On Comparing Ad Hoc Detectors with Statistical Hypothesis Tests

Eli Dworetzky, Edgar Kaziakhmedov, and Jessica Fridrich
Binghamton University
Department of Electrical and Computer Engineering
Binghamton, NY 13850
{edworet1,ekaziak1,fridrich}@binghamton.edu

ABSTRACT

This paper addresses how to fairly compare ROCs of ad hoc (or data driven) detectors with tests derived from statistical models of digital media. We argue that the ways ROCs are typically drawn for each detector type correspond to different hypothesis testing problems with different optimality criteria, making the ROCs uncomparable. To understand the problem and why it occurs, we model a source of natural images as a mixture of scene oracles and derive optimal detectors for the task of image steganalysis. Our goal is to guarantee that, when the data follows the statistical model adopted for the hypothesis test, the ROC of the optimal detector bounds the ROC of the ad hoc detector. While the results are applicable beyond the field of image steganalysis, we use this setup to point out possible inconsistencies when comparing both types of detectors and explain guidelines for their proper comparison. Experiments on an artificial cover source with a known model with real steganographic algorithms and deep learning detectors are used to confirm our claims.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Image manipulation**; **Neural networks**;

KEYWORDS

Ad hoc, detector, steganalysis, benchmarking, ROC, cover source, LRT

ACM Reference Format:
Eli Dworetzky, Edgar Kaziakhmedov, and Jessica Fridrich. 2023. On Comparing Ad Hoc Detectors with Statistical Hypothesis Tests. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '23), June 28–30, 2023, Chicago, IL.*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 MOTIVATION

A steganography detector is a mapping from the space of cover objects to the set of real numbers. The detector output, which is called the test statistic, can be thresholded to reach a binary decision on the analyzed object – it is either cover or stego (containing a secret message). Depending on how the test statistic is computed, detectors can be roughly divided into two categories: tests derived from models within the theory of statistical hypothesis testing and ad hoc detectors designed using heuristics [8, 10, 11, 14, 16, 34, 35] or learned on a training set using some machine learning strategy.

The advantage of the former is that the tests can be shown to be optimal within the adopted statistical model with guaranteed error rates. They also provide feedback to the steganographer. The obvious downside is that one often needs to adopt modeling assumptions that are too simple to closely describe digital media objects, which may lead to poor detection performance in practice. This is especially true for complex content-adaptive steganographic schemes.

On the other hand, ad hoc detectors, in particular the kind built with machine learning, can be easily obtained in a fully automatized fashion simply by generating many examples of cover and stego images even without knowing the details of the embedding algorithm. Most importantly, they can often achieve significantly better detection accuracy, albeit at the cost of giving up on insight and losing the ability to control error rates for images outside of the sources on which the detector has been trained.

Both detector types are important, and they have been the subject of intense research ever since the birth of the field of digital media steganography in 1990s. It is thus not surprising that researchers desire to fairly compare (benchmark) these detectors against each other. This comparison is usually based on some scalar quantity derived from the Receiver Operating Characteristic (ROC) curve, such as the Area Under the Curve (AUC), its weighted version wAUC [9], the total error probability under equal priors P_E, missed detection at a fixed false-alarm rate [9], and false-alarm rate for a fixed probability of detection [9]. In many cases, the entire ROC curve is drawn to highlight the trade off between false alarms and missed detections so that practitioners can adjust the decision threshold for specific application-dependent requirements.

A test derived using hypothesis testing usually starts by adopting a statistical model for the specific cover image at hand either by modeling content [32], noise residuals [13], or the acquisition noise [36]. The test is then derived as some form of the Likelihood Ratio Test (LRT) or versions of the Universally Most Powerful (UMP) test [11, 16]. The test statistic is often normalized to follow the same known statistical distribution across cover images so that a fixed threshold guarantees the same false-alarm rate across the cover source. We can also draw an ROC *for each* cover image. In contrast, the ROC for ad hoc detectors can only be drawn empirically from a set of unique cover images and their stego counterparts. As shown

IH&MMSec '23, June 28–30, 2023, Chicago, IL.

ACM ISBN 978-1-XXXX-XXXX…$15.00
https://doi.org/10.1145/XXXXXXX.XXXXXXX

in this paper, ROCs drawn empirically by evaluating the normalized test or the ad hoc detector on a set of cover and stego images are incomparable as they correspond to different hypothesis tests with different optimality criteria. And this stays true *even when the covers exactly follow the adopted statistical model.*

To explain why this happens and to understand how the comparison should be properly executed, we start this paper in the next section by refining the concept of a cover source as a mixture of scene oracles. In Section 3, we describe three different hypothesis testing setups and argue that ROCs typically drawn for ad hoc detectors and for hypothesis tests correspond to detectors equipped with different knowledge and derived for different optimality criteria. Adopting simplifying modeling assumptions on the scene oracle and the associated optimal tests in Section 4, in Section 5 we reveal a relationship among the different hypothesis tests and properties of their ROCs. Based on this analysis, we formulate guidelines on how to fairly compare the ROCs of both detector types. To confirm the validity of our analysis, in Section 6 we include the results of experiments with a real steganographic method and deep learning detectors on an artificial cover source. The paper is concluded in Section 7.

2 SCENE ORACLE

Without loss of generality and for simplicity, we will assume that cover objects are grayscale digital images of natural scenes represented as $W \times H$ matrices of real numbers. We denote the set of cover objects by $\mathcal{X} = \mathbb{R}^{W \times H}$ with $N = W \times H$ the total number of cover elements. The choice to use continuous-valued images is to simplify the analysis to avoid having to deal with quantization. This is feasible since statistical models are usually imposed not on pixels but on some transformed quantities, such as DCT coefficients or noise residuals.

In this paper, random variables are denoted with capital letters with lowercase letters reserved for their realizations. Boldface is used for vectors and matrices.

Technically, a cover source is a distribution (or measure) on $\mathcal{X}$. Since it is unlikely that a reasonable statistical model could be adopted for this high-dimensional distribution [4], we instead postulate the existence of a two-step statistical process that can be used to acquire an abitrary number of images for a dataset. One possibility to visualize this process is to think of it as one or more photographers who take pictures by first selecting a scene and then creating its digital representation by acquiring it with a camera. Both actions involve randomness. We will first discuss the scene selection and then describe the acquisition.

A scene $\mathbf{c} \in \mathcal{X}$ is a digital representation of the physical reality one wishes to photograph in the absence of any imperfections or noise. For now, we will merely assume that there exists a measure ν on $\mathcal{X}$ describing the distribution of the random scene C. Hence, scene selection amounts to obtaining a sample or realization of $C \sim \nu$.

Next, acquisitions of a scene $\mathbf{c}$ follow a distribution over $\mathcal{X}$ conditioned on $C = \mathbf{c}$, denoted by $X \sim s_0(\mathbf{x}|\mathbf{c})$. Meaningful modeling assumptions can be adopted on this conditional distribution, which we call in this paper the *scene oracle*, by considering the properties of imaging sensors, various noise sources, such as the photonic (shot) noise or the readout noise, and a specific development and processing pipeline.[1] We note that modeling frameworks other than those rooted in acquisition models are possible. For instance, in the embedding algorithm MiPOD [32], noise residuals of pixels are modeled as independent realizations of a Gaussian random variable whose variance depends on local content complexity. As for another example, the RJCA [6] models the rounding errors of pixels after JPEG decompression as wrapped Gaussian random variables.

