typically entails a loss in detection performance. Such decision boundaries are also difficult to parameterize and, consequently, to put to work in practice. Other solutions rely on *zero-knowledge* detectors, which output the one-bit binary decision without revealing any additional information (e.g., on the presence of the watermark [27], [28]). Noticeably, even this minimum disclosure of information is enough for BNSA to succeed.

III. Metadetection of Line Search attacks

In this section we revisit the problem, originally proposed in [10], of detecting LS attacks when either all the queries in the batch considered by the metadetector are produced by a legal user, or all of them come from a dishonest user unaware of the metadetector existence (i.e., no misleading strategies against the metadetector are implemented by the attacker). Then, in Sect. IV we extend this analysis to the case where the attacker is aware of the metadetector, and he performs anti-counterattacking strategies aimed at misleading the metadetector.

A. Notation

Throughout the paper, we use bold letters to denote vectors, e.g., $\mathbf{x}$. Random variables will be denoted by capital letters, e.g., X, whereas bold capital letters will denote random vectors, e.g., $\mathbf{X}$. Given a sequence of random vectors $\mathbf{X}_i \in \mathbb{R}^L$, $i = 1,..,N$, their realizations are denoted by $\mathbf{x}_i$, and we further define $\mathbf{x}^N \in \mathbb{R}^{LN}$ as the vector built by arranging the components of each of those vectors as $\mathbf{x}^N = (x_{1,1},...,x_{N,1},...,x_{1,L},....,x_{N,L})^T$, and $\Upsilon_{\mathbf{x}}$ as the $L \times N$-sized matrix containing the concatenation of the N L-dimensional column vectors $\mathbf{x}_i, i = 1,\dots,N$. Similarly, we denote by $\bar{\Upsilon}_{\mathbf{x}}$ an $L \times N$ matrix which contains the sample mean of $\mathbf{x}_i$, itself denoted as $\bar{\mathbf{x}}_i$, $i = 1,...,N$; further details on the computation of $\bar{\mathbf{x}}_i$ are provided in Sect. IV-B. Consequently, $\Sigma_{\mathbf{x}} = \frac{1}{L-1}(\Upsilon_{\mathbf{x}} - \bar{\Upsilon}_{\mathbf{x}})^T(\Upsilon_{\mathbf{x}} - \bar{\Upsilon}_{\mathbf{x}})$ stands for the $N \times N$ sample covariance matrix between vectors, where each entry (i,j) of $\Sigma_{\mathbf{x}}$ corresponds to the sample covariance between vector $\mathbf{x}_i$ and $\mathbf{x}_j$, that is $\Sigma_{\mathbf{x}}(i,j) = \frac{1}{L-1}(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T(\mathbf{x}_j - \bar{\mathbf{x}}_j)$. Let $\Upsilon_{\mathbf{X}}$ denote the random version of the matrix. Then, $\Sigma_{\mathbf{X}} = \frac{1}{L}\mathrm{E}\left\{(\Upsilon_{\mathbf{X}} - \mathrm{E}\{\Upsilon_{\mathbf{X}}\})^T(\Upsilon_{\mathbf{X}} - \mathrm{E}\{\Upsilon_{\mathbf{X}}\})\right\}$ stands for its statistical counterpart of the covariance matrix. We denote with σ_{ij} the (i,j)-th entry of the matrix, namely the statistical covariance between $\mathbf{X}_i$ and $\mathbf{X}_j$, which has the expression $\sigma_{ij} = \frac{1}{L}\mathrm{E}\left\{(\mathbf{X}_i - \mathrm{E}\{\mathbf{X}_i\})^T(\mathbf{X}_j - \mathrm{E}\{\mathbf{X}_j\})\right\}$. Then, $\Sigma_{\mathbf{X}} = [\sigma_{ij}]_{i,j=1}^N$. The $L \times L$ identity matrix is denoted by $I_{L\times L}$. For any given pair of vectors $\mathbf{x}$ and $\mathbf{y}$, we denote by $\langle \mathbf{x}, \mathbf{y} \rangle$ the scalar product between them, i.e., $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$. Given a sequence of random variables X_i, $i = 1,..,n$, and a random variable X with probability distribution $f(X)$, we use the notation $X_i \xrightarrow{d} f(X)$ to indicate convergence in distribution (or weak convergence) of X_i to X. For two positive sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ stands for asymptotic equality, i.e., $\lim_{n\to\infty} a_n/b_n = 1$. Finally, we denote by $\mathcal{I}$ the indexing set $\{1,2,...,N\}$, and $\mathcal{I}_j$ stands for the set of all the possible subsets of $\mathcal{I}$ of size j, i.e., an element $I \in \mathcal{I}_j$ is a set of j indices chosen from $\mathcal{I}$.

For ease of reading, we list the notation used throughout the paper in Table I.

TABLE I
TABLE OF SYMBOLS.

$\mathbf{x}$	host signal, $\mathbf{x} \in \mathbb{R}^L$
$\sigma_X{}^2$	variance of the host signal
$\mathbf{w}$	watermark sequence, $\mathbf{w} \in \{-\gamma, +\gamma\}^L$
γ	watermark strength
$\mathbf{y}_i$	i-th observed query, $\mathbf{y}_i \in \mathbb{R}^L$ $\mathbf{y}_i = (y_{i,1},...,y_{i,L})^L$
$\bar{\mathbf{y}}_i$	sample mean of $\mathbf{y}_i$
$\Sigma_{\mathbf{y}}$	sample covariance matrix of N queries, $\Sigma_{\mathbf{y}} \in \mathbb{R}^{N\times N}$
$\Sigma_{\mathbf{Y}}$	statistical covariance matrix of N queries, $\Sigma_{\mathbf{Y}} \in \mathbb{R}^{N\times N}$
$\mathbf{y}^N$	vector of N observed queries, $\mathbf{y}^N \in \mathbb{R}^{LN}$ $\mathbf{y}^N = (y_{1,1},...,y_{N,1},...,y_{1,L},....,y_{N,L})^T$
$\mathbf{s}$	indicator sequence ($s_i = 1$, if $\mathbf{y}_i$ is watermarked), $\mathbf{s} \in \{0,1\}^N$
$H_{w,0}/H_{w,1}$ ϕ_w $\mathcal{R}_{w,0}$, $\mathcal{R}_{w,1}$	null/alternative hypothesis of the watermark test decision function of the watermark test decision regions of the watermark test
$H_{q,0}/H_{q,1}$ ϕ_q $\mathcal{R}_{q,0}$, $\mathcal{R}_{q,1}$ $P_{F,q}/P_{M,q}$	null/alternative hypothesis of the metatest decision function of the metatest decision regions of the metatest false positive/negative error probability of the metatest
$\boldsymbol{\mu}_0$	statistical mean of $\mathbf{Y}^N$ under $H_{q,0}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^{LN}$
$\boldsymbol{\mu}_1$	statistical mean of $\mathbf{Y}^N$ under $H_{q,1}$, $\boldsymbol{\mu}_1 \in \mathbb{R}^{LN}$
Σ_0	statistical covariance matrix of $\mathbf{Y}^N$ under $H_{q,0}$, $\Sigma_0 \in \mathbb{R}^{LN\times LN}$
Σ_1	statistical covariance matrix of $\mathbf{Y}^N$ under $H_{q,1}$, $\Sigma_1 \in \mathbb{R}^{LN\times LN}$
$\boldsymbol{\psi}$	vector of the combination weights of the line search $\boldsymbol{\psi} \in [0,1]^N$
σ_N^2	variance of the noise in the line search model
D / $N-D$	number of dummies/ aligned queries
T	threshold value of the metatest
$\lambda_i(A)$	i-th eigenvalue of matrix A

B. Formalization of the metatest

We are interested in constructing a smart detector that is able to decide whether an attacker is launching an oracle-attack to the system. Without any loss of generality, in this work, we formalize the primary test by referring to the watermark detection problem. Subindices w and q are respectively used to make distinction between primary detector and metadetector.

Formally, given a sequence under test $\mathbf{y} \in \mathbb{R}^L$ (where L is the dimensionality of the feature space) and the watermark sequence $\mathbf{w}$, a watermark detector has to decide whether the sequence $\mathbf{y}$ contains the watermark $\mathbf{w}$ (alternative hypothesis $H_{w,1}$ of the watermark detector binary hypothesis test) or not ($H_{w,0}$, null hypothesis). The watermark decision, based on a decision function ϕ_w, splits the space of sequences into two disjoint regions, $\mathcal{R}_{w,0}$ and $\mathcal{R}_{w,1}$. In this paper, we assume that Add-SS is used,[1] so $\mathbf{x}_w \triangleq \mathbf{x}+\mathbf{w}$, where $\mathbf{x}_w$ is the watermarked

[1]The reason for this assumption is mathematical tractability; however, the metadetectors proposed in this paper can be quite straightforwardly extended to most other embedding/detection methods.

signal and the watermark sequence $\mathbf{w}$, independent of $\mathbf{x}$, takes values in $\{-\gamma, +\gamma\}^L$, where γ determines the watermark strength. The optimal detector for Add-SS in the Gaussian i.i.d. case relies on the correlation between the observed sequence $\mathbf{y}$ and the watermark $\mathbf{w}$, that is, on the decision function $\phi_w(\mathbf{y}) \triangleq \langle \mathbf{y}, \mathbf{w} \rangle$. As any binary hypothesis test, the performance of watermark detection is quantified in terms of the probability of false alarm and the probability of missed detection (or equivalently, the false positive and false negative error probabilities, respectively).

Given N consecutive observed queries $\mathbf{y}_i$, $i = 1, .., N$, the metadetector decides whether $\mathbf{y}^N$ is a legitimate (sometimes also referred to as *legal*) batch of queries coming from honest users, or instead the queries $\mathbf{y}^N$ have been generated by an attacker by using a line search (see [10]). The first case corresponds to the null hypothesis of the metadetector test ($H_{q,0}$), where the batch can contain both watermarked and non-watermarked contents, while the second case is the alternative hypothesis ($H_{q,1}$). The corresponding probability density functions will be denoted by $f_{\mathbf{Y}^N|H_{q,0}}$ and $f_{\mathbf{Y}^N|H_{q,1}}$, respectively.

Let ϕ_q be the decision function of the metatest; the detector will output $\phi_q(\mathbf{y}^N) = 0$ if $\mathbf{y}^N$ is deemed a legitimate batch of queries, and 1 otherwise. We denote by $\mathcal{R}_{q,0}$ (respectively, $\mathcal{R}_{q,1}$) the acceptance (resp. rejection) region of the metatest. The false positive and false negative error probabilities of the metatest are denoted by $P_{F,q}$ ad $P_{M,q}$, respectively. By adopting a Neyman-Pearson approach, the metadetector puts a constraint on the false positive error probability (or, alternatively, on the false negative, depending on the application and purpose of the test)

$$P_{F,q} = \int_{\mathcal{R}_{q,1}} f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N|H_{q,0}) d\mathbf{y}^N \leq P^*_{F,q},$$

for some prescribed value $P^*_{F,q}$, and tries to maximize $P_{M,q}$. To get a full statistical characterization of the test we need to model the behavior of the attacker.

In [10], two types of metadetectors are presented, depending on the assumptions made on the attacking strategy. The only assumption made by the CTB-based metadetector is that, in order to succeed, the attacker must submit to the detector a number of queries close to the detection boundary. This consideration leads to a metatest of wide-applicability. In contrast, the LS metadetector exploits the knowledge of the specific methodology adopted by the attacker (that is, a line search), to develop an even more powerful defense. In this paper, we consider this second type of metadetector introduced in [10], i.e., the LS metadetector. Although the applicability of this test is confined to LS attacks (implying a loss of generality with respect to the CTB-based metadetector), it must be noticed that virtually all oracle attacks perform line searches.

C. Line Search metadetector

For the sake of mathematical tractability, hereafter we assume the host signal $\mathbf{x}$ to be Gaussian distributed. Of course more complicated models can be treated, probably by means of numerical analysis.

