Detection of Content Adaptive LSB Matching (a Game Theory Approach)

Tomáš Denemark and Jessica Fridrich

Department of ECE, SUNY Binghamton, NY, USA

ABSTRACT

This paper is an attempt to analyze the interaction between Alice and Warden in Steganography using the Game Theory. We focus on the modern steganographic embedding paradigm based on minimizing an additive distortion function. The strategies of both players comprise of the probabilistic selection channel. The Warden is granted the knowledge of the payload and the embedding costs, and detects embedding using the likelihood ratio. In particular, the Warden is ignorant about the embedding probabilities chosen by Alice. When adopting a simple multivariate Gaussian model for the cover, the payoff function in the form of the Warden's detection error can be numerically evaluated for a mutually independent embedding operation. We demonstrate on the example of a two-pixel cover that the Nash equilibrium is different from the traditional Alice's strategy that minimizes the KL divergence between cover and stego objects under an omnipotent Warden. Practical implications of this case study include computing the loss per pixel of Warden's ability to detect embedding due to her ignorance about the selection channel.

Keywords: Adaptive steganography, game theory, steganalysis, LSB matching, distortion

1. INTRODUCTION

Content-adaptive steganography constrains its embedding changes to those parts of the image where one expects their detection to be hard(er). In steganography that minimizes an additive embedding distortion,[1] each pixel is changed with probability (change rate)

$$\beta_i = \frac{\exp(-\lambda\rho_i)}{1+\exp(-\lambda\rho_i)}, \tag{1}$$

where $\rho_i \geq 0$ is the cost (distortion) of changing pixel i. The parameter $\lambda \geq 0$ is determined from the payload constraint

$$\frac{1}{n}\sum_{i=1}^{n} h(\beta_i) = \alpha, \tag{2}$$

where n is the total number of pixels and α is the relative payload expressed in bits per pixel (bpp) or nats per pixel (npp) depending on the logarithm base of the binary entropy function $h(x) = -x\log x - (1-x)\log(1-x)$.

The costs ρ_i are usually obtained using some deterministic rule applied to the cover image. They could be based, e.g., on the local pixel variance, on how much a change in pixel i affects some feature-based representation of the cover image,[2] or on the relative change to transform coefficients.[3–5] Embedding costs can also be determined based on the rounding error when computing the cover from a precover source[5–9] or by imposing some heuristic rules.[10] Alternatively, the costs can be optimized to minimize the distortion in some model space[11] or analytically computed to minimize the KL divergence between the cover and stego distributions for a fixed cover model.[?]

Since the stego image is a slightly modified version of the cover, the Warden could in theory estimate the set of change rates β_i, $i = 1, \ldots, n$, which we call in this paper the probabilistic selection channel. Granted, she would need to know the payload α or λ. The accuracy with which the Warden can estimate β_i depends on the payload (she will be more accurate when the payload is small) and, primarily, on the rule used to compute ρ_i.[12]

E-mail: {tdenema1,fridrich}@binghamton.edu; http://dde.binghamton.edu

Since the introduction of content-adaptive stego schemes, it has long been hypothesized that any information about the selection channel given to the Warden could be used to her advantage to improve her detector. However, despite the monumental effort by teams participating in the BOSS competition,[13] whose aim was to attack the content-adaptive algorithm HUGO,[2] no one was able to utilize HUGO's adaptivity to improve their attacks. The authors of [12], however, showed that, at least for naive adaptive LSB replacement, when Alice only embeds in $n\alpha$ pixels with the largest β_i, when the Warden approximately knows the embedding probabilities, the steganography becomes (almost) sequential, and for sequential LSBR one can build better detectors than when the embedding changes are randomly spread.[?] This indicates a potential gain in security for Alice should she embed with β_i's that are not available to the Warden.

Recently, Böhme et al. [**?**] introduced the Game Theory as a possible framework to incorporate Warden's ignorance in Steganography.* In this article, we investigate this intriguing research direction for the modern embedding paradigm in which the sender minimizes an additive distortion function instead of the naive LSBR, while executing the embedding changes using LSB matching (LSBM) rather than LSBR.† We start by pointing out that the knowledge of the selection channel is only one side of the coin. The other is *how detectable* the embedding changes are. Consider a cover image whose one half is composed of random noise while the other half is a completely flat content. In this case, it is far better for the steganographer to embed in the random part even though the Warden knows it. In fact, the sender could even hide data with perfect security using naive embedding if she knew the cover model. And even if she did not and used a mutually independent embedding operation,[16] the statistical spread of Warden's detection statistic will be much larger in the noisy part of the image than if Alice spread her message by utilizing the flat part, where she is totally detectable. Obviously, the information about the selection channel available to the Warden may be a weakness only to a degree depending on how detectable the changes are at each pixel. This is exactly the problem that we focus on in this paper.

We consider the following two options for Alice:

1. **[Omnipotent Warden]** Assuming an omnipotent Warden, she also knows Alice's actions (her embedding probabilities β_i). Alice approaches the problem using information theory and determines the costs (β_i's) to communicate her payload with minimal KL divergence between cover and stego objects (see, e.g., [**?**]) or in some other manner discussed above. In this case, the Warden knows the β_i's exactly and uses the likelihood-ratio test to optimally detect Alice's embedding. The usual argument justifying this scenario relies on the Kerckhoffs' principle and the fact that the Warden can estimate the costs from the stego image with high enough accuracy, which is reasonable at least when the payload is small and the individual changes are more spread out.

