

Towards Optimal Design of Data Hiding Algorithms Against Nonparametric Adversary Models

Alvaro A. Cárdenas
 Department of Electrical Engineering
 and Computer Science
 University of California
 Berkeley, CA, 94720, USA
 cardenas@eecs.berkeley.edu

George V. Moustakides
 Department of Electrical and
 Computer Engineering
 University of Patras
 26500 Rio, Greece
 moustaki@ece.upatras.gr

John S. Baras
 Department of Electrical and
 Computer Engineering
 University of Maryland
 College Park, MD, 20742, USA
 baras@isr.umd.edu

Abstract—This paper presents a novel zero-sum watermarking game between a detection algorithm and a data hiding adversary. Contrary to previous research, the detection algorithm and the adversary we consider are both nonparametric in a continuous signal space, and thus they have no externally imposed limitations on their allowed strategies except for some distortion constraints. We show that in this framework no deterministic detection algorithm is optimal. We then find optimal randomized detection algorithms for different distortion levels and introduce a new performance tradeoff between completeness and accuracy when a detection algorithm does not have enough evidence to make an accurate decision.

I. INTRODUCTION

To formally characterize the behavior of data hiding algorithms there has recently been a growing interest in trying to prove analytical properties of data hiding codes. In this general framework one usually assumes a performance metric $\Psi(\mathcal{DH}, \mathcal{A}^f)$ (such as the probability of error) and describe it in terms of the data hiding algorithm \mathcal{DH} (which include the embedding and detection algorithms) and a fixed attack \mathcal{A}^f (e.g., a Gaussian noisy channel with known parameters). An ideal data hiding code in this framework optimizes the given performance metric, e.g.,

$$\mathcal{DH}^* = \arg \min_{\mathcal{DH}} \Psi(\mathcal{DH}, \mathcal{A}^f).$$

An example of this type of approach can be found in [1], where the adversary model is fixed as an average collusion attack with additive Gaussian noise, and several parameters of the data hiding code are optimized in order to find good performance guarantees.

Fixing the adversary model to follow any given parametric model is a good approach for particular data hiding applications that are not subject to a significant threat, since several parametric models represent accurately the end results of standard *non-adversarial* signal processing algorithms applied to a signal. However in several other fields such as multimedia fingerprinting for traitor tracing, the objective of the attacker is directly opposite to the one of the data hiding code. Any limitation on the capabilities

of the adversary might therefore be unrealistic. Any data hiding algorithm designed under the assumption of a limited adversary will provide very weak security guarantees.

In order to limit the number of assumptions that might be violated easily by an adversary, there is a growing interest in trying to understand the notion of an intelligent opponent against data hiding algorithms. A basic objective in this framework is to find provable performance guarantees for a fixed data hiding algorithm \mathcal{DH}^f . The problem is therefore to find the performance of the algorithm against the worst-type attacks:

$$\Psi^* = \max_{\mathcal{A}} \Psi(\mathcal{DH}^f, \mathcal{A}).$$

A solution to this formulation shows that there is no other attack that will make \mathcal{DH}^f perform worse, i.e., $\forall \mathcal{A}, \Psi(\mathcal{DH}^f, \mathcal{A}) \leq \Psi^*$. A recent example of this kind of approach can be found in [2], where the detection algorithm is a fixed correlation detector, but where the adversary's strategy is optimized in order to find the attack probability density function (pdf) that maximizes the probability of error.

The question however remains on how to design optimal detection algorithms to achieve the smallest possible Ψ^* . For this end we need to find an optimal hiding strategy \mathcal{DH}^* and a least favorable attack strategy \mathcal{A}^* to the following problem:

$$(\mathcal{DH}^*, \mathcal{A}^*) = \arg \min_{\mathcal{DH}} \max_{\mathcal{A}} \Psi(\mathcal{DH}, \mathcal{A}). \quad (1)$$

A solution to this formulation tells us that there is no other \mathcal{DH} that will have better worst performance: $\forall \mathcal{DH}, \max_{\mathcal{A}} \Psi(\mathcal{DH}, \mathcal{A}) \leq \max_{\mathcal{A}} \Psi(\mathcal{DH}^*, \mathcal{A})$.

