Joint Detection-Estimation Games for Sensitivity Analysis Attacks

Maha El Choubassi[a] and Pierre Moulin[b]

University of Illinois, Beckman Institute and ECE Department,
405 N. Mathews, Urbana, IL 61801, USA

ABSTRACT

Sensitivity analysis attacks aim at estimating a watermark from multiple observations of the detector's output. Subsequently, the attacker removes the estimated watermark from the watermarked signal. In order to measure the vulnerability of a detector against such attacks, we evaluate the fundamental performance limits for the attacker's estimation problem. The inverse of the Fisher information matrix provides a bound on the covariance matrix of the estimation error. A general strategy for the attacker is to select the distribution of auxiliary test signals that minimizes the trace of the inverse Fisher information matrix. The watermark detector must trade off two conflicting requirements: (1) reliability, and (2) security against sensitivity attacks. We explore this tradeoff and design the detection function that maximizes the trace of the attacker's inverse Fisher information matrix while simultaneously guaranteeing a bound on the error probability. Game theory is the natural framework to study this problem, and considerable insights emerge from this analysis.

1. INTRODUCTION

Sensitivity analysis attacks belong to the family of watermark removal attacks. Assume that the attacker has a watermarked signal and unlimited access to the watermark detector. When repetitively probed with signals chosen by the attacker, the detector leaks information about the watermark. Hence the attacker may construct test signals that use the leaked information to estimate the watermark. Once the watermark is estimated, it can be removed from the watermarked signal. A multitude of attack algorithms already exist[1–12] for this purpose. In general, a detector is believed to be secure till someone develops an attack that breaks it. Obviously, this approach to security is questionable. One would prefer to quantify the vulnerability of a detector independently of any specific algorithm. In this paper, we provide a model for sensitivity analysis attacks and a complete framework based on Fisher information[14,15] to understand the behavior of the attacker, the detector, and the interaction between them. As a result, we obtain a systematic method to design watermark detectors in the presence of attackers.

There are two problems of interest in the context of sensitivity analysis attacks. The first one is the **watermark detection problem** and represents the basic function of the detector: deciding about the absence or presence of the watermark in its input signal. The measure of detection performance is error probability and classical detection theory and large deviation tools[14] are quite handy in analyzing performance. The second problem is the **watermark estimation problem** and represents the attacker's basic task. The natural measure of estimation performance is the covariance matrix of the estimation error. Moreover, Fisher information matrix inverse[14,15] is a classical algorithm-independent lower bound on the variance of any estimator in particular the *maximum a posteriori* (MAP) estimator. Clearly, the two problems are coupled. The accuracy of the watermark estimation directly depends on the detector, since it is from its output that the attacker gains information about the watermark. Dually, the system designer should take into account the attacker's strategies when choosing a detector. Therefore, the design criteria must include estimation performance along with detection performance. We model the coupling between the two problems as a game between two players: the attacker and the detector. With such a model, the system designer can determine the best attacker's strategy, and then choose the detector accordingly. Our analysis is quite general and widely applicable. For illustration, we present an image watermarking example.

The organization of the paper is as follows. In Section 2, we describe the watermark detection problem and give bounds on the error probability. The watermark estimation problem is presented in Section 3, where we define our model for sensitivity analysis attacks and specify the necessary assumptions. In Section 4, we

formulate the game between the attacker and the detector and derive their optimal strategies. Finally, we conclude in Section 5.

2. WATERMARK DETECTION PROBLEM

In this section, we briefly review some of the material from.[11,12]

2.1 Detection model

The watermark $\mathbf{W} = [W_1, \ldots, W_n]$ is a random vector with iid components following a *symmetric, unimodal* pdf p_W. The watermark is embedded additively into a host signal $\mathbf{S} = [S_1, \ldots, S_n]$ whose samples are iid with pdf p_S. Moreover, $\mathbf{S}$ and $\mathbf{W}$ are independent. The detector knows the watermark $\mathbf{W}$, receives a signal $\mathbf{X} = [X_1, \ldots, X_n]$, and decides whether $\mathbf{W}$ is present or absent in $\mathbf{X}$. More explicitly, the detector decides which of the following two hypotheses is correct:

$$\begin{aligned} \mathrm{H}_1 &: \quad \mathbf{X} = \mathbf{S} + \mathbf{W}: \quad \text{Watermark is present,} \\ \mathrm{H}_0 &: \quad \mathbf{X} = \mathbf{S}: \quad \text{Watermark is absent.} \end{aligned}$$

