

# A Framework for Evaluating the Data-Hiding Capacity of Image Sources \*

Pierre Moulin<sup>†</sup> and M. Kivanç Mihçak  
University of Illinois  
Beckman Inst., Coord. Sci. Lab. & ECE Department  
405 N. Mathews Ave., Urbana, IL 61801

May 20, 2000.

Revised, June 18, 2001, and April 29, 2002

## Abstract

An information-theoretic model for image watermarking and data hiding is presented in this paper. Recent theoretical results are used to characterize the fundamental capacity limits of image watermarking and data-hiding systems. Capacity is determined by the statistical model used for the host image, by the distortion constraints on the data hider and the attacker, and by the information available to the data hider, to the attacker, and to the decoder. We consider autoregressive, block-DCT and wavelet statistical models for images and compute data-hiding capacity for compressed and uncompressed host-image sources. Closed-form expressions are obtained under sparse-model approximations. Models for geometric attacks and distortion measures that are invariant to such attacks are considered.

**Keywords** — image watermarking, data hiding, minimax techniques, autoregressive processes, discrete cosine transform, wavelets, image modeling, information theory.

**EDICS Category** : 5—AUTH

---

\*Work presented in part at the IEEE International Conference on Image Processing, Vancouver, Canada, Oct. 2000. This research was supported by NSF grants MIP-97-07633, CCR 00-81268, and CDA 96-24396.

<sup>†</sup>**Corresponding Author**: Tel +1 217 244-8366, fax +1 217 244-8371, email: *moulin@isp.uiuc.edu*.

# 1 Introduction

Data hiding refers to the nearly invisible embedding of information within a host data set such as text, audio, image, or video [1, 2, 3, 4, 5, 6, 7]. There is a broad variety of data one may want to hide in such data sets. In watermarking applications, the hidden data represent authorship information, a time stamp, or copyright information. In steganographic applications, the hidden data are a secret message whose mere presence within the host data set should be undetectable; a classical example is that of a prisoner communicating with the outside world under the supervision of a prison warden. In image database applications, the hidden data could represent a textual description of image features, or some complementary information (words, sound, etc.) about the original scene. In this context, data hiding represents a useful alternative to the construction of a hypermedia document, which may be less convenient to manipulate. A related application is in-band captioning of data such as foreign-language movie subtitles, stock-market data, etc.

In some of these applications, a malicious opponent may apply various data processing operations to interfere with a decoder's ability to recover the hidden data. In other applications, there is no malicious opponent, but various manipulations of the modified data set may result in similar unfortunate effects. Hence there is a need not only to develop ways to embed information in a nearly invisible fashion, but also to do so in a way that is robust against a broad variety of data processing operations, whether they are intentional or not. In this paper, we assume the presence of an attacker that attempts to disrupt the communication of hidden data, subject to certain constraints. We tend to use the terminology *data hiding* over *watermarking*, as the latter often suggests that a very limited amount of data is to be hidden within the data set; whereas our focus is on communicating significant amounts of information to a decoder.

Most research papers in the watermarking and data-hiding literature have focused on novel ways to hide data and to remove hidden data [1, 8, 2, 3, 4, 5, 6, 7]. However, the information-theoretic analysis that describes the fundamental limits of any watermarking or data-hiding system and can ultimately guide the design of efficient watermarking algorithms has been emerging only recently [9, 10, 11, 12]. Information theory yields results that are substantially different from those obtained so far in the data-hiding literature. A typical approach in this literature is to assume that the data-embedding and -decoding functions take a particular form, and that a particular type of attack is used [13, 14, 15]. Restrictions on the embedding and decoding functions can have a severe impact on performance, as evidenced by the large gap to capacity for commonly-used systems [11]. Dually, a system analysis that assumes a suboptimal attack yields overly optimistic results. To derive the fundamental limits of watermarking and data-hiding systems, one should avoid making a priori assumptions about the embedding and decoding functions, analogously to Shannon's analysis of the fundamental limits of communication systems [16]; and one may want to assume an intelligent opponent, as is done in some game-theoretic analyses of jamming systems [17, 18].

The goal of this paper is to develop estimates of data-hiding capacity, or maximum rates of reliable transmission, for host-image sources. To this end, we use several recent theoretical results, which are not yet well known in the data-hiding and image-watermarking communities, and are summarized in three sections. Sec. 2 describes our basic mathematical model; Sec. 4 gives capacity results for Gaussian channels; and Sec. 5 presents results for parallel Gaussian channels, which are used extensively in the remainder of the paper. While the results in Secs. 2, 4 and 5 apply to abstract probabilistic models, application of these concepts to image processing models presents substantial challenges. One needs to develop suitable models for attack channels, for distortion metrics, and for image statistics. The class of attack channels considered is described in Sec. 3. In Sec. 6, we use the squared-error distortion metric and develop capacity expressions for images under autoregressive, block-DCT (discrete cosine transform) and more advanced inhomogeneous wavelet models. The role of sparsity of the host image is identified. This theme is also underlying in Sec. 7, which analyzes the effects of host-image compression on data-hiding capacity. In Sec. 8, we reexamine capacity under weighted squared-error metrics and under distortion metrics that do not penalize geometric attacks. Conclusions are presented in Sec. 9.

**Notation.** We use capital letters to denote random variables, small letters to denote their individual values, calligraphic fonts to denote sets, and a superscript  $N$  to denote length- $N$  vectors.

## 2 Basic Mathematical Model

A theory has recently been developed to establish the fundamental limits of the fairly general data hiding problem depicted in Fig. 1 [11, 19]. A message  $M$  is to be communicated to a decoder. The message is embedded in a length- $N$  sequence  $S^N = (S_1, \dots, S_N)$  termed *host data set*, typically data from an host image, video, or audio signal. The embedding is done using side information (e.g., a cryptographic key)  $K^N = (K_1, \dots, K_N)$  that is also available at the decoder. The resulting *watermarked data* or *composite data*  $X^N = (X_1, \dots, X_N)$  are subject to *attacks* that attempt to remove any trace of  $M$  from  $X^N$ . The data-hiding process should be *transparent*:  $X^N$  should be similar to  $S^N$ , according to a suitable distortion measure. The system should also be *robust*: the hidden message should survive the application of any attack (within a certain class) to  $X^N$ . There is a limit on the amount of distortion that an attacker is willing to introduce.

This system can be analyzed by defining a statistical model for  $M, S^N$  and  $K^N$ , a distortion function, specifying constraints on the admissible distortion levels for the data hider and the attacker, and specifying the information available to all parties. Then one can seek the *maximum rate of reliable transmission* for  $M$ , over *any* possible data-hiding strategy and *any* attack that satisfy the specified constraints.

The statistical model for  $(M, S^N, K^N)$  is as follows.

Fig. 1  
goes  
here.

- In typical image data-hiding problems, each datum  $S_i$  is a block of data or transform data from an host image. For instance,  $S_i$  could be a single pixel value, an  $8 \times 8$  block of DCT coefficients, or a subtree of wavelet coefficients. These coefficients could be discrete-valued (following quantization) or real-valued. The *host data*  $S^N = (S_1, \dots, S_N)$  is a set of independent and identically distributed (i.i.d.) samples from a probability mass function <sup>1</sup>  $p(s), s \in \mathcal{X}$ . For data hiding in a  $512 \times 512$  image where  $S$  is an  $8 \times 8$  block of DCT coefficients, we would have  $N = 4096$ .
- The individual letters  $K_i$  of the side information  $K^N$  are i.i.d.  $p(k), k \in \mathcal{K}$ . The side information  $K^N$  potentially plays two roles. First, provide a source of randomness (cryptographic key) that is known to the decoder and enables the use of randomized codes. Second, provide side information about  $S$  to the decoder. The dependencies between  $S$  and  $K$  are modeled using a joint distribution  $p(s, k)$ . For instance,  $S$  may be a function of  $K$ , meaning that  $S$  is fully available at the decoder; this scenario is sometimes called *private watermarking*. In blind watermarking (*public watermarking*) applications,  $K^N$  does not convey any information about  $S^N$ , so the decoder does not know the host signal. In some systems, no side information is used at all.
- The message  $M$  of interest is uniformly distributed over the message set  $\mathcal{M}$  and is to be reliably transmitted to the decoder.  $M$  is independent of  $(S, K)$ .

The embedding and attack are subject to distortion constraints. Consider a nonnegative, bounded distortion function  $d(\cdot, \cdot)$  between elements of  $\mathcal{X}$ . The distortion function is extended to a distortion on  $N$ -vectors  $x^N$  by

$$d^N(x^N, y^N) = \frac{1}{N} \sum_{k=1}^N d(x_k, y_k). \quad (1)$$

A *length- $N$  data-hiding code subject to distortion*  $D_1$  is defined as a triple  $(\mathcal{M}, f_N, \phi_N)$ , where the message set  $\mathcal{M}$  is defined as above, and

- $f_N$  is the encoder mapping a sequence  $s^N$ , a message  $m$ , and a key  $k^N$  to a sequence  $x^N = f_N(s^N, m, k^N) \in \mathcal{X}^N$ . This mapping is subject to the expected-distortion constraint

$$Ed^N(S^N, X^N) \leq D_1; \quad (2)$$

- $\phi_N$  is the decoder mapping the received sequence  $y^N$  and the key  $k^N$  to a decoded message  $\hat{m} = \phi_N(y^N, k^N) \in \mathcal{M}$ .

---

<sup>1</sup>Assume here that  $\mathcal{X}$  is a discrete set.

An *attack channel with memory*, subject to distortion  $D_2$ , is defined as a sequence of conditional probability mass functions  $A^N(y^N|x^N)$  from  $\mathcal{X}^N$  to  $\mathcal{X}^N$ , such that

$$Ed^N(S^N, Y^N) \leq D_2, \quad N \geq 1. \quad (3)$$

This is a constraint on the expected distortion with respect to the host signal  $S^N$  that the attacker is willing to introduce. Related versions of this problem include the case of expected distortions relative to the watermarked image  $X^N$  [11] and the use of pointwise distortions [20, 11]. Deterministic attacks are a special case of (3).

A rate  $R = \frac{1}{N} \log |\mathcal{M}|$  is said to be achievable for distortions  $(D_1, D_2)$ , if there is a sequence of codes  $(\mathcal{M}, f_N, \phi_N)$  subject to distortion  $D_1$ , with respective rates  $R_N > R$ , such that the probability of error  $P_{e,N} = Pr[\hat{M} \neq M]$  tends to zero as  $N \rightarrow \infty$ , for any attack subject to distortion  $D_2$ . The *data-hiding capacity*  $C(D_1, D_2)$  is then defined as the supremum of all achievable rates for distortions  $(D_1, D_2)$ .

### 3 Modeling of Attack Channel

The class of all attack channels  $A^N(y^N|x^N)$  that satisfy distortion constraints such as (3) is extremely large. To develop a capacity analysis, it is useful to first consider restricted classes of attack channels which include many attack channels used in practice. Developing a capacity analysis for the broadest possible class of attack channels is still an open problem, analogous to the study of arbitrarily varying channels in classical communication problems [21, 22, 18]. Capacity can be zero in some cases [20, 21, 22, 18].

In Secs. 4–7, we first consider memoryless, time-invariant attack channels of the form  $A^N(y^N|x^N) = \prod_{i=1}^N A(y_i|x_i)$ . Such attack channels include addition of i.i.d. noise (in case  $\mathcal{X}$  is a field), erasures, and JPEG-like <sup>2</sup> attacks (in case  $X$  is a block of  $8 \times 8$  pixels). It has been shown that memoryless attacks are optimal in a qualified sense [10, 11]. A heuristic explanation for this effectiveness is that *memoryless attacks introduce more randomness than attacks with memory*.