Solving the min-max problem is usually very difficult, and for that reason most of the research has restricted the adversary or the detection algorithm to follow parametric models. This simplifies the problem of finding the most damaging attacks and optimal detection algorithms, to one of finding the most damaging *parameters* of a specific kind of attack and the optimal *parameters* of a specific data hiding algorithm. Examples include optimizing the parameters

of Spread Spectrum or QIM watermarking algorithms with their corresponding detection structures (e.g., correlation detectors) [3], [4] or parametric versions of the adversary distributions such as Gaussian attacks [5]. Very little work has been directed into finding *optimal nonparametric detection algorithms against least-favorable nonparametric attack distributions*. Furthermore notice that by solving Eq. (1) there is no guarantee that we cannot do better, or that \mathcal{DH}^* is optimal against \mathcal{A}^* . More specifically we cannot guarantee that $\forall \mathcal{DH}, \Psi(\mathcal{DH}^*, \mathcal{A}^*) \leq \Psi(\mathcal{DH}, \mathcal{A}^*)$. Note that this optimality property can be achieved for example, by formulating the max-min problem, which does not always yield the same solution as the min-max formulation.

In this paper: (i) we introduce a nonparametric adversary model in order to claim stronger security properties for our detection algorithm (i.e., the performance of our algorithm does not depend on assuming a limited adversary). (ii) We introduce a nonparametric data-hiding detection algorithm in order to obtain lower bounds for Ψ than could not have been achieved if we had been using a parametric data hiding algorithm. (iii) We find algorithms that satisfy the following saddle-point equilibrium:

$$\Psi(\mathcal{DH}^*, \mathcal{A}) \leq \Psi(\mathcal{DH}^*, \mathcal{A}^*) \leq \Psi(\mathcal{DH}, \mathcal{A}^*) \quad (2)$$

and thus we are guaranteed that our solution satisfies not only the min-max formulation but also the max-min formulation of the problem.

Since we consider *continuous* signal spaces, finding a saddle point solution requires a double optimization over infinite dimensional spaces. In this paper we focus therefore on the scalar case and emphasize the insights on how to obtain saddle point solutions with the aim of generalizing the results to multi-dimensional spaces.

II. THE NON-BLIND WATERMARK VERIFICATION PROBLEM

In the watermark verification problem the data hiding code \mathcal{DH} consists of two main algorithms, an embedding algorithm \mathcal{E} and a detection algorithm \mathcal{D} , where

- The embedder \mathcal{E} receives as inputs a *host signal* s and a bit m . The objective of the embedder is to produce a *marked signal* x with no perceptual loss of information or major differences from s , but that carries the information about m . The general formulation of this required property is to force the embedder to satisfy a distortion constraint, since x should be perceptually similar to s . In order to facilitate the detection process, the embedder and the detector usually share a random secret key k (e.g., the seed for a pseudorandom number generator that is used to create an embedding pattern). This key can be initialized in the devices or exchanged via a secure out of band channel.

- The adversary \mathcal{A} can intercept the marked signal x produced by \mathcal{E} , and can send in its place, a *degraded signal* y to the detector. This degraded signal should satisfy a distortion constraint, since y should be perceptually similar to s (and x).
- The detection algorithm \mathcal{D} receives the degraded signal y and has to determine if y was embedded with $m = 1$ or $m = 0$. In non-blind watermarking the detector is also assumed to have access to the original signal s .

III. NONPARAMETRIC DETECTION AND ADVERSARY MODELS

In order to better understand the problem while keeping the computation tractable we work with a simplified formulation which we believe can give insights into how to solve more general formulations. However, although we restrict the complexity of the embedding algorithm and of the signal space, we still allow for a very general adversary model and detection algorithm. In particular we assume:

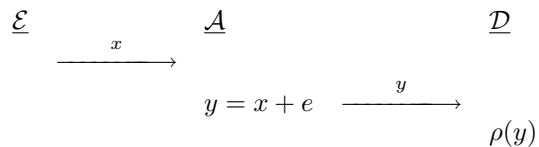
- The host signal s , the marked signal x and the degraded signal y are in \mathbb{R} .
- The embedding algorithm $\mathcal{E}(m, s)$ is fixed and parameterized by a publicly known distance d between different embeddings: that is, for all s ; $|\mathcal{E}(1, s) - \mathcal{E}(0, s)| = d$.
- The adversary \mathcal{A} , on input x , creates the attack $y = x + e$, where e is sampled from a distribution h chosen by the adversary from the set of pdfs that satisfy the distortion constraint

$$\mathcal{F}_D = \{h : \mathbb{E}[\mathcal{D}(y - x)] = \mathbb{E}[\mathcal{D}(e)] \leq D\},$$

where \mathbb{E} denotes the expected value over h . We also assume the adversary knows d , but does not know the value of m .

