

# ML Detection of Steganography

Mark T. Hogan, Neil J. Hurley, Gu enol e C.M. Silvestre, F elix Balado and Kevin M. Whelan.

Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland.

## ABSTRACT

Digital steganography is the art of hiding information in multimedia content, such that it remains perceptually and statistically unchanged. The detection of such covert communication is referred to as steganalysis. To date, steganalysis research has focused primarily on either, the extraction of features from a document that are sensitive to the embedding, or the inference of some statistical difference between marked and unmarked objects. In this work, we evaluate the statistical limits of such techniques by developing asymptotically optimal tests (Maximum Likelihood) for a number of side informed embedding schemes. The required probability density functions (pdf) are derived for Dither Modulation (DM) and Distortion-Compensated Dither Modulation (DC-DM/SCS) from an steganalyst's point of view. For both embedding techniques, the pdfs are derived in the presence and absence of a secret dither key. The resulting tests are then compared to a robust blind steganalytic test based on feature extraction. The performance of the tests is evaluated using an integral measure and receiver operating characteristic (ROC) curves.

**Keywords:** Non-blind steganalysis, Distortion-Compensated Dither Modulation, maximum likelihood detection.

## 1. INTRODUCTION

The term steganography refers to the family of techniques used to hide data within a *covert* multimedia signal. Ideally, the corresponding modified signal, referred to as *stegotext*, is perceptually *and* statistically indistinguishable from the covertext. The classical representation of steganographic communication is given by the prisoners' problem.<sup>1</sup> Alice produces a stegotext using the message that she wants to communicate and a given covertext, and sends it to Bob through an insecure communications channel. Usually, Alice and Bob make use of secret keys for their covert communication. The warden Wendy monitors the channel between Alice and Bob, and performs a detection test to decide if the signal being sent includes hidden information by exploiting potential imperfections of the steganographic method used. In an analogous way to cryptanalysis, this detection procedure is known as *steganalysis*.

Steganalysis approaches may be blind or non-blind. In blind (or universal) steganalysis the detection test is independent of the steganographic method actually employed. On the other hand, non-blind steganalysis targets specific embedding techniques. In this work we will concentrate on non-blind steganalysis of continuous signals. We will consider that Wendy is a "passive warden", that is, that she never modifies the stegotext but just interrupts the communication in the case of a positive detection test. The possible existence of an active warden is of course not relevant for the detection test, but this case raises robustness issues which certainly influence the performance (and then, the choice) of the steganographic method used by Alice and Bob, as we discuss later.

The main problem of non-blind steganalysis is the uncertainty of Wendy about that steganographic method. Nevertheless, we may argue that the interest of investigating non-blind steganalysis is three-fold. First, non-blind steganalysis allows to determine the reliability of the covert channel in the worst-case scenario for Alice and Bob in which Wendy knows the details of the covert communications procedure. This case resembles the situation that happens in cryptography when a determined standard becomes widespread, and then security, following Kerckhoffs' principle, depends only on the secret key. Secondly, it is the only way of rigorously establishing the

---

Further author information: (Send correspondence to M.H.)

- M.H.: E-mail: markhogan@ihl.ucd.ie, Telephone: +353 (0)1 716 2454.

- This work is kindly supported by Enterprise Ireland under research grant ATRP-2002/230 and the European Commission through the IST Programme under contract IST-2002-507609 SIMILAR.

performance limits of blind steganalytic methods—which are usually heuristically designed—when applied to a given steganographic technique. Lastly, the results of non-blind steganalysis may also have applicability in other practical scenarios, as for instance in the bank-based setting studied by Chandramouli<sup>2</sup> in which the reliability of the detection for different methods is necessary to obtain the actual payoffs of the game between Wendy and Alice/Bob.

The non-blind analysis presented here is devoted to study the detectability of Distortion-Compensated Dither Modulation (DC-DM) with uniform scalar quantizers,<sup>3</sup> or, equivalently, SCS (Scalar Costa Scheme).<sup>4</sup> The justification for choosing this particular method as the steganographic technique used by Alice and Bob is as follows. If we assume that, after the detection test, the stegotext undergoes a certain noise channel before reaching Bob, it is clear that steganographic algorithms able to withstand such distortion are required. This is actually the main concern—rather than undetectability—of watermarking methods. In this context, it is well known that side-informed quantization-based watermarking algorithms such as DC-DM, are able to approach the achievable rate of additive and independent noise channels for the case in which Bob also possesses the same side information. Therefore, the potential of these methods in scenarios with a passive warden and unintentional channel distortions makes worthwhile their study under non-blind steganalysis.

Our approach will be based on obtaining maximum likelihood (ML) detectors for deciding between covertext and DC-DM stegotext in several particular cases, for which the statistical distributions observed by Wendy have to be obtained. This problem was partially tackled already by Sullivan *et al.*,<sup>5</sup> who presented ML detection tests restricted to DM (a particular case of DC-DM). In this paper we extend that work to full DC-DM, considering a number of different situations for Wendy and removing the necessity of previous assumptions<sup>5</sup> which limited the applicability of those results. We have to remark that other works closely related to the approach here are those oriented to obtain the Kullback-Leibler distance ( $D_{\text{KL}}$ ) between the distributions of the stegotext and the covertext for DC-DM. This owes to the fact that the performance of an optimal binary hypothesis test is determined by the  $D_{\text{KL}}$  between the corresponding distributions involved (Stein’s lemma,<sup>6</sup> see also Wang and Moulin<sup>7</sup>). Among those works there is the one by Guillon *et al.*,<sup>8</sup> who empirically assess the security of SCS using the  $D_{\text{KL}}$ , and the one by Wang and Moulin,<sup>7</sup> who give numerical evaluations of this distance for DC-DM (DC-QIM), therefore dealing with some of the distributions that we will employ here.

Detection results are presented in this work in terms of the receiver operating characteristic (ROC) curves for all four cases studied, namely, the detection of DC-DM and DM when Wendy does not know the secret key, and when the key has been leaked to her. Finally, we evaluate the blind technique by Farid,<sup>9</sup> which has been empirically shown to be one of the best blind steganalytic methods, against the derived non-blind detectors, and we propose a measure by which the performance of such blind detection techniques can be gauged.

