Detection of Hiding in the Least Significant Bit

Onkar Dabeer, *Member, IEEE*, Kenneth Sullivan, Upamanyu Madhow, *Senior Member, IEEE*, Shivakumar Chandrasekaran, and B. S. Manjunath, *Senior Member, IEEE*

***Abstract*—In this paper, we apply the theory of hypothesis testing to the *steganalysis*, or detection of hidden data, in the least significant bit (LSB) of a host image. The hiding rate (if data is hidden) and host probability mass function (PMF) are unknown. Our main results are as follows.**

- **a) Two types of tests are derived: a universal (over choices of host PMF) method that has certain asymptotic optimality properties and methods that are based on knowledge or estimation of the host PMF and, hence, an appropriate likelihood ratio (LR).**
- **b) For a known host PMF, it is shown that the composite hypothesis testing problem corresponding to an unknown hiding rate reduces to a worst-case simple hypothesis testing problem.**
- **c) Using the results for a known host PMF, practical tests based on the estimation of the host PMF are obtained. These are shown to be superior to the state of the art in terms of receiver operating characteristics as well as self-calibration across different host images. Estimators for the hiding rate are also developed.**

***Index Terms*—Approximate log-liklihood ratio test, hypothesis testing, LSB hiding, steganalysis, universal asymptotic optimality.**

I. Introduction

A. Background and Motivation

DRIVEN by applications such as watermarking and document authentication, there has been a spurt of activity in the area of data hiding in multimedia objects such as image, video, and audio (see, for example, [1]–[6], and the references therein). Unfortunately, applications such as steganography (that is, hidden communication) also have the potential of being misused. Many steganography tools are/were also available in the public domain (see [7] and [8]), and many are easy to create. Naturally, there is an interest in knowing if such hiding can be reliably detected. This detection problem is referred to as steganalysis and is the focus of this paper.

Data hiding is used to convey information by making imperceptible changes to a host object, which can be deciphered by a receiver that knows the hiding scheme and the specific parameters used by the encoder. While research in data hiding is well advanced, steganalysis is still in its infancy. The main reason for this is that in its full generality, steganalysis is an ill-posed problem: The original host in which information is embedded is unknown, the rate of hiding (if data is hidden) is not known, and the number of steganography schemes is large. Even a convincing statistical characterization of a "natural" image (i.e., one without hidden data) is not available. Despite the intrinsic difficulty of the problem of steganalysis, its importance has led to a number of attempts at developing steganalysis tools. These attempts have focused mainly on the detection of the simple yet popular technique of hiding in the least significant bit (LSB) of the host, either in the pixel or transform domain, or its variants such as Outguess [9]. See [10] for a survey of the few steganalysis methods available in the open literature. These methods can be roughly divided into two categories.

1) Intuition regarding the characteristics of a natural image is employed to develop statistics that can discriminate between images with and without hidden data ([7] and [10] fall into this category). Aside from the issue of whether the statistics employed by these methods are the "best" ones, the important question of how to calibrate these methods by choosing parameters such as decision thresholds (e.g., to guarantee a certain probability of false alarm) is difficult to answer within a purely intuitive framework. Indeed, the "right" parameter choice for a given method may often depend on the data itself.
2) Standard supervised learning methods are employed, with intuition regarding the characteristics of a natural image and the disturbance induced by hiding guiding the selection of the feature set. Such a scheme is exemplified in [11] and is perhaps the first published attempt at employing learning for steganalysis. As in all applications of supervised learning, the difficult problem here is to choose an appropriate feature set.

Prior work on mathematical derivation and analysis of steganalysis for LSB hiding includes [12] and [13]. In [13], reasonable properties for the values taken by pairs of pixels in natural images are postulated, and the embedding rate is estimated. The authors of [13] also employ their framework to explain the intuition behind the RS scheme in [10]. The key feature of the RS scheme in [10] and [13] is the exploitation of spatial continuity (or memory) in the images.

We consider an alternative approach to steganalysis, using the classical tools of hypothesis testing [14], [15]. If good statistical models for image characteristics and the hiding process are available, then, in principle, this approach can have the following benefits.

a) An optimal decision statistic can be obtained in the form of a likelihood ratio test (LRT).
b) The performance of an optimal decision rule, even in an idealized setting, can provide benchmarks that indicate

Manuscript received July 25, 2003; revised February 19, 2004. This work was supported by the Office of Naval Research under Grant N00014-01-1-0380. The associate editor coordinating the review of this paper and approving it for publication was Prof. Pierre Moulin.

O. Dabeer is with the Qualcomm Inc., San Diego, CA 92121 USA (e-mail: onkar@ieee.org).

K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

Digital Object Identifier 10.1109/TSP.2004.833869

the intrinsic difficulty of the steganalysis problem (e.g., if it is impossible to achieve a probability of error better than 30% for an optimal rule in an idealized setting, then we cannot expect the performance of suboptimal rules in practical settings to be better).

c) The hypothesis testing approach allows for a degree of self-calibration. For example, if we want to minimize a convex combination of the false alarm and miss probabilities, then Bayesian hypothesis testing can be employed to set a threshold, *independent of the data*, against which we can compare the optimal decision statistic.

As we will see, however, even for the simple setting of LSB hiding, the application of hypothesis testing is not straightforward since the original host statistics and the rate of hiding (if any data is hidden) are unknown. To fit within the standard hypothesis testing framework, we have to constrain our design. We employ the histogram of the host data as our starting point. This is optimal only if the host coefficients are independent and identically distributed (i.i.d.), which does not hold for natural images in either the pixel or the transform domain. Despite this suboptimal choice, we are able to get some of the benefits a)-c) mentioned above. The decision rules obtained outperform the state of the art in histogram-based steganalysis, represented by Stegdetect [7], and can be calibrated to work well over a large class of images. In addition, we also consider the idealized model of an i.i.d. host with a known density to motivate the structure of the decision rule when the hiding rate is unknown and to obtain insight into the achievable limits of steganalysis for LSB hiding. As shown in our numerical results, however, there is a price to be paid for not using the memory inherent in natural images: Our histogram-based scheme does not perform as well as the RS scheme [10], which exploits spatial continuity in the image. Nevertheless, our success in outperforming histogram-based schemes motivates further research into systematic application of detection-theoretic techniques to exploit spatial memory.

B. Outline of Paper

In Section II, we describe a common statistical model for LSB hiding and the host. We assume the host symbols to be i.i.d., although most of our results extend to block i.i.d. (that is M-dependent data) host. The hiding scheme is also assumed to be memoryless. Since the host probability mass function (PMF) is not known, we can take two approaches: Develop hypothesis tests that are optimal uniformly over the possible host PMFs, or develop tests based on perfect knowledge of host PMF and, in practice, use estimates in place of the true PMF. We explore both of these possibilities in this paper. In Section III-A, we cast the steganalysis problem in Hoeffding's framework [16] and develop a scheme, which does not assume any knowledge about the host PMF and the hiding rate. This detector has some asymptotic optimality properties uniformly over the unknown host PMF, and in this sense, it may be termed universal. This formulation is in fact valid for any memoryless hiding scheme, although in this paper, we analyze the detector (see Proposition 1) for the case of LSB hiding only. While this formulation of steganalysis is widely applicable, it ignores the fact that in many cases, we have good models (or we can get good estimates) of the host PMF. With this in mind, in Section III-B, we consider a composite hypothesis testing problem associated with steganalysis when the host PMF is known but the hiding rate is unknown; practical tests based on this formulation and estimates of the host PMF are developed in Section IV. In Proposition 2 of Section III-B, we prove that the composite hypothesis testing problem is solved by a worst-case simple hypothesis testing problem. We use this fact to develop practical tests in Section IV. In Section III-C, we also provide a performance analysis for a class of detection schemes, which is used in Section III-D to compare various detection methods. A number of practical tests and their performance on a database of 4000 digital ortho quarter quad (DOQQ) images are reported in Section IV. In this section, we also develop estimators of the rate of hiding and demonstrate performance improvements over Stegdetect [7]. The conclusions are given in Section V, and all the theoretical results are proved in Appendixes A and B.

Notation: We use capital letters to denote random variables and corresponding lowercase letters to denote their realizations. We use bold letters to denote vectors and matrices. Most of the data we consider takes values in the finite set $\mathcal{A} := \{0, 1, \ldots, 255\}$. For a random variable X taking values in $\mathcal{A}$, we denote the PMF by the 256-dimensional column vector $\boldsymbol{p}^{(X)} = \left[p_0^{(X)}, \ldots, p_{255}^{(X)}\right]^t$, where the superscript t denotes transpose. The set of all the PMFs on $\mathcal{A}$ is denoted by

$$\mathcal{P} := \Bigg\{ \boldsymbol{p} = [p_0 p_1 \cdots p_{255}]^t \in \mathbb{R}^{256} \text{ such that } p_i \geq 0 \quad i = 0, 1, \ldots, 255, \text{ and } \sum_{i=0}^{255} p_i = 1 \Bigg\}.$$

By $Q(t)$, we denote the complementary Gaussian function

$$Q(t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) du.$$

II. Statistical Model for LSB Hiding

In this section, we provide a probabilistic description of the host and the LSB hiding mechanism, which is central to the study of statistical steganalysis tools. As a first step, we consider the case of independent and identically distributed (i.i.d.) data samples. This model is commonly used in steganography [6]. For steganography schemes that hide data in the DCT or wavelet domain (see, for example, [2]), this is a good model as these transforms are known to significantly decorrelate the data. Since the host samples are assumed to be i.i.d., without loss of generality, we assume the data to be one-dimensional. Suppose the i.i.d. host is $\{H_n\}_{n=1}^N$, where the intensity values H_n are represented by 8 bits, that is, $H_n \in \mathcal{A} = \{0, 1, \ldots, 255\}$. We wish to model the situation where data is being hidden in the LSB of the host at a rate R bits per host sample. We assume that over different data sets, the hider changes the host samples where data is hidden to make the task of the detector tougher. Thus, for the detector, the locations of the hidden bits appear

random. We capture this situation by assuming that the hidden data $\{D_n\}_{n=1}^N$ is i.i.d., and

$$P(D_n = 0) = \frac{R}{2}, \quad P(D_n = 1) = \frac{R}{2}$$
$$P(D_n = \text{NULL}) = (1 - R), \quad 0 < R \leq 1.$$

The hider does not hide in host sample H_n if $D_n = \text{NULL}$; otherwise, the hider replaces the LSB of H_n with D_n. With this model for rate R LSB hiding, if the PMF of H_n is $\boldsymbol{p}^{(H)}$, then the PMF $\boldsymbol{p}_R^{(H)}$ of the data after LSB hiding at rate R is given by

$$p_{R,2l}^{(H)} = \left(1 - \frac{R}{2}\right) p_{2l}^{(H)} + \frac{R}{2} p_{2l+1}^{(H)}$$
$$p_{R,2l+1}^{(H)} = \frac{R}{2} p_{2l}^{(H)} + \left(1 - \frac{R}{2}\right) p_{2l+1}^{(H)}$$
$$l = 0, 1, \ldots, 127. \tag{1}$$

In vector notation, we write $\boldsymbol{p}_R^{(H)} = \boldsymbol{Q}_R \boldsymbol{p}^{(H)}$, where $\boldsymbol{Q}_R$ is a 256×256 matrix corresponding to the above linear operation.