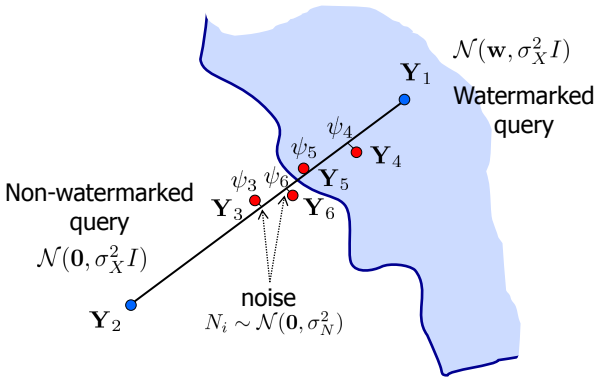


Fig. 1. Graphical illustration of the line search procedure of generation of the attacking queries (metadetector-unaware attack). To ease the graphical representation, the case $L = 2$ is considered.

Definition 1 (Model of legal queries): Legitimate users can submit two kinds of queries, corresponding to watermarked and non-watermarked signals. We model the former by $\mathcal{N}(\mathbf{w}, \sigma_X^2 \mathbf{I}_{L\times L})$, where the watermark $\mathbf{w}$ is known at the detector. On the other hand, non-watermarked signals are assumed to follow a $\mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{L\times L})$. In both cases, we assume the variance $\sigma_X^2 \in \mathbb{R}^+$ to be the same for all the queries, and known by the detector. Reasonably, query signals are assumed to be mutually independent.

We will also find it useful to define the indicator $\mathbf{s}$, whose components are $s_i = 1$ if $\mathbf{y}_i$ is watermarked and $s_i = 0$ otherwise, $i = 1, \dots N$, and the corresponding random variable S_i. Therefore, the null hypothesis is formalized as $H_{q,0} : \mathbf{Y}^N \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, where $\boldsymbol{\mu}_0$ and Σ_0 are, respectively, the mean and the covariance matrix of $\mathbf{Y}^N$ under that hypothesis. Then, each component of the mean vector $\boldsymbol{\mu}_0$ is set to 0 or to the corresponding component of $\mathbf{w}$, depending on whether the query corresponds to a watermarked signal or not, and $\Sigma_0 \triangleq \sigma_X^2 I_{NL\times NL}$. Consequently, the null hypothesis distribution can be parameterized by the indicator $\mathbf{s}$.

By exploiting the knowledge of the attacker strategy, we can define the model of queries under $H_{q,1}$.

Definition 2 (Model of attacking queries (metadetector-unaware attacker)): Query signals corresponding to dishonest users are noisy convex combinations of a watermarked signal, which we denote by $\mathbf{Y}_W \sim \mathcal{N}(\mathbf{w}, \sigma_X^2 \mathbf{I}_{L\times L})$, and a non-watermarked signal, denoted by $\mathbf{Y}_{NW} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{L\times L})$. Specifically, we let $\mathbf{Y}_1 \triangleq \mathbf{Y}_W$ and $\mathbf{Y}_2 \triangleq \mathbf{Y}_{NW}$; then, the i-th query ($i \geq 3$) can be written as $\mathbf{Y}_i \triangleq \psi_i \mathbf{Y}_1 + (1-\psi_i)\mathbf{Y}_2 + \mathbf{N}_i$, where $0 < \psi_i < 1$, $\mathbf{N}_i \sim \mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I}_{L\times L})$, $i = 3, \dots, N$. Since $\mathbf{N}_i$ is typically used for modeling quantization effects in a transform domain, we will assume that: 1) $\mathbf{N}_i$ is independent of $\mathbf{Y}_1$ and $\mathbf{Y}_2$; 2) the $\mathbf{N}_i$ are mutually independent, and identically distributed; and 3) both σ_X^2 and σ_N^2 are known by the detector. We further assume that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are mutually independent. The generation of attacking queries according to the line search procedure is graphically depicted in Fig. 1.

We must remark that in a real framework the location of the watermarked and non-watermarked signals within the batch is

unknown to the detector; therefore, under $H_{q,1}$ the indices corresponding to the line extremes should be included in the set of unknown parameters, jointly with ψ_i, $i = 3, \ldots, N$. Nevertheless, for the sake of notational simplicity we do not model the line extremes differently,[2] but we consider the N aligned queries to follow the same model $\mathbf{Y}_i \sim \mathcal{N}(\psi_i \mathbf{w}, \left[(\psi_i^2 + (1-\psi_i)^2)\sigma_X^2 + \sigma_N^2\right] \mathbf{I}_{L\times L})$, $i = 1, ..., N$. Besides, $\sigma_{ij} = \frac{1}{L}\mathrm{E}\{(\mathbf{Y}_i - \psi_i \mathbf{w})^T \cdot (\mathbf{Y}_j - \psi_j \mathbf{w})\} = (\psi_i \psi_j + (1-\psi_i)(1-\psi_j))\sigma_X^2 + \sigma_N^2 \delta[i-j]$, $i, j = 1, ..., N$, where $\delta[\cdot]$ is the Kronecker delta function ($\delta[a] = 1$ if $a = 0$, 0 otherwise).

Then, the alternative hypothesis can be formalized as $H_{q,1} : \mathbf{Y}^N \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ where mean $\boldsymbol{\mu}_1$ and covariance matrix Σ_1 of Y^N under $H_{q,1}$ and are built as follows: $\boldsymbol{\mu}_1$ contains the mean vectors $\psi_i \mathbf{w}$, $i = 1, \ldots, N$, with the component arranged as in $\mathbf{Y}^N$, that is $\boldsymbol{\mu}_1 \triangleq (\psi_1 w_1, \ldots, \psi_N w_1, \ldots, \psi_1 w_L, \ldots, \psi_N w_L)$; the $LN \times LN$ covariance matrix Σ_1 is a block diagonal matrix, made up of the $N \times N$ statistical covariance matrix $\Sigma_{\mathbf{Y}} = [\sigma_{ij}]_{i,j=1}^N$ repeated L times. Therefore, the alternative hypothesis can be parameterized by $\psi_i \in [0, 1]$, $i = 1, \ldots, N$.

Due to the presence of nuisance parameters, both the null and the alternative hypothesis are *composite hypotheses*, so the Neyman-Pearson criterion cannot be directly applied, and, instead, the *generalized likelihood ratio test* (GLRT) is customarily used. Such test becomes

$$\Lambda(\mathbf{y}^N) \triangleq \log\left(\max_{\boldsymbol{\psi}} f_{\mathbf{Y}^N|H_{q,1}}(\mathbf{y}^N|\boldsymbol{\psi})\right) - \log\left(\max_{\mathbf{s}} f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N|\mathbf{s})\right) \gtrless \tau, \quad (1)$$

and the decision function of the metatest is

$$\phi_q(\mathbf{y}^N) = \begin{cases} 0 & \text{if } \Lambda(\mathbf{y}^N) < \tau \\ 1 & \text{if } \Lambda(\mathbf{y}^N) \geq \tau \end{cases}. \quad (2)$$

By exploiting the nature of the pdf's involved in the current problem, the test statistic $\Lambda(\mathbf{y}^N)$ in (1) can be rewritten as

$$\min_{\mathbf{s}\in\{0,1\}^N} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T \Sigma_0^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|)\right] - \min_{\boldsymbol{\psi}\in[0,1]^N} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T \Sigma_1^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|)\right]. \quad (3)$$

The only remaining issue is to determine the value of the decision threshold τ in order to verify that the false positive error probability when the GLRT is used is smaller than or equal to a target value $P_{F,q}^*$, i.e., $P_{F,q} = P(\Lambda(\mathbf{Y}^N) \geq \tau | H_{q,0}) \leq P_{F,q}^*$. This can be done by upperbounding the test statistic by a function $\Lambda'(\mathbf{y}^N)$ which does not depend on $\mathbf{s}$, i.e., irrespectively of whether a legitimate query corresponds to a watermarked or non-watermarked content. From $\Lambda'(\mathbf{X}^N)$, the threshold τ can be estimated by using Monte Carlo integration (see [10] for the details and the computations).

Experiments in [10] show that such a detector is able to reveal the presence of an oracle attack by observing very few queries. For instance, for a setting with $L = 2 \cdot 10^4$ and $\sigma_X^2/\sigma_N^2 \sim U(8, 12)$, by considering only 3 observed queries (i.e., $N = 3$), Monte Carlo simulations yield a probability of missed detection $P_{M,q} \approx 10^{-50}$ with a prescribed false alarm error probability $P_{F,q}^* \approx 10^{-20}$.

[2]The proposed formalization can be generalized to deal with the detection of the two extreme line queries, but we consider that the substantial increase in notational complexity only provides a minor improvement in the problem insight, so we skip it.

IV. Dummy-aware metadetection

In this section we extend the previous analysis by accounting for the possibility that the smart attacker reacts to the countermeasure adopted against him by the metadetector; in order to do so, the attacker designs a query pattern explicitly thought to reduce the detectability of his attack. Although finding the optimal counterattacking strategy is a hard task, general reasonings on the possible reactions of the attacker allow us to refine the metadetection. We will show both theoretically (Section V) and experimentally (Section VI) that the attacker's task of confusing the metadetector becomes extremely hard.

A. Counterattacking the metadetector

As a consequence of the introduction of a smart detector, we might expect that the attacker will adapt his strategy to react to the coutermeasures adopted by the detector. Then, the interplay between attacker and metadetector should be studied in order to evaluate the final achievable detection performance. Clearly, the best strategy for the attacker depends on the knowledge he has of the metadetector, but even assuming that the attacker knows the specific metadetector he wants to mislead, it is difficult to figure out what would be the best reaction for the attacker. Moreover, being now both parties (detector and attacker) intelligent players, a counterattack will be presumably followed by a higher level of metadetection, that is, the smart detector will, in turn, refine its strategy by assuming that proper countermeasures are taken by the attacker, thus falling into a never-ending loop.

A common and elegant way to address this problem in detection applications is to resort to a game theoretic formulation, which models the interaction between detector and adversary, and study the existence of equilibria (e.g., this is done in [29] for the watermark detection problem and in [30], [31] for a general binary detection problem). A drawback of most existing game-theoretic approaches is that they tend to be overly conservative: the system accounts for the worst case of an attacker trying to minimize in some way the detection performance; then, when an attacker is not present, the performance is highly suboptimal, as a higher correct detection probability could be achieved for a given false alarm rate. A way to escape this problem is to define a metagame, where the presence or absense of the adversary is taken into account within the game formulation. This, however, complicates significantly the analysis. Besides, in our case, the dynamic interaction between the players should be formalized, which accounts for the fact that the attacker may gain knowledge about secret information of the system by repeatedly invoking the detector (dynamic games, [32]). Then, without significantly limiting the freedom of attacker and smart detector, deriving a proper game definition is a

rather complicated task. Instead, in this paper, we make some reasonable considerations about the reaction strategy of the attacker in a possible second iteration of the game (tug of war), in order to refine the metadetection.

Specifically, we consider that a reasonable strategy for the attacker, who does not want to be detected, is to try to mimic as close as possible the behavior of legal users. From this perspective, the attacker will mix his attacking queries with as many legal queries as possible. This is a simple yet effective way for the attacker to make the batch of N queries $\mathbf{y}^N$ to look like a honestly generated batch. To be more specific, such a strategy allows to increase the probability of the attacking batch under $H_{q,0}$, i.e., $f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N)$, and decreases its probability under $H_{q,1}$, i.e., $f_{\mathbf{Y}^N|H_{q,1}}(\mathbf{y}^N)$ with respect to the case in which all the N queries are attacking ones, thus reducing the value of the likelihood ratio that purely attacking sequences would achieve. This would allow the attacker to increase the probability of fooling the metatest, at the price of reducing the oracle attack effectiveness, as a number of queries in the batch will not gather any information. Indeed, one must also consider that the larger the number of queries performed by the oracle attack, the higher the accuracy of the detection boundary; in other words, the average distortion introduced by the attacker in order to illegally remove (or introduce) the watermark, will be smaller for a larger number of queries of the oracle attack. Therefore, for a given target accuracy of the attack (namely, the worst case accuracy of estimation of the detection boundary), introducing legal queries implies that the attacker has to spread out its line search attack over a large number of queries. Then, the attacker has to face a trade off between the distortion introduced by his subsequent watermark removal/embedding attack, the probability of evading the metadetection (not being caught), and the time spread of the attack. [3] It is worth noting that for the attacker the choice of mixing useful with legal queries also avoids the introduction of a new particular pattern that could be, in turn, easily detected by the metadetector in a further iteration of the tug of war. In addition, the idea of concealing the aligned queries within a set of honestly generated queries has some ties with the case where members of a rare class of data are hidden within noise. Indeed, for the problem of detecting large-mean variables within a pool of white Gaussian noise (a variant of the so-called 'needle-in-a-haystack problem' [33], [34]), it has been shown that the detection performance only depends on the number of these large-mean variables relative to the number of total samples; specifically, reducing the relative number of large-mean variables is what makes the detection fail. This connection with the needle-in-a-haystack problem gives further rationale for the attacker's mimicking strategy discussed above. As a final observation, we point out that such a strategy has also some ties with the evasion attacks in intrusion detection systems, where the attacker tries to match the normal behaving profile (e.g., the polymorphic blending attack (PBA) [8] and the mimicry attack [9]).