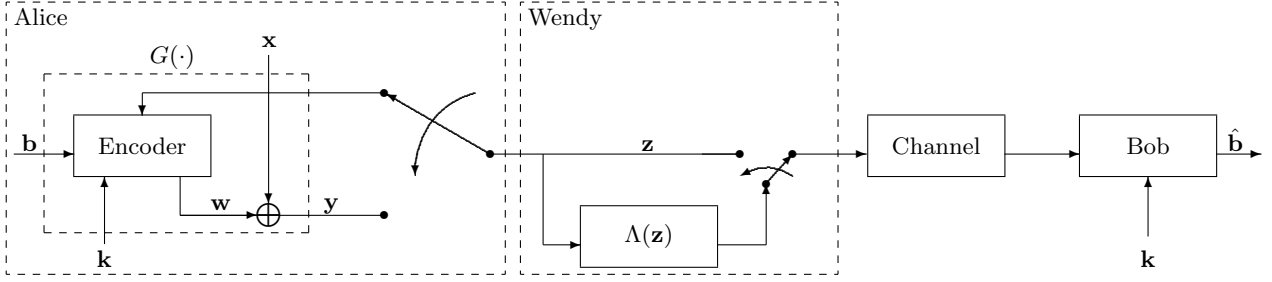
The paper is organized as follows. Section 2 presents the framework and the generic detection approach used. Next, Sections 3 and 4 detail the development of the particular likelihood ratios for DM and DC-DM in the different cases considered. In Section 5 results are presented for ML detection alongside the blind steganalysis algorithm. Finally, conclusions are drawn in Section 6.

## 2. PROBLEM SET-UP

### 2.1. Notation and Preliminaries

In this work capital letters refer to random variables, e.g.  $X$ , while lower case are realizations of random variables, e.g.  $x$ . Bold, capital,  $\mathbf{X}$ , and lower case letters,  $\mathbf{x}$ , refer to random vectors and their realizations, respectively. Individual elements of  $\mathbf{x}$  are indexed as  $x_j$ . All vectors are of length  $N$ . The probability density function (pdf) of a random variable  $X$  is denoted as  $f_X(x)$ , and the corresponding cumulative distribution function (cdf) is denoted by  $F_X(x)$ . The statistical expectation of  $X$  is denoted  $E_X\{X\}$ . Subscripts will be dropped for notational simplicity wherever there is no ambiguity in doing this.

We assume that the covertext,  $\mathbf{x} = [x_1, \dots, x_N]$ , consists of a realization of a random vector  $\mathbf{X}$  formed by independent identically distributed (iid), zero-mean random variables. Alice may send either this covertext to Bob, or modify it before transmission to embed  $\mathbf{b} = [b_1, \dots, b_N]$ ,  $b_i \in \{0, 1\}$ , a sequence of message bits drawn from a uniform distribution as shown in Figure 1. This produces a stegotext (watermarked) vector  $\mathbf{y} = G(\mathbf{x}, \mathbf{b})$ , and the watermark  $\mathbf{w}$  is then given as  $\mathbf{w} \triangleq \mathbf{y} - \mathbf{x}$ . We assume that only one information symbol  $b_j$  is embedded



**Figure 1.** Diagram of steganographic communications using side information at the encoder, with passive warden and non-blind steganalysis.

in one corresponding cocontext sample  $x_j$ . The embedding process may be secured by using a pseudorandom symmetric key  $\mathbf{k}$ , shared by Alice and Bob, and then  $\mathbf{y} = G_{\mathbf{k}}(\mathbf{x}, \mathbf{b})$ . In this proposed setting, we may consider the random vectors  $\mathbf{Y}$  and  $\mathbf{W}$  to be respectively iid.

From Wendy's viewpoint the assumption that the message bits are equally likely is tantamount to assuming that Alice embeds a maximum entropy message. Arguably, Alice will not want to send redundant information which could be exploited by Wendy, and so she will likely compress the message before embedding. Furthermore, if Alice encrypts her message with perfectly secure encryption, the resulting bit stream will also be viewed by Wendy as having maximum entropy.

An important parameter for establishing the working point of the steganographic method is the *watermark to content* power ratio (WCR). This is the average power of the watermark normalized by the content (cocontext) energy, which can be written as

$$\text{WCR} \triangleq \frac{\mathbb{E}\{\|\mathbf{W}\|^2\}}{\mathbb{E}\{\|\mathbf{X}\|^2\}} = \frac{\sigma_W^2}{\sigma_X^2},$$

where  $\sigma_X^2$  and  $\sigma_W^2$  refer to the variances of the cocontext (host signal) and watermark, respectively, assuming that  $W$  has zero mean. If  $X$  and  $Y$  are independent,  $\sigma_W^2 = \sigma_Y^2 - \sigma_X^2$ . Notice that the perceptual constraints in any data hiding problem impose very low values for the WCR. It is intuitively clear that lowering the WCR will lead to worse detection performance.

**Detection Test.** According to the preceding section, Wendy must decide if a given monitored signal  $\mathbf{z}$  sent to Bob by Alice has been drawn either from  $f_{\mathbf{X}}(\cdot)$  or from  $f_{\mathbf{Y}}(\cdot)$ . Assuming that  $f_{\mathbf{X}}(\cdot)$  is known, which as we will discuss later is plausible in certain domains of interest, then, given  $G(\cdot)$ ,  $f_{\mathbf{Y}}(\cdot)$  is also known. This detection problem is then a hypothesis testing problem with two choices, denoted as the null hypothesis  $H_0$  ( $\mathbf{z}$  is a cocontext), and the alternative hypothesis,  $H_1$  ( $\mathbf{z}$  is a stegotext). To make a decision on  $\mathbf{z}$ , the optimal test based on the Bayes likelihood ratio,<sup>10</sup> and is given by

$$\Lambda(\mathbf{z}) \triangleq \frac{f_{\mathbf{X}}(\mathbf{z})}{f_{\mathbf{Y}}(\mathbf{z})} \underset{H_0}{\overset{H_1}{\geq}} \frac{P_0 C_{10} - C_{00}}{P_1 C_{01} - C_{11}}, \quad (1)$$

with  $P_i$ ,  $i \in \{0, 1\}$  the *a priori* probabilities for the null and alternative hypotheses respectively, and  $C_{ij}$  the cost of choosing  $H_i$  when the true hypothesis is  $H_j$ . Letting  $C_{ij} = \delta_{ij}$ , with  $\delta_{ij}$  the Kronecker delta function, and choosing the *a priori* probabilities to be uniformly distributed, gives the ML test (see Figure 1). This test is asymptotically optimal, and, as such, provides a limit to the performance of blind detection techniques for the particular embedding method considered. Using the iid assumptions, the test (1) can be rewritten more conveniently as

$$\log \Lambda(\mathbf{z}) = \log \frac{\prod_{j=1}^N f_X(z_j)}{\prod_{j=1}^N f_Y(z_j)} = \sum_{j=1}^N \log f_X(z_j) - \sum_{j=1}^N \log f_Y(z_j) \underset{H_0}{\overset{H_1}{\geq}} 0. \quad (2)$$