The above statistical model can be easily extended to take higher order dependence into consideration. Consider, for example, the joint PMF of neighboring pixels. If we denote this by the 256×256 matrix $\boldsymbol{P}$, then upon i.i.d. LSB hiding with rate R as described above, the joint PMF is $\boldsymbol{P}_R = \boldsymbol{Q}_R \boldsymbol{P} \boldsymbol{Q}_R$. Clearly, this extends to any arbitrary order of dependence. In this paper, however, we only consider the case of i.i.d. observations.

III. Hypothesis Testing Formulation of Steganalysis

The theory of hypothesis testing (see, for example, [15]), which has been successfully applied in many areas such as communications and signal processing, also provides a natural framework for steganalysis. In this approach, the observed data (say, an image) is viewed as a realization of a random process. A random process is completely characterized by its probability law, and therefore, the two hypotheses (presence or absence of hidden data) can be distinguished by estimating the probability law of the observed data. An advantage of this approach is that it enables us to study the limits of steganalysis. In this section, for the i.i.d. host and i.i.d. LSB hiding described in Section II, we study two hypothesis testing formulations of steganalysis.

1) In Section III-A, we study asymptotically (as number of data samples $N \to \infty$) optimal hypothesis tests without any knowledge about the host PMF. Our goal here is to set up a universal framework for steganalysis, which is applicable for a number of hiding schemes, and exemplify it with LSB hiding.
2) Another engineering approach is to first derive optimal tests when the host PMF is known and in practice use an estimate in place of the true host PMF. With this in mind, we study the composite hypothesis testing formulation of steganalysis when the host PMF is known in Section III-B; the lessons learned are used in Section IV to design practical steganalysis schemes, which do not assume knowledge of the host PMF.

Further analysis and comparison of these detection scheme are given in the remainder of the section.

A. Asymptotically Optimal Steganalysis with Unknown Host PMF

Suppose the observed data $\{X_n\}_{n=1}^N$ is i.i.d. with PMF $\boldsymbol{p}^{(X)}$ and takes values in the alphabet $\mathcal{A} = \{0, 1, \ldots, 255\}$. We wish to decide between two possibilities: No data is hidden (hypothesis $\mathcal{H}_0$) or data is hidden at some rate R, where $R_0 \leq R \leq R_1$ (hypothesis $\mathcal{H}_1$). The parameters $0 < R_0 \leq R_1 \leq 1$ are specified by the user. In this section, we consider the case where the host PMF $\boldsymbol{p}^{(H)}$ is unknown. Recall that $\mathcal{P}$ is the set of PMFs on $\mathcal{A}$. Let $\mathcal{P}_R := \boldsymbol{Q}_R \mathcal{P}$ be the image of $\mathcal{P}$ under the linear map $\boldsymbol{Q}_R$. Since LSB hiding at rate R results in a transformation of the host PMF by the linear map $\boldsymbol{Q}_R$, $\mathcal{P}_R$ is the set of possible PMFs for the data after hiding at rate R. It is easy to see that $\mathcal{P}_R \subset \mathcal{P}_{R'}$ for $R > R'$ [see (12) in Appendix A]. Thus, the hypothesis that data is hidden at a rate R, where $R_0 \leq R \leq R_1$ is the composite hypothesis

$$\mathcal{H}_1 : \boldsymbol{p}^{(X)} \in \bigcup_{R_0 \leq R \leq R_1} \mathcal{P}_R = \mathcal{P}_{R_0}.$$

The hypothesis that data is not hidden is

$$\mathcal{H}_0 : \boldsymbol{p}^{(X)} = \boldsymbol{p}^{(H)}, \text{ where } \boldsymbol{p}^{(H)} \in \mathcal{P} \setminus \mathcal{P}_{R_0} \text{ is unknown.}$$

We next derive detectors that do not depend on $\boldsymbol{p}^{(H)}$.

A detector δ_N is characterized by the acceptance region $A_N \in \mathcal{A}^N$ of hypothesis $\mathcal{H}_1$:

$$\delta_N(x_1, \ldots, x_N) = \mathcal{H}_1, \text{ if } (x_1, \ldots, x_N) \in A_N$$
$$= \mathcal{H}_0, \text{ otherwise}$$

where $\{x_n\}_{n=1}^N$ is a realization of $\{X_n\}_{n=1}^N$. The performance of the detector is given by the two error probabilities:

$$P_1(\delta_N) := P(\text{Miss}) = P(\delta_N(X_1, \ldots, X_N) = \mathcal{H}_0 | \mathcal{H}_1)$$
$$= \sup_{\boldsymbol{p}^{(X)} \in \mathcal{P}_{R_0}} P\left((X_1, \ldots, X_N) \in \mathcal{A}^N \setminus A_N\right)$$

and

$$P_2(\delta_N) := P(\text{False alarm})$$
$$= P(\delta_N(X_1, \ldots, X_N) = \mathcal{H}_1 | \mathcal{H}_0)$$
$$= P((X_1, \ldots, X_N) \in A_N)$$

where the PMF of X_1 under $\mathcal{H}_0$ is some unknown $\boldsymbol{p}^{(H)} \in \mathcal{P} \setminus \mathcal{P}_{R_0}$. In the Neyman–Pearson formulation of the optimal detection problem, for given $\alpha > 0$, we minimize $P_2(\delta_N)$ over detectors δ_N that satisfy $P_1(\delta_N) \leq \alpha$. Unfortunately, the solution to this problem in general depends on the host PMF $\boldsymbol{p}^{(H)}$, which is not known in our case. However, if instead we cast the problem in terms of the rate of decay of the error probabilities (as $N \to \infty$), then we do get an asymptotically optimal detector that works for all $\boldsymbol{p}^{(H)} \in \mathcal{P} \setminus \mathcal{P}_{R_0}$. This formulation is due to Hoeffding [16], which we describe next. (A generalization of Hoeffding's formulation was recently proposed in [17] and [18], but we do not consider it here.) We are seeking detectors that maximize the rate of exponential decay of P_2 subject to a minimum guarantee on the exponential rate of decay of P_1:

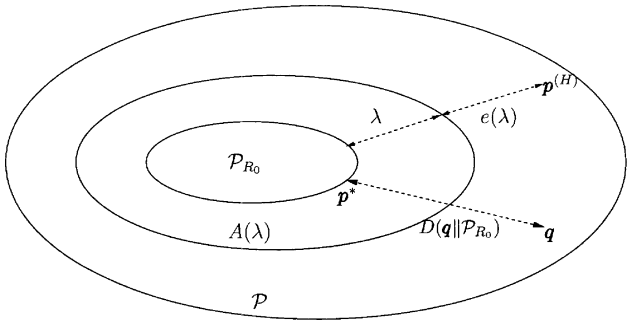

Fig. 1. Schematic of the test (2).

Given $\lambda > 0$, amongst all sequences of detectors $\{\delta_N\}$ satisfying

$$\liminf_{N\to\infty} -\frac{1}{N}\log(P_1(\delta_N)) \geq \lambda$$

we seek a sequence which maximizes

$$\liminf_{N\to\infty} -\frac{1}{N}\log(P_2(\delta_N))$$

for *all* $\boldsymbol{p}^{(H)} \in \mathcal{P}\setminus\mathcal{P}_{R_0}$. Let $\boldsymbol{q}$ be the empirical PMF (normalized histogram) of the observed data and $D(\boldsymbol{q}\|\boldsymbol{p})$ denote the Kullback–Leibler (K-L) divergence between the PMFs $\boldsymbol{q}$ and $\boldsymbol{p}$:

$$D(\boldsymbol{q}\|\boldsymbol{p}) = \sum_k q_k \log\left(\frac{q_k}{p_k}\right).$$

From [16, Th. 3.1], we know that an optimal (in the above asymptotic sense) test declares data to be hidden whenever

$$\boxed{D(\boldsymbol{q}\|\mathcal{P}_{R_0}) := \min_{\boldsymbol{p}\in\mathcal{P}_{R_0}} D(\boldsymbol{q}\|\boldsymbol{p}) \leq \lambda.} \tag{2}$$

Thus, the decision statistic simply computes the K-L distance of the empirical PMF $\boldsymbol{q}$ from the set of PMFs $\mathcal{P}_{R_0}$ corresponding to the feasible PMFs after hiding at rate $R \geq R_0$ (see Fig. 1). This detector can be termed *universal* since it optimizes the error exponent corresponding to P_2 for all $\boldsymbol{p}^{(H)} \in \mathcal{P}\setminus\mathcal{P}_{R_0}$. To implement the optimal test, we need to solve the optimization problem in (2). We note that $\mathcal{P}_{R_0}$ is convex and closed, and $D(\boldsymbol{q}\|\boldsymbol{p})$ is convex in $\boldsymbol{p}$. Therefore, a unique solution exists and is given by the following proposition.