The sufficient statistic for detection is the log likelihood ratio

$$\begin{aligned} t(\mathbf{X}, \mathbf{W}) &= \ln \frac{p_1(\mathbf{X})}{p_0(\mathbf{X})} \\ &= \sum_{i=1}^{n} \ln \frac{p_S(X_i - W_i)}{p_S(X_i)}. \end{aligned}$$

This detection statistic minimizes the probability of error under the above model for H_0 and H_1. However, if the attacker is able to test the detector with signals of his own choosing, the above model for H_0 and H_1 will be invalid. Therefore, we deliberately assume a mismatched distribution q on the host signal instead of p_S in order to increase the detector's security against such attacks. In,[11,12] we have shown how to choose a mismatched distribution to obtain randomized watermark detectors.[11,12] The detection statistic is

$$t(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{n} \phi_q(X_i, W_i), \tag{1}$$

$$\text{with } \phi_q(X_i, W_i) = \ln \frac{q(X_i - W_i)}{q(X_i)}. \tag{2}$$

Whenever the detector receives a signal $\mathbf{X}$, it evaluates the test statistic and compares it to a threshold τ to make a decision:

$$t(\mathbf{X}, \mathbf{W}) \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

The set of signals $\mathbf{X}$ such that

$$t(\mathbf{X}, \mathbf{W}) = \tau$$

is called the detection boundary. The attacker is assumed to know the distributions p_S and p_W as well as the function $t(\cdot, \cdot)$ and the threshold τ. However, he does not know $\mathbf{W}$.

2.2 Example

Consider the image watermarking application. In this case, the host signal $\mathbf{S}$ is an image. As mentioned before, the family of generalized Gaussian distributions is a good model[16,17] for the DCT coefficients of images. These coefficients are modeled as iid samples from a generalized Gaussian distribution

$$q_\mu(s) = \frac{1}{2b\Gamma\left(1 + \mu^{-1}\right)} \exp\left(-\left|\frac{s}{b}\right|^{\mu}\right) \quad \text{with } b = \sigma\sqrt{\frac{\Gamma(\mu^{-1})}{\Gamma(3\mu^{-1})}}, \tag{3}$$

where $\mu > 0$ is the exponent and σ is the standard deviation. It is known[16,17] that distributions with $\mu \leq 1$ provide a reasonably good model for images. Let μ_s be the actual parameter of the distribution. Instead, the system designer chooses a value μ from the set $0 < \mu \leq 1$ and evaluates the mismatched log likelihood ratio:

$$t(\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{n} \phi_{q_\mu}(X_i, W_i)$$

where

$$\begin{aligned} \phi_{q_\mu}(X_i, W_i) &= \ln \frac{q_\mu(X_i - W_i)}{q_\mu(X_i)} && (4) \\ &= \frac{1}{b^\mu} \left(|X_i|^\mu - |X_i - W_i|^\mu \right). && (5) \end{aligned}$$

The choice of μ is based on detection performance and security requirements of the detector.

2.3 Error exponent

Assume that the detection hypotheses H_0 and H_1 are equiprobable. For a given $\mathbf{W}$, the probability of error $P_e(\mathbf{W})$ is the average of the probability of false alarm, $P_F(\mathbf{W})$, and the probability of miss, $P_M(\mathbf{W})$:

$$\begin{aligned} P_e(\mathbf{W}) &= \frac{1}{2} P_F(\mathbf{W}) + \frac{1}{2} P_M(\mathbf{W}) \\ P_F(\mathbf{W}) &= P\left(t(\mathbf{X}, \mathbf{W}) \geq \tau | H_0, \mathbf{W} \right) && (6) \\ P_M(\mathbf{W}) &= P\left(t(\mathbf{X}, \mathbf{W}) \leq \tau | H_1, \mathbf{W} \right). && (7) \end{aligned}$$

The expected value of $P_e(\mathbf{W})$ is denoted by

$$\begin{aligned} P_e &= \mathbb{E}\left[P_e(\mathbf{W}) \right] \\ &= \frac{1}{2} \mathbb{E}\left[P_F(\mathbf{W}) \right] + \frac{1}{2} \mathbb{E}\left[P_M(\mathbf{W}) \right]. \end{aligned}$$

