Rethinking Optimal Embedding

Andrew D. Ker
Dept. of Computer Science
University of Oxford
Oxford OX1 3QD, UK
adk@cs.ox.ac.uk

Tomáš Pevný
CTU in Prague
Cisco R&D Center in Prague
Prague, Czech Republic
pevnak@gmail.com

Patrick Bas
Univ. Lille, CNRS, Centrale Lille, UMR 9189
CRIStAL Lab, Lille, France,
Patrick.Bas@ec-lille.fr

ABSTRACT

At present, almost all leading steganographic techniques for still images use a distortion minimization paradigm, where each potential change is assigned a cost c_i and the change probabilities π_i chosen to minimize the average total cost $\sum_i \pi_i c_i$. However, some detectors have exploited knowledge of this adaptivity and the embedding cannot be considered optimal. In this work we prove a theoretical result suggesting that, against a knowing attacker, the embedder should simply minimize $\sum_i \pi_i^2 c_i$ instead, for the same costs c_i, which is the minimax and equilibrium strategy. This aligns with some special case results that have appeared in recent literature. We then test some simple steganographic methods in theoretical and real settings, showing that naive (average cost) adaptivity is exploitable, but the equilibrium probabilities cannot be exploited. However, it is essential to determine statistically well-founded costs c_i.

Keywords

Steganography, game theory, distortion, optimal embedding

1. INTRODUCTION

Even the first steganographers knew that not all embedding changes are equal: some are more detectable than others. Early steganographic literature [27, 15] tried various approaches to what we now call *adaptive embedding*, but it was with the discovery of Syndrome Trellis Codes [6] that adaptive steganography became practical. In the theory of additive *optimal embedding*, each change is independent and has some distortion cost c_i[1] that can be computed from the cover, and aims to make that change with probability π_i to minimize the average total cost

$$\sum_i \pi_i c_i. \tag{1}$$

[1]A common notation for a cost is ρ_i, but we prefer to reserve Greek letters for the strategies of the embedder and detector.

IH&MMSec 2016, June 20-23, 2016, Vigo, Spain

DOI: http://dx.doi.org/10.1145/2909827.2930797

When embedding is not binary, the sum becomes $\sum_{i,j} \pi_i^j c_i^j$, where c_i^j is the cost of applying change option j to location i, and π_i^j the corresponding probability. Note that the hypothesis of additive distortion can be extended to local interactions between neighbouring distortions, and computed using Gibbs sampling over disjoint sublattices [5].

At present, the leading steganographic embedding methods for digital images employ this method: HUGO [18], WOW [9], the UNIWARD family [10], HILL [17]. But in the last two years some detectors have been published that exploit the adaptivity: by placing more weight on parts of the image where it is estimated more likely to contain payload, the performance of the detector can be improved [25, 4, 3]. The first version of UNIWARD [10] was exploited to the extent that it would have been better to use an older embedding method instead, and the cost function had to be altered [3]. A somewhat similar bug was present in the first version of HUGO, exploited in the BOSS contest [8], but this can better be explained by the cost function omitting important information rather than the detector exploiting its values.

If circumstances exist under which additive adaptive embedding can be exploited by a *knowing* detector (one who has knowledge of the distortion costs) then the adaptivity cannot be optimal. In part the suboptimality is because detectability is a property of cover and stego *distributions*, and cannot be determined from minimizing distortion in a single images; in part because the knowing detector can exploit the adaptivity. Hence we call embedding that optimizes (1) *naive adaptivity*[2]. We will argue that the embedder should minimize an alternative total cost,

$$\sum_i \pi_i^2 c_i \tag{2}$$

for the same c_i as before. When embedding is not binary, this becomes $\sum_{i,j,k} \pi_i^j c_i^{j,k} \pi_i^k$, where $c_i^{j,k}$ is a matrix defined by cost interactions at location i. We call this *equilibrium adaptivity*, because it is the equilibrium strategy of a zero sum game between the embedder and detector (in this work also called the *attacker*), in a suitably simple theoretical setting. And since it is a minimax strategy, the embedder has optimized against the worst case: the knowing attacker.

Our result certainly has limitations: it assumes independence of pixels or pixel groups, and that each embedding changes only one element. The former is common in the

[2]Note that in some other work [21], *naive adaptivity* was used for an even weaker embedding that always picks the location with lowest cost.

theory of steganography and has not prevented theoretical results from predicting practical performance [14], but the latter is an important consideration for further work on practical equilibrium adaptivity, as changing one pixel typically affects multiple measurements [4].

The structure of the paper is as follows. In Section 2 we give the theoretical justification for equilibrium adaptivity in models of covers where the pixels are independent. We make comparisons with existing literature using embedding costs in Section 3 and note that our result generalizes some special cases that have appeared recently, where squared probabilities can also be found; we also discuss how (2) can be optimized in practice. In Section 4 we test the theory against artificially-generated binary covers matching the binary model, and in Section 5 we examine the original version of S-UNIWARD, whose adaptivity was exploitable, and show that equilibrium mitigates this effect. However, since UNIWARD costs are not statistically founded, we do not optimize the embedding in this case. We draw conclusions in Section 6.

2. THEORETICAL RESULTS

In the game theory of steganography, the detector wishes to optimize some performance metric of their hypothesis test for

H_0: object under consideration is a cover, vs.

H_1: object under consideration is stego.

In the results below, we will be able to make a large-sample approximation, so that the detection is based on a statistic ℓ with Gaussian distribution under either hypothesis. We will define

$$\mu_0 = \mathrm{E}_{H_0}[\ell], \qquad \mu_1 = \mathrm{E}_{H_1}[\ell],$$
$$\sigma_0^2 = \mathrm{Var}_{H_0}[\ell], \qquad \sigma_1^2 = \mathrm{Var}_{H_1}[\ell].$$

In this work we choose the detector's payoff as the *true negative rate when the false negative rate is 50%*. The embedder's payoff is the inverse, the false positive rate when the true positive rate is 50%, which is a benchmark we have advocated in [19] for high-accuracy steganalysis. The game is zero sum (i.e. the gain on the detector's side is exactly balanced by the losses on the embedder's side).

Without loss of generality we may assume $\mu_0 < \mu_1$. For 50% true positives, the detection threshold for ℓ is the median value of ℓ in hypothesis H_1, which is μ_1 (and does not depend on σ_1). The detector's payoff is $\mathrm{P}_{H_0}[\ell < \mu_1] = \Phi(\delta)$, where Φ is the Gaussian cumulative density function and δ is the deflection

$$\delta = \frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2}}. \tag{3}$$

Since Φ is increasing, we can assume that the detector aims to maximize, and the embedder to minimize, δ.