To summarize, sampling from the cover source involves sampling $C = \mathbf{c}$ from ν and then obtaining the actual digital image from the scene oracle $s_0(\mathbf{x}|\mathbf{c})$. The cover source distribution is thus the mixture

$$p_0(\mathbf{x}) = \int_{\mathbf{c}\in\mathcal{X}} s_0(\mathbf{x}|\mathbf{c})\mathrm{d}\nu(\mathbf{c}). \tag{1}$$

In this paper, we frequently consider the conditional probability $\mathbb{P}(E|C)$ of some event E and conditional expectation $\mathbb{E}[Z|C]$ or variance $\mathrm{Var}[Z|C]$ of some random variable Z. Recall that $\mathbb{P}(E|C)$, $\mathbb{E}[Z|C]$, and $\mathrm{Var}[Z|C]$ are random variables since they are functions of C and so it makes sense to consider, e.g., $\mathbb{E}[\mathbb{P}(E|C)] = \int_{\mathbf{c}\in\mathcal{X}} \mathbb{P}(E|C=\mathbf{c})\mathrm{d}\nu(\mathbf{c})$. Additionally, to make mathematical expressions more concise we use $\mathbb{P}_k$, $\mathbb{E}_k$, and Var_k to denote probability, expectation, and variance assuming hypothesis H_k is true ($k = 0$ or 1).

3 TEST FOR MIXTURE OR MIXTURE OF TESTS?

Equipped with the cover source model as a mixture of scene oracles, we now formulate three types of hypothesis test setups the steganalyst may face depending on the available information and optimality criteria. We argue that the way ROCs are typically drawn for ad hoc detectors corresponds to a very different hypothesis testing setup than when drawing an ROC for a normalized test statistic.

Assuming a known fixed payload α in bits per pixel (bpp) for simplicity, we denote the stego distribution conditioned on $\mathbf{c}$ by $s_\alpha(\mathbf{x}|\mathbf{c})$. The stego distribution is thus the mixture

$$p_\alpha(\mathbf{x}) = \int_{\mathbf{c}\in\mathcal{X}} s_\alpha(\mathbf{x}|\mathbf{c})\mathrm{d}\nu(\mathbf{c}), \tag{2}$$

with the same prior distribution on scenes ν. There are two hypothesis tests we can consider when testing between the cover and stego classes: testing a mixture (Case I in Section 3.1) and a mixture of tests for a fixed scene (Section 3.2). Moreover, the latter can be considered for two different optimality criteria (Cases II and III).

3.1 Testing a mixture (Case I)

Given an observable (image) $\mathbf{y} \in \mathcal{X}$, a realization of Y, we consider the hypothesis test of the mixtures in their entirety,

$$\begin{aligned} \mathrm{H}_0 &: Y \sim p_0(\mathbf{x}) = \int_{\mathcal{X}} s_0(\mathbf{x}|\mathbf{c})\mathrm{d}\nu(\mathbf{c}) \\ \mathrm{H}_1 &: Y \sim p_\alpha(\mathbf{x}) = \int_{\mathcal{X}} s_\alpha(\mathbf{x}|\mathbf{c})\mathrm{d}\nu(\mathbf{c}), \end{aligned} \tag{3}$$

where we denote the scene oracle as s_0 to highlight the fact that covers are stego images with nothing embedded.

[1] Acquisition noise models have been used for both steganography [1, 19, 20, 33] and forensics [36].

The test above requires the prior measure ν on the scenes to be known in order for the test to be simple. In this case, the most powerful test is the LRT

$$L(\mathbf{y}) = \log \frac{\int_{\mathcal{X}} s_\alpha(\mathbf{y}|\mathbf{c})\mathrm{d}\nu(\mathbf{c})}{\int_{\mathcal{X}} s_0(\mathbf{y}|\mathbf{c})\mathrm{d}\nu(\mathbf{c})} > \gamma, \tag{4}$$

where γ is a fixed threshold chosen to satisfy a desired probability of false alarm. This general form of the hypothesis test, which we call Case I, corresponds to how ad hoc and data driven detectors are used and trained. In particular, the way we draw the ROC for L,

$$\begin{aligned} P_{\mathrm{D}}(\gamma) &= \mathbb{P}_1\left(L(Y) > \gamma\right) \\ P_{\mathrm{FA}}(\gamma) &= \mathbb{P}_0\left(L(Y) > \gamma\right), \end{aligned} \tag{5}$$

and the empirical ROC for an ad hoc detector $d : \mathcal{X} \to \mathbb{R}$ are consistent in the sense of using a single threshold γ on the output of L and d to partition $\mathcal{X}$ into cover and stego classes.

3.2 Mixture of tests

On the other hand, we can consider a hypothesis test conditioned on the realization $C = \mathbf{c}$. That is, we face the *random* hypothesis test (or a mixture of hypotheses)

$$\begin{aligned} \mathrm{H}_0 &: Y \sim s_0(\mathbf{x}|\mathbf{c}) \\ \mathrm{H}_1 &: Y \sim s_\alpha(\mathbf{x}|\mathbf{c}) \end{aligned} \tag{6}$$

according to the prior distribution ν. Note that the test does not require ν to be known for the test to be simple—only the realization $\mathbf{c}$ needs to be known to prevent any error due to estimating $\mathbf{c}$, and consequently, mismatching tests. The most powerful test is the LRT

$$\ell_{\mathbf{c}}(\mathbf{y}) = \log \frac{s_\alpha(\mathbf{y}|\mathbf{c})}{s_0(\mathbf{y}|\mathbf{c})} > \gamma_{\mathbf{c}}, \tag{7}$$

where the chosen threshold $\gamma_{\mathbf{c}}$ now dependents on $\mathbf{c}$. In this paper, we consider two perspectives for constructing a Neyman–Pearson (NP) detector explained below.

3.2.1 Fixed false alarm (Case II). For each $\mathbf{c}$, one could choose $\gamma_{\mathbf{c}}$ to maximize the conditional probability of correct detection $\mathbb{P}_1\left(\ell_{\mathbf{c}}(Y) > \gamma_{\mathbf{c}}\right)$ while constraining the conditional probability of false alarm to be bounded above by a pre-determined value P_{FA}:

$$\mathbb{P}_0\left(\ell_{\mathbf{c}}(Y) > \gamma_{\mathbf{c}}\right) \le P_{\mathrm{FA}}. \tag{8}$$

Observe that no knowledge of ν is needed to compute the thresholds. Conditional constraints of this form are reasonable to impose if repetitions of the hypothesis testing experiment are potential rather than actual or if the main interest is the particular event $(C = \mathbf{c})$ that occurs (c.f. Chapter 10 of [28]). This constraint is typically adopted when drawing the ROC of $\ell_{\mathbf{c}}(Y)$ in situations when a simple model for $s_\alpha(\mathbf{x}|\mathbf{c})$ is adopted or is easily computable [6, 15, 32] (also, see Section 4).

Assuming $\mathbb{P}_k\left(\ell_{\mathbf{c}}(Y) > t\right)$ is continuous and strictly monotone in t for all $\mathbf{c}$ and k, the optimal threshold $\gamma_{\mathbf{c}}$ uniquely satisfies $\mathbb{P}_0\left(\ell_{\mathbf{c}}(Y) > \gamma_{\mathbf{c}}\right) = P_{\mathrm{FA}}$, meaning we can express it as a function of P_{FA} by $\gamma_{\mathbf{c}}(P_{\mathrm{FA}})$. Therefore, we have the following functional form for the ROC

$$\overline{P}_{\mathrm{D}}(P_{\mathrm{FA}}) = \mathbb{E}\left[\mathbb{P}_1\left(\ell_C(Y) > \gamma_C(P_{\mathrm{FA}})|C\right)\right]. \tag{9}$$

In plain language, the ROC for Case II is computed by drawing the ROCs for each scene and then vertically averaging the ROCs so that P_{FA} is fixed across all scenes. An illustrated example of this is provided in Section 6.1.