Proposition 1: Let $\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}\in\mathcal{P}_{R_0}} D(\boldsymbol{q}\|\boldsymbol{p})$. Then, for $k = 0, 1, \ldots, 127$

a) if $(R_0/(2-R_0)) \leq (q_{2k+1}/q_{2k}) \leq ((2-R_0)/R_0)$, then

$$p^*_{2k} = q_{2k}, \quad p^*_{2k+1} = q_{2k+1}$$

b) if $(q_{2k+1}/q_{2k}) < (R_0/(2-R_0))$, then

$$p^*_{2k} = \left(1-\frac{R_0}{2}\right)(q_{2k}+q_{2k+1})$$
$$p^*_{2k+1} = \frac{R_0}{2}(q_{2k}+q_{2k+1})$$

c) if $(q_{2k+1}/q_{2k}) > ((2-R_0)/R_0)$, then

$$p^*_{2k} = \frac{R_0}{2}(q_{2k}+q_{2k+1})$$
$$p^*_{2k+1} = \left(1-\frac{R_0}{2}\right)(q_{2k}+q_{2k+1})$$

where we interpret 0/0 to be 1.

The proof is given in Appendix A.

The above detection problem is well-posed only if $\boldsymbol{p}^{(H)}$ is not in $\mathcal{P}_{R_0}$. It is easy to see that $\boldsymbol{p}^{(H)} \in \mathcal{P}_{R_0}$ if and only if

$$\frac{R_0}{2-R_0} \leq r_k^{(H)} \leq \frac{2-R_0}{R_0}$$

where

$$r_k^{(H)} := \frac{p_{2k+1}^{(H)}}{p_{2k}^{(H)}}, \quad k = 0, 1, \ldots, 127.$$

Equivalently, the test (2) makes sense if and only if

$$R_0 > \frac{2r}{1+r}, \quad r := \min\left\{\min_k r_k^{(H)}, \frac{1}{\max_k r_k^{(H)}}\right\}.$$

Thus, for a Binomial(255, 1/2) host, the above test is meaningful for $R_0 > 1/128$ only. This is the hit we have to take for not knowing anything about the host PMF.

Even though the test (2) does not depend on the knowledge of the host PMF, its performance depends on the host PMF. Consider the error exponent of the probability of false alarm, which is given by [16, Th. 3.1]

$$e(\lambda) = \liminf_{N\to\infty} -\frac{1}{N}\log(P_2(\delta_N)) = \min_{\boldsymbol{u}\in A(\lambda)} D(\boldsymbol{u}\|\boldsymbol{p}^{(H)})$$

where

$$A(\lambda) := \{\boldsymbol{u} \in \mathcal{P} : D(\boldsymbol{u}\|\mathcal{P}_{R_0}) \leq \lambda\}.$$

Thus, the further away $\boldsymbol{p}^{(H)}$ is from $A(\lambda)$, the better the performance will be. To get an asymptotically reliable test (for which the probability of errors decays to zero as $N \to \infty$), we want both λ and $e(\lambda)$ to be positive (see Fig. 1). A necessary and sufficient condition for $e(\lambda)$ to be positive is that $\boldsymbol{p}^{(H)} \notin A(\lambda)$. As we increase the rate of decay λ of the probability of a miss, $A(\lambda)$ grows to fill up $\mathcal{P}$, and hence, the rate of decay $e(\lambda)$ of the probability of false alarm decreases. Therefore, the test (2) is reliable only for $\lambda \in \left(0, \lambda_*(\boldsymbol{p}^{(H)})\right)$ for some $\lambda_*(\boldsymbol{p}^{(H)}) > 0$.

We note that the above formulation of steganalysis is applicable to any memoryless hiding scheme—only the definition of the set $\mathcal{P}_{R_0}$ in the optimization problem (2) changes as per the hiding scheme. If the hiding scheme is such that $\mathcal{P}_{R_0}$ is convex, then (2) is a convex optimization problem. While Proposition 1 gives a closed-form solution to the optimization problem in the case of LSB hiding, in general, we have to use numerical techniques to solve (2) [19].

Even though the approach we have taken so far is promising, it has a limitation. Note that in this approach, we always compete $\boldsymbol{p}^{(H)}$ against *all* of $\mathcal{P}_{R_0}$. However, for steganalysis, it seems

natural that $\boldsymbol{p}^{(H)}$ should compete only with $\boldsymbol{p}_{R_0}^{(H)}$. This, however, leads to tests based on $\boldsymbol{p}^{(H)}$. Although $\boldsymbol{p}^{(H)}$ is not known in practice, in specific situations, we may have access to good models/estimates for it. With this in mind, we next consider the other extreme case when the host PMF is known exactly (but the hiding rate is still unknown).

B. Optimal Composite Hypothesis Testing with Known Host PMF

In this section, we consider the case when the host PMF is known to the detector. The analysis of optimal schemes in this section is used in Section IV to design practical steganalysis schemes that do not assume knowledge of the host PMF.

We use H_R to represent the hypothesis that data is hidden at rate R. The steganalysis problem in this notation is to distinguish between H_0, and

$$K(R_0, R_1) := \{H_R : R_0 \leq R \leq R_1\}.$$

The hypothesis that data is hidden is thus *composite*, whereas the hypothesis that nothing is hidden is *simple*. In the absence of an *a priori* distribution on R when data is hidden, we can take the following two approaches.

a) We can use the generalized LRT (GLRT) ([15]), which declares data to be hidden whenever

$$\boxed{\min_{R_0 \leq R \leq R_1} D\left(\boldsymbol{q} \| \boldsymbol{p}_R^{(H)}\right) - D\left(\boldsymbol{q} \| \boldsymbol{p}^{(H)}\right) \leq T} \tag{3}$$

where T is the threshold to be chosen, and $\boldsymbol{q}$ is the empirical PMF of the observed data. The GLRT is known to have some asymptotically optimal properties [20].

b) We can derive optimal detectors in a Neyman–Pearson framework: For given $\alpha > 0$, minimize

$$P(\text{Miss}) = \sup_{R_0 \leq R \leq R_1} P(\delta(X_1, \ldots, X_N) = H_0 | H_R)$$

over detectors δ that satisfy

$$\begin{aligned} P(\text{False Alarm}) &= P(\delta(X_1, \ldots, X_N) = K(R_0, R_1) | H_0) \\ &\leq \alpha. \end{aligned}$$

In general, the minimization in (3) has to be carried out numerically. In contrast, approach b) leads to simple tests in our case, and we pursue it in detail below. Practical tests based on both these approaches are compared in Section IV.

From [14, Th. 7, pp. 91], we know the following for framework b) above.

1) An optimal detector exists.
2) Consider an apriori probability distribution π on $[R_0, R_1]$. With this *a priori* distribution, the optimal detector is the well-known LRT. If for a fixed $P(\text{False alarm})$ level α, this test results in a lower $P(\text{Miss})$ for any other distribution π' on $[R_0, R_1]$, then π is said to be the *least favorable* distribution. If a least favorable distribution exists, then [14, Th. 7 pp. 91] tells us that the corresponding LRT is the optimal detector.

Therefore, one way to find the optimal detector is to find the least favorable distribution. Intuitively, for steganalysis, the worst-case corresponds to the smallest hiding rate. The following proposition shows that this intuition is accurate for sufficiently large data lengths N and sufficiently small hiding rates. A detailed discussion is given after the statement of the proposition.

Proposition 2: Suppose

$$p_l^{(H)} > 0, \quad l = 0, \ldots, 255$$

and

$$r_k^{(H)} = \frac{p_{2k+1}^{(H)}}{p_{2k}^{(H)}} \neq 1, \text{ for some } k = 0, 1, \ldots, 127.$$

Consider the composite hypothesis testing problem for distinguishing between H_0 and $K(R_0, R_1)$. We restrict our attention to detectors that operate in the region $P(\text{Miss}) \leq 0.5$, $P(\text{False Alarm}) \leq 0.5$. Then, there exists $N_0, R_* > 0$ such that for $N \geq N_0, R_1 \leq R_*$, the unique least favorable distribution is a unit mass at R_0. Therefore, if $\boldsymbol{q}$ denotes the empirical PMF of the observed data, the optimal detector for $N \geq N_0$ and $R_1 < R_*$ is the corresponding LRT, which accepts $K(R_0, R_1)$ if

$$\begin{aligned} S_{\text{LLRT}}(\boldsymbol{q}) &:= D\left(\boldsymbol{q} \| \boldsymbol{p}_{R_0}^{(H)}\right) - D(\boldsymbol{q} \| \boldsymbol{p}^{(H)}) \\ &= \sum_{k=0}^{255} q_k \log\left(\frac{p_k^{(H)}}{p_{R_0,k}^{(H)}}\right) \leq T(\alpha) \end{aligned} \tag{4}$$

where $T(\alpha)$ is a threshold chosen to obtain $P(\text{False Alarm}) = \alpha$.

The proof is given in Appendix B.

The main conclusion of the above result is that the composite hypothesis testing problem associated with steganalysis can be replaced by the simple hypothesis testing problem: test H_0 versus H_{R_0}. Even though we have established the result only for sufficiently small rates of hiding, in simulations, we have found the results to be meaningful for higher rates as well.

By Stein's lemma [21, Th. 12.8.1], we know that for small α, the error exponent of $P(\text{Miss})$ (as $N \to \infty$) is (approximately) given by $D\left(\boldsymbol{p}^{(H)} \| \boldsymbol{p}_{R_0}^{(H)}\right)$. Using the definition of $\boldsymbol{p}_{R_0}^{(H)}$, it is easy to see that $D\left(\boldsymbol{p}^{(H)} \| \boldsymbol{p}_{R_0}^{(H)}\right)$ is of the order of R_0^2. This shows that as R_0 decreases, the performance of even the optimal test degrades rapidly: For a two-fold decrease in R_0, we need a four-fold increase in data size to maintain the same performance. **Thus, for low hiding rates, the steganalysis problem is inherently difficult**.

C. Asymptotic Performance of Hypothesis Tests

In Sections III-A and B, we used error exponents to make qualitative remarks about the performance of the tests. In this section, our goal is to provide approximate expressions for $P(\text{Miss})$ and $P(\text{False Alarm})$ for large N and for a wide class of decision statistics. Our motivation is two-fold.

- Evaluating the error probabilities by simulation is time consuming. On the other hand, the approximate expressions (8) and (9) can be computed with hardly any effort

and provide a quick way of comparing different schemes for a fixed host PMF.
- Our proof of Proposition 2 is based on these approximations.

