Information–Theoretic Analysis of Information Hiding [*]

Pierre Moulin
University of Illinois
Beckman Inst., Coord. Sci. Lab & ECE Dept.
405 N. Mathews Ave.
Urbana, IL 61801
moulin@ifp.uiuc.edu

Joseph A. O'Sullivan
Washington University
Dept. of Electrical Engineering
One Brookings Drive
St Louis, MO 63130
jao@ee.wustl.edu

October 1, 1999

Abstract

An information–theoretic analysis of information hiding is presented in this paper, forming the theoretical basis for design of information–hiding systems. Information hiding is an emerging research area which encompasses applications such as copyright protection for digital media, watermarking, fingerprinting, and steganography. In these applications, information is hidden within a host data set and is to be reliably communicated to a receiver. The host data set is purposely corrupted, but in a covert way, designed to be imperceptible to a casual analysis. Next, an attacker may seek to destroy this hidden information, and for this purpose, introduce additional distortion to the data set. Cryptographic keys may be available to the information hider and to the decoder.

We formalize these notions and evaluate the *hiding capacity*, which upper–bounds the rates of reliable transmission and quantifies the fundamental tradeoff between three quantities: the achievable information–hiding rates and the allowed distortion levels for the information hider and the attacker. The hiding capacity is the value of a game between the information hider and the attacker. The optimal attack strategy is the solution of a particular rate-distortion problem, and the optimal hiding strategy is the solution to a channel coding problem. Explicit formulas for capacity are given in several cases, including the important special case of small distortions. In some cases, including the one above, the hiding capacity is the same whether or not the decoder knows the host data set. It is shown that existing information–hiding systems in the literature operate far below capacity.

Index terms: Information hiding, watermarking, steganography, fingerprinting, cryptography, game theory, channel capacity, rate-distortion theory, optimal jamming, randomized codes, network information theory.

[*] Work presented in part at the IEEE International Symposium on Information Theory, Boston, MA, Aug. 1998. P. Moulin was supported by NSF grant MIP-97-07633.

1 Introduction

Information hiding is an emerging research area which encompasses applications such as copyright protection for digital media, watermarking, fingerprinting, and steganography. In particular, watermarking is now a major activity in audio, image, and video processing, and standardization efforts for JPEG-2000, MPEG-4, and Digital Video Disks are underway. Commercial products are already being developed. The first three International Workshops on Information Hiding were held in 1996, 1998, and 1999. Special issues of the IEEE Journal on Selected Areas in Communications and of the Proceedings of IEEE were recently devoted to copyright and privacy protection [1, 2]. An excellent review of the current state of the art appears in [3], and comprehensive surveys of image and multimedia watermarking techniques are available from [4, 5]. The majority of the papers to date have focused on novel ways to hide information and to detect and/or remove hidden information. However, these papers have lacked a guiding theory describing the fundamental limits of any information–hiding system. The need for practitioners and system designers to understand the nature of these fundamental limits has been recognized [1, 3, 5, 6, 7]. We help to close this gap by providing a theoretical basis for a generic version of the information-hiding problem. We formulate the information–hiding problem as a communication problem and seek the maximum rate of reliable communication through the communication system. Related aspects of this problem have also been recently explored by Merhav [8] and Chen and Wornell [9]. Also see the recent study by Hernández and Pérez-González on decision-theoretic aspects of the watermarking problem [10].

In our generic information–hiding problem, a message M is to be embedded in a host data set $\tilde{X}$, and the resulting data set X may be subject to data processing operations (*attacks*) that attempt to remove any trace of M from X. The information–hiding system should satisfy two basic requirements. The first requirement is usually referred to as *transparency*, or *unobtrusiveness*: the data set X should be similar to $\tilde{X}$, according to a suitable distortion measure. The second requirement is referred to as *robustness*: the hidden message should survive the application of any data processing technique (within a certain class) to X. Often there is a limit on the amount of distortion that an attacker is willing to introduce. A precise statement of the information–hiding problem is proposed in Sec. 2.

Applications of information hiding are quite diverse [3]. In watermarking applications, the message contains information such as owner identification and a digital time stamp. The goal here is usually copyright protection. The message itself is not secret, but it is desired that it permanently resides within the host data set. Similar requirements exist for systems that embed object identification in audio, image, or video databases. Closely related to watermarking is the fingerprinting, or traitor tracing, problem, where in addition to copyright information, the owner of the data set embeds a serial number, or fingerprint, that uniquely identifies the user of the dataset and makes it possible to trace any unauthorized use of the data set back to the user [11, 12]. This application is particularly challenging as it opens up the possibility of a collusion between different users to remove these fingerprints. A different type of application is the embedding of data such as multilingual soundtracks in pay–per–view television programs. Here the message is secret in the sense that it should not be decipherable by unauthorized decoders. Other applications of information hiding to television broadcasting are described in [13]. Another, more classical application of information hiding is steganography. Here not only is the message secret, but its very presence within the host data set should be undetectable. Steganography and related applications have a long, sometimes romantic history dating from ancient times [3, 6, 14, 15].

This brief discussion suggests that information hiding borrows from a variety of areas, including signal processing, communications, game theory, and cryptography. Indeed, a vast array of techniques from signal processing and communications have been used to design algorithms for hiding information (e.g., the spread–spectrum methods popularized by Cox *et al.* [11], or the dithered quantization methods developed by Chen and Wornell [16]) and for attempting to remove that information (by means of techniques such as compression, signal warping, and addition of noise.) Perceptual models for audio, imagery, and video have helped quantify the distortions introduced by information–hiding and attack algorithms. Game–theoretic aspects of information hiding have been explored for special cases by Ettinger et al. [17, 18]. Cryptographic aspects of information hiding include the use of secret keys to protect the message. It should however be clearly recognized that the functional requirements of cryptography and information hiding as described above are very different, as secrecy of the message is the main objective in cryptography but is often not a requirement in information hiding, where reliable embedding of the message within the host data is often the single most important requirement. Moreover, while cryptography has received significant attention in the Information Theory community, following Shannon's landmark paper [19], information hiding today is an immature subject, both on a mathematical and a technological level. For instance, there is no consensus today about the formulation of system requirements; and there is considerable uncertainty about the eventual performance of such systems, as all published algorithms in current audio and image watermarking literature can be defeated by attacks such as those in the popular freeware package *Stirmark* [6, 20].

Our analysis of information hiding as a communication problem begins in Sec. 2. There we propose a precise formulation of the information–hiding problem and introduce the notions of *covert channel* and *attack channel* that are respectively designed by the information hider and by the attacker, subject to average distortion constraints. We then seek the maximum rate of reliable transmission through the communication system. This maximum rate is termed *hiding capacity* and is the value of a game played between the information hider and the attacker. In Sec. 3, we characterize the value of this game under various assumptions about the knowledge available to the attacker and to the decoder. These assumptions include possible knowledge of the information–hiding strategy by the attacker, and possible knowledge of the attack strategy by the decoder. It is not our intent to provide an analysis of every possible scenario that may be encountered in practice, but we do investigate a few scenarios that lead to insightful results. To this end, we assume that a cryptographic key is available to the decoder and investigate two extreme but important special cases in some depth: in the first case, the host data themselves are available to the decoder [11]; and in the second case, which we term *blind information hiding*[1], no side information at all is available [21, 22]. The theory is illustrated using an example based on a Bernoulli process and a Hamming distortion function.

In Sec. 4, we extend these results to the case of infinite alphabets. This allows us to treat the case of squared error distortion in Euclidean spaces, which provides considerable insight into the information–hiding problem. In this case, we are able to give explicit formulas for hiding capacity under the assumption of Gaussian host data. These formulas are also upper bounds on hiding capacity for non–Gaussian host data sets. These results show that existing information–hiding schemes in the literature operate far below capacity.

In Sec. 5, we investigate the case of small distortion. This is a problem of considerable interest, as many information–hiding schemes are precisely designed to operate at small distortion levels. We show that the upper bound on hiding capacity in Sec. 4 is in fact tight for non–Gaussian processes.

The results above were derived under the assumption that information hiding strategies and attacks are memoryless, which greatly simplifies the presentation of the main ideas. In Sec. 6, we show that such memoryless strategies are in fact optimal in a certain sense, and we extend our results to a simple but useful class of channels with memory, namely, *blockwise memoryless channels*. We also study extensions of our basic setup to steganography problems and to problems in which the attacker knows the information–hiding codes used. We also present an information–theoretic formulation of the fingerprinting problem and characterize its solution. Conclusions are presented in Sec. 7. The proof of some technical results is given in the appendix.

2 Statement of the Problem

Notation. We use the following notation. Random variables are denoted by capital letters (e.g., X), and their individual values by lowercase letters (e.g., x). The domains over which random variables are defined are denoted by script letters (e.g., $\mathcal{X}$). Sequences of N random variables are denoted with a superscript N (e.g., $X^N = (X_1, X_2, \cdots, X_N)$ takes its values on the product set $\mathcal{X}^N$.) The probability mass function (p.m.f.) of a random variable $X \in \mathcal{X}$ is denoted by $p_X(x), x \in \mathcal{X}$. When no confusion is possible, we drop the subscript in order to simplify the notation. Special letters such as Q and $\tilde{Q}$ are reserved for p.m.f.'s of special interest. Given random variables X, Y, Z, we denote the entropy of X by $H(X)$, the mutual information between X and Y by $I(X;Y)$, and the conditional mutual information between X and Y, conditioned on Z, by $I(X;Y|Z)$ [23]. The Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$.

[1] By analogy to the established terminology "*blind watermarking*".

2.1 Description of the Problem

There are various formulations of information–hiding problems. We consider the following generic version in this paper. Referring to Fig. 1, suppose there is a host–data source producing random variables $\tilde{X}$ taking values in $\mathcal{X}$ according to a known probability mass function (p.m.f.) $p(\tilde{x})$, a cryptographic–key source producing random variables $K \in \mathcal{K}$ distributed as $p(k)$, and a message source producing a message M from a message set $\mathcal{M}$. Specifically, $\tilde{X}$, K and M are as follows.

- In typical problems, $\tilde{X}$ is a block of data or transform data (such as discrete cosine transform coefficients or wavelet coefficients) from an image, video, audio signal, or some other *host data set* in which the information hider embeds information. The set $\mathcal{X}$ could be a continuum (such as the l–dimensional cube $[0,1]^l$) or a discrete set (such as a set of quantized transform coefficients). In Sections 3 and 6, we assume that $\mathcal{X}$ has finite cardinality. The *host data* are a sequence $\tilde{X}^N = (\tilde{X}_1, \cdots, \tilde{X}_N)$ of independent and identically distributed (i.i.d.) samples from $p(\tilde{x})$.

- A cryptographic key $K^N = (K_1, \cdots, K_N)$ is available both at the encoder and the decoder. The individual letters K_i are i.i.d. $p(k), k \in \mathcal{K}$. Even though we refer to K^N as a cryptographic key, the main role of K^N is to provide a source of randomness that is known to the decoder and enables the use of randomized codes. This is a standard communication technique which generally leads to improved transmission performance and is often used to design combat jamming [24, 25]. In information–hiding applications, it is also useful to allow for dependencies between the host data and the key. We model these dependencies using a joint distribution $p(\tilde{x}, k)$. An example of such dependencies arises when the host data are available at the decoder, a common assumption in the watermarking literature [3, 4, 11]. In this case, $\tilde{X}$ is part of the key. Other examples of keys include hash functions [26], location of watermarks [27, 28], and seeds for modulating pseudo–noise sequences in spread–spectrum systems [15, 22]. In blind–information–hiding applications, the decoder is not allowed access to any secret key, so anyone can decode the message [21, 16].