Then, we assume that the attacker, instead of adopting an 'eager' strategy by generating N queries according to the line search, refines his approach in order to counter the metadetection. He does so by generating a batch of N queries, where $(N-D)$ 'useful' attacking queries are mixed with D queries generated according to $H_{q,0}$.[4] We will refer to such 'no information gathering' queries as *dummies*. Alternatively, the attacker might share the oracle with legitimate users, both problems leading to the same formulation.

B. Dummy-aware Line Search metadetector

We generalize the analysis carried out in Sect. III-C to the case in which some of the queries made by the attacker are generated according to the line search procedure, and some follow the distribution of legal queries. Accordingly, we must redefine the model under the alternative hypothesis (model of attacking queries) introduced in Sect. III-C:

Definition 3 (Model of attacking queries (metadetector-aware attacker)): D out of N queries follow $\mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I}_{L\times L})$ (non-watermarked) or $\mathcal{N}(\mathbf{w}, \sigma_X^2\mathbf{I}_{L\times L})$ (watermarked); they are mutually independent, and independent of the signals we define next. Out of the remaining $(N-D)$ queries, one of them, denoted by $\mathbf{Y}_{NW}$, follows $\mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I}_{L\times L})$, while another, denoted by $\mathbf{Y}_W$, follows $\mathcal{N}(\mathbf{w}, \sigma_X^2\mathbf{I}_{L\times L})$; they respectively correspond to a non-watermarked and a watermarked independent signals that define the extremes of the line search. The remaining $(N-D-2)$ attacking queries are generated according to the line search procedure. Specifically, a generic non-extreme attacking query $\mathbf{Y}_i$ follows the model $\mathbf{Y}_i \triangleq \psi_i\mathbf{Y}_W + (1-\psi_i)\mathbf{Y}_{NW} + \mathbf{N}_i$, where $0 < \psi_i < 1$, and the noise random variables $\mathbf{N}_i \sim \mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I}_{L\times L})$ are i.i.d. and independent of $\mathbf{Y}_{NW}$, $\mathbf{Y}_W$, and ψ_i. Consequently, $\mathbf{Y}_i \sim \mathcal{N}(\psi_i\mathbf{w}, \left[(\psi_i^2+(1-\psi_i)^2)\sigma_X^2+\sigma_N^2\right]\mathbf{I}_{L\times L})$, and $\sigma_{ij} = \frac{1}{L}\mathrm{E}\{(\mathbf{Y}_i-\psi_i\mathbf{w})^T\cdot(\mathbf{Y}_j-\psi_j\mathbf{w})\} = L(\psi_i\psi_j+(1-\psi_i)(1-\psi_j))\sigma_X^2+\sigma_N^2\delta[i-j]$.

Note that for the sake of simplicity we focus on the detection of a single line search with dummy queries.

Let $I \in \mathcal{I}_{N-D}$ be the index set of the $(N-D)$ attacking queries; hence, $\overline{I} = \mathcal{I}\setminus I$ is the set of the indices of the honest queries. Similarly to Definition 2, in a real scenario the location of the line extremes within the batch is not known to the metadetector. Therefore, in the subsequent GLRT analysis, and for the sake of notational simplicity, the pdf under the alternative hypothesis is parameterized by the number j of aligned queries, $3 \le j \le N$, the set of indices corresponding to those queries $I \in \mathcal{I}_j$, the combination parameters $\psi_i \in [0,1]$ for $i \in I$, and the indicators $s_i \in \{0,1\}$ for $i \in \bar{I}$, where $\bar{I}$ is the complementary set of I, i.e., the index set of the dummies. We remark that, according to the proposed formalism, we skip the determination of the line search extremes in the GLRT analysis, which would have significantly increased the notational complexity. Note that legal and aligned queries are interspersed in an arbitrary way; more formally, each observed

[3] In many practical situations, as in many applications of digital audio/video watermarking, the real-time nature of the applications imposes constraints on the delay of the attack. In all these cases, having to spread the attack over many batches of queries may be critical.

[4] Note that our analysis assumes that the attacker knows N, thus considering a favorable situation for him; in any case, the strategy of mixing attacking queries with dummy ones does not require the attacker to know that value.

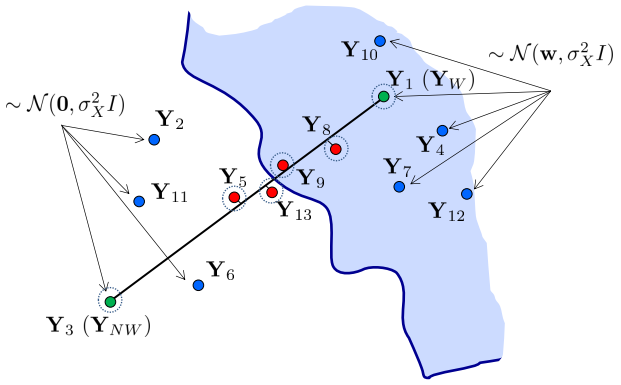


Fig. 2. Graphical illustration of the attacking strategy according to the model in Definition 3 (metadetector-aware attack) with $L = 2$. In the depicted case, $\mathbf{Y}_d^D = \{\mathbf{Y}_2, \mathbf{Y}_4, \mathbf{Y}_6, \mathbf{Y}_7, \mathbf{Y}_{10}, \mathbf{Y}_{11}, \mathbf{Y}_{12}\}$ and $\mathbf{Y}_a^{N-D} = \{\mathbf{Y}_1, \mathbf{Y}_3, \mathbf{Y}_5, \mathbf{Y}_8, \mathbf{Y}_9, \mathbf{Y}_{13}\}$.

query $\mathbf{Y}_i$ is an attacking query if $i \in I$, for some arbitrary $I \in \mathcal{I}_{N-D}$, and a dummy query otherwise. Let $\mathbf{Y}_d^D$ be the vector containing the legal queries, and $\mathbf{Y}_a^{N-D}$ the vector containing the aligned queries. Fig. 2 illustrates the procedure of generation of attacking queries according to the model given in Definition 3. In the example considered, $D = 7$, $N-D = 6$, $\mathbf{Y}_W = \mathbf{Y}_1$, $\mathbf{Y}_{NW} = \mathbf{Y}_3$ and $I = \{1, 3, 5, 8, 9, 13\}$. The DL-length vector $\mathbf{Y}_d^D$ is then modeled according to Definition 1, that is $\mathbf{Y}_d^D \sim \mathcal{N}(\boldsymbol{\mu}_{1d}, \sigma_X^2 I_{DL\times DL})$, where each component of $\boldsymbol{\mu}_{1d}$ is set to 0 or to the corresponding component of $\mathbf{w}$, depending on whether the query corresponds to a watermarked signal or not, whereas $\mathbf{Y}_a^{N-D}$ is a vector of length $(N-D)L$ which follows the model introduced in Sect. III-C. Hence, $\mathbf{Y}_a^{N-D} \sim \mathcal{N}(\boldsymbol{\mu}_{1a}, \Sigma_{1a})$ where the mean vector $\boldsymbol{\mu}_{1a}$ contains the mean values $\psi_i \mathbf{w}$, $i = 1, ..., N-D$ and Σ_{1a} is a block diagonal matrix where each block of size $(N-D)\times(N-D)$, repeated L times, corresponds to the covariance matrix of the aligned queries, namely, $\Sigma_{\mathbf{Y}_a} = [\sigma_{ij}]_{i,j\in I}$.

Accordingly, $\mathbf{Y}^N \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ where $\boldsymbol{\mu}_1$ is obtained from $\boldsymbol{\mu}_{1d}$ and $\boldsymbol{\mu}_{1a}$ by picking the elements according to the query ordering in $\mathbf{Y}^N$ (i.e., according to the index sets I and $\overline{I}$), and matrix Σ_1 is a block diagonal matrix where each block of size $N \times N$ corresponds to the covariance matrix of the queries in $\mathbf{Y}^N$ and it is repeated L times. Therefore, each block of Σ_1 is a reordering of the $D \times D$ block diagonal matrix made up of the subblock $\sigma_X^2 I_{D\times D}$ (i.e., covariance matrix of the dummy queries) and the $(N-D)\times(N-D)$ matrix $\Sigma_{\mathbf{Y}_a} = [\sigma_{ij}]_{i,j\in I}$ (i.e., covariance matrix of the attacking queries), where the reordering goes according to the position of the queries in $\mathbf{Y}^N$.

Consequently, the GLRT can be written in this case as

$$\Lambda(\mathbf{y}^N) = \log\left(\max_{\boldsymbol{\theta}_1} f_{\mathbf{Y}^N|H_{q,1}}(\mathbf{y}^N|\boldsymbol{\theta}_1)\right) - \log\left(\max_{\mathbf{s}\in\{0,1\}^N} f_{\mathbf{Y}^N|H_{q,0}}(\mathbf{y}^N|\mathbf{s})\right) \gtrless \tau,$$

where $\boldsymbol{\theta}_1$ is the concatenation of j (with $j \in \{3, \dots, N\}$), the index set $I \in \mathcal{I}_j$, the combination parameters ψ_i, for $i \in I$, and the indicators s_i, for $i \in \overline{I}$.

Thus, for the Gaussian case

$$\Lambda(\mathbf{y}^N) = \tag{4}$$
$$\min_{s_i\in\{0,1\}, i=1,\dots,N} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|)\right]$$
$$- \min_{j=3,\dots,N} \min_{I\in\mathcal{I}_j} \min_{s_i\in\{0,1\}, i\in\overline{I}} \min_{\psi_i\in[0,1], i\in I}$$
$$\frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|)\right].$$

Recall that $\boldsymbol{\mu}_0$ depends on $s_i, i = 1, \dots, N$ (but Σ_0 does not), while Σ_1 depends on j, I and $\psi_i, i \in I$, and $\boldsymbol{\mu}_1$ depends on the three latter, and on $s_i, i \in \overline{I}$. It is worth mentioning that (4) is designed in order to cope with a single line search; this approach can be generalized to the case where multiple line searches are simultaneously performed.

Note that the value of (4) does not depend on $s_i \in \{0, 1\}, i \in \overline{I}$, as the contribution of those s_i is the same in both target functions and, consequently, they cancel each other out. Hence, (4) can be rewritten as

$$\Lambda(\mathbf{y}^N) = \max_{j=3,\dots,N} \max_{I\in\mathcal{I}_j} \tag{5}$$
$$\min_{s_i\in\{0,1\}, i\in I} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|)\right]$$
$$- \min_{\psi_i\in[0,1], i\in I} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|)\right],$$

which indeed can be solved in two steps by first computing

$$I^* \triangleq \arg \min_{I\in\mathcal{I}_j, j=3,..,N} \tag{6}$$
$$\min_{\psi_i\in[0,1], i\in I} \left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|)\right],$$

and then

$$\min_{s_i\in\{0,1\}, i\in I^*} \left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|)\right]. \tag{7}$$

Furthermore, note that if I were known at the metadetector, then (5) would be equivalent to

$$\min_{s_i\in\{0,1\}, i\in I} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_0) + \log(|\Sigma_0|)\right]$$
$$- \min_{\psi_i\in[0,1], i\in I} \frac{1}{2}\left[(\mathbf{y}^N - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{y}^N - \boldsymbol{\mu}_1) + \log(|\Sigma_1|)\right],$$

which, as one would expect, is equivalent to (3) once the indices in the optimization set used there (i.e., $\{1, \dots, N\}$) are replaced by I. This means that the line search detection problem with dummy queries, *when the observation indices on the line are known to the metadetector*, is the same as the line search detection problem with no dummy queries, but taking into account that now there are $|I| = N - D$ observations, instead of N. Of course, this is not a realistic situation, and in practice the two-step minimization in (6)-(7) must be performed.