If the distribution of $\ell_{\mathbf{c}}(Y)$ only has shift and scale parameters, $\ell_{\mathbf{c}}(Y)$ can be normalized—denoted by $\overline{\ell}_{\mathbf{c}}(Y)$—to follow the same conditional distribution under H_0 for all $\mathbf{c}$, guaranteeing that a *uniform threshold* γ achieves P_{FA} across all scenes. For example, when the test is mean-shifted Gauss-Gauss [23] the normalized test decides H_1 when

$$\overline{\ell}_{\mathbf{c}}(\mathbf{y}) = \frac{\ell_{\mathbf{c}}(\mathbf{y}) - \mathbb{E}_0\left[\ell_{\mathbf{c}}(Y)\right]}{\sqrt{\mathrm{Var}_0\left[\ell_{\mathbf{c}}(Y)\right]}} > \gamma, \tag{10}$$

and its distribution is independent of the scene $\mathbf{c}$ under H_0, i.e., $\overline{\ell}_{\mathbf{c}}(Y) \sim \mathcal{N}(0,1)$ for all $\mathbf{c}$.

Note that the ROC for Case I is in general different than for Case II. This observation is elaborated upon in more detail in Section 5.

3.2.2 Fixed expected false alarm (Case III). Alternatively, one could relax the constraint so that only the *expected* probability of false alarm across scenes is bounded from above:

$$\overline{P}_{\mathrm{FA}} = \mathbb{E}\left[\mathbb{P}_0\left(\ell_C(Y) > \gamma_C|C\right)\right] \le P_{\mathrm{FA}} \tag{11}$$

while maximizing the expected probability of detection

$$\overline{P}_{\mathrm{D}} = \mathbb{E}\left[\mathbb{P}_1\left(\ell_C(Y) > \gamma_C|C\right)\right] \tag{12}$$

$$= \int_{\mathbf{c}\in\mathcal{X}} \mathbb{P}_1(\ell_C(Y) > \gamma_C|C = \mathbf{c})\mathrm{d}\nu(\mathbf{c}). \tag{13}$$

In this case, one would need to know ν in order to choose the thresholds $\gamma_{\mathbf{c}}$. The requirement (11) is reasonable to adopt if this random test is performed a large number of times or if average performance is the main interest [28]. The ROC for this case plots $\overline{P}_{\mathrm{D}}$ as a function of P_{FA}.

In general, the ROC for Case III must bound the ROC for Case II since the optimization constraint for Case III is weaker (i.e. Case III's feasible set of thresholds is a superset of Case II's). The reason for adding Case III will become apparent later as it will help us establish that, under some mild modeling assumptions, the ROC for Case I coincides with the ROC for Case III and thus must bound the ROC for Case II (see Section 5).

4 TYPICAL MODELING SETUP

Among many possibilities for the oracle, the following general model is often adopted in steganalysis [13, 26, 32]:

$$s_0(\mathbf{x}|\mathbf{c}) = \prod_{i=1}^{N} p_0^{(i)}(x_i;\mathbf{c}), \tag{14}$$

which essentially assumes that the individual pixels x_i (cover elements in general) are independent realizations of random variables following distributions $p_0^{(i)}(x;\mathbf{c})$ on their specific ranges. Note that these distributions are allowed to vary with i.

Embedding relative payload α bpp into the cover, each pixel ends up being modified with probability $\beta_i(\mathbf{c})$ with β_i's determined by the particular embedding algorithm and payload α. Computing the change rates β_i from $\mathbf{c}$ rather than the particular realization $\mathbf{x}$ ensures that the ensuing stego distribution is factorizable:

$$s_\alpha(\mathbf{x}|\mathbf{c}) = \prod_{i=1}^{N} p_{\beta_i}^{(i)}(x_i;\mathbf{c}). \tag{15}$$

Without this assumption, stego image pixels would be dependent and it would not in general be tractable to work out the stego distribution for typical content-adaptive stego algorithms, such as the UNIWARD family [21], HILL [29], or MiPOD.

The LRT (7) is

$$\ell_{\mathbf{c}}(\mathbf{y}) = \sum_{i=1}^{N} \log \frac{p^{(i)}_{\beta_i}(y_i;\mathbf{c})}{p^{(i)}_{0}(y_i;\mathbf{c})}. \tag{16}$$

As shown in Part I of the appendix, by the Lindeberg version of the CLT and Taylor expansion w.r.t. β_i at 0, for a large number of pixels N and for small payloads (change rates β_i), the distribution of the non-normalized test is approximately Gaussian under both hypotheses

$$\ell_{\mathbf{c}}(Y) \stackrel{\cdot}{\sim} \begin{cases} \mathcal{N}\left(-\frac{1}{2}\delta^2_{\mathbf{c}}, \delta^2_{\mathbf{c}}\right) & \mathrm{H}_0 \\ \mathcal{N}\left(\frac{1}{2}\delta^2_{\mathbf{c}}, \delta^2_{\mathbf{c}}\right) & \mathrm{H}_1. \end{cases} \tag{17}$$

The quantity $\delta^2_{\mathbf{c}}$ is the deflection coefficient

$$\delta^2_{\mathbf{c}} = \sum_{i=1}^{N} \beta_i^2 F_i \tag{18}$$

and

$$F_i = \int \frac{1}{p^{(i)}_0(y;\mathbf{c})} \left(\frac{\partial p^{(i)}_{\beta}(y;\mathbf{c})}{\partial \beta}\bigg|_{\beta=0} \right)^2 \mathrm{d}y, \tag{19}$$

is the steganographic Fisher information at pixel i [17, 25]. The asymptotic approximation (17) allows us to write the ROC for scene $\mathbf{c}$ as

$$P_{\mathrm{D}}(P_{\mathrm{FA}}) = Q\left(Q^{-1}(P_{\mathrm{FA}}) - \delta_{\mathbf{c}}\right), \tag{20}$$

where $Q(x) = \int_x^\infty (2\pi)^{-1/2} e^{-t^2/2} \mathrm{d}t$ is the standard normal tail probability function.

5 COMPARISON IN PRACTICE

The hypothesis tests above will in practice be formulated for finite mixtures with images from a dataset formed by sampling ν n-times, obtaining scenes $\mathbf{c}_1, \ldots, \mathbf{c}_n$, and then sampling each scene oracle $X_i \sim s_0(\mathbf{x}|\mathbf{c}_i)$ once to form a dataset of n cover images, $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

In order to simplify the analysis and to be able to express our claims in a closed-form, the rest of this paper will assume that the LRT for each scene $\mathbf{c}$, $\ell_{\mathbf{c}}(Y)$, follows the asymptotic Gaussian distribution (17) under each hypothesis exactly. This approximation is often very tight even for small 256×256 images typically used in steganography experiments as illustrated in Figure 8 in Section 5 and also as apparent in numerous prior art [5, 11, 32, 36].

5.1 Case I

Typical datasets of natural images (even with a size on the order of millions) do not contain multiple acquisitions of the same scene, or even similar scenes—images taken in the same physical location and camera settings will differ due to changes in lighting conditions, camera shake, etc. Taking multiple acquisitions of the same scene is only feasible in laboratory conditions. Therefore, we assume that for any realistic ν, given scenes $\mathbf{c}_1, \ldots, \mathbf{c}_n$ independently sampled according to ν, the log-likelihood $\log \sum_{i=1}^n s_\alpha(\mathbf{x}|\mathbf{c}_i)$ will be numerically close to $\log s_\alpha(\mathbf{x}|\mathbf{c}_i)$ for some i since the other $n-1$ terms in the sum should be approximately zero.[2] Thus, the distribution of the LRT for Case I, $L(Y) = \log \frac{\sum_{i=1}^n s_\alpha(Y|\mathbf{c}_i)}{\sum_{i=1}^n s_0(Y|\mathbf{c}_i)}$, is well approximated by a mixture of the distributions of $\ell_i(Y) \triangleq \ell_{\mathbf{c}_i}(Y) = \log \frac{s_\alpha(Y|\mathbf{c}_i)}{s_0(Y|\mathbf{c}_i)}$. In other words, under both hypotheses ($k = 0$ or 1) we have that

$$\mathbb{P}_k\left(L(Y) > \gamma\right) \approx \sum_{i=1}^{n} \mathbb{P}_k\left(\ell_i(Y) > \gamma\right). \tag{21}$$

The LRT $L(Y)$ (4) will thus be a Gaussian mixture under the simplifying assumption described in the beginning of this section.