We restrict our attention here to distinguishing between H_0 and H_R. We note that for i.i.d. data, the empirical PMF $\boldsymbol{q}$ is a sufficient statistic for testing hypothesis H_0 versus H_R. Therefore, we are interested in tests that accept H_R if $S(\boldsymbol{q}) < T$. We note that $\boldsymbol{q} = N^{-1}\sum_{n=1}^{N} \boldsymbol{Z}_n$, where $\boldsymbol{Z}_n$ is a 256-dimensional column vector whose kth entry is 1 if the data $X_n = k$ and is zero otherwise. The $\{\boldsymbol{Z}_n\}$ are i.i.d., and straightforward computations show that $\mathbb{E}[\boldsymbol{Z}_1] = \boldsymbol{p}_R^{(H)}$, and the covariance matrix of $\boldsymbol{Z}_1$ is $\boldsymbol{\Sigma}_R := \text{diag}\left(\boldsymbol{p}_R^{(H)}\right) - \boldsymbol{p}_R^{(H)}\left(\boldsymbol{p}_R^{(H)}\right)^t$ under hypothesis H_R. An application of the law of large numbers then gives us

$$\lim_{N\to\infty} \boldsymbol{q} = \boldsymbol{p}_R^{(H)} \text{ almost surely under } H_R. \tag{5}$$

In addition, applying the central limit theorem, we get that under H_R

$$\sqrt{N}\left(\boldsymbol{q} - \boldsymbol{p}_R^{(H)}\right) \Longrightarrow \mathcal{N}(0, \boldsymbol{\Sigma}_R). \tag{6}$$

Here, "$\Longrightarrow$" denotes convergence in distribution, and $\mathcal{N}(0, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with zero mean and correlation matrix $\boldsymbol{\Sigma}$. Now, suppose $S : \mathbb{R}^{256} \to \mathbb{R}$ is differentiable at $\boldsymbol{p}_R^{(H)}$. Then, we know from an old result of Mann and Wald (see [22, Prop. 6.4.3, pp. 211]) that

$$\begin{aligned}\sqrt{N}(S(\boldsymbol{q}) - \mu(R)) &\Longrightarrow \mathcal{N}(0, \sigma^2(R))\\ \mu(R) &:= S\left(\boldsymbol{p}_R^{(H)}\right)\\ \sigma^2(R) &:= \boldsymbol{u}_R^t \boldsymbol{\Sigma}_R \boldsymbol{u}_R, \quad \boldsymbol{u}_R := \nabla S\big|_{\boldsymbol{p}_R^{(H)}}.\end{aligned} \tag{7}$$

We are interested in finding expressions for the probabilities of errors for the detector based on $S(\boldsymbol{q})$. Unfortunately, for large N and fixed $R \in (0, 1]$, this goal cannot be met by the above asymptotic normality result; such goals lie in the large deviations regime [23], which is too complicated for our purpose. Fortunately, in practice, we are interested in small R, typically around 0.05. In this case, the two alternative hypotheses are close, and the asymptotic normality result provides good approximations to the error probabilities. (Rigorously, this is established by choosing $R = O(1/\sqrt{N})$ as $N \to \infty$; see our proof of Proposition 2.) Therefore, using the Gaussian approximation for N large and R small, we have

$$\begin{aligned}P(\text{False Alarm}) &= P(S(\boldsymbol{q}) < T \big| H_0)\\ &\approx Q\left(\frac{\sqrt{N}(\mu(0) - T)}{\sigma(0)}\right)\end{aligned} \tag{8}$$

and

$$\begin{aligned}P(\text{Miss}) &= P(S(\boldsymbol{q}) \geq T \big| H_R)\\ &\approx Q\left(\frac{\sqrt{N}(T - \mu(R))}{\sigma(R)}\right)\end{aligned} \tag{9}$$

where $Q(t)$ is the complementary Gaussian function defined previously.

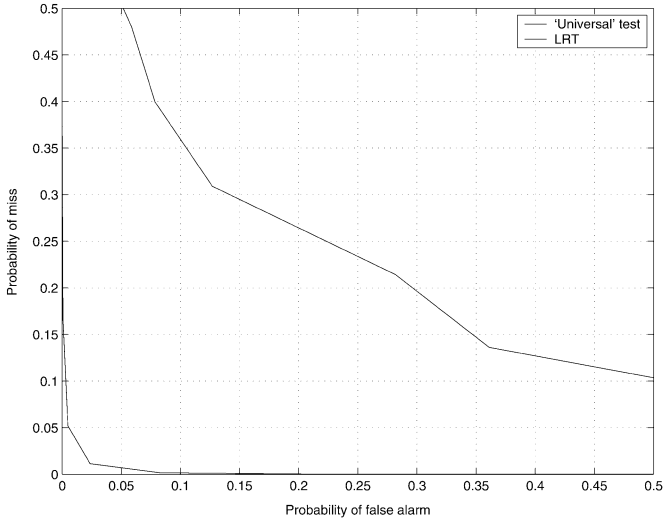

Fig. 2. Comparison between the universal test ($R_0 = 0.5$) and the LRT test for binomial(15,1/2) data of length 1024 and hiding rate $R = 0.5$. Any practical scheme would perform between these two limits.

D. Comparison of Detectors

In this section, we compare three schemes for detecting LSB hiding: the universal scheme (2), the LRT (4), and Stegdetect [7].

In Fig. 2, we compare the test (2) with the LRT (4) when the host is i.i.d. Binomial(15,1/2). We note that these tests represent two extremes: Equation (2) does not assume any knowledge of the host PMF and competes $\boldsymbol{p}^{(H)}$ with all of $\mathcal{P}_R$, whereas (4) assumes exact knowledge of the host PMF and competes $\boldsymbol{p}^{(H)}$ only with $\boldsymbol{p}_R^{(H)}$. Therefore, we expect any practical test to perform between these two tests. Not surprisingly, (2) is worse than (4). In addition, as stated after Proposition 1, in this case, (2) is useful only for hiding rates greater than 1/128. The price for generality is that in specific instances where we may have good models, the performance is far from optimal.

Next, consider the statistic of a test called Stegdetect [7]:

$$S_{\text{stegdetect}}(\boldsymbol{q}) := \sum_{k=0}^{127} \frac{(q_{2k+1} - q_{2k})^2}{q_{2k} + q_{2k+1}}.$$

Strictly speaking, Stegdetect [7] is also tuned to specifics of particular hiding schemes such as Outguess [9]. However, its main element is the above statistic, which does not depend on the host PMF. If $S_{\text{stegdetect}}(\boldsymbol{q})$ is less than a threshold T, then Stegdetect declares data to be hidden; otherwise, no data is hidden. We know that after hiding at $R = 1$

$$p_{R,2k}^{(H)} = p_{R,2k+1}^{(H)} = \frac{p_{2k}^{(H)} + p_{2k+1}^{(H)}}{2}, \quad 0 \leq k \leq 127.$$

The Stegdetect statistic is a measure of closeness of the adjacent bins $\{2k, 2k+1\}$—the smaller the statistic, the closer these bins are, and the higher the chances that data is hidden. We next compare Stegdetect with the optimal LRT (4). By choosing the host PMF to be Binomial(255, θ), $\theta \in (0, 1)$ and using (8) and (9) for these two tests, we have observed that Stegdetect performs very close to the optimal LRT; we have

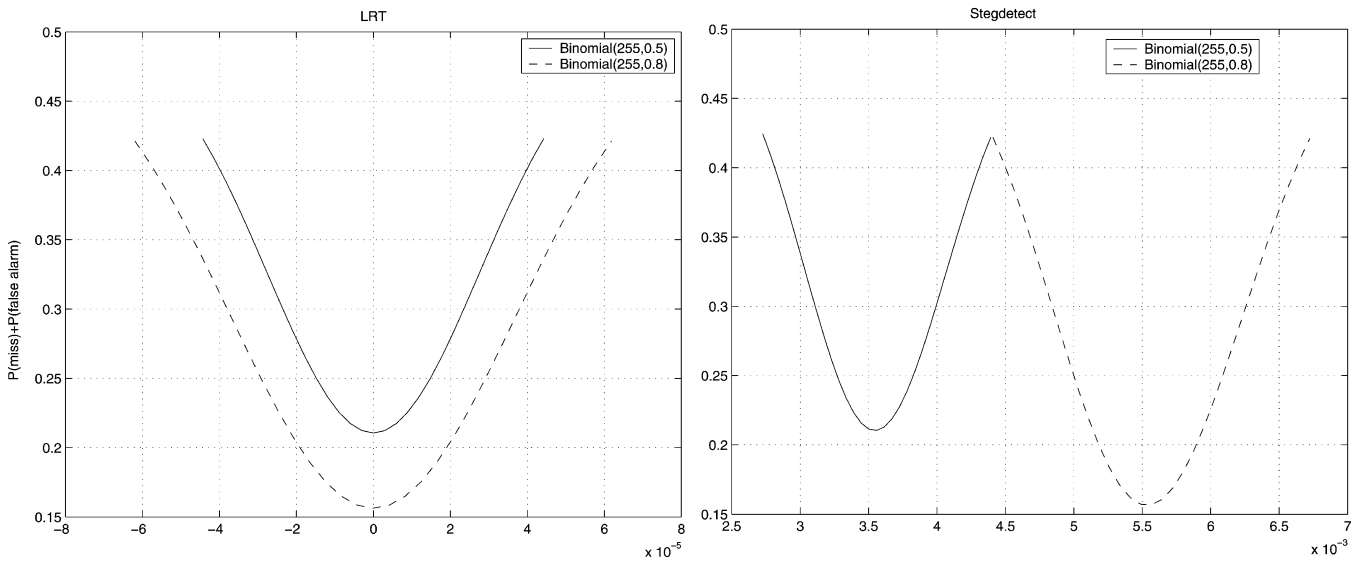

Fig. 3. Stegdetect statistic does not depend on host PMF, but the threshold is sensitive to the host PMF, unlike LRT.

observed the same result for host PMFs randomly chosen to be a mixture of binomials. (However, we do not have a theoretical proof that this is the case whenever the PMF is a mixture of binomials.) While the Stegdetect statistic does not depend on the host PMF and, for a given host PMF, its performance appears to be close to the LRT, the choice of threshold required to guarantee a target performance depends on the host PMF, as shown by the following example. Suppose our target is to minimize $P(\text{Miss}) + P(\text{False Alarm})$. In Fig. 3, we plot the sum of the $P(\text{Miss}) + P(\text{False Alarm})$ as a function of the threshold for the LRT and Stegdetect for the Binomial(255,0.5) and Binomial(255,0.8) host PMFs. For the LRT, the threshold $T = 0$ minimizes $P(\text{Miss}) + P(\text{False Alarm})$ for *any* host PMF, whereas for Stegdetect, the minimizing threshold T depends on the host PMF.