- The detection algorithm ρ outputs an estimate of m ($\hat{m} = 0$ or $\hat{m} = 1$) on input y .

The problem can be represented as:



Our objective is to find a pair (ρ^*, h^*) such that for all possible detection algorithms ρ and for all $h \in \mathcal{F}_D$

$$\Psi(\rho^*, h) \leq \Psi(\rho^*, h^*) \leq \Psi(\rho, h^*) \quad (3)$$

where $\Psi(\rho, h)$ is the probability of error: $\Pr[\rho(x + e) \neq m]$.

A. Conditions for Optimal Detection Rules

Let $x_i = \mathcal{E}(i, s)$. Recall that in order to satisfy $\Psi(\rho^*, h^*) \leq \Psi(\rho, h^*)$, ρ^* should be a *likelihood test*. In particular, ρ^* should select the largest between the

likelihood of y given x_1 : $h(y - x_1)$, and the likelihood of y given x_0 : $h(y - x_0)$, and should randomly flip a coin to decide if both likelihoods are equal. This decision rule is called *Bayes optimal*.

Note also that this optimal decision function assumes equal priors, i.e., $\Pr[m = 0] = \Pr[m = 1]$. However this can be easily generalized in the case where the priors are different by comparing $\Pr[m = 1]h(y - x_1)$ with $\Pr[m = 0]h(y - x_0)$. In the remaining of this paper we assume for simplicity of exposition that we have equal priors.

B. On the Necessity of Randomized Decisions

Since the likelihood test depends on the unknown distribution h , the challenge we face is to design the decision boundaries for ρ so that the optimal pdf h^* makes ρ a Bayes optimal decision.

The most intuitive decision functions are usually non-randomized decisions (except for sets of measure zero), where the decision space \mathbb{R} is divided into two *open* sets: R and its complement R^c . If $y \in R$ then $\rho(y) = 1$, otherwise $\rho(y) = 0$ (randomization is only used when y falls in the boundary of the two sets). However as we show next, it is impossible to obtain saddle point equilibria with deterministic decision functions.

Before we proceed let us define $a = y - x_1$ and assume without loss of generality that $d = x_0 - x_1 > 0$. Since there is no loss of information for ρ by taking as input a instead of y , in the remaining of this paper we assume ρ receives as input a for notational simplicity. Note that in this case the likelihood of a being generated under the hypothesis $m = 1$ is $h(a)$, and the likelihood of a being generated under hypothesis $m = 0$ is $h(a - d)$. Therefore the Bayes optimal decision function is:

$$\rho^*(a) = \begin{cases} 1 & \text{if } h(a) > h(a - d) \\ 0 & \text{if } h(a) < h(a - d) \\ \gamma & \text{if } h(a) = h(a - d) \end{cases} \quad (4)$$

where γ represents an arbitrary decision (e.g., a random decision) if the two likelihoods are equal. This can be allowed because the objective function is not affected for the case when the likelihoods are the same. This is in contrast to other optimality criteria, such as the Neyman-Pearson formulation, where the randomization γ must be selected with care.

Theorem 1: Let the distortion function $\mathcal{D}()$ be continuous, symmetric and monotone increasing for $a > 0$. Also assume that ρ is a deterministic decision function function (except for a set of measure zero). Then there is no ρ that satisfies a saddle point solution for any distortion function \mathcal{D} , distortion bound D and embedding distance d .