Let μ be the distribution of the deflection coefficient δ^2_C as introduced in Section 4, $C \sim \nu$. Notice that, unlike ν, μ can be feasibly estimated from a large dataset of images because it is a one-dimensional distribution. From (17), we have for the probability of detection and false alarm probability

$$\begin{aligned} P_{\mathrm{D}}(\gamma) &= \mathbb{P}_1\left(L(Y) > \gamma\right) \\ &\approx \int_{t\in\mathbb{R}} \mathbb{P}\left(\mathcal{N}(t/2, t) > \gamma\right) \mathrm{d}\mu(t) && (22) \\ &= \int_{t\in\mathbb{R}} \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi t}} \exp\left(\frac{-(x-t/2)^2}{2t}\right) \mathrm{d}x \mathrm{d}\mu(t), && (23) \\ P_{\mathrm{FA}}(\gamma) &= \int_{t\in\mathbb{R}} \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi t}} \exp\left(\frac{-(x+t/2)^2}{2t}\right) \mathrm{d}x \mathrm{d}\mu(t). && (24) \end{aligned}$$

We now argue that empirically drawing the ROC using only *one sample per scene* is asymptotically equivalent to drawing the ROC using many acquisitions. This can be argued in the same style as proofs of the Glivenko–Cantelli Theorem [28]. Consider the realizations $\mathbf{c}_1, \ldots, \mathbf{c}_n$ independently sampled according to ν and their deflections $\delta^2_1, \delta^2_2, \ldots, \delta^2_n$. By the strong law of large numbers and realizing that $\delta^2_C \sim \mu$, we have the following convergence for the random empirical distribution under H_1

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_k\left(\ell_i(Y_i) > \gamma\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\mathcal{N}(\mp\delta^2_i/2, \delta^2_i) > \gamma\right) \\ &\to \mathbb{E}\left[\mathbb{P}\left(\mathcal{N}(\mp\delta^2_C/2, \delta^2_C) > \gamma \mid \delta^2_C\right)\right] \\ &\approx \mathbb{P}_k\left(L(Y) > \gamma\right) \end{aligned} \tag{25}$$

almost surely in γ as $n \to \infty$ under both hypotheses $k = 0$ and 1 with the signs of the Gaussian means equal to $-$ and $+$, respectively.[3] The LHS amounts to sampling one acquisition from a finite n number of scenes. Since a formula similar to (25) holds for $\mathbb{P}_0\left(\ell_i(Y_i) > \gamma\right)$ with the opposite means of the Gaussians, the ROC for the dataset of n images $\mathbf{x}_i$ with deflections δ^2_i is in a parametric form

$$P_{\mathrm{D}}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\mathcal{N}(\delta^2_i/2, \delta^2_i) > \gamma\right) \tag{26}$$

$$P_{\mathrm{FA}}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\mathcal{N}(-\delta^2_i/2, \delta^2_i) > \gamma\right). \tag{27}$$

[2] The validity of this assumption relies on N being large, which is a reasonable assumption given typical image sizes produced by modern cameras and phones.

[3] Additionally, the convergence is uniform in γ due to the monotonicity of the right-tail probability function.

Thus, to fairly compare the ROC of an ad hoc detector with a statistical test, we need to draw the ROC for the test from the *non-normalized* LRT (7). Note that in this case, due to (6) the ROC will be *symmetrical* about the minor diagonal in agreement with the observation that the ROCs of ad hoc detectors built using machine learning are indeed approximately symmetrical (also see Section 6.3).

5.2 Case II

In this section, we show that the ROC for Case II given by Eq. (9) is asymmetrical unlike the ROC for Case I. Given scene $\mathbf{c}$, the difficulty of detecting steganography is quantified by the deflection coefficient $\delta_{\mathbf{c}}^2$ associated with the UMP test for the *random* hypothesis. For a fixed false-alarm P_{FA}, the probability of correct detection is (see Eq. (20))

$$\overline{P}_{\text{D}}(P_{\text{FA}}) = \mathbb{E}[Q(Q^{-1}(P_{\text{FA}}) - \delta_C)], \tag{28}$$

where the expectation is taken over scenes.

Denoting $m = \mathbb{E}[\delta_C]$, using Taylor expansion of $Q(x)$ at $Q^{-1}(P_{\text{FA}})-m$, the expected ROC (28) can be written as

$$\begin{aligned}\overline{P}_{\text{D}}(P_{\text{FA}}) &= \mathbb{E}\left[Q(Q^{-1}(P_{\text{FA}}) - \delta_C)\right] \\ &= \mathbb{E}\Big[\sum_{k=0}^{\infty} \frac{(m-\delta_C)^k}{k!} Q^{(k)}(Q^{-1}(P_{\text{FA}}) - m)\Big] \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k c_k}{k!} Q^{(k)}(Q^{-1}(P_{\text{FA}}) - m),\end{aligned} \tag{29}$$

where c_k is the kth central moment of δ_C. We note that the Taylor expansion converges rather quickly due to the following bound on the kth derivative of the Q function

$$|Q^{(k)}(x)| \le \sqrt{\frac{k!}{2\pi}} \text{ for all } x, \tag{30}$$

which follows from the Cramér inequality [22], and the fact that central moments of a bounded random variable[4] can grow only polynomially fast.

The analysis shown above highlights an important property of ROCs drawn under Case II; the ROC can be highly *asymmetric*, bending toward the y-axis that allows $\overline{P}_{\text{D}}(P_{\text{FA}})$ to be large for small P_{FA}. This is because only the first term for $k = 0$ in expansion (29) is symmetrical about the minor diagonal. Ultimately, the shape of the ROC depends on the central moments c_k, which when contrasted with Case I is a very different ROC. [5]

5.3 Case III

We wish to determine the thresholds γ_i for each conditional test ℓ_i in order to maximize average P_{D}

$$\overline{P}_{\text{D}}(\gamma_1, \ldots, \gamma_n) \triangleq \sum_{i=1}^{n} \frac{1}{n} \mathbb{P}_1\left(\ell_i(Y) > \gamma_i\right), \tag{31}$$

while satisfying the constraint for the average P_{FA}

$$\overline{P}_{\text{FA}}(\gamma_1, \ldots, \gamma_n) \triangleq \sum_{i=1}^{n} \frac{1}{n} \mathbb{P}_0\left(\ell_i(Y) > \gamma_i\right) \le P_{\text{FA}}. \tag{32}$$

Using the method of Lagrange multipliers as shown in Part II of the appendix, the thresolds $\gamma_i = \gamma$ will be the same across scenes. Setting $\gamma_i = \gamma$ in Eq. 31 and using Eq. 21, we have

$$\begin{aligned}\mathbb{P}_k\left(L(Y) > \gamma\right) &\approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_k\left(\ell_i(Y) > \gamma\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_k\left(\ell_i(Y) > \gamma_i\right),\end{aligned} \tag{33}$$

and therefore, the ROCs for LRT in Eq. (3) (Case I) and LRT in Eq. (6) with thresholds chosen to satisfy P_{FA} only on average (Case III) are *equal*. Since in general the UMP test for Case III should upper bound the UMP for Case II, Case I must also upper bound Case II when under the assumptions of Section 5.

At first, the result of the UMP of Case I dominating the UMP of Case II may seem counter-intuitive since Case II fundamentally knows the exact scene while Case I is scene-agnostic. However, Case II is ultimately crippled by the stringency of the constraint even though it is afforded detection power in the sense of having knowledge of the scene.

6 EXAMPLES AND PRACTICAL IMPLICATIONS

In general, one cannot guarantee any relationship between an ROC drawn from outputs of an ad hoc detector and a test derived from a model. This is, of course, because real images do not necessarily follow the model but, as argued in this paper, also because an ROC drawn from outputs of an ad hoc detector on cover and stego images corresponds to Case I, while drawing the ROC for a normalized test corresponds to Case II.