This payoff function is a Neyman-Pearson criterion: the detector optimizes one type of error against a fixed limit on the other. In case the reader would prefer a different choice, for example the equal-prior error rate which is often used in steganographic game theory [20], we mention that this provides an asymptotically equivalent ranking since in the large sample limit the square root law [14] would force $\pi_i \to 0$, hence $\sigma_0^2 \sim \sigma_1^2$. Then equal-prior error has a monotone relationship with any Neyman-Pearson criterion. The deflection has appeared in other work (in which the cover models are restricted) on optimality against a knowing attacker [23, 22], and can be found as the payoff in steganographic games even as far back as [11]. It has also been justified by empirical evidence from steganographic likelihood ratio tests, in [2].

2.1 Binary Embedding in Binary Covers

To illustrate the calculations we begin with the simplest possible cover, consisting of n independent binary pixels. If they had equal probability distribution there would be no adaptivity, so we define p_i as the probability that pixel i takes value 1. We assume that these are known to the detector. Embedding must flip pixels, and we define the embedder's strategy as $(\pi_1, \dots, \pi_n)$ where π_i is the probability of flipping pixel i; this is the adaptivity that we may (or may not) grant to the detector. Therefore in a stego object the probability that pixel i takes value 1 is

$$(1-\pi_i)p_i + \pi_i(1-p_i) = p_i + \pi_i(1-2p_i).$$

We know the form of an optimal detector for these hypotheses; the Neyman-Pearson Lemma states that we should reject the null hypothesis (give a positive detection) if the log-likelihood ratio

$$\log \prod_i \frac{\big(p_i + \pi_i(1-2p_i)\big)^{x_i}\big(1 - p_i - \pi_i(1-2p_i)\big)^{1-x_i}}{p_i^{x_i}(1-p_i)^{1-x_i}}$$
$$= c + \sum_i x_i \Big[\log\Big(1 + \pi_i\big(\tfrac{1-2p_i}{p_i}\big)\Big) - \log\Big(1 - \pi_i\big(\tfrac{1-2p_i}{1-p_i}\big)\Big)\Big],$$

where x_i denotes the observed value of pixel i, exceeds a threshold.

We do not need to analyze the likelihood ratio; it is sufficient to note that it is a constant plus[3]

$$\ell(x_i, \omega_i) = \sum_i x_i \omega_i$$

where (ω_i) defines the detector's strategy (the *weight* they give to each observation). Because likelihood ratio tests are an optimal subclass of all hypothesis tests under a Neyman-Pearson criterion (such as the false positive rate at 50% true positives), it is sufficient to consider detectors of this type, seeking equilibrium between (π_i) and (ω_i).

Asymptotically for large n and constant payload size, the Central Limit Theorem says that the distribution of $\ell(x_i, \omega_i)$ is Gaussian in either null or alternative hypothesis, with $\mu_0 = \sum p_i\omega_i$, $\mu_1 = \sum\big(p_i + \pi_i(1-2p_i)\big)\omega_i$, and $\sigma_0^2 = \sum p_i(1-p_i)\omega_i^2$. So by (3), the detector's payoff is monotone increasing in the deflection

$$\frac{\sum_i \pi_i(1-2p_i)\omega_i}{\sqrt{\sum_i p_i(1-p_i)\omega_i^2}} = \delta(\pi_i, \omega_i, d_i, e_i)$$

where $d_i = (1-2p_i)$, $e_i = p_i(1-p_i)$, and

$$\delta(\pi_i, \omega_i, d_i, e_i) = \frac{\sum_i \pi_i d_i \omega_i}{\sqrt{\sum_i e_i \omega_i^2}}. \tag{4}$$

It may easily be verified that

$$\frac{\partial \delta}{\partial \omega_j} \propto (\textstyle\sum_i e_i\omega_i^2)\pi_j d_j - (\sum_i \pi_i d_i \omega_i) e_j \omega_j. \tag{5}$$

[3]We use $\ell(x_i, \omega_i)$ as shorthand to mean a function of all x_i's and ω_i's.

The detector wants to maximize the value of δ in (4). An *ignorant* detector[4] must proceed as if all π_i are equal to a constant π. Therefore (5) yields in this case

$$\omega_j \propto \pi d_j e_j^{-1}.$$

Note that multiplying the weights ω_i by a constant does not change the detector. The stationary point can be verified to be a maximum as long as all p_i are not equal to zero or one, but we omit the routine calculation. Substituting into (4) gives

$$\delta(\pi_i, d_i, e_i) = \frac{\sum_i \pi\pi_i d_i^2 e_i^{-1}}{\sqrt{\sum_i \pi^2 d_i^2 e_i^{-1}}},$$

the π terms cancel (hence it does not matter whether the ignorant detector is granted knowledge of the payload size, which would reveal π), and the denominator is constant, leaving

$$\delta \propto \sum_i \pi_i c_i, \tag{6}$$

where:

$$c_i = d_i^2 e_i^{-1} = \frac{1-2p_i}{p_i(1-p_i)}, \tag{7}$$

turns out to be the true statistical cost of flipping pixel i.

The best counter-strategy for the embedder minimizes this value of δ, subject to a payload constraint $\sum_i H(\pi_i) = m$, where H is the binary entropy function. This is the familiar "optimal embedding" scenario, where the total costs are linear in the embedding probabilities; the standard solution can be found using the method of Lagrange multipliers, which gives $\lambda c_j = H'(\pi_j)$ and hence the well-known solution

$$\pi_i = \frac{e^{-\lambda c_i}}{1+e^{-\lambda c_i}} \tag{8}$$

for some constant λ determined by the payload constraint.

A *knowing* detector wants to maximize the value of δ, given complete knowledge of π_i. From (5), δ is maximized when

$$\omega_j \propto \pi_j d_j e_j^{-1} \tag{9}$$

in which case the numerator of (4) is the square of the denominator, so that

$$\delta^2 \propto \sum_i \pi_i^2 c_i \tag{10}$$

where again $c_i = d_i^2 e_i^{-1}$. Note that the payoff depends on the squared embedding probabilities, but the costs are identical to the standard optimal embedding scenario. This time the solution satisfies $\lambda c_j = H'(\pi_j)/\pi_j$, which does not have a closed form. See Subsection 3.2 for a discussion of how (π_i) should be found; for now we simply draw attention to the difference between optimization for an ignorant attacker (6) and a knowing attacker (10).

Since the two player game is zero sum, the minimax solution (π_i, ω_i) is an equilibrium [26].

2.2 q-ary Embedding in Arbitrary Covers

Now consider a model where n pixels take values in some finite alphabet $\{a_1, \ldots, a_l\}$. We still assume that they are independent of each other, and a fixed independent embedding operation, but this can be arbitrary, and the pixels can take arbitrary and different distributions.

We define $p_i^k = \mathrm{P}[X_i = a_k]$ as the distribution of cover pixel i, and gather it into a vector $\boldsymbol{p_i}$.[5] When a pixel is used for embedding, it changes from value a_j to a_k with probability q_{kj} gathered into a matrix $\boldsymbol{Q}$ (for convenience, our matrix is organized in columns). Such matrices can describe LSB Replacement or Matching, Ternary or Pentary Embedding, etc. Note that $\boldsymbol{p_i}$ and $\boldsymbol{Q}$ are parameters of the game, known to the detector.