Proof: The probability of error, assuming equal prior probabilities for $m = 0$ or $m = 1$, can be

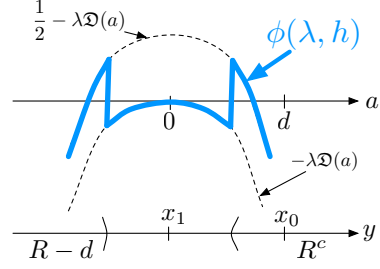


Fig. 1. Since $a = y - x_1$ we can see how the shape of $\phi(\lambda, h)$ determines where the adversary tries to distribute the density h such that $L(\lambda, h)$ is maximized while satisfying the constraints (i.e., while minimizing $L(\lambda, h)$ over λ).

expressed as:

$$\begin{aligned} \Psi(\rho, h) &= \frac{1}{2} (\Pr[\rho = 1|m = 0] + \Pr[\rho = 0|m = 1]) \\ &= \frac{1}{2} \left(\int_R h(y - x_0) dy + \int_{R^c} h(y - x_1) dy \right) \\ &= \frac{1}{2} \left(\int_R h(a - d) da + \int_{R^c} h(a) da \right) \\ &= \frac{1}{2} \left(\int_{R-d} h(a) da + \int_{R^c} h(a) da \right) \\ &= \frac{1}{2} \int_{\mathbb{R}} (1_{R-d}(a) + 1_{R^c}(a)) h(a) da \end{aligned}$$

where 1_R is the indicator function for the set R (i.e., $1_R(a) = 1$ if $a \in R$ and $1_R(a) = 0$ otherwise) and where $R - d$ is defined as the set $\{a - d : a \in R\}$.

The objective function is therefore:

$$\min_{R \subseteq \mathbb{R}} \max_{h \in \mathcal{F}_D} \frac{1}{2} \int (1_{R-d}(a) + 1_{R^c}(a)) h(a) da$$

Let us first fix R and proceed with the optimization of h . Since the maximization is a constrained optimization problem we form the Lagrangian for h :

$$L(\lambda, h) = \int \phi(\lambda, a) h(a) da + \lambda D \quad (5)$$

where λ is a Lagrange multiplier and where

$$\phi(\lambda, a) = \frac{1}{2} (1_{R-d}(a) + 1_{R^c}(a)) - \lambda \mathcal{D}(a).$$

The problem we need to solve now is

$$(\lambda^*, h^*) = \arg \min_{\lambda \geq 0} \max_{h \in \mathcal{F}_D} L(\lambda, h). \quad (6)$$

Note that $L(\lambda^*, h^*) = \Psi(\rho, h^*)$ for fixed R .

By looking at the form of $\phi(\lambda, a)$ in Fig. 1 (for $\lambda > 0$, i.e., when the distortion constrains on h are active) it is clear that a necessary condition for optimality is $R - d \cap R^c = \emptyset$, since otherwise, the adversary will put all the mass of h in this interval, achieving a probability of error of one. Under this condition we assume the very intuitive decision function specified by $R - d = [-\infty, \frac{-d}{2}]$ (the same results are obtained for non-connected sets R).

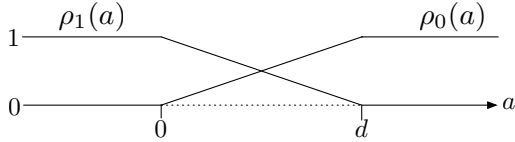


Fig. 2. Piecewise linear decision function, where $\rho(0.5d) = \rho_0(0.5d) = \rho_1(0.5d) = \frac{1}{2}$

Assuming $D < \mathfrak{D}(\frac{d}{2})$, the optimal adversarial strategy is¹

$$h^*(a) = p\delta\left(\frac{-d}{2} - \epsilon\right) + (1 - 2p)\delta(0) + p\delta\left(\frac{d}{2} + \epsilon\right)$$

where the adversary chooses ϵ to be arbitrarily small (and take care of the discontinuity of the decision function), and then selects $p = \frac{D}{2\mathfrak{D}(\frac{d}{2} + \epsilon)}$, which turns out to be also the probability of error; i.e., $\Psi(\rho, h^*) = p$.