In Sec. 8, we also consider attacks that have been difficult to cope with in practice, such as geometric attacks on images. A geometric attack, for instance, can be modeled as a global warping operation of the form  $y^N = W(x^N, \theta)$ , where  $\theta$  is a parameter taking values in a set  $\Theta$ . The mappings  $W(\cdot, \theta)$  and  $W(x^N, \cdot)$  are *invertible* for all  $\theta \in \Theta$  and for all  $x^N \in \mathcal{X}^N$ , respectively<sup>3</sup>. Typically  $\theta$  would be a shifting, scaling or rotation parameter, or any set of parameters representing

<sup>2</sup>Because DC coefficients of each block are encoded using a predictive technique, JPEG attacks are not memoryless but “almost” memoryless.

<sup>3</sup>This model is generally valid for images defined over a compact domain. For digital images, inverting the mapping  $W(\cdot, \theta)$  implies an error (due to the limited accuracy of image interpolation algorithms) and so our model is only an approximation.

a more complicated geometric warping operation. Even if the warping function  $W$  used by the attacker is known, the fact that the decoder may be unable to identify  $\theta$  *apparently* makes such attacks very effective. While this is a commonly-held belief, a recent theoretical result shows that an optimal encoding/decoding system should be able to do very well even if  $\theta$  is unknown [11, Prop. 6.6]. Specifically, consider warping attacks of the form above, applied to individual blocks of  $L$  data (rather than one single block of  $N$  data), where  $\theta$  is a random variable taking independent values for each block ( $\theta$  is i.i.d.  $p(\theta), \theta \in \Theta$ ).<sup>4</sup> Then the data-hiding capacity  $C$  satisfies the lower and upper bounds

$$C^* - \frac{1}{L}H(\theta|Y^L, K^L) \leq C \leq C^*, \quad (4)$$

where  $C^*$  (typically a large number) is the capacity obtained *in the absence of any attack* or equivalently, obtained using a decoder that *knows*  $\theta$ , and  $H(\theta|Y^L, K^L) \leq H(\theta)$  is the conditional entropy of  $\theta$  given  $Y^L, K^L$ . Hence the loss in capacity is at most  $\frac{1}{L}H(\theta)$ , which tends to zero for large  $L$ . The worst-case scenario  $H(\theta|Y^L, K^L) = H(\theta)$  occurs when the data convey no information about  $\theta$ . The basic reason why  $C \approx C^*$  is that *invertible attacks that introduce little randomness are ineffective against an optimal decoder*.

The mathematical model  $y^N = W(x^N, \theta), \theta \in \Theta$  also applies to attacks such as *tone-scale attacks*, which modify (e.g., multiply) pixel intensities using a slowly varying function parameterized by  $\theta$ . Other attacks that depend on a few parameters can be modeled similarly. What is important is not so much the specific form of  $W(x^N, \theta)$ , but its basic mathematical properties: invertibility and dependence on a low-dimensional parameter  $\theta$ . However, to make our presentation more concrete, we will refer to  $y^N = W(x^N, \theta)$  as a warping attack.

Of course classical random attacks and warping attacks can be combined. This can be done as shown in Fig. 2, where the attack channel is the cascade of a memoryless channel  $A(z|x)$ , and a global warping operation  $y^N = W(z^N, \theta), \theta \in \Theta$ . A result similar to (4) applies to the blockwise version of this problem, namely, capacity is nearly the same as if the warping attack did not take place [11].

**Channel Identification.** The problem of channel identification consists of estimating the attack channel from the data  $y^N$  and the key  $k^N$ . The channel may or may not be identifiable. If the channel is identifiable, the decoder can reliably estimate it (for large  $N$ ) and perform asymptotically as well as a decoder that knows the attack channel. To ensure that the channel is identifiable, one may have to transmit channel identification codes, e.g., embed a synchronization pattern in the host signal  $S^N$  [23]. In some cases, the data-hiding codes themselves may be good identification codes.

---

<sup>4</sup>Such blockwise warping attacks are a variation on the theme of global warping attacks. This variation results in a blockwise memoryless attack, which makes information-theoretic asymptotic techniques applicable. This is the main condition under which [11, Prop. 6.6] holds.

Fig. 2  
goes  
here.

## 4 Data-Hiding Capacity

In Sec. 2, we have given an operational definition of data-hiding capacity. Recent research has shown that this capacity is the value of a mutual-information game between the data hider and the attacker. For memoryless attack channels, capacity takes a form similar to the capacity derived by Gel'fand and Pinsker twenty years ago for a communication problem with a fixed noisy channel, and side information at the encoder [24]. The main result is stated in [11, Thm 3.3] [19]. The payoff function in the data-hiding game is a difference between two mutual informations.

**Gaussian Channels.** The case of Gaussian  $S$  and squared-error distortion measure  $d(x, y)$  is of considerable interest, because these models are useful in image processing, explicit solutions can be derived, and the results give *upper bounds* on capacity for non-Gaussian  $S$  [11, 25, 26]. Here we have  $\mathcal{X} = \mathbb{R}$ ,  $d(x, y) = (x - y)^2$ , and  $S \sim \mathcal{N}(0, \sigma^2)$ , meaning that  $S$  follows a Gaussian distribution with mean 0 and variance  $\sigma^2$ . Remarkably, *the data-hiding capacity is the same for both blind and nonblind data-hiding problems*. Under the distortion constraints (2) and (3), we obtain [25]

$$C = \Gamma(\sigma^2, D_1, D_2) \triangleq \begin{cases} \frac{1}{2} \log \left( 1 + \frac{D_1}{D} \right) & : \text{if } D_1 < D_2 < \sigma^2 \\ 0 & : \text{if } D_2 \geq \sigma^2 \end{cases} \quad (5)$$

where  $D \triangleq \sigma^2 \frac{D_2 - D_1}{\sigma^2 - D_2}$ . The optimal attack is the Gaussian test channel from rate-distortion theory. In the case of small distortions ( $\sigma^2 \gg D_1, D_2$ ), which is common in practical data-hiding applications, we have  $D \sim D_2 - D_1$  and  $C \sim \frac{1}{2} \log \left( 1 + \frac{D_1}{D_2 - D_1} \right)$ , i.e., the capacity expression is *asymptotically independent* of  $\sigma^2$ .

**Achievability of Capacity Bound.** In principle, the capacity can be approached using a *random binning* coding technique [27, 28, 24, 11]. Research on structured data-hiding codes that approach this capacity bound is currently underway [9, 29, 30, 31]. The optimal decoder for Gaussian channels is a certain minimum-distance decoder.

## 5 Parallel Gaussian Channels

In this section, we review recent results from [25] on data-hiding capacity under MSE constraints; these results are essential to understanding the remainder of this paper. We use 1-D notation for simplicity. A simple extension of this setup to the case of weighted MSE constraints is presented in Sec. 8.1. Referring to Fig. 3, consider a decomposition of the real-valued signal (image)  $S(n), 1 \leq n \leq N$  into  $K$  channels using a multirate transform  $\mathbf{T}$ . The signal  $S_k(n), 1 \leq n \leq N_k$  in channel  $k$  is termed subsignal and has a total of  $N_k$  samples. The mapping  $\mathbf{T}$  from  $S$  to the subsignals  $\{S_k\}$  is one-to-one but need not be linear, so the multirate transform is not necessarily a conventional linear transform such as a subband transform or a block-DCT transform.

Fig. 3  
goes  
here.

In our parallel-Gaussian model, the subsignals  $\{S_k\}$  are independent, and the samples  $S_k(n), 1 \leq n \leq N_k$  are i.i.d.  $\mathcal{N}(0, \sigma_k^2)$ . It is assumed that  $\sum_{k=1}^K r_k = 1$ , where  $r_k = N_k/N$  is the inverse subsampling factor in channel  $k$ .

Define the per-sample distortion between two signals  $x^N$  and  $x'^N \in \mathbb{R}^N$  as

$$d^N(x^N, x'^N) = \frac{1}{N} \sum_{k=1}^K \sum_{n=1}^{N_k} |x_k(n) - x'_k(n)|^2. \quad (6)$$

If  $\mathbf{T}$  is unitary,  $d^N$  is the normalized Euclidean distance in  $\mathbb{R}^N$ . Consider the following distortion constraints. The data hider designs the data-hiding scheme so that  $Ed(S, X) \leq D_1$ , and the attacker designs the attack so that  $Ed(S, Y) \leq D_2$ . What is the data-hiding capacity under these constraints?

### 5.1 Capacity for Parallel Gaussian Sources

As in the Gaussian data-hiding game, the same value of capacity is obtained whether or not  $S^N$  is available at the decoder [25]. Let  $d_{1k}$  and  $d_{2k}$  be the distortions introduced in channel  $k$  by the data hider and the attacker, respectively. The optimal data-hiding and attack strategies are memoryless (as in the single-channel case), and the channels are also independent. The solution is shown in Fig. 4. The optimal data-hiding and attack strategies in each channel are those for a single Gaussian channel with host-signal variance  $\sigma_k^2$  and squared-error distortions  $d_{1k}$  and  $d_{2k}$  for the data hider and the attacker, respectively. The allocation of powers  $d_1 = \{d_{1k}\}$  and  $d_2 = \{d_{2k}\}$  between channels, satisfies the overall distortion constraints

$$\sum_{k=1}^K r_k d_{1k} \leq D_1, \quad (7)$$

$$\sum_{k=1}^K r_k d_{2k} \leq D_2, \quad (8)$$

and the  $3K$  inequality constraints

$$0 \leq d_{1k} \quad (9)$$

$$d_{1k} \leq d_{2k} \quad (10)$$

$$d_{2k} \leq \sigma_k^2, \quad (11)$$

for  $1 \leq k \leq K$ .

The equilibrium of the data-hiding game is the solution of a constrained optimization problem involving optimal allocation of resources by the data hider and by the attacker (resp. signal and noise power) to the parallel channels. This is reminiscent of optimal power-allocation problems for parallel Gaussian channels in rate-distortion theory [28, Ch. 13.3.3] and in channel coding [28,

Fig. 4  
goes  
here.

Ch. 10.4]. However, the data-hiding problem is more involved, because there is a maxmin game (as opposed to a simple minimization or maximization as in the aforementioned problems). The main result is stated in Theorem 1 below <sup>5</sup>. Using the notation of (5), let  $\Gamma(\sigma_k^2, d_{1k}, d_{2k})$  be the capacity of channel  $k$  when  $E|S_k(n) - X_k(n)|^2 = d_{1k}$  and  $E|Y_K(n) - X_k(n)|^2 = d_{2k}$ .