As before, the embedder's strategy is the probability that each pixel is used (π_i). The unconditional distribution of stego pixels is therefore given by $\boldsymbol{q_i} = (1-\pi_i)\boldsymbol{p_i} + \pi_i \boldsymbol{Q}\boldsymbol{p_i}$.

This time the log-likelihood ratio is of the form

$$c + \sum_i \sum_k [x_i = k] \log(1 + \pi_i \tfrac{q_i^k - p_i^k}{p_i^k})$$

where $[A]$ is the Iverson bracket taking value 1 when A is true. For the same reasons as before it is sufficient to consider detectors, parameterized by the detector's strategy ω_i^k $(i = 1, \ldots, n,\ k = 1, \ldots, l)$, of the form

$$\ell(x_i, \omega_i^k) = \sum_i \sum_k [x_i = k]\omega_i^k. \tag{11}$$

We collect the detector's strategy into n vectors $\boldsymbol{\omega_i}$. Routine calculation gives

$$\mu_1 - \mu_0 = \sum_i \pi_i \boldsymbol{d_i}^T \boldsymbol{\omega_i},$$

where $\boldsymbol{d_i} = (\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{p_i}$, and

$$\sigma_0^2 = \sum_i \boldsymbol{p_i}^T \boldsymbol{E_i} \boldsymbol{p_i},$$

where $\boldsymbol{E_i} = \Delta_{\boldsymbol{p_i}} - \boldsymbol{p_i}\boldsymbol{p_i}^T$, $\Delta_{\boldsymbol{p_i}}$ representing a diagonal matrix. Thus the detector's payoff is monotone increasing in

$$\delta(\pi_i, \boldsymbol{\omega_i}, \boldsymbol{d_i}, \boldsymbol{E_i}) = \frac{\sum_i \pi_i \boldsymbol{d_i}^T \boldsymbol{\omega_i}}{\sqrt{\sum_i \boldsymbol{\omega_i}^T \boldsymbol{E_i} \boldsymbol{\omega_i}}}. \tag{12}$$

Employing some vector calculus,

$$\frac{\partial \delta}{\partial \boldsymbol{\omega_j}} \propto (\textstyle\sum_i \boldsymbol{\omega_i}^T \boldsymbol{E_i} \boldsymbol{\omega_i})\pi_j \boldsymbol{d_j} - (\textstyle\sum_i \pi_i \boldsymbol{d_i}^T \boldsymbol{\omega_i})\boldsymbol{E_j}\boldsymbol{\omega_j}. \tag{13}$$

For an *ignorant* detector, δ is maximized when $\boldsymbol{E_j}\boldsymbol{\omega_j} \propto \boldsymbol{d_j}$. Then (12) simplifies to

$$\delta \propto \sum_i \pi_i c_i, \tag{14}$$

where $c_i = \boldsymbol{d_i}^T \boldsymbol{E_i}^{-1} \boldsymbol{d_i}$.

For a *knowing* detector, (13) implies that they should choose $(\boldsymbol{\omega_i})$ so that

$$\boldsymbol{E_j}\boldsymbol{\omega_j} \propto \pi_j \boldsymbol{d_j}$$

which, similarly to before, leads to

$$\delta^2 \propto \sum_i \pi_i^2 c_i \tag{15}$$

for the same c_i. Again, compare (14) with (15).

[4]Recall that the detector is ignorant of the individual values of π_i, but still granted knowledge of p_i, hence d_i and e_i.

[5]We will use boldface lowercase letters for vectors, and boldface uppercase for matrices.

We will not solve these optimization problems, but proceed to a more general problem in the following subsection.

(There is a little wrinkle that has been obscured by our use of vector notation. The matrices $\boldsymbol{E_i}$, which are covariance matrices of a multinomial distribution, are all deficient. This does not affect the correctness of the calculations, as long as all p_i^k are positive, since then $\text{rank}(\boldsymbol{E_i}) = l - 1$ and $\sum_k d_i^k = \sum_k p_i^k - q_i^k = 0$ also has one degree of redundancy. Thus $\boldsymbol{E_j \omega_j} \propto \pi_j \boldsymbol{d_j}$ does define ω_j uniquely up to a constant multiple, and $c_i = \boldsymbol{d_i}^T \boldsymbol{E_i}^{-1} \boldsymbol{d_i}$ is well-defined.)

2.3 Arbitrary Embedding

Consider now a fully arbitrary embedder, who can apply different embedding operations to each pixel. The cover model is still defined by $\boldsymbol{p_i}$, but it makes sense to consider the embedder's strategy to be exactly their output at pixel i, the vectors $\boldsymbol{\pi_i} = (\pi_i^1, \ldots, \pi_i^l)$; we do not need to know what the embedder does with the cover pixel, merely what they output.

All pixels are potentially used (this does not stop the embedder from outputting cover pixels, of course). The calculations are therefore the same as the previous subsection with $\pi_i = 1$, but now

$$\boldsymbol{d_i} = \boldsymbol{\pi_i} - \boldsymbol{p_i}$$

as what was previously called $\boldsymbol{q_i}$ becomes the embedder's move in the game, rather than a parameter known to the detector.

The detector's strategy is of the same form as before, $(\boldsymbol{\omega_i})$, but there is no optimal detector who is ignorant of $\boldsymbol{\pi_i}$. They cannot perform a likelihood ratio test without knowing the ratio, and we postpone to future work investigation of a generalized likelihood ratio test in this model. However, against *any fixed* detector the $\boldsymbol{\omega_i}$ can be considered constants; therefore so is the denominator of the deflection (12), which in this case reduces to

$$\delta \propto \sum_i (\boldsymbol{\pi_i} - \boldsymbol{p_i})^T \boldsymbol{\omega_i}.$$

The payload constraint is the familiar

$$\sum_i H(\boldsymbol{\pi_i}) = m$$

where $H(\pi_i)$ represents the entropy of the vector $\boldsymbol{\pi_i}$, with the additional constraint that $\sum_k \pi_i^k = 1$. Using vector calculus[6] and Lagrange multipliers (omitting routine calculation of this standard result) we reach

$$\lambda \boldsymbol{\omega_i} = \ln(\mu_i \boldsymbol{\pi_i})$$

where $\ln(\mu_i \boldsymbol{\pi_i})$ applies the logarithm pointwise to the vector. This has the well-known q-ary optimal embedding solution

$$\pi_i^k = \frac{e^{-\lambda \omega_i^k}}{\sum_k e^{-\lambda \omega_i^k}}. \tag{16}$$

On the other hand, for a *knowing* detector, we still have

$$\boldsymbol{E_j \omega_j} \propto \boldsymbol{\pi_j} - \boldsymbol{p_j},$$

and so

$$\delta^2 \propto \sum_i (\boldsymbol{\pi_i} - \boldsymbol{p_i})^T \boldsymbol{E_i}^{-1} (\boldsymbol{\pi_i} - \boldsymbol{p_i}). \tag{17}$$

[6]Note that $\frac{\partial}{\partial \boldsymbol{\pi_i}} H(\boldsymbol{\pi_i}) \propto \mathbf{1} + \ln \boldsymbol{\pi_i}$.

