TWO PRIVATE, PERCEPTUAL DATA-HIDING GAMES

Anil Kumar Goteti and Pierre Moulin

University of Illinois at Urbana-Champaign
Beckman Inst., Coord. Sci. Lab & ECE Dept.
405 N. Mathews Ave., Urbana, IL 61801, USA
email: {goteti, moulin}@ifp.uiuc.edu

ABSTRACT

Perceptual watermarking methods are designed to be transparent and robust to attacks. A perceptual model based on Just Noticeable Difference levels introduces amplitude constraints on the watermark and the noise generated by an attacker. Two problems are considered in this paper: (1) detection performance for embedding a single bit in n data, and (2) Shannon capacity. In both cases the original host data are known to the receiver. Both problems are formulated as games involving a suitable cost function (Bhattacharyya distance and mutual information, respectively). The watermarker and the attacker design probability distributions in order to respectively maximize and minimize the cost function. The optimal distributions are quite different from the uniform distributions that have been previously used in the watermarking literature.

1. INTRODUCTION

Data-hiding codes designed for applications such as copyright protection and authentication must satisfy two important properties. The embedding process should be *transparent*, in the sense that marking the original host data should not introduce perceptually noticeable degradations. Thus watermarks should be weak signals, whose statistical characteristics depend on perceptual models. The embedding process should also be *robust* against various types of attacks. Thus different watermarks (codewords) should be statistically distinguishable by a receiver who may know little about the attack process. The watermarking literature contains several instances of this problem when signal degradations are measured using models of human perception [1, 2, 3].

One may ask whether/how these designs could be improved. A natural question in this spirit is: What are the fundamental performance limits for watermarking under perceptual distortion models? Recent studies (e.g., [4, 5]) have provided a framework for such analysis, but so far practical solutions have only been obtained for distortion models such as squared-error distortion and Hamming distortion. The goal of this paper is to build on this framework and quantify fundamental performance limits under a simple perceptual image model, based on Watson's work [6].

The papers [4, 5] used a game-theoretic framework to characterize fundamental performance limits. In this framework, one specifies distortion constraints for the watermarker and the attacker and formulates an appropriate cost function to be maximized and minimized, respectively [4, 5]. For a capacity analysis, the appropriate cost function is a certain mutual information (in the *private* game where the host data are known to the receiver) or a difference between two mutual informations (in the *public* game where the host data are unknown to the receiver). In some data hiding applications, only a few bits are to be embedded. Then capacity is not an issue, and error probability is the appropriate cost function. Both problems are considered in this paper when the embedding and attack are subject to perceptual distortion constraints.

2. JND-BASED WATERMARKING

Data can be embedded into images (compressed or uncompressed) using Human Visual System (HVS) models [1]. These models can be formulated in various spatio-frequency domains. For instance, the image can be decomposed into DCT blocks and Watson's model [6] for the HVS can be used to estimate a Just Noticeable Difference (JND) level for each DCT coefficient. Denote a coefficient by $S(u)$, where u represents coefficient coordinates. The message to be embedded is mapped onto a normalized *watermark sequence* $Z(u)$, where each $Z(u) \in [-1, 1]$. Next, let $J(u)$ be the JND coefficient for coefficient $S(u)$. If $|S(u)| > J(u)$ (i.e., the coefficient is significant), the watermarked coefficient $X(u)$ is defined as

$$X(u) = S(u) + J(u)Z(u).$$

WORK SUPPORTED BY NSF GRANTS CCR 00-81268, CCR 02-08809, AND CDA 96-24396.

According to the JND model, the resulting marked image is perceptually indistinguishable from the original.

An attacker who tries to distort the image while keeping the attack transparent can exploit the JND model. Indeed the JND coefficients can be learnt by the attacker from the watermarked image. The attacker can produce an image with DCT coefficients

$$Y(u) = X(u) + W(u)$$

where $W(u)$ is referred to as the attacker's noise. The severity of the attack depends on the size of the $W(u)$'s relative to the JND levels. For instance, the attacker may want to select a noise sequence that satisfies $|W(u)| \leq \alpha J(u)$, where $\alpha \geq 1$ controls the severity of the attack. [1] In this paper, we find the worst-case attack noise distribution for different cost functions under such JND-based constraints. Previous work has used uniform noise distributions [2] to model JPEG quantization noise, but as we shall see, such attacks can be severely suboptimal.