2. **[Ignorant Warden]** This scenario is more realistic in that we assume that the Warden has no information about the probabilities with which Alice changes each pixel. In this case, Alice may choose to deviate from the optimal embedding probabilities derived under the omnipotent Warden and instead embed suboptimally with a different set of probabilities with a hope that she can communicate more securely because an uninformed Warden will likely have a mismatched and thus less powerful detector. Postulating Warden's detection error as the payoff function, the question is whether there exists a Nash equilibrium – a set of embedding probabilities for Alice and a set of detection probabilities for the Warden such that it will not be advantageous for either to change their strategy.

In the next section, we describe the cover model used in this paper as well as the adaptive, mutually independent LSBM as an example of the most common embedding operation used today. In Section 3, we define the Warden's detector, the payoff function, and the players' strategies. Due to the smoothness of the payoff function, in Section 4 it is argued that the game admits a solution in pure strategies, which can be found using a gradient search. Analysis of the game-theoretic formulation of the interplay between Alice and the Warden is the subject of Section 4, where we find the Nash equilibrium numerically and assess the impact of the above-mentioned two embedding strategies on statistical detectability. For repeatability of the results, in Section 6 we provide some technical details regarding the approximations and methods for controlling the numerical error in our experiments. The paper is summarized in Section 7.

*The first game-theoretic approach to steganography appeared in [14].

†This is how essentially all current most secure steganographic schemes for empirical covers[15] work.

1.1 Notation

For better readability, we introduce the following notation for a Gaussian density with mean μ and variance σ^2:

$$f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{3}$$

and a $\{-1,0,1\}$-mixture of Gaussian densities with a parameter $0 \le \beta \le 1/2$:

$$f_\beta(x;\sigma^2) = \frac{\beta}{2}f(x;-1,\sigma^2) + (1-\beta)f(x;0,\sigma^2) + \frac{\beta}{2}f(x;1,\sigma^2). \tag{4}$$

Capital letters and symbols will be reserved for random variables whose realizations will be denoted with the corresponding lower-case letters. Vectors will be typeset in boldface.

2. COVER MODEL AND EMBEDDING METHOD

2.1 Cover model

As in [17], the cover model we use in this article is a simplified model for one channel of a raw imaging sensor output. Assuming one takes multiple images of the same scene, the images will only differ in the image acquisition noise, which is a conglomerate of random processes as well as fixed distortions. They include the shot noise (also known as the photonic noise), the readout noise, electronic noise, charge transfer noise, dark current, fixed-pattern noise, cross-talk, blooming, defective pixels, and others.[18] Ignoring the noise components that stay consistently the same when taking the same picture multiple times with the same camera settings (the fixed pattern noise, dark current, hot and dead pixels), the remaining noise components are random in nature and well modeled as an independent (but not necessarily identically distributed) Gaussian noise. Even though the mean of each pixel output is a non-negative quantity $\mu_i \ge 0$ representing the light intensity of the noise-free real scene, when giving the knowledge of μ_i's to both Alice and the Warden, the results derived in this paper will not depend on μ_i, which allows us to adopt the following simple model for a cover consisting of n pixels:

$$\mathbf{X} = (X_1,\dots,X_n), \quad X_i \sim \mathrm{N}(0,\sigma_i^2), \quad i = 1,\dots,n. \tag{5}$$

Due to the independence of pixels, without any loss on generality we can assume that $\sigma_i \le \sigma_{i+1}$, e.g., the first pixel is the least "noisy" while the last pixel is the most noisy.

We note that (5) is not a realistic model for natural images due to complex dependencies among pixels that are inevitably introduced during postprocessing inside the camera, which may include demosaicking, white balance adjustment, color correction, gamma correction, various types of filtering, lens distortion correction, and lossy JPEG compression.

2.2 Embedding method

Although the analysis in this paper is easily extendable to any mutually independent embedding,[16] for simplicity we selected LSB matching. The LSBM will be adaptive to the content – we expect it to prefer changing the more noisy pixels (pixels with a higher variance). Alice changes pixel x_i by ±1 with probability $\beta_i^{(\mathrm{A})}$ and leaves it unchanged with probability $1-\beta_i^{(\mathrm{A})}$. Denoting the stego image $\mathbf{Y} = (Y_1,\dots,Y_n)$,

$$\Pr(Y_i = x_i + s_i) = \begin{cases} \beta_i^{(\mathrm{A})}/2 & \text{for } s_i = -1, \\ 1-\beta_i^{(\mathrm{A})} & \text{for } s_i = 0, \\ \beta_i^{(\mathrm{A})}/2 & \text{for } s_i = 1. \end{cases} \tag{6}$$

Therefore, each stego pixel follows a Gaussian mixture:

$$Y_i \sim f_{\beta_i^{(\mathrm{A})}}(x,\sigma_i^2). \tag{7}$$

When Alice embeds α npp, the embedding probabilities must satisfy the payload constraint

$$\sum_{i=1}^{n} h(\beta_i^{(\mathrm{A})}) = \alpha n. \tag{8}$$

Thus, Alice's action is captured with $n-1$ parameters: $\beta_i^{(\mathrm{A})}$, $i = 1, \ldots, n-1$ as $\beta_n^{(\mathrm{A})}$ is determined from the payload constraint (8). Note that this embedding paradigm is quite realistic and is used almost solely in all modern content adaptive steganographic schemes.

3. WARDEN'S DETECTOR, PAYOFF FUNCTION, AND STRATEGIES

3.1 Warden's detector

The Warden will be running a simple binary hypothesis test:

$$\mathrm{H}_0: \; X_i \sim f(x, 0, \sigma_i^2), \; \forall i, \tag{9}$$
$$\mathrm{H}_1: \; X_i \sim f_{\beta_i^{(\mathrm{W})}}(x, \sigma_i^2), \; \forall i, \tag{10}$$

where $\beta_i^{(\mathrm{W})}$ are the change rates *assumed by the Warden* that satisfy the same payload constraint:

$$\sum_{i=1}^{n} h(\beta_i^{(\mathrm{W})}) = \alpha n. \tag{11}$$

The null hypothesis corresponds to observing a cover image, while the alternative hypothesis corresponds to a stego object.