In this situation the embedder should minimize a quadratic form, the multidimensional analogue of the squared probabilities in (10). The solution satisfies

$$\lambda \boldsymbol{E_i}^{-1} \boldsymbol{\pi_i} = \ln(\mu_i \boldsymbol{\pi_i}) \tag{18}$$

for a constant λ determined by the payload constraint, and μ_i to ensure that $\sum_k \pi_i^k = 1$. For a brief discussion of how this is solved, see Subsection 3.2.

3. RELATION TO OTHER WORK

There have been a few contributions on embedding against a knowing detector, and some recent work has produced results that have superficial similarities with ours. In this section we make the connections more explicit.

3.1 Related Work on Optimality

Most distortion costs in the literature are derived directly from the cover content. They are built on the rationale that the distortion should be low for samples that are difficult to predict (e.g. pixels located in textures of an image) and high for samples that can be easily predicted (e.g. edges or homogeneous areas). For example the schemes UNIWARD [10] and WOW [9] use high-pass wavelet subbands, while HILL [17] uses a succession of high-pass and low-pass filters. Theses schemes approximate the local noise power in the image, and the cost is akin to a (stego-)signal to (image self-)noise ratio. For all these schemes the distortion minimized is of the form $\sum_i \pi_i c_i$.

The cost derived in HUGO [18] is a "hybrid" in the sense that its distortion uses weights directly computed from the image content (the successive differences between neighbouring pixels), but the algorithm also attempts to preserve a global distortion model between the entire cover and stego image. The model is derived from co-occurrence matrices of image residuals. Again, the distortion minimized is of the form $\sum_i \pi_i c_i$.

Note that these distortion measures are blind to the knowledge of the steganalyser, and more particularly to the fact that a detector might use an estimation of the embedding probabilities $\hat{\pi}_i$ to derive detectors that will be more sensitive to embedding changes. Recently, knowing detectors (sometimes called *omniscient* in the literature) have appeared with tSRM [25] or the maxSRM [4] feature sets, which explicitly use $\hat{\pi}_i$ and can offer up to 10% improvement on the classification accuracy for very small payloads [4]. Note also that these feature sets are derived by weighting each occurrence by its probability of change, exactly like the optimal detection strategy proposed in equation (9).

The following steganographic schemes are interesting because they implicitly or explicitly assume that the steganalyser may have knowledge or estimates of π_i, and result in distortions depending on π_i^2.

The first distortion measure comes from [12] which is set in the context of batch steganography, and measures statistical distortion using Kullback-Leibler Divergence (KLD). The total distortion of the batch is then a weighted sum of the squares of the number of embedding changes in each image; the latter is proportional to the probability of embedding in any given location in that image. For example, in [12, §2.2] total distortion is proportional to $\sum_i c_i \pi_i^2$, where c_i is called the "Q-factor". Even though this theoretical analysis did not generate a practical embedding method, it appears to be the first that proposed a distortion where the cost is

weighted by π_i^2. Measuring detectability by KLD implicitly models the worst-case opponent, who would indeed be a knowing detector.

Ref. [7] is another that measures detectability by KLD. It is approximated by $\sum \frac{1}{2} I_i(0) \pi_i^2$ where $I_i(0)$ denotes the Fisher information of the stego content with respect to π_i. The authors use a local Gaussian approximation of the image pixels in a neighbourhood to derive $I_i(0)$ (which is independent of π_i) and to perform the minimization. It is interesting to note that the "Q-factor" mentioned in [12] is in fact also half of Fisher's information.

Instead of minimizing KLD, two recent works have optimized embedding with respect to a likelihood ratio detector.

In [23] the cover is modelled as a discretized univariate generalized Gaussian model with varying variance (similar to the model of [7]), and the embedding is pentary. The probabilities of ± 1 and ± 2 changes constitute the embedder's strategy. Similarly to our analysis, the authors argue that the embedder should minimize a deflection coefficient, and by approximating the likelihood ratio explicitly they derive an optimization [23, Eq. (11)] which is a quadratic form like (18).

In [22] the cover is modelled as a discretized Gaussian with varying variance, and the embedding is ternary. The square deflection of the likelihood ratio is given by $\sum \pi_i^2/\sigma_i^4$ where σ_i^2 denotes the variance of the underlying local Gaussian model at location i [22, Eq. (11)]. They also compute the deflection against an ignorant (there called indifferent) detector, which is proportional to $\sum \pi_i/\sigma_i^4$ [22, Eq. (12)].

This confluence of results is encouraging. Our contribution is to show that the conclusions are not dependent on a particular cover or embedding model, and apply to arbitrary discrete covers and arbitrary embedding operations.

3.2 Related Work on Implementation

We have not yet explained how the optimization problems (10) and (17) can be solved. In fact, some of the other literature mentioned above derives equations of a similar form, and the generalizations are simple enough that we need not go into much detail about them.

For linear additive distortions $\sum \pi_i c_i$ it is well-known that we can optimize without even finding the solution (8), via Syndrome Trellis Codes (STCs). For arbitrary embedding against a fixed detector, $\sum \boldsymbol{\pi_i \omega_i}$ can be optimized using l-ary STCs, though their complexity grows very rapidly with the alphabet size l, or nested STCs [6].

In [22] the authors explain how a function that fits the form $\sum \pi_i^2 c_i$ can be optimized: numerically solve the relationship $\lambda c_j = H'(\pi_j)/\pi_j$, inside an interval bisection search for λ meeting the payload constraint. Then reverse-engineer costs $d_j = H'(\pi_j)/\lambda$ would have given the same solution for naive embedding and λ that would leave these probabilities on the rate-distortion bound. Applying STCs to the adjusted costs (d_i) should make changes with approximately the optimal probabilities (π_i). However the probabilities will not be exact, because STCs do not meet the rate-distortion bound exactly.

We should mention that, although $\lambda c_j = H'(\pi_j)/\pi_j$ does not have a closed-form inverse, it would be trivial to tabulate a few million values for locating π_j from λc_j, and this only needs to be done once. Then, if necessary, the inverse could be refined with one or two steps of the Newton-Raphson method. In our experiments the calculation of embedding probabilities was negligible compared with calculating the costs in the first place.

In [23] the authors arrive at a system of $2n+1$ variables that is similar to the nk simultaneous equations we reached in (18); their system is simpler because the embedding operation is limited to ± 1 and ± 2 changes. We advocate the same approach that they take: it is a convex system, so Newtonian numerical methods should converge quickly to a solution. Again, these can be reverse-engineered into costs for a l-ary or nested STC.