> Since the ROC for Case I bounds the one for Case II, ad hoc detectors are thus portrayed in better light or one can say that the normalized statistical test is disadvantaged.

As explained in the introduction, our goal is to provide guidelines on comparing both detector types so that if the model is true the ROC of the most powerful detector indeed bounds the one for the ad hoc detector. To guarantee this, we need to either make sure that both detector types are evaluated as in Case I or in Case II. The former should be used when working with a dataset without access to the scene oracles, which pretty much covers all practical cases. The latter is an option when working with an artificial dataset that exactly follows the model.

Below, we start with a simple illustrative example to reinforce how the ROC averaging differs between Case I and II. We then point out cases of prior art with comparisons that involve both Case I and II. To solidify the theoretical claims made above, we include experiments on an artificial dataset to demonstrate the consequences of conflating Case I and Case II and point out the properties of ROCs corresponding to both cases.

[4] The boundedness of δ_C follows from the fact that the Fisher information F_i is bounded since natural images contain certain minimum amount of acquisition noise. Unbounded or infinite F_i would indicate deterministic pixels and in general pixels where the embedding changes are extremely detectable and thus should be avoided by the embedding algorithm anyway.

[5] In practice, we use Eq. (28) to draw the ROC. The form given in (29) is strictly used for analysis in this paper.

6.1 Toy Example

Suppose the modeling assumptions in Section 4 and (21) hold. Consider a cover source consisting of three scenes whose deflections are $\delta_1^2 = 1$, $\delta_2^2 = 6$, $\delta_3^2 = 12$ and prior probabilities are $\nu(\mathbf{c}_1) = 0.3$, $\nu(\mathbf{c}_2) = 0.2$, $\nu(\mathbf{c}_3) = 0.5$. Figure 1 (left) shows the distribution of $\ell_i(Y)$ under both hypotheses for each scene as well as the distribution of $L(Y)$ under both hypotheses. On the right, the figure shows the distributions of the normalized LRTs $\bar{\ell}_i(Y)$ per Eq. (10). Figure 2 illustrates on the left that the ROCs for each scene are vertically averaged in order to draw the ROC for Case II. On the right, Figure 2 demonstrates that the ROC for Case I bounds the ROC for Case II. Additionally, note that the ROC for Case I is symmetric while the ROC for Case II bends towards the y-axis.

6.2 Examples from prior art

The ROCs of three Asymptotically Universally Most Powerful (AUMP) normalized tests of non-adaptive Least Significant Bit (LSB) Matching are compared in Fig. 6 of [11] with ROCs of two ad hoc detectors – the center of mass of a two-dimensional histogram characteristic function [24] and a heuristic measure of histogram smoothness (ALE) [37]. In [34], normalized tests derived for the Jsteg algorithm are contrasted with Zhang's quantitative detector [38]. The ROC of AUMP test derived in [16] is compared with a version of the Generalized Category Attack [27].

Another interesting example worth mentioning appears in [32]. The ROC of an hoc detector in the form of a spatial rich model [18] with the low complexity linear classifier [12] is compared with the most powerful test in Fig. 6 of [32]. Here, the comparison is consistent because the stego images are embedded with variable payloads to guarantee the same deflection coefficient across all images. Thus, the Gaussian mixture of Case I becomes a single Gaussian distribution and, at the same time, the ROC of Case II is also a Gaussian ROC since the expectation in (28) can be removed as $\delta_{\mathbf{c}} = \delta$ for all $\mathbf{c}$. For this deflection-limited sender, the ROCs for Case I and II coincide.

The most glaring example of improper comparison (mismatching Case I and II hypothesis tests) appears in the "Are We There Yet" paper [5]. The authors investigated the limits of machine learning detectors by working with an artificial cover source and embedding modified so that a closed form for the most powerful detector (a LRT) exists. The ROCs of ad hoc detectors, however, appear to intersect the ROCs of the corresponding optimal detectors (see Fig. 5 for LSBM and Fig. 6 for S-UNIWARD in [5]). This is due to the fact that the ROCs of ad hoc detectors were based on Case I while the LRT's ROCs corresponded to Case II, making their ROCs highly non-symmetrical and weaker in comparison to the ad hoc detectors. To paraphrase the authors of [5], "we are less there than we thought." Below, we used the same datasets and embedding algorithms to show that when both detectors are compared either based on testing in Case I or both as in Case II, the LRT's ROC indeed bounds the ROCs of the ad hoc detectors as this must be the case for such artificial datasets that follow the modeling assumptions exactly.

6.3 Experiments on an artificial dataset

To illustrate the impact of various comparison methods on ROCs in real-world scenarios, we conduct experiments using an artificial dataset which is derived from natural images. The dataset of natural images is the union of BOSSbase 1.01 [2] and BOWS2 [3], each containing 10,000 grayscale images resized to 256×256 pixels using Matlab's `imresize` with default parameters. The images have been stored in an uncompressed format and split into three subsets: training (TRN), validation (VAL), and testing (TST). The training set includes all 10,000 images from BOWS2, as well as 4,000 randomly selected images from BOSSbase. The remaining images from BOSSbase were randomly divided into a validation set with 1,000 images and a testing set with 5,000 images. This dataset of images is referred to as the 'raw image dataset.'

To enable access to a scene oracle, we create an artificial version of the dataset in the same manner as in [5]. This artificial dataset assumes that pixels are independent realizations of a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, N$. To create this dataset, each raw image was first denoised using the wavelet denoising filter [31] with $\sigma_{\text{Den}} = 10$. The dynamic range of the image was then narrowed and the values rounded to the nearest integers to ensure that the pixel values after adding noise fall within the range $[0, 255]$ with high probability. The denoised images represent the means μ_i of the Gaussian distributions, while the variances σ_i^2 were obtained using MiPOD's variance estimator [32] from the original raw images. Finally, cover image pixels are sampled from $\mathcal{N}(\mu_i, \sigma_i^2)$, and the values are rounded to the closest integer and clipped to the range $[1, 254]$ to enable ternary embedding in all pixels. Further details on the creation of this dataset can be found in [5].

To generate stego images, we use the content-adaptive scheme S-UNIWARD [21] and non-adaptive Least Significant Bit Matching (LSBM). As already pointed out in Section 4, to have a closed form for the LRT, the embedding change probabilities (the selection channel) were computed from raw images. Since our covers are curbed to $[1, 254]$, no wet costs need to be assigned to pixels at the boundary of the dynamic range. S-UNIWARD's embedding simulator was used to simulate embedding payload $\alpha = 0.6$ bpp, while the impact of LSBM was simulated by fixing the global change rate $\beta = 0.01$.

As an ad-hoc detector, we used the SRNet pre-trained on a binary task of steganalyzing J-UNIWARD [21] (the so-called JIN pre-training exactly as described in [7]). This pre-trained SRNet was used to train detectors for both stego sources. Each model was trained for 100 epochs with a batch size of 64. The Adamax optimizer was used in order to optimize the model's performance with the initial learning rate set to 10^{-3}. A cosine scheduler was employed to reduce the learning rate smoothly down to 2×10^{-5} during training.