With $\tau_0 = \frac{\tau}{n}$ and $r > 0$, the large deviation bounds on $\mathbb{E}\left[P_F(\mathbf{W}) \right]$ and $\mathbb{E}\left[P_M(\mathbf{W}) \right]$ are

$$\begin{aligned} \mathbb{E}\left[P_F(\mathbf{W}) \right] &\leq \exp\left[-n \left\{ r\tau_0 - \ln \mathbb{E}\left[\left(\frac{q(S-W)}{q(S)} \right)^r \right] \right\} \right], && (8) \\ \mathbb{E}\left[P_M(\mathbf{W}) \right] &\leq \exp\left[-n \left\{ -r\tau_0 - \ln \mathbb{E}\left[\left(\frac{q(S-W)}{q(S)} \right)^r \right] \right\} \right]. && (9) \end{aligned}$$

Moreover, performing an analysis similar to our previous work,[11,12] we can show the following:

- The necessary and sufficient condition for exponential decay of the probability of error is

$$\mathbb{E}\left[\ln \frac{q(S-W)}{q(S)} \right] < 0.$$

- The threshold that maximizes the error exponent is $\tau_0^* = 0$.
- The error exponent for P_e is

$$\begin{aligned} \beta(q) &= -\inf_{r>0} \ln \mathbb{E}\left[\left(\frac{q(S-W)}{q(S)} \right)^r \right] && (10) \\ \text{and } P_e &\leq e^{-n\beta(q)}. \end{aligned}$$

The expectations are with respect to p_S and p_W. For the image watermarking example given in Section 2.2, the pdf q is replaced by the pdf q_μ, and the pdf p_S by the pdf q_{μ_s}.

3. WATERMARK ESTIMATION PROBLEM

In this section, we present the watermark estimation problem, we define our model for sensitivity analysis attacks with the necessary assumptions for the validity of the model, and finally we evaluate the asymptotic estimation error.

3.1 Attack model

For sensitivity analysis attacks,[8–12] the attacker tries to generate signals on the detection boundary based on the binary decisions at the detector's output. Let L be the number of these signals, denoted by $\mathbf{Y}^m = [Y_1^m, \ldots, Y_n^m]$, $1 \leq m \leq L$. For each test signal $\mathbf{Y}^m$, $1 \leq m \leq L$, the attacker first estimates the test statistic $t(\mathbf{Y}^m, \mathbf{W})$ using a search algorithm, e.g., binary search. For reasons to be clarified below, the attacker randomizes the initial step of the search algorithm. Let K be the number of steps of the search algorithm; in each step the attacker probes the detector once. Therefore, the total number of detector's probes is KL. The precision in the measurements of $t(\mathbf{Y}^m, \mathbf{W})$ improves exponentially with K. For example, for binary search, if the width of the search interval is $A > 0$, after K steps, the precision is $2^{-K}A$. This would imply that an infinite number of search steps yields perfect measurements. However, physical devices have finite precision arithmetic, and there is always noise in the measurements taken by the attacker. Let σ^2_{Floor} denote the average energy of that noise. The variance of the total measurement noise is equal to $\sigma_N^2 = \sigma^2_{\text{Floor}} + 4^{-K}A^2$ for binary search. It is enough for the attacker to use

$$K = a + \log_2 \frac{A}{\sigma_{\text{Floor}}},$$

for σ_N^2 to be approximately equal to σ^2_{Floor}, where $a > 1$ is some constant. We therefore propose the following model for sensitivity analysis attacks. We first define ρ as the number of measurements per component of $\mathbf{W}$:

$$\rho \triangleq \frac{L}{n}. \tag{11}$$

The attacker generates signals $\mathbf{Y}^m$, $1 \leq m \leq L$, and takes noisy measurements of the detection statistic:

$$T_m \quad = \quad t(\mathbf{Y}^m, \mathbf{W}) + N_m, \quad 1 \leq m \leq L = \rho n, \tag{12}$$

where N_m is the measurement noise whose statistics are modeled below. By randomizing the initialization step of the search algorithm, the attacker also randomizes the noise due to both sources, finite number of steps of the search algorithm and finite precision arithmetic of real devices. Note that the attacker has interest in random instead of deterministic noise, to obtain diverse measurements T_m in (12) even for the same probing signal $\mathbf{Y}_m = \mathbf{Y}$. Combining these measurements, the attacker reduces the noise and results in a finer measurement of $t(\mathbf{Y}, \mathbf{W})$. Finally, having all the measurements in (12), the attacker then estimates the watermark. This model naturally lends itself to analysis based on classical estimation theory tools.