Nonetheless, in this work we will not try to implement equilibrium embedding, since it is sufficient for our purposes to simulate it by applying changes with the optimal probabilities.

4. ARTIFICIAL BINARY COVERS

Our initial experiments are on artificial "images" fitting exactly the model in Subsection 2.1, which ensures that most of the assumptions of the theory are true (in particular, independence of pixels). These experiments test whether the use of the large sample approximations is valid. Our covers will contain $n = 2^{18}$ pixels, the size of images in the BOSSBase library [1].

In order to generate covers according to the model, we must select the pixel probabilities $p_i \in (0,1)$. In the first experiment we drew p_i independently from a Beta distribution with parameters $(5,5)$; this means that they are somewhat bell shaped and symmetrical around $p_i = 0.5$. We generated 10 000 cover images using this model.

We then simulated three types of embedding on the rate-distortion curve: non-adaptive embedding where π_i is constant, naive adaptive embedding where $\sum_i \pi_i c_i$ is minimized, and equilibrium adaptive embedding where $\sum_i \pi_i^2 c_i$ is minimized. In each case the payload constraint was $\sum H(\pi_i) = 0.1 \cdot 2^{18}$, simulating a payload of 0.1 bits per pixel, a size chosen so that the detectors will be neither practically-perfect nor near-random. We generated 10 000 stego images for each case.

In Figure 1 (top) we show the histogram of the values p_i, and the relationship between the cost c_i (computed from p_i using the results in section 2.1) and the naive or equilibrium embedding probabilities π_i. Throughout the figure, red denotes the naive adaptivity and blue the equilibrium. Observe that equilibrium embedding assigns lower probabilities to low costs, but slightly higher probabilities to high costs: it is less aggressive in its adaptivity (which is the reason that it cannot be exploited). We also show histograms of the binary entropy of the naive and equilibrium probabilities: $H(\pi_i)$ indicates how many bits of payload (under perfect coding) are placed in location i. Observe that naive adaptive embedding prefers to hide almost one bit in each of a few locations (in this case over 58% of the payload is placed into just under 6% of the locations) and zero or almost zero in the rest. Note that in the Bernoulli model, the costs dip sharply for p_i close to 0.5, since $c_i = (1-2p_i)^2/p_i(1-p_i)$ (see Eq.(7)), so these locations appear attractive for payload. However, equilibrium embedding places at least a tiny amount of payload in every location, and very few contain as much as nearly one bit.

Because we know exactly the cover model and the embedding probabilities, we can build an optimal detector directly from the likelihood ratio test (LRT), rejecting the null hy-

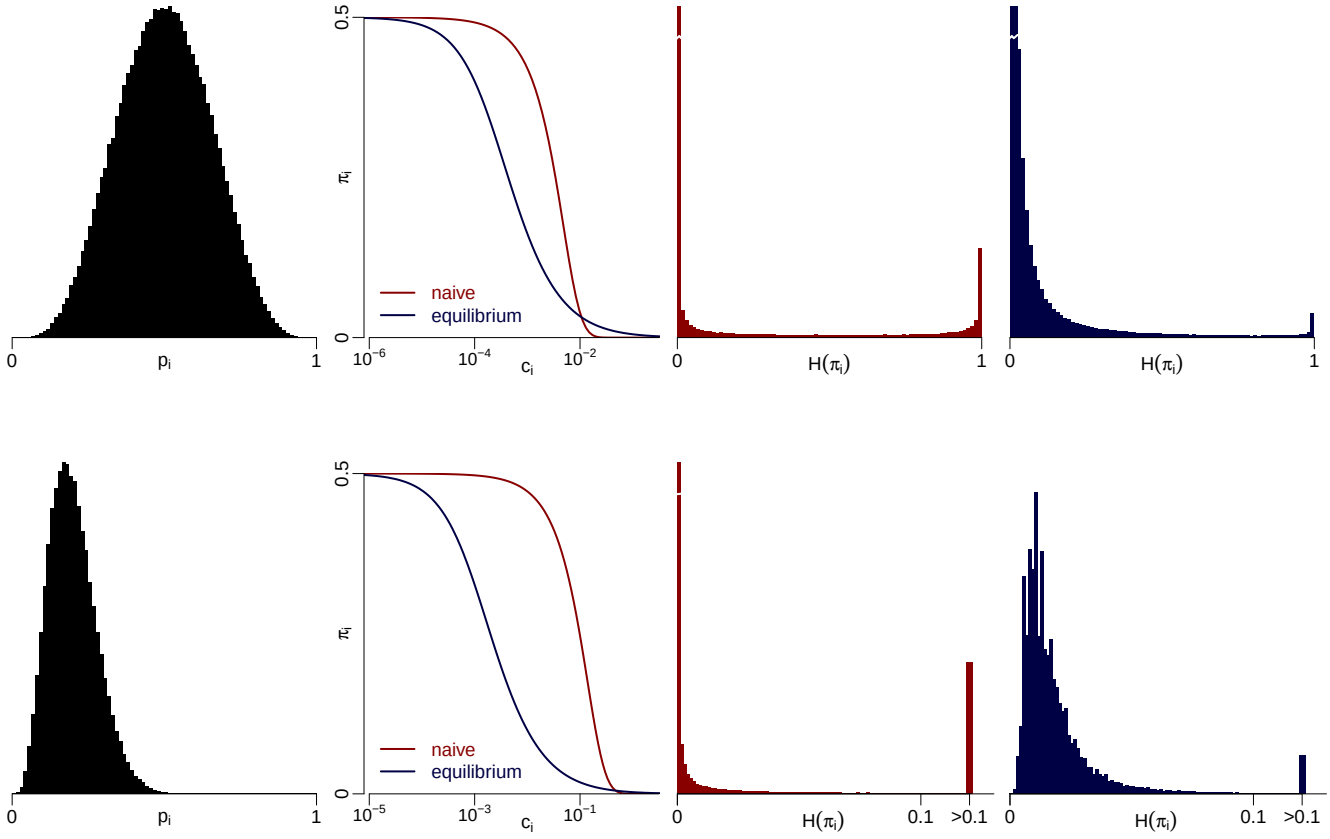

Figure 1: Embedding in artificial binary covers. From left to right: histogram of the probabilities p_i that pixel i takes value 1; relationship between embedding cost and probability when the adaptivity is naive (red) and in equilibrium (blue); histogram of the binary entropy of the embedding probabilities for naive adaptivity (note that the zero bin is truncated); for equilibrium adaptivity, displayed on the same scale. Above, p_i is drawn from the $\mathrm{Beta}(5,5)$ distribution, with a payload of 0.1 bits per pixel in 2^{18} pixels. Below, p_i is drawn from the $\mathrm{Beta}(5,20)$ distribution, with a payload of 0.02 bits per pixel in 2^{18} pixels, and the histogram of $H(\pi_i)$ focuses on low entropy locations.

pothesis if

$$\sum_i x_i \Big[\log\Big(1+\pi_i\big(\tfrac{1-2p_i}{p_i}\big)\Big) - \log\Big(1-\pi_i\big(\tfrac{1-2p_i}{1-p_i}\big)\Big)\Big]$$

exceeds a threshold. This can be done with each (π_i), to create the detector's optimal counter-strategy to each of the embedder's three strategies. We then computed the empirical *embedder's* payoff, the false positive rate when the true positive rate is 50%, for each detector against each embedding adaptivity. The results are shown in Table 1 (top), where each row corresponds to the optimal detector for each type of embedding, and each column to that embedding.