Solving (6) is computationally demanding, as the first minimization requires an exhaustive search over all the possible sets of j elements out of N possibilities (there are $\binom{N}{j}$ of those sets), and also one exhaustive search over j, for a total of

$$\sum_{j=3}^{N} \binom{N}{j} = 2^N - 1 - N - \frac{N(N-1)}{2}$$

choices. Then, we consider that j-dimensional optimizations should be performed for each set in $\mathcal{I}_j$, and that a j-dimensional optimization problem is harder to solve than j one-dimensional optimization problems (as the multidimensional problem has to deal with the combined effect of those variables); consequently, we can state that the total complexity of (6) is larger than the complexity of solving

$$\sum_{j=3}^{N} j\binom{N}{j} = N/2(2^N - 2N)$$

real scalar optimizations in $[0, 1]$. Just for the sake of illustration, note that for $N = 50$ (which is a very small value for practical applications)[5] such number is about 2.815×10^{16}; even assuming that the rate of solving those problems is 10^6 problems per second, this means that more than 892 years of computation would be required for solving this toy example. Consequently, for practical values of N this is not a computationally feasible problem.

Since a metatest based on the GLRT is hard to implement, alternative metadetectors will be proposed next. A common characteristic of these metadetectors is that they try to exploit the fact that if $\sigma_X^2 \gg \sigma_N^2$, then under $H_{q,1}$ the $\mathbf{Y}_i$, $i = 1, \ldots, N$ will approximately lie on a $(D+2)$-dimensional subspace, while under $H_{q,0}$ they will lie on an N-dimensional one. In order to quantify this effect, we focus on the sample covariance matrix of the queries $\Sigma_{\mathbf{y}}$. For computing $\bar{\Upsilon}_{\mathbf{y}}$, the detector exploits that $\text{sign}(w_j)y_{ij} \sim \mathcal{N}(\gamma, \sigma_X^2)$ for watermarked queries, and $\text{sign}(w_j)y_{ij} \sim \mathcal{N}(0, \sigma_X^2)$, $j = 1, ..., L$, for non-watermarked queries. Then, the sample mean of $\mathbf{y}_i$ is calculated as $\bar{\mathbf{y}}_i \triangleq \text{sign}(\mathbf{w})\frac{1}{L}\sum_{j=1}^{L}\text{sign}(w_j)y_{ij}$, which is used in the computation of the sample covariance matrix between queries. Note that, in the above calculations, we assume that the watermark strength γ does not depend on L.

Specifically, we propose to use the following metadetectors based on statistics obtained from the sample covariance matrix $\Sigma_{\mathbf{y}}$:

- *Determinant-based Metadetector (DM)*: defined as $\Lambda_{\text{DM}}(\mathbf{y}^N) \triangleq |\Sigma_{\mathbf{y}}|$. Under $H_{q,1}$, and due to the closeness of some of the $\mathbf{y}_i$ to a linear subspace discussed above, one would expect $\Lambda_{\text{DM}}(\mathbf{y}^N)$ to be much smaller than under $H_{q,0}$. Specifically, by letting $s(n, k) \triangleq n(n-1)\cdots(n-k+1)$, according to [35] (Section V) and [36], as long as $\lim_{L\to\infty}\frac{N}{L} = p$, $0 < p < 1$, the statistic
$$\frac{|\Sigma_{\mathbf{y}}|}{|\Sigma_{\mathbf{Y}}|}\frac{(L-1)^N}{s(L-2, N)}$$
converges in distribution to a log-normal distribution with null mean and variance $-2\log(1-p)$, as $L \to \infty$. Note that under $H_{q,0}$, $|\Sigma_{\mathbf{Y}}| = \Sigma_0^{1/L} = \sigma_X^{2N}$, while under $H_{q,1}$ $|\Sigma_{\mathbf{Y}}| = \Sigma_1^{1/L} = \sigma_X^{2(D+2)}\sigma_N^{2(N-D-2)}$ (cf. Sect. V-B).
- *Sphericity-based Metadetector (SM)*: defined as $\Lambda_{\text{SM}}(\mathbf{y}^N) \triangleq \log(|\Sigma_{\mathbf{y}}|) - N\log\left[\frac{\text{tr}(\Sigma_{\mathbf{y}})}{N}\right]$ [37], in such a way that under $H_{q,0}$ $\Lambda_{\text{SM}}(\mathbf{y}^N)$ will be close to 0, while under $H_{q,1}$ it will be negative.
- *Smallest Eigenvalue-based Metadetector (SEM)*: defined as $\Lambda_{\text{SEM}}(\mathbf{y}^N) \triangleq \min_{i=1,\ldots,N}\lambda_i(\Sigma_{\mathbf{y}})$, where $\lambda_i(A)$ is the ith eigenvalue of square matrix A, i.e., the metadetector considers the smallest eigenvalue of the sample covariance matrix. Under $H_{q,0}$ that value is expected to be larger than under $H_{q,1}$. A straightforward way of implementing this metadetector is by computing the Singular Value Decomposition (SVD) of the sample covariance matrix. Nevertheless, since not all the eigenvalues, but just the smallest one, is needed, alternative implementations exist that allow to alleviate the computational burden of the SVD implementation; this is achieved, for example, by computing the inverse of the sample covariance matrix and then running an off-the-shelf algorithm (e.g., the power method [38]) for numerically computing the largest eigenvalue of the resulting matrix.[6] Alternative designs might provide larger computational savings.
Be aware that the statistical characterization of this test (in terms of false positive and false negative probabilities) requires to model the distribution of the smallest eigenvalue of the sample covariance matrix of non-independent observations, which, to the best of our knowledge, is an open problem. Therefore, for the sake of analytical tractability, the next metadetector is also proposed.
- *Genie-Aided Metadetector (GAM)*: defined as $\Lambda_{\text{GAM}}(\mathbf{y}^N) \triangleq \min\{\text{diag}(U^T \cdot \Sigma_{\mathbf{y}} \cdot U)\}$, where $\text{diag}(A)$ stands for the diagonal elements of matrix A, and U is the orthogonal matrix that diagonalizes Σ_1. Indeed, if one calculates $U^T \cdot \Sigma_{\mathbf{Y}} \cdot U$ under both hypotheses, then a diagonal matrix is obtained with the eigenvalues of $\Sigma_{\mathbf{Y}}$. Instead, if we consider the diagonal elements of $U^T \cdot \Sigma_{\mathbf{y}} \cdot U$, it is known that their limiting distribution (as L tends to infinity) is normal with mean $\lambda_i(\Sigma_{\mathbf{Y}})$ and variance $\frac{2}{(L-1)}\lambda_i^2(\Sigma_{\mathbf{Y}})$ and the obtained diagonal values are mutually independent ([37, Theorem 3.4.4]). Be aware, in any case, that U depends on which queries are involved in the line search (and also on the corresponding ψ_i), and in practice that information would not be available to the metadetector. For this reason, we term this metadetector *Genie-aided*. Although this metadetector is not realistic, we find it useful in order to derive theoretical performance bounds on eigenvalue-based metadetectors (as SEM).

V. Parameter analysis and limiting performance of the metadetection

A question that arises when defining the metadetector is how to choose the batch size N. Since the defender does not have information on how the attacker distributes the useful attacking

[5]Note that in practical applications the dimensionality of the watermark detection features L is in the order of thousands or tens of thousands to achieve robustness [2]; consequently, for a given fixed accuracy (i.e., for a fixed value of $N - D$), the detection of the search lines will be an easy task for the metadetector. Therefore, the attacker must introduce a large number of dummy queries D, producing large values of N, in order to mislead the metadetector decision. A more formal analysis regarding this issue can be found in Sect. V.

[6]For the sake of illustration, in a Core i5-4670 3.4GHz 20 GB RAM, the average computing time of a Monte Carlo simulation for $L = 100$, $N = 80$, $\sigma_X^2 = 10$, and $\sigma_N^2 = 1$, is about 0.623 ms for SVD and 0.260 ms for the alternative method, while the obtained detection results are virtually the same.

queries among the dummy ones, choosing a proper value for N is a difficult and critical issue: with a too large N, the defender risks that the oracle attack may succeed within those N queries, thus failing in timely detecting it; on the other hand, a too small N might cause that only few attacking queries are made in each batch, so the attack can be pursued over a number of batches. To address this problem, it is interesting to study the behavior of the metatest with respect to N.

In this section we analyze some of the methods proposed in Section IV-B for the dummy-aware line search metadetector. Specifically, we provide a theoretical analysis of the performance for DM and GAM. Interestingly, for the latter we will prove that even a line search of only three queries is detectable within the pool of dummy queries, for any N below a critical value which grows exponentially with the dimensionality L.

A. DM analysis

In the case of DM, a theoretical analysis can be carried out by exploiting the knowledge of the statistics of the determinant of the sample covariance matrix. According to [36, Theorem 1], for the case $\lim_{L\to\infty} N/L = p$, $0 < p < 1$, the following limit holds as $L \to \infty$:

$$\frac{\log|\Sigma_{\mathbf{y}}| - \log|\Sigma_{\mathbf{Y}}| - \tau_{L,N}}{\sigma_{L,N}} \xrightarrow{d} \mathcal{N}(0,1), \tag{8}$$

where $\tau_{L,N} \triangleq \sum_{k=1}^{N}\log(1-k/L)$ and $\sigma_{L,N} \triangleq \sqrt{-2\log(1-N/L)}$. From the knowledge of the value of the determinant of the statistical covariance matrix under the two hypotheses, it is easy to get an asymptotic evaluation for the two error probabilities. Focusing on the false positive probability, for large L we can write:[7]

$$\begin{aligned} P_{F,q} =& \Pr\{|\Sigma_{\mathbf{y}}| < T | H_{q,0}\} \\ =& \Pr\{\log|\Sigma_{\mathbf{y}}| < \log T | H_{q,0}\} \\ \doteq& 1 - Q\left(\frac{\log T - \left[\log\sigma_X^{2N} + \sum_{k=1}^{N}\log(1-k/L)\right]}{\sqrt{-2\log(1-N/L)}}\right) \\ =& 1 - Q\left(\frac{\log T' - \log\sigma_X^{2N}}{\sqrt{-2\log(1-N/L)}}\right), \end{aligned} \tag{9}$$

where $\log T' \triangleq \log T - \sum_{k=1}^{N}\log(1-k/L)$, and $Q(\cdot)$ is the Q-function (i.e., the tail probability of the standard normal distribution).

Similarly, for the false negative probability we have:

$$\begin{aligned} P_{M,q} =& \Pr\{|\Sigma_{\mathbf{y}}| \geq T | H_{q,1}\} \\ =& \Pr\{\log|\Sigma_{\mathbf{y}}| \geq \log T | H_{q,1}\} \\ \doteq& Q\left(\frac{\log T' - \log[\sigma_X^{2(D+2)}\sigma_N^{2(N-D-2)}]}{\sqrt{-2\log(1-N/L)}}\right). \end{aligned} \tag{10}$$

[7]Throughout this section, we exploit the fact that the error probabilities can be considered as sequences indexed by L; then, we consider the asymptotic behavior of those sequences when L goes to infinity.

From the above expressions, we can derive necessary and sufficient conditions for the two error probabilities tending to zero as $L \to \infty$, which are:

$$\lim_{L\to\infty} \frac{\log\sigma_X^{2N} - \log T'}{\sqrt{-2\log\left(1-\frac{N}{L}\right)}} = \infty, \quad \text{and} \tag{11}$$

$$\lim_{L\to\infty} \frac{\log T' - \log\sigma_X^{2(D+2)}\sigma_N^{2(N-D-2)}}{\sqrt{-2\log\left(1-\frac{N}{L}\right)}} = \infty. \tag{12}$$

Since we assume that for large L the number of observations N grows linearly with L, the denominator tends to a constant (i.e., $\sqrt{-2\log(1-p)}$). Then, under that assumption, from (11) and (12) it is easy to check that the simultaneous verification of the two following conditions

$$\lim_{L\to\infty} N - D = \infty, \quad \text{and} \tag{13}$$

$$\sigma_N^2 < \left(\frac{T'}{\sigma_X^{2(D+2)}}\right)^{1/(N-D-2)} < \sigma_X^2 \tag{14}$$

is indeed a sufficient condition for both error probabilities to asymptotically go to 0. Then, a proper choice of the threshold T can be made such that an asymptotical powerful test is achieved if (13) holds. Note that the derived condition is sufficient, and it was obtained under the assumption of linear growth of N as a function of L; consequently, less restrictive conditions might be derived. In any case, from the perspective of the metadetector this condition is less demanding than the condition found for the CTB-based metadetector in [16], illustrating the advantage in terms of detectability of using LS instead of CTB-based metadetectors whenever a Line Search is indeed run; in fact, while here the sufficient condition only requires that the number of aligned queries goes to infinity with the dimensionality, the sufficient condition for the CTB-based metadetector establishes a minimum growth rate for the number of attacking queries. In fact, unless the number of attacking queries increases at least logarithmically with L, the attack is not guaranteed to be discovered by the CTB-based metadetector.