Given an image $\mathbf{x} = (x_1, \ldots, x_n)$, the Warden uses the Likelihood Ratio Test (LRT) as her detector:[‡]

$$T(\mathbf{x}; \boldsymbol{\beta}^{(\mathrm{W})}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} \frac{f_{\beta_i^{(\mathrm{W})}}(x_i, \sigma_i^2)}{f(x_i, 0, \sigma_i^2)}. \tag{12}$$

3.2 Payoff function

As a payoff function, we need some scalar characteristic of the performance of the Warden's detector. In steganalysis, it is customary to use the minimal total error probability under equal priors,

$$P_{\mathrm{E}} = \min_{P_{\mathrm{FA}}} \frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}}(P_{\mathrm{FA}})), \tag{13}$$

which we also adopt as the payoff function. In (13), P_{FA} and P_{MD} are the probabilities of false alarms and missed detection.

To evaluate the payoff function, we first need to compute the distribution of the Warden's statistic (12) under both hypotheses. Using (3) and (4), after a straightforward simplification the logarithm of the LRT (12) can be put into the following form:

$$\ln T = \sum_{i=1}^{n} L_i, \quad L_i = \ln\left(1 - \beta^{(\mathrm{W})} + \beta^{(\mathrm{W})} \exp\left(-1/(2\sigma^2)\right) \cosh\left(x/\sigma^2\right)\right). \tag{14}$$

Since the distribution under the null hypothesis is a special case of embedding with zero change rates, $\beta_i^{(\mathrm{A})} = 0$, we only need to work out the distribution under the alternative hypothesis.

[‡]In (12), we defined the following vector quantities: $\boldsymbol{\beta}^{(\mathrm{W})} = (\beta_1^{(\mathrm{W})}, \ldots, \beta_n^{(\mathrm{W})})$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_n^2)$.

The distribution of each L_i can be obtained by first computing its cumulative distribution function (c.d.f.):

$$F_{L_i}(y) \triangleq \Pr\{\phi(X_i) \leq y\}, \tag{15}$$

where $\phi(x) : \mathbb{R} \to \mathbb{R}$,

$$\phi(x) = \ln\left(1 - \beta^{(\mathrm{W})} + \beta^{(\mathrm{W})} \exp\left(-1/(2\sigma^2)\right) \cosh\left(x/\sigma^2\right)\right). \tag{16}$$

The condition $\phi(x) \leq y$ is equivalent with

$$\cosh\left(x/\sigma^2\right) \leq \left(1 + (e^y - 1)/\beta^{(\mathrm{W})}\right) \exp\left(1/(2\sigma^2)\right), \tag{17}$$

which implies

$$x \in [x_-, x_+],\ x_\pm = \pm\sigma^2 \cosh^{-1}(A + Be^y), \tag{18}$$

where

$$\cosh^{-1}(t) = \ln(t + \sqrt{t^2 - 1}), \tag{19}$$

$$A = (1 - 1/\beta^{(\mathrm{W})}) \exp\left(1/(2\sigma^2)\right), \quad B = (1/\beta^{(\mathrm{W})}) \exp\left(1/(2\sigma^2)\right). \tag{20}$$

Thus, the c.d.f. is

$$F_{L_i}(y) = \int_{x_-}^{x_+} f_{\beta^{(\mathrm{A})}}(x, \sigma_i^2)\mathrm{d}x, \tag{21}$$

and the p.d.f. $h_L(y)$ is obtained by differentiating (21) w.r.t y:

$$h_L(y) = f_{\beta^{(\mathrm{A})}}(x_+, \sigma^2)x'_+(y) - f_{\beta^{(\mathrm{A})}}(x_-, \sigma^2)x'_-(y) \tag{22}$$

$$= \frac{2B\sigma^2 e^y f_{\beta^{(\mathrm{A})}}\left(\sigma^2 \ln\left(A + Be^y + \sqrt{(A + Be^y)^2 - 1}\right), \sigma^2\right)}{\sqrt{(A + Be^y)^2 - 1}}, \tag{23}$$

using the fact that $x'_\pm(y) = \pm\sigma^2 Be^y / \sqrt{(A + Be^y)^2 - 1}$ and the fact that $f_\beta(x; \sigma^2)$ is even for all σ^2.

The distribution of the Warden's statistic is thus a convolution

$$\ln T \sim h_{L_1}(y) \star \cdots \star h_{L_n}(y). \tag{24}$$

Due to the form of the densities $h_{L_i}(y)$, no closed-form expression exists for (24) under either hypothesis and both densities must be sampled numerically. Once the densities are sampled with sufficient accuracy, the payoff function (13) can be evaluated by a numerical integration.