- The message M of interest is uniformly distributed over the message set $\mathcal{M}$ and is to be reliably transmitted to the decoder. M may be a cryptographically encrypted message, in which case an additional cryptographic key is required at the encoder and the decoder.

The information hider passes $\tilde{X}^N$, K^N, and the message m through a function, producing *composite data* X^N that are made publicly available[2].

[2] In watermarking applications, X^N is often referred to as the *watermarked signal*. In the information–hiding literature, M, K, $\tilde{X}$ and X are usually referred to as the mark, the stego–key, the cover–data, and the stego–data, respectively. We prefer not to use the *stego–** terminology to avoid possible confusion: undetectability of the message is a requirement in steganographic applications only.

Next, an attacker passes X^N through a random attack channel $Q^N(y^N|x^N)$ to produce corrupted data Y^N, in an attempt to remove traces of the message M. Clearly this setup includes deterministic attacks of the form $y^N = g_N(x^N)$, where g_N is a deterministic map, as a special case. For deterministic attacks, the only possible values of $Q^N(y^N|x^N)$ are zero and one. An important point is that the attacker would in general benefit from knowing the information–hiding code used by the encoder, and use this knowledge to design a more effective channel $Q^N(y^N|x^N)$. In particular, if the information hiding system did not use a key, the attacker would be able to decode the message and might then be able to remove it from the data X^N. From this point of view, it is desirable to keep the information–hiding code secret. This can be done by including the description of the code as part of the secret key. A similar technique is often used in the design of randomized codes for antijamming systems [25]. Our working assumption is that the attacker knows the distributions of all random variables in the problem but does not know the actual information–hiding code used. For convenience of notation, we shall not explicitly include the description of the code as part of K^N. In some applications, it would be more realistic to assume that the attacker knows or is able to learn the information–hiding code; this scenario is briefly considered in Sec. 6.2.

We assume that X and Y take their values in the same set $\mathcal{X}$ as $\tilde{X}$ does [3]. The decoder computes an estimate $\hat{m}$ for the message that was originally transmitted.

Constrained Information–Hiding Systems. The encoders used almost universally in the watermarking literature use a constrained encoder shown in Fig. 2 [7, 11, 21, 22]. First, a mapping from M to codewords $U^N(M) \in \mathcal{X}^N$ is defined. This mapping is independent of $\tilde{X}^N$. The composite data are obtained as $x^N = f_N(\tilde{x}^N, u^N, k^N)$. A simple example would be $x_i = \tilde{x}_i + u_i$ for $1 \leq i \leq N$, when $\mathcal{X}$ is a finite field. More elaborate designs of f_N in the image watermarking literature exploit the perceptual characteristics of the human visual system and are not additive [7, 4]. While such designs make it convenient for the information hider to satisfy distortion constraints, heuristic choices can be largely suboptimal in terms of achievable rates. Moreover, the maps $f_N(\cdot, u^N, k^N)$ and $f_N(\tilde{x}^N, \cdot, k^N)$ are often invertible, meaning that the original host data $\tilde{x}^N$ (resp. u^N) can be recovered if u^N (resp. $\tilde{x}^N$) and k^N are known. As our subsequent analysis shows, the above restrictions on the alphabet for $U^N(m)$ and on the information–hiding code f_N can substantially reduce the maximum rate of reliable transmission.

2.2 Distortion Constraints

We now formally define the constraints on the information–hiding and attack strategies. This completes the mathematical description of the information hiding problem and allows us to derive achievable rates of reliable transmission for the communication system in Fig. 1.

Definition 2.1 *A distortion function is a nonnegative function* $d : \mathcal{X} \times \mathcal{X} \to I\!\!R_+$.

[3] Our results can be extended without difficulty to the important practical case where $\tilde{X}$, X, and Y takes their values on different subsets of $\mathcal{X}$.

The distortion functions considered are bounded: $d_{max} = \max_{(x,y)\in\mathcal{X}\times\mathcal{X}} d(x,y) < \infty$. Other properties of interest are stated explicitly where applicable, including symmetry: $d(x,y) = d(y,x)$ for all $x, y \in \mathcal{X}$, and the condition $d(x,y) = 0 \Leftrightarrow x = y$. The distortion function is extended in a classical way to a distortion on N–tuples $x^N = (x_1, x_2, \ldots, x_N)$ by $d^N(x^N, y^N) = \frac{1}{N}\sum_{k=1}^{N} d(x_k, y_k)$. The theory developed in this paper applies to classical distortion functions such as the Hamming distance, as well as to any arbitrarily complicated perceptual (auditory or visual) distortion function that would satisfy the technical conditions above. Note that the condition $d(x,y) = 0 \Leftrightarrow x = y$ is not satisfied by perceptual distortion functions in image processing, due to the presence of threshold effects in the human visual system [4].

Definition 2.2 *A length–N information–hiding code subject to distortion D_1 is a triple $(\mathcal{M}, f_N, \phi_N)$, where:*

- *$\mathcal{M}$ is the message set of cardinality $|\mathcal{M}|$;*
- *$f_N : \mathcal{X}^N \times \mathcal{M} \times \mathcal{K}^N \to \mathcal{X}^N$ is the encoder mapping a sequence $\tilde{x}^N$, a message m, and a key k^N to a sequence $x^N = f_N(\tilde{x}^N, m, k^N)$. This mapping is subject to the distortion constraint*

$$\sum_{\tilde{x}^N\in\mathcal{X}^N}\sum_{k^N\in\mathcal{K}^N}\sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{M}|}\, p(\tilde{x}^N, k^N)\, d^N(\tilde{x}^N, f_N(\tilde{x}^N, m, k^N)) \le D_1;$$

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

Note there are distortion constraints but no rate constraints on X^N. Typically D_1 is small, as the embedding of information within the host data set is intended to be imperceptible to a casual analysis. (In watermarking applications, this is known as *transparent watermarking.*) Also note that the definition of the distortion constraints involves an averaging with respect to the distribution $p(\tilde{x}^N, k^N)$ and with respect to the uniform distribution on the messages. This choice is made for convenience as it allows us to use classical tools from Shannon theory. Also possible but more difficult to analyze would be the use of maximum distortion constraints, where the maximum is with respect to $\tilde{x}^N, k^N$, and m. The distribution $p(\tilde{x}^N, k^N)$ and the choice of the encoder mapping f_N induce a distribution $p(x^N)$ on the composite data set.

Definition 2.3 *An attack channel with memory, subject to distortion D_2, is a sequence of conditional p.m.f.'s $Q^N(y^N|x^N)$ from $\mathcal{X}^N$ to $\mathcal{X}^N$, such that*

$$\sum_{x^N\in\mathcal{X}^N}\sum_{y^N\in\mathcal{X}^N} d^N(x^N, y^N)\, Q^N(y^N|x^N)\, p(x^N) \le D_2 \tag{2.1}$$

for all $N \ge 1$.

If the distortion function is symmetric, it is reasonable to also require that $D_2 \geq D_1$, as the range of options offered to the attacker should include recovering the host data. Under the definition (2.1), the attack channel is subject to a constraint on the average distortion between X^N and Y^N. Another possibility, briefly considered further in this paper, is to constrain the average distortion between the host data $\tilde{X}^N$ and Y^N:

$$\sum_{m,k^N,\tilde{x}^N,y^N} d^N(\tilde{x}^N, y^N)\, Q^N(y^N|f_N(\tilde{x}^N, m, k^N))\, p(\tilde{x}^N, k^N) \leq D_2. \tag{2.2}$$

This constraint only makes senses if the attacker knows f_N. Having defined the information–hiding code f_N and the attack channel $Q^N(y^N|x^N)$ as in (2.1), we can now define the game between the information hider and the attacker.

Definition 2.4 *An information–hiding game subject to distortions (D_1, D_2) consists of an information–hiding code subject to distortion D_1 and an attack channel subject to distortion D_2.*

The rate of the code is $R = \frac{1}{N} \log |\mathcal{M}|$. The average probability of error is

$$P_{e,N} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} P(\phi_N(Y^N, K^N) \neq m \mid M = m) \tag{2.3}$$

where M is a random variable in $\mathcal{M}$ with a uniform probability distribution. That is, $P_{e,N}$ equals the probability that an attack successfully removes the message, averaged over all messages. Rates of reliable transmission and hiding capacity are defined as follows.

Definition 2.5 *A rate R is achievable for distortions (D_1, D_2), if there is a sequence of codes subject to distortion D_1, with rates $R_N > R$ such that $P_{e,N} \to 0$ as $N \to \infty$, for any attack subject to distortion D_2.*

Definition 2.6 *The information–hiding capacity $C(D_1, D_2)$ is the supremum of all achievable rates for distortions (D_1, D_2).*

3 Main Results

We have just given an operational definition of information–hiding capacity. In this section, we show that this capacity is the value of a game between the information hider and the attacker. The payoff function in this game is a difference between mutual informations. In order to maximize the payoff, the information hider optimally designs the *covert channel* defined in Sec. 3.1. In order to minimize the payoff, the attacker designs an optimal memoryless attack channel, defined in Sec. 3.2. An expression for the information–hiding capacity is derived in terms of the covert and attack channels in Sec. 3.3 and is applied to a simple example in Sec. 3.4. Game–theoretic aspects of the information–hiding problem are studied in Sec. 3.5.

3.1 Covert Channel

Consider an auxiliary random variable U defined over an arbitrary, finite set $\mathcal{U}$. The role of U will become apparent in Sec. 3.3. We now define the notion of covert channel. Note that the distortion constraint below only involves the marginal distribution $\sum_u \tilde{Q}(x, u|\tilde{x}, k)$.

Definition 3.1 *A covert channel subject to distortion D_1 is a conditional p.m.f. $\tilde{Q}(x, u|\tilde{x}, k)$ from $\mathcal{X} \times \mathcal{K}$ to $\mathcal{X} \times \mathcal{U}$, such that*

$$\sum_{x,\tilde{x},k,u} d(\tilde{x}, x)\tilde{Q}(x, u|\tilde{x}, k)p(\tilde{x}, k) \leq D_1. \tag{3.1}$$

The length–N memoryless extension of the channel is the conditional p.m.f.

$$\tilde{Q}^N(x^N, u^N|\tilde{x}^N, k^N) = \prod_{i=1}^{N} \tilde{Q}_i(x_i, u_i|\tilde{x}_i, k_i), \quad \forall N \geq 1.$$

Definition 3.2 *The compound covert channel subject to distortion D_1 is the set $\tilde{\mathcal{Q}}$ of all covert channels satisfying (3.1).*

To analyze the information–hiding system, we find it convenient to write $\tilde{Q}$ in the cascaded form

$$\tilde{Q}(x, u|\tilde{x}, k) = p(x|\tilde{x}, u, k)p(u|\tilde{x}, k), \tag{3.2}$$

where we refer to $p(x|\tilde{x}, u, k)$ as the *main covert channel* and to $p(u|\tilde{x}, k)$ as the *auxiliary covert channel.* Note that the distortion constraint (3.1) involves $\tilde{Q}(x, u|\tilde{x}, k)$ via the main covert channel only.