## 2.2. Distortion-Compensated Dither Modulation

DC-DM with uniform scalar quantizers is a practical implementation of Distortion-Compensated Quantization Index Modulation (DC-QIM), proposed by Chen and Wornell.<sup>3</sup> The embedding technique is based on the quantization of the covertext samples with a dithered version of a uniform scalar quantizer  $Q_\Delta(\cdot)$ . We assume that the quantization step  $\Delta$  is the same for all covertext samples, and that it is disclosed to Wendy. As already mentioned, we consider binary messages and so the embedding takes place with two quantizers shifted by  $\Delta/2$ . In order to embed a binary symbol  $b_j$  at the covertext sample  $x_j$  the corresponding stegotext sample  $y_j$  is obtained in DC-DM as

$$y_j = G_{k_j}(x_j, b_j) = x_j + \alpha \left[ Q_\Delta \left( x_j - k_j - \Delta \frac{b_j}{2} \right) + k_j + \Delta \frac{b_j}{2} - x_j \right], \quad (3)$$

where the additional dither  $k_j \in (-\frac{\Delta}{2}, \frac{\Delta}{2}]$  is the secret key shared by Alice and Bob at the  $j$ th sample. The distortion compensation factor  $0 < \alpha \leq 1$  allows for tuning the method for optimal robustness to channel noise, assuming that its power is known in advance,<sup>4</sup> or alternatively, for tuning its detectability properties.<sup>7,8</sup> As may be verified from (3), the distortion compensation consists in adding a fraction  $1 - \alpha$  of the quantization error to the quantized covertext.

As a digression consider that, despite the discussion in the preceding section about the near-optimality of this method under additive white Gaussian noise, it is somewhat paradoxical that capacity may only be achieved by encoding the message. The statistical dependencies thus introduced between the information symbols could then be exploited by Wendy to detect that covert communication is under way (cf. the corresponding discussion in the preceding section), indicating that the real achievable rate of DC-DM in a steganographic context has to be lower than expected. In any case, we can expect from the method a reasonable degree of self-noise cancellation. We will not study the coded case here; for prior discussions on steganographic capacity see for instance.<sup>11</sup>

Assuming that the quantization error is approximately uniform and independent from  $\mathbf{X}$ , the WCR is for this embedding method is given by

$$\text{WCR} = \frac{\alpha^2 \Delta^2}{12\sigma_X^2}. \quad (4)$$

DM is a particular case of DC-DM for which  $\alpha = 1$ . As we will see in Section 3, this case will require a particular treatment due to the peculiar discrete distribution of the stegotext associated.

## 2.3. Statistical Models

According to (2) it is now necessary to obtain the statistical models needed for tackling ML detection. Sections 3 and 4 will be devoted to derive general expressions of  $f_Y(y)$  for DM and DC-DM, respectively, which may be accomplished from  $f_X(x)$  using (3). In order to undertake the actual test, Wendy will need to know  $f_X(x)$ . We will assume that it is known that the embedding takes place in a transformed domain, and then that the covertext may be reasonably well modelled by a Laplacian random variable, whose pdf is given by

$$f_X(x; \lambda) = \frac{\lambda}{2} \exp(-\lambda|x|). \quad (5)$$

This is a common choice to represent the transform coefficients (e.g. discrete cosine and wavelet transforms) of natural images.<sup>12</sup> Although it is out of the scope of the worst-case framework considered here, it is interesting to note that the use of domains where statistical modelling is difficult —such as the spatial domain— is actually advantageous for hampering detection, as in this case it is not possible to undertake the test (1).

Before the test, Wendy has to undertake the estimation of the pdf parameter  $\lambda$  for both  $f_{\mathbf{X}}(\mathbf{x}; \lambda)$  and  $f_{\mathbf{Y}}(\mathbf{y}; \lambda)$ . Let  $\hat{\lambda}_X$  denote the estimate of  $\lambda$  under  $H_0$  and  $\hat{\lambda}_Y$  the estimate under  $H_1$ . Now, assuming that  $\mathbf{z}$  is drawn from (5), this estimation is carried out as,<sup>10</sup>

$$\hat{\lambda}_X = \arg \max_{\tilde{\lambda}} f_{\mathbf{X}}(\mathbf{z}; \tilde{\lambda}). \quad (6)$$

Here  $\hat{\lambda}_X = N / \sum_{j=1}^N |z_j|$  is the ML estimate of  $\lambda$ . Using  $f_Y(\mathbf{z}; \tilde{\lambda})$  for the ML estimate gives, similarly, the estimate for  $\hat{\lambda}_Y$ . However, in general for the stegotext pdfs examined here, we will see that explicit expressions for  $\hat{\lambda}_Y$  are not available so  $\hat{\lambda}_Y$  must be obtained numerically. For the actual test, (2), both estimates  $\hat{\lambda}_X$  and  $\hat{\lambda}_Y$  are used.

Some more considerations on the pdf of  $Y$  are necessary. Depending on the knowledge available to Wendy about the secret key, we may distinguish two cases in the statistical modelling. In the first one,  $\mathbf{k}$  has been stealthily leaked to Wendy. Then, she “sees” the same pdf of  $Y$  as Bob would for  $H_1$ , that is,  $f_Y(\mathbf{y}|\mathbf{k})$ , and she can exploit this knowledge to her advantage in (2). For analytical purposes, we may equivalently see this case as *unkeyed*, and we may use any particular key, as for instance  $\mathbf{k} = \mathbf{0}$ , without loss of generality. An alternative detection approach in this case would be decoding  $\hat{\mathbf{b}}$  from  $\mathbf{z}$  using  $\mathbf{k}$ , and then deciding if  $\hat{\mathbf{b}}$  is a valid message. A possible way to accomplish this could be estimating blindly (i.e., without reference to  $\mathbf{b}$ ) the probability of error at the decoder, but this can only be easily done when the message is known to be encoded using a given channel code (cf. previous discussion about coding).

The situation is completely different in the second case in which Wendy has no deterministic information about  $\mathbf{k}$ . Lacking any further statistical knowledge, Wendy may only assume that the key is a uniform random vector  $\mathbf{K}$  with iid elements such that  $K_j \sim U(-\frac{\Delta}{2}, \frac{\Delta}{2})$ . One possible approach in this case is assuming that Wendy uses the expected pdf, that is,

$$\tilde{f}_Y(y) = E_K\{f(y|k)\} = \int f_Y(y|k) \cdot f_K(k) dk. \quad (7)$$

Notice that an alternative approach such as the estimation of  $\mathbf{k}$  by Wendy is a completely different problem, because it requires first determining that  $\mathbf{z}$  is a stegotext. This could lead to an iterative approach, that we do not study here.

### 3. DM DETECTION

In this section we obtain the distributions necessary for DM detection in the aforementioned cases.