Notice also that for $D \geq \mathfrak{D}(\frac{d}{2})$, $\lambda^* = 0$, and therefore there will always be an h^* such that $\Psi(\rho, h^*) = \frac{1}{2}$. The interpretation for this case is that the distortion constraints are not strict enough, and the adversary can create error rates up to 0.5 by choosing

$$h^*(a) = \frac{1}{2}\delta\left(\frac{-d}{2} - \epsilon\right) + \frac{1}{2}\delta\left(\frac{d}{2} + \epsilon\right)$$

where $0 < \epsilon \leq \mathfrak{D}^{-1}(D) - \frac{d}{2}$.

Therefore for any fixed decision regions R , and any D , the adversary is able to find h^* maximizing the probability of error and at the same time making the decision region suboptimal, since any ρ will not be Bayes optimal, and if it is Bayes optimal then h^* is not a maximizing distribution. ■

C. Saddle Point Equilibria for $D = (\frac{d}{2})^2$

Having shown that there are no saddle point equilibrium solutions for deterministic decision functions, we now show three saddle point equilibria that can be achieved with randomized decision functions and by assuming the very common quadratic distortion constraint $\mathbb{E}[\mathfrak{D}(a)] = \mathbb{E}[a^2] \leq D$. First we show a saddle point solution for $D = (\frac{d}{2})^2$, which has probability of error of $\frac{1}{4}$. We then show how to obtain a saddle point solution for $0 \leq D \leq (\frac{d}{2})^2$ with probability of error of $\frac{D}{d^2} \leq \frac{1}{4}$ and finally a saddle point solution for $(\frac{d}{2})^2 < D \leq 3(\frac{d}{2})^2$, using an expected cost function as a generalization from the probability of error.

Let $\rho(a) = 0$ with probability $\rho_0(a)$ and $\rho(a) = 1$ with probability $\rho_1(a)$. In order to have a well defined decision function we require $\rho_0 = 1 - \rho_1$. Consider now the decision function given in Fig. 2. For this case,

¹For notational simplicity we denote throughout this paper the Dirac delta function $\delta(a - a_0)$ as $\delta(a_0)$. The fact that it is a function of a will remain implicit.

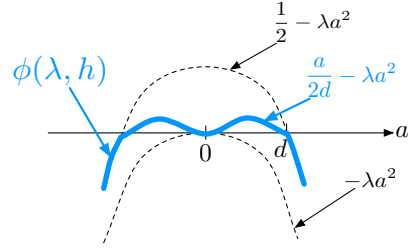


Fig. 3. The discontinuity problem in the Lagrangian is solved by using piecewise linear *continuous* decision functions. It is now easy to shape the Lagrangian such that the maxima created form a saddle point equilibrium.

the Lagrangian has again the same form as in Eq. (5), but this time

$$\phi(\lambda, a) = \frac{1}{2}(\rho_1(a + d) + \rho_0(a)) - \lambda a^2. \quad (7)$$

In order to have active distortion constraints, the maxima of $L(\lambda, h)$ (with respect to h) should be in the interval $a \in [-d, d]$. Looking at Fig. 3 we see that $\phi(\lambda, a)$ achieves its maximum value as a function of a for $a = \pm \frac{1}{4\lambda d}$ (assuming λ is fixed). Therefore

$$h^*(a) = \frac{1}{2}\left(\delta\left(-\frac{1}{4\lambda d}\right) + \delta\left(\frac{1}{4\lambda d}\right)\right).$$

Notice however that under h^* , ρ will only be Bayes optimal if and only if $\frac{1}{4\lambda d} = \frac{d}{2}$, or equivalently, if and only if

$$\lambda = \frac{1}{2d^2}, \quad (8)$$

since at $a = \frac{d}{2}$, $h^*(a) = h^*(a - d)$ and therefore the optimal decision can be randomized.

Notice that $\lambda^* = \frac{1}{4d\sqrt{D}}$ minimizes the Lagrangian, therefore from this constraint and Eq. (8) we can solve for D to obtain the condition where the saddle point equilibrium holds: $D = (\frac{d}{2})^2$.