3.2 Attack Channel

Definition 3.3 *An attack channel subject to distortion D_2 is a conditional p.m.f. $Q(y|x)$ from $\mathcal{X}$ to $\mathcal{X}$, such that*

$$\sum_{x,y} d(x, y)Q(y|x)p(x) \leq D_2. \tag{3.3}$$

The length–N memoryless extension of the attack channel is the conditional p.m.f.

$$Q^N(y^N|x^N) = \prod_{i=1}^{N} Q_i(y_i|x_i).$$

Definition 3.4 *The compound attack channel subject to distortion D_2 is the set $\mathcal{Q}$ of all attack channels of the form (3.3).*

3.3 Hiding Capacity

We now introduce Theorem 3.3, which we view as the basic theorem of information hiding. For any arbitrarily complicated encoding scheme and memoryless attack, this theorem upper bounds the rate of reliable transmission for the information hider, under the assumptions that the attacker knows the covert channel, and that the decoder knows both the covert channel and the attack channel. These assumptions are weaker than they appear at first sight. Even if the transition probabilities of the covert and attack channels are not given, it is to be expected that the attacker and the decoder will be able to learn them, provided that N is large enough. While these assumptions are reasonable, they might not always be satisfied in practice. We return to this issue, and to game–theoretic aspects of this problem, in Sec. 3.5.

The proof of Theorem 3.3 relies on a proof of achievability (Prop. 3.1) and a converse (Prop. 3.2), for a fixed attack channel. Achievability is proved using a random bin coding technique [23, p. 410]. The decoder uses joint typical set decoding. The proof of Prop. 3.1 is an adaptation of techniques used by Gel'fand and Pinsker [29] to derive the capacity of a discrete memoryless channel with random parameters (state of the channel) that are known at the encoder but not at the decoder, see Fig. 3. The capacity of this channel is given by $C = \max_{p(x,u|s)}[I(U;Y) - I(U;S)]$ where S is the random channel parameter ("state" of the channel), and U is an auxiliary random variable. These results have been extended by Heegard and El Gamal to find the capacity of computer memory with defects [30]. In the information–hiding problem, the host data $\tilde{X}$ plays the role of the random parameter S in Gel'fand and Pinsker's work. The analogy between Gel'fand and Pinsker's work and watermarking problems was first identified by Chen [9]. The converse theorem (Proposition 3.2) cannot be proved using the same technique as in [29, 31], because the attack channel is not known to the encoder, so we need the Markov chain property $(U, \tilde{X}, K) \to X \to Y$ (see proof of Theorem 3.3.) Other differences between our setup and Gel'fand and Pinsker's include the presence of distortion constraints, and the availability of side information (key) at the encoder and the decoder. Proofs of both propositions can be found in the appendix.

Proposition 3.1 *(Achievability for Fixed Memoryless Attack Channel).*
Fix the attack channel $Q(y|x)$ and select $\tilde{Q}(x,u|\tilde{x},k)$ that maximizes

$$J(\tilde{Q}, Q) = I(U;Y|K) - I(U;\tilde{X}|K) \tag{3.4}$$

over $\tilde{\mathcal{Q}}$. For any $\epsilon > 0$ and sufficiently large N, there exists a length–N information–hiding code $(\mathcal{M}, f_N, \phi_N)$ with $P_{e,N} < \epsilon$, and $|\mathcal{M}| > 2^{N[I(U;Y|K) - I(U;\tilde{X}|K) - \epsilon]}$.

Proposition 3.2 *(Converse for Fixed Memoryless Attack Channel).*
Consider a length–N information–hiding code $(\mathcal{M}, f_N, \phi_N)$ with rate $R = \frac{1}{N}\log|\mathcal{M}|$, and an attack channel $Q(y|x)$. If for any $\epsilon > 0$ we have $P_{e,N} < \epsilon$ as $N \to \infty$, then there exists a finite alphabet $\mathcal{U}$ and a covert channel $\tilde{Q}(x,u|\tilde{x},k) \in \tilde{\mathcal{Q}}$ such that $R \le I(U;Y|K) - I(U;\tilde{X}|K)$.

Theorem 3.3 *Assume the attacker knows the covert channel and the decoder knows the covert and attack channels. For any information hiding game subject to distortions (D_1, D_2), a rate R is achievable if and only if $R < \underline{C}$, where $\underline{C}$ is given by*

$$\underline{C} = \max_{\tilde{Q}(x,u|\tilde{x},k) \in \tilde{\mathcal{Q}}} \min_{Q(y|x) \in \mathcal{Q}} J(\tilde{Q}, Q) \tag{3.5}$$

where U is a random variable defined over an arbitrary finite alphabet $\mathcal{U}$, $(U, \tilde{X}, K) \to X \to Y$ is a Markov chain, and $J(\tilde{Q}, Q)$ is given by (3.4).

Proof. Fix the attack channel $Q \in \mathcal{Q}$. Proposition 3.1 shows that all rates below $\max_{\tilde{Q} \in \tilde{\mathcal{Q}}} J(\tilde{Q}, Q)$ are achievable. Proposition 3.2 implies the converse result, namely, no reliable transmission is possible above that rate. Since the attacker knows $\tilde{Q}$, he is able to choose Q so as to minimize the rate of transmission, under any possible information–hiding strategy. The information hider uses a minimum-size message set $\mathcal{M}$ (whose rate is given by (3.5)), and his encoding strategy is independent of the actual attack channel used (see proof of Prop. 3.1.) Hence the rate of reliable transmission is given by (3.5). Note that the minimum over Q can be attained because (3.4) is a bounded, continuous function of Q, and the admissible set $\mathcal{Q}$ is compact. Dually, the maximum of $\min_Q J(\tilde{Q}, Q)$ over $\tilde{Q}$ can be attained because $\min_Q J(\tilde{Q}, Q)$ is a nonnegative, continuous function of $\tilde{Q}$, and the admissible set $\tilde{\mathcal{Q}}$ is compact. □

Corollary 3.4 below shows that in the important special case $K = \tilde{X}$ (host data available at the decoder), there is no loss in optimality in restricting the encoder to be in the constrained class of Fig. 2. The solution becomes a saddlepoint of $I(X; Y|\tilde{X})$.

Corollary 3.4 *In the special case $K = \tilde{X}$, U is optimal if and only if X can be written in the form $X = f(\tilde{X}, U)$, where the map $f(\tilde{x}, \cdot)$ is invertible for all $\tilde{x}$. In particular, the design $U = X$ is optimal. The hiding capacity is given by*

$$C = \max_{p(x|\tilde{x})} \min_{Q(y|x)} I(X; Y|\tilde{X}) = \min_{Q(y|x)} \max_{p(x|\tilde{x})} I(X; Y|\tilde{X}). \tag{3.6}$$

Proof. When $K = \tilde{X}$, the payoff function (3.4) can be written as

$$I(U; Y|\tilde{X}) - I(U; \tilde{X}|\tilde{X}) = I(U; Y|\tilde{X}) = I(U, \tilde{X}; Y|\tilde{X}) \leq I(X; Y|\tilde{X}), \tag{3.7}$$

where the inequality follows from the data processing inequality applied to the Markov chain $(U, \tilde{X}) \to X \to Y$. The inequality is satisfied with equality when $U = X$. Hence the capacity is given by the first inequality in (3.6). The objective function $I(X; Y|\tilde{X})$ is convex in $Q(y|x)$ (for fixed $p(x|\tilde{x})$) and concave in $p(y|\tilde{x}) = \sum_x Q(y|x) p(x|\tilde{x})$, and hence concave in $p(x|\tilde{x})$ (for fixed $Q(y|x)$). Moreover, the sets $\tilde{\mathcal{Q}}$ and $\mathcal{Q}$ are convex. By application of the Von Neuman's min-max theorem [34], the order of maximization and minimization can be switched. □

The hiding capacity (3.5) clearly depends on (D_1, D_2), which can be emphasized by writing $C(D_1, D_2)$. The function $C(D_1, D_2)$ satisfies the following elementary properties.

1. $C(D_1, D_2)$ is monotonically increasing in D_1 and monotonically decreasing in D_2.

2. $C(D_1, D_2)$ is convex in D_2.

3. $C(D_1, D_2)$ is upper–bounded by the entropy of Y:

$$C(D_1, D_2) \leq \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \min_{Q \in \mathcal{Q}} H(Y) \leq \log |\mathcal{X}|.$$

   Informally, it is easier to hide information in complex datasets than in simple ones.

4. If the distortion function satisfies the condition $d(x, y) = 0 \Leftrightarrow x = y$, we have $X = \tilde{X}$ when $D_1 = 0$, i.e., no information can be transmitted. In this case, $C(0, D_2) = 0$ for all D_2.

5. For any fixed D_1, if D_2 is large enough, there exists an admissible attack that generates Y^N independently of X^N, thereby eliminating all traces of the message. More precisely, let $D_2^* = \min_y \sum_x d(x, y)p(x)$ and y^* be the value of y that achieves this minimum. For all $D_2 \geq D_2^*$, the deterministic attack $Q(y|x) = \delta(y - y^*)$ is admissible, and $C(D_1, D_2) = 0$ for all D_1. Our previous assumptions imply that D_2^* is finite but depends on D_1.

Comments.

1. Theorem 3.3 states that the problem reduces to a rate-distortion–constrained capacity game. The optimal attack is the solution to a rate-distortion problem, indicating the important role of data compression in the theory. The optimal information–hiding strategy is the solution to a constrained capacity problem.

2. The information–hiding problem is also related to the wiretapping problem in cryptography [24, p. 407]. The wiretapping problem involves communication of data to a receiver via a direct channel, and presence of a wiretapper that observes these data through a second channel, and knows the codes used by the transmitter. The secrecy capacity is the maximum rate of reliable transmission under the constraint that the equivocation rate for the wiretapper is maximum. The secrecy capacity is zero if the wiretapper's channel is less noisy than the direct channel. In the information–hiding problem, the attacker does more than being a simple wiretapper as he maliciously degrades the direct channel from M to Y^N, which is at least as noisy as the wiretapper's channel (from M to X^N). The secrecy capacity is zero unless cryptographic keys are used.

3. Choosing U independent of $\tilde{X}$, as in conventional watermarking algorithms (see Fig. 2), would in general be suboptimal. Evaluation of (3.4) shows that rates of reliable transmission in that case are upper bounded by $I(U; Y|K)$.

4. Assume a symmetric distortion measure, and the condition $D_2 \geq D_1$. If the attacker somehow knew the host signal $\tilde{X}^N$, the optimal attack would consist in setting $Y^N = \tilde{X}^N$. In this case, the output of the attack no longer contains any trace of the message, and the hiding capacity is zero. This simple observation motivates a potentially powerful attack, in which the attacker attempts to construct an estimate of $\tilde{X}^N$. Specifically, if the attacker is able to construct Y such that $H(Y|\tilde{X}) < \epsilon$, the payoff is upper–bounded by

$$\begin{aligned} & I(U;Y|K) - I(U;\tilde{X}|K) \\ & \quad = [I(U;\tilde{X},Y|K) - I(U;\tilde{X}|Y,K)] - [I(U;\tilde{X},Y|K) - I(U;Y|\tilde{X},K)] \\ & \quad \leq I(U;Y|\tilde{X},K) \leq H(Y|\tilde{X},K) \leq H(Y|\tilde{X}) < \epsilon, \end{aligned} \tag{3.8}$$

for all U. Hence $C < \epsilon$.