#### 3.1. DM with a Leaked Key

Firstly we examine the case where  $\mathbf{k} = \mathbf{0}$  (see Section 2.3). Under our assumptions, it is straightforward to show that in this case the model of  $Y$  is given as a train of weighted Dirac  $\delta$ -functions. We thus have

$$f_Y(y|k) = \sum_{i=-\infty}^{\infty} w_i \delta(y - i\Delta/2), \quad (8)$$

where the weights  $w_i$  on each of the  $\delta$ -functions are given by the following expression

$$w_i = \frac{1}{2} \int_{\frac{i\Delta}{2} - \frac{\Delta}{2}}^{\frac{i\Delta}{2} + \frac{\Delta}{2}} f_X(x) dx = \frac{1}{2} \left( F_X \left( \frac{i\Delta}{2} + \frac{\Delta}{2} \right) - F_X \left( \frac{i\Delta}{2} - \frac{\Delta}{2} \right) \right), \quad \forall i \in \mathbb{Z}.$$

This gives an exact expression for  $f_Y(y|k)$  but this model does not provide full support over  $y$ . This poses problems when the log-likelihood ratio is calculated if the received document is a covertext (or, indeed, in the case of a stegotext subject to additive noise attack). In order to sort out this difficulty, we propose to extend the support set of the pdf to the entire real line by replacing the  $\delta$ -functions with approximations to them. To do this we convolve the train with another function  $g(t; \lambda_n)$  such that  $g(t; \lambda_n)$  integrates to one over  $\mathbb{R}$ , where  $\lambda_n$  is an empirically chosen shape parameter. Our model now becomes  $f(y|k) \approx \sum_{i=-\infty}^{\infty} w_i [g(y; \lambda_n) * \delta(y - \frac{i\Delta}{2})]$ . Let  $g(y; \lambda_n) = \frac{\lambda_n}{2} \exp(-\lambda_n|y|)$ , i.e. a Laplacian function. This gives,

$$f(y|k) \approx \sum_{i=-\infty}^{\infty} \frac{w_i \lambda_n}{2} e^{-\lambda_n |y - \frac{i\Delta}{2}|}.$$

This function has full support over  $\mathbb{R}$  and as such avoids the potential problems involved with the previous model. The parameter  $\lambda_n$  should be kept large to maintain a good approximation to the true pdf. In Sullivan *et*

al.<sup>5</sup> the authors develop an hypothesis test for the same scenario examined here. However they make no reference to the problem of the support issues of the stegotext pdf. Instead they treat the case of detection of DM against a host pdf which is also quantized. Thus only certain regions of  $y$  are permitted in their test. Assumptions are also made about the content of  $\mathbf{b}$ . Our proposed approximation overcomes both of these problems as the test has full support and makes no reference to  $\mathbf{b}$ .

### 3.2. DM with a Secret Key

As detailed in Section 2.3, when Wendy does not know the key she can design her test for the expected pdf given that  $K$  can be viewed as a uniform random variable. We will see that for this particular case,  $\tilde{f}(y)$  is continuously valued unlike (8).

To begin the analysis assume that the value of  $K = k$ . Then, as before, we have that the stegotext pdf consists of a train of  $\delta$ -functions (delayed by a value  $k$ ) which is given by

$$f(y|k) = \sum_{i=-\infty}^{\infty} w_i^k \delta\left(y - \left(\frac{i\Delta}{2} + k\right)\right), \quad (9)$$

with,

$$w_i^k = \frac{1}{2} \int_{\frac{i\Delta}{2} + k - \frac{\Delta}{2}}^{\frac{i\Delta}{2} + k + \frac{\Delta}{2}} f_X(x) dx = \frac{1}{2} \left( F_X\left(\frac{i\Delta}{2} + k + \frac{\Delta}{2}\right) - F_X\left(\frac{i\Delta}{2} + k - \frac{\Delta}{2}\right) \right), \quad \forall i \in \mathbb{Z}.$$

Following (7) the average pdf is calculated as the expectation of (9) over  $K$ . The details of this calculation can be found in Appendix A.  $\tilde{f}(y)$  is given as

$$\tilde{f}(y) = \frac{1}{\Delta} \left( F_X\left(y + \frac{\Delta}{2}\right) - F_X\left(y - \frac{\Delta}{2}\right) \right). \quad (10)$$

The result can be seen to be the convolution of the host signal pdf  $f_X(x)$  with a uniform random variable,  $U\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$ . Therefore we have the following expression for the pdf of a stegotext

$$\tilde{f}_Y(y) = f_X(y) * U\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right). \quad (11)$$

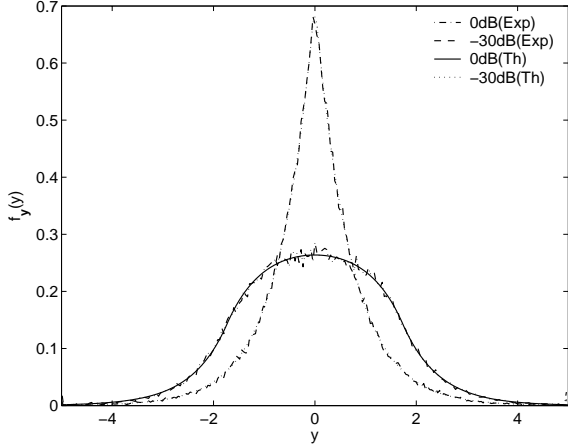
In Wang and Moulin<sup>7</sup> the authors obtain the same result for the pdf of keyed DM, albeit with different reasoning. In Figure 2, (11) is shown in comparison to the empirical pdf for secure DM for two separate WCRs. It can be seen that the theoretical result closely matches the empirical histograms for any WCR.

## 4. DC-DM DETECTION

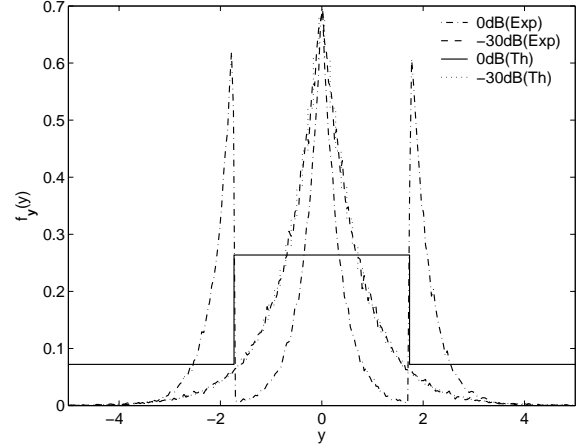
We undertake next the derivation of the distributions necessary for DC-DM detection in the two cases considered.