B. GAM analysis

We now focus on the genie-aided metadetector. The test statistic is obtained from matrix $U^T\Sigma_{\mathbf{y}}U$, by considering the diagonal elements and looking for the smallest. Since the statistical matrix $U^T\Sigma_{\mathbf{Y}}U$ is diagonal under both hypotheses, the diagonal elements of matrix $U^T\Sigma_{\mathbf{y}}U$ are asymptotically mutually independent and normally distributed [37, Theorem 3.4.4]. Formally, for large L,

$$\{U^T\Sigma_{\mathbf{y}}U\}_{ii} \xrightarrow{d} \mathcal{N}\left(\lambda_i(\Sigma_{\mathbf{Y}}), \frac{2}{(L-1)}\lambda_i^2(\Sigma_{\mathbf{Y}})\right). \tag{15}$$

From the above discussion, the evaluation of the error probabilities for large values of L becomes an easy task. Let

$V \triangleq U^T \Sigma_{\mathbf{y}} U$. Then, the asymptotic behavior with respect to L of $P_{F,q}$ is:[8]

$$\begin{aligned} P_{F,q} =& \Pr\{\min_i v_{ii} \leq T | H_{q,0}\} \\ =& 1 - \Pr\{v_{11} > T, v_{22} > T, ..., v_{NN} > T | H_{q,0}\} \\ \doteq& 1 - \prod_{i=1}^{N} \Pr\{v_{ii} > T | H_{q,0}\} \\ \doteq& 1 - Q\left(\frac{\sqrt{L-1}(T-\sigma_X^2)}{\sqrt{2}\sigma_X^2}\right)^N . \end{aligned} \quad (16)$$

In order to achieve an asymptotically zero false positive error probability, $T < \sigma_X^2$ must hold. For any given threshold satisfying this condition, going on from (16) we can write

$$\begin{aligned} P_{F,q} &\doteq 1 - \left(1 - Q\left(\frac{\sqrt{L-1}(|T-\sigma_X^2|)}{\sqrt{2}\sigma_X^2}\right)\right)^N \\ &\leq 1 - \left(1 - e^{-\frac{(L-1)(T-\sigma_X^2)^2}{4\sigma_X^4}}\right)^N \\ &= \sum_{k=1}^{N} \binom{N}{k} e^{-\frac{k(L-1)(T-\sigma_X^2)^2}{4\sigma_X^4}} \\ &\leq \sum_{k=1}^{N} \left(N e^{-\frac{(L-1)(T-\sigma_X^2)^2}{4\sigma_X^4}}\right)^k \\ &\doteq e^{-\frac{(L-1)(T-\sigma_X^2)^2}{4\sigma_X^4} + \log N}, \end{aligned} \quad (17)$$

where we have exploited the Chernoff bound of the Q function, i.e., $Q(x) \leq e^{-x^2/2}$, the binomial expansion, and the fact that in the case of interest $\lim_{L\to\infty} \exp\{-(L-1)(T-\sigma_X^2)^2/4\sigma_X^4 + \log N\} = 0$, and consequently the sum over k is asymptotically equivalent to consider just $k = 1$. Therefore, a sufficient condition for $P_{F,q}$ to go to 0 is that

$$T < \sigma_X^2, \quad \text{and} \quad (18)$$

$$\lim_{L\to\infty} \left(\frac{(L-1)(T-\sigma_X^2)^2}{4\sigma_X^4} - \log N\right) = \infty \quad (19)$$

simultaneously hold.

Let us now focus on the probability of a false negative error $P_{M,q}$. In order to evaluate $P_{M,q}$, we need to determine the eigenvalues of the statistical covariance matrix between queries under H_1, i.e., $\Sigma_{\mathbf{Y}}$. We know that D of those eigenvalues will be equal to σ_X^2. To characterize the remaining $N - D$ eigenvalues, corresponding to the aligned queries, we can resort to Weyl's inequality [39]. According to Definition 3, the covariance matrix for the aligned queries, that is $\Sigma_{\mathbf{Y}_a} = [\sigma_{ij}]_{i,j\in I}$, is given by

- $\sigma_{1,1} = \sigma_{2,2} = \sigma_X^2$.
- $\sigma_{1,2} = \sigma_{2,1} = 0$.
- $\sigma_{1,j} = \sigma_{j,1} = \psi_j \sigma_X^2$, $3 \leq j \leq N - D$.
- $\sigma_{2,j} = \sigma_{j,2} = (1 - \psi_j)\sigma_X^2$, $3 \leq j \leq N - D$.
- $\sigma_{i,i} = [\psi_i^2 + (1 - \psi_i)^2]\sigma_X^2 + \sigma_N^2$, $3 \leq i \leq N - D$.
- $\sigma_{i,j} = \sigma_{j,i} = [\psi_i \psi_j + (1 - \psi_i)(1 - \psi_j)]\sigma_X^2$, $3 \leq i \neq j \leq N - D$.

[8]Note that the independence between v_{ii} only holds asymptotically, when L goes to infinity.

We observe that, since $\Sigma_{\mathbf{Y}_a}$ is symmetric, it can be rewritten as $\Sigma_{\mathbf{Y}_a} = RDR^T$, where D is the diagonal matrix

$$D \triangleq \left(\begin{array}{c|c} \sigma_X^2 I_{2\times 2} & 0_{2\times(N-D-2)} \\ \hline 0_{(N-D-2)\times 2} & \sigma_N^2 I_{(N-D-2)\times(N-D-2)} \end{array}\right) \quad (20)$$

and R, which diagonalizes $\Sigma_{\mathbf{Y}_a}$, is the lower triangular matrix

$$R \triangleq \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ \psi_3 & (1-\psi_3) & 1 & 0 & \dots & 0 \\ \psi_4 & (1-\psi_4) & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_{N-D} & (1-\psi_{N-d}) & 0 & 0 & \dots & 1 \end{pmatrix} \quad (21)$$

From matrix theory, the determinant of $\Sigma_{\mathbf{Y}_a}$ corresponds to that of the inner diagonal matrix D, which is $\sigma_X^4 \sigma_N^{2(N-D-2)}$. In order to exploit Weyl's inequality, we write $\Sigma_{\mathbf{Y}_a}$ as the sum of two covariance matrices, namely Γ and Θ (relative to the signal and noise contributions, respectively). Specifically, $\Sigma_{\mathbf{Y}_a} = \Gamma + \Theta$, where the only non-zero elements of Θ are $\theta_{ii} = \sigma_N^2$, $3 \leq i \leq N - D$, and Γ is the denoised version of $\Sigma_{\mathbf{Y}_a}$ (i.e., where we set σ_N^2 to 0).

From the above rewriting, it is easy to infer that $N - D - 2$ eigenvalues of Γ are null; the other two are equal to σ_X^2. Also straightforwardly, Θ has two null eigenvalues and $N - D - 2$ eigenvalues equal to σ_N^2. Ordering the eigenvalues in decreasing order, and applying Weyl's inequality, we get:

- $\lambda_i(\Sigma_{\mathbf{Y}_a}) \geq \lambda_i(\Gamma) + \lambda_{N-D}(\Theta) = \lambda_i(\Gamma)$ and $\lambda_i(\Sigma_{\mathbf{Y}_a}) \leq \lambda_i(\Gamma) + \lambda_1(\Theta) = \lambda_i(\Gamma) + \sigma_N^2$, so
  - $\sigma_X^2 \leq \lambda_1(\Sigma_{\mathbf{Y}_a}) \leq \sigma_X^2 + \sigma_N^2$,
  - $\sigma_X^2 \leq \lambda_2(\Sigma_{\mathbf{Y}_a}) \leq \sigma_X^2 + \sigma_N^2$, and
  - $0 \leq \lambda_i(\Sigma_{\mathbf{Y}_a}) \leq \sigma_N^2$, $3 \leq i \leq N - D$.
- $\lambda_i(\Sigma_{\mathbf{Y}_a}) \geq \lambda_i(\Theta) + \lambda_{N-D}(\Gamma) = \lambda_i(\Theta)$, so
  - $\lambda_i(\Sigma_{\mathbf{Y}_a}) \geq \sigma_N^2$, $1 \leq i \leq N - D - 2$.

By combining the inequalities, it is straightforward to conclude that $\lambda_i(\Sigma_{\mathbf{Y}_a}) = \sigma_N^2$ for $3 \leq i \leq N - D - 2$ (i.e., $N - D - 4$ eigenvalues are equal to σ_N^2); two of the remaining ones are smaller than or equal to σ_N^2, and the last two lie in the interval $[\sigma_X^2, \sigma_X^2 + \sigma_N^2]$. By exploiting the value of the determinant of $\Sigma_{\mathbf{Y}_a}$ discussed above, we have that $\lambda_1(\Sigma_{\mathbf{Y}_a}) \cdot \lambda_2(\Sigma_{\mathbf{Y}_a}) \cdot \lambda_{N-D-1}(\Sigma_{\mathbf{Y}_a}) \cdot \lambda_{N-D}(\Sigma_{\mathbf{Y}_a}) = \sigma_X^4 \sigma_N^4$.

Let μ_1 and μ_2 denote the two small eigenvalues and μ_3 and

μ_4 denote the large ones. Then, from (15) we can write:

$$\begin{aligned}
P_{M,q} =&\Pr\{\min_i v_{ii} \geq T|H_{q,1}\}\\
=&\Pr\{v_{11} \geq T, v_{22} \geq T, ..., v_{NN} \geq T|H_{q,1}\}\\
\doteq&\prod_{i=1}^{N}\Pr\{v_{ii} \geq T|H_{q,1}\}\\
\doteq&Q\left(\frac{\sqrt{L-1}(T-\sigma_X^2)}{\sqrt{2}\sigma_X^2}\right)^D \cdot \prod_{i=1}^{4} Q\left(\frac{\sqrt{L-1}(T-\mu_i)}{\sqrt{2}\mu_i}\right)\\
&\cdot Q\left(\frac{\sqrt{L-1}(T-\sigma_N^2)}{\sqrt{2}\sigma_N^2}\right)^{N-D-4}\\
\leq&Q\left(\frac{\sqrt{L-1}(T-\sigma_X^2)}{\sqrt{2}\sigma_X^2}\right)^D\\
&\cdot Q\left(\frac{\sqrt{L-1}(T-\sigma_N^2)}{\sqrt{2}\sigma_N^2}\right)^{N-D-2}\\
\leq&Q\left(\frac{\sqrt{L-1}(T-\sigma_N^2)}{\sqrt{2}\sigma_N^2}\right)^{N-D-2}\\
\leq&e^{-\frac{(N-D-2)(L-1)(T-\sigma_N^2)^2}{4\sigma_N^4}},
\end{aligned} \tag{22}$$

where in the first inequality we exploited the fact that the Q terms of the product are monotonically increasing with the eigenvalue μ_i; hence, the two terms corresponding to the small eigenvalues μ_1 and μ_2 can be upper-bounded by replacing them by σ_N^2. The remaining two terms corresponding to the large eigenvalues are upper bounded by 1.