3.3 Strategies

As mentioned in the introduction, Alice's and Warden's strategies are the following sets of $n - 1$ real values, $\beta_i \in [0, 1/2]$,

$$S_{\mathrm{A}} = \{\beta_1^{(\mathrm{A})}, \ldots, \beta_{n-1}^{(\mathrm{A})}\}, \tag{25}$$

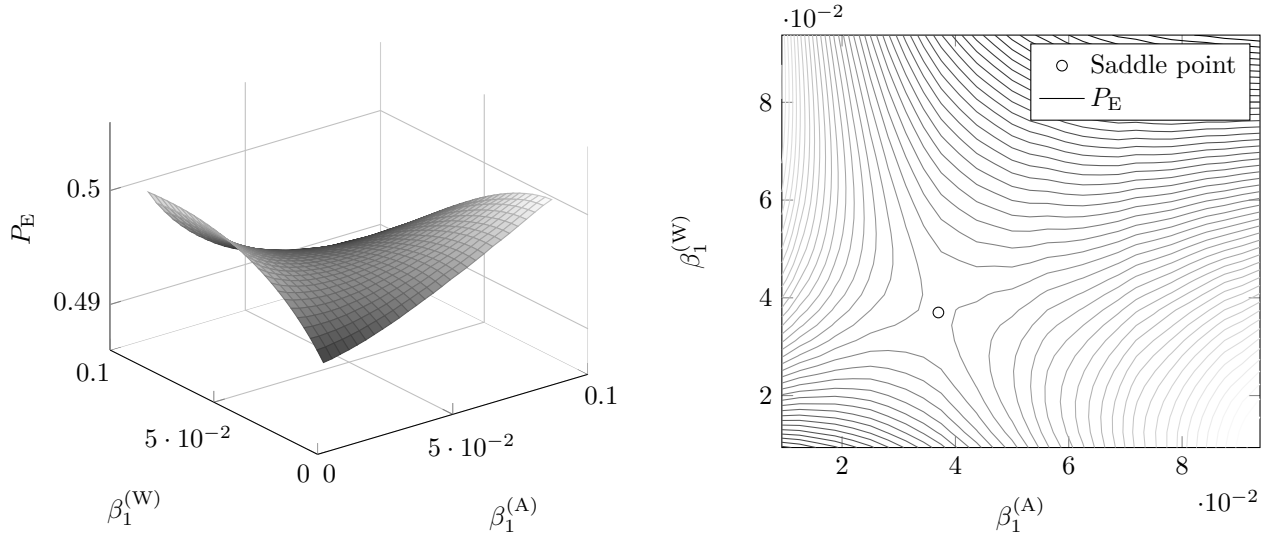
$$S_{\mathrm{W}} = \{\beta_1^{(\mathrm{W})}, \ldots, \beta_{n-1}^{(\mathrm{W})}\}, \tag{26}$$

because $\beta_n^{(\mathrm{A})}$ and $\beta_n^{(\mathrm{W})}$ are determined from their corresponding payload constraints (2) and (11). The constraints will further narrow down the range of possible values of β_i for both players (see Section 4).

To summarize, our game is formulated in mixed and continuous-valued strategies with both players playing simultaneously.

4. SOLVING THE GAME

Realizing how the Gaussian mixture (4) and the test statistic (14) depend on the strategies $\beta_i^{(\mathrm{A})}, \beta_i^{(\mathrm{W})}$, $i = 1 \ldots, n-1$, the payoff function (13) is a smooth function of the strategies. A game with continuous strategies and a smooth payoff function admits solution in pure strategies (see Chapter 4, Theorem 30 and 31 in Ref. [**?**]), which coincides with the saddle point, the Nash equilibrium. The solution can be determined numerically using a gradient search in which the payoff function is minimized over the Warden's strategies and maximized over Alice's strategies. The complexity of the search for the saddle point increases polynomially (but rather quickly due to the need to sample the test statistic distributions) with the number of pixels, n. Moreover, one needs to proceed with extra care due to accumulating numerical errors introduced by sampling the test statistic densities (24) and the payoff function (13), and evaluating the partial derivatives during the gradient search for the saddle. Section 6 contains some of the essential details regarding the various numerical approximations in our implementation.



(a) Surface graph (b) Contour graph

Figure 1: Payoff function $P_{\mathrm{E}}(\beta_1^{(\mathrm{A})}, \beta_1^{(\mathrm{W})})$ (13) for $\alpha = 0.2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1.2$.

5. EXPERIMENTS WITH TWO-PIXEL COVERS

Due to the difficulties with controlling the numerical error, we limited our experiments to covers consisting of only two pixels, $n = 2$, with variances σ_1^2, $\sigma_2^2 \lesssim 2$. Despite the simplicity, the results already provide an interesting insight.

For a two-pixel cover, the one-dimensional strategies must lie in a range determined by the payload, $\beta_1^{(\mathrm{A})}$, $\beta_1^{(\mathrm{W})} \in [\beta_{\min}, \beta_{\max}]$, where

$$\beta_{\max} = \min\left(0.5, h^{-1}\left(2\alpha\right)\right), \tag{27}$$

$$\beta_{\min} = h^{-1}\left(2\alpha - h(\beta_{\max})\right), \tag{28}$$

where $h^{-1}(x)$ is the inverse binary entropy on $[0, 1/2]$.[§] The payload constraint determines the remaining change rates:

$$\beta_2^{(\mathrm{A})} = h^{-1}\left(2\alpha - h(\beta_1^{(\mathrm{A})})\right), \tag{29}$$

$$\beta_2^{(\mathrm{W})} = h^{-1}\left(2\alpha - h(\beta_1^{(\mathrm{W})})\right). \tag{30}$$

[§]Recall that we work with a natural logarithm.

Figure 1 shows how the payoff function, $P_{\mathrm{E}}(\beta_1^{(\mathrm{A})}, \beta_1^{(\mathrm{W})}; \alpha, \sigma_1^2, \sigma_2^2)$, depends on Alice's and Warden's strategies, $\beta_1^{(\mathrm{A})}$ and $\beta_1^{(\mathrm{W})}$. It confirms our arguments presented in Section 4 that the payoff function is smooth in its arguments and that it also exhibits a saddle point – the Nash equilibrium.