### 4.1. DC-DM with a Leaked Key

Again we make the assumption that  $\mathbf{k} = \mathbf{0}$  for the case where Wendy has access to the key. The analysis of DC-DM in<sup>4</sup> is simplified by assuming that the pdf of  $\mathbf{x}$  is uniform within each quantization bin. The problem is then reduced to the analysis of one bin. This is equivalent to stating that the host energy  $\sigma_X^2 \rightarrow \infty$  (WCR tends to zero). Here we adopt the same approach but do not make the assumption of the WCR tending to zero. Instead the pdf is assumed to be uniform within each quantization bin with a weight given by (9). It has been suggested that the best performance from a steganographic point of view (i.e. secrecy and not noise performance) is obtained when  $\alpha = 0.5$ ,<sup>7,8</sup> which is the value we adopt for our tests. The width of the weighted uniform around each quantization point is given as  $\Delta(1 - \alpha)$ , which for  $\alpha = 1/2$  gives a width of  $\Delta/2$ . This is, of course, the distance between the quantization points. As such this particular value of  $\alpha$  means that there are no



**Figure 2.** Theoretical  $\tilde{f}(y)$  (Th) for DM (see (11)) shown alongside experimental histograms (Exp) for host given by (5) with  $\lambda = \sqrt{2}$ . WCRs of 0 dB and -30 dB are shown.



**Figure 3.** Theoretical  $f(y|k)$  (Th) for DC-DM with  $\alpha = 0.5$  (see (12)) shown alongside experimental histograms (Exp) for host given by (5) with  $\lambda = \sqrt{2}$ . WCRs of 0 dB and -30 dB are shown.

areas of zero probability within the overall pdf and allows for zero error decoding in the absence of noise. The expression for the conditional pdf required is given as,

$$f(y|k) = \frac{1}{\Delta(1-\alpha)} \sum_{i=-\infty}^{\infty} w_i \pi \left( \frac{y}{\Delta(1-\alpha)} - \frac{i\Delta}{2} \right), \quad (12)$$

where the  $w_i$  are weights given by (9) and  $\pi(z)$  is a function given as,

$$\pi(z) \triangleq \begin{cases} 1 & z \in \left(-\frac{1}{2}, \frac{1}{2}\right], \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3 illustrates the chosen model for two separate WCRs. It is clear that for low enough WCRs the model is representative of the true pdf but that as the WCR increases the model deviates significantly from the true pdf. This has been noted in<sup>13</sup> where the “flat host assumption” is found to be inaccurate when the watermark has high power. Despite the shortcomings of this model, it will be shown in Section 5 that the detection of high power watermarks with this model is adequate.

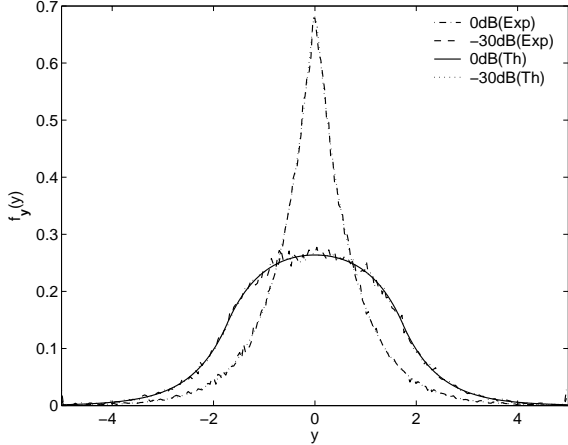
#### 4.2. DC-DM with a Secret Key

In the development of  $\tilde{f}_Y(y)$  for DM, the average pdf was taken over  $f_Y(y|k)$  given by (8) with  $K = 0$  replaced with  $K = k$ . However the same procedure applied to (12) for DC-DM is not so straightforward. This expectation proves difficult to solve, so a different approach is adopted here. Again, assume an initial value of  $K = k$ . Then an exact formulation of  $f(y|k)$  will be developed<sup>13,14</sup> from which the expectation can be taken. Firstly assume that the pdf is dependent on one symbol, i.e. we will examine  $f(y|k, b = 0)$ , without loss of generality. Let the centroids of this quantizer be given by  $i\Delta \forall i \in \mathbb{Z}$ . For any given  $k$  these bins are simply shifted by an amount equal to  $k$ . Then (3) corresponds to the following transformation for the  $i$ th centroid,

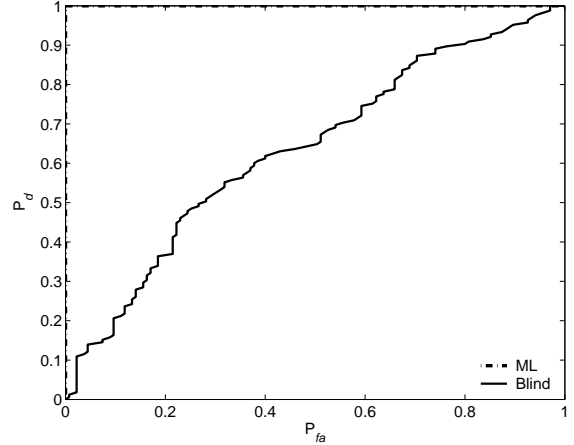
$$y = G_k(x, b = 0) = (1-\alpha)x + \alpha(i\Delta + k), \quad i\Delta + k - \frac{\Delta}{2} \leq x < i\Delta + k + \frac{\Delta}{2}, \quad (13)$$

where  $|G'_k(x, b = 0)|$  is given by  $1-\alpha$ . The only root of (13) is given as  $x = (y - \alpha(i\Delta + k))/(1-\alpha)$ . This gives

$$f(y|k, b = 0) = \frac{1}{1-\alpha} f_X \left( \frac{y - \alpha(i\Delta + k)}{1-\alpha} \right), \quad i\Delta + k - \frac{(1-\alpha)\Delta}{2} \leq y < i\Delta + k + \frac{(1-\alpha)\Delta}{2}. \quad (14)$$



**Figure 4.** Theoretical  $\tilde{f}(y)$  (Th) for DC-DM with  $\alpha = 0.5$  (see (15)) shown alongside experimental histograms (Exp) for host given by (5) with  $\lambda = \sqrt{2}$ . WCRs of 0 dB and -30 dB are shown.