Therefore, a sufficient condition for achieving an asymptotically zero false negative error probability is that the inequalities

$$T > \sigma_N^2, \quad \text{and} \tag{23}$$

$$N - D > 2 \tag{24}$$

(we remind that the necessary condition $T < \sigma_X^2$ was set in order to have $P_{F,q}$ asymptotically going to zero) simultaneously hold.[9] Consequently, if (24) and

$$\frac{\log N}{L-1} < \frac{1}{4}\left(\frac{1}{\text{SNR}} - 1\right)^2, \tag{25}$$

(where SNR stands for the Signal-to-Noise Ratio, i.e., SNR $\triangleq \sigma_X^2/\sigma_N^2$) simultaneously hold, then it is possible to fix a value for the threshold T in the interval (σ_N^2, σ_X^2) such that the resulting test is asymptotically powerful, as both the sufficient conditions in (18) and (19), and the sufficient conditions in (23) and (24) simultaneously hold. In particular, from (25), for a fixed L and N, we can derive a minimum SNR for which an attack is surely correctly detected. More specifically, we can argue that if

$$\text{SNR}_{N,L} > \left(1 - 2\sqrt{\frac{\log N}{L-1}}\right)^{-1}, \tag{26}$$

[9]Note that less demanding but more involved conditions on T might be derived in order to have asymptotically zero false negative error probability.

then the test correctly detects the line search attack asymptotically (i.e., it is asymptotically powerful). Expectedly, such minimum sufficient SNR decreases by increasing L.

Furthermore, from the previous analysis we can also derive bounds on the value of N in order to ensure correct detection. Specifically,

$$D + 2 < N < e^{\frac{(L-1)}{4}\left(\frac{1}{\text{SNR}}-1\right)^2}. \tag{27}$$

Hence, 3 aligned queries (i.e., $N - D = 3$) are sufficient to detect the line search attack. On the other hand, for keeping a low probability of detecting an alignment under the null hypothesis, the growth rate for N must be at most sub-exponential in L. Therefore, these conditions are by far less demanding than the ones found for DM, where the number of attacking queries is required to go to infinity with the dimensionality for a successful detection.

We point out that, thanks to the knowledge that the metadetector has on the attacking strategy, the conditions for an asymptotical successful detection with the LS metadetection are much less strict with respect to the case of CTB-based metadetection analyzed in [16], where in order to ensure the detection of the attack the number of attacking queries $(N-D)$ is required to grow with L at a super-logarithmic rate.

VI. Experimental results

A. Synthetic signals

In this section we pseudorandomly generate signals following the statistical models proposed in Sect. IV-B for both the null and alternative hypotheses. Specifically, in our experiments, the N honest queries under the null hypothesis are randomly chosen to be watermarked or not. Concerning the alternative hypothesis, we consider 3 aligned queries with $\boldsymbol{\psi} = (0, 1, 0.5)$, while the remaining $D = N - 3$ (dummies) are generated as honest queries. In all the reported experiments we set $\sigma_N^2 = 1$, whereas several values are considered for L, N and σ_X^2. It is proper to say that we considered small values of N and L, with respect to the typical values for these parameters in real frameworks, in order to be able to illustrate the behavior of the detectors and compare the methods (for larger values of N and L, we would get smaller error detection probability, which would not only hinder the comparison between different scenarios, but in turn make it more difficult to validate our analytical results).

For all the experiments, we evaluate the false positive rate (FPR), i.e., the fraction of legal batches of queries classified as illegal, and true positive rate (TPR), i.e., the fraction of illegal batches of queries correctly classified as illegitimate.

Figs. 3 and 4 show the Equal Error Rate (EER), i.e., the metadetection error rate when the false positive and false negative error rates take the same value. While Fig. 3 reflects the dependence with L and N of the probability of error for all four proposed metadetectors, Fig. 4 focuses on its behavior with N and σ_X^2. Concerning Fig. 3, as one would expect, for a given size N of the batch of queries, the error probability decreases when the dimensionality L of the input signals is increased, i.e., the larger the number of components per signal we have, the smaller the error probability. Conversely, for a

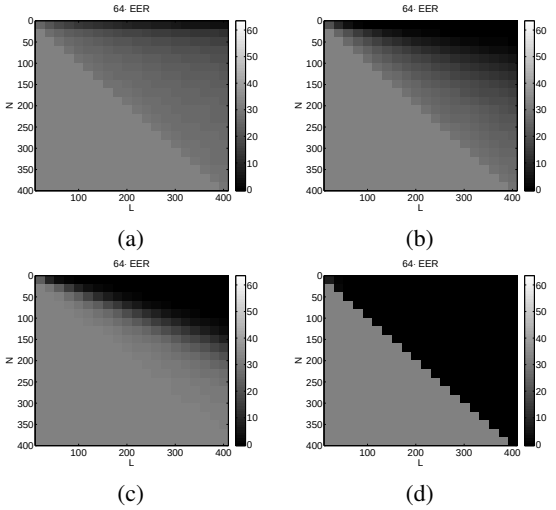


Fig. 3. EER for (a) DM, (b) SM, (c) SEM, and (d) GAM. $L \in [20, 400]$, $N \in [10, L-10]$, $\sigma_X^2 = 2$. For the sake of graphical representation, those points in the plots where $N \geq L$ were assigned EER = 0.5.

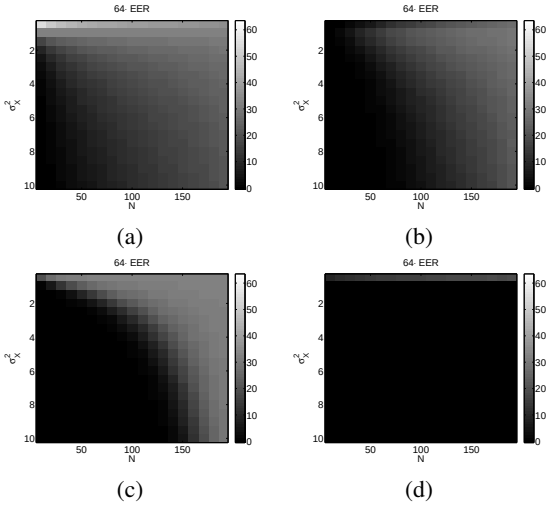


Fig. 4. EER for (a) DM, (b) SM, (c) SEM, and (d) GAM. N goes from 10 to 190, while σ_X^2 ranges from 0.5 to 10 and $L = 200$.

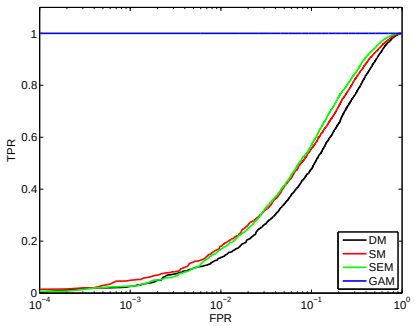


Fig. 5. ROC curves for the four metadetectors. $L = 100$, $N = 80$, $\sigma_X^2 = 10$.

fixed value of the dimensionality, the larger the number of queries grouped in a batch, the more difficult will be to make the right decision (i.e., the error probability increases). Note that this performance behavior is consistent with the analysis made in Sect. V. The comparison between the different metadetectors shows that SM outperforms DM, and both of them improve upon SEM when N is close to L. Nevertheless, when L is significantly larger than N, the latter shows a better performance than both DM and SM. On the other hand, GAM clearly outperforms them all; this is expected, since it exploits information that is not available in practice (recall that GAM is not a feasible detector, but we analyze it to bound the performance of the proposed covariance matrix-based metadetectors). Fig. 4 shows the performance of the four metadetectors as a function of N and σ_X^2. First of all, note that the proposed metadetectors establish the detection regions under the assumption that $\sigma_X^2 \geq \sigma_N^2$, and in those plots $\sigma_N^2 = 1$. This explains why DM provides error probabilities larger than 0.5 for $\sigma_X^2 = 0.5$, and also the behavior of GAM for the same case. Besides this degenerate case, we can see that, as a general rule, for a given batch size N, the EER decreases with σ_X^2 (i.e., the larger σ_X^2, the easier it will be to detect the oracle attack, as the deviation of the obtained covariance matrix with respect to the covariance matrix corresponding to the null hypothesis is larger); on the other hand, for a given σ_X^2, the larger N, the more difficult it will be to make a correct decision (similar conclusions were derived from Fig. 3). Again, this behavior is consistent with the analysis in Sect. V. Concerning the comparison among metadetectors, once again SM outperforms DM, and SEM shows a sharper transition from small to large error probabilities than the former ones. Finally, GAM shows significantly better performance than their counterparts.

Figs. 5 through 9 show the Receiver Operating Characteristic (ROC) curve for several parameter settings. For the sake of visualization, since we are interested in small values of the error probabilities, the FPR is plotted in logarithmic scale. Specifically, Figs. 5, 6, and 7 show the ROCs for $L = 100$, $N = 80$, with $\sigma_X^2 = 10$, $\sigma_X^2 = 30$, and $\sigma_X^2 = 1.2$, respectively. The comparison confirms that the larger σ_X^2, the easier is to make the right decision. Improvements in the metadetection performance can be also achieved by decreasing N (cf. Fig. 8), or increasing L (cf. Fig. 9; note that in this figure the GAM curve overlaps the SEM one), validating the conclusions in Sect. V. Furthermore, those figures show that most of the time the metadetectors can be sorted by increasing performance as DM, SM, SEM, and GAM, although the performance loss of SEM is much sharper than that observed for DM and SM. As we have discussed above with regard to Figs. 3 and 4, this seems to indicate that SEM is much more sensitive to the values of SNR, N, and L than the other practical metadetectors; specifically, if for some given N and L values the SNR is not sufficiently high, then it may be the case that the performance of SEM is even worse than that of the other two metadetectors. Such behavior of SEM is shown in Fig. 7 where we can also see that when the SNR is very low, the performance of all the practical metadetectors is nearly equivalent to a random guess. Moreover, according to the analysis in Sect. V-B GAM is expected to be asymptotically powerful as long as (26) holds; for $L = 100$ and $N = 80$ the threshold value is 1.7264; therefore, for $\sigma_X^2 = 1.2$ one should no longer expect the performance of GAM to be perfect, as it is confirmed by the experimental results shown in Fig. 7.

Another relevant measure regarding the proposed metade-

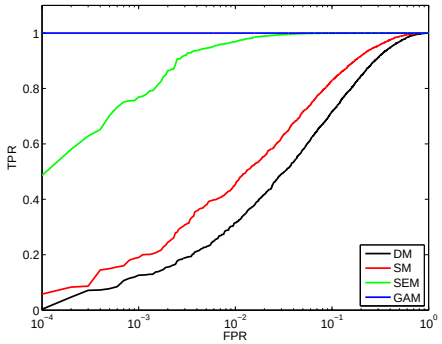


Fig. 6. ROC curves for the four metadetectors. $L = 100$, $N = 80$, $\sigma_X^2 = 30$.

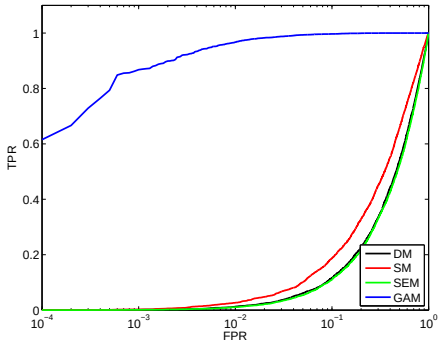


Fig. 7. ROC curves for the four metadetectors. $L = 100$, $N = 80$, $\sigma_X^2 = 1.2$.

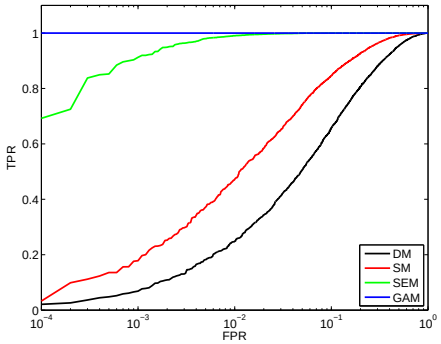


Fig. 8. ROC curves for the four metadetectors. $L = 100$, $N = 60$, $\sigma_X^2 = 10$.