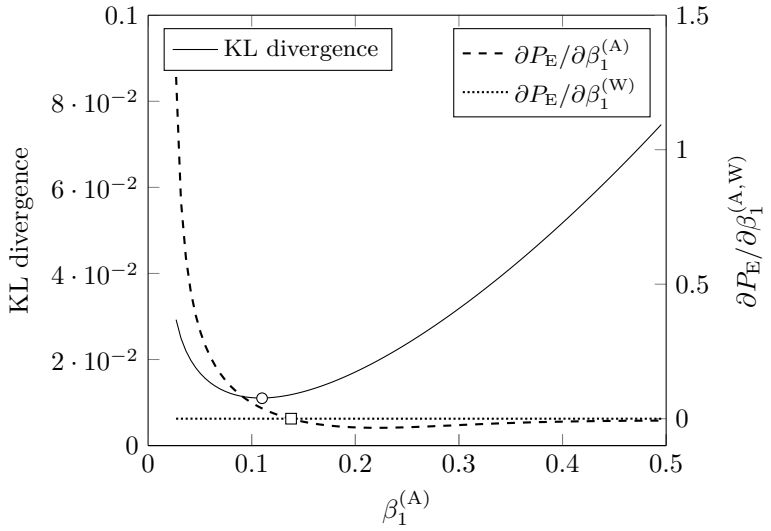


Figure 2: Left y axis: KL divergence between the distributions of cover and stego images $D_{\mathrm{KL}}(\mathbf{X}||\mathbf{Y})$ for an omnipotent Warden ($\beta_1^{(\mathrm{A})} = \beta_1^{(\mathrm{A},1)} = \beta_1^{(\mathrm{W})}$). The circle indicates the minimum of the KL divergence. Right y axis: Partial derivatives of the payoff function w.r.t. $\beta_1^{(\mathrm{A})}$ and $\beta_1^{(\mathrm{W})}$ for $\beta_1^{(\mathrm{A})} = \beta_1^{(\mathrm{W})}$ (see the text for more details). The square indicates the location of the Nash equilibrium.

The purpose of the second experiment is to show the difference between Alice's strategy that minimizes the KL divergence under an omnipotent Warden (the classical approach in steganography) and the strategy when Alice embeds at the Nash equilibrium (for an uninformed Warden). We denote Alice's strategies corresponding to both scenarios as $\beta_1^{(\mathrm{A},1)}$ and $\beta_1^{(\mathrm{A},2)}$. Note that an omnipotent Warden always chooses the same strategy for detection as Alice uses for embedding, $\beta_1^{(\mathrm{W})} = \beta_1^{(\mathrm{A},1)}$. Thus, one can plot the KL divergence between cover and stego images, $D_{\mathrm{KL}}(\mathbf{X}||\mathbf{Y})$, as a function of $\beta_1^{(\mathrm{A})}$ and determine $\beta_1^{(\mathrm{A},1)}$ as the strategy that minimizes the KL divergence. In Figure 2, the minimum is marked with a circle.

To find the saddle point, we first note that our numerical experiments indicate that the saddle point always satisfies $\beta_1^{(\mathrm{A},2)} = \beta_2^{(\mathrm{A},2)}$ (in other words, the saddle always seems to be on the axis of the first quadrant in the space of strategies). Thus, we plot in the same figure (shown on y-axis on the right) the partial derivatives $\partial P_{\mathrm{E}}(x, \beta_1^{(\mathrm{A})})/\partial x$ at $x = \beta_1^{(\mathrm{A})}$ and $\partial P_{\mathrm{E}}(\beta_1^{(\mathrm{A})}, y)/\partial y$ at $y = \beta_1^{(\mathrm{A})}$ as a function of $\beta_1^{(\mathrm{A})}$. The intersection of this curve with the x-axis marks Alice's strategy at the saddle, $\beta_1^{(\mathrm{A},2)}$ (shown as a square in Figure 2. The figure clearly shows the difference in Alice's strategies for the two scenarios. In other words, for an ignorant Warden it pays off for Alice to embed with change rates that lead to a slightly higher KL divergence under omnipotent Warden as she benefits from the mismatched detector of the Warden. Both players converge to a set of embedding and detection change rates, $\beta_1^{(\mathrm{A},2)}, \beta_1^{(\mathrm{W},2)}$ that correspond to a Nash equilibrium.

Next, we assess how the difference in strategies depends on the relative payload. To this end, in Figure 3 we plot $\beta_1^{(\mathrm{A},1)}$ and $\beta_1^{(\mathrm{A},2)}$ as a function of α. Note that the difference between the optimal strategies under both scenarios (omnipotent and ignorant Warden) monotonically increases with payload. For an ignorant Warden, it is always more convenient for Alice to put a slightly larger payload into the less noisy pixel. This way Alice lowers the payoff function despite increasing the KL divergence between cover and stego sources.

The aim of the next experiment is to study the difference in Alice's embedding strategies as a function of the content diversity (the pixel variances). The intention is to reveal how the difference in both strategies is affected

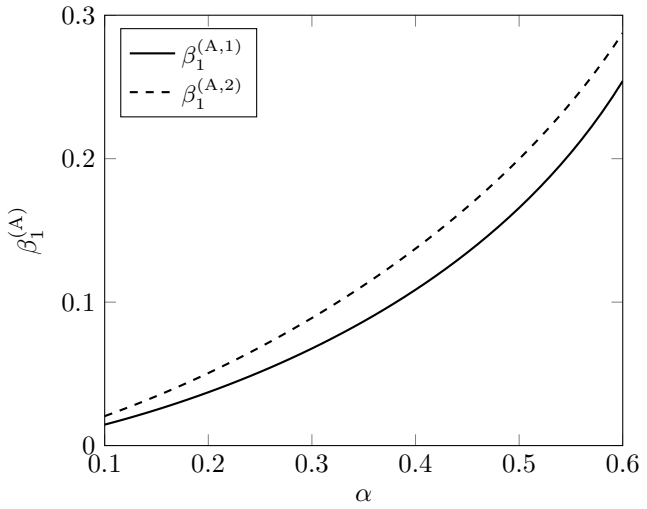


Figure 3: Alice's strategies under both scenarios $\beta_1^{(\mathrm{A},1)}$, $\beta_1^{(\mathrm{A},2)}$ as a function of relative payload α for $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.2$.