As a summary, for $D = (\frac{d}{2})^2$, (ρ^*, h^*) is a saddle point equilibrium when ρ^* is defined as in Fig. 2 and

$$h^*(a) = \frac{1}{2}\left(\delta\left(-\frac{d}{2}\right) + \delta\left(\frac{d}{2}\right)\right).$$

Furthermore the probability of error is $\Psi(\rho^*, h^*) = L(\lambda^*, h^*) = \frac{1}{16\lambda d^2} + \lambda D = \frac{1}{8} + \frac{D}{2d^2} = \frac{1}{4}$.

D. Saddle Point Equilibria for $0 \leq D \leq (\frac{d}{2})^2$

For $0 \leq D \leq \frac{d^2}{4}$ we can obtain a saddle point by considering the decision function shown in Fig. 4. Following the same procedure as in the previous section we form the Lagrangian, which again is the same as in Eq. (5) with $\phi(\lambda, a)$ as in Eq. (7) but with ρ defined in Fig. 4.

For this new decision function we find that for $z \in (3, \infty)$, the local maxima of $\phi(\lambda, a)$ as a function of a , occur at $a = 0$, and $a = \pm \frac{z}{4d(z-2)\lambda}$.

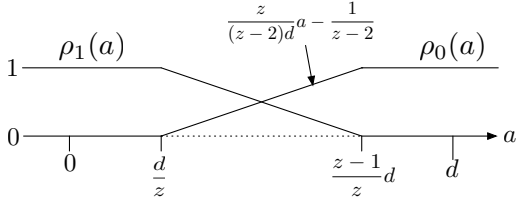


Fig. 4. Bayes optimal decision function for $0 \leq D \leq \left(\frac{d}{2}\right)^2$ when $z = 4$

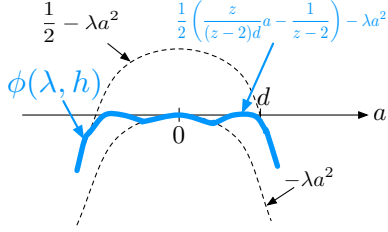


Fig. 5. With ρ defined in Figure 4 the Lagrangian is able to exhibit three local maxima, one of them at the point $a = 0$, which implies that the adversary will use this point whenever the distortion constraints are too severe

In order to force the optimal attack distribution h to place its mass in regions where ρ is optimal, we need to find the value of λ^* that makes all the local maxima of ϕ the same. That is, we need to find λ^* as the solution to:

$$\phi\left(\lambda^*, \pm \frac{z}{4d(z-2)\lambda^*}\right) = \phi(\lambda^*, 0) = 0$$

or more specifically, the solution to:

$$\frac{z^2 - 8d^2(z-2)\lambda^*}{16d^2(z-2)^2\lambda^*} = 0$$

which is $\lambda^* = \frac{z^2}{8d^2(z-2)}$. Any other λ would have implied inactive constraints (D too large or $D = 0$). See Fig. 5.

Replacing this optimal value of λ^* for the optimal values of a^* , we conclude that $a^* = \pm \frac{2d}{z}$. The optimal adversary has therefore the form:

$$h^*(a) = p_0 \delta\left(-\frac{2d}{z}\right) + (1 - 2p_0) \delta(0) + p_0 \delta\left(\frac{2d}{z}\right)$$

With this optimal attack distribution ρ can only form a saddle point (i.e., be Bayes optimal) if $z = 4$, since at $a = \pm \frac{d}{2}$, $h^*(a) = h^*(a - d)$ and the optimal decision can be randomized.

We observe that h^* is an optimal strategy for the adversary as long as

$$\mathbb{E}[a^2] = 2p_0 \left(\frac{d}{2}\right)^2 = D,$$

that is, $p_0 = \frac{2D}{d^2}$. Since the maximum value that p_0 should attain is $\frac{1}{2}$, this implies that this is the

optimal strategy for the adversary for any $D \leq \frac{d^2}{4}$. The probability of error for this saddle point equilibrium is

$$\Psi(\rho^*, h^*) = L(\lambda^*, h^*) = \lambda^* D = \frac{D}{d^2} \leq \frac{1}{4}$$

Note also that we have now two saddle point solutions for the case $D = \frac{d^2}{4}$, with decision functions defined in Fig. 2 and Fig. 4. In fact we can show that we have an infinite set of optimal solutions for this case, since a saddle point is satisfied for any $z \geq 4$. The case $z \rightarrow \infty$ corresponds to Fig. 2.