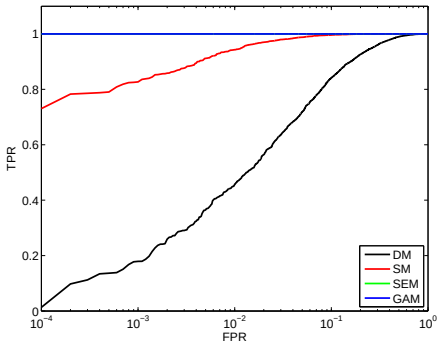


Fig. 9. ROC curves for the four metadetectors. $L = 200$, $N = 80$, $\sigma_X^2 = 10$.

tectors is their computational cost. For the sake of illustration, we consider the framework reported in Fig. 5 (i.e., $L = 100$, $N = 80$, $\sigma_X^2 = 10$); the average time required for computing each metadector in a *2xXeon E5-2690v3 2.6 GHz* with 24 cores and 256 GB of RAM is around 0.17 ms, 0.21 ms, 1.4 ms, and 0.17 ms for DM, SM, SEM, and GAM, respectively.

Finally, Fig. 10 illustrates the good performance of the proposed metadetectors when the dimensionality is larger, although still reasonable (even small) for practical scenarios, and the batch size is large too. Specifically, we consider $L = 400$, $N = 300$, and $\sigma_X^2 = 10$; the obtained plots show that the metadetectors will be able to detect the 3 aligned queries, even if the attacker hides them in a pool of as many as 297 dummy queries. Therefore, the attacker would have to significantly reduce the efficiency of his attacks (in terms of the ratio between the number of aligned queries and dummy queries) if he wants to go undetected. As a consequence, for reaching a given target final accuracy without being caught, the attacker has to spread out its attack over a large number of queries and then is forced to significantly delay his action.

B. Real images

We also performed experiments with real images. Note that the assumptions made in Sect. IV-B, e.g., the Gaussianity of both the host signals and the noise, and the fact that their samples are i.i.d., are far from holding when we work with real images. In spite of this, we will illustrate that the qualitative conclusions derived in Sect. IV are still applicable when real images are considered.

The database used in these experiments is UCID [40] which consists of 1338 images of sizes 512×384 or 384×512. In order to show the behavior of the proposed metadetectors, we have split the images into non-overlapping blocks, yielding a database with a larger number of smaller images. This allows us to have a larger number of realizations for plotting the results, but also to get a wider overlapping between the pdfs under the two tested hypotheses; as a consequence of both factors, the reported comparisons are more illustrative.

Specifically:

- Each image in UCID is converted to grayscale.
- The resulting grayscale image is split into non-overlapping blocks of size $M \times M$. In the reported results $M = 16, 32, 64$, so there is an integer number of blocks per image.
- Flat areas in the original UCID database images will produce a covariance matrix with a very small determinant, even if an oracle attack is not performed. Note that this problem comes from considering small blocks, which may have almost no texture; in practice, this issue will not arise with full images. To sidestep it, we disregarded the 5% blocks of the new database with the smallest variance.
- For each Monte Carlo simulation, the images in the new database (i.e., the 95% of $M \times M$ blocks from the UCID database with the largest variance) are pseudorandomly permuted. Half of them are watermarked with probability $1/2$, else are left unaltered, simulating the null hypothesis. For the other half, we consider disjoint sets of N images; D of them are watermarked with probability $1/2$ (the dummy queries), while the remaining $N - D$ images are generated from 2 images:
  - One of those 2 images is not watermarked.

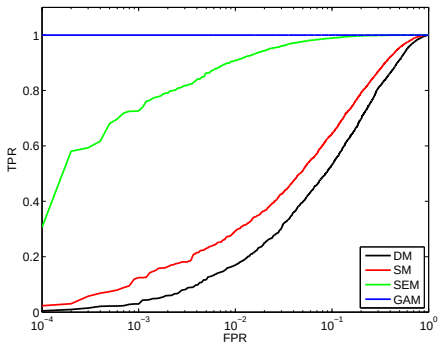


Fig. 10. ROC curves for the four metadetectors. $L = 400$, $N = 300$, $\sigma_X^2 = 10$.

- The other one is watermarked.
- A bisection algorithm is performed from the two previous images, yielding $N - D - 2$ images which are convex combinations of them. The detector driving this oracle attack sets the detection boundary at $\phi_w(\mathbf{y}) = \frac{<\mathbf{w},\mathbf{w}>}{2} = \frac{L\gamma^2}{2}$.
- The resulting images are quantized to 8 bits.

Unless otherwise explicitly mentioned, in this section we consider $\gamma = 3$. Due to the small values of the detection error probabilities when large values of L are considered, in this section we perform a more qualitative analysis, based on the histograms of the metadetection statistics, and complement it with the use of the Jensen-Shannon Divergence (JSD) [41] of the involved histograms, which is defined as

$$\begin{aligned} \mathrm{JSD}(p_0, p_1) \quad &\triangleq \quad \frac{1}{2}\sum_k p_0(k)\left[\log(p_0(k)) - \log(p(k))\right] \\ &+ \quad \frac{1}{2}\sum_k p_1(k)\left[\log(p_1(k)) - \log(p(k))\right], \end{aligned}$$

where $p_0(k)$ and $p_1(k)$ are the relative frequencies of the kth histogram bin under the null and alternative hypotheses, respectively, $p \triangleq \frac{p_0+p_1}{2}$, and we use the base-$e$ logarithm (consequently the units of the reported JSD values are *nats*).

For example, Fig. 11 shows the histograms of the metadetection statistics for each metadetector under both hypotheses. Similarly to the results reported for synthetic signals, the results achieved for DM and SM are very similar. In this case is striking the good behavior of SEM (note the logarithmic scale in the X-axis), although the best performance is still achieved by GAM. In the subsequent experiments we check the behavior of the metadetection statistics with some of the system parameters. In particular, in Fig. 12 we consider the histograms of SEM statistics when $N = 8$ and $D = 4$ (as in Fig. 11(c)), but we change L to $16 \times 16 = 256$, and $64 \times 64 = 4096$; in terms of JSD, the obtained values are 0.1527 ($L = 256$), 0.1990 ($L = 1024$), and 0.2620 ($L = 4096$). Comparing those three plots we can verify that the larger L, the easier will be to make the right decision, thus validating the conclusions given in Sect. V and further checked in Sect. VI-A. Similarly, in Fig. 13 we consider the impact of the batch size on the system performance. We can also compare those plots with Fig. 11(a). In view of those

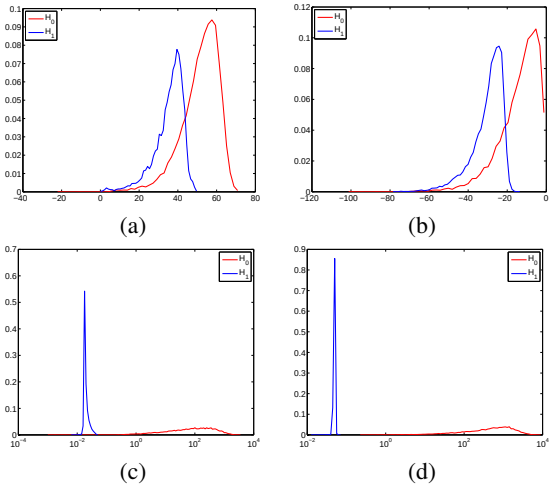


Fig. 11. Histograms of the metadetection statistics for (a) DM, (b) SM, (c) SEM, and (d) GAM. The red histogram corresponds to the null hypothesis (no attack), and the blue one to the alternative hypothesis (oracle attack). $L = 32 \times 32 = 1024$, $N = 8$, $D = 4$.

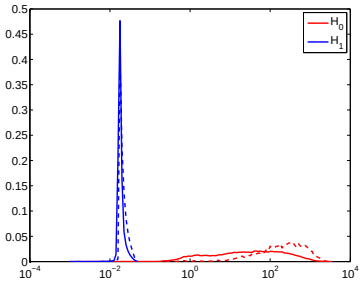


Fig. 12. Histograms of the metadetection statistics for SEM for $L = 256$ (16×16, solid lines), and 4096 (64×64, dashed lines). The red histograms correspond to the null hypothesis (no attack), and the blue ones to the alternative hypothesis (oracle attack). $N = 8$, $D = 4$.

results, we confirm that larger values of N make it harder to make the right decision. Quantitatively, in terms of JSD the obtained values are 0.1389, 0.0986, 0.0666, and 0.0478 for $N = 12, 16, 20$, and 24, respectively. Finally, in Fig. 14 we analyze the behavior of DM when the watermark strength γ is modified. In particular in Fig. 14 we consider both $\gamma = 1$ and $\gamma = 10$; furthermore, we can compare both plots with Fig. 11(a), where $\gamma = 3$. As we can observe, the obtained histograms are virtually the same; only small differences, due to the random nature of the histograms, can be found. Indeed, the JSD for $\gamma = 1$ is 0.1780, while for $\gamma = 10$ we obtain 0.1781. This result validates again our conclusions in Sect. V. Although it might seem counterintuitive at first sight, this result is reasonable if one thinks that the proposed detectors are based on the covariance matrix (i.e., they disregard the signal mean, which is the statistic modified by the watermark embedding strength).

VII. Conclusions

In this paper we have addressed the problem of metadetection of oracle attacks based on line searches when an aware-attacker takes some countermeasures to avoid being discovered by the smart detector. We have proposed several

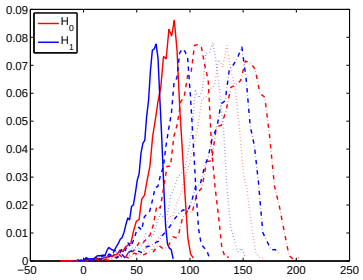


Fig. 13. Histograms of the metadetection statistics for DM for $N = 12$ (solid lines), 16 (dashed lines), 20 (dotted lines), 24 (dash-dotted lines). The red histograms correspond to the null hypothesis (no attack), and the blue ones to the alternative hypothesis (oracle attack). $L = 32 \times 32 = 1024$, $D = 4$.

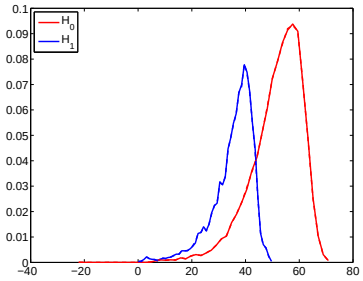


Fig. 14. Histograms of the metadetection statistics for DM for $\gamma = 1$ (solid lines), and $\gamma = 10$ (dashed lines). The red histograms correspond to the null hypothesis (no attack), and the blue ones to the alternative hypothesis (oracle attack). $L = 32 \times 32 = 1024$, $N = 8$, $D = 4$.

practical metadetectors and assessed their theoretical performance. Then, we have used an asymptotic analysis to find the critical values of the parameters of the system allowing for a correct detection of the oracle attacks. Experiments on both synthetic signals and images confirm the power of the smart detector, by showing that excellent detection performance can be achieved even when few attacking queries are hidden in a large pool of dummy queries. Note again that the LS metadetectors developed in this paper are general purpose; we applied them to the watermark detection problem just to illustrate their practical usefulness. Although the LS metadetectors proposed are applied to the watermark detection problem, this is only a case study and our arguments about the metadetection are general purpose. Since adversarial binary decision is one of the core problems in adversarial signal processing [1], the techniques developed in this paper find application in many different fields, such as multimedia forensics, biometrics, network intrusion detection, reputation systems, and many others.

As a future work, we plan to investigate the most suitable strategy to be implemented by the smart detector once an oracle attack is detected. Among the possible directions of research, it would be also interesting to focus on a very simple setup where it would be possible to study the interplay between the smart detector and the adversary as a dynamic game. Another very interesting direction would be to study the metagame between the oracle and the attacker as an inspection game [42]. Inspection games have been recently advocated as a possible way to extend the classical statistical decision problem when the distribution of the random variable observed by the statistician, or 'inspector', is strategically controlled by another player, namely the 'inspectee'. In this way, such models account for the fact that the 'inspectee' can behave either legally (H_0) or illegally (H_1), in which case he also chooses a violation procedure.

VIII. Acknowledgment

This work was partially funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under projects TACTICA and COMPASS (TEC2013-47020-C2-1-R), by the Galician Regional Government and ERDF under projects "Consolidation of Research Units" (GRC2013/009), REdTEIC (R2014/037) and AtlantTIC, and by the EU under project NIFTy (HOME/2012/ISEC/AG/INT/4000003892).

References

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 8682 – 8686.

[2] I. J. Cox and J.-P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *IEEE International Conference on Image Processing*, vol. 3, Santa Barbara, CA, USA, October 1997, pp. 0.3–0.6.