**Figure 5.** Sample ROC curves for DM detection in the case where  $\mathbf{k}$  is known. The WCR is set at -20 dB.

The dependence of  $y$  on the bin can be removed by summing over  $i$ , giving,

$$f(y|k, b=0) = \sum_{i=-\infty}^{\infty} \frac{1}{1-\alpha} f_X \left( \frac{y - \alpha(i\Delta + k)}{1-\alpha} \right).$$

Finally the message dependence can be removed by averaging over the message alphabet. This pdf is then substituted into (7) to give  $\tilde{f}(y)$ . This analysis is outlined in Appendix B. The final result is given as

$$\tilde{f}_Y(y) = \frac{1}{\alpha\Delta} \left[ F_X \left( y + \frac{\alpha\Delta}{2} \right) - F_X \left( y - \frac{\alpha\Delta}{2} \right) \right].$$

This can be seen as the convolution of the host density  $f(x)$  with a uniform density given as  $U \left( -\frac{\alpha\Delta}{2}, \frac{\alpha\Delta}{2} \right)$ , i.e.,

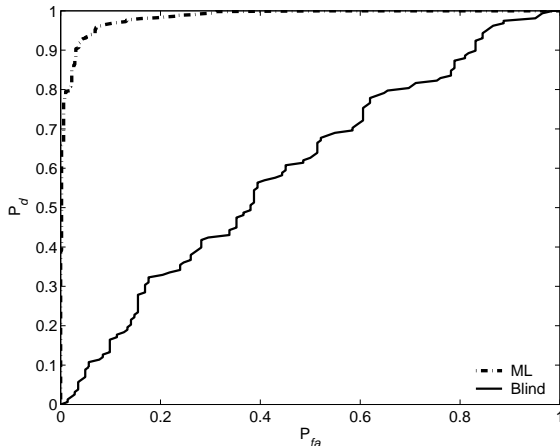
$$\tilde{f}_Y(y) = f_X(y) * U \left( -\frac{\alpha\Delta}{2}, \frac{\alpha\Delta}{2} \right). \quad (15)$$

Some example pdfs are shown in Figure 4. It should be noted that for  $\alpha = 1$  this result is the same as (11), as expected.

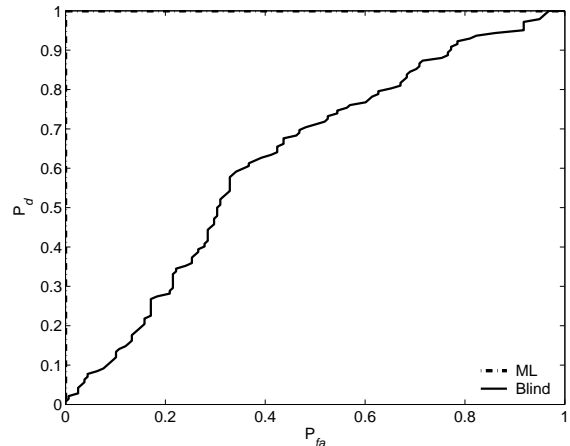
We could think that different values of  $\alpha$  could be exploited by Alice to improve the secrecy of her communication. However, the WCR for DC-DM is set by the product of  $\alpha$  and  $\Delta$  in (4), which indicates that if Alice sets her embedding to be at a certain WCR she has a tradeoff between the two parameters to make. Also, the average stegotext pdf (15) is governed by the product of these same parameters. Therefore, under the approach given by (7), the choice of  $\alpha$  is not important in respect of the statistical transparency of the communication. This can be seen when Figures 2 and 4 are compared. It is apparent the pdfs, seen by Wendy, for both DM and DC-DM are identical at given WCRs. From Wendy's point of view this leads to the somewhat surprising result that the ML test should perform equally well on both DM and DC-DM in the absence of knowledge about  $\mathbf{k}$ .

## 5. SIMULATION RESULTS

The distinction between blind and non-blind steganalysis has been noted in Section 1. Here we show the limits of blind steganalysis as applied to the four embedding techniques analyzed in Sections 3 and 4. For comparative purposes we also show the performance of a popular blind technique.



**Figure 6.** Sample ROC curves for DM detection in the case where no knowledge of  $\mathbf{k}$  is available. The WCR is set at  $-20$  dB.



**Figure 7.** Sample ROC curves for DC-DM detection in the case where  $\mathbf{k}$  is known. The WCR is set at  $-20$  dB with  $\alpha = 0.5$ .

### 5.1. Blind Steganalysis

The principal advantage of blind steganalysis over non-blind is that the test is independent of the particular embedding scheme used. From a practical point of view this is a highly desirable feature as it means that the *a priori* information required by Wendy is minimised. However, to date, there has been no thorough examination of how good these blind techniques can get. One feasible answer to this problem is to compare the ROC curves of a given blind technique with that of the statistically optimal test. To this end we propose an integral measure over the ROC curves. The ROC curves plot the probability of detection,  $P_d$ , against the probability of false alarm,  $P_{fa}$ . As such let the ROC curve of the ML test be represented as  $P_{d_{ML}}(P_{fa})$  and that of the sub-optimal test be  $P_{d_B}(P_{fa})$ , where the subscript “B” refers to “blind”. We then have the following,

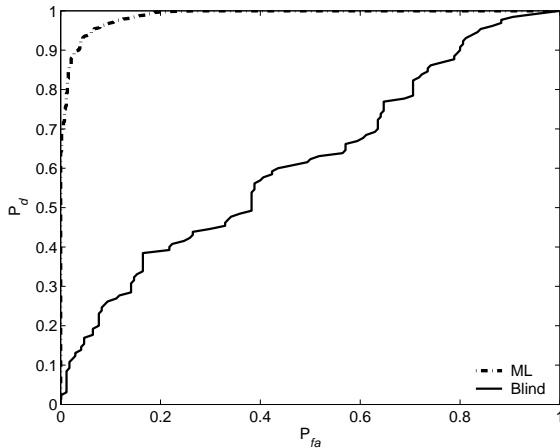
$$\eta \triangleq \int P_{d_{ML}}(P_{fa}) - P_{d_B}(P_{fa}) dP_{fa}. \quad (16)$$

Thus for two tests that are equally good (or poor),  $\eta$  has a value of zero while for two tests which are maximally separated  $\eta$  will be equal to 0.5 (due to the fact that the area under the ML curve will be 1 while the other area will be equal to 0.5). This parameter is non-negative as the optimal test will always outperform or equal the blind test.

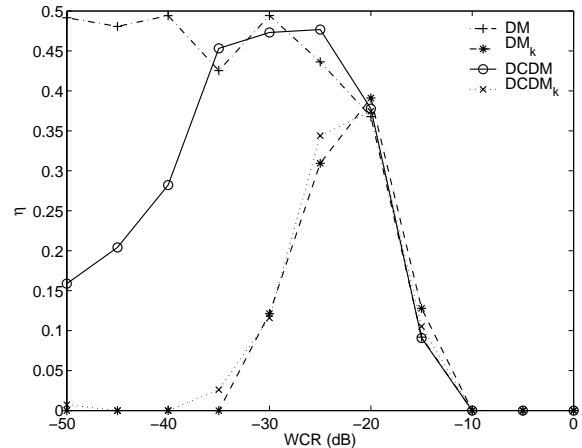
Several blind steganalysis techniques have been proposed in the literature, see e.g.<sup>15</sup> but for our purposes we choose to compare our technique to that of Farid<sup>9</sup> which has shown excellent detection results on a number of embedding techniques. The basis of the technique is to extract a feature vector from images (both stegotexts and coverttexts) using the discrete wavelet transform. These features are then used to train a classifier which in this case is a Fisher linear discriminant (FLD). A random document can be then tested by extracting the feature vector and using this in the trained classifier. The authors have made several improvements to their scheme (see<sup>16</sup> and references therein) but here we will concentrate on the simpler test based on the FLD.