To evaluate SRNet's performance in Case II, 5000 covers were sampled with the scene oracle for each image and then embedded to create the same number of stego images. Both cover and stego images were fed into the network trained for the corresponding stego algorithm, and the output logits were stored. The process was repeated for each scene from the testing set of the artificial dataset. To plot the SRNet ROC in Case I in Figure 3 (shown as the red dashed line), the logits computed for all images in the test set

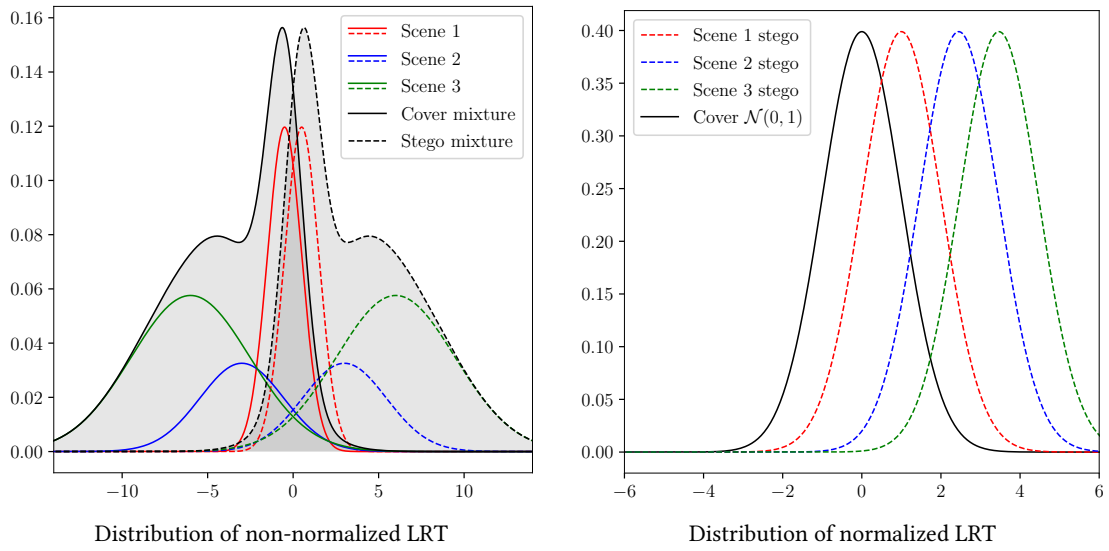

Figure 1: Distributions of LRTs in the case of a cover source with three scenes whose deflections and priors are provided in Section 6.1. Left: The densities of $\ell_i(Y)$ under H_0, $\mathcal{N}(-\frac{1}{2}\delta_i^2, \delta_i^2)$, are drawn in solid color, and the densities of $\ell_i(Y)$ under H_1, $\mathcal{N}(\frac{1}{2}\delta_i^2, \delta_i^2)$, are drawn with dashed colored lines. Their heights are adjusted to convey the priors in the mixture. The density of $L(Y)$, which is a Gaussian mixture under both hypotheses assuming (21) holds, is drawn in black. Right: Distribution of normalized LRT $\overline{\ell}_i(Y)$ under H_0, which is $\mathcal{N}(0,1)$ for all scenes, and under H_1, which is $\mathcal{N}(\delta_i, 1)$ for each scene.

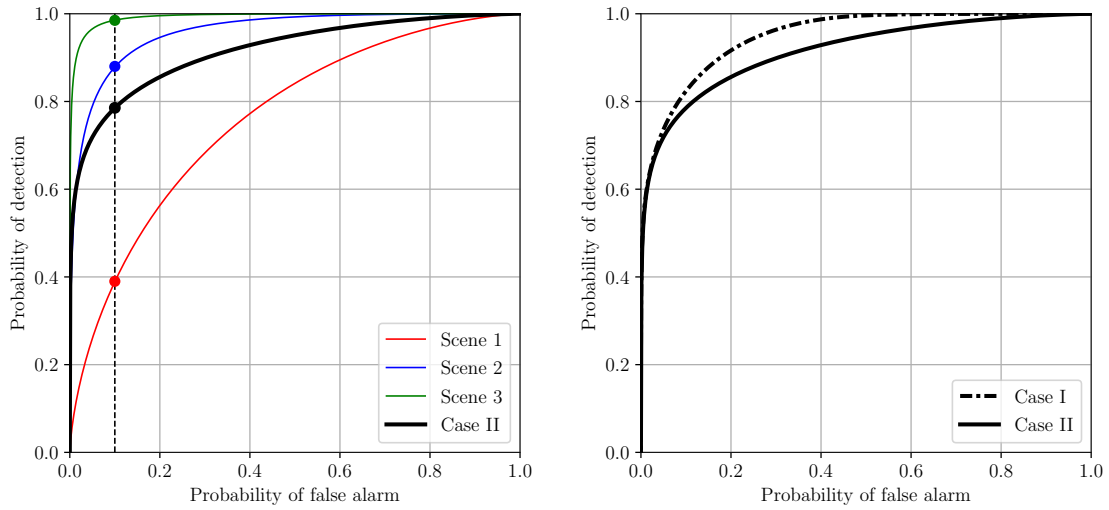

Figure 2: Left: Case II ROC $\overline{P}_{\mathrm{D}}(P_{\mathrm{FA}})$ in relation to the ROCs drawn for each individual scene. To compute $\overline{P}_{\mathrm{D}}(P_{\mathrm{FA}})$ at, e.g., $P_{\mathrm{FA}} = 0.1$ the y-coordinate of the black point is computed by taking the average (weighted by the priors) of the y-coordinates of the colored points. Right: ROC for Case I drawn next to ROC for Case II.

were aggregated into a single vector and fed to the ROC plotter. To plot SRNet's ROC in Case II in Figure 3 (blue dashed line), the ROC for each test image (scene) was computed separately from the 2×5000 images, and the values of P_{D} were averaged across the scenes for the same P_{FA}value.

In the same figure, we also plot the ROCs of the LRT on the test set for both cases. Since the deflection coefficient δ_{c}^2 can be computed from our model (18), the ROCs for both Case I and II can be computed analytically from Eqs. (26)–(27) and Eq. (28). Since these analytic forms are based on asymptotic approximations, we verify their tightness by including the empirically sampled ROCs of the LRT (dotted lines) computed from cover / stego images in the same manner as SRNet (see previous paragraph).

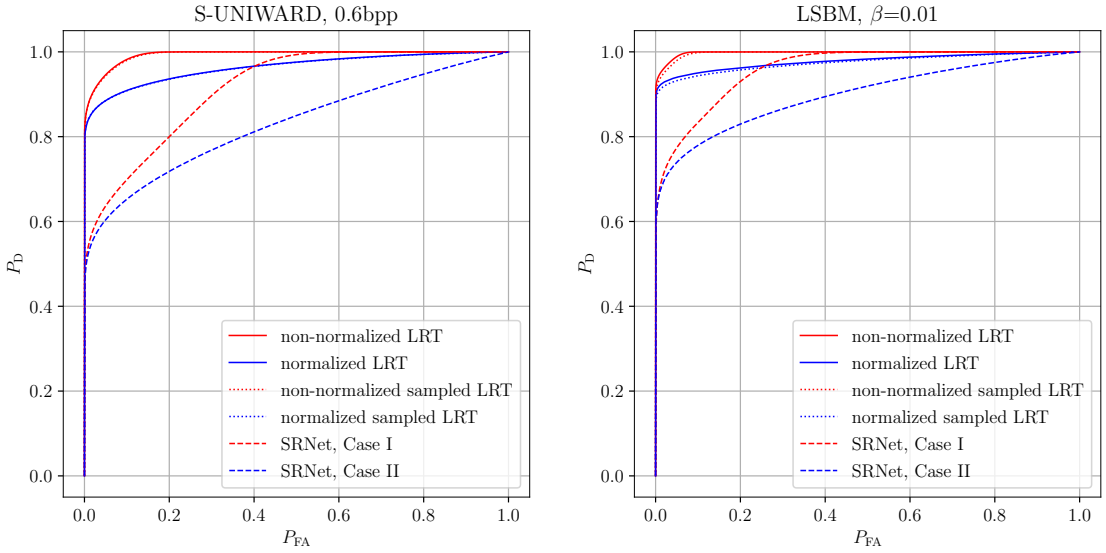

Figure 3: The analytic ROC of the LRT (solid line) and the sampled ROCs of the LRT (dotted line) and SRNet (dashed line) for S-UNIWARD (left) and non-adaptive LSBM (right). Red and blue corresponds to Case I and II, respectively.