**Theorem 1** [25] *The data-hiding capacity for parallel Gaussian channels is equal to*

$$C = \max_{d_1} \min_{d_2} \sum_{k=1}^K r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k}), \quad (12)$$

where the maximization and minimization are subject to the  $3K + 2$  constraints above. If  $\sum_{k=1}^K r_k \sigma_k^2 \leq D_2$ , the optimal attack is  $d_{2k} = \sigma_k^2 \forall k$ ,  $d_1$  is arbitrary, and  $C = 0$ . If  $\sum_{k=1}^K r_k \sigma_k^2 > D_2$ , the data-hiding game admits a unique solution. For each  $1 \leq k \leq K$ , the optimal  $d_{1k}$  and  $d_{2k}$  are zero if  $\sigma_k^2 = 0$ ; otherwise  $d_{1k}$  is the unique root of the third-order polynomial

$$p_k(d_{1k}) = \frac{1}{\sigma_k^4} d_{1k}^3 + d_{1k}^2 \left( \frac{1}{\sigma_k^4} \frac{1}{\lambda_1 + \lambda_2} - \frac{2}{\sigma_k^2} \right) + d_{1k} \left( \frac{1}{\sigma_k^4} \frac{1}{4\lambda_1(\lambda_1 + \lambda_2)} + \frac{1}{\sigma_k^2} \frac{\lambda_2 - 2\lambda_1}{2\lambda_1(\lambda_1 + \lambda_2)} + 1 \right) - \frac{\lambda_2}{2\lambda_1(\lambda_1 + \lambda_2)} \quad (13)$$

in the open interval  $\left(0, \frac{2\lambda_2\sigma_k^4}{1+2\lambda_2\sigma_k^2}\right)$ . The optimal  $d_{2k}$  is given by

$$d_{2k} = \frac{d_{1k}}{2} + \sqrt{\frac{d_{1k}^2}{4} + \frac{d_{1k}}{2\lambda_2}}. \quad (14)$$

The Lagrange multipliers  $\lambda_1 < 0$  and  $\lambda_2 > 0$  achieve  $C = \max_{\lambda_2 > 0} \min_{\lambda_1 < 0} r(\lambda_1, \lambda_2)$ . The distortion constraints (7) and (8) are satisfied with equality.

The maxmin optimization problem above can be converted into a convex Lagrangian optimization problem [25]. The execution time of the algorithm is approximately two seconds on a Sparc 20 workstation for problems involving 100 channels.

### Properties of Solution

1. (*Low power allocation to weak channels*). The capacity  $C_k \triangleq \Gamma(\sigma_k^2, d_{1k}, d_{2k})$  for channel  $k$  tends to zero as  $\sigma_k^2 \rightarrow 0$ . The power allocations  $d_{1k}$  and  $d_{2k}$  also tend to zero, and satisfy the asymptotic expressions  $d_{1k} \sim 2\lambda_2\sigma_k^4$  and  $d_{2k} \sim \sigma_k^2 + 2\lambda_1\sigma_k^4$ . Moreover,  $C_k \sim -2\lambda_1\lambda_2\sigma_k^4$ . As expected, the data hider should allocate fewer resources to weak channels.
2. (*Uniform power allocation to strong channels*). If  $\sigma_k^2 \rightarrow \infty$ , then (13) shows that the optimal  $d_{1k}$  depends only on  $\lambda_1$  and  $\lambda_2$ . Equations (14) and (5) respectively imply that the same holds for  $d_{2k}$  and  $C_k$ . These asymptotic expressions are the same for all  $k$ , provided that  $\sigma_k^2 \rightarrow \infty$ .

---

<sup>5</sup>Theorem 1 can be extended to the case of stationary, Gaussian host-signals with bounded spectral density by replacing sums with integrals over the frequency domain [25]. Capacity can then be computed by using a discretization technique and the algorithm mentioned below Theorem 1.

## 5.2 Spike Models

Recently Weidmann and Vetterli have developed rate-distortion bounds for still image compression, under a so-called *spike model* which captures the sparsity of wavelet image representations [32]. It appears that this model is very useful in data hiding as well. Under a spike model, there are two types of channels: those with large variance, say  $\sigma_k^2 \gg D_1, D_2$ , and those with low variance,  $\sigma_k^2 \ll D_1, D_2$ . The signal components are independent. Assume that  $\sigma_k^2 \gg D_1, D_2$  for  $1 \leq k \leq K^*$  and  $\sigma_k^2 \ll D_1, D_2$  for  $K^* < k \leq K$ . Then it follows from Properties 1 and 2 that

$$C = \frac{1}{2} r^* \log \left( 1 + \frac{D_1}{D_2 - D_1} \right), \quad (15)$$

where  $r^* = \sum_{k=1}^{K^*} r_k \in [0, 1]$  is the fraction of strong signal components. In conclusion, for spike models:

1. The optimal power allocations by the data hider and the attacker are independent of the signal variances  $\{\sigma_k^2\}$  in the strong channels, provided that these variances are large relative to  $D_1$  and  $D_2$ .
2. The optimal data-hiding strategy equalizes the power among the strong channels, and likewise, the optimal attack strategy equalizes the noise power among strong channels. Negligible power is allocated to weak channels <sup>6</sup>.
3. The (per-sample) capacity  $C_k$  is the same for all strong channels and is negligible for the weak channels. The capacities  $\{C_k\}$  in the strong channels and  $C = \sum_k r_k C_k$  depend only on the distortion levels  $D_1$  and  $D_2$ , and not on the variances  $\sigma_k^2$ .

## 5.3 Upper Bounds on Capacity

Parallel Gaussian models are useful in that they are reasonably tractable and provide capacity expressions for realistic signal models. Interestingly, they also provide upper bounds on capacity if the actual distribution of  $S$  differs from the model. For instance, any correlation between subsignals  $S_k$  would decrease capacity, and so would any deviation from a Gaussian distribution with the same second-order statistics [25]. For non-Gaussian  $S$ , the Gaussian upper bound (12) on capacity is asymptotically tight as  $D_1$  and  $D_2$  tend to zero, and in this case the capacity-achieving distributions are the same as in the parallel-Gaussian case. A fundamental implication of this result is that *the exact distribution of the source plays only a second-order effect in a small-distortion scenario*.

---

<sup>6</sup>As is done in most data-hiding schemes used in current practice.

## 6 Data-Hiding Capacity for Typical Image Sources

Our goal now is to apply the theory above to image processing. The main difficulties are to formulate tractable but realistic statistical models for the source, and to formulate appropriate distortion constraints on the data hider and the attacker. From there, we derive expressions for the data-hiding capacity. The basic model of Sec. 2 suggests that the data-hiding problem should be formulated in a domain where signal components are approximately i.i.d. and the distortion measure is approximately additive. This suggests the use of block-DCT or wavelet transforms, as mentioned earlier, and models of the form given in Fig. 3. The data-hiding capabilities of the resulting signal components plays a central role in the analysis. In particular, DCT and wavelet representations of images are typically *sparse*, meaning that only a small number of components are able to convey significant hidden information. If compressed images are to be watermarked, many transform coefficients are already zero, and a sparse model for the host image becomes even more appropriate.

### 6.1 AR(1) Models

We begin our analysis with a classical (but simplistic) AR(1) Gaussian model for images. Assume the host-image source is a separable AR(1) Gaussian process [33, Sec. 6.9]. The distribution of this process is parameterized by four quantities: mean  $\mu$ , variance  $\sigma^2$ , and horizontal and vertical correlation coefficients  $\rho_x$  and  $\rho_y$ , respectively. The decoder is assumed to know a priori (or be able to learn from the data  $y^N, k^N$ ) the values of all four source parameters. The 2-D spectral density of  $S$  is given by

$$S(f_x, f_y) = \frac{\sigma^2(1 - \rho_x^2)(1 - \rho_y^2)}{|1 - \rho_x e^{-j2\pi f_x}|^2 |1 - \rho_y e^{-j2\pi f_y}|^2}, \quad -\frac{1}{2} \leq f_x, f_y \leq \frac{1}{2}.$$

We can compute data-hiding capacity estimates for image sources characterized by different values of  $\rho_x$ ,  $\rho_y$ , and  $\sigma^2$  (the value of  $\mu$  has no bearing on capacity). We chose parameters that are representative of images such as *Lena*. We let  $\rho_x = \rho_y = 0.95$ ,  $\sigma^2 = 2500$ , and fix  $D_1 = 10$ , corresponding to a subjectively determined just-noticeable visibility threshold for a reference i.i.d. white noise pattern. Fig. 5a show the original  $512 \times 512$  *Lena* image. Fig. 5b and c (resp. d and e) show the output of the data-hiding channel and the attack channel respectively, using the optimal power allocations derived for  $D_1 = 10$  and  $D_2 = 20$  (resp. for  $D_1 = 10$  and  $D_2 = 50$ ). We let  $D_2$  vary between 10 and 50.

The capacities were computed using the numerical algorithm mentioned in Sec. 5.1 (see footnote therein) with a uniform discretization of the  $\log S(f_x, f_y)$  range into  $K = 256$  channels. Fig. 6 shows capacity as a function of  $D_2/D_1$ . Considering values of  $D_2$  that are representative of mild and strong distortions that an attacker may introduce, we obtain  $C = 0.1897$  and  $C = 0.0330$  bit/pixel

when  $D_2 = 20$  and  $D_2 = 50$ , respectively. Fig. 6 also shows capacity computed using spike approximations with several values of the threshold that separates “strong” from “weak” channels. For this experiment and many others, we have found a threshold of  $2D_2$  to be satisfactory in the sense that the approximation to  $C$  is quite accurate (except for very low values of  $D_2/D_1 - 1$ ).

Fig. 5

The capacity per sample in each channel is displayed in Fig. 7. This graph illustrates the two properties mentioned in Sec. 5.1: there is a saturation of the capacity  $C_k$  in a given channel when the variance  $\sigma_k^2$  in that channel increases, and  $C_k$  is proportional to  $\sigma_k^2$  for small  $\sigma_k^2$ .

goes here.

Fig. 6

goes here.

The influence of the correlation parameter on  $C$  is illustrated in Fig. 8, where  $\rho_x = \rho_y$  varies between 0 and 1. For relatively low values of  $\rho_x = \rho_y$  (say less than 0.8), capacity is essentially the same as in the i.i.d. Gaussian case:  $C \approx \frac{1}{2} \log(1 + D_1/(D_2 - D_1))$ . As  $\rho_x = \rho_y$  approaches 1, capacity tends to zero, as more and more channels (frequencies) become weak and hence are unable to hide significant information.

Fig. 7

goes here.

Fig. 8

goes here.

## 6.2 Block-DCT Models

An  $8 \times 8$  block DCT gives rise to an image representation that consists of 64 equal-size channels<sup>7</sup>. Capacity estimates can be computed from Theorem 1, assuming that the data in each channel are i.i.d. Gaussian, and that the channels are independent. The independence assumption is a simplification, but is reasonable when applied to the 63 AC coefficients. The independence assumption applied to the DC channel is more crude. JPEG, for instance, uses a simple predictive model to capture correlations between DC coefficients of adjacent blocks. Hence the analysis below could be refined by using a dependent-process model for the sequence of DC coefficients. Finally, the actual distribution of the coefficients is more heavy-tailed than a Gaussian; Laplacian models are often used for DCT coefficients.

Fig. 9 shows the capacity obtained using a typical set of channel variances  $\{\sigma_k^2\}$  (solid curve in Fig. 15). These variances were evaluated for the  $512 \times 512$  *Lena* image. Also note that the spike approximation is quite accurate here too. Fig. 10 shows the optimal power allocations  $\{d_{1k}\}$  and  $\{d_{2k}\}$  between channels. The two asymptotic regimes discussed in Sec. 5.1 are clearly seen on these plots. First, a saturation phenomenon occurs for large  $\sigma_k^2$ . Second, the  $\ln d_{1k}$  and  $\ln d_{2k}$  vs  $\ln \sigma_k^2$  curves tend to straight lines with slopes equal to 2 and 1, respectively, as  $\sigma_k^2 \rightarrow 0$ .