3. DETECTION PROBLEM

We study a generic version of the detection problem. The coefficients $S(u)$ are group into spatio-frequency bands with similar perceptual characteristics and same JND $J(u)$. We consider a length-n host sequence $\mathbf{s} = \{s_i, 1 \leq i \leq n\}$ made of the coefficients in a given band; denote by b the JND for that band. There is one watermark bit to be embedded into the sequence $\mathbf{s}$, producing a marked sequence $\mathbf{x}$. The bit values are equally likely. Next, the attacker adds a noise sequence $\mathbf{w} = \{w_i, 1 \leq i \leq n\}$ to $\mathbf{x}$, giving rise to the degraded sequence $\mathbf{y} = \mathbf{x} + \mathbf{w}$. In our analysis, we assume the samples w_i are independent and identically distributed (i.i.d.) according to some distribution $p(w)$. The attacker limits the magnitude of the noise to a (typically $a > b$).

Under this setup, for any choice of $p(w)$, the optimal signaling scheme is binary antipodal: select an arbitrary signaling waveform $\mathbf{z} = \{z_i, 1 \leq i \leq n\}$, where each $z_i \in \{\pm 1\}$. Then mark the host sequence according to $\mathbf{x} = \mathbf{s} \pm b\mathbf{z}$, depending on the bit value. To simplify subsequent notation (but without loss of generality), we select $z_i \equiv 1$.

We have assumed the receiver knows the sequences $\mathbf{s}$ (private watermarking) and $\mathbf{z}$, so it can also learn the noise distribution $p(w)$ from the sequence $\mathbf{y} - \mathbf{s} = \mathbf{w} \pm b\mathbf{z}$. The rival pdf's for $y_i - s_i$ are shown in Fig. 1. Based on $p(w)$, the detector can implement the optimal likelihood ratio test (comparing the sufficient statistic $\prod_{i=1}^{n} p(y_i - s_i - b) \,/\, p(y_i - s_i + b)$ with an appropriate threshold), incurring a probability of error P_e.

[1] A similar approach is often used to select quantizer step sizes in perceptual signal compression [7].

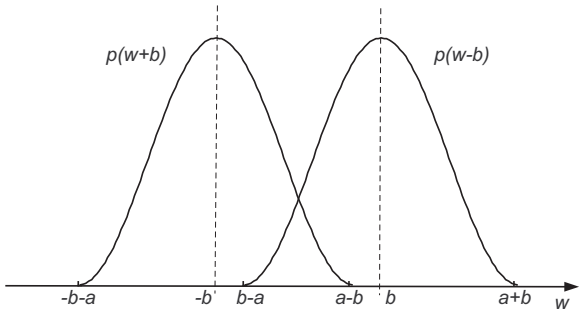


Fig. 1. Rival pdfs for the binary hypothesis testing problem.

Now, we have the following problem at hand – from the attacker's perspective. The attacker would like to design a distribution $p(w)$ which maximizes P_e. Unfortunately this problem appears intractable as soon as n is moderately large. However, the attacker can determine $p(w)$ which minimizes the Chernoff distance $\max_t d_t(p)$, between the binary hypotheses [10], where

$$d_t(p) = -\ln \int_{b-a}^{a-b} p^t(w-b)p^{1-t}(w+b)\, dw.$$

The solution to this problem is a symmetric $p(w)$ and $t = \frac{1}{2}$. Recall that $d_{1/2}(p)$ is the Bhattacharyya distance,

$$B(p) = -\ln \int_{b-a}^{a-b} \sqrt{p(w-b)p(w+b)}\, dw$$

which is convex in p. The error probability satisfies the upper bound $P_e \leq \frac{1}{2}e^{-nB(p)}$; moreover, for large n, we have $-n^{-1}\ln P_e \sim B(p)$. The attacker's problem is to find