Assume* that the watermark components W_i, $1 \leq i \leq n$, can be estimated one at a time using the following procedure. The attacker generates a series of signals $\mathbf{Y}^m$, $1 \leq m \leq \rho$. Each signal $\mathbf{Y}^m$ has only one nonzero component at location i, i.e., $\mathbf{Y}^m = [0, \ldots, 0, Y^m, 0, \ldots, 0]$. From (1) and (12), the attacker takes noisy measurements of the form

$$T_m \quad = \quad \phi_q(Y^m, W) + N_m, \quad 1 \leq m \leq \rho. \tag{13}$$

Here, we denoted the i^{th} component of the watermark by W instead of W_i for notational simplicity. The signal-to-noise ratio of the measurement channel in (13) is given by

$$\frac{1}{\sigma_N^2} \mathbb{E}\left[\left(\phi_q(Y, W) \right)^2 \right]. \tag{14}$$

Additionally, we make the following assumptions:

*We address the more general case in Section 5 and we discuss it in details in our journal paper,[13] currently in preparation.

(A1) The random variables $Y^m,\ 1 \le m \le \rho$, are independent of $\mathbf{W}$ and iid generated from a pdf p_Y with variance σ_Y^2.

(A2) The measurement noise N_m, $1 \le m \le \rho$ is iid Gaussian, with mean zero and variance σ_N^2. Moreover, the noise is independent of $t(\mathbf{Y}^m, \mathbf{W})$, $1 \le m \le \rho$.

The randomness of the initial search step justifies the independence assumption of the noise in (A2).

3.2 Estimation error

Define $T^{1:\rho} \triangleq \{T_m, 1 \le m \le \rho\}$ and $Y^{1:\rho} \triangleq \{Y^m, 1 \le m \le \rho\}$. The MAP estimator of W given $T^{1:\rho}$ and $Y^{1:\rho}$ is

$$\begin{aligned} \hat{W}_\rho &= \arg\max_w \ln p(w|T^{1:\rho}, Y^{1:\rho}) \\ &= \arg\min_w \left(\frac{1}{2\sigma_N^2} \sum_{m=1}^{\rho} \left(T_m - \phi_q(Y^m, w)\right)^2 - \ln p_W(w) \right), \end{aligned} \tag{15}$$

and is the solution to a regularized nonlinear least squares problem. The variance of the estimation error for any estimator of W is bounded below by the inverse of the total Fisher information,[14,15] denoted as $J_{total,\rho}$. In particular, the variance of the MAP estimation error $\hat{W}_\rho - W$ is bounded as

$$\mathbb{V}\mathrm{ar}\left[\hat{W}_\rho - W\right] \ge \frac{1}{J_{\mathrm{total},\rho}}. \tag{16}$$

As $\rho \to \infty$, the normalized total Fisher information converges to the limit

$$\lim_{\rho\to\infty} \frac{1}{\rho} J_{\mathrm{total},\rho} = J, \tag{17}$$

$$J \triangleq \frac{1}{\sigma_N^2} \mathbb{E}\left[\left(\frac{d}{dW} \phi_q(Y, W) \right)^2 \right]. \tag{18}$$

Therefore, for a large number ρ of measurements per component, we approximately have from (16) and (17)

$$\mathbb{V}\mathrm{ar}\left[\hat{W}_\rho - W\right] \ge \frac{1}{\rho J}. \tag{19}$$

The lower bound in (19) is achieved by the MAP and the ML estimators in the limit as $\rho \to \infty$, in which case precise estimation of W is guaranteed. Hence, J is a fundamental measure of the accuracy of the estimate $\hat{W}_\rho$. The attacker and the detector have conflicting goals. The attacker wants J to be as large as possible, while the detector wants the opposite.

4. GAME BETWEEN ATTACKER AND DETECTOR

In this section, we investigate the game[18] between the attacker and the detector and derive their optimal strategies. Under the necessary regularity assumptions, we explicitly express the dependency of J in (18) on the distributions p_Y and q and we denote it as

$$J(p_Y, q) = \frac{1}{\sigma_N^2} \mathbb{E}\left[\left(\frac{d}{dW} \phi_q(Y, W) \right)^2 \right], \tag{20}$$

which is viewed as the utility function for the watermark estimation problem. The utility function for the watermark detection problem is the error exponent of the detector in (10):

$$\beta(q) = -\inf_{r>0} \ln \mathbb{E}\left[\left(\frac{q(S-W)}{q(S)} \right)^r \right]. \tag{21}$$

The attacker plays a role only in the watermark estimation problem and affects directly the first utility function, $J(p_Y, q)$. The choice of the detector is critical for both the watermark estimation and detection problems. We model the dynamics of the relation between the attacker and the detector as a zero sum game with cost functions $J(p_Y, q)$ and $\beta(q)$. We formulate the game as

$$\min_{q:\ \beta(q)\geq\beta^*} \max_{p_Y} J(p_Y, q), \tag{22}$$

where $\beta^* > 0$ is a lower bound on the error exponent. The detector is the leader in this game: he chooses and reveals q publicly, and the attacker responds by choosing his strategy which depends on q. The attacker wants to construct the best estimator of W and chooses $p_Y^*(q)$ that maximizes $J(p_Y, q)$. All choices of $p_Y^*(q)$ satisfying the assumptions in Section 3.1 are feasible. The system designer wants the opposite. He therefore chooses q that minimizes $J(p_Y^*(q), q) = \max_{p_Y} J(p_Y, q)$ under the constraint $\beta(q) \geq \beta^*$. The bound on the error exponent guarantees a specified level of detection performance.