Fig. 9

goes here.

In case the block-DCT coefficients are not independent and/or do not follow a Gaussian distribution, the capacity results above (using the correct values of  $\{\sigma_k^2\}$ ) are strict upper bounds on actual capacity. This follows directly from the properties mentioned in Sec. 5.3. Note that a similar conclusion may not be drawn from the study of autoregressive models in Sec. 6.1: the

Fig. 10

goes here.

<sup>7</sup>If the actual image process was wide-sense-stationary with spectral density  $S(f_x, f_y)$ , the variance of each channel would be an approximation to the average  $S(f_x, f_y)$  in the corresponding frequency bin.

actual variances are generally not consistent with an autoregressive model, and using the capacity expressions of Sec. 6.1 would yield misleading results.

### 6.3 Wavelet Models

Two influential papers in recent image compression literature are those by Joshi *et al.* [34] and LoPresto *et al.* [35], which constructed state-of-the-art wavelet image coders based on closely related statistical models. These papers respectively assumed that the wavelet coefficients are Gaussian and generalized-Gaussian distributed, with zero means and variances that depend on the coefficient location within each subband. Assume these variances take their values in a finite set  $\{\sigma_k^2, 1 \leq k \leq K\}$ . Typically  $K$  is equal to eight times the number of subbands, say  $K = 80$  [34]. If  $\{\sigma_k^2\}$  are treated as known deterministic quantities, the model is again of the form shown in Fig. 3. Otherwise one may assume that the variance field is slowly varying [35] and can be reliably estimated by the decoder. This model [35] for the wavelet coefficients is commonly referred to as the EQ model. We note that such adaptive Gaussian models have been successfully used in image denoising and restoration as well [36, 37, 38].

To obtain representative values of the variances  $\sigma_k^2, 1 \leq k \leq K$  for typical images, we proceeded as follows:

1. Select a discrete wavelet transform – say a 5-level decomposition using Daubechies’ length-8 filters [39]. Apply this transform to an image that is representative of some hypothetical class of images.
2. Estimate the local variance in a  $5 \times 5$  window centered at each wavelet coefficient.
3. Quantize the natural logarithm of each of these real numbers using a uniform quantizer with  $K$  reproduction levels and quantizer step size  $\Delta$ . We used  $K = 256$  in our experiments. All coefficients within a subband that have the same quantized variance are said to form one channel.

The subsampling factors  $r_k$  for the channels are given in the top panel of Fig. 11 for *Lena*. The capacity per sample in each channel is displayed in the bottom panel of Fig. 11. The product  $K\Delta$  is fixed by the range of the log-variances, and the estimate of data-hiding capacity converges to a limiting value if  $\Delta$  is made small enough. The value  $K = 256$  was chosen accordingly. Fig. 12 shows the corresponding power allocations  $d_{1k}$  and  $d_{2k}$  on a log scale. Again, the two asymptotic regimes discussed in Sec. 5.1 are clearly seen on these plots. Fig. 13 shows capacity as a function of  $D_2/D_1$  when  $D_1 = 10$ .

Similar experiments were conducted for other images. Table 1 gives the data-hiding capacity for variances  $\{\sigma_k^2\}$  computed from the images *Baboon*, *Lena* and *Peppers*. Again  $D_1$  was selected

Fig. 11  
goes  
here.  
Fig. 12  
goes  
here.  
Fig. 13  
goes  
here.

to be just below the visibility threshold of an i.i.d. white noise pattern. (For *Baboon*,  $D_1 = 25$ ; for *Lena* and *Peppers*,  $D_1 = 10$ .) Results are provided for  $D_2 = 2D_1$  and for  $D_2 = 5D_1$ . It is seen from Table 1 that an image like *Baboon*, which contains significant texture and is characterized by large  $\{\sigma_k^2\}$  and a large just-noticeable noise threshold  $D_1$ , has considerably larger data-hiding capacity than simpler images such as *Lena*.

Table 1  
goes  
here.

As Fig. 13 and Table 1 show, the spike-model approximation to capacity is quite accurate. Using a spike model amounts to classifying wavelet coefficients into two classes: those with very small variance and those with very large variance (relative to  $D_1$  and  $D_2$ ). This classification can be done by using the quantization method above with a two-level quantizer, with a very large value of  $\Delta$ .

**Block-DCT vs Wavelet EQ Model.** Comparison of Fig. 9 and Fig. 13 reveals that capacity estimates under the wavelet EQ model are lower than those under the block-DCT model. Since both expressions are upper bounds on actual capacity (where equality can be achieved only if the channels are independent and Gaussian), we conclude that the upper bound given by the wavelet EQ model is tighter. This is consistent with the sparsity properties of both image representations. Fig. 14 displays a measure of sparsity, namely, the fraction of transform coefficients whose variance is greater than  $t$ , plotted as a function of  $t$ . As this plot shows, the EQ model is indeed significantly sparser than the DCT model. It is seen for instance that under the EQ (resp. DCT) model, only 24% (resp. 44%) of the samples are in channels such that  $\sigma_k^2 > 40$ , and 16% (resp. 27%) of the samples are in channels such that  $\sigma_k^2 > 100$ . The threshold  $t = 40$  (resp.  $t = 100$ ) is our spike-approximation threshold  $t = 2D_2$  when  $D_2 = 20$  (resp.  $D_2 = 50$ ). This explains why the spike approximation to capacity under the EQ model is also lower than under the DCT model.

Fig. 14  
goes  
here.

## 6.4 Color Images

The study above can be extended to color images, by choosing a color space in which the color components are approximately independent (such as  $YC_rC_b$ ) and formulating parallel-Gaussian models on these images. The number of channels is three times larger than in the monochrome case. Chrominance channels are sparser than luminance channels, as indicated by the variances of block-DCT channels in Fig. 15, and hence contribute less to data-hiding capacity. One could similarly compute capacity estimates for hyperspectral images with arbitrarily many hyperspectral planes.

Fig. 15  
goes  
here.

Images with  $N$  color pixels may be viewed as vectors in  $\mathbb{R}^{3N}$  (obtained by stacking  $N$  subvectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \in \mathbb{R}^3$ , each representing an individual color pixel). Referring to Fig. 3, consider the decomposition of one such color image  $\mathbf{x}^N$  using a transform  $\mathbf{T}$  with  $3N$  output real-valued coefficients  $\{x_k(n)\}$  distributed among  $K$  channels. The *per-pixel* distortion between two color

images  $\mathbf{x}^N$  and  $\mathbf{x}'^N \in \mathbb{R}^{3N}$  is defined by extension of (6) as

$$d_{color}^N(\mathbf{x}^N, \mathbf{x}'^N) = \frac{1}{N} \sum_{k=1}^K \sum_{n=1}^{N_k} |x_k(n) - x'_k(n)|^2. \quad (16)$$

where  $\sum_{k=1}^K N_k = 3N$ . Hence  $d_{color}^N(\mathbf{x}^N, \mathbf{x}'^N) = 3d^N(\mathbf{x}^N, \mathbf{x}'^N)$ , where  $d^N$  is the per-sample distortion used in Sec. 5. Data-hiding capacity can be computed under the constraints  $Ed_{color}^N(\mathbf{S}^N, \mathbf{X}^N) \leq D_1$  and  $Ed_{color}^N(\mathbf{S}^N, \mathbf{Y}^N) \leq D_2$  for the data hider and the attacker, respectively. Hence we can again use the algorithm of Sec. 5.1, with distortion constraints  $D_1/3$  and  $D_2/3$ , respectively. The capacity per color pixel is three times the capacity per sample.

Fig. 16 compares the capacities (per pixel) of monochrome and color images for the *Lena* DCT model using  $D_1 = 10$ . The increase in capacity due to the availability of chrominance channels is somewhat less than threefold, as expected.

Fig. 16  
goes  
here.

## 7 Compressed Host Signals

Often data are to be hidden within compressed images. Of course one expects that a side effect of compression is to reduce data-hiding capacity, but by how much? Below we derive the solution under a parallel Gaussian model for the image source.

Denote by  $v(n), 1 \leq n \leq N$ , the uncompressed host signal and by  $s(n), 1 \leq n \leq N$  the compressed signal to be watermarked. The signal  $v(n)$  satisfies the statistical model of Sec. 5: the subsignals  $V_k$  are independent, and the samples  $V_k(n), 1 \leq n \leq N_k$  are i.i.d.  $\mathcal{N}(0, \sigma_{v,k}^2)$ . The transform  $\mathbf{T}$  is assumed to be orthonormal. Let  $D = E(V - S)^2$  be the distortion of the host signal due to compression, and assume that rate-distortion-optimal codes for parallel Gaussian channels are used [28, Sec. 13.3.3]. In this case, the subsignals  $S_k$  are independent, and the samples  $S_k(n), 1 \leq n \leq N_k$  are i.i.d.  $\mathcal{N}(0, \sigma_k^2)$ . Here  $\sigma_k^2$  is given by the classical reverse water-filling theorem:

$$\sigma_k^2 = \max(\sigma_{v,k}^2 - \lambda, 0), \quad (17)$$

the distortion in channel  $k$  due to compression is

$$\delta_k \triangleq E(V_k - S_k)^2 = \min(\lambda, \sigma_{v,k}^2), \quad (18)$$

and the nonnegative parameter  $\lambda$  is such that  $\sum_k \delta_k = D$ . The joint distribution of  $V_k$  and  $S_k$  is given by the Gaussian test channel with distortion  $\delta_k$ . The bit rate for  $S$  is given by

$$R_S = \frac{1}{2} \sum_{k=1}^K r_k \max\left(0, \log \frac{\sigma_k^2}{\delta_k}\right).$$

Of course, compression tends to increase the sparsity of the host signal because some of the channels may be set to zero, according to (17). Under our compression model,  $S$  still satisfies

a parallel-Gaussian model, so Theorem 1 still gives the data-hiding capacity  $C$  for  $S$ . We are interested in comparing  $C$  with the data-hiding capacity  $C_V = \max_{d_1} \min_{d_2} \sum_{k=1}^K r_k \Gamma(\sigma_{v,k}^2, d_{1k}, d_{2k})$  for the uncompressed signal  $V$ . By (17), we have  $\sigma_k^2 \leq \sigma_{v,k}^2$  for all  $k$ , and hence  $C \leq C_V$ . In low bit rate applications,  $S$  could be considerably more sparse than  $V$ , and the reduction in data-hiding capacity could be significant.

To obtain a quick but useful estimate of the loss of data-hiding capacity due to compression of  $V$ , consider the spike model approximation for both  $V$  and  $S$ , with respective sparsity factors  $r_V^*$  and  $r^* \leq r_V^*$ . In this case, the data-hiding capacity is reduced from  $C_V = \frac{1}{2} r_V^* \log \left( 1 + \frac{D_1}{D_2 - D_1} \right)$  to  $C = \frac{1}{2} r^* \log \left( 1 + \frac{D_1}{D_2 - D_1} \right) = \frac{r^*}{r_V^*} C_V$ .

The effects of host-image compression on data-hiding capacity are illustrated in Table 2 and Fig. 17. Assume that rate-distortion codes for the EQ model are used. We generate compressed image models by letting the Lagrange multiplier  $\lambda$  in (17) sweep a range of values. Fig. 17 shows the data-hiding capacity for the *Lena* model as a function of the bit rate for the host image, for several values of  $D_2/D_1$ . It is interesting to observe that compression down to approximately 1 bit per pixel has a relatively small effect on data-hiding capacity. This is not too surprising, because power is mostly allocated to strong channels (especially if  $D_2/D_1$  is large, see Fig. 12), and those channels remain strong even after mild compression. At rates below 1 bit per pixel, some of the moderately strong channels become weak (or even zero), and hence the loss in data-hiding capacity becomes more significant. Similar effects are observed for other images, see Table 2.