To summarize, if the host PMF is known, then there is little loss in using the suboptimal Stegdetect. In practice, this means that if we have good models for the host PMF and lookup tables for choosing the threshold (depending on the host PMF), then Stegdetect performs close to the LRT. However, the host PMF usually varies substantially over image databases, and hence, we are more interested in completely data driven tests that attain the target performance—both the statistic and the threshold have to be chosen based on the data to achieve the desired performance. With this in mind, we note two points, which motivate our work in Section IV.

- The LR depends on the host PMF, but the threshold T can be chosen independent of the host PMF. For example, to minimize $P(\text{Miss}) + P(\text{False Alarm})$, we can choose $T = 0$.
- The Stegdetect statistic does not depend on the host PMF, but to obtain a target performance, T has to be chosen depending on the host PMF. Thus, Stegdetect does *not* resolve the problem of not knowing the host PMF—it simply transfers it to the choice of the threshold. This aspect does not appear to have been realized in the literature (see, for example, [7]).

IV. Practical Tests

Our goal in this section is to develop practical hypothesis tests for steganalysis and evaluate their performance for a database of 4000 DOQQ images. (We also briefly state the performance for scanned images and digital camera images.) We note that the images in the database have different histograms—thus, in probabilistic terms, we are evaluating the performance of the hypothesis tests for a nonergodic source. Based on our discussion so far, we can take three directions.

- We can use the universal test (2).
- We can estimate the host PMF and use (4) with this estimate in place of the true PMF.
- We can use Stegdetect.

While the test (2) is applicable in a wider range of scenarios than (4) and Stegdetect, we have found that the latter two approaches give better performance for the specific case of LSB hiding. Hence, we only report the experiments for them. In particular, we exhibit new tests based on the estimation of the LR and demonstrate their superiority over Stegdetect. We also compare with the RS analysis scheme, which, unlike our focus here, uses memory in the image.

A. Estimating LR

We note from Proposition 2 that we only need to develop tests for testing H_0 versus H_R, where R is the smallest rate amongst the possible rates for which the user is testing. A problem with the optimal LRT is that we do not know the host PMF in practice. However, there are two factors that help us to develop good practical tests based on the optimal LRT.

1) The hiding rate in practice is very low, and therefore, we can estimate the host PMF well. We show below that a number of simple estimates of the host PMF based on the assumption that the host PMF is "smooth" work well.
2) For the optimal LRT, the threshold that minimizes $aP(\text{Miss}) + (1 - a)P(\text{False Alarm})$ for $a \in [0, 1]$ does not depend on the host.

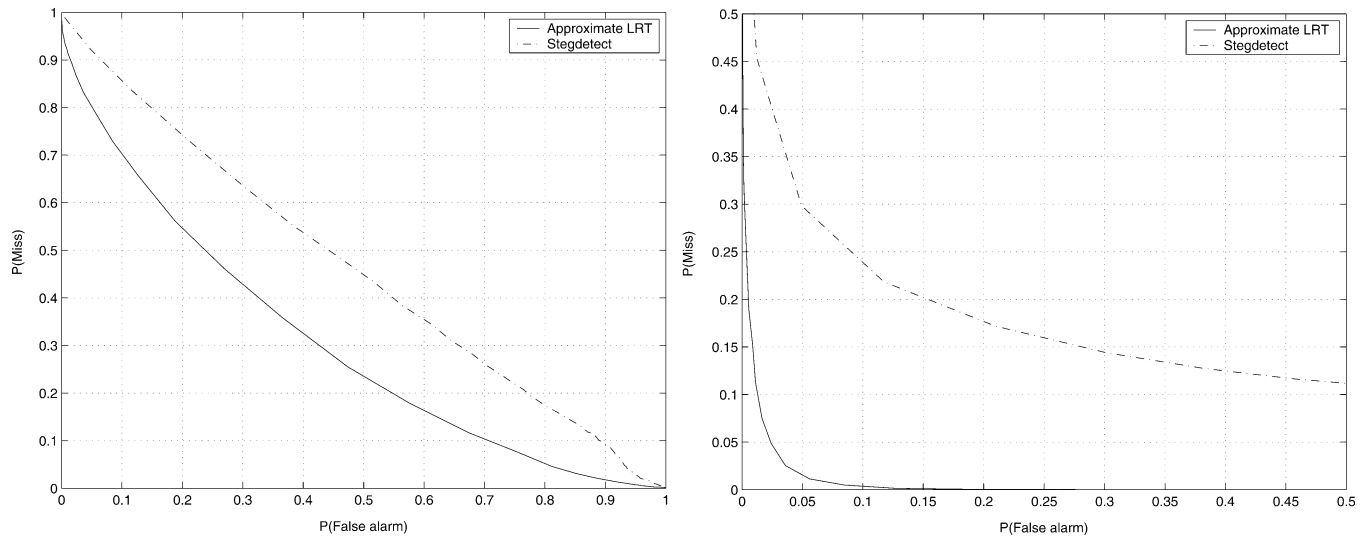

Fig. 4. Approximate LRT with half–half filter estimate versus stegdetect: At high rates (right side; $R = 0.5$) as well as low rates (left side; $R = 0.05$), the approximate LRT is superior. The same hiding rate r was used for all the test images with hidden data.

With the above motivation, we propose to form an estimate $\hat{\boldsymbol{p}}^{(H)}$ of the host PMF $\boldsymbol{p}^{(H)}$ and then form the decision statistic

$$\boxed{S(\boldsymbol{q}) = D\left(\boldsymbol{q}\|\boldsymbol{Q}_R\hat{\boldsymbol{p}}^{(H)}\right) - D\left(\boldsymbol{q}\|\hat{\boldsymbol{p}}^{(H)}\right).} \tag{10}$$

1) For natural images, the PMF is usually low pass. On the other hand, random LSB hiding introduces high-frequency components in the histogram. Hence, one simple estimate $\hat{\boldsymbol{p}}^{(H)}$ is to pass the empirical PMF $\boldsymbol{q}$ through a lowpass two-tap FIR filter with taps (0.5, 0.5), which we refer to as the half–half filter below. We note that normalization is required after the filtering.
2) Another regularity constraint that we can impose on the host PMF is that the local slope is preserved, that is

$$p_{k+3}^{(H)} - p_k^{(H)} = 3\left(p_{k+2}^{(H)} - p_{k+1}^{(H)}\right), \quad k = 0, 4, 8, \ldots, 252.$$

This regularity constraint can be written as $\boldsymbol{A}\boldsymbol{p}^{(H)} = 0$ for a suitable 64×256 matrix. Under this regularity constraint, a natural estimate of $\boldsymbol{p}^{(H)}$ is to project $\boldsymbol{q}$ onto the null space of $\boldsymbol{A}$. We again need normalization and removal of negative components after this filtering.
3) We also propose a nonlinear approach that, unlike the above two approaches, adapts to the underlying host PMF. We note that LSB hiding only affects the eighth bit. Suppose we only consider the "coarse" image with 7-bit pixels corresponding to the seven most significant bits. By appending the remaining eighth bit, we can go from this coarse image to the original image. We assume that the addition of the eighth bit is such that it preserves the shape of the histogram of the coarse 7-bit image. More precisely, we impose the regularity constraint that the host PMF is such that we can obtain it by spline interpolation of the PMF of the coarse image. The estimate $\hat{\boldsymbol{p}}^{(H)}$ corresponding to this regularity constraint is obtained by first subsampling $\boldsymbol{q}$, then interpolating using splines, and then normalizing.

We refer to all these tests as the approximate LRT.

B. Simulation Results

We next state a number of simulation results for 4000 images from a DOQQ image set and discuss them in light of the results of Section III.

Comparison with Stegdetect: In Fig. 4, we compare the approximate LRT based on the half-half filter for estimating p with Stegdetect. For each point on the curve, the threshold has been fixed over the entire database. Clearly, our test significantly outperforms Stegdetect for small as well as high rates; at $R = 0.05$, Stegdetect is as bad as random guessing. Based on the discussion in Section III-D, it appears that for a fixed host PMF, both these tests perform closely. However, for the database of images we have used, the host PMF varies substantially from image to image. Thus, these simulations suggest that Stegdetect is more sensitive to the choice of the threshold than the approximate LRT. This is not surprising since we know that to attain a target performance, the choice of the threshold in the LRT does not depend on the host PMF. For example, by choosing $T = 0$ for the approximate LRT in the case when the hiding rate is 0.05, we found the operating point to be $P(\text{Miss}) = 0.4043$ and $P(\text{False Alarm}) = 0.3219$. From Fig. 4, we can verify that the tangent to the operating curve at this point is of slope approximately 1, as predicted by the theory. To summarize, our test is definitely closer to the goal of obtaining data driven tests than Stegdetect. In particular, with $T = 0$, for small hiding rates typically encountered in practice, we obtain performance that is close to that expected theoretically. We have observed that the story remains unchanged if we hide in the LSB of the DCT coefficients of JPEG compressed images (with quality factor 75).