4.1 Optimal attacker's strategy

First, we assume that the following expectation is finite surely (p_W):

$$\begin{aligned} g_q(Y) &\triangleq \frac{1}{\sigma_N^2}\,\mathbb{E}\left[\left(\frac{d}{dW}\phi_q(Y,W)\right)^2\middle| Y\right] < \infty \\ &= \frac{1}{\sigma_N^2}\mathbb{E}\left[\left(\frac{d}{dW}\ln q(Y-W)\right)^2\middle| Y\right], \end{aligned} \tag{23}$$

where the last equality follows from (2). Defining

$$h_q(t) \triangleq \frac{1}{\sigma_N^2}\left(\frac{d}{dt}\ln q(t)\right)^2 = \frac{1}{\sigma_N^2}\left(\frac{\dot{q}(t)}{q(t)}\right)^2, \tag{24}$$

we have from (2) and (23) that

$$\begin{aligned} g_q(Y) &= \mathbb{E}\left[h_q(Y-W)|\,Y\right] \\ &= \left(p_W * h_q\right)(Y). \end{aligned} \tag{25}$$

If q is symmetric, then h_q and g_q are also symmetric. We rewrite (20) as

$$J(p_Y, q) = \int_{-\infty}^{\infty} p_Y(y) g_q(y) dy. \tag{26}$$

From (26), the utility function is the expected value of the function $\tilde{g}_q$ with respect to p_Y and the attacker's maximization problem in (22) takes the form

$$\max_{p_Y} \int_{-\infty}^{\infty} p_Y(y) g_q(y) dy.$$

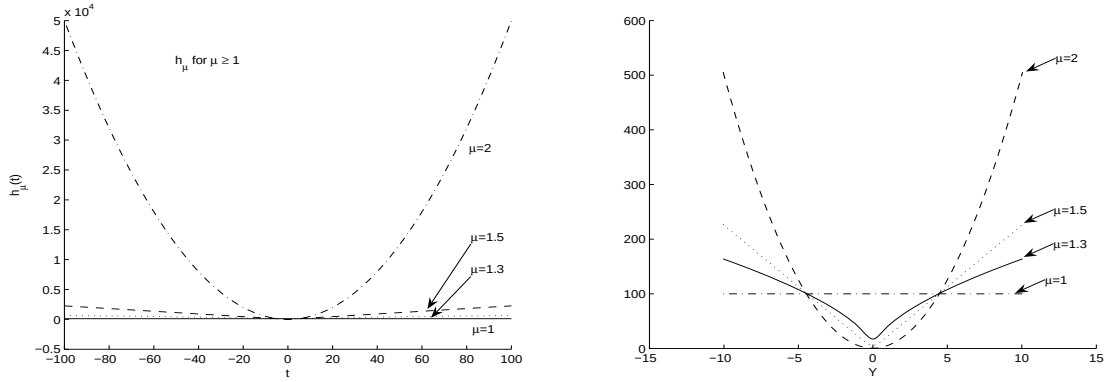
The function g_q is the result of the convolution of the unimodal symmetric function p_W with the symmetric function h_q. When h_q is unimodal, respectively, g_q is unimodal and the optimal distribution $p_Y^*(q)$ for the attacker in the game (22) is equal to an impulse at zero. The resulting value of $J(p_Y^*(q), q)$ is $g_q(0)$.