### 5.2. Simulation Conditions

Vectors of iid elements of length 16384 are formed according to (5), corresponding to images of size  $128 \times 128$  pixels, assumed to be represented in the wavelet domain. Subsequently stegotexts are created using the four quantization embedding techniques outlined in Sections 3 and 4 for a wide range of WCRs ( $-50$  dB to 0 dB). For DC-DM, the value of  $\alpha$  is set to 0.5. This completes a set of 2000 images of which 1000 are randomly drawn for testing with equal probabilities for each type of image. For the ML tests the vectors are separately tested and the resulting decision values are used in the formation of the ROC curves. For the blind technique 800 of the 1000 images were first drawn, 400 from each class, and used to train the classifier, the remaining 200 images then being classified. The value of the parameter  $\lambda_n$  in the DM test was empirically set to be  $\sqrt{2} \times 10^4$  which was found to give suitably accurate results.



**Figure 8.** Sample ROC curves for DC-DM detection in the case where no knowledge of  $\mathbf{k}$  is available. The WCR is set at  $-20$  dB with  $\alpha = 0.5$ .



**Figure 9.** The integral parameter  $\eta$ , (see (16)) for the wavelet feature technique, plotted against WCR for DM, DM with a key ( $DM_k$ ), DC-DM and DC-DM with a key ( $DCDM_k$ ). For the DC-DM curves  $\alpha = 0.5$ .

### 5.3. Results

Some sample ROCs are shown in Figures 5, 6, 7 and 8. All the ROCs are given for a WCR =  $-20$  dB. It is immediately clear that the ML test outperforms the blind test for all embedding schemes. It should be noted however that the performance of the blind scheme is not dependent on the particular embedding algorithm used and it remains approximately consistent across the schemes. An examination of the ROCs for the blind steganalysis shows that the decision is random at all WCRs lower than approximately  $-20$  dB. This is not the case for the ML tests however, for which the range that the tests perform well is much larger. It was noted in Section 4.2 that the detection performance on secret key DM and DC-DM should be equal. This is indeed the case as the ROCs for both tests are approximately the same across all the WCRs investigated.

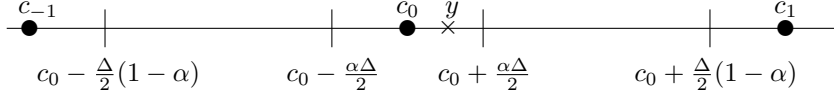
In the sample ROCs shown, it is apparent that the ML tests for  $f(y|k)$  perform better than those for  $\tilde{f}(y)$ . This is to be expected as Wendy has less information in the latter case. However, as the WCR increases this loss of information becomes less important and the performance of the tests with and without  $\mathbf{k}$  converges. This is an important result as it means that the security of high power DC-DM cannot be significantly improved by using a secret key.

In Figure 9 the performance of the tests across the WCRs is presented, the results being condensed using (16). The four plots show that the difference between the ML and blind tests is quite large at intermediate WCRs. For secret key DM and DC-DM  $\eta$  approaches zero as the WCR decreases and the blind test approaches ML. However, the performance of all these tests is poor over this WCR range. More noteworthy is the fact that  $\eta \rightarrow 0$  at high WCRs. This indicates that, regardless of the embedding technique, the blind tests performance tends to ML. At WCR  $\approx -10$  dB the blind test begins to fail and is no longer close to the ML performance.

## 6. CONCLUSION

ML tests based on DC-DM are presented. The particular case of DM ( $\alpha = 1$ ) is treated separately due to the nature of the pdf involved. The case of DM without a key presents difficulties due to the limited support set of the stegotext pdf. Using an approximation to the pdf, this problem is circumvented. For the tests in which no knowledge of the key is available to Wendy the average pdf for the stegotexts is obtained. At given embedding strengths the pdfs for DM and DC-DM are equivalent. The detection of each type of embedding is therefore also equivalent.

Blind steganalytic methods use as little *a priori* information to make decisions as possible. However the limit to the performance of such techniques has not, to date, been thoroughly investigated. Viewing steganalysis as a binary hypothesis testing problem allows these limits to be obtained by obtaining the statistically optimal test



**Figure 10.** The pdf regions for DC-DM with a key dependent dither centered at  $c_0$ .

for a given embedding technique. The performance of any blind technique can then be gauged against the ML tests using an integral measure defined over the ROC curves of the tests.

The capacity of DC-DM can only be achieved through the use of suitable channel codes which, in general, introduce dependencies between samples. These dependencies could be exploited to further improve the detection performance, but this case has not been examined here. Another open problem is that of developing an embedding method that achieves the constrained capacity of the steganography framework.

## APPENDIX A. AVERAGE PDF FOR DM

Substituting (9) into (7) we obtain:

$$\begin{aligned}\tilde{f}(y) &= \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \sum_{i=-\infty}^{\infty} \frac{1}{2} \left( F_X \left( \frac{i\Delta}{2} + k + \frac{\Delta}{2} \right) - F_X \left( \frac{i\Delta}{2} + k - \frac{\Delta}{2} \right) \right) \delta \left( y - \left( \frac{i\Delta}{2} + k \right) \right) \frac{1}{\Delta} dk \\ &= \sum_{i=0}^1 \frac{1}{2\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \left( F_X \left( \frac{i\Delta}{2} + k + \frac{\Delta}{2} \right) - F_X \left( \frac{i\Delta}{2} + k - \frac{\Delta}{2} \right) \right) \delta \left( y - \left( \frac{i\Delta}{2} + k \right) \right) dk.\end{aligned}$$

Now treating each term of the summation separately starting with  $i = 0$ . Denoting this integral as  $I_0$  we have,

$$I_0 = \frac{1}{2\Delta} \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \left( F_X \left( k + \frac{\Delta}{2} \right) - F_X \left( k - \frac{\Delta}{2} \right) \right) \delta(y - k) dk = \frac{1}{2\Delta} \left( F_X \left( y + \frac{\Delta}{2} \right) - F_X \left( y - \frac{\Delta}{2} \right) \right).$$

The case for  $i = 1$  is similar. Summing  $I_0$  and  $I_1$  gives (10).