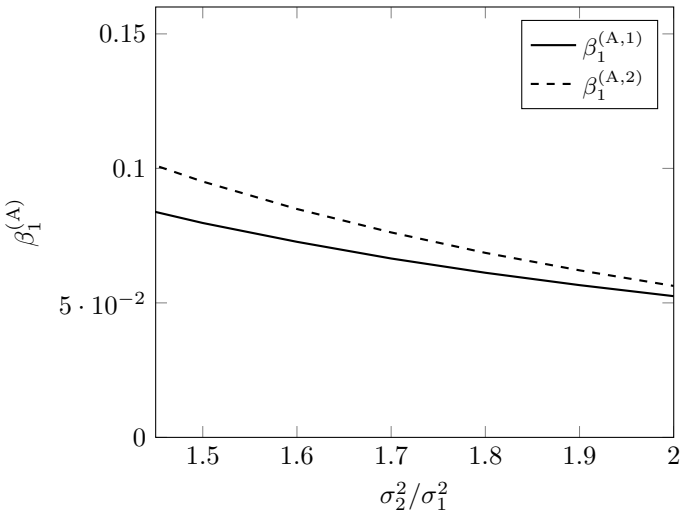


Figure 4: Alice's strategies under both scenarios $\beta_1^{(\mathrm{A},1)}$, $\beta_1^{(\mathrm{A},2)}$ as a function of the content diversity measured by the ratio σ_2^2/σ_1^2 for $\alpha = 0.4$ and $\sigma_1^2 = 1$.

by diverse content. Figure 4 shows that with increasing content diversity the difference between both strategies becomes smaller. This is intuitive as detecting embedding in a very noisy pixel will be increasingly more difficult.

Finally, we wish to assess the impact of different scenarios on the statistical detectability. Figure 5 shows the loss of Warden's ability to detect embedding due to her ignorance of Alice's actions. We express this loss in terms of an information-theoretic measure per pixel to obtain a quantity that can be scaled to larger covers and to provide some meaning for practitioners. To this end, we compute the KL divergence between the distributions of the Warden's statistic (14) under both hypotheses for each scenario and then take their difference:

$$\begin{aligned}\triangle D_{\mathrm{KL}}(\ln T|\mathrm{H}_0||\ln T|\mathrm{H}_1) &= D_{\mathrm{KL}}\left(\ln T^{(2)}|\mathrm{H}_0||\ln T^{(2)}|\mathrm{H}_1\right) - D_{\mathrm{KL}}\left(\ln T^{(1)}|\mathrm{H}_0||\ln T^{(1)}|\mathrm{H}_1\right)\\ &\triangleq D_{\mathrm{KL}}^{(2)} - D_{\mathrm{KL}}^{(1)}. \end{aligned} \tag{31}$$

This difference informs us about the change in the error exponent that controls the missed detection proba-

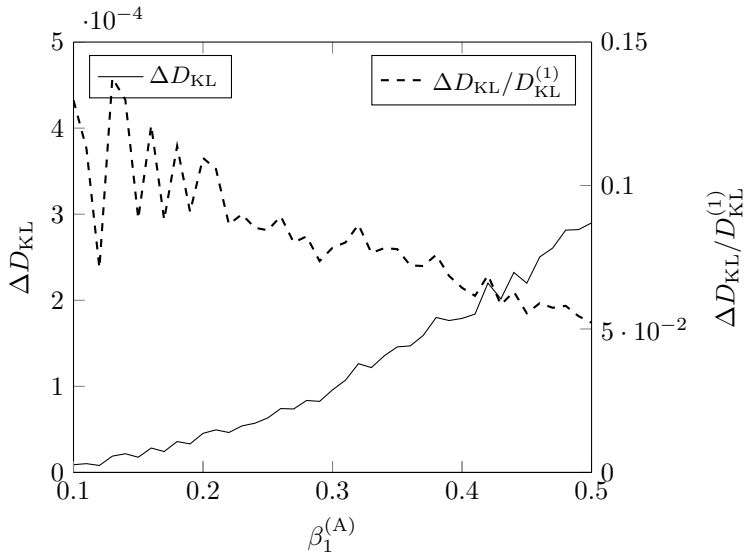


Figure 5: Left y axis: Warden's loss in her ability to detect Alice's embedding, $\triangle D_{\mathrm{KL}}(\ln T|\mathrm{H}_0||\ln T|\mathrm{H}_1)$, as a function of α for $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.2$. Right y axis: The relative change, $\triangle D_{\mathrm{KL}}/D_{\mathrm{KL}}^{(1)}$. The variations seen in the graphs are due to numerical errors. See the text for more details.

bility in Nayman–Pearson hypothesis testing. In (31), $\ln T^{(1)}|\mathrm{H}_0$ stands for the distribution of the test statistic $\ln T$ when Alice embeds with $\beta_1^{(\mathrm{A},1)}$ and the Warden detects with $\beta_1^{(\mathrm{W},1)} = \beta_1^{(\mathrm{A},1)}$ (the first, classical scenario) while $\ln T^{(2)}|\mathrm{H}_0$ stands for the scenario when both players embed at the Nash equilibrium $\beta_1^{(\mathrm{A},2)}, \beta_1^{(\mathrm{W},2)}$.