$$p^* = \arg\min_p B(p).$$

A notationally simpler version of the problem is where all variables $a, b, z, w \in \{0, \pm\Delta, \pm 2\Delta, \ldots\}$ and $\Delta \leq 1$ can be made arbitrarily small. Assume without significant loss of generality that $\frac{a}{\Delta}$ and $\frac{b}{\Delta}$ are integers. Then, $\mathbf{p}$ is a probability mass function with N mass points $(p_1, p_2, \ldots, p_N)$, where p_1 is the mass at $w = -a$ and p_N is the mass at $w = a$. We have, $N = \frac{2a}{\Delta} + 1$ and if we define $L = \frac{2b}{\Delta} = \frac{b(N-1)}{a}$, our problem now reduces to the convex programming problem of finding a $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} B(\mathbf{p})$$

where,

$$B(\mathbf{p}) = -\ln \sum_{k=1}^{N-L} \sqrt{p_k p_{k+L}}$$

subject to $\sum_{k=1}^{N} p_k = 1$ and $p_k \geq 0$ for $k = 1, 2, \ldots, N$.

We solve this problem numerically using an interior point method [11]. For our simulations, we fix $b = 1$ and vary a. The optimal Bhattacharyya distance depends only on $\lfloor a \rfloor$, the integer part of a (Fig. 2). The staircase effect in the Bhattacharyya distance is more pronounced when a is small. Further, if we use an *uniform* discrete distribution for the noise, the Bhattacharyya distance is $-\ln(1 - \frac{2}{2a+\Delta})$. It decays asymptotically only as a^{-1} whereas the optimal Bhattacharyya distance decays as a^{-2} (Fig. 2).

In Fig. 3 we plot the optimal $p(w)$ for different integer values of a. Small values of a are more realistic in watermarking applications. Two observations should be made here. First, for integer a, the optimal noise is a *lattice noise*, with lattice spacing equal to 2. A similar phenomenon has been observed in related detection problems, under different sets of assumptions [8, 9]. Second, the curves resemble a truncated subsampled Gaussian curve for large a. For non-integer a, the curves are more complicated but they too exhibit the truncated Gaussian nature.

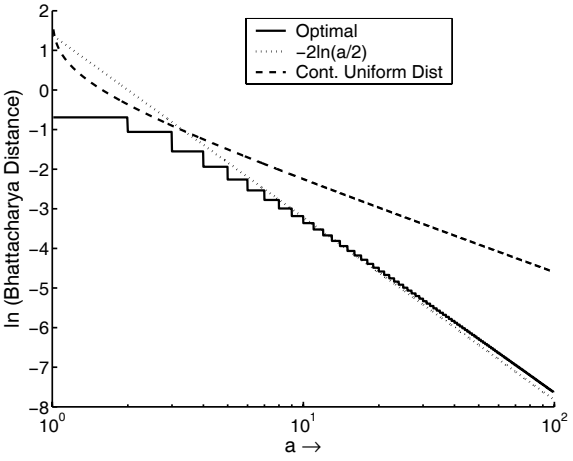
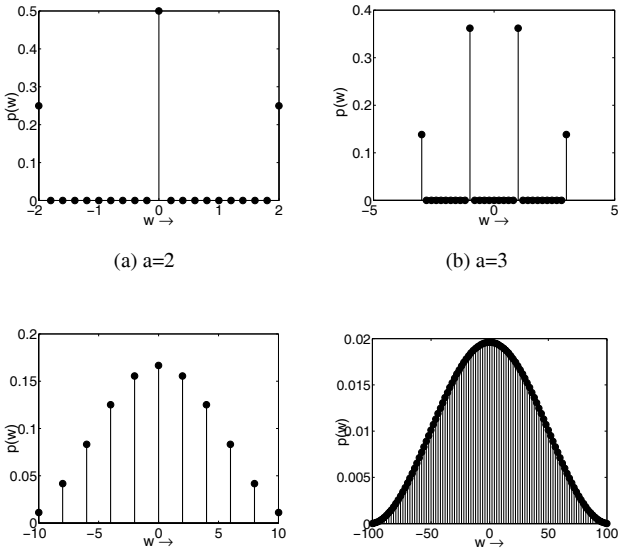


Fig. 2. Optimal Bhattacharyya distance as a function of a.