[3] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," in *International Workshop on Digitial Watermarking*, Siena, Italy, September 2005, pp. 260–274.

[4] M. Martínez-Díaz, J. Fiérrez-Aguilar, F. Alonso-Fernández, J. Ortega-García, and J. A. Sigüenza, "Hill-climbing and brute-force attacks on biometric systems: A case study in match-on-card fingerprint verification," in *Annual IEEE International Carnahan Conferences Security Technology*, Lexington, KY, USA, October 2006, pp. 151–159.

[5] J. Galbally, J. Fiérrez, J. Ortega-García, C. McCool, and S. Marcel, "Hill-climbing attack to an eigenface-based face verification system," in *IEEE International Conference on Biometrics, Identity and Security*, Tampa, FL, USA, September 2009, pp. 1–6.

[6] E. Maiorana, G. E. Hine, and P. Campisi, "Hill-climbing attacks on multibiometrics recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 900–915, May 2015.

[7] D. Lowd and C. Meek, "Adversarial learning," in *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, IL, USA, August 2005, pp. 641–647.

[8] P. Fogla and W. Lee, "Evading network anomaly detection systems: Formal reasoning and practical techniques," in *ACM Conference on Computer and Communications Security*, Alexandria, VA, USA, October-November 2006, pp. 59–68.

[9] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems," in *ACM Conference on Computer and Communications Security*, Washington, DC, USA, November 2002, pp. 255–264.

[10] M. Barni, P. Comesaña-Alfaro, F. Pérez-González, and B. Tondi, "Are you threatening me?: Towards smart detectors in watermarking," in *SPIE Electronic Imaging*, San Fransciso, CA, USA, February 2014.

[11] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *International Conference on Machine Learning*, Pittsburgh, PA, USA, June 2006, pp. 353–360.

[12] B. Biggio, I. Corona, Z.-M. He, P. P. K. Chan, G. Giacinto, D. S. Yeung, and F. Roli, "One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time," in *International Workshop on Multiple Classifier Systems*, Günzburg, Germany, June-July 2015, pp. 168–180.

[13] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, and A. D. Keromytis, "Casting out demons: Sanitizing training data for anomaly sensors," in *IEEE Symposium on Security and Privacy*, Oakland, CA, USA, May 2008, pp. 81–95.

[14] M. El Choubassi and P. Moulin, "Sensitivity analysis attacks against randomized detectors," in *IEEE International Conference on Image Processing*, vol. 2, San Antonio, TX, USA, September 2007, pp. 129–132.

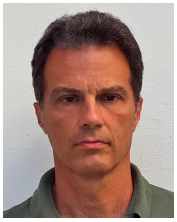
[15] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," *IEE Proceedings on Information Security*, vol. 153, no. 3, pp. 115–125, September 2006.

[16] B. Tondi, P. Comesaña-Alfaro, F. Pérez-González, and M. Barni, "On the effectiveness of meta-detection for countering oracle attacks in watermarking," in *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 2015, pp. 1 – 6.
[17] I. Cox, M. Miller, and J. Bloom, *Digital watermarking*. Morgan Kaufmann, 2002.
[18] P. Moulin and R. Koetter, "Data-hiding codes," *Proceedings of the IEEE*, vol. 93, no. 12, pp. 2083–2126, 2005.
[19] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, December 1997.
[20] M. Barni and F. Bartolini, *Watermarking systems engineering: enabling digital assets security and other applications*. CRC Press, 2004.
[21] J.-P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *International Workshop on Information Hiding*, Portland, OR, USA, April 1998, pp. 258–272.
[22] M. F. Mansour and A. H. Tewfik, "LMS-based attack on watermark public detectors," in *IEEE International Conference on Image Processing*, vol. 3, Rochester, NY, USA, September 2002, pp. 649–652.
[23] M. E. Choubassi and P. Moulin, "Noniterative algorithms for sensitivity analysis attacks," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 2, pp. 113–126, June 2007.
[24] P. Comesaña and F. Pérez-González, "Breaking the BOWS watermarking system: key guessing and sensitivity attacks," *EURASIP Journal on Information Security*, 2007.
[25] http://bows2.ec-lille.fr/index.php?mode=VIEW&tmpl=resPrevEp#ResEp2.
[26] M. Barreno, B. Nelson, A. D. Joseph, and J. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, November 2010.
[27] A. Adelsbach and A.-R. Sadeghi, "Zero-knowledge watermark detection and proof of ownership," in *International Workshop on Information Hiding*, Pittsburgh, PA, USA, April 2001, pp. 273–288.
[28] J. Troncoso-Pastoriza and F. Pérez-González, "Zero-knowledge watermark detector robust to sensitivity attacks," in *ACM Workshop on Multimedia and Security*, Portland, OR, USA, April 2006, pp. 97–107.
[29] P. Moulin and I. Ivanovic, "The zero-rate spread-spectrum watermarking game," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1098–1117, April 2003.
[30] 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.
[31] ——, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, August 2014.
[32] R. Gibbons, *A primer in game theory*. Harvester Wheatsheaf, 1992.
[33] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: asymptotic theory*. Springer-Verlag, 1981, vol. 2.
[34] I. A. Ibragimov, A. S. Nemirovskii, and R. Z. Khas'minskii, "Some problems on nonparametric estimation in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 31, no. 3, pp. 391–406, September 1987.
[35] D. Jonsson, "Some limit theorems for the eigenvalues of a sample covariance matrix," *Journal of Multivariate Analysis*, vol. 12, no. 1, pp. 1–38, March 1982.
[36] T. T. Cai, T. Liang, and H. H. Zhou, "Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions," *arXiv preprint arXiv:1309.0482*, 2013.
[37] T. W. Anderson, *An introduction to multivariate statistical analysis*. Wiley New York, 1958, vol. 2.
[38] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1996.
[39] R. A. Horn, N. H. Rhee, and S. Wasin, "Eigenvalue inequalities and equalities," *Linear Algebra and its Applications*, vol. 270, no. 1-3, pp. 29–44, February 1998.
[40] G. Schaefer and M. Stich, "UCID - an uncompressed colour image database," in *SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*, San Jose, CA, USA, January 2004, pp. 472–480.
[41] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, January 1991.
[42] R. Avenhaus, B. Von Stengel, and S. Zamir, "Inspection games," *Handbook of game theory with economic applications*, vol. 3, pp. 1947–1987, 2002.

Benedetta Tondi (SM'13) received the master degree (*cum laude*) in Electronics and Communications Engineering at the University of Siena, Siena, Italy, in 2012 with a thesis on the adversary-aware source identification in the area of multimedia forensics and the PhD degree at the Department of Information Engineering and Mathematics (DIISM) of the University of Siena in 2016. She is currently Research Associate at the the DIISM, University of Siena. She is a member of the Visual Information Processing and Protection (VIPP) Group in the DISM. She is assistant for the course of Information Theory and Coding and Multimedia Security, led by Mauro Barni. She is a member of the National Inter-University Consortium for Telecommunications (CNIT). She is a Student member of the IEEE Young Professionals and IEEE Signal Processing Society. Her research interest focuses on the application of Information theory and Game theory concepts to forensics and counter-forensics analysis and more in general on the Adversarial Signal Processing. From October 2014 to February 2015 she has been a visiting student at the University of Vigo at the Signal Processing in Communications Group (GPSC). She has been designated reviewer on the technical program committee for the IEEE GlobalSIP14-Workshop on Information Forensics and Security (WIFS) 2014, the IEEE International Conference on Multimedia and Expo (IEEE ICME) 2015 and 2016, and the IEEE International Conference on Image Processing (IEEE ICIP) 2015 and 2016. She is winner of the Best Student Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS) 2014, and the Best Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS) 2015.

Pedro Comesaña-Alfaro (M08-SM15) received the Telecommunication Engineer degree from the University of Vigo, Vigo, Spain in 2002 and the Ph. D. degree in telecommunications engineering from the same institution in 2006. During his Ph. D. studies Dr. Comesaña stayed at the Technische Universiteit Eindhoven (The Netherlands, 2004); then, he hold post-doc positions at the University College Dublin (Ireland, 2006), University of Siena (Italy, 2007-2008), and University of New Mexico (Albuquerque, NM, USA, 2010-2011); and he was visiting scholar at the State University of New York at Binghamton (NY, USA, 2015). In 2008 Dr. Comesaña joined the faculty of the School of Telecommunication Engineering, University of Vigo, as an Assistant Professor, where he is currently an Associate Professor. His research interests lie in the areas of multimedia security (including watermarking and forensics), and digital communications. He has coauthored several international patents related to watermarking for video surveillance, and fingerprinting of audio signals. Dr. Comesaña has co-authored over 50 papers in leading international journals and peer-reviewed conferences; he was recipient of IEEE-WIFS 2014 Best Paper Award. He has participated in the European projects ECRYPT, REWIND, and NIFTY. Currently, Dr. Comesaña serves as an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology, IEEE Signal Processing Letters, and IET Information Security, and is a member of the IEEE SPS Student Services Committee; furthermore, he was a member of the IEEE SPS Information Forensics and Security-Technical Committee, Technical co-Chair of ACM IH&MMSec 2015, and Area Chair of IEEE ICIP 2015.

Fernando Pérez-González (M'90-SM'09-F'16) received the Telecommunication Engineer degree from the University of Santiago, Santiago, Spain in 1990, and the Ph.D. degree in telecommunications engineering from the University of Vigo, Vigo, Spain, in 1993. In 1990, he became an Assistant Professor with the School of Telecommunication Engineering, University of Vigo. From 2007 to 2010, he was Program Manager of the Spanish National R&D Plan on Electronic and Communication Technologies, Ministry of Science and Innovation. From 2009 to 2011, he was the Prince of Asturias Endowed Chair of Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. From 2007 to 2014, he was the Executive Director of the Galician Research and Development Center in Advanced Telecommunications. He has been the Principal Investigator of the University of Vigo Group, which participated in several European projects, including CERTIMARK, ECRYPT, REWIND, NIFTY, and WITDOM. He is currently a Professor in the School of Telecommunication Engineering, University of Vigo, Vigo, Spain, and a Research Professor in Information Science and Technology, University of New Mexico, Albuquerque, NM, USA. He has coauthored more than 60 papers in leading international journals and 160 peer-reviewed conference papers. He has coauthored several international patents related to watermarking for video surveillance, integrity protection of printed documents, fingerprinting of audio signals, and digital terrestrial broadcasting systems. His research interests include the areas of digital communications, adaptive algorithms, privacy enhancing technologies, and information forensics and security. Prof. Pérez-González was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (from 2005 to 2009) and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (from 2006 to 2010). He is currently is an Associate Editor of the LNCS Transactions on Data Hiding and Multimedia Security, and the EURASIP International Journal on Information Forensics and Security.



Mauro Barni (M'92-SM'06-F'12) graduated in Electronic Engineering at the University of Florence in 1991. He received the PhD in Informatics and Telecommunications in October 1995. He has carried out his research activity for over 20 years first at the Department of Electronics and Telecommunication of the University of Florence, then at the Department of Information Engineering and Mathematics of the University of Siena where he works as associate Professor. During the last decade he has been studying the application of image processing techniques to copyright protection and authentication of multimedia, and the possibility of processing signals that has been previously encrypted without decrypting them (digital watermarking, multimedia forensics, signal processing in the encrypted domain). Lately he has been working on theoretical and practical aspects of adversarial signal processing. He is author/co-author of about 270 papers published in international journals and conference proceedings, and holds four patents in the field of digital watermarking and image authentication. He is co-author of the book 'Watermarking Systems Engineering: Enabling Digital Assets Security and other Applications', published by Dekker Inc. in February 2004. He participated to several National and European research projects on diverse topics, including computer vision, multimedia signal processing, remote sensing, digital watermarking, IPR protection. Currently he is involved in the REWIND project focusing on theoretical and practical aspects of Multimedia Forensics (FET program). He was the funding editor of the EURASIP Journal on Information Security. He is currently part of the editorial board of the IEEE Signal Processing Magazine. Prof. Barni has been the chairman of the IEEE Information Forensic and Security Technical Committee (IFS-TC) from 2010 to 2011. He is a fellow member of the IEEE and a member of EURASIP. He was appointed DL of the IEEE SPS for the years 2013-2014.