Proposition 3.5 below shows that the payoff (3.4) is convex in Q but is neither convex nor concave in $\tilde{Q}$. However, it is useful to decompose $\tilde{Q}$ as indicated in (3.2). We then have the following properties. See the appendix for a proof.

Proposition 3.5 *(Convexity Properties of Payoff Function)*

(i) *The sets $\tilde{\mathcal{Q}}$ and $\mathcal{Q}$ are convex.*

(ii) *For fixed $p(\tilde{x},k)$ and $\tilde{Q}(x,u|\tilde{x},k)$, the payoff (3.4) is convex in $Q(y|x)$.*

(iii) *For fixed $p(\tilde{x},k)$, $p(x|\tilde{x},u,k)$ and $Q(y|x)$, (3.4) is concave in $p(u|\tilde{x},k)$.*

(iv) *For fixed $p(\tilde{x},k)$, $p(u|\tilde{x},k)$ and $Q(y|x)$, (3.4) is convex in $p(x|\tilde{x},u,k)$.*

(v) *To evaluate the maximum of (3.4) over $\tilde{Q}$, one can restrict $\mathcal{U}$ to sets of cardinality $2|\mathcal{X}| + |\mathcal{K}|$.*

(vi) *For fixed $Q(y|x)$, (3.4) is maximized by $\tilde{Q}(x,u|\tilde{x},k) = p(x|\tilde{x},u,k)p(u|\tilde{x},k)$ that has the following property. The cardinality of the support set $\mathcal{X}(\tilde{x},u,k)$ of $p(x|\tilde{x},u,k)$ is equal to one*[4] *for all $(\tilde{x},u,k)$, except perhaps for one single value of $(\tilde{x},u,k)$, for which the cardinality of $\mathcal{X}(\tilde{x},u,k)$ is equal to two. If the latter condition is not satisfied, the distortion constraint (3.1) is inactive.*

3.4 Example: Binary Channel

We illustrate Theorem 3.3 and Corollary 3.4 using a simple problem involving a binary alphabet $\mathcal{X} = \{0,1\}$ and a Bernoulli source $\tilde{X}$ with parameter $\frac{1}{2}$. The distortion function is Hamming distance: $d(x,y) = 0$ if $x = y$ and $d(x,y) = 1$ otherwise. We assume the host data are available at the decoder: $K = \tilde{X}$. The problem is illustrated in Fig. 4, and its solution is stated in Prop. 3.6. One can think of the random variable Z as a direct representation of the message M itself.

[4] In other words, x is a deterministic function of $(\tilde{x},u,k)$.

Proposition 3.6 *For $D_1, D_2 \leq \frac{1}{2}$, the hiding capacity for the information–hiding game above is given by*

$$C = \underline{H}(D_1 \star D_2) - \underline{H}(D_2)$$

where we let $\underline{H}(t) = -t \log t - (1-t)\log(1-t)$, and $D_1 \star D_2 = D_1(1-D_2) + D_2(1-D_1)$. The capacity–achieving distributions are as follows. The composite data is given by $X = \tilde{X} \oplus Z$, where $\oplus$ denotes exclusive or*, and Z is a Bernoulli(D_1) random variable. Both $U = X$ and $U = Z$ are optimal. The optimal attack is given by $Y = X \oplus W$, where W is a Bernoulli(D_2) random variable. For $D_1 \geq \frac{1}{2}$ and $D_2 < \frac{1}{2}$, the hiding capacity is $C = 1 - \underline{H}(D_2)$. For $D_2 \geq \frac{1}{2}$, the hiding capacity is equal to zero.*

Proof. By application of Corollary 3.4, we need to verify that the distributions above for $Z = \tilde{X} \oplus X$ and $W = X \oplus Y$ are a saddlepoint of the payoff $I(X;Y|\tilde{X})$. First we fix $Q(y|x)$. Assuming $D_1, D_2 \leq \frac{1}{2}$, we have

$$\begin{aligned} I(X;Y|\tilde{X}) &\overset{(a)}{=} H(Y|\tilde{X}) - H(Y|X,\tilde{X}) \\ &\overset{(b)}{=} H(Y|\tilde{X}) - H(Y|X) \\ &= H(Y - \tilde{X}|\tilde{X}) - H(W) \\ &= H(Z \oplus W|\tilde{X}) - \underline{H}(D_2) \\ &\overset{(c)}{\leq} H(Z \oplus W) - \underline{H}(D_2) \\ &\overset{(d)}{\leq} \underline{H}(D_1 \star D_2) - \underline{H}(D_2) \end{aligned}$$

where (a) is by definition of the conditional mutual information, (b) is because $\tilde{X} \to X \to Y$ is a Markov chain, and inequality (c) holds because conditioning reduces entropy. Equality is achieved in (c) if and only if $Z \oplus W$, hence Z, is independent of $\tilde{X}$. Inequality (d) holds because Z and W are independent (as $Z \to X \to W$ forms a Markov chain), and $Pr[Z=1] \leq D_1$. Equality is achieved if Z is Bernoulli(D_1). The distribution $p(x|\tilde{x})$ above satisfies both inequalities with equality and hence maximizes $I(X;Y|\tilde{X})$.

The second step is to fix $p(x|\tilde{x})$ and minimize $I(X;Y|\tilde{X})$ over $Q(y|x)$. With $p(x|\tilde{x})$ specified as above, Z and X are independent. Since $Z \to X \to W$ forms a Markov chain, Z and W are also independent. We have

$$\begin{aligned} I(X;Y|\tilde{X}) &= I(X \oplus \tilde{X}; Y \oplus \tilde{X}|\tilde{X}) \\ &= H(Z) - H(Z|Z \oplus W, \tilde{X}) \\ &\overset{(a)}{\geq} H(Z) - H(Z|Z \oplus W) \\ &= I(Z; Z \oplus W) \\ &\overset{(b)}{\geq} \underline{H}(D_1 \star D_2) - \underline{H}(D_2) \end{aligned}$$

where inequality (a) holds because conditioning reduces entropy, and inequality (b) holds because Z and W are independent, and $Pr[W = 1] \leq D_2$, with equality if W is Bernoulli(D_2).

The derivation of hiding capacity in the cases $D_1 \geq \frac{1}{2}$ or $D_2 \geq \frac{1}{2}$ is straightforward. □

The encoding system for X above is the same as Vernam's *one–time pad* encryption system [32]. The distributions $p(Z)$ and $p(Z|X)$ are identical, which means that this system satisfies Shannon's *perfect secrecy* condition [19, 32]: observing the data X does not provide the attacker with any information about the message Z. Moreover, the encoding system above satisfies the basic requirement of a steganographic system: the distributions $p(\tilde{x})$ and $p(x)$ are identical, so it is impossible for an observer (such as the attacker) to determine whether the data x^N were drawn from $p(\tilde{x}^N)$ or from $p(x^N)$.

3.5 Game–Theoretic Formulation of the Information Hiding Problem

The information hider selects an alphabet $\mathcal{U}$ and a covert channel $\tilde{Q}(x, u|\tilde{x}, k)$ from $\tilde{\mathcal{Q}}$. The attacker selects an attack channel $Q(y|x)$ from $\mathcal{Q}$. In Theorem 3.3, it is assumed that the attacker knows $\tilde{Q}$, and the decoder knows both $\tilde{Q}$ and Q. These assumptions are reasonable but may not be satisfied in practice. Let us examine some alternative assumptions the information hider and the attacker could make.

Information Hider. The information hider wants a guaranteed rate of reliable transmission under any attack that satisfies the distortion constraint (3.3). Therefore he may want to design the covert channel under the conservative assumption that the attacker knows the covert channel used. Such an approach is often referred to as a secure strategy in game theory [34]. The rate the information hider is able to guarantee is given by (3.5), which is restated below for convenience:

$$\underline{C} = \max_{\tilde{Q}(x,u|\tilde{x},k)\in\tilde{\mathcal{Q}}} \min_{Q(y|x)\in\mathcal{Q}} J(\tilde{Q}, Q).$$

For the rate (3.5) to be achievable, the decoder must know the attack channel used, or be able to estimate it reliably from the received data. (Likewise, the attacker must know the covert channel used, or be able to estimate it reliably from the composite data.) In order to reliably estimate $Q \in \mathcal{Q}$, a *universal decoder* over the class $\mathcal{Q}$ would have to be available. There is a well–developed theory of universal decoding for compound channels [24, 33], but extension of this theory and development of universal decoding algorithms for information hiding is still an open problem.

Attacker. The attacker wants to minimize the rate of reliable transmission for any information–hiding strategy that satisfies the distortion constraint (3.1). Therefore he may design the attack channel under the conservative assumption that the information hider and the decoder know the attack channel used. Under this assumption, the attacker can guarantee that information cannot be reliably transmitted at a rate greater than

$$\overline{C} = \min_{Q(y|x)\in\mathcal{Q}} \max_{\tilde{Q}(x,u|\tilde{x},k)\in\tilde{\mathcal{Q}}} J(\tilde{Q}, Q). \tag{3.9}$$

This assumption would be overly conservative in many information–hiding problems.

Saddlepoint. The capacities in (3.5) and (3.9) are respectively lower and upper values of the game [34]. If these capacities are equal, as in Corollary 3.4, the common value is a saddlepoint of the game. The information hider and the attacker will respectively choose a channel $\tilde{Q}^*(x,u|\tilde{x},k)$ and a channel $Q^*(y|x)$ that satisfy the saddlepoint condition

$$J(\tilde{Q},Q^*) \le J(\tilde{Q}^*,Q^*) \le J(\tilde{Q}^*,Q), \quad \forall\tilde{Q} \in \tilde{\mathcal{Q}}, \forall Q \in \mathcal{Q}. \tag{3.10}$$

This is a stable Nash equilibrium of the game. It is in the interest of neither party to deviate from a saddlepoint strategy [34]. For other examples of saddlepoint strategies in information–theoretic games, see [35, 36] [23, p. 263].

Mixed Strategies. According to Proposition 3.5, the payoff $J(\tilde{Q},Q)$ is always convex in Q, and in some cases, is concave in $\tilde{Q}$ (see for instance Corollary 3.4.) In such cases, the game admits a unique saddlepoint, and the capacities in (3.5) and (3.9) are equal. Consider now the general case when the payoff is not concave in $\tilde{Q}$, and neither player knows the strategy the opponent will be using. This suggests the use of a randomized strategy (or *mixed strategy* in game–theoretic terminology) as an alternative to the *pure strategies* above. In a randomized strategy, the information hider and the attacker respectively select strategies $\tilde{Q}$ and Q according to properly selected probability distributions $\tilde{P}(\tilde{Q})$ and $P(Q)$, and use these strategies throughout the transmission. In other words, the covert and attack channels are no longer ergodic. For instance, a set of possible strategies for the attacker may include attacks from the *Stirmark* software package [20] (when the host data set is an image), such as an additive white Gaussian noise attack, a signal-dependent random attack, and a JPEG compression attack; and a set of possible strategies for the information hider may include a spread-spectrum encoding strategy [11] as well as a dithered-quantization strategy [16]. For mixed strategies, the payoff is $j(\tilde{P},P) = \int\int J(\tilde{Q},Q)\,d\tilde{P}(\tilde{Q})\,dP(Q)$, to be maximized over $\tilde{P}$ by the information hider and minimized over P by the attacker. Pure strategies are a special case of mixed strategies in which $\tilde{P}(\tilde{Q})$ and $P(Q)$ are point mass functions. We impose the (nonergodic) average distortion constraints

$$\int \sum_{x,\tilde{x},u,k} \tilde{Q}(x,u|\tilde{x},k)p(\tilde{x},k)d(\tilde{x},x)\,d\tilde{P}(\tilde{Q}) \quad \le \quad D_1, \tag{3.11}$$

$$\int \sum_{x,y} Q(y|x)p(x)d(x,y)\,dP(Q) \quad \le \quad D_2, \tag{3.12}$$

on $\tilde{P}(\tilde{Q})$ and $P(Q)$ instead of the constraints (3.1) and (3.3). The advantage of (3.11) and (3.12) is that only two distortion constraints need to be considered, instead of one for every possible pair $(\tilde{Q},Q)$, as in (3.1) (3.3). However, a meaningful formulation of the game is difficult due to the fact that the admissible set for $P(Q)$ depends on $\tilde{P}(\tilde{Q})$ via the distribution $p(x)$. In game–theoretic terminology, the constrained sets are coupled, or nonrectangular [34, p. 175—177].