At this payload rate, non-adaptive embedding is highly detectable by the conventional LRT (0.000 false positives), but if the embedder switches to standard naive adaptivity this detector is almost random. However, a knowing attacker can regain good performance by using the adaptivity in the LRT (0.023 false positives). To avoid being exploited, the embedder should use equilibrium adaptivity, and the optimal detector still makes false positives at a rate of 0.145. Amongst the three embedder/detector strategies shown here, the table shows that equilibrium embedding and the LRT detector for it do indeed form an equilibrium, since the detector can only degrade their performance by deviating from it, and the embedder can only become more detectable by deviating.

The bell-shaped distribution of pixel probabilities p_i leads to many costs very close to zero, which is probably not a good model for steganography in real images. Therefore we repeated the previous experiments with a different distribution of p_i, this time from the Beta distribution with parameters $(20,5)$, in which the costs do not get too close to zero. The higher costs indicate that payload will be more detectable, so we reduced the payload constraint to 0.02 bits per pixel, again so that the detectors' performance is neither too good nor too poor. The distribution of p_i, the relationship between costs and embedding probabilities under naive and equilibrium adaptivity, and the distribution of payload are shown in Figure 1 (bottom) and the performance of the detectors in Table 1 (bottom). We observe similar features: naive adaptive embedding is more secure against an ignorant opponent, but can be exploited by a knowing opponent to the extent that it becomes even more detectable than no adaptivity at all. Equilibrium embedding spreads the payload rather more evenly than naive adaptivity, and cannot be exploited.

(a) p_i taken from $Beta(5,5)$, 0.1 bits per pixel embedded in 2^{18} pixels

LRT detector for	Embedding adaptivity: None	Naive	Equilibrium
No adaptivity	0.000	0.492	0.335
Naive adaptivity	0.443	0.023	0.225
Equilibrium	0.038	0.081	0.145

(b) p_i taken from $Beta(5,20)$, 0.02 bits per pixel embedded in 2^{18} pixels

LRT detector for	Embedding adaptivity: None	Naive	Equilibrium
No adaptivity	0.040	0.433	0.198
Naive adaptivity	0.472	0.000	0.305
Equilibrium	0.099	0.023	0.116

Table 1: The embedder's payoff (false positive rate when true positive rate is 50%) for artificial binary covers. Likelihood ratio tests for each adaptivity type were tested against embedding with each adaptivity type.

5. EXPLOITABILITY OF S-UNIWARD

Finally, we attempt to apply the theory of equilibrium to the contemporary steganographic algorithm S-UNIWARD [10]. This calculates the distortion of changing pixel i from a Wavelet-transformed image:

$$c_i = \sum_{k=1}^{3}\sum_{u=1}^{m}\sum_{v=1}^{n} \frac{W_{uv}^{(k)}(\boldsymbol{X}) - W_{uv}^{(k)}(\boldsymbol{X}^{\boldsymbol{i}})}{\sigma + W_{uv}^{(k)}(\boldsymbol{X})},$$

where $\boldsymbol{X}$ is the matrix of $m \times n$ cover pixels, $\boldsymbol{X}^{\boldsymbol{i}}$ the same image with only change i applied and $W_{uv}^{(k)}(\cdot)$ denote the (u,v) wavelet coefficients in the first-level undecimated Daubechies 8-tap wavelet decomposition. The index $k = 1, 2, 3$ corresponds to the LH, HL, and HH subbands, respectively.

The original version of UNIWARD used $\sigma = 10^{-15}$ to avoid division by zero, but this caused wide variation in the costs, and hence a strong preference for certain embedding locations. In [3] the authors demonstrated the vulnerability of S-UNIWARD to an attacker who could estimate the costs via a feature set called Content-Selective Residuals (CSR); the same work proposed the simple heuristic fix of setting $\sigma = 1$ to moderate the costs. The effect on embedding probabilities is similar to that of an equilibrium strategy: it increases the likelihood of embedding in previously unused pixels, and reduces it in the more commonly-used areas. Since then, further (less catastrophic) attacks on S-UNIWARD have been accomplished by weighting features according to the probability of embedding, which our theory predicts to be optimal (9), in feature sets tSRM [25] and maxSRM [4].

In these experiments, we investigate whether an equilibrium strategy provides an alternative to changing the value of σ. For just one image, the histogram of the costs with $\sigma = 10^{-15}$, the relationship between costs and embedding probabilities, and histograms of the binary entropy of those probabilities are displayed in Figure 2. We also display binary entropy of the naive embedding probabilities for some other values of σ; in this case it appears that equilibrium embedding probabilities are quite close to those that could be obtained naively using $\sigma = 20$.

(a) CSR features; payload constraint of 0.2 bits per pixel

Detector trained on	$\sigma = 10^{-15}$: Embedding ad. Naive	Equil.	$\sigma = 1$: Embedding ad. Naive	Equil.
Naive ad.	0.0076	0.4999	0.4452	0.4975
Equil. ad.	0.5053	0.2949	0.4025	0.3425

(b) CSR features; payload constraint of 0.3 bits per pixel

Detector trained on	$\sigma = 10^{-15}$: Embedding ad. Naive	Equil.	$\sigma = 1$: Embedding ad. Naive	Equil.
Naive ad.	0.0073	0.5000	0.3997	0.4950
Equil. ad.	0.5022	0.1297	0.3593	0.3013

(c) maxSRM features; payload constraint of 0.3 bits per pixel

Detector trained on	$\sigma = 10^{-15}$: Embedding ad. Naive	Equil.	$\sigma = 1$: Embedding ad. Naive	Equil.
Naive ad.	0.2406	0.3229	0.2773	0.3109
Equil. ad.	0.4719	0.2442	0.3463	0.1580

Table 2: Equal-prior error rates for S-UNIWARD, both the original version with $\sigma = 10^{-15}$ and the updated version with $\sigma = 1$, with naive and equilibrium adaptivity.

For both the original S-UNIWARD with $\sigma = 10^{-15}$, and the modern version with $\sigma = 1$, we computed the naive and equilibrium optimal embedding probabilities; we also computed CSR and maxSRM features[7].

Our experiments used BOSSBase [1]: 10 000 grayscale images of size 512×512; although a single cover corpus would not be appropriate for significant experiments [24], it is sufficient to explore our theory in practice.