The following can be inferred from Figure 3. In agreement with our analysis, while the ROCs for Case I are indeed symmetrical for both types of detectors, for Case II they bend towards the y-axis. Second, for large false alarms SRNet's ROC for Case I can be above LRT's ROC for Case II, creating a seeming contradiction with the LRT being the most powerful detector. Comparing both detectors correctly (always both either for Case I or both for Case II), LRT's ROC bounds the SRNet's (the dashed line is always below the solid line for the same color). Additionally, the non-normalized LRT (Case I) bounds the normalized one (Case II). While we also see this effect for the ad-hoc detector, it is not anticipated in theory. These observations hold for both embedding algorithms, which further supports the significance of the problem and confirms the validity of our results.

7 CONCLUSIONS

This paper deals with the problem of how to fairly compare ROCs of ad hoc (data driven) detectors and normalized tests derived from statistical models. In general, ROCs of ad hoc detectors the way they are usually drawn correspond to a test between mixtures (Case I) while ROCs drawn from normalized statistics correspond to a mixture of tests while guaranteeing constant false alarm rate across images (Case II). When the datasets exactly follow the adopted statistical model, the ROCs of Case I are symmetrical and bound the (highly asymmetrical) ROCs for Case II. Thus, when comparing a test statistic against an ad hoc detector, one should use the LRT (or some other version of the most powerful detector) in its *non-normalized* form.

On artificial datasets, a second option is available for a fair comparison since the ad hoc detector can produce an ROC for each image (scene), allowing thus a comparison via Case II – by averaging the ROCs of individual images. This paper shows examples of prior art where the comparison is inconsistent in that ROCs for Cases I and II are shown in the same graph. For artificial datasets, this can lead to a paradoxical (impossible) situation when the ad hoc detector ROC intersects the ROC of the most powerful detector, making the procedural mistake in [5] obvious.

8 ACKNOWLEDGEMENTS

The work on this paper was supported by the National Science Foundation under grant No. 2028119.

APPENDIX

Part I: Asymptotic form of LRT

Here, we explain the asymptotic form of the LRT from Section 4. First, we compute the leading terms of the expectations and variances of the log LRT for the ith pixel

$$\Lambda_i(x) \triangleq \log \frac{p^{(i)}_{\beta_i}(x;\mathbf{c})}{p^{(i)}_{0}(x;\mathbf{c})} \tag{34}$$

under both hypotheses while expanding the distribution of the stego pixels at $\beta = 0$:

$$p^{(i)}_{\beta_i}(x;\mathbf{c}) = p^{(i)}_{0}(x;\mathbf{c}) + \beta_i \left.\frac{\partial p^{(i)}_{\beta}(x;\mathbf{c})}{\partial \beta}\right|_{\beta=0} + O(\beta_i^2). \tag{35}$$

For the expectations,

$$\mathbb{E}_0[\Lambda_i(x)] = -D_{\mathrm{KL}}(p^{(i)}_0 || p^{(i)}_{\beta_i}) = -\frac{1}{2}F_i\beta_i^2 + O(\beta_i^3) \tag{36}$$

$$\mathbb{E}_1[\Lambda_i(x)] = D_{\mathrm{KL}}(p^{(i)}_{\beta_i} || p^{(i)}_0) = \frac{1}{2}F_i\beta_i^2 + O(\beta_i^3), \tag{37}$$

because the leading term of both KL divergence terms is the same.

For the variances under either hypothesis ($k \in \{0, 1\}$)

$$\begin{aligned}
\mathrm{Var}_k[\Lambda_i(x)] &\doteq \mathrm{Var}_k\left[\log \frac{p_0^{(i)}(x;\mathbf{c}) + \beta_i \frac{\partial}{\partial\beta} p_\beta^{(i)}(x;\mathbf{c})\Big|_{\beta=0}}{p_0^{(i)}(x;\mathbf{c})}\right] \\
&\doteq^{(a)} \beta_i^2 \mathrm{Var}_k\left[\frac{\frac{\partial}{\partial\beta} p_\beta^{(i)}(x;\mathbf{c})\Big|_{\beta=0}}{p_0^{(i)}(x_i;\mathbf{c})}\right] \\
&=^{(b)} \beta_i^2 \left(\mathbb{E}_k\left[\left(\frac{\frac{\partial}{\partial\beta} p_\beta^{(i)}(x;\mathbf{c})\Big|_{\beta=0}}{p_0^{(i)}(x_i;\mathbf{c})}\right)^2\right] - O(\beta^4)\right) \\
&\doteq \beta_i^2 F_i, \qquad (38)
\end{aligned}$$

where F_i is the Fisher information (19). The approximation $\doteq^{(a)}$ is due to Taylor expansion of the natural logarithm, $\log(1+x) = x + O(x^2)$, while the equality $=^{(b)}$ uses $Var[X] = E[X^2] - E[X]^2$. The leading term of the expectation on the third line is F_i under both hypotheses.

Using the Lindeberg version of the CLT, as $N \to \infty$ the following normalized LRT converges in distribution

$$\frac{\sum_{i=1}^N \Lambda_i(Y) - \mathbb{E}_k[\Lambda_i(Y)]}{\sqrt{\sum_{i=1}^N \mathrm{Var}_k[\Lambda_i(Y)]}} \rightsquigarrow \mathcal{N}(0,1) \text{ under } \mathrm{H_k}, \qquad (39)$$

where $\delta_{\mathbf{c}}^2 = \sum_{i=1}^N \beta_i^2 F_i$. Note the expectation and variance terms in (39) depend on k. The convergence under both hypotheses motivates the Gaussian approximation of $\ell_{\mathbf{c}}(Y)$ used in (17).

The authors wish to point out that the above result was obtained under the assumptions that the CLT can be applied and the Taylor expansions converge. These assumptions would need to be verified for each particular combination of image representation and embedding method. Just as importantly, one should verify the tightness of the asymptotic results. A specific example of this for Gaussian pixels and Gaussian stego mixture appears in the appendix of [32].

Part II: Optimizing error rates for expected false alarm

In this section, we derive the thresholds γ_i that maximize average P_{D} (31) while satisfying a desired average P_{FA} (32) under the testing scenario described in Section 4 and 5.3. For each scene indexed by $i = 1, \ldots, n$, we write its associated deflection $\delta_i^2 > 0$ (18). Using the method of Lagrange multipliers and the approximation in Eq. (17), the Lagrangian

$$\begin{aligned}
\mathcal{L}(\gamma_1, \ldots, \gamma_n, \lambda) &= \sum_{i=1}^n \mathbb{P}_1\left(\ell_i(Y) > \gamma_i\right) \qquad (40) \\
&\quad - \lambda\left(\sum_{i=1}^n \mathbb{P}_0\left(\ell_i(Y) > \gamma_i\right) - nP_{\mathrm{FA}}\right)
\end{aligned}$$

with $\lambda \geq 0$ gives us the following necessary conditions for $\gamma_1, \ldots, \gamma_n$ to be a maximizer:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \gamma_i} &= -\frac{1}{\sqrt{2\pi\delta_i^2}} \exp\left(\frac{-(\gamma_i - \frac{1}{2}\delta_i^2)^2}{2\delta_i^2}\right) \\
&\quad + \frac{\lambda}{\sqrt{2\pi\delta_i^2}} \exp\left(\frac{-(\gamma_i + \frac{1}{2}\delta_i^2)^2}{2\delta_i^2}\right) = 0,
\end{aligned}$$

for all i. If $\lambda = 0$, then $\gamma_i = \infty$ for all i, implying $P_{\mathrm{D}} = 0$. Thus, $\lambda > 0$, meaning the maximizer achieves the average P_{FA} exactly. Solving for γ_i, observe that we must equivalently satisfy

$$\begin{aligned}
&\exp\left(\frac{-(\gamma_i - \frac{1}{2}\delta_i^2)^2}{2\delta_i^2}\right) = \lambda \exp\left(\frac{-(\gamma_i + \frac{1}{2}\delta_i^2)^2}{2\delta_i^2}\right) \\
\Leftrightarrow \quad & 2\delta_i^2 \log(\lambda) = (\gamma_i + \frac{1}{2}\delta_i^2)^2 - (\gamma_i - \frac{1}{2}\delta_i^2)^2 \\
\Leftrightarrow \quad & \gamma_i = \log(\lambda)
\end{aligned}$$

for each i. In other words, all n thresholds are equal to some constant $\gamma_i = \gamma$.