To illustrate the usefulness of this approach, consider $b = 1$ and $a = 3$ (mildly agressive attacker) and a target $P_e \approx 10^{-3}$. From Fig. 2, we obtain $B(p^*) = 0.2119$. We then select $n = \lceil \frac{-\ln 10^{-3}}{0.2119} \rceil = 33$. The actual P_e based on this value of n is 1.8×10^{-4}, as determined from 10^6 Monte-Carlo experiments. The probability of error for a suboptimal uniform-noise attack would be three orders of magnitude lower : $P_e = 7.72 \times 10^{-7}$.

4. CAPACITY PROBLEM

Our second problem is a capacity problem, in which we wish to reliably embed as many bits as possible in a length-n host sequence $\mathbf{s}$. No statistical model is needed for $\mathbf{s}$. There is an amplitude constraint $|x_i - s_i| \leq b$ *on each sample* $i = 1, 2, \ldots, n$ of the marked sequence $\mathbf{x}$. The sequence



(a) a=2 (b) a=3

(c) a=10 (d) a=100

Fig. 3. Worst case noise pdf $p(w)$ for different values of a using Bhattacharyya distance as the cost function.

$\mathbf{x}$ is attacked with i.i.d. noise $\mathbf{w}$ whose samples have magnitude at most equal to a and are distributed as $p(w)$. The decoder receives $\mathbf{y} = \mathbf{x} + \mathbf{w}$.

Assuming the decoder knows $p(w)$ and the host signal $\mathbf{s}$, we evaluate the capacity C of this channel. We have [4]

$$C = \max_{p(z)} \min_{p(w)} I(z; bz + w) = \min_{p(w)} \max_{p(z)} I(z; bz + w) \quad (1)$$

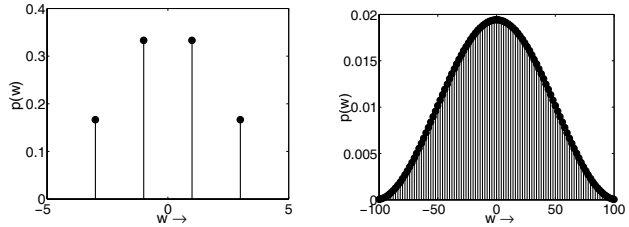
where the support set of $p(z)$ is $[-1, 1]$, and the mutual-information function $I(z; bz + w) = \frac{1}{n} I(\mathbf{z}; \mathbf{y}|\mathbf{s})$ is concave in $p(z)$ and convex in $p(w)$. The second equality in (1) holds because the mutual-information game admits a saddle point.

To evaluate C, we again use numerical optimization methods by discretizing the range $[-1, 1]$ for z and the range $[-a, a]$ for w. If we restrict optimization of $p(z)$ to the binary alphabet $\{\pm 1\}$, then the optimal $p(z)$ is symmetric with equal masses at $-b$ and b, and the worst-case $p(w)$ for large a is a truncated subsampled Gaussian-looking curve (see Fig. 4). The optimal $p(w)$ obtained here are similar, but not identical, to those obtained in Sec. 3. Further, as in Sec. 3, the value of the game (here capacity) depends only on $\lfloor a \rfloor$.

For a quaternary alphabet, i.e., $z \in \{\pm\frac{1}{3}, \pm 1\}$, we see that capacity improvement over the binary case is significant *only for small* a (Fig. 5). For larger a, the optimal $p(z)$ for the quaternary alphabet tends to a symmetric mass distribution at $\pm b$ (binary alphabet). This is typical of problems of transmission over very noisy channels: if the attacker is ag-

gressive, there are few signals we can reliably transmit, and they need to be as much separated as possible. Conversely, when the attack noise is very low, the capacity-achieving distribution $p(z)$ is nearly uniform.