In each repetition of the experiment, two binary classifiers were trained: one for stego images embedded with naive adaptivity, and the other for stego images embedded with equilibrium adaptivity. Both steganalyzers were implemented as an ensemble of FLDs and trained on 8000 cover and 8000 stego images. Since we are now performing real steganalysis, we adopt the standard benchmark $P_{\text{err}} = 0.5(P_{\text{fp}} + P_{\text{fn}})$, estimated on the remaining 2000 cover and stego images. The experiment was repeated ten times with different partitions of training and testing data.

[7] maxSRM features were calculated by estimating the embedding strategy from the stego image, as proposed in [4], rather than assuming omniscience of the detector.

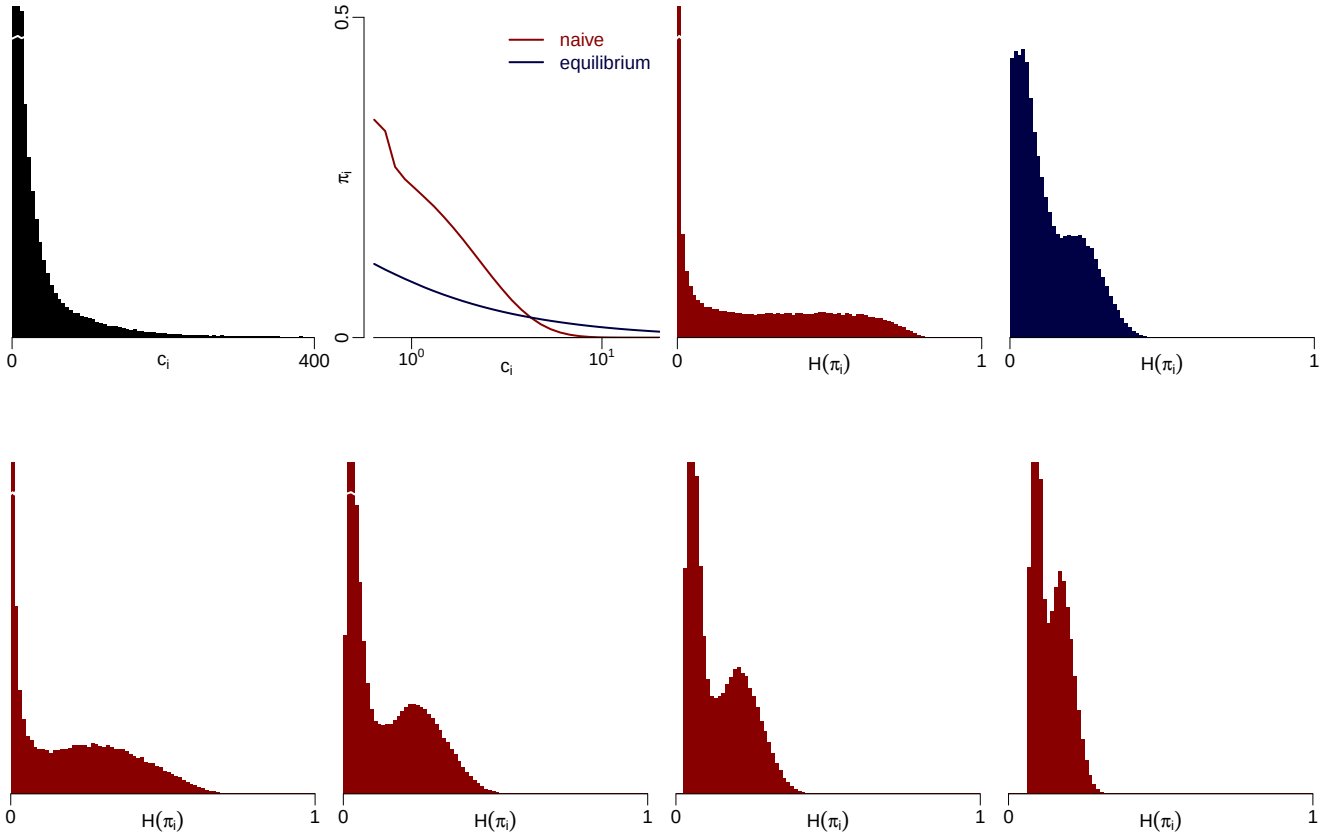

Figure 2: Above, equilibrium embedding with the original version of S-UNIWARD ($\sigma = 10^{-15}$), in one of the BOSSBase images. From left to right: histogram of the distortion costs c_i; relationship between embedding cost and probability when the adaptivity is naive (red) and in equilibrium (blue); histogram of the binary entropy of the embedding probabilities for naive adaptivity; for equilibrium adaptivity, displayed on the same scale. The image has 2^{18} pixels and the payload is 0.4 bits per pixel. Below, histograms of $H(\pi_i)$ on the same scale for naive embedding with (left to right) $\sigma = 1, 10, 20, 50$.

The error rates for the various combinations of σ, embedder, and detector, are shown in Table 2. We include two payloads tested against CSR features. For the original S-UNIWARD with $\sigma = 10^{-15}$, the equilibrium strategy is much less exploitable than naive adaptive embedding. However, the "equilibrium" is not really an equilibrium, since against such a detector it would be advantageous for the embedder to switch to the other embedding probabilities. Furthermore, naive embedding with $\sigma = 1$ is less detectable than equilibrium embedding with $\sigma = 10^{-15}$. These results confirm that equilibrium embedding makes the flaw in the original S-UNIWARD less exploitable, but we should not be surprised that the so-called equilibrium is suboptimal: the costs are heuristic rather than founded on statistical optimality, and the theory does not directly apply. When $\sigma = 1$ it is downright worse to use equilibrium embedding, but that is only to be expected when $\sigma = 1$ already moderates the embedding probabilities.

We also include one payload detected by maxSRM features. Here there is no significant benefit in switching to equilibrium embedding even at $\sigma = 10^{-15}$, and disadvantages at $\sigma = 1$. Again we should not be surprised, because the costs were not optimal to begin with.

6. CONCLUSIONS

An adaptivity criterion is an example of side information. Given asymptotically perfect coding, it does not matter whether the receiver possesses this information. But this is not at all true for the detector, and optimizing an embedder against a worst-case (likelihood ratio test) detector, which is the minimax strategy for the embedder, has received recent attention. Solving such game theory problems in steganography was identified as an important open problem in [13].

In this work we have substantially generalized prior results, to the case of arbitrary nonstationary (discrete) cover distributions and arbitrary embedding. Thus we can justify theoretically certain emerging themes: quadratic forms in the embedding probabilities [23], and weighting features by an estimated embedding probabilities [4]. For binary and fixed q-ary embedding, our result can be stated simply: optimize $\sum \pi_i^2 c_i$ instead of $\pi_i c_i$, but do not change the costs. We have seen empirical evidence that equilibrium strategies are less aggressive than naive adaptivity, making even high-cost changes with non-negligible probability and less certain to make low-cost changes. Heuristically, something similar was applied to counter exploitation of the original S-UNIWARD.