E. Saddle Point Solutions for $D \geq \frac{d^2}{4}$

As the distortion bound is relaxed for the adversary, the performance of the data hiding algorithm degrades severely in the class $D > \frac{d^2}{4}$. For example for $D = \frac{2}{3}d^2$ the optimum decision function ρ_0 is linear from $-0.5d$ until $1.5d$ and the least favorable $h^*(a)$ puts masses of $\frac{1}{3}$ at 0 , d , and $-d$. However, the probability of error for this case is $\frac{1}{3}$, which might be too large for some applications.

A very interesting formulation to increase the confidence of the decisions made by a classifier is the idea of using a “no decision” output to help the decision function in regions where deciding between the two hypothesis is prone to errors. In this framework we allow $\rho(y)$ to output \neg when not enough information is given in y in order to decide between $m = 0$ or $m = 1$.

To characterize the performance of the detection algorithm we need to analyze the tradeoff between the *completeness* and the *accuracy* of ρ . The *accuracy* of ρ is defined as the probability of correct classification $\Pr[\rho = m]$, while the *completeness* of ρ is defined to be the probability of making a classification $\Pr[\rho \neq \neg]$ (that is, the probability of classifying as either 0 or 1).

We quantify this tradeoff formally using the Bayes risk and using a tradeoff parameter α . Let $C(i, j)$ represent the cost of deciding for i when the true hypothesis is $m = j$. By using as an evaluation metric the probability of error, we have been so far minimizing the expected cost $\mathbb{E}[C(\rho, m)]$ when $C(0, 0) = C(1, 1) = 0$ and $C(1, 0) = C(0, 1) = 1$. We now extend this evaluation metric by incorporating the cost of not making a decision: $C(\neg, 0) = C(\neg, 1) = \alpha$.

It can be shown that if $\alpha \geq \frac{1}{2}$, then the no-decision region can be ignored (the test is complete) if we want to minimize $\mathbb{E}[C(\rho, m)]$. Therefore we only need to concentrate on the case $\alpha < \frac{1}{2}$. It is easy to show that a detection algorithm that minimizes $\Psi(\rho, h) = \mathbb{E}[C(\rho, m)]$ has the following form:

$$\rho^*(a) = \begin{cases} 1 & \text{if } h(a) > \frac{1-\alpha}{\alpha} h(a-d) \\ \neg & \text{if } \frac{1-\alpha}{\alpha} h(a-d) > h(a) > \frac{\alpha}{1-\alpha} h(a-d) \\ 0 & \text{if } h(a) < \frac{\alpha}{1-\alpha} h(a-d) \end{cases}$$

and whenever $h(a)$ equals either $\frac{\alpha}{1-\alpha} h(a-d)$ or $\frac{1-\alpha}{\alpha} h(a-d)$ the decision is randomized between 1

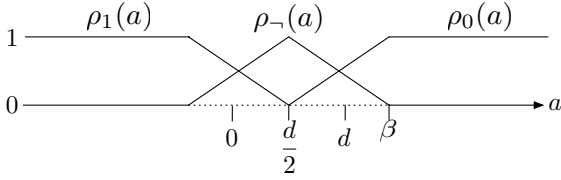


Fig. 6. ρ_{-} represents a decision stating that we do not possess enough information in order to make a reliable selection between the two hypotheses.

and \neg and between \neg and 0 (respectively).

Under our non-blind watermarking model the expected cost is:

$$\begin{aligned} \Psi(\rho, h) = & \\ & \frac{1}{2} \int [\rho_1(x+d) + \alpha \rho_{-}(x+d)] h(x) dx + \\ & \frac{1}{2} \int [\rho_0(x) + \alpha \rho_{-}(x)] h(x) dx \end{aligned}$$

where ρ_i is the probability of deciding for i and where $\rho_0(x) + \rho_{-}(x) + \rho_1(x) = 1$ for all x .