Table 2  
goes  
here.  
Fig. 17  
goes  
here.

## 8 Perceptual Distortion Metrics

Another major difficulty arises when trying to apply the theory above to image processing. The MSE distortion metric is not well matched to the human visual system [33, 40, 7]. Moreover, in a game with the attacker, the MSE metric may be quite inappropriate, because spatial shifts, rotations, and other geometric warping operations introduce negligible loss in subjective signal quality, even though the resulting MSE may be large. Ideally one would like to use a perceptual distortion metric, but there is currently no universally accepted such metric, and analysis based on perceptual metrics is likely to be intractable. How strongly does the capacity formula in Theorem 1 depend on the distortion metric used? To gain some insight into that problem, we first consider classical weighted squared-error metrics. Then we study a modified MSE metric that is invariant to geometric attacks.

## 8.1 Weighted MSE

In some problems, a weighted mean-squared error (WMSE) criterion [33, 40] might be more appropriate than the MSE used so far:

$$d^N(x^N, x'^N) = \frac{1}{N} \sum_{k=1}^K w_k \sum_{n=1}^{N_k} |x_k(n) - x'_k(n)|^2 \quad (19)$$

where  $w_k, 1 \leq k \leq K$  are positive weights, and  $x^N, x'^N$  are any two images in  $\mathbb{R}^N$ . Observe that the data-hiding problem expressed in terms of a constraint on the WMSE (19) can be reduced to a problem stated in terms of the MSE (6), if the samples in channel  $K$  are multiplied by  $\sqrt{w_k}$  after application of the multirate transform  $\mathbf{T}$ , see Fig. 18. These scale factors can be absorbed into  $\mathbf{T}$ ; hence the capacity formulas of Secs. 5.1–7 apply directly, with weighted variances  $\{w_k \sigma_k^2\}$  in lieu of the original host-signal variances  $\{\sigma_k^2\}$ . A similar analytic technique has been used in the context of signal compression [40].

Fig. 18  
goes  
here.

For some channels we could have  $w_k = 0$ ; such channels have zero data-hiding capacity and may be arbitrarily tampered with by the attacker, without his incurring any penalty. For color images for instance, the attacker might be satisfied with a monochrome version of the watermarked image. If the weighting factors  $w_k$  for the chrominance channels are zero, converting the color image to a monochrome image would be a legitimate attack. It is then impossible to reliably hide information in the chrominance components of the host image.

We experimented with WMSE metrics, by selecting a set of nonuniform weights  $\{w_k\}$  for the model in Fig. 3 and computing the resulting capacities for monochrome and color image models. We used the block-DCT model with weights given by  $w_k = \gamma/q_k^2$ , where  $q_k$  is the default JPEG quantizer step [41] in channel  $k$  (64 and 192 channels for monochrome and color images, respectively.) The motivation for this choice is that noise with variance proportional to  $q_k^2$  would be perceptually white. The arbitrary constant  $\gamma$  was adjusted so that the just noticeable distortion level for the reference i.i.d. noise stimulus in Sec. 6 would be equal to that in the unweighted case. Hence  $D = \sum_k r_k w_k D$ , and so  $\gamma = 1/\sum_k r_k q_k^{-2}$ .

The results using these weights are shown in Fig. 16. There appears to be relatively little difference between capacities computed using weighted and unweighted squared-error distortion measures. Channels with low weights  $w_k$  are high-frequency channels, which already have low hiding capacity under unweighted distortion measures. Hence the effect of weighting, as seen from Fig. 3, is generally to make weak channels even weaker and strong channels stronger. Such operations have little effect on capacity, so discrepancies between capacity estimates for weighted and unweighted distortion measures are primarily due to a few strong channels becoming weak and vice-versa. These results suggest it may be an overkill to use complex perceptual models to refine a simpler analysis of data-hiding capacity based on squared-error distortion measures. Similar

conclusions have been obtained in image compression using rate-distortion-optimized coders where MSE (and not weighted MSE) as the cost function used to select various coder parameters [42].

For color images, both the variances  $\sigma_k^2$  (see Fig. 15) and the weighting factors  $w_k$  for chrominance channels are lower than those for luminance channels; both of these factors reduce the chrominance channels' data-hiding capability.

## 8.2 Invariant Distortion Metrics

Consider a distortion metric that does not penalize warping attacks in the diagram of Fig. 2:

$$d_W^N(s^N, y^N) \triangleq d^N(s^N, z^N) \quad (20)$$

for all  $z^N$  and  $\theta \in \Theta$  such that  $y^N = W(z^N, \theta)$ . Here  $d^N(\cdot, \cdot)$  is the squared-error distortion metric. The definition (20) implies that the new distortion  $d_W^N(\cdot, \cdot)$  is invariant to warping operations with parameters  $\theta \in \Theta$ .

Suppose first that the parameters  $\theta$  of the geometric attack are identifiable (see Sec. 3). In practice this may require that the host signal  $S^N$  contain a synchronizing pattern [23]. Then our invertibility assumption on  $W(\cdot, \theta)$  implies that the decoder can retrieve  $z^N$  in Fig. 2, and so the data-hiding capacity under the distortion metric (20) is the same as that given by Theorem 1, under the MSE metric.

If  $\theta$  is nonidentifiable, the same issue mentioned in Sec. 3 arise. If the warping attack is applied independently on blocks of length  $L$ , the capacity is essentially the same as that of a decoder that knows  $\theta$  used for each block. Hence the capacity result of Theorem 1 essentially applies again.

## 8.3 Threshold Vision

Due to threshold effects in the human visual system, it is possible to modify an image without introducing any perceptual degradation. In other terms,  $d(x^N, x'^N) = 0$  does not necessarily imply  $x^N = x'^N$ . There is a substantial body of literature quantifying threshold effects in the frequency domain; in particular, Just Noticeable Difference (JND) thresholds have been measured experimentally as a function of frequency. Moreover, JND models can be locally adapted. While JNDs have been used extensively in the design of practical watermarking systems [7], evaluating capacity formulas for this particular distortion model (with  $D_1 = 0$ ) remains to be done.

## 9 Conclusion

We have characterized data-hiding capacity for realistic image sources, by application of recent theoretical results on parallel Gaussian channels. Several statistical image models have been considered, ranging from conventional ones such as autoregressive processes and block-DCT models, to more advanced wavelet models that capture the spatial clustering of wavelet coefficients. The results obtained are guaranteed to be upper bounds on actual capacity if the true model deviates from the assumed parallel Gaussian model. The analysis demonstrates the essential role of image sparsity in determining capacity. The EQ wavelet model we considered captures sparsity much better than does the simpler block-DCT and autoregressive models, and yields better capacity estimates.

Our basic capacity results have been derived under MSE constraints on the data hider and on the attacker. We have also considered two types of perceptual distortion metrics: a weighted MSE, and a modified MSE metric that is invariant to geometric attacks. It appears that the dependence of capacity on the actual distortion metric used is quite weak. Moreover, we have shown that the loss in capacity due to geometric attacks may be quite benign.

Research is already underway on data-hiding codes that approach capacity for Gaussian channels. The results presented in this paper should likewise help design data-hiding codes that approach the capacity limits for images. The performance of several watermarking codes has been studied in several recent papers [30, 43, 44, 45].

Closely related to these efforts is the development of optimal decoders. This requires moving beyond the conventional correlation rules and normalized correlation rules that have often been used in the image watermarking literature, due to their simplicity. An optimal decoder needs to be designed jointly with the encoder, and may need to be provided with the means to reliably estimate various unknown source and channel parameters.

## A Outline of the Proof of Theorem 1

The main ideas used in the proof of Theorem 1 are:

1. The payoff function  $\sum_{k=1}^K r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k})$  is additive over  $k$ , and so are the distortion constraints (7) and (8). The other  $3K$  constraints (9) (10), and (11) apply to each channel separately.
2. For any  $d_1$ , the constrained minimization problem is reformulated as the dual maximization problem  $\max_{\lambda_2 \geq 0} q(d_1, \lambda_2)$ , where the dual variable  $\lambda_2 \geq 0$  corresponds to the distortion constraint (8).
3. A closed-form solution for each optimal  $d_{2k}$  is derived in terms of  $d_{1k}$  and  $\lambda_2$ .
4. The function  $q(d_1, \lambda_2)$  is concave in  $d_1$  and the constraint set is convex. The maximization problem is converted to a dual minimization problem  $\min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2)$ , where the dual variable  $\lambda_1 \leq 0$  corresponds to the distortion constraint (7).
5. We have  $C = \max_{\lambda_2 \geq 0} \min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2)$  where  $r$  is strictly convex in  $\lambda_1$ . This maxmin problem is solved using a standard numerical algorithm.

A numerical optimization algorithm based on these properties is described in [25]. The dual variables  $\lambda_1$  and  $\lambda_2$  represent sensitivity parameters with respect to changes in distortion levels  $D_1$  and  $D_2$ .

## References

- [1] W. Bender, D. Gruhl and N. Morimoto, “Techniques for Data Hiding,” *IBM Syst. J.*, Vol. 35, 1996.
- [2] F. Hartung and M. Kutter, “Multimedia Watermarking Techniques,” pp. 1079—1107 in [5].
- [3] *IEEE Journal on Selected Areas in Communications*, Special Issue on Copyright and Privacy Protection, Vol. 16, No. 4, May 1998.
- [4] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Information Hiding – A Survey,” pp. 1062—1078 in [5].
- [5] *Proceedings IEEE*, Special Issue on Identification and Protection of Multimedia Information, Vol. 87, No. 7, July 1999.
- [6] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, “Multimedia Data–Embedding and Watermarking Strategies,” *Proc. IEEE*, Vol. 86, No. 6, pp. 1064—1087, June 1998.
- [7] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, “Perceptual Watermarks for Digital Images and Video,” pp. 1108—1126 in [5].
- [8] I. J. Cox, J. Killian, F. T. Leighton and T. Shamoan, “Secure Spread Spectrum Watermarking for Multimedia,” *IEEE Trans. Image Proc.*, Vol. 6, No. 12, pp. 1673—1687, Dec. 1997.
- [9] B. Chen and G. W. Wornell, “Preprocessed and Postprocessed Quantization Index Modulation Methods for Digital Watermarking,” *Proc. SPIE*, Vol. 3971, San Jose, CA, Jan. 2000.
- [10] N. Merhav, “On Random Coding Exponents of Watermarking Codes,” *IEEE Trans. Info. Theory*, Vol. 46, No. 2, pp. 420—430, March 2000.
- [11] P. Moulin and J. A. O’Sullivan, “Information–Theoretic Analysis of Information Hiding,” *submitted to IEEE Trans. Information Theory*, Oct. 1999; revised, June 2001, and December 2001. Available from [www.ifp.uiuc.edu/~moulin/Papers/IThiding99r.ps.gz](http://www.ifp.uiuc.edu/~moulin/Papers/IThiding99r.ps.gz). Presented at *IEEE Int. Symp. on Info. Thy*, Sorrento, Italy, June 2000.
- [12] J. A. O’Sullivan, P. Moulin, and J. M. Ettinger, “Information–Theoretic Analysis of Steganography,” *Proc. IEEE Int. Symp. on Info. Thy*, Cambridge, MA, p. 297, Aug. 1998.
- [13] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, “Capacity of the Watermark Channel: How Many Bits Can be Hidden Within a Digital Image?,” *Proc. SPIE*, Vol. 3657, pp. 437—448, San Jose, CA, Jan. 1999.
- [14] M. Ramkumar and A. N. Akansu, “Information Theoretic Bounds for Data Hiding in Compressed Images,” *IEEE 2nd Workshop on Multimedia Signal Processing*, P. W. Wong et al., Eds., Redondo Beach, CA, Dec. 1998.
- [15] S. D. Servetto, C. I. Podilchuk and K. Ramchandran, “Capacity Issues in Digital Image Watermarking,” *Proc. ICIP’98*, Chicago, IL, Vol. I, pp. 445—449, Oct. 1998.
- [16] C. E. Shannon, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, 1948.
- [17] T. Başar, “The Gaussian Channel with an Intelligent Jammer,” *IEEE Trans. Info. Thy*, Vol. 29, No. 1, pp. 152—157, Jan. 1983.
- [18] A. Lapidoth and P. Narayan, “Reliable Communication Under Channel Uncertainty,” *IEEE Trans. Info. Thy*, Vol. 44, No. 6, pp. 2148—2177, Oct. 1998.
- [19] P. Moulin and J. A. O’Sullivan, “Information–Theoretic Analysis of Watermarking,” *Proc. Int. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, Istanbul, Turkey, June 2000.
- [20] A. Cohen and A. Lapidoth, “On the Gaussian Watermarking Game,” *Proc. Conf. on Info. Science and Systems*, Princeton, NJ, pp. TA4/21—26, March 2000.