Fig. 5 also shows capacity under uniform $p(w)$, for binary z. Capacity is equal to $1/a$; the suboptimality of uniform $p(w)$ is particularly evident for large a. Finally, note that problems where z is restricted to a small alphabet (binary, etc.) are applicable to problems of data hiding in JPEG and JPEG-2000 images.



(a) a=3 (b) a=100

Fig. 4. Capacity-minimizing noise pdf $p(w)$ for different values of a.

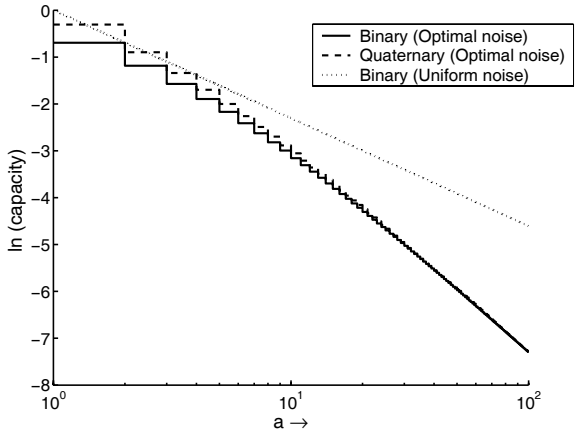


Fig. 5. Log-capacity for binary and quaternary alphabets with varying a.

5. CONCLUSION

In this paper, we have looked at two private watermarking problems where the watermark and noise are amplitude constrained. For the 1-bit transmission problem, binary antipodal signaling is optimal. We found using convex programming techniques, the noise which minimizes the Bhattacharyya distance between the two binary hypotheses. The optimal Bhattacharyya distance depends only on the integer value of the amplitude ratio a/b. For severe attacks, the noise pdf looks like a subsampled truncated Gaussian noise. In the capacity problem, where the objective is to find the capacity-maximizing input distribution and the worst-case attack noise, we found similar results. Future research will apply these results to data hiding in images.

6. REFERENCES

[1] R. B. Wolfgang and C. I. Podilchuk and E. J. Delp, "Perceptual watermarks for digital images and video," *Proc. IEEE*, vol. 87, no. 7, pp. 1108-1126, 1999.

[2] D. Kundur, "Implications for High Capacity Data Hiding in the Presence of Lossy Compression," *Proc. IEEE Int. Conf. on Information Technology: Coding and Computing*, Las Vegas, NV, pp. 16-21, March 2000.

[3] K. Solanki, N. Jacobsen, C. Chandrasekaran, U. Madhow and B. S. Manjunath, "High-Volume Data Hiding in Images: Introducing Perceptual Criteria into Quantization Based Embedding," *Proc. ICASSP*, Orlando, FL, May 2002.

[4] P. Moulin and J. A. O'Sullivan, "Information-Theoretic Analysis of Information Hiding," *IEEE Trans. on Info. Theory*, vol. 49, no. 3, pp. 563-593, 2003.

[5] A. Somekh-Baruch and N. Merhav, "On the Error Exponent and Capacity Games of Private Watermarking Systems," *IEEE Trans. on Info. Theory*, vol. 49, no. 3, pp. 537-562, 2003.

[6] A. B. Watson, "DCT quantization matrices optimized for individual images," *Human Vision, Visual Processing, and Digital Display IV*, vol. SPIE-1913, pp. 202-216, 1993.

[7] N. Jayant, J. Johnston and R. Safranek, "Signal Compression Based Models on Human Perception," *Proc. IEEE*, vol. 81, pp. 1385-1422, 1993.

[8] A. L. McKellips and S. Verdù, "Maximin Performance of Binary-Input Channels with Uncertain Probability Distributions," *IEEE Trans. on Info. Theory*, vol. 44, no. 3, pp. 947-972, 1998.

[9] J. Morris, "On Single-Sample Robust Detection of Known Signals with Additive Unknown-Mean Amplitude-Bounded Random Interference," *IEEE Trans. on Info. Theory*, vol. 26, no. 2, pp. 199-209, 1980.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, 1994.

[11] http://www-neos.mcs.anl.gov/neos/, NEOS.