Example. Let us revisit our image watermarking example in Section 2.2. The utility function in (20) becomes

$$J(p_Y, q_\mu) = \mathbb{E}\left[g_\mu(Y)\right] \tag{27a}$$

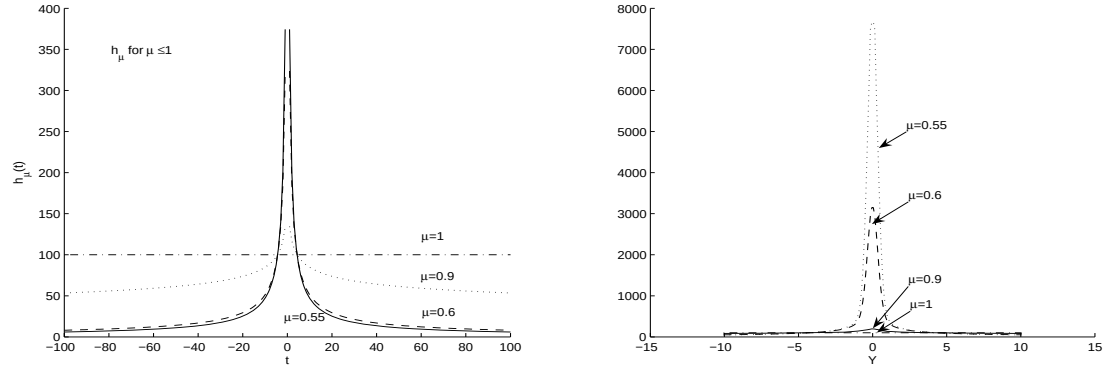
$$\text{where } g_\mu(Y) \triangleq \mathbb{E}\left[h_\mu(Y-W)|\,Y\right], \quad \mu > 0.5 \tag{27b}$$

$$h_\mu(t) = \frac{\mu^2}{\sigma_N^2 b^{2\mu}}|t|^{2\mu-2}, \quad \exists\, \forall t \neq 0. \tag{27c}$$



(a) (b)

Figure 1: (a) Plots of h_μ for $\mu \geq 1$. (b) Plots of $\tilde{g}_\mu$ for $\mu \geq 1$, Gaussian $p_W \sim N(0, 0.11)$.



(a) (b)

Figure 2: (a) Plots of h_μ for $\mu \leq 1$. (b) Plots of $\tilde{g}_\mu$ for $\mu \leq 1$, Gaussian $p_W \sim N(0, 0.11)$.

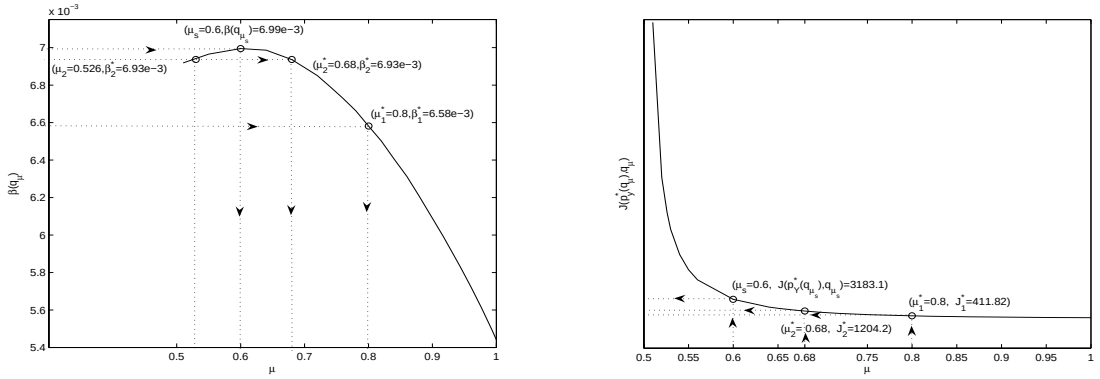
Clearly, both g_μ and h_μ are positive and symmetric.

For $\mu = 1$, it is straightforward to show that $g_\mu(Y) = J(p_Y, q) = \frac{2}{\sigma_N^2 \sigma^2}$ (constant) independently of the attacker's choice of p_Y. For $\mu > 1$, $h_\mu(t)$ is symmetric and strictly increasing for $t > 0$ (please see Figure 1a). The resulting $g_\mu(Y)$ is nondecreasing for $Y > 0$, and therefore is not guaranteed to be bounded. For example, it is straightforward to check that $\sup g_\mu(Y) = \infty$, for Gaussian p_W (see Figure 1b). In this case, $J(p_Y, q_\mu)$ does not admit a maximum since it is unbounded. For $0.5 < \mu < 1$, h_μ is symmetric unimodal (see Figure 2a), and the best attacker's strategy is $p_Y^*(q)$ is an impulse at zero, and $J(p_Y^*, q_\mu) = g_\mu(0)$.