- [21] I. Csiszár and J. Körner, *Information Theory: Coding Theory for Discrete Memoryless Systems*, Academic Press, NY, 1981.
- [22] B. Hughes and P. Narayan, "Gaussian Arbitrarily Varying Channels," *IEEE Trans. Info. Thy*, Vol. 33, No. 2, pp. 267—284, March 1987.
- [23] M. Kutter, F. Jordan and F. Bossen, "Digital Signature of Color Images Using Amplitude Modulation," available from [ltswww.epfl.ch/~jordan/watermarking.html](http://ltswww.epfl.ch/~jordan/watermarking.html), 1996.
- [24] S. I. Gel'fand and M. S. Pinsker, "Coding for Channel with Random Parameters," *Problems of Control and Information Theory*, Vol. 9, No. 1, pp. 19—31, 1980.
- [25] P. Moulin and M. K. Mihçak, "The Parallel-Gaussian Watermarking Game," *UIUC Coord. Sci. Lab Tech. Report*, June 2001, revised Jan. 2002. Available from [www.ifp.uiuc.edu/~moulin/Papers/pg01.ps.gz](http://www.ifp.uiuc.edu/~moulin/Papers/pg01.ps.gz).
- [26] P. Moulin, "The Parallel-Gaussian Watermarking Game," *Proc. 35th Conf. on Information Sciences and Systems*, Baltimore, MD, March 2001. Available from [www.ifp.uiuc.edu/~moulin/Papers/gauss-ciss01.ps.gz](http://www.ifp.uiuc.edu/~moulin/Papers/gauss-ciss01.ps.gz).
- [27] M. Costa, "Writing on Dirty Paper," *IEEE Trans. Info. Thy*, Vol. 29, No. 3, pp. 439—441, May 1983.
- [28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.
- [29] J. Chou, S. Pradhan, L. El Ghaoui and K. Ramchandran, "A Robust Optimization Solution to the Data Hiding Problem using Distributed Source Coding Principles," *Proc. SPIE*, Vol. 3971, San Jose, CA, Jan. 2000.
- [30] M. Kesal, M. K. Mihçak, R. Kötter and P. Moulin, "Iteratively Decodable Codes for Watermarking Applications," *Proc. 2nd Symposium on Turbo Codes and Their Applications*, Brest, France, Sep. 2000.
- [31] P. Moulin, M. K. Mihçak and G.-I. Lin, "An Information-Theoretic Model for Image Watermarking," *Proc. Int. Conf. on Image Proc.*, Vancouver, B.C., Oct. 2000.
- [32] C. Weidmann and M. Vetterli, "Rate-Distortion Analysis of Spike Processes," *Proc. Data Compression Conf.*, Snowbird, UT, March 1999.
- [33] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, NJ, 1989.
- [34] R. L. Joshi, H. Jafarkhani, J. H. Kasner, T. R. Fischer, N. Farvardin, M. W. Marcellin and R. H. Bamberger, "Comparison of Different Methods of Classification in Subband Coding of Images," *IEEE Trans. on Image Processing*, Vol. 6, No. 11, pp. 1473-1486, Nov. 1997.
- [35] S. LoPresto, K. Ramchandran and M. T. Orchard, "Image Coding based on Mixture Modeling of Wavelet Coefficients and a Fast Estimation-Quantization Framework," *Proc. Data Compression Conference 97*, Snowbird, Utah, pp. 221—230, 1997.
- [36] J. Liu and P. Moulin, "Statistical Image Restoration Based on Adaptive Wavelet Models," *Proc. ICIP'01*, Thessaloniki, Greece, Oct. 2001.
- [37] M. K. Mihçak, I. Kozintsev, K. Ramchandran and P. Moulin, "Low-Complexity Image Denoising Based on Statistical Modeling of Wavelet Coefficients," *IEEE Signal Processing Letters*, Vol. 6, No. 12, pp. 300—303, Dec. 1999.
- [38] E. P. Simoncelli, "Modeling the Joint Statistics of Images in the Wavelet Domain," *Proc. SPIE*, Vol. 3813, Denver, CO, July 1999.
- [39] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- [40] N. S. Jayant, J. D. Johnston and R. J. Safranek, "Signal Compression Based on Models of Human Perception," *Proc. IEEE*, Vol. 81, No. 10, pp. 1385—1422, Oct. 1993.
- [41] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*, Van Nostrand Reinhold, 1993.

- [42] A. Ortega and K. Ramchandran, "Rate-Distortion Methods for Image and Video Compression," *IEEE Sig. Proc. Magazine*, Special Issue on Rate-Distortion Techniques for Image/Video Compression, Vol. 15, No. 6, pp. 51—73, Nov. 1998.
- [43] J. J. Eggers, J. K. Su and B. Girod, "Robustness of a Blind Image Watermarking Scheme," *Proc. IEEE Int. Conf. on Image Proc.*, Vancouver, B.C., Oct. 2000.
- [44] J. R. Hernandez, J.-F. Delaigle and B. M. M. Macq, "Improving Data Hiding by using Convolutional Codes and Soft-Decision Decoding," *Proc. SPIE*, Vol. 3971, San Diego, CA, Jan. 2000.
- [45] M. K. Mihçak and P. Moulin, "Information Embedding Codes Matched to Locally Stationary Gaussian Image Models," to appear in *Proc. ICIP'02*, Rochester, NY, Sep. 2002.

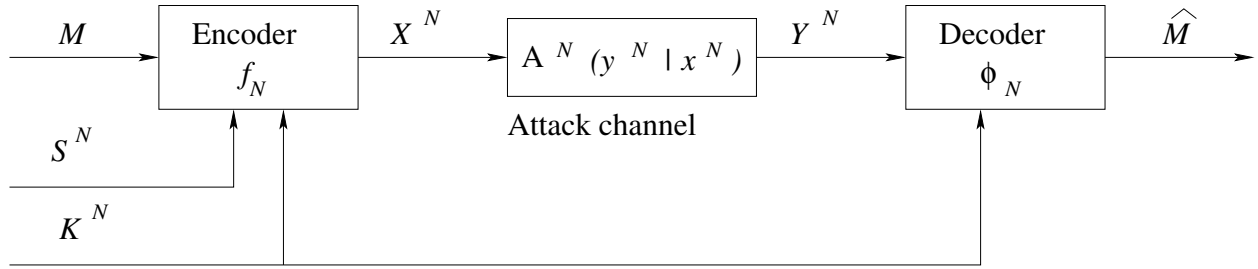


Figure 1: The watermark communication problem.

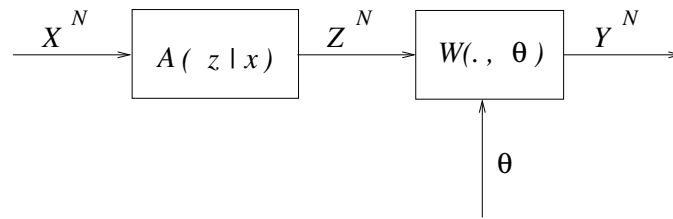


Figure 2: Cascade of a memoryless attack channel and a warping attack channel.

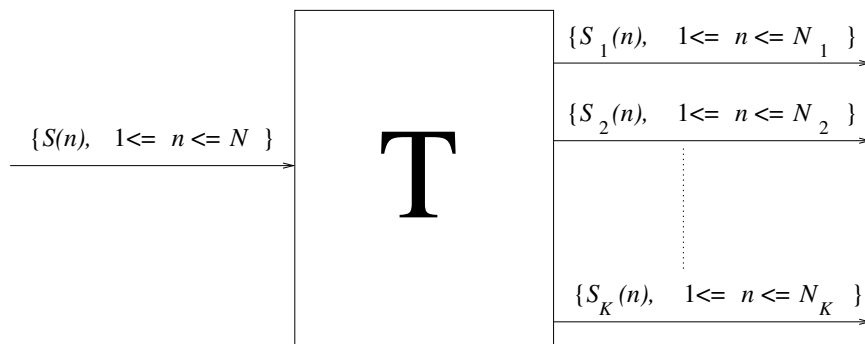


Figure 3: Decomposition of host signal  $S$  into  $K$  channels, using a (possibly nonlinear) multirate transform  $\mathbf{T}$ .