Performance at small hiding rates: We note from Fig. 4 that for low hiding rates such as $R = 0.05$, the performance is not good. In fact, this is true even in the ideal case when the host is generated i.i.d. from a known PMF, the hiding rate is known, and the optimal LRT is employed. Based on the discussion after Proposition 2, we know that as the hiding rate decreases, the performance of even the optimal LRT degrades

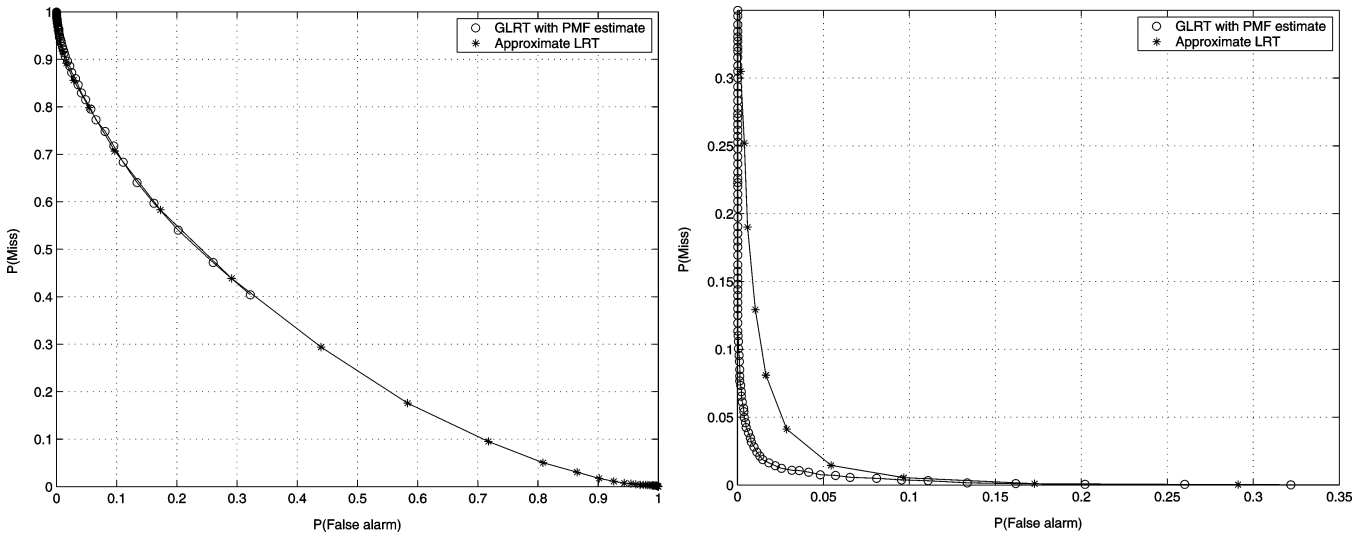

Fig. 5. Difference between the approximate GLRT and our worst case approximate LRT is small, especially at low hiding rates. ($R = 0.05$ case is on left side, whereas the $R = 0.5$ case is on the right side. The same hiding rate was used for all the test images with hidden data.)

rapidly. Evaluating the error exponent $D\left(\boldsymbol{p}^{(H)}\|\boldsymbol{p}_R^{(H)}\right)$ given by Stein's lemma [21, Theorem 12.8.1], for a Binomial(255, 0.5) host PMF, we see that to obtain the same performance as the $R = 0.5$ case, for $R = 0.05$, we need images that are about 99.43 times larger.

Comparison with approximate GLRT: In principle, instead of the simple hypothesis tests as above, we could use the GLRT (3) with the host PMF replaced by its estimate: Data is declared hidden if

$$\boxed{\min_{R_0 \le R \le R_1} D\left(\boldsymbol{q}\|\hat{\boldsymbol{p}}_R^{(H)}\right) - D(\boldsymbol{q}\|\hat{\boldsymbol{p}}^{(H)}) \le T.} \tag{11}$$

The minimization in (11) is carried out by searching over the entire range of R. As shown in Fig. 5, this approximate GLRT performs very close to the (simple) approximate LRT we have developed (which uses R_0 instead of the above minimization), especially at low hiding rates. This is not surprising given Proposition 2, which states that for small hiding rates, the optimal composite hypothesis testing problem considered in Section III-B is solved by the simple hypothesis testing problem. Due to the numerical minimization required in the approximate GLRT, it is also more computationally intensive than the approximate LRT. However, we note that the GLRT also furnishes an estimate of the hiding rate: We can use the argument R that minimizes (11) as an estimate of the actual embedding rate. We find that this works reasonably well in practice (see Fig. 6).

Effect of different PMF estimates: We have compared the approximate LRT based on spline estimates of $\boldsymbol{p}^{(H)}$ and based on the half–half lowpass filter. There is very little difference in performance. We have observed that the local slope-preserving filter is slightly worse.

Comparison with RS analysis: Our focus in this paper is to develop a fundamental understanding of steganalysis by considering schemes that do not exploit host memory. In order to understand the possible scope for improvement by exploiting host memory, in Fig. 7, we compare the approximate LRT with RS analysis [10]. The RS scheme provides an estimate of the unknown rate of hiding. To obtain the performance curve in Fig. 7,

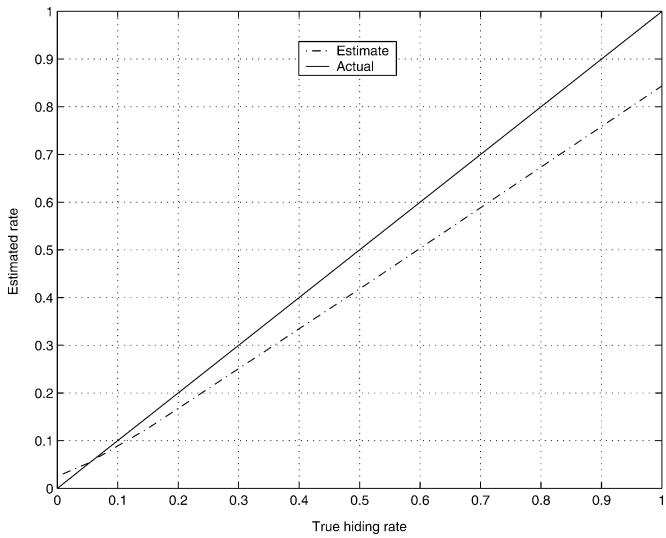

Fig. 6. Minimizing rate in (11) serves as an estimate of the true hiding rate.

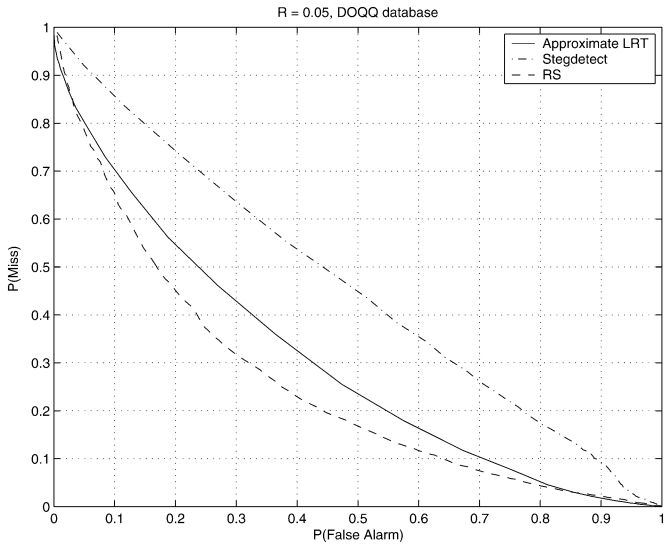

Fig. 7. RS analysis, which uses host memory, is slightly better than our memoryless approximate LRT on the DOQQ database. A hiding rate of 0.05 was used for all the test images with hidden data.

we compare this threshold with different thresholds (fixed over the database at each point on the performance curve). Since the RS scheme takes advantage of host memory, it performs better than our scheme, which is close to optimal only if we decide not to use memory. The gain of the RS scheme over our method varies over image databases; for the DOQQ database of aerial images used for our simulation, the performance gain is relatively small, whereas larger gains are observed for other image types such as scanned images and digital camera photographs. This strongly motivates further investigation into developing systematic parametric and nonparametric frameworks for steganalysis using host memory.

Performance for other types of images: We have also tested the approximate LRT on 128 scanned images and 3000 digital camera images. We have found that the performance for $R = 0.05$ degrades for these databases (compared with the DOQQ database), although it is much better than Stegdetect. As expected, the performance is worse than RS analysis, and the gap in difference in performance is greater than in Fig. 7 for the DOQQ database.

V. Conclusion

In this paper, we studied steganalysis of LSB hiding using hypothesis testing. The main conclusions are as follows.

1) The steganalysis problem can be cast in the framework of [16] to obtain a universal detector. However, in specific situations where the host PMF can be estimated well, the approach based on the estimation of the LR is preferable; this is the case with LSB hiding for the DOQQ database with which we experimented.
2) When the host PMF is known, the optimal composite hypothesis testing problem reduces to a worst case simple hypothesis testing problem.
3) For low hiding rate, the host PMF can be estimated well under a smoothness assumption by a number of simple methods. These can then be plugged into the optimal LRT to obtain approximate LRT. These tests perform close to the computationally intensive GLRT, especially at low hiding rates.
4) For a fixed host PMF, Stegdetect performance is close to the optimal LRT and to our approximate LRT, but for most image databases, the PMF varies significantly from image to image. Since the threshold required in Stegdetect to attain a target performance depends on the host PMF, and for image databases, the performance of Stegdetect degrades substantially. In comparison, the approximate LRT significantly outperforms Stegdetect since the threshold can be chosen independent of the host PMF.
5) For low hiding rates (less than 0.05), even in the ideal case with known PMF, i.i.d. simulated host, and known hiding rate, the performance of even the optimal LRT is not good; accordingly, practical tests applied to real data also do not give good performance for low hiding rates. This demonstrates the intrinsic difficulty of steganalysis for low hiding rates. Of course, the performance for natural images can be improved by using memory, as in [10] and [13].

An important area for future investigation is the exploitation of memory in the host samples for steganalysis since image coefficients have substantial local dependencies in both the pixel and transform domains. In principle, one could approximate such dependence by modeling blocks of host samples as i.i.d. and then applying methods similar to those in this paper. However, this would result in substantial increase in computational complexity. It is also unclear whether multidimensional densities can be estimated accurately with the amount of data available and what kind of smoothness assumptions, if any, would apply to such densities. Clearly, much more effort is required in models that lead to more economical representations of host memory that are amenable to a hypothesis testing framework.

Finally, a major effort is required to develop systematic approaches to steganalysis of other prevalent hiding strategies, such as variants of LSB hiding that employ memory in the hiding process [9] or methods such as quantization index modulation [1], [2].