4.2 Optimal detector's strategy

As mentioned in the previous section, for symmetric unimodal $h_q(t)$, the maximum value of $J(p_Y, q)$ the attacker can attain is $J(p_Y^*(q), q) = g_q(0)$. Therefore, the detector chooses q^* minimizing $g_q(0)$ subject to $\beta(q) \geq \beta^*$. Moreover, we know that the error exponent $\beta(q)$ is maximized at $q = p_S$, the matching distribution. Therefore, $\beta(q)$ is concave for q in a small neighborhood around p_S. In the following example, we consider the particular family of GGD detectors.

Example. For the image watermarking example of Section 2.2, the detector's designer seeks the exponent μ^* that minimizes the cost $J(p_Y^*(q_\mu), q_\mu)$, subject to $\beta(q_\mu) \geq \beta^*$. As explained in Section 4.1, the optimal exponent μ^* must be in the range $0.5 < \mu^* \leq 1$. To illustrate what $J(p_Y^*(q_\mu), q_\mu)$ and $\beta(q_\mu)$ look like, we consider Gaussian distribution $p_W \sim N(0, \sigma_W^2)$. Obviously, p_W is unimodal symmetric. From Section 4.1, the best attacker's strategy is $p_Y^*(q)$ is a point mass concentrated at zero and the attacker's value of the game is



(a) (b)

Figure 3: (a) Error exponent versus μ for $\mu_s = 0.6$. (b) Attacker's Fisher information $J(p_Y^*(q_\mu), q_\mu)$ versus μ.

$g_\mu(0)$. From (25) and (27c), $g_\mu(0)$ is given by

$$\begin{aligned} g_\mu(0) &= (p_W * h_\mu)(0) \\ &= \frac{\mu^2}{\sigma_N^2 b^{2\mu}\sqrt{2\pi\sigma_W^2}} \int_{-\infty}^{\infty} |t|^{2\mu-2} \exp\left(-\frac{t^2}{2\sigma_W^2}\right) dt, \end{aligned}$$

for $0.5 < \mu \leq 1$. We evaluate the above integral[19] and we conclude that the payoff is

$$J(p_Y^*(q_\mu), q_\mu) = \frac{2\mu^2}{\sqrt{2\pi}\sigma_N^2 b^{2\mu}} \frac{\Gamma(2\mu-1)}{\sigma_W^{2(1-\mu)}} D_{-2\mu+1}(0), \tag{28}$$

where $0.5 < \mu \leq 1$, and $D_{-2\mu+1}(x)$ is a parabolic cylinder function.[19] Assume that the actual GGD exponent is $\mu_s = 0.6$ and for an embedding signal-to-noise ratio of approximately 20 dB, let $\sigma^2 = 10$ and $\sigma_W^2 = 0.11$. Also let the variance of the measurement noise be $\sigma_N^2 = 0.2 \times 10^{-2}$. From Figure 3a, the largest error exponent $\beta(q_{\mu_s}) = 6.99 \times 10^{-3}$ is achieved by the matched detector, i.e., with $\mu = \mu_s = 0.6$. Regarding estimation accuracy, Figure 3b shows that $J(p_Y^*(q_\mu), q_\mu)$ decreases with μ. Hence the system designer is motivated to select the largest possible μ as μ^*, not necessarily the same $\mu = \mu_s$. The floor β^* on the error exponent $\beta(q_\mu)$ results in an upper bound on μ as seen in Figure 3a. Therefore, if $\beta^* < \beta(q_{\mu_s})$, then the designer sets μ^* equal to the above upper bound. For β^* equal to the maximum exponent $\beta(q_{\mu_s})$, we have $\mu^* = \mu_s$, and the payoff of the game is $J(p_Y^*(q_{\mu_s}), q_{\mu_s}) = 3183.1$ (see Figures 3a and 3b). However, for smaller exponents $\beta_1^* = 6.58 \times 10^{-3}$ and $\beta_2^* = 6.93 \times 10^{-3}$, the best choices for the detector are respectively $q_{\mu_1^*}$ and $q_{\mu_2^*}$ with $\mu_1^* = 0.8$ and $\mu_2^* = 0.68$, and the payoffs of the game are respectively $J_1^* = 411.2$ and $J_2^* = 1204.2$.