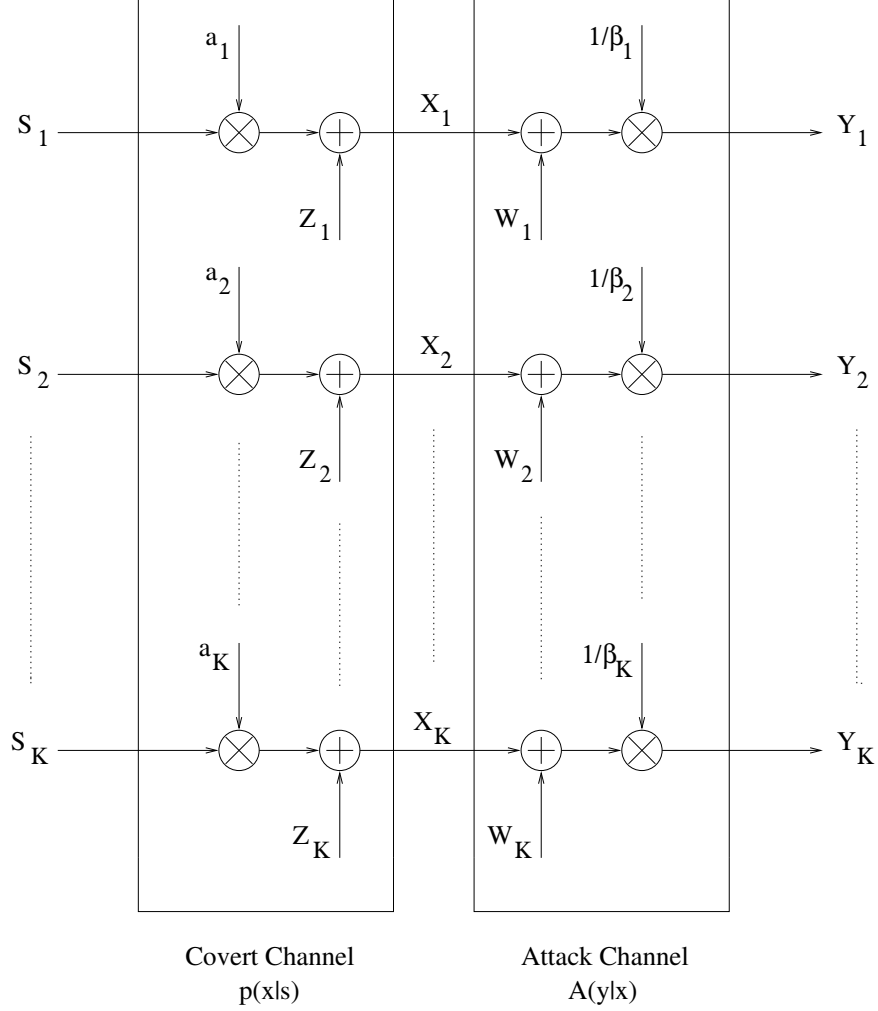


Figure 4: Optimal data-hiding and attack strategies for parallel Gaussian channels  $S_k \sim \mathcal{N}(0, \sigma_k^2), 1 \leq k \leq K$ . The channels are decoupled, with optimal power allocations  $\{d_{1k}\}, \{d_{2k}\}$  between channels is given in Theorem 1. Here  $a_k = 1 - \frac{d_{1k}}{\sigma_k^2}$ ,  $Z_k \sim \mathcal{N}(0, ad_{1k})$ ,  $W_k \sim \mathcal{N}\left(0, (d_{2k} - d_{1k}) \frac{\sigma_k^2 - d_{1k}}{\sigma_k^2 - d_{2k}}\right)$ ,  $\beta_k = \frac{\sigma_k^2 - d_{1k}}{\sigma_k^2 - d_{2k}}$ ;  $Z_k$  and  $W_k$  are independent of  $S_k$  and  $X_k$  respectively.



(a)



(b)



(c)



(d)



(e)

Figure 5: (a) Original *Lena*, (b) watermarked for  $D_1 = 10$ ,  $D_2 = 20$  ( $PSNR = 38.2$  dB), (c) attacked for  $D_1 = 10$ ,  $D_2 = 20$  ( $PSNR = 35.2$  dB), (d) watermarked for  $D_1 = 10$ ,  $D_2 = 50$  ( $PSNR = 38.2$  dB), (e) attacked for  $D_1 = 10$ ,  $D_2 = 50$  ( $PSNR = 31.3$  dB).

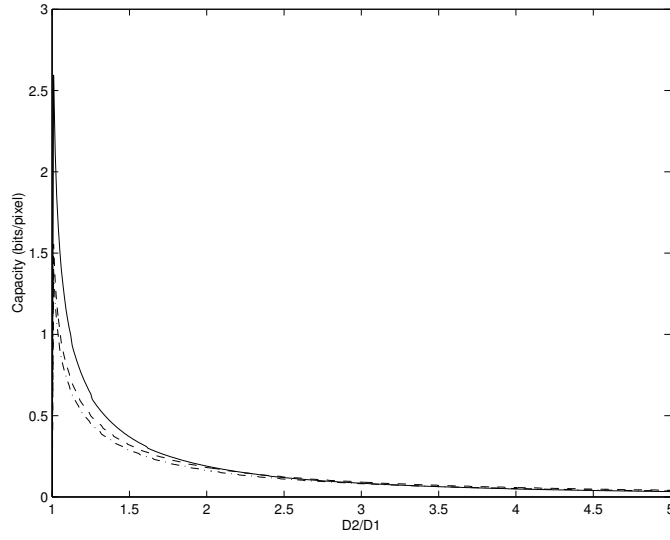


Figure 6: Capacity in bit/pixel (solid line) versus  $D_2/D_1$  for 2-D separable AR(1) process with  $\sigma^2 = 2500$ ,  $\rho_x = \rho_y = 0.95$ , and  $D_1 = 10$ . Spike approximation with threshold equal to  $1.5D_2$  (dashed line) and  $2D_2$  (dashed-dotted line).

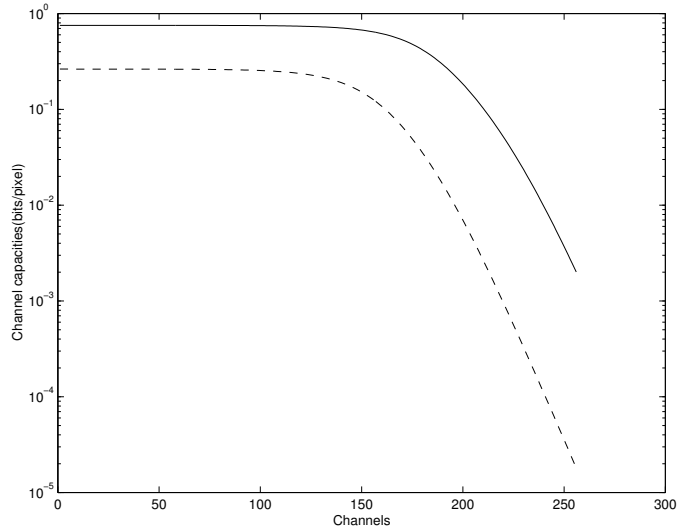


Figure 7: Contribution (on a log scale) of individual frequency channels to capacity for separable 2-D AR(1) process with  $\rho_x = \rho_y = 0.95$ ,  $\sigma^2 = 2500$ , for  $D_1 = 10$  and  $D_2 = 20$  (solid curve) and  $D_2 = 50$  (dashed curve).

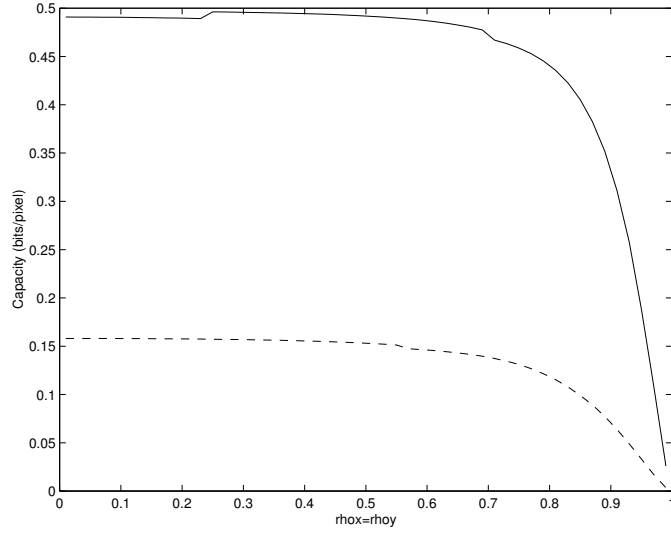


Figure 8: Capacity (in bit/pixel) versus  $\rho_x = \rho_y$  for 2-D separable AR(1) process with  $\sigma^2 = 2500$ ,  $D_1 = 10$ , and  $D_2 = 20$  (solid curve) and  $D_2 = 50$  (dashed curve).

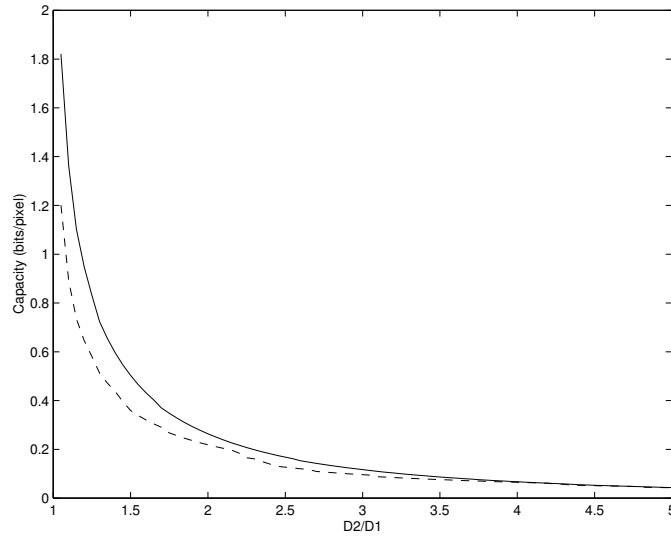


Figure 9: Capacity in bit/pixel (solid line) versus  $D_2/D_1$  for block-DCT model of *Lena*,  $D_1 = 10$ . Spike approximation with threshold equal to  $2D_2$  (dashed line).

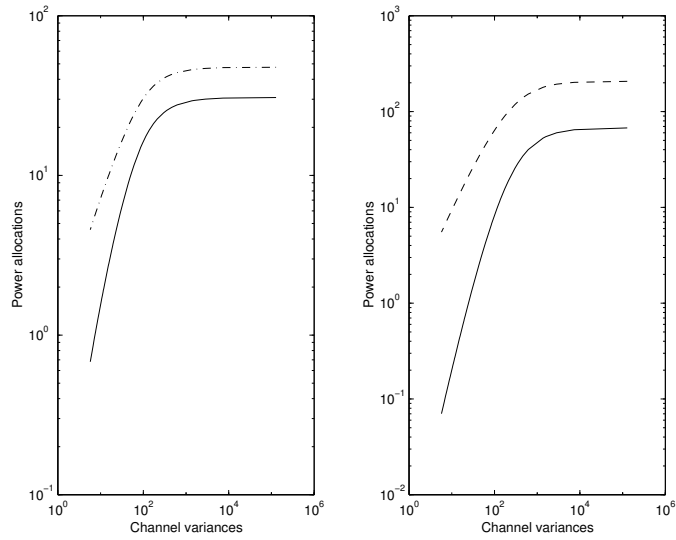


Figure 10: Optimal power allocations for data hider (solid line) and attacker (dashed line) under block-DCT model of *Lena*, for  $D_1 = 10$  and (a)  $D_2 = 20$  and (b)  $D_2 = 50$ .

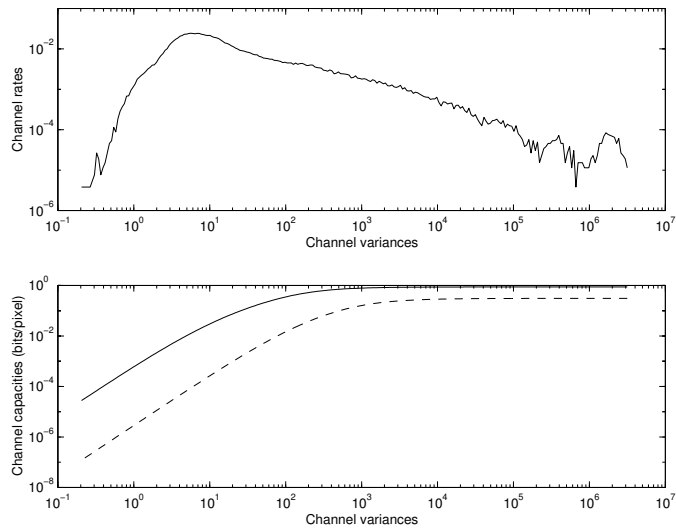


Figure 11: Top plot: subsampling factors (on a log scale) for individual channels of *Lena* under EQ model ( $K = 256$ ). Bottom plot: contribution (on a log scale) of individual channels to capacity for  $D_1 = 10$  and  $D_2 = 20$  (solid line) and  $D_2 = 50$  (dashed line).