Given ρ , the Lagrangian for the optimization problem of the adversary is:

$$L(\lambda, h) = \frac{1}{2} \int \phi(\lambda, a) h(a) da + \lambda D,$$

where in this case

$$\begin{aligned} \phi(\lambda, a) = & \\ & \rho_1(a+d) + \alpha \rho_{-}(a+d) + \rho_0(a) + \alpha \rho_{-}(a) - \lambda a^2. \end{aligned}$$

Consider now the decision function given in Fig. 6. Following the same reasoning as in the previous chapter, it is easy to show that for $\beta = \frac{3}{2}d$, the maximum values for $L(\lambda, h)$ occur for $a = 0$ and $a = \pm d$. The optimal distribution h has the following form:

$$h^*(a) = \frac{p}{2} \delta(-d) + (1-p) \delta(0) + \frac{p}{2} \delta(d).$$

The decision function ρ is Bayes optimal for this attack distribution only if the likelihood ratio for $a = 0$ is equal to $\frac{1-\alpha}{\alpha}$, (i.e., if $\frac{1-p}{p/2} = \frac{1-\alpha}{\alpha}$) and if the likelihood ratio for $a = \pm d$ is equal to $\frac{\alpha}{1-\alpha}$ (i.e., $\frac{p/2}{1-p} = \frac{\alpha}{1-\alpha}$).

This optimality requirement places a constraint on the completeness of the test $\alpha = 2p - 1$. Furthermore, the distortion constraint implies the adversary will select $\mathbb{E}[a^2] = pd^2 = D$. Since we need $\alpha < \frac{1}{2}$ in order to make use of the no-decision region, the above formulation is thus satisfied for $D \leq \frac{3}{4}d^2$.

IV. CONCLUSIONS AND FUTURE WORK

In this paper we have presented some initial results for the optimal design of detection algorithms against nonparametric adversary models. The natural extension of this work is to consider multi-dimensional signal spaces with embeddings separated by a vector

d. Our early results in this area are very similar to the results in this paper with one key insight; the optimal attack distribution has memory. In particular we cannot condition on any event but rather must consider the full vector distribution $h^*(e_1, e_2, \dots, e_n)$.

Another important extension for this work should be in considering more general and *secure* data embedding algorithms, where the embedding of the signal is randomized with the secret key k , and not a publicly known distance between the different fingerprints.

We believe the idea of a tradeoff between the accuracy and the completeness of a classifier is also of particular importance. It can be shown for example that for the case $D = \frac{2}{3}d^2$ we are able to reduce the probability of error of a complete classifier from $\frac{1}{3}$ to $\frac{1}{6}$ by using a classifier with $\frac{2}{3}$ -completeness (by following the results of subsection E). Future work in understanding this tradeoff and its practical relevance to several applications appears to be very promising.

Finally, another major point to mention is the sensitivity of the optimal solution to the specific distortion function \mathfrak{D} , and whereas the constraints are average distortion constraints $\mathbb{E}[\mathfrak{D}(a)]$ or hard constraints $\mathfrak{D}(a)$. The assumption of a specific distortion function seems to be prevalent in the data hiding literature. However this restricts the usability of any data hiding model, since in practice the adversary will again not be confined to follow a specific distortion constraint. We have explored saddle point solutions for other distortion functions and the solutions change dramatically for each different distortion function considered. In future work we also plan to explore the question of designing saddle point solutions that are optimal for a very large class of distortion constraints.

ACKNOWLEDGEMENTS

Research supported by the U.S. Army Research Office under CIP URI grant No. DAAD19-01-1-0494 and by the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011, to the University of Maryland, College Park.

REFERENCES

- [1] W. Trappe, M. Wu, J. Wang, and K. J. R. Liu, "Anti-collusion fingerprinting for multimedia," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1069–1087, April 2003.
- [2] N. Kiyavash and P. Moulin, "On optimal collusion strategies for fingerprinting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, May 2006, pp. v-405–408.
- [3] T. Liu and P. Moulin, "Error exponents for watermarking game with squared-error constraints," in *Proceedings of the IEEE 2003 International Symposium on Information Theory (ISIT 03)*, July 2003, p. 190.
- [4] A. K. Goteti and P. Moulin, "QIM watermarking games," in *International Conference on Image Processing (ICIP 04)*, vol. 2, October 2004, pp. 717–720.
- [5] P. Moulin and A. Ivanović, "The zero-rate spread-spectrum watermarking game," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1098–1117, April 2003.