6. IMPLEMENTATION DETAILS

For reproducibility, in this section we include implementation details used in the numerical evaluation of formulas from Sections 3.2 and 5 as well as the techniques used to control the numerical error.

When inverting the binary entropy, minimizing the KL divergence for the omnipotent Warden option strategy, or finding the P_{E} for the payoff function in the ignorant Warden option, the optimization is done numerically. All optimized functions have a single global optimum that can be determined within the machine precision.

Numerical integration was carried out in matlab using the adaptive Gauss-Kronrod quadrature, which conveniently outputs an approximate upper bound on the integration error. This error can also be controlled but making it very small (e.g., of the order of machine precision) is very computationally expensive.

Most of the integrations need to be carried out on an infinite interval. Even though the Gauss-Kronrod quadrature can handle unbounded integration intervals, it can be unstable. Fortunately, since the integrands fall to zero exponentially quickly, we integrate on a compact interval instead and introduce a new error, which can be made arbitrarily small.

The supports of all distributions of the test statistics are intervals (T, ∞), where T is a parameter depending on $\beta_1^{(\mathrm{W})}$ and the choice of variances of pixels. We integrate them on the interval $(T + \epsilon, \infty)$, where ϵ can be made sufficiently small to achieve an arbitrarily small approximation error.

The distribution of the test statistic for more pixels is a convolution of the test statistic densities for single pixels. When integrating this convolution, we take note of the bounds on the error with which the convolution has been evaluated to obtain an error bound for the integral of the convolution (when computing the P_{E}).

When finding the saddle point, which must satisfy

$$\left(\partial P_{\mathrm{E}}/\partial\beta_1^{(\mathrm{A})}\right)^2 + \left(\partial P_{\mathrm{E}}/\partial\beta_1^{(\mathrm{W})}\right)^2 = 0, \tag{32}$$

we noticed in our experiments that the function $\partial P_{\mathrm{E}}/\partial\beta_1^{(\mathrm{W})}$ tends to be zero on or near the diagonal line $\beta_1^{(\mathrm{W})} = \beta_1^{(\mathrm{A})}$. To save on the computational time, we restrict ourselves only to this diagonal and find $\hat{\beta}_1^{(\mathrm{A})}$ that solves $\partial P_{\mathrm{E}}(\hat{\beta}_1^{(\mathrm{A})}, \hat{\beta}_1^{(\mathrm{A})})/\partial\beta_1^{(\mathrm{A})} = 0$ instead of (32). We then inspect the values of $\partial P_{\mathrm{E}}/\partial\beta_1^{(\mathrm{W})}$ on a line perpendicular to the diagonal going through the point $(\hat{\beta}_1^{(\mathrm{A})}, \hat{\beta}_1^{(\mathrm{A})})$ to determine the size of the segment where we are certain that the condition (32) is satisfied within a prescribed accuracy (5×10^{-3} was used in our work). We take the size of this segment as the approximate upper bound on the error with which we determine the saddle point location. The approximation $\partial P_{\mathrm{E}}/\partial\beta_1^{(\mathrm{W})} = 0$ holds reasonably accurately near the line $\beta_1^{(\mathrm{W})} = \beta_1^{(\mathrm{A})}$ only for $\beta_1^{(\mathrm{W})} \lesssim 0.3$, depending on α and the pixel variances. Fortunately, in all cases we inspected in this work this approximation introduced a small enough error for locating the saddle point.

7. CONCLUSION

The vast majority of publications on steganography is cast within the information theoretic framework for a computationally unbounded Warden who has a complete access to the steganographic channel, which comprises of the steganographic method, the cover source, message source, and stego key source. This choice is usually justified by evoking the Kerckhoffs' principle, which is the golden standard in information security. In steganography in practice, however, the Warden rarely has a full access to the steganographic channel, and the Kerckhoffs' principle seems overly pessimistic, which may lead to conclusions that are too conservative. For example, it has been argued by Böhme that the cover source is fundamentally incognizable and thus is in fact unavailable to either party. Likewise, Alice is free to incorporate randomness or any side information when embedding her message. It thus seems inevitable to consider more realistic scenarios in which the Warden is ignorant about certain components of the steganographic channel.

In this paper, we lift the assumption that the Warden knows the probabilistic selection channel used by Alice. In fact, we make the channel a strategy in a game played by Alice and the Warden. The Warden uses the channel to construct a likelihood ratio detector of steganography (here, we give the Warden the knowledge of the payload and the cover distribution). The cover source is modeled as a sequence of independent Gaussian variables with unequal variances. The payoff function driving the game is a scalar measure of the performance of Warden's detector – the total error probability under equal priors. For a mutually independent embedding operation (LSB matching), we show numerically that the game admits a solution in pure strategies for a cover consisting of two elements. This limitation has been imposed by the quickly growing complexity of numerical approximations to the underlying distributions of Warden's test statistic. However, even this simple case already reveals some interesting phenomena.

First, in terms of Warden's total detection error, it is advantageous for Alice to trade off the optimality of her embedding strategy w.r.t. KL divergence between cover and stego distributions for a mismatched detector at the Warden's end. The Nash equilibrium was numerically shown to be different from the strategy that minimizes the KL divergence. The difference in both strategies decreases with increased differences between the variances of both cover elements, which is to be expected as in the limit of an Gaussian element with infinite variance, the entire payload should be embedded in the more noisy element. Surprisingly, it is always advantageous for Alice to embed a slightly larger payload into the element with a smaller variance rather than vice versa. Finally, we quantified the impact of embedding at the Nash equilibrium as opposed to embedding with minimal KL divergence by evaluating the change in the KL divergence between Warden's statistics per cover element (the error exponent).