Fortunately, in some instances, the coupling between the constrained sets can be neglected. This is the case for the small–distortion problem in Sec. 5, where $p(x)$ asymptotically tends to $p(\tilde{x})$ which is independent of the information–hiding strategy. If the coupling between constrained sets can be neglected, game–theoretic analysis yields the following results. First note that the function $J(\tilde{Q}, Q)$ is continuous and bounded from above and below, and its arguments range over a compact subset (probability simplex) of the finite–dimensional Euclidean space $I\!R^{|\mathcal{X}|}$. Using the extension of Von Neuman's min-max theorem to such function sets [34, Sec. 4.3], we obtain $\sup_{\tilde{P}} \inf_P j(\tilde{P}, P) = \inf_P \sup_{\tilde{P}} j(\tilde{P}, P) = C$. The supremum and infimum are achieved by some pair of strategies $\tilde{P}(\tilde{Q})$ and $P(Q)$, and the value of the game is C. These strategies are optimal for both players, and we have $\underline{C} \leq C \leq \overline{C}$. As discussed above, if $J(\tilde{Q}, Q)$ is concave/convex, then the optimal strategies are *pure strategies*, i.e., $\tilde{P}(\tilde{Q})$ and $P(Q)$ are point mass functions. In general, $J(\tilde{Q}, Q)$ is convex in Q but not concave in $\tilde{Q}$. This implies the optimal strategy for the attacker is a pure strategy, while the optimal strategy for the information hider is a mixed strategy. At this point, we recall that the use of a mixed strategy for the communicator is a classical one in jamming problems. Randomized codes are an example of such an approach [25] [33, p. 2163]. Here randomization is due to the use of a probabilistic distribution over $\tilde{Q}$ rather than to the use of cryptographic keys, as discussed in the Introduction. Fig. 5 illustrates the game between the information hider and the attacker. We reemphasize that mixed strategies are useful only if neither player knows the strategy used by the adversary.

4 Continuous Alphabets

The results above can be extended to the case of infinite alphabets $\mathcal{X}$,$\mathcal{U}$,$\mathcal{K}$. This is done by extending the definition of mutual information to such alphabets [37, Ch. 7.2]:

$$I(U; Y|K) = \sup I(U_d; Y_d|K_d) \quad \text{and} \quad I(U; \tilde{X}|K) = \sup I(U_d, \tilde{X}_d|K_d)$$

where the supremum is over all partitions $\tilde{X}_d$, K_d, X_d and Y_d of the channel input and output alphabets, and over all finite–alphabet variables U_d. If all probability density functions are "well behaved", the results from Sec. 3 still apply, provided that sums are replaced with integrals.

The case of Gaussian $\tilde{X}$ and squared–error distortion measure $d(x, y)$ is of considerable interest, as it becomes possible to explicitly compute the distributions that achieve capacity. We refer to this case as the Gaussian channel. Here $\mathcal{X}$ is the set $I\!R$ of real numbers, $\tilde{X} \sim \mathcal{N}(0, \sigma^2)$, and $d(x, y) = (x - y)^2$. We consider two special cases. In the first one, the key K is the host data $\tilde{X}$ itself. In the second case (blind information hiding), no key is available at the decoder.

The case of non–Gaussian $p(\tilde{x})$ is of course more difficult, but useful results can still be obtained. In particular, lower bounds on hiding capacity can be computed by evaluating the optimal attack under a particular, generally suboptimal, information–hiding strategy $\tilde{Q}^*$. Likewise, upper bounds can be computed by evaluating the optimal information–hiding strategy under a particular (generally suboptimal) attack Q^*:

$$C_{lb} = \min_{Q(y|x) \in \mathcal{Q}} J(\tilde{Q}^*, Q) \leq \underline{C} \leq C_{ub} = \max_{\tilde{Q}(x,u|\tilde{x},k) \in \tilde{\mathcal{Q}}} J(\tilde{Q}, Q^*). \tag{4.1}$$

We found this bounding technique to be very useful for non–Gaussian channels, provided that $\tilde{Q}^*$ and Q^* are suitably selected (see Sec. 5). Of course, if the lower and upper bounds in (4.1) are equal, the pair $(\tilde{Q}^*, Q^*)$ is a saddlepoint of the payoff (3.4).

4.1 Host Signal is Available at the Decoder

Theorem 4.1 gives the hiding capacity when the host data is available at the decoder. The capacity–achieving distributions for Gaussian $p(\tilde{x})$ are given in Fig. 6.

Theorem 4.1 *Let $\mathcal{X} = I\!R$ and $d(x, y) = (x - y)^2$ be the squared–error distortion measure. Assume that $K = \tilde{X}$. Then we have*
(i) If $\tilde{X} \sim \mathcal{N}(0, \sigma^2)$, the hiding capacity is given by

$$\underline{C} = \overline{C} = C = \begin{cases} \frac{1}{2}\log\left(1 + \frac{D_1}{\beta D_2}\right) & : \textit{ if } D_2 < \sigma^2 + D_1, \\ 0 & : \textit{ otherwise} \end{cases} \tag{4.2}$$

where $\beta = \left(1 - \frac{D_2}{\sigma^2 + D_1}\right)^{-1}$. The optimal covert channel is given by $X = \tilde{X} + Z$, where $Z \sim \mathcal{N}(0, D_1)$ is independent of $\tilde{X}$. The optimal attack is the Gaussian test channel,

$$Q^*(y|x) = \begin{cases} \mathcal{N}(\beta^{-1}x, \beta^{-1}D_2) & : \textit{ if } D_2 < \sigma^2 + D_1 \\ \delta(y) & : \textit{ else.} \end{cases} \tag{4.3}$$

(ii) If $\tilde{X}$ is non–Gaussian with mean zero and variance σ^2, (4.2) is an upper bound on hiding capacity.

Proof: See appendix.

In the small–distortion scenario $D_1 = D_2 << \sigma^2$, the capacity (4.2) would be nearly $\frac{1}{2}$ bit per sample. In contrast, the capacity is zero if $D_2 \geq \sigma^2 + D_1$, corresponding to the case of a host signal that is weak relative to the specified distortion levels. The zero–capacity problem is precisely the main limitation of some early watermarking schemes, in which watermarks were hidden in the least significant bit(s) of the host signal in a straightforward attempt to introduce minimal distortion. Unfortunately, such watermarks were easy to eliminate by an attacker that would simply randomize the least significant bit(s), and cause a minimal amount of distortion in the process. Cox *et al.* [11] were among the first researchers to realize that watermarks should be hidden in significant components of the host signal instead.

4.2 Blind Information Hiding

Rates of reliable transmission for blind information hiding clearly cannot be higher than the rates in the case the decoder has access to side information (such as the host data themselves, or a cryptographic key). Hence hiding capacity is upper-bounded by (4.2) for any $p(\tilde{x})$.

Theorem 4.2 below gives the optimal blind–information–hiding strategy and the optimal attack for Gaussian $p(\tilde{x})$. This pair of optimal strategies forms a saddlepoint solution. The optimal attack $Q(y|x)$ is again the Gaussian test channel (4.3). The optimal distribution $\tilde{Q}(x,u|\tilde{x})$ the same optimal distribution that achieves capacity in a problem studied by Costa [38]. Costa's result is an elegant extension of Gel'fand and Pinsker's results (see Fig. 3) to the case of additive white Gaussian noise channels with input power constraints. The hiding capacity for the blind–information–hiding problem is given by (4.2) again! In other words, *the achievable rate of reliable transmission is the same whether or not the host data are known at the decoder* [5].

Theorem 4.2 *Let $\mathcal{X} = I\!R$ and $d(x,y) = (x-y)^2$ be the squared–error distortion measure. Assume that $\tilde{X}$ is $\mathcal{N}(0,\sigma^2)$. Then the following distribution yields a saddlepoint of the payoff (3.4): $X = \tilde{X} + Z$, $U = Z + \alpha\tilde{X}$, where $\alpha = \frac{D_1}{D_1+\beta D_2}$, β is given below (4.3), $Z \sim \mathcal{N}(0,D_1)$ is independent of $\tilde{X}$, and $Q(y|x)$ is the Gaussian test channel (4.3). The hiding capacity is given by (4.2).*

Proof: See appendix.

Comments.

1. Consider the use of a covert channel of the form $U = Z + \alpha\tilde{X}$, where α is not given by the optimal value $\frac{D_1}{D_2+\beta D_2}$ in Theorem 4.2. The rate of reliable transmission is given by (E.1) in the appendix. A special case of this design is the constrained encoding strategy of Fig. 2, in which a codebook is designed for $U = Z$ (corresponding to $\alpha = 0$ above.) In the small–distortion case, $\sigma^2 >> D_1, D_2$, the host signal acts as a strong interference, and the capacity of the cascaded covert/attack channel is limited by this interferer. From (E.1), we obtain the rate of reliable transmission for that suboptimal class of encoders as

$$R(0) = \frac{1}{2}\log\left(1 + \frac{D_1}{\sigma^2 + \beta D_2}\right) \approx \frac{D_1}{2(\ln 2)\sigma^2}. \tag{4.4}$$

   Comparison of (4.2) and (4.4) shows that dramatic improvements are possible by using an optimal encoding strategy.

2. Consider the design of a codebook for $U = X$ (corresponding to $\alpha = 1$ above). One practical such system, based on dithered quantizers, is described in [16]. Equation (E.1) shows that the

[5] This conclusion was originally obtained by Chen for a closely related problem that assumes a fixed additive white Gaussian noise attack channel [9].

achievable rates for this system are given by $R(1) = \frac{1}{2}\log\frac{D_1(D_1+\sigma^2+\beta D_2)}{\beta D_2(D_1+\sigma^2)}$. After some algebra, it can be seen that these rates are zero for $D_2 \geq D_1$. Hence such systems are unreliable if $D_2 \geq D_1$. This discussion emphasizes the need to design a codebook for $U = Z + \alpha\tilde{X}$ rather than for Z or for X.

3. Consider now two plausible but suboptimal attacks. The additive white Gaussian noise attack $Q(y|x) \sim \mathcal{N}(x, D_2)$ is suboptimal but is asymptotically optimal for $\sigma^2 >> D_1, D_2$, because $\beta \to 1$ in this case. In contrast, the attack $y = \text{argmax}_{\tilde{x}}\, p(\tilde{x}|x)$, which attempts to degrade the message by recovering $\tilde{X}$ using the Maximum A Posteriori (MAP) estimation rule, is completely inefficient. Here $Y = \frac{\sigma^2}{\sigma^2+D_1}X$, so X is an invertible function of Y. Hence the data processing inequality $I(U;Y) \leq I(U;X)$ is satisfied with equality, and the attack fails to remove any information. However, an "invertible attack" can be very effective if a suboptimal decoder is used; for instance, a simple scaling of pixel intensities has been known to defeat some image watermarking decoders.