Appendix A
Proof of Proposition 1

For simplicity, we denote R_0 by R. Recall that our goal is to minimize the function $D(\boldsymbol{q}\|\boldsymbol{p})$, which is convex in $\boldsymbol{p}$, over the convex set $\mathcal{P}_R$. We assume $R < 1$; the proof for $R = 1$ is straightforward. From the definition of $\mathcal{P}_R$, it is easy to verify that

$$\boldsymbol{p} \in \mathcal{P}_R \text{ iff } \frac{R}{2-R} \leq \frac{p_{2k+1}}{p_{2k}} \leq \frac{2-R}{R}, \quad k = 0, 1, \ldots, 127. \tag{12}$$

If both q_{2k} and q_{2k+1} are zero, then the corresponding terms from $D(\boldsymbol{q}\|\boldsymbol{p})$ drop out, and therefore, in the following, we assume $q_{2k} + q_{2k+1} \neq 0$. From the definition of $D(\boldsymbol{q}\|\boldsymbol{p})$, our problem is the same as maximizing

$$L(\boldsymbol{p}) = \sum_{k=0}^{255} q_k \log(p_k) \text{ subject to } \sum_{k=0}^{255} p_k = 1$$
$$\frac{R}{2-R} \leq \frac{p_{2l+1}}{p_{2l}} \leq \frac{2-R}{R}, \quad l = 0, 1, 2, \ldots, 127.$$

The trick to solving this problem is to solve it for each pair p_{2k}, p_{2k+1}. To this end, consider the following parametrization of $\mathcal{P}_R$:

$$\mathcal{P}_R = \bigcup \left\{ \{\boldsymbol{p} \geq 0 : p_{2k} + p_{2k+1} = \alpha_k (q_{2k} + q_{2k+1})\} \times \bigcap \left\{ \boldsymbol{p} \geq 0 : \frac{R}{2-R} \leq \frac{p_{2l+1}}{p_{2l}} \leq \frac{2-R}{R} \right\} \right\}$$

where the union is over all $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_{127}]^t > \mathbf{o}$ such that

$$\sum_{k=0}^{127} \alpha_k (q_{2k} + q_{2k+1}) = 1.$$

We first maximize $L(\boldsymbol{p})$ for $\boldsymbol{\alpha}$ fixed and then over $\boldsymbol{\alpha}$ satisfying these constraints. We write

$$L(\boldsymbol{p}) = \sum_{k=0}^{127} [q_{2k}\log(p_{2k}) + q_{2k+1}\log(p_{2k+1})].$$

To carry out the maximization, we need the following lemma.

Lemma 1: Suppose $a, b \geq 0$, $a+b \neq 0$, $c > 1$, and $\beta > 0$. Consider the problem of maximizing $F(x_1, x_2) = a\log(x_1) + b\log(x_2)$ subject to the conditions $x_1 + x_2 = \beta(a+b)$ and $1/c \leq x_1/x_2 \leq c$. The solution is given by

$$\begin{aligned} x_1^* &= \beta a, \quad x_2^* = \beta b, \text{ if } \frac{1}{c} \leq \frac{b}{a} \leq c \\ x_1^* &= \frac{\beta(a+b)}{1+c}, \quad x_2^* = \frac{c\beta(a+b)}{1+c}, \text{ if } \frac{b}{a} > c \\ x_1^* &= \frac{c\beta(a+b)}{1+c}, \quad x_2^* = \frac{\beta(a+b)}{1+c}, \text{ if } \frac{b}{a} < c. \end{aligned}$$

The proof of this lemma is simple: We substitute $x_2 = \beta(a+b) - x_1$ in $F(x_1, x_2)$ and then maximize with respect to x_1 under the remaining constraints. We skip the details.

Now, from Lemma 1, under the constraints

$$\begin{aligned} &p_{2k} + p_{2k+1} = \alpha_k(q_{2k} + q_{2k+1}) \\ &\text{and } \frac{R}{(2-R)} \leq \frac{p_{2k+1}}{p_{2k}} \leq \frac{(2-R)}{R} \end{aligned}$$

we get

$$\begin{aligned} &q_{2k}\log(p_{2k}) + q_{2k+1}\log(p_{2k+1}) \\ &\quad \leq (q_{2k} + q_{2k+1})\log(\alpha_k) + q_{2k}\log\left(p_{2k}^*\right) \\ &\qquad + q_{2k+1}\log\left(p_{2k+1}^*\right) \end{aligned}$$

where $\boldsymbol{p}^*$ is as specified in the statement of Proposition 1. Therefore

$$L(\boldsymbol{p}) \leq \sum_{k=0}^{127}(q_{2k} + q_{2k+1})\log(\alpha_k) + \sum_{l=0}^{255} q_l \log\left(p_l^*\right).$$

Under the constraints on $\boldsymbol{\alpha}$, it is easy to see that the above right-hand side is maximized for $\alpha_k = 1$ for all k, and this completes the proof.

Appendix B
Proof of Proposition 2

For simplicity of notation, we denote $\boldsymbol{p}^{(H)}$ by $\boldsymbol{p}$ and $r_k^{(H)}$ by r_k. Consider the log of the LR

$$S_{\mathrm{LLRT}}(\boldsymbol{q}) = D(\boldsymbol{q}\|\boldsymbol{p}_{R_0}) - D(\boldsymbol{q}\|\boldsymbol{p}) = \boldsymbol{a}^t\boldsymbol{q}$$

where $\boldsymbol{a}$ is a column vector whose kth entry is $\log(p_k/p_{R_0,k})$. In this proof, we repeatedly use the following estimates for small R_0:

$$a_{2k} = R_0\left(\frac{p_{2k} - p_{2k+1}}{2p_{2k}}\right) + O\left(R_0^2\right) \tag{13a}$$

$$a_{2k+1} = R_0\left(\frac{p_{2k+1} - p_{2k}}{2p_{2k+1}}\right) + O\left(R_0^2\right). \tag{13b}$$

We note that the false alarm probability $P(S_{\mathrm{LLRT}}(\boldsymbol{q}) < T|H_0)$ does not depend on the unknown hiding rate under H_1. Therefore, to prove the result, we wish to show that for a given threshold T, $F_N(T|H_\theta) = P(S_{\mathrm{LLRT}}(\boldsymbol{q}) > T|H_\theta)$ is decreasing for $R_0 \leq \theta \leq R_1$ so that the unit mass at R_0 is the least favorable distribution. (We note that $F_N(t|H_\theta)$ is the complementary distribution function of $S_{\mathrm{LLRT}}(\boldsymbol{q})$ under hypothesis H_θ.) To do so, we first obtain an approximation for $F_N(T|H_\theta)$ when R_0, R_1 are small and N is large; this is a rigorous derivation of the discussion in Section III-C. By the Berry–Esseen estimate for the rate of convergence in the central limit theorem ([24, Th. 4.9, pp. 126]), under hypothesis H_θ

$$\left|F_N(T|H_\theta) - Q\left(\sqrt{N}\frac{(T-\mu(\theta))}{\sigma(\theta)}\right)\right| \leq \frac{3}{\sqrt{N}}\frac{\mathbb{E}\left[|W_1|^3\right]}{\left(\mathbb{E}\left[W_1^2\right]\right)^{3/2}}.$$

where $W_1 = \boldsymbol{a}^t(\boldsymbol{Z}_1 - \boldsymbol{p}_\theta)$, and where $\boldsymbol{Z}_1$ is as defined in Section III-C. At the end of the proof, we show that

$$\frac{\mathbb{E}\left[|W_1|^3\right]}{\left(\mathbb{E}\left[W_1^2\right]\right)^{3/2}} \leq \text{constant} \tag{14}$$

where the constant does not depend on θ. Now, if R_0, $R_1 = O(1/\sqrt{N})$, then it is easy to check using (13) that $\mu(\theta) = O(1/N)$, and $\sigma(\theta) = O(1/\sqrt{N})$. Therefore, choosing $T = O(1/N)$, we get that

$$\gamma_N(\theta) := \frac{\sqrt{N}(T-\mu(\theta))}{\sigma(\theta)} = O(1).$$

It follows that for sufficiently large N and sufficiently small R_0, R_1

$$|F_N(T|H_\theta) - Q(\gamma_N(\theta))| \leq \frac{\text{constant}}{\sqrt{N}}. \tag{15}$$

Thus, for sufficiently small R_0, R_1, and N sufficiently large, we can approximate $P(\text{Miss}|H_\theta)$ by $Q(\gamma_N(\theta))$. Next, we show that the similar approximations hold for derivatives with respect to θ. From the above, we know that

$$F_N(T|H_\theta) = Q(\gamma_N(\theta)) + \frac{B(\theta,T)}{\sqrt{N}}.$$

Since we have N data samples, $\boldsymbol{q} = [K_0/N, \ldots, K_{255}/N]^t$ for some random positive integers $K_0, \ldots, K_{255}$, that is, $\boldsymbol{q}$ is of type N. Furthermore

$$\begin{aligned} P\left[\boldsymbol{q} = \left[\frac{k_0}{N}, \ldots, \frac{k_{255}}{N}\right]^t |H_\theta\right] \\ = \frac{N!}{k_0! \cdots mk_{255}!} p_{\theta,0}^{k_0} \cdots p_{\theta,255}^{k_{255}} \end{aligned}$$

which is infinitely differentiable in θ. Since

$$F_N(T|H_\theta) = \sum_{\{\text{PMFs}\boldsymbol{u}:S_{\mathrm{LLRT}}(\boldsymbol{u})>T\}} P[\boldsymbol{q} = \boldsymbol{u}|H_\theta]$$

and there are only finite number of type-N PMFs, we get that $F_N(T|H_\theta)$ is also infinitely differentiable. From the continuous differentiability of $\gamma_N(\theta)$ and $Q(t)$, we get that $B(\theta,T)$

is continuously differentiable in θ. Under the assumption $R_1 = c_0/\sqrt{N}$, $R_0 = c_1/\sqrt{N}$, and $T = T_0/N$ for some constants c_0, c_1, T_0, it can be checked from the Edgeworth expansion [25, pp. 229], using estimates like (14) for higher moments in place of the third moment, that $dB(\theta,T)/d\theta$ is bounded independently of T and θ. Thus, we get that for sufficiently large N

$$\left|\frac{dF_N(T|H_\theta)}{d\theta} - \frac{dQ(\gamma_N(\theta))}{d\theta}\right| \le \frac{\text{constant}}{\sqrt{N}}. \tag{16}$$

From approximations (15) and (16), it suffices to show that $Q(\gamma_N(\theta))$ is decreasing with θ.