The value of this work lies primarily in shedding more light on the problem of optimal steganography under an ignorant Warden. In particular, we confirm the conclusion already reached in [**?**] that the KL divergence is no longer an appropriate measure of security and Alice's optimal embedding strategy should be determined from a framework based on the Game Theory.

8. ACKNOWLEDGMENTS

The work on this paper was supported by Air Force Office of Scientific Research under the research grant number FA9950-12-1-0124. The U.S. Government is authorized to reproduce and distribute reprints for Governmental

purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of AFOSR or the U.S. Government.

REFERENCES

1. Filler, T., Judas, J., and Fridrich, J., "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security* **6**, 920–935 (September 2011).
2. Pevný, T., Filler, T., and Bas, P., "Using high-dimensional image models to perform highly undetectable steganography," in [*Information Hiding, 12th International Conference*], Böhme, R. and Safavi-Naini, R., eds., Lecture Notes in Computer Science **6387**, 161–177, Springer-Verlag, New York, Calgary, Canada (June 28–30, 2010).
3. Huang, F., Huang, J., and Shi, Y.-Q., "New channel selection rule for JPEG steganography," *IEEE Transactions on Information Forensics and Security* **7**, 1181–1191 (August 2012).
4. F. Huang, W. Luo, J. H. and Shi, Y.-Q., "Distortion function designing for JPEG steganography with uncompressed side-image," in [*1st ACM IH&MMSec. Workshop*], Puech, W., Chaumont, M., Dittmann, J., and Campisi, P., eds. (June 17–19, 2013).
5. Holub, V. and Fridrich, J., "Digital image steganography using universal distortion," in [*1st ACM IH&MMSec. Workshop*], Puech, W., Chaumont, M., Dittmann, J., and Campisi, P., eds. (June 17–19, 2013).
6. Kim, Y., Duric, Z., and Richards, D., "Modified matrix encoding technique for minimal distortion steganography," in [*Information Hiding, 8th International Workshop*], Camenisch, J. L., Collberg, C. S., Johnson, N. F., and Sallee, P., eds., Lecture Notes in Computer Science **4437**, 314–327, Springer-Verlag, New York, Alexandria, VA (July 10–12, 2006).
7. Sachnev, V., Kim, H. J., and Zhang, R., "Less detectable JPEG steganography method based on heuristic optimization and BCH syndrome coding," in [*Proceedings of the 11th ACM Multimedia & Security Workshop*], Dittmann, J., Craver, S., and Fridrich, J., eds., 131–140 (September 7–8, 2009).
8. Wang, C. and Ni, J., "An efficient JPEG steganographic scheme based on the block–entropy of DCT coefficents," in [*Proc. of IEEE ICASSP*], (March 25–30, 2012).
9. Fridrich, J., Goljan, M., and Soukal, D., "Perturbed quantization steganography using wet paper codes," in [*Proceedings of the 6th ACM Multimedia & Security Workshop*], Dittmann, J. and Fridrich, J., eds., 4–15 (September 20–21, 2004).
10. Guo, L., Ni, J., and Shi, Y.-Q., "An efficient JPEG steganographic scheme using uniform embedding," in [*Fourth IEEE International Workshop on Information Forensics and Security*], (December 2–5, 2012).
11. Filler, T. and Fridrich, J., "Design of adaptive steganographic schemes for digital images," in [*Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics III*], Alattar, A., Memon, N. D., Delp, E. J., and Dittmann, J., eds., **7880**, OF 1–14 (January 23–26, 2011).
12. Schöttle, P., Korff, S., and Böhme, R., "Weighted stego-image steganalysis for naive content-adaptive embedding," in [*Fourth IEEE International Workshop on Information Forensics and Security*], (December 2–5, 2012).
13. Bas, P., Filler, T., and Pevný, T., "Break our steganographic system – the ins and outs of organizing BOSS," in [*Information Hiding, 13th International Conference*], Filler, T., Pevný, T., Ker, A., and Craver, S., eds., Lecture Notes in Computer Science **6958**, 59–70 (May 18–20, 2011).
14. Ettinger, J. M., "Steganalysis and game equilibria," in [*Information Hiding, 2nd International Workshop*], Aucsmith, D., ed., Lecture Notes in Computer Science **1525**, 319–328, Springer-Verlag, New York, Portland, OR (April 14–17, 1998).
15. Böhme, R., [*Advanced Statistical Steganalysis*], Springer-Verlag, Berlin Heidelberg (2010).
16. Filler, T., Ker, A. D., and Fridrich, J., "The Square Root Law of steganographic capacity for Markov covers," in [*Proceedings SPIE, Electronic Imaging, Media Forensics and Security*], Memon, N. D., Delp, E. J., Wong, P. W., and Dittmann, J., eds., **7254**, 08 1–08 11 (January 18–21, 2009).

17. Cogranne, R., Zitzmann, C., Fillatre, L., Retraint, F., Nikiforov, I., and Cornu, P., "A cover image model for reliable steganalysis," in [*Information Hiding, 13th International Conference*], Filler, T., Pevný, T., Ker, A., and Craver, S., eds., Lecture Notes in Computer Science, 178–192 (May 18–20, 2011).
18. Janesick, J. R., [*Scientific Charge-Coupled Devices*], vol. Monograph PM83, Washington, DC: SPIE Press - The International Society for Optical Engineering (January 2001).