4. Fig. 7 summarizes the various designs for Gaussian channels.

5. Consider again the small–distortion case $\sigma^2 >> D_1, D_2$. To achieve the capacity bound using the random bin technique, we need a codebook of size not 2^{NR} but $2^{N[R+I(U;\tilde{X})]}$, where $I(U;\tilde{X}) = \frac{1}{2}\log(1+\frac{\alpha^2\sigma^2}{D_1})$ is typically very large relative to $R = \frac{1}{2}\log(1+\frac{D_1}{\beta D_2})$. This suggests it may be difficult to approach the capacity bound using practical information–hiding codes.

4.3 Optimal Decoder

Optimal decoding performance is obtained using the MAP decoding rule $\hat{u}^N = \text{argmax}_{u^N\in\mathcal{C}}\, p(u^N|y^N, k^N)$, where $\mathcal{C}$ denotes the codebook for U^N. For the optimal information–hiding and attack strategies of Theorem 4.2, we let $R_{UY} = \begin{pmatrix} E(U^2) & E(UY) \\ E(UY) & E(Y^2) \end{pmatrix}$ and obtain

$$\hat{u}^N = \underset{u^N\in\mathcal{C}}{\text{argmin}} \sum_{i=1}^{N} \begin{pmatrix} u_i \\ y_i \end{pmatrix}^T R_{UY}^{-1} \begin{pmatrix} u_i \\ y_i \end{pmatrix} = \underset{u^N\in\mathcal{C}}{\text{argmin}} \sum_{i=1}^{N} (u_i - \gamma y_i)^2 \tag{4.5}$$

where $\gamma = \frac{E[UY]}{E[Y^2]} = \frac{D_1+\alpha\sigma^2}{D_1+\sigma^2+D_2}$ is approximately equal to $\alpha \approx \frac{D_1}{D_1+D_2}$ if $\sigma^2 >> D_1, D_2$. Hence the decoder simply scales the received y^N by a factor of γ and finds the codeword closest to γy^N in the Euclidean–distance sense. A practical watermarking system based on this concept is described in Lin's thesis [39]. For the conventional but suboptimal design $U = Z$ in Sec. 4.2, $\sum_{i=1}^N u_i^2$ is approximately the same for all $u^N \in \mathcal{C}$, and the MAP decoding rule (4.5) is approximately equivalent to the maximum–correlation rule

$$\hat{u}^N = \underset{u^N\in\mathcal{C}}{\text{argmax}} \sum_{i=1}^{N} u_i y_i. \tag{4.6}$$

If the pair of random variables (U, Y) is non–Gaussian, or if $\sum_{i=1}^{N} u_i^2$ is not the same for all $u^N \in \mathcal{C}$, the maximum–correlation rule (4.6) is suboptimal. In current watermarking literature, a maximum–correlation method similar to (4.6) is often used to assess the performance of watermark detection algorithms, by testing an hypothesis $u^N = u^{*N}$ against the alternative $u^N \neq u^{*N}$, for some fixed watermark u^{*N} [21, 22]. These tests consist of comparing the correlation statistic $\sum_{i=1}^{N} u_i^* y_i$ against some threshold, and verifying that the empirical probability of error is low. Another commonly used decision statistic is the normalized correlation coefficient between u^{*N} and y^N [7, 11].

5 Small–Distortion Analysis

The case of small distortions D_1 and D_2 is typical of many information–hiding problems. One may wonder whether some simplifications occur in the theory, possibly like in rate–distortion theory [40].

We show that this is indeed the case. We consider the squared error distortion metric over the real line and show that the hiding capacity is *independent of the statistics of* $\tilde{X}$*, asymptotically as* $D_1, D_2 \to 0$. This complements the analogous remarkable Costa–type result for Gaussian channels in Sec. 4.2, which was however valid for all distortion levels. Our result is formally stated in Theorem 5.1 below. An intuitive explanation for this apparently surprising result is that hiding capacity is essentially determined by the geometry of small distortion balls, as there are many such small balls within regions where $p(\tilde{x})$ is essentially flat.

Theorem 5.1 *Let* $\mathcal{X} = I\!R$ *and* $d(x, y) = (x - y)^2$ *be the squared–error distortion measure. Assume that no key is used and that* $p(\tilde{x})$ *has zero mean and variance* σ^2 *and is bounded and continuous. Then the capacity* $C(D_1, D_2)$ *is asymptotic to the hiding capacity in the Gaussian case:* $\frac{1}{2} \log \left(1 + \frac{D_1}{D_2}\right)$*, as* $D_1, D_2 \to 0$*. The distribution that asymptotically achieves capacity is the same as in the Gaussian case:* $X = \tilde{X} + Z$*,* $U = Z + \alpha \tilde{X}$*, where* $\alpha = \frac{D_1}{D_1 + D_2}$*,* $Z \sim \mathcal{N}(0, D_1)$ *is independent of* $\tilde{X}$*, and* $Q(y|x)$ *is the Gaussian test channel (4.3) (with* $\beta = 1$*).*

Proof: See appendix.

6 Further Extensions

The framework developed in this paper can be used to analyze the performance of a variety of information–hiding systems. Some useful extensions are considered below.

6.1 Steganography

The purpose of steganography is to convey a message to a receiver in such a way that the very presence of the message is undetectable by a third party. Other restrictions may apply. In particular, we assume that the distortion constraints (3.1) and (3.3) are imposed on the information hider and the attacker, respectively.

Cachin [14] introduced a natural requirement for steganographic systems. The information hider can guarantee a certain level of undetectability provided that the relative entropy $D(p_{\tilde{X}^N}||p_{X^N})$ is no greater than some small, specified value ϵ. If ϵ is small enough, it becomes impossible for the attacker to determine whether x^N was drawn from the distribution $p(x^N)$ or from $p(\tilde{x}^N)$. The requirement $D(p_{\tilde{X}^N}||p_{X^N}) \leq \epsilon$ introduces an additional convex constraint on p_{X^N} and hence on the covert channel $\tilde{Q}^N(x^N, u^N|\tilde{x}^N, k^N)$.

For memoryless sources, the relative entropy increases linearly with N: $D(p_{\tilde{X}^N}||p_{X^N}) = ND(p_{\tilde{X}}||p_X)$. It follows that, unless $D(p_{\tilde{X}}||p_X) = 0$, information–hiding schemes are always detectable for large enough N. The undetectability constraint is very strong. For *perfect undetectability*, we would need $D(p_{\tilde{X}}||p_X) = 0$, hence $p_{\tilde{X}} = p_X$ (as in the Bernoulli example of Sec. 3.4). One can easily show the following result. Let $\mathcal{P}_0$ be the set of all $p_{\tilde{X}}$ such that $p_{\tilde{X}}(\tilde{x}) = p_{\tilde{X}}(\tilde{x}')$ for at least two distinct $\tilde{x}, \tilde{x}' \in \mathcal{X}$. For finite alphabets $\mathcal{X}$, the set $\mathcal{P}_0$ has measure zero. Then there exists D_1^* such that for all $p_{\tilde{X}} \notin \mathcal{P}_0$ and $D_1 < D_1^*$, the only $p(x|\tilde{x})$ that satisfies the constraints is $\delta(x - \tilde{x})$, whence the hiding capacity is zero. In other words, perfect undetectability and reliable transmission are essentially incompatible requirements.

6.2 Attacker Knows Message

Some information–hiding systems may not use cryptographic keys, and moreover, the attacker may know the information–hiding code. Hence the attacker is able to reliably decode the message. This does not mean that he is able to remove traces of the message from X. Clearly, the hiding capacity is upper–bounded by the capacity from Theorem 3.3, because the attacker uses more information. But can the hiding capacity still be strictly positive?

The solution of this problem is as follows. We define an attack channel as a conditional p.m.f. $Q(y|x, u)$, and let $\mathcal{Q}$ be the set of all such channels satisfying the distortion constraint

$$\sum_{y,x,u,\tilde{x},k} d(x, y)Q(y|x, u)\tilde{Q}(x, u|\tilde{x}, k)p(\tilde{x}, k) \leq D_2. \tag{6.1}$$

Then we have the following theorem, which takes the same form as Theorem 3.3, except that the set $\mathcal{Q}$ is larger than in Theorem 3.3.

Theorem 6.1 *Assume the attacker knows the information-hiding code and the particular codeword u^N used, and the decoder knows the attack channel. For any attack subject to distortion D_2, a rate R is achievable if and only if $R < \underline{C}$, where*

$$\underline{C} = \max_{\tilde{Q}(x,u|\tilde{x},k) \in \tilde{\mathcal{Q}}} \min_{Q(y|x,u) \in \mathcal{Q}} J(\tilde{Q}, Q). \tag{6.2}$$

Proof. The proof parallels that of Theorem 3.3. For achievability, the encoder and decoder use the same technique as in Proposition 3.1. But now the attacker is able to decode the codeword $u(j,m)$ used. He is then able to generate y^N from the distribution $Q^N(y^N|x^N, u(j,m))$. In the converse theorem (Prop. 3.2), the cardinality of the set $\mathcal{U}$ is set equal to $2|\mathcal{X}| + |\mathcal{K}|$. This does not compromise the optimality of the design of the covert channel, per Prop. 3.5(v), and makes it possible to optimize the covert channel while keeping the attack channel fixed. □

Corollary 6.2 *In the special case $K = \tilde{X}$, the hiding capacity $\underline{C}$ in (6.2) is the same as C in (3.6) and is achieved by the design $U = X$.*

Proof. If the decoder knows $\tilde{X}$, then by Corollary 3.4, $U = X$ is an optimal design. If $U = X$, the value of the added information U to the attacker is nil. □

Gaussian channels. If the decoder does not know $\tilde{X}$ and does not have access to any key, what is the performance of the encoding schemes in Sec. 4.2? Consider the design $U = Z + \alpha\tilde{X}$ in Fig. 6 again. Under this design, we have $X = U + (1-\alpha)\tilde{X}$. Conventional information hiding systems use $\alpha = 0$ [11]. For any $\alpha \neq 1$, the deterministic attack $Y = \frac{1}{1-\alpha}(X-U) = \tilde{X}$ is admissible if $D_2 \geq D_1$. But the rate of reliable transmission under this attack is zero! Can better performance be obtained by choosing $\alpha = 1$? In that case, $U = X$, and the value of the added information for the attacker (U) is nil. Then the rate formulas of Sec. 4.2 apply directly. From (E.1), we see that the rate of reliable transmission under the Gaussian test channel attack is $R = \frac{1}{2}\log\frac{D_1(D_1+\sigma^2+\beta D_2)}{\beta D_2(D_1+\sigma^2)}$. This rate is zero for $D_2 \geq D_1$ again. So for any value of α, the rate of reliable transmission is zero for $D_2 \geq D_1$. Such systems would be unreliable in many applications. Hence the information–hiding code should not be publicly known, or at least the system should be designed using cryptographic keys. In the special case $K = \tilde{X}$, Corollary 6.2 shows there is no loss of performance due to the attacker knowing the information–hiding code used.