5. DISCUSSION AND CONCLUSION

In Section 3, the attacker generates one-dimensional signals $\mathbf{Y}^m$. A more general case is when the signals $\mathbf{Y}^m, 1 \leq m \leq L$, have dimension n. More specifically, the attacker draws the n components of $\mathbf{Y}^m = [Y_1^m, \ldots, Y_n^m]$ for $1 \leq m \leq L$, iid from a distribution p_Y. Combining (1) and (12), the attacker's measurements are

$$T_m = \sum_{i=1}^{n} \phi_q(Y_i^m, W_i) + N_m, \quad 1 \leq m \leq L = \rho n. \tag{29}$$

Denoting $\hat{\mathbf{W}}_L$ as the MAP estimator, we conclude that in the asymptotics, the variance of the estimation error per component is given by

$$\frac{1}{n}\text{trace}\left(\mathbb{C}\text{ov}\left[(\hat{\mathbf{W}}_L - \mathbf{W})\right]\right) \simeq \frac{1}{\rho J'} \tag{30}$$

$$\text{where } J' \triangleq \frac{1}{\sigma_N^2}\mathbb{V}\text{ar}\left[\frac{d}{dW}\phi_q(Y, W)\right].$$

Comparing J' with J in (18), the only difference is that J' has the variance of $\frac{d}{dW}\phi_q(Y,W)$ in its expression while J has the second order moment. Consequently, the variance of estimation error for one-dimensional signals is less or equal than the variance for multidimensional signals. This result is expected: For the first type of signal, the only degradation in the measurement channel (13) for a watermark component W is the noise N_m. For the second type of signal, there is interference in (29) from the other watermark components W_j, $1 \leq j \neq i \leq n$, in addition to the noise N_m, when estimating W_i for some $i \in \{1, \ldots, n\}$. Refer to our paper[13] for the a more thorough analysis.

In conclusion, we designed a model for sensitivity analysis attacks. Leveraging tools from estimation, detection, and game theory, we developed a framework that includes the goals and the strategies of the attacker and the detector, in addition to the interaction between these players. The analysis in the paper is fairly general and is not constrained to specific scenarios. With the results of our analysis, the trade-off between security and detection performance can be well understood. Moreover, the detector's designer can expect the worst behavior of the attacker, and choose the detector's specifications accordingly.

REFERENCES

[1] I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *Proc. International Conference on Image Processing (ICIP)*, only CD version of proceedings available, Santa Barbara, CA, 1997.

[2] J. P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proceedings of the Workshop of Information Hiding*, Portland, OR, April 1998, pp. 258-272.

[3] T. Kalker, J. P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *Proc. International Conference on Image Processing (ICIP)*, vol. 1, pp. 425-429, Chicago, IL, October 1998.

[4] A. Tewfik and M. Mansour, "LMS-based attack on watermark public detectors," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Rochester, NY, September 2002, pp. 649-652.

[5] P. Comesãna, L. Pérez-Freire, and F. Pérez-Gonzaléz, "The return of the sensitivity attack," in *Proc. IWDW*, Siena, Italy, 2005, pp. 260-274.

[6] R. Venkatesan and M.H. Jakubowski, "Randomized detection for spread-spectrum watermarking: Defending Against Sensitivity and Other Attacks," *Proc. ICASSP*, Philadelphia PA, 2005.

[7] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. San Francisco: Morgan Kaufmann, 2001.

[8] M. El Choubassi and P. Moulin, "A new sensitivity analysis attack," in *Proc. SPIE Conf.*, San Jose, CA, January 2005, pp. 734-745.

[9] M. El Choubassi and P. Moulin, "Noniterative algorithms for sensitivity analysis attacks," *IEEE TIFS*, vol. 2, no. 2, pp. 113-126, 2007.

[10] M. El Choubassi and P. Moulin, "Sensitivity analysis attacks against randomized detectors," in *Proc. International Conference on Image Processing (ICIP)*, San Antonio, TX, 2007.

[11] M. El Choubassi and P. Moulin, "On the fundamental tradeoff between watermark detection performance and robustness against sensitivity analysis attacks," in *Proc. SPIE*, San Jose, CA, 2006, pp. 575-586.

[12] M. El Choubassi and P. Moulin, "On reliability and security of randomized detectors against sensitivity analysis attacks," *to appear in IEEE TIFS*.

[13] M. El Choubassi and P. Moulin, "Joint Estimation and Detection Games for Sensitivity analysis attacks," *IEEE TIFS*, in preparation.

[14] H.V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1988.

[15] H.L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley & Sons, New York, 1968.

[16] F. Müller, "Distribution shape of two-dimensional DCT coefficients of natural images," *Electron. Lett.*, vol. 29, no. 22, pp. 1935–1936, Oct. 1993.

[17] J. R. Hernández, M. Amado, and F. Pérez-González, "DCT-domain watermarking techniques for still images: detector performance analysis and a new structure," *IEEE Trans. Signal Processing*, vol. 9, no. 1, pp. 55—68, January 2000.

[18] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1995.

[19] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th. ed. Orlando, FL: Academic Press Inc., 1980.