The critical point is verified to be a maximizer by seeing that the signs of leading principal minors of the bordered Hessian alternate with the smallest being positive (see Ch. 7, Theorem 12 of [30]).

REFERENCES

[1] P. Bas. Steganography via cover-source switching. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, December 4-7 2016.

[2] P. Bas, T. Filler, and T. Pevný. Break our steganographic system – the ins and outs of organizing BOSS. In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, volume 6958 of Lecture Notes in Computer Science, pages 59–70, Prague, Czech Republic, May 18–20, 2011.

[3] P. Bas and T. Furon. BOWS-2. http://bows2.ec-lille.fr, July 2007.

[4] R. Böhme. *Improved Statistical Steganalysis Using Models of Heterogeneous Cover Signals*. PhD thesis, Faculty of Computer Science, Technische Universität Dresden, Germany, 2008.

[5] M. Boroumand, J. Fridrich, and R. Cogranne. Are we there yet? In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2019*, San Francisco, CA, January 26–30, 2019.

[6] J. Butora and J. Fridrich. Reverse JPEG compatibility attack. *IEEE Transactions on Information Forensics and Security*, 15:1444–1454, 2020.

[7] J. Butora, Y. Yousfi, and J. Fridrich. How to pretrain for steganalysis. In D. Borghys and P. Bas, editors, *The 9th ACM Workshop on Information Hiding and Multimedia Security*, Brussels, Belgium, 2021. ACM Press.

[8] R. Cogranne. Selection-channel-aware reverse JPEG compatibility for highly reliable steganalysis of JPEG images. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*, pages 2772–2776, Barcelona, Spain, May 4–8, 2020.

[9] R. Cogranne, Q. Giboulot, and P. Bas. The ALASKA steganalysis challenge: A first step towards steganalysis "Into the wild". In R. Cogranne and L. Verdoliva, editors, *The 7th ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, July 3–5, 2019. ACM Press.

[10] R. Cogranne and F. Retraint. Application of hypothesis testing theory for optimal detection of LSB matching data hiding. *Signal Processing*, 93(7):1724–1737, July, 2013.

[11] R. Cogranne and F. Retraint. An asymptotically uniformly most powerful test for LSB Matching detection. *IEEE Transactions on Information Forensics and Security*, 8(3):464–476, 2013.

[12] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich. Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19, 2015.

[13] R. Cogranne, C. Zitzmann, L. Fillatre, F. Retraint, I. Nikiforov, and P. Cornu. A cover image model for reliable steganalysis. In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, Lecture Notes in Computer Science, pages 178–192, Prague, Czech Republic, May 18–20, 2011.

[14] R. Cogranne, C. Zitzmann, F. Retraint, I. Nikiforov, L. Fillatre, and P. Cornu. Statistical detection of LSB Matching using hypothesis testing theory. In M. Kirchner

and D. Ghosal, editors, *Information Hiding, 14th International Conference*, volume 7692 of Lecture Notes in Computer Science, pages 46–62, Berkeley, California, May 15–18, 2012.

[15] E. Dworetzky, E. Kaziakhmedov, and J. Fridrich. Advancing the JPEG compatibility attack: Theory, performance, robustness, and practice. In Y. Yousfi, C. Pasquini, and A. Bharati, editors, *The 11th ACM Workshop on Information Hiding and Multimedia Security*, Chicago, IL, June 28–30, 2023. ACM Press.

[16] L. Fillatre. Adaptive steganalysis of least significant bit replacement in grayscale images. *IEEE Transactions on Signal Processing*, 60(2):556–569, 2011.

[17] T. Filler and J. Fridrich. Fisher information determines capacity of ϵ-secure steganography. In S. Katzenbeisser and A.-R. Sadeghi, editors, *Information Hiding, 11th International Conference*, volume 5806 of Lecture Notes in Computer Science, pages 31–47, Darmstadt, Germany, June 7–10, 2009. Springer-Verlag, New York.

[18] J. Fridrich and J. Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, June 2011.

[19] Q. Giboulot, P. Bas, and R. Cogranne. Multivariate side-informed Gaussian embedding minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 17:1841–1854, 2022.

[20] Q. Giboulot, R. Cogranne, and P. Bas. Detectability-based JPEG steganography modeling the processing pipeline: The noise-content trade-off. *IEEE Transactions on Information Forensics and Security*, 16:2202–2217, 2021.

[21] V. Holub, J. Fridrich, and T. Denemark. Universal distortion design for steganography in an arbitrary domain. *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, 2014:1, 2014.

[22] J. Indritz. An inequality for Hermite polynomials. 12(6):981–983, 1961.

[23] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, volume II. Upper Saddle River, NJ: Prentice Hall, 1998.

[24] A. D. Ker. Steganalysis of LSB matching in grayscale images. *IEEE Signal Processing Letters*, 12(6):441–444, June 2005.

[25] A. D. Ker. Estimating steganographic fisher information in real images. In S. Katzenbeisser and A.-R. Sadeghi, editors, *Information Hiding, 11th International Conference*, volume 5806 of Lecture Notes in Computer Science, pages 73–88, Darmstadt, Germany, June 7–10, 2009. Springer-Verlag, New York.

[26] A. D. Ker. On the relationship between embedding costs and steganographic capacity. In M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017. ACM Press.

[27] K. Lee and A. Westfeld. Generalized category attack – improving histogram-based attack on JPEG LSB embedding. In T. Furon, F. Cayre, G. Doërr, and P. Bas, editors, *Information Hiding, 9th International Workshop*, volume 4567 of Lecture Notes in Computer Science, pages 378–392, Saint Malo, France, June 11–13, 2007. Springer-Verlag, Berlin.

[28] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses, Third Edition*. Springer, 3rd edition, 2005.

[29] B. Li, M. Wang, and J. Huang. A new cost function for spatial image steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP*, Paris, France, October 27–30, 2014.

[30] J. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 3rd edition, 2007.

[31] M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, December 1999.

[32] 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.

[33] T. Taburet, P. Bas, W. Sawaya, and J. Fridrich. A natural steganography embedding scheme dedicated to color sensors in the JPEG domain. In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2019*, San Francisco, CA, January 26–30, 2019.

[34] T. Thai, R. Cogranne, and F. Retraint. Statistical model of quantized DCT coefficients: Application in the steganalysis of Jsteg algorithm. *Image Processing, IEEE Transactions on*, 23(5):1–14, May 2014.

[35] T. H. Thai, R. Cogranne, and F. Retraint. Optimal detection of OutGuess using an accurate model of DCT coefficients. In *Sixth IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[36] Thanh Hai Thai, R. Cogranne, and F. Retraint. Camera model identification based on the heteroscedastic noise model. *Image Processing, IEEE Transactions on*, 23(1):250–263, Jan 2014.

[37] J. Zhang, I. J. Cox, and G. Doerr. Steganalysis for LSB matching in images with high-frequency noise. In *IEEE 9th Workshop on Multimedia Signal Processing*, pages 385–388, 2007.

[38] T. Zhang and X. Ping. A fast and effective steganalytic technique against Jsteg-like algorithms. In *Proceedings of the ACM Symposium on Applied Computing*, pages 307–311, Melbourne, FL, March 9–12, 2003.