6.3 Blockwise Memoryless Information Hiding and Attack Strategies

In this section, we extend the basic results of Sec. 3 to a simple class of attack channels with memory. Practical applications of this setup would include problems in image processing, where image data are partitioned into blocks, and the image is subject to some complex attack on each block, rather than to independent attacks on the constituent pixels. A JPEG compression attack would nearly fit this model [6]. Using a boldface notation $\mathbf{x} = \{x_1, \cdots, x_L\} \in \mathcal{X}^L$ for blocks, we let $Q^L(\mathbf{y}|\mathbf{x})$ be a conditional p.m.f. from $\mathcal{X}^L$ to $\mathcal{X}^L$ which satisfies the distortion constraint (2.1) for $N = L$. Consider blockwise memoryless attack channels, which are memoryless extensions of $Q^L(\mathbf{y}|\mathbf{x})$:

$$Q^N(y^N|x^N) = \prod_{i=1}^{N/L} Q^L(\mathbf{y}_i|\mathbf{x}_i), \quad \forall N = jL,\ j \geq 1,$$

[6] In the JPEG standard, the DC coefficients are encoded using a DPCM technique. This introduces a small dependency between blocks.

where $\mathbf{x}_i$ refers to the i–th block of data, $x_{Li}, \cdots, x_{Li+L-1}$, and $x^N = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N/L}\}$. We refer to L as the attack–channel blocklength. (The information–hiding codelength N is a multiple of L.) Similarly, we define

Definition 6.1 *A blockwise memoryless covert channel, subject to distortion D_1, is a sequence of conditional p.m.f.'s*

$$\tilde{Q}^N(x^N, u^N | \tilde{x}^N, k^N) = \prod_{i=1}^{N/L} \tilde{Q}^L(\mathbf{x}_i, \mathbf{u}_i | \tilde{\mathbf{x}}_i, \mathbf{k}_i), \quad \forall N = jL,\, j \geq 1.$$

from $\mathcal{X}^N \times \mathcal{K}^N$ to $\mathcal{X}^N \times \mathcal{U}^N$, such that

$$\sum_{\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{u}, \mathbf{k}} d^L(\tilde{\mathbf{x}}, \mathbf{x})\, \tilde{Q}^L(\mathbf{x}, \mathbf{u} | \tilde{\mathbf{x}}, \mathbf{k})\, p(\tilde{\mathbf{x}}, \mathbf{k}) \leq D_1. \tag{6.3}$$

Definition 6.2 *The compound blockwise memoryless covert channel subject to distortion D_1 is the set $\tilde{\mathcal{Q}}^L$ of all channels of the form (6.3).*

The pair $(\tilde{X}^N, K^N)$ is now assumed to be *blockwise i.i.d.*:

$$p(\tilde{x}^N, k^N) = \prod_{i=1}^{N/L} p(\tilde{\mathbf{x}}_i, \mathbf{k}_i) \quad \forall N = jL,\, j \geq 1. \tag{6.4}$$

This is a generalization of the model in Sec. 3, which assumed the individual symbols $(\tilde{X}_i, K_i)$ to be i.i.d. A block-diagram of the system is shown in Fig. 8. For the model (6.4), we obtain the following result, which is a straightforward consequence of Theorem 3.3, using the alphabets $\mathcal{K}^L$ and $\mathcal{X}^L$ in place of $\mathcal{K}$ and $\mathcal{X}$. The auxiliary alphabet can be taken to have a product form $\mathcal{U}^L$ without loss of generality.

Theorem 6.3 *Assume the attacker knows the covert channel, and the decoder knows the covert and attack channels. For any attack subject to distortion D_2, a rate R is achievable if and only if $R < C_L$, where*

$$C_L = \max_{\tilde{\mathcal{Q}}^L} \min_{\mathcal{Q}^L} J^L(\tilde{Q}^L, Q^L) \tag{6.5}$$

where

$$J^L(\tilde{Q}^L, Q^L) = \frac{1}{L}[I(U^L; Y^L | K^L) - I(U^L; \tilde{X}^L | K^L)], \tag{6.6}$$

and $(U^L, \tilde{X}^L, K^L) \to X^L \to Y^L$ is a Markov chain.

It follows directly from Proposition 3.1 that, under the assumptions of a time–invariant, memoryless attack and a memoryless source $(\tilde{X}, K)$, the optimal (capacity–achieving) information–hiding strategy is time–invariant and memoryless. Proposition 6.4 below establishes a dual result, namely, under a memoryless information–hiding strategy, the worst attack channel is memoryless and time–invariant.

Proposition 6.4 *Assume the source $(\tilde{X}, K)$ is memoryless. Then we have*
(i) *Assume $\tilde{Q}^L(x^L, u^L|\tilde{x}^L, k^L) = \prod_{i=1}^L \tilde{Q}_i(x_i, u_i|\tilde{x}_i, k_i)$. Let $Q_i(y_i|x_i)$ be the marginals of $Q(y|x)$, and $\overline{Q}^L(y|x) = \prod_{i=1}^L Q(y_i|x_i)$ be the product of these marginals. Then $\overline{Q}^L \in \mathcal{Q}^L$, and $J^L(\tilde{Q}^L, \overline{Q}^L) \leq J^L(\tilde{Q}^L, Q^L)$.*
(ii) *If the N covert channels $\tilde{Q}_i$ in Part (i) are independent of i, then the N optimal attack channels Q_i must also be independent of i.*

Proof. See appendix.

6.4 Fingerprinting

In fingerprinting applications, the information hider makes several copies of the host data $\tilde{X}$ available to different users. However, a different message is embedded in each copy. The message is a fingerprint, or serial number, which can be used to trace any unauthorized use of the signal back to the user. The user should not be able to remove traces of the fingerprint without seriously degrading the signal, and the fingerprint itself should be imperceptible. It is well known that developing a successful fingerprinting system is difficult, because of possible *collusion* between multiple users [3, 11, 12]. We show below that collusion allows users to compute a good estimate of the host signal, which contains little residual information about the individual messages.

We propose the following mathematical formulation of the fingerprinting problem. Referring to Fig. 9, assume there are L users, all potential colluders. Let $x_{l,i} \in \mathcal{X}$ be the data sent to user $l \in \{1, \cdots, L\}$ at time $i \in \{1, \cdots, N\}$. Also let $\mathbf{x}_i = \{x_{1,i}, \cdots, x_{L,i}\} \in \mathcal{X}^L$, $x^{l,N} = \{x_{l,1}, \cdots, x_{l,N}\} \in \mathcal{X}^N$, and $\mathbf{x}^N = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$. Each user's sequence is individually encoded according to $x^{l,N} = f_N(\tilde{x}^N, m_l, k^N)$, $1 \leq l \leq L$, where $m_l \in \mathcal{M}$ is the fingerprint for user l, and f_N is the encoder defined in Sec. 2. In other words, the same host signal and the same key are used for encoding all L messages. The messages m_l are assumed to be independent and uniformly distributed over $\mathcal{M}$. Next, the attack channel is defined as a conditional p.m.f. $Q(\mathbf{y}|\mathbf{x})$ from $\mathcal{X}^L$ to $\mathcal{X}^L$, and its memoryless extension as $Q^N(\mathbf{y}^N|\mathbf{x}^N) = \prod_{i=1}^N Q(\mathbf{y}_i|\mathbf{x}_i)$. The sets of admissible covert and attack channels are denoted by $\tilde{\mathcal{Q}}^L$ and $\mathcal{Q}^L$, respectively. Each message m_l is decoded as $\hat{m}_l = \phi_N(y^{l,N}, k^N)$, where ϕ_N is the decoder defined in Sec. 2. The average probability of error is

$$P_{e,N} = \frac{1}{|\mathcal{M}|} \sum_{l=1}^{L} \sum_{m \in \mathcal{M}} P(\phi_N(Y^{l,N}, K^N) \neq m \mid M_l = m).$$

Code rate and capacity can then defined as in Sec. 2. We obtain the following result.

Theorem 6.5 *For any attack subject to distortion D_2, a rate R is achievable if and only if $R < C_L$, where C_L is given in (6.5). Assume that the distortion function $d(x, y)$ is symmetric, $D_2 > D_1$, and $d_{\min} = \min_{\tilde{x},\tilde{x}'} d_s(\tilde{x}, \tilde{x}') > 0$ for some $s \in (0, 1)$, where $d_s(\tilde{x}, \tilde{x}') = -\log \sum_{x \in \mathcal{X}} p(x|\tilde{x}) \left(\frac{p(x|\tilde{x})}{p(x|\tilde{x}')}\right)^s$ is the Chernoff distance between distributions $p(\cdot|\tilde{x})$ and $p(\cdot|\tilde{x}')$. Then the capacity C_L tends to zero exponentially fast (with rate lower-bounded by Ld_{min}) as $L \to \infty$.*

Proof. Assume the information hider uses the capacity–achieving distribution $\tilde{Q}$. Let each user implement the memoryless attack $\hat{X}_i = \mathrm{argmax}_{\tilde{x}}\, p(\tilde{x}|\mathbf{x}_i)$, the MAP estimate of $\tilde{X}_i$ from $\mathbf{X}_i$. The attack $Y_{l,i} = \hat{X}_i$ for all $1 \leq l \leq L$ defines an attack channel $Q^L(\mathbf{y}|\mathbf{x})$ and yields an upper bound $J^L(\tilde{Q}^L, Q^L) \geq C$ on capacity. Next we use Fano's inequality and Chernoff bounds to show that $H(\hat{X}|\tilde{X})$ tends to zero exponentially fast with L, verify that the distortion constraint is satisfied, and invoke a simple extension of (3.8) to conclude the proof.

The error probability for estimating $\tilde{X}$ is given by $P_e(\tilde{X}) = Pr[\hat{X} \neq \tilde{X}|\tilde{X}]$. Using the union–of–events bound and Chernoff bounds on the probability of error of binary hypothesis tests [23], we write

$$P_e(\tilde{X}) \leq \sum_{\tilde{x},\tilde{x}'} p(\tilde{x})^{1-s} p(\tilde{x}')^s 2^{-Ld_s(\tilde{x},\tilde{x}')} \leq \frac{|\mathcal{X}|^2}{4} 2^{-Ld_{\min}} \tag{6.7}$$

for any $0 < s < 1$. Since $d_{\min} > 0$ by assumption, the MAP estimate converges a.s. to $\tilde{X}$ as $L \to \infty$. Hence $E_{Q^L} d^L(\mathbf{x},\mathbf{y}) \to E_{Q^L} d^L(\mathbf{x},\tilde{\mathbf{x}}) = D_1$, and the distortion constraint $E_{Q^L} d^L(\mathbf{x},\mathbf{y}) \leq D_2$ is satisfied for L large enough.

Now Fano's lemma gives $H(\hat{X}|\tilde{X}) \leq \delta(P_e(\tilde{X}))$ where $\delta(P_e) = -(1-P_e)\log(1-P_e) - P_e \log \frac{P_e}{|\mathcal{X}|-1}$ is monotonically increasing over the interval $[0, \frac{1}{2}]$ and is asymptotic to $P_e \log P_e^{-1}$ as $P_e \to 0$. Hence the Fano and Chernoff bounds can be combined to yield the upper bound

$$\log H(\hat{X}|\tilde{X}) \overset{<}{\sim} -Ld_{\min}, \quad \text{as } L \to \infty. \tag{6.8}$$

The technique used to establish (3.8) can be used again to show that $C \leq J^L(\tilde{Q}^L, Q^L) \leq H(\mathbf{Y}|\tilde{\mathbf{X}}) = H(\hat{X}|\tilde{X})$. The asymptotic inequality (6.8) implies that $\log C \overset{<}{\sim} -Ld_{\min}$ as $L \to \infty$. □

Observe that the optimal attack channel is not memoryless, and that the exponential decrease in capacity holds for any $p(\tilde{x})$. The papers [11, 12] contain examples of specific fingerprinting algorithms that are victims of the curse of large L. According to Theorem 6.5, these results hold for a broad class of fingerprinting algorithms.