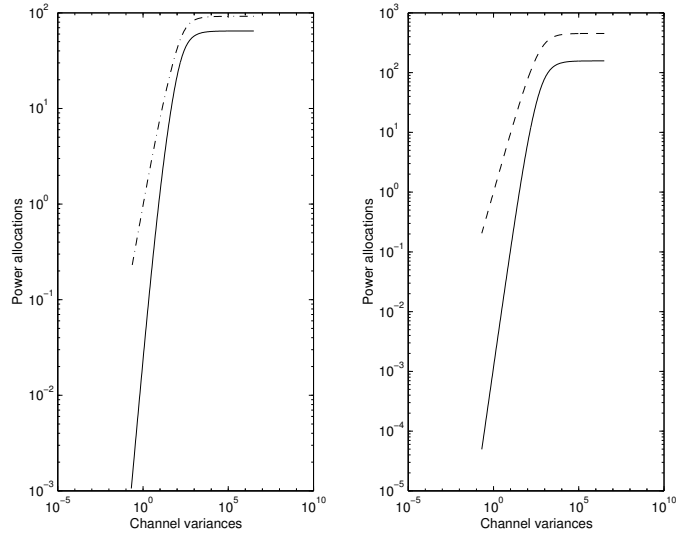


Figure 12: Optimal power allocations for data hider (solid line) and attacker (dashed line) under EQ wavelet model of *Lena*, for  $D_1 = 10$  and (a)  $D_2 = 20$ ; (b)  $D_2 = 50$ .

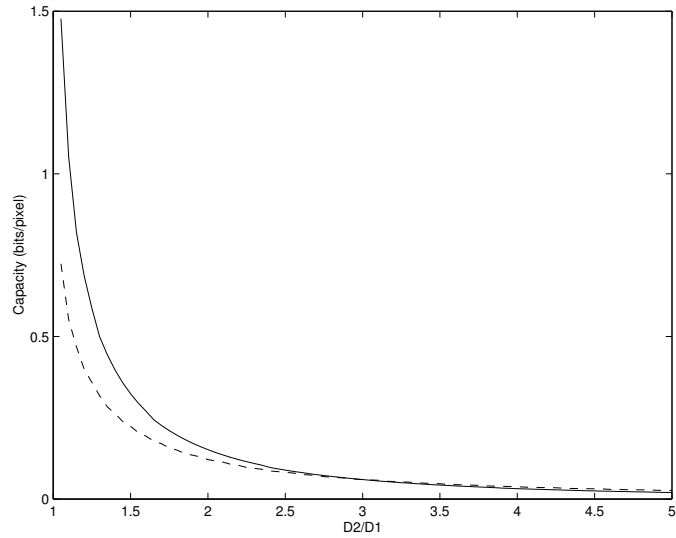


Figure 13: Capacity in bit/pixel (solid line) versus  $D_2/D_1$  for EQ wavelet model of *Lena*, for  $D_1 = 10$ . Spike approximation with threshold equal to  $2D_2$  (dashed line).

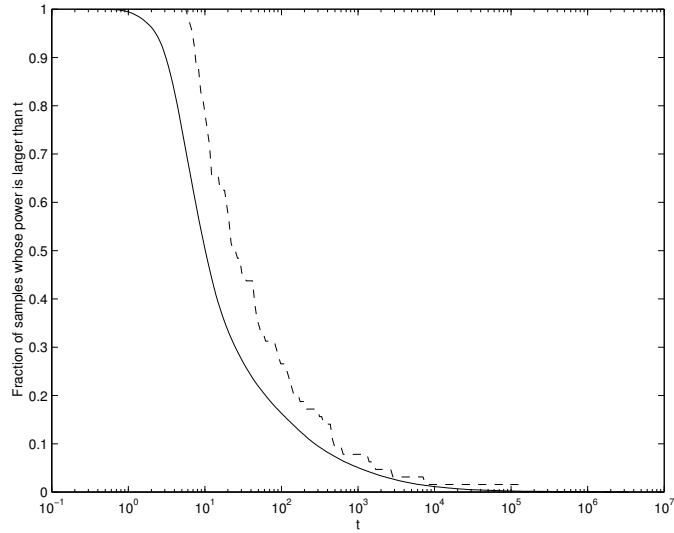


Figure 14: Sparsity of block-DCT (dashed curve) and EQ (solid curve) wavelet representations of *Lena*. The fraction of transform coefficients with variance greater than  $t$  is plotted *vs t*.

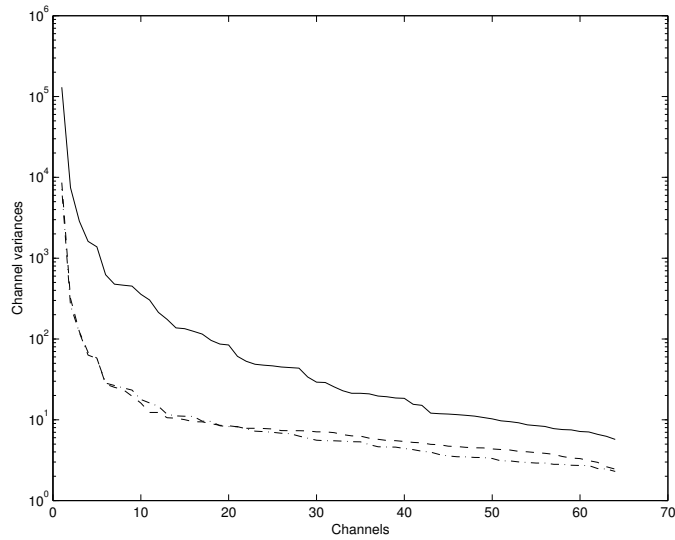


Figure 15: Variance of block-DCT coefficients of  $Y$  (solid),  $C_r$  (dashed) and  $C_b$  (dashed-dotted) components of *Lena*.

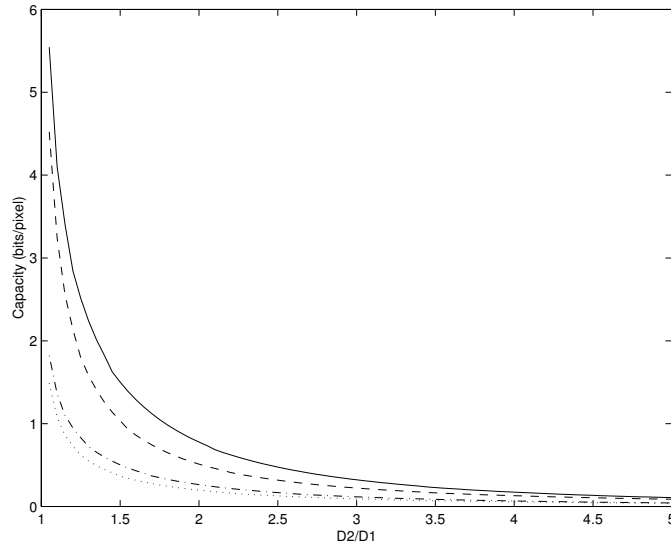


Figure 16: Comparison of capacity per pixel under monochrome (dashed-dotted line) and color (solid line) block-DCT models of *Lena*, for  $D_1 = 10$ . Also shown is the capacity under weighted MSE distortion measures, for monochrome (dotted line) and color models (dashed line).

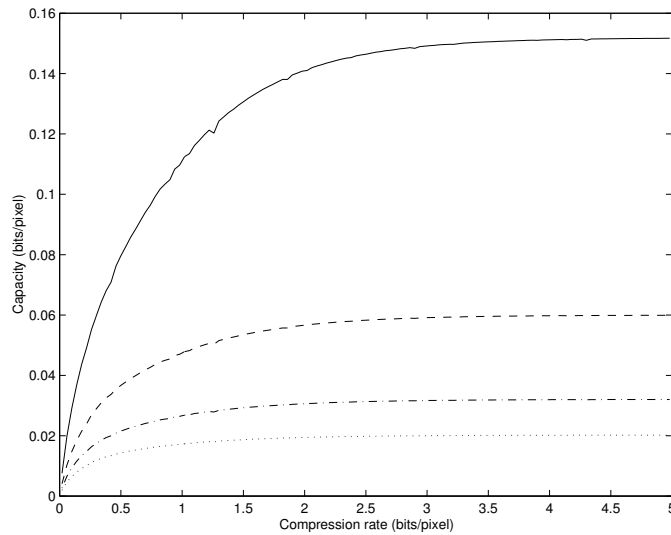


Figure 17: Reduction in data-hiding capacity of *Lena* due to compression. Rate-distortion codes for the EQ model are used, and  $D_1 = 10$ . Data-hiding capacities are given for four values of  $D_2$ . From top to bottom:  $D_2 = 20, 30, 40, 50$ .

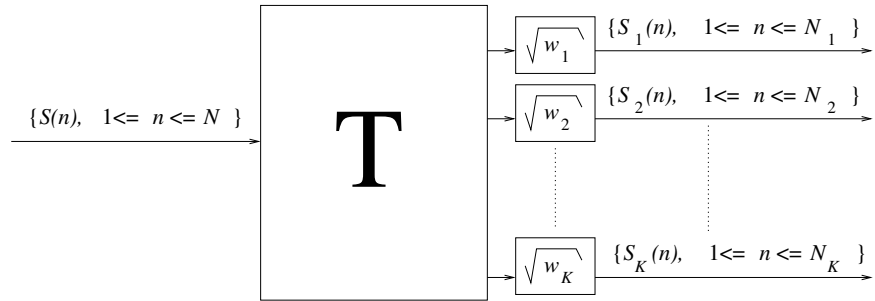


Figure 18: Decomposition of host signal  $S$  into  $K$  channels, using a (possibly nonlinear) multirate transform  $\mathbf{T}$  and a set of nonnegative weights  $\{\sqrt{w_k}\}$ . The weighted MSE at the output of  $\mathbf{T}$  is equal to the unweighted MSE at the output of the system.

Image	$D_1$	$D_2 = 2D_1$		$D_2 = 5D_1$	
		NC	NC(spike)	NC	NC(spike)
<i>Baboon</i>	25	96352	93048	21837	20146
<i>Lena</i>	10	39817	31856	5297	6952
<i>Peppers</i>	10	52034	34078	6241	6568

Table 1: Total data-hiding capacities (in bits) for images of size  $N = 512^2$ , for just noticeable  $D_1$ . The underlying EQ image model assumes Gaussian wavelet coefficients with location-dependent variances. The first estimate  $NC$  uses a fine quantization of variances ( $K = 256$  channels), and the second  $NC(\text{spike})$  uses a coarse quantization with threshold equal to  $2D_2$  (two channels).

Image	$D_2 = 2D_1$				$D_2 = 5D_1$			
	Rate of $S(\text{bpp})$							
	0.1	0.5	1	$\infty$	0.1	0.5	1	$\infty$
<i>Baboon</i>	15287	47406	71914	96352	3222	11044	16383	21837
<i>Lena</i>	7641	20880	29228	39817	1712	3752	4529	5297
<i>Peppers</i>	7383	21627	35265	52034	1753	3811	4937	6241

Table 2: Total data-hiding capacities (in bits) for compressed images, under the same assumptions as in Table 1.