Since the Q-function is monotonically decreasing, we need to establish that $G(\theta) := \gamma_N/\sqrt{N} = (T-\mu(\theta))/\sigma(\theta)$ is increasing, that is, its derivative is non-negative for $\theta \in [R_0, R_1]$. Taking the derivative of $G(\theta)$

$$G'(\theta) = -\frac{\mu'(\theta)\sigma^2(\theta) + (T-\mu(\theta))\left(\frac{\sigma^2(\theta)}{2}\right)'}{\sigma^3(\theta)} =: -\frac{V(\theta)}{\sigma^3(\theta)}.$$

Our goal is to show that $V(\theta)$ is nonpositive in the desired region. We begin by obtaining an expression for $V(\theta)$.

Let $\boldsymbol{w}$ be the vector such that $w_{2k} = p_{2k} - p_{2k+1}$ and $w_{2k+1} = p_{2k+1} - p_{2k}$. Then, it is easy to see that $\boldsymbol{p}_\theta = \boldsymbol{p} - \theta\boldsymbol{w}/2$, and hence, $\mu(\theta) = \mu(0) + \beta\theta$, where $\beta = -\boldsymbol{a}^t\boldsymbol{w}/2$. Substituting for $\boldsymbol{a}$ and $\boldsymbol{w}$, we get

$$\beta = \frac{1}{2}\sum_{k=0}^{127} p_{2k}(r_k - 1)\log\left(\frac{\left(\frac{R_0}{2}\right) + \left(\frac{1-R_0}{2}\right)r_k}{\left(\frac{R_0}{2}\right)r_k^2 + \left(\frac{1-R_0}{2}\right)r_k}\right).$$

We note that each summand is negative, so that $\beta < 0$ and $\mu(\theta)$ is decreasing. Hence, the restriction $P(\text{Miss}) \le 0.5$ and $P(\text{False Alarm}) \le 0.5$ implies that

$$\mu(R_0) \le T \le \mu(0). \tag{17}$$

Similarly, we obtain

$$\sigma^2(\theta) = \sigma^2(0) + b\theta + c\theta^2$$

where

$$b = -\frac{1}{2}\boldsymbol{a}^t(\text{diag}(\boldsymbol{w}) + 2\boldsymbol{p}\boldsymbol{w}^t)\boldsymbol{a}, \quad c = \beta^2.$$

Putting $\phi = \mu(0) - T$, we obtain

$$V(\theta) = \left(\frac{\beta b}{2} - \phi c\right)\theta + \left(\beta\sigma^2(0) - \frac{\phi b}{2}\right).$$

From (17), we know that $\phi \ge 0$. We showed above that $\beta < 0$ and $c = \beta^2 > 0$. Therefore, to prove that $V(\theta) < 0$, it suffices to show that $b > 0$. We note that

$$\begin{aligned} b &= -\frac{1}{2}\boldsymbol{a}^t\text{diag}(\boldsymbol{w})\boldsymbol{a} - 2(\boldsymbol{a}^t\boldsymbol{p})\beta \\ &= -\frac{1}{2}\sum_{k=0}^{127}(p_{2k+1} - p_{2k})\left(a_{2k+1}^2 - a_{2k}^2\right) - 2D(p\|p_{R_0})\beta. \end{aligned}$$

Using (13), we know that $D(\boldsymbol{p}\|\boldsymbol{p}_{R_0}) = O(R_0^2)$, $\beta = O(R_0)$, and

$$b = \frac{R_0^2}{8}\sum_{k=0}^{127}(p_{2k+1} - p_{2k})^4\frac{(p_{2k+1} + p_{2k})}{p_{2k}^2 p_{2k+1}^2} + O\left(R_0^3\right).$$

Thus, for R_0 sufficiently small, b is positive, and this proves that $V(\theta) \le 0$. The proof is complete, except that we still have to show (14).

Proof of (14): We note that

$$|W_1| \le 2\max_k |a_k| \le \text{constant} \cdot R_0$$

where we have used (13), and the constant does not depend on θ since a_k does not depend on θ. Therefore, we get that $\mathbb{E}\left[|W_1|^3\right] \le \text{constant} \cdot R_0^3$. To prove (14), we now show that $\mathbb{E}\left[W_1^2\right] \ge \text{constant} \cdot R_0^2$. Since $c \ge 0$ and $b > 0$ for sufficiently small R_0, $\sigma^2(\theta) \ge \sigma^2(0)$ for sufficiently small R_0. Using (13)

$$\mathbb{E}\left[W_1^2\right] = \sigma^2(\theta) \ge \sigma^2(0) = \frac{U(p)R_0^2}{4} + O\left(R_0^3\right)$$

where

$$U(\boldsymbol{p}) = \sum_{k=0}^{127}\left\{(p_{2k} + p_{2k+1})\left(r_k + \frac{1}{r_k} - 2\right)\right\}.$$

Since $r_k \ne 1$ for at least some k, $U(p) > 0$, and we get $\mathbb{E}[W_1^2] \ge \text{constant} \cdot R_0^2$.

ACKNOWLEDGMENT

The authors are thankful to the reviewers for their careful reading and detailed comments.

REFERENCES

[1] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.

[2] N. Jacobsen, K. Solanki, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Image-adaptive high-volume data hiding based on scalar quantization," in *Proc. IEEE Military Commun. Conf.*, Anaheim, CA, Oct. 2002.

[3] K. Solanki, O. Dabeer, B. Manjunath, U. Madhow, and S. Chandrasekaran, "A joint source-channel coding scheme for image-in-image hiding," in *Proc. ICIP*, Sept. 2003.

[4] J. Chou and K. Ramachandran, "Robust turbo-based data hiding for image and video sources," in *Proc. ICIP*, Oct. 2002.

[5] J. Eggers, R. Buml, R. Tzschoppe, and B. Girod, "Scalar costa scheme for information embedding," *IEEE Trans. Signal Processing*, vol. 51, pp. 1003–1019, Apr. 2003.

[6] P. Moulin and J. A. O'sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, pp. 563–593, Mar. 2003.

[7] N. Provos and P. Honeyman. Detecting steganographic content on the internet. presented at ISOC NDSS. [Online]. Available: http://www.outguess.org/

[8] [Online]. Available: http://hacktivismo.com/projects/camerashy/

[9] N. Provos, "Defending against statistical steganalysis," in *Proc. 10th USENIX Security Symp.*, Washington, DC, 2001.

[10] J. Fridrich and M. Golijan, "Practical steganalysis of digital images—State of the art," *Proc. SPIE*, vol. 4675, 2002.

[11] H. Farid, "Detecting Steganographic Messages in Digital Images," Comput. Sci. Dept., Dartmouth College, Hanover, NH, 2001.

[12] R. Chandramouli and N. Memon, "Analysis of LSB based image steganography techniques," in *Proc. ICIP*, Oct. 2001.
[13] S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *IEEE Trans. Signal Processing*, vol. 51, pp. 1995–2007, July 2003.
[14] E. Lehmann, *Testing Statistical Hypothesis*. New York: Wiley, 1959.
[15] V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 1994.
[16] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.
[17] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitive minimax approach," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1504–1517, June 2002.
[18] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2215–2229, Aug. 2002.
[19] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, CA: Athena Pacific, 1999.
[20] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1597–1602, Sept. 1992.
[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[22] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Second ed, ser. Springer Series in Statistics. New York: Springer-Verlag, 1991.
[23] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Second ed. New York: Springer, 1998.
[24] R. Durrett, *Probability: Theory and Examples*, Second ed. Belmont, CA: Duxbury, 1996.
[25] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1961.

Onkar Dabeer (S'99–M'02) was born in Solapur, India, on September 15, 1974. He received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology, Bombay, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of California at San Diego, La Jolla, in June 2002.

From June 2002 to July 2003, he was a postdoctoral researcher at the University of California, Santa Barbara. Currently, he is with Qualcomm Inc., San Diego. His research interests include wireless communication systems, adaptive and stochastic approximation algorithms, information theory, and steganalysis.

Kenneth Sullivan received the B.S. degree in electrical engineering from the University of California at San Diego, La Jolla, in 1998 and the M.S. degree in 2002 at the University of California at Santa Barbara, where he is currently pursuing the Ph.D. degree.

His research interests include communications, image processing, steganalysis, and steganography.

Upamanyu Madhow (SM'96) received the bachelor's degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1985 and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois, Urbana-Champaign, in 1987 and 1990, respectively.

From 1990 to 1991, he was a Visiting Assistant Professor at the University of Illinois. From 1991 to 1994, he was a research scientist at Bell Communications Research, Morristown, NJ. From 1994 to 1999, he was with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, first as an Assistant Professor and, since 1998, as an Associate Professor. Since December 1999, he has been with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. His research interests are in communication systems and networking, with current emphasis on wireless communication, sensor networks, and data hiding.

Dr. Madhow received the National Science Foundation CAREER award. He has served as Associate Editor for Spread Spectrum for the IEEE TRANSACTIONS ON COMMUNICATIONS and as Associate Editor for Detection and Estimation for the IEEE TRANSACTIONS ON INFORMATION THEORY.

Shivkumar Chandrasekaran received the M.Sc. degree in physics from B.I.T.S., Pilani, India, and the Ph.D. in computer science from Yale University, New Haven, CT.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of California, Santa Barbara. His research interests include numerical evaluation of structured systems of linear equations, differential and integral equations, and inverse scattering and computer vision problems.

Dr. Chandrasekaran received a 1998 National Science Foundation CAREER award.

B. S. Manjunath (SM'01) received the B.E. degree in electronics (with distinction) from Bangalore University, Bangalore, India, in 1985, the M.E. degree (with distinction) in systems science and automation from the Indian Institute of Science, Bangalore, in 1987, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1991.

He is now a Professor of electrical and computer engineering and Director of the Center for Bio-Image Informatics at the University of California, Santa Barbara. His current research interests include data mining, computer vision, learning algorithms, image/video databases, and bio-image informatics. He is a co-editor of the book *Introduction to MPEG-7* (New York: Wiley, 2002).

Dr. Manjunath received the National Merit Scholarship (for 1978 to 1985) and was awarded the university gold medal for the best graduating student in electronics engineering in 1985 from Bangalore University. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.