## APPENDIX B. AVERAGE PDF FOR DC-DM

For clarity assume that the centroid around which the expectation is taken, is denoted  $c_0$  and not 0. From (14) it can be seen that, as  $\alpha$  decreases in value towards zero, the pdf conditioned to  $b = 0$  has a growing support set over  $\mathbb{R}$ . At values of  $\alpha < 0.5$  the pdfs conditioned to  $b = 0$  and  $b = 1$  overlap. This is illustrated in Figure 10 where five separate regions are indicated over the range  $(c_0 - \Delta/2, c_0 + \Delta/2]$ . The ranges immediately around the quantization points are conditioned to a single bit whereas the mid range values have overlapping pdfs from each bit. Assume, for the purposes of this development, that  $\alpha < 0.5$ , without loss of generality. Assume also that  $y \in (c_0 - \alpha\Delta/2, c_0 + \alpha\Delta/2]$ . For this range of  $y$ , there are five ranges of  $k$  over which to integrate. For example if  $k \in (-\Delta/2, y - c_0 - \Delta/2 + \alpha\Delta/2]$  the integration is over  $(c_{-1}, c_{-1} + \alpha\Delta/2]$  in Figure 10. The following results,

$$\begin{aligned}\tilde{f}(y) &= \frac{1}{2\Delta(1-\alpha)} \left\{ \int_{c_0 - \frac{\Delta}{2}}^{y - c_0 - \frac{\Delta}{2} + \frac{\alpha\Delta}{2}} f_X \left( \frac{y - \alpha(c_0 - \frac{\Delta}{2} + k)}{1 - \alpha} \right) dk + \int_{y - c_0 - \frac{\alpha\Delta}{2}}^{y - c_0 + \frac{\alpha\Delta}{2}} f_X \left( \frac{y - \alpha(c_0 + k)}{1 - \alpha} \right) dk \right. \\ &+ \int_{y - c_0 - \frac{\alpha\Delta}{2}}^{y - c_0 - \frac{\alpha\Delta}{2}} \left[ f_X \left( \frac{y - \alpha(c_0 - \frac{\Delta}{2} + k)}{1 - \alpha} \right) + f_X \left( \frac{y - \alpha(c_0 + k)}{1 - \alpha} \right) \right] dk \\ &+ \int_{y - c_0 + \frac{\alpha\Delta}{2}}^{y - c_0 - \frac{\alpha\Delta}{2} + \frac{\Delta}{2}} \left[ f_X \left( \frac{y - \alpha(c_0 + \frac{\Delta}{2} + k)}{1 - \alpha} \right) + f_X \left( \frac{y - \alpha(c_0 + k)}{1 - \alpha} \right) \right] dk \\ &\left. + \int_{y - c_0 + \frac{\Delta}{2} - \frac{\alpha\Delta}{2}}^{c_0 + \frac{\Delta}{2}} f_X \left( \frac{y - \alpha(c_0 + \frac{\Delta}{2} + k)}{1 - \alpha} \right) dk \right\}, \quad y \in (c_0 - \alpha\Delta/2, c_0 + \alpha\Delta/2].\end{aligned}\quad (17)$$

After some algebraic manipulation and translation (17) can be rewritten as the following single integral,

$$\tilde{f}(y) = \frac{1}{\Delta(1-\alpha)} \int_{y-c_0-\frac{\Delta}{2}(1-\alpha)}^{y-c_0+\frac{\Delta}{2}(1-\alpha)} f_X \left( \frac{y-\alpha(c_0+k)}{1-\alpha} \right) dk, \quad y \in (c_0 - \alpha\Delta/2, c_0 + \alpha\Delta/2].$$

Using the substitution,  $u = (y - \alpha(c_0 + k))/(1 - \alpha)$ , this integral solves to give,

$$\tilde{f}(y) = \frac{1}{\alpha\Delta} \left[ F_X \left( y + \frac{\alpha\Delta}{2} \right) - F_X \left( y - \frac{\alpha\Delta}{2} \right) \right], \quad y \in (c_0 - \alpha\Delta/2, c_0 + \alpha\Delta/2].$$

This result is only valid for one region of  $y$  and  $\alpha < 0.5$  but it can be shown that the result holds for all values of  $y \in (c_0 - \Delta/2, c_0 + \Delta/2]$  and  $\alpha$ .

## REFERENCES

1. G. Simmons, "The prisoner's problem and the subliminal channel," in *Advances in Cryptology, Crypto '83*, **20**, pp. 51–67, Plenum Press, 1984.
2. R. Chandramouli, "On information hiding with incomplete information about steganalysis," in *IEEE Intl. Conf. on Image Processing*, pp. 1161–1164, (Singapore), October 2004.
3. B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory* **47**, pp. 1423–1443, May 2001.
4. J. Eggers and B. Girod, *Informed watermarking*, Kluwer Academic Publishers, 2002.
5. K. Sullivan, Z. Bi, U. Madhow, S. Chandrasekaran, and B. Manjunath, "Steganalysis of quantization index modulation data hiding," in *IEEE International Conference on Image Processing*, Oct 2004.
6. T. Cover and J. Thomas, *Elements of Information Theory*, J. Wiley & Sons, 1 ed., 1991.
7. Y. Wang and P. Moulin, "Steganalysis of block structured stegotext," in *Security, Steganography and Watermarking of Multimedia Contents, Proc. Electronic Imaging* **5306**, SPIE, January 2004.
8. P. Guillon, T. Furon, and P. Duhamel, "Applied public-key steganography," in *Security and Watermarking of Multimedia Contents, Proc. Electronic Imaging* **4675**, pp. 38–49, SPIE, January 2002.
9. H. Farid, "Detecting steganographic messages in digital images," in *Report TR2001-412*, 2001. Dartmouth College, Hanover, NH.
10. H. L. Van Trees, *Detection, Estimation and Modulation Theory*, J. Wiley & Sons, 1968.
11. P. Moulin and Y. Wang, "New results on steganographic capacity," in *Proc. CISS Conference*, (Princeton, USA), March 2004.
12. J. Price and M. Rabbani, "Biased reconstruction for JPEG decoding," *Signal Processing Letters* **6**, December 1999.
13. L. Pérez-Freire, F. Pérez-González, and S. Voloshinovskiy, "Revisiting scalar quantization-based data hiding: Exact analysis and results," *IEEE Transactions on Signal Processing*, 2004. submitted.
14. A. Papoulis and S. U. Lillai, *Probability, Random Variables and Stochastic Processes*, Mc Graw Hill, 2002.
15. M. Hogan, G. Silvestre, and N. Hurley, "Performance evaluation of blind steganalysis classifiers," in *Security, Steganography and Watermarking of Multimedia Contents, Proc. Electronic Imaging* **5306**, SPIE, January 2004.
16. S. Lyu and H. Farid, "Steganalysis using color wavelet statistics and one-class support vector machines," in *Security, Steganography and Watermarking of Multimedia Contents, Proc. Electronic Imaging* **5306**, SPIE, January 2004.