Ironically, compared with the discretized nonstationary univariate Gaussian [22] or generalized Gaussian [23] special cases, the more powerful theoretical results do not lead to easily-implementable optimal steganography. The advantage of restricted cover models is that the parameters are few enough that they can be estimated; that would be quite impossible for arbitrary nonstationary covers with arbitrary embedding. We consider our contribution more to the theory of optimal embedding than the practice; the works [23, 22] have already demonstrated that there is practical value in this approach.

Although the theoretical results are general, they have important limitations. They grant the detector considerable knowledge, including the exact distortion costs (even though they do not possess the cover) and cover model. They require the detection space to be identical to that in which distortion is calculated: this is not the case for current leading steganalysis and adaptive steganography, but the optimality of the likelihood ratio test suggests that these domains should eventually converge. It is for further work to adapt these results to the practical situation where one embedding change affects multiple features. Finally, we have assumed independence of pixels in the cover model, and independent embedding. It would be valuable to consider Gibbs embedding [5] against a knowing opponent, but for fully-general distortions we will not be able to appeal to the law of large numbers, making analysis difficult. We speculate that a similar result could at least be proved for costs that are sums of local potentials.

Finally, we should stress that swapping $\sum \pi_i c_i$ to $\sum \pi_i^2 c_i$ is not a panacea in steganography. It is only correct if the costs c_i were statistically justified, and does not necessarily apply to the vast majority of steganography cost functions such as WOW, UNIWARD, or HILL, which are purely heuristic. And they have to be at least somewhat heuristic, because to know the true costs requires impossible knowledge of the exact parameters of the cover model (the probability that each pixel takes a particular value). Our theoretical results exist in something of a vacuum because neither player will in practice know the costs. However, it remains possible that good estimates of the costs will suffice for a near-equilibrium (costs perhaps obtained form empirical learning about the cover source), and this is something for future work.

Development of less-heuristic distortion functions is an important area for future research in steganography, and even though MiPOD [22] is based on a very simple cover model it makes a valuable contribution in this direction. As proposed in [16], one possible solution to derive a practical distortion having statistical meaning would be to use the output of a classifier. Perhaps statistical methods can also lead to the development of distortion functions for domains where they are currently lacking, such as audio or video.

Acknowledgements

Patrick Bas thanks the *Centre National de la Recherche Scientifique* for specific support on this topic.

7. REFERENCES

[1] P. Bas, T. Pevný, and T. Filler. BOSSBase. `http://webdav.agents.fel.cvut.cz/data/projects/stegodata/BossBase-1.01-cover.tar.bz2`, May 2011.

[2] R. Cogranne and J. Fridrich. Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory. *IEEE Transactions on Information Forensics and Security*, 10(12):2627–2642, Dec. 2015.

[3] T. Denemark, J. Fridrich, and V. Holub. Further study on the security of S-UNIWARD. In *Proceedings of SPIE/IS&T International Symposium on Electronic Imaging*, volume 9028. International Society for Optics and Photonics, 2014.

[4] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich. Selection-channel-aware rich model for steganalysis of digital images. In *Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014.

[5] T. Filler and J. Fridrich. Gibbs construction in steganography. *IEEE Transactions on Information Forensics and Security*, 5(4):705–720, 2010.

[6] T. Filler, J. Judas, and J. Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3):920–935, 2011.

[7] J. Fridrich and J. Kodovský. Multivariate gaussian model for designing additive distortion for steganography. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2949–2953, May 2013.

[8] G. Gul and F. Kurugollu. A new methodology in steganalysis: breaking highly undetectable steganograpy (HUGO). In *Proceedings of International Conference on Information Hiding (IH)*, pages 71–84. Springer, 2011.

[9] V. Holub and J. Fridrich. Designing steganographic distortion using directional filters. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 234–239. IEEE, 2012.

[10] V. Holub, J. Fridrich, and T. Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014(1):1–13, 2014.

[11] A. D. Ker. Batch steganography and the threshold game. In *Proceedings of SPIE/IS&T International Symposium on Electronic Imaging*, volume 6505. International Society for Optics and Photonics, 2007.

[12] A. D. Ker. Steganographic strategies for a square distortion function. In *Proceedings of SPIE/IS&T International Symposium on Electronic Imaging*, volume 6819. International Society for Optics and Photonics, 2008.

[13] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevnỳ. Moving steganography and steganalysis from the laboratory into the real world. In *Proceedings of ACM Workshop on Information Hiding and Multimedia Security (IHMMSec)*, pages 45–58. ACM, 2013.

[14] A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich. The square root law of steganographic capacity. In *Proceedings of ACM Workshop on Multimedia and Security (MMSec)*, pages 107–116. ACM, 2008.

[15] Y. Kim, Z. Duric, and D. Richards. Modified matrix encoding technique for minimal distortion steganography. In *Proceedings of International Workshop on Information Hiding (IH)*, pages 314–327, 2007.

[16] S. Kouider, M. Chaumont, and W. Puech. Adaptive steganography by oracle (ASO). In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

[17] B. Li, M. Wang, J. Huang, and X. Li. A new cost function for spatial image steganography. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 4206–4210. IEEE, 2014.

[18] T. Pevný, T. Filler, and P. Bas. Using high-dimensional image models to perform highly undetectable steganography. In *Proceedings of the International Conference on Information Hiding (IH) 2010*, pages 161–177, 2010.

[19] T. Pevný and A. D. Ker. Towards dependable steganalysis. In *Proceedings of SPIE/IS&T International Symposium on Electronic Imaging*, volume 9409. International Society for Optics and Photonics, 2015.

[20] P. Schöttle and R. Böhme. Game theory and adaptive steganography. *IEEE Transactions on Information Forensics and Security*, 11(4):760–773, April 2016.

[21] P. Schöttle, S. Korff, and R. Böhme. Weighted stego-image steganalysis for naive content-adaptive embedding. In *Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 193–198. IEEE, 2012.

[22] V. Sedighi, R. Cogranne, and J. Fridrich. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 11(2):221–234, 2016.

[23] V. Sedighi, J. Fridrich, and R. Cogranne. Content-adaptive pentary steganography using the multivariate generalized gaussian cover model. In *Proceedings of SPIE/IS&T International Symposium on Electronic Imaging*, volume 9409. International Society for Optics and Photonics, 2015.

[24] V. Sedighi, J. Fridrich, and R. Cogranne. Toss that bossbase, Alice! In *Proceedings of IS&T International Symposium on Electronic Imaging*. Society for Imaging Science and Technology, 2016.

[25] W. Tang, H. Li, W. Luo, and J. Huang. Adaptive steganalysis against WOW embedding algorithm. In *Proceedings of ACM Workshop on Information hiding and Multimedia Security (MMSec)*, pages 91–96. ACM, 2014.

[26] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[27] A. Westfeld. F5—a steganographic algorithm. In *Proceedings of International Workshop on Information Hiding (IH)*, pages 289–302. Springer Berlin Heidelberg, 2001.