It may seem surprising that the MAP attack, which was so ineffective in problems of information hiding using Gaussian data (Sec. 4.2) is so effective here. This is because the fingerprint attack, while still deterministic, is many–to–one. Also observe that the bound (6.7) becomes looser as $|\mathcal{X}|$ increases.

If the message to be embedded contains two parts: a message m that is common to all users (say copyright information) and user–dependent messages m_l (fingerprints), then the results above suggest a two–stage encoding technique, where in the first stage the common message m is embedded in the data set $\tilde{x}$ to produce an intermediary data set $\overline{x}$, and in the second stage the fingerprints m_l are embedded in the data set $\overline{x}$ to produce the fingerprinted data $\mathbf{x}^N$.

7 Summary and Conclusions

We have presented an information–theoretic analysis of information hiding, including the fundamental theorem of information hiding that characterizes the communication rate achievable for the information hider in the face of an optimal attack. The goal of the information hider is to maximize the rate of reliable transmission; the goal of the attacker is to minimize that rate. The payoff function in this game is a difference between two mutual informations. Different expressions for the hiding capacity are obtained depending on the knowledge available to the information hider, to the attacker, and to the decoder. We have primarily focused on a scenario involving the following assumptions. The information hider does not know the attack that will be implemented. The attacker does not know the information–hiding code and hence is unable to decode the message. He designs an attack channel $Q(y|x)$ based on the data X available to him. The decoder knows the information–hiding code and knows (or is able to learn) the attack channel. The allowed distortion levels (under a specified distortion function $d(x, y)$) for the information hider and the attacker are D_1 and D_2, respectively. These distortion levels respectively characterize the transparency of the information–hiding scheme and the severity of the attacks.

Our analysis shows the fundamental role of optimal data compression strategies in the attack and of channel coding in the hiding. Under the assumptions stated above, we have shown that the hiding capacity is the solution to a maxmin optimization problem, and that saddlepoint solutions are obtained in some cases. For instance, saddlepoint solutions are obtained when the host data are available as side information to the decoder, as well as in blind–information–hiding problems with Gaussian host data and a squared–error distortion function. In the latter case, the optimal attack is the Gaussian test channel, and the hiding capacity is the same as if the host data were available at the decoder. This result may seem surprising but is analogous to results by Costa [38] and Chen [9]. The hiding capacity for non–Gaussian host data is upper–bounded by the capacity for Gaussian host data with the same variance.

We have also conducted an analysis of the information–hiding problem in the case of small squared–error distortion. Remarkable simplifications arise in this case. The hiding capacity is asymptotic to $C \sim \frac{1}{2}\log(1 + D_1/D_2)$, in the limit as $D_1, D_2 \to 0$, independently of the statistics of the host data set, whether or not the decoder knows the host data.

While reasonable assumptions have been made about the knowledge available to the information hider, the attacker and the decoder, other assumptions may be preferable in some applications, and would require an extension of our basic theory. We have briefly described some of these extensions. For instance, the attacker might not know the covert channel that has been implemented, in which case the optimal strategy is the solution to a different game. The solution under this optimistic scenario is still a pure strategy for the attacker, but becomes a mixed strategy for the information hider. A more pessimistic assumption is that the attacker knows or is able to learn the information–hiding code used. In a system without cryptographic keys, the attacker would then be able to decode the message and design a more effective attack. We have derived an expression for the hiding capacity in this case too, and shown that encoding strategies that are capacity–achieving when the attacker is ignorant of the information–hiding code can be severely suboptimal (possibly zero) otherwise. Our results suggest that the real test for information–hiding systems is when $D_2 \geq D_1$, which corresponds to most practical problems of interest in the information–hiding literature. Other extensions of our work considered in this paper include the use of steganography constraints, that guarantee that the very presence of a hidden message is undetectable, and multipler–user scenarios, such as in fingerprinting applications. In this case, collusion between L users can have a devastating effect on decoding performance, as hiding capacity is reduced by a factor 2^L under some mild technical conditions.

Much work remains to be done in designing practical information–hiding codes that approach capacity. We anticipate this will be an active research area in the future. Our analysis has outlined the potential benefits of using randomized codes. Other practical problems include the choice of a suitable distortion measure, which is a holy grail in audio, image and video processing. Theoretical problems include computation of error exponents and reliability function [8], and design of universal decoders that perform well for a broad class of attack channels.

Acknowledgments We are deeply indebted to several colleagues who helped us at different stages during the preparation of this paper. In particular, we would like to thank J. Mark Ettinger and Tamer Basar for their help regarding game–theoretic aspects of our problem, and Richard E. Blahut, Brian Chen, Ralf Koetter, Kannan Ramchandran, and Andrew C. Singer for their helpful comments. We would also like to thank S.-M. William Koon and W. Gunawan for preparing the figures.

A Proof of Proposition 3.1

Here we prove achievability. The proof parallels the proof of Proposition 2 in [29]. First recall some definitions [23]. The type, or empirical probability distribution $P(a|x^N)$ of the sequence x^N is N^{-1} times the number of occurrences of the symbol $a \in \mathcal{X}$ in the sequence x^N. The strongly typical set $T_X^N(\eta)$ is the set of sequences in $\mathcal{X}^N$ whose type does not differ from its expected value by more than $\frac{\eta}{|\mathcal{X}|}$ in any component:

$$T_X^N(\eta) = \left\{ x^N \in \mathcal{X}^N \; : \; |P(a|x^N) - p_X(a)| < \frac{\eta}{|\mathcal{X}|}, \; \forall a \in \mathcal{X} \right\}.$$

The strong law of large numbers implies that for all $\eta > 0$, $Pr[x^N \in T_X^N(\eta)] \to 1$ as $N \to \infty$. The definition of strongly typical sets $T_{XY}^N(\eta)$ for pairs of random variables (X^N, Y^N), for triples, etc. is similar to the definition of $T_X^N(\eta)$ above.

Let $\mathcal{M} = \{1, 2 \cdots, 2^{N[I(U;Y|K) - I(U;\tilde{X}|K) - 2\epsilon]}\}$ and $R = \frac{1}{N} \log |\mathcal{M}|$. Using the same type of identity as in (3.8), we have $I(U;Y|K) - I(U;\tilde{X}|K) = I(U;Y,K) - I(U;\tilde{X},K)$. We generate codewords $u^N \in \mathcal{U}^N$ whose letters $u_n, 1 \leq n \leq N$ are independently drawn from the distribution $p(u)$. For every message $m \in \mathcal{M}$, we generate $J = 2^{N[I(U;\tilde{X},K)+\epsilon]}$ independent codewords from the distribution $p(u^N)$. We denote each codeword by $u(j,m), 1 \leq j \leq J$, $m \in \mathcal{M}$. The size of the codebook is $J|\mathcal{M}| = 2^{N[R+I(U;\tilde{X},K)+\epsilon]}$. The number η in the definition of the strongly typical sets below is a function of ϵ such that $\eta \to 0$ as $\epsilon \to 0$. The encoder and decoder perform the following operations.

Encoder. Given any $\tilde{x}^N \in \tilde{\mathcal{X}}^N$, $k^N \in \mathcal{K}^N$ and $m \in \mathcal{M}$, the encoder seeks a codeword in $\{u(j,m), 1 \leq j \leq J\}$ that is jointly typical with $(\tilde{x}^N, k^N)$. Specifically, the encoder seeks the smallest j such that $(u(j,m), \tilde{x}^N, k^N) \in T_{U\tilde{X}K}^N(\eta)$. (An encoding error is declared and j is set equal to zero if no such j can be found, but the strong law of large numbers ensures this event has vanishing probability p_{ee}.) Let $j(\tilde{x}^N, k^N, m)$ be the corresponding value of j. Given $\tilde{x}^N$, k^N and $u(j(\tilde{x}^N, k^N, m), m)$, the encoder then randomly generates x^N from the distribution $\tilde{Q}(x^N|\tilde{x}^N, u(j(\tilde{x}^N, k^N, m), m), k^N)$. The expected distortion is upper–bounded by $(1 - p_{ee})D_1 + p_{ee} d_{max}$. If p_{ee} tends to zero, the pair $(\tilde{x}^N, x^N)$ is distortion typical [23].

Decoder. Given the received sequence y^N and the key k^N, the decoder seeks a codeword $u(j,m), 1 \leq j \leq J, m \in \mathcal{M}$ in the codebook, such that $(u(j,m), y^N, k^N) \in T_{UYK}^N(\eta)$, i.e., $u(j,m)$ is jointly typical with (y^N, k^N). A decoding error is declared if no such codeword can be found or if several codewords indexed by different m's are found. Proceeding as in [29], it is seen that the probability of error is dominated by the event that there exist j' and $m' \neq m$ such that $(u(j', m'), y^N, k^N) \in T_{UYK}^N(\eta)$. The probability of this event is upper–bounded by $J|\mathcal{M}|2^{-N[I(U;Y,K)-\epsilon/2]} = 2^{-N\epsilon/2}$. The total probability of error is less than ϵ. If a codeboook chosen at random from the distribution $p(u^N)$ has probability of error lower than ϵ, then there exists at least one codebook with probability of error lower than ϵ. This proves achievability of all rates below $I(U;Y|K) - I(U;\tilde{X}|K)$.

B Proof of Proposition 3.2

Here we prove the converse coding theorem.

We show that, for a given attack, any sequence of codes with rate $\log|\mathcal{M}| = NR$ and error probability $P_{e,N} = P[\hat{M} \neq M|Y^N, K^N] \to 0$ must satisfy the following condition: there exists a covert channel $\tilde{Q}(x, u|\tilde{x}, k)$ such that $\frac{1}{N}\log|\mathcal{M}| = R \leq \max_{\tilde{Q}}[I(U;Y|K) - I(U;\tilde{X}|K)]$, and the distortion constraints (3.1) and (3.3) are satisfied.

We have

$$\begin{aligned} NR = H(M) = H(M|K^N) &= H(M|Y^N, K^N) + I(M;Y^N|K^N) \\ &\leq 1 + P_{e,N}NR + I(M;Y^N|K^N) \end{aligned} \tag{B.1}$$

where the first equality is because the messages M are drawn from a uniform distribution, the second is because M and K^N are independent, the third follows from the definition of mutual information, and the inequality is Fano's inequality. If $P_{e,N} \to 0$ as $N \to \infty$, then (B.1) implies $NR \leq I(M;Y^N|K^N)$ as $N \to \infty$.

Because M and $(\tilde{X}^N, K^N)$ are independent, we have

$$\begin{aligned} I(M;Y^N|K^N) &= I(M;Y^N|K^N) - I(M;\tilde{X}^N|K^N) \\ &= H(M|\tilde{X}^N, K^N) - H(M|Y^N, K^N). \end{aligned}$$

Because X^N is a function of $(\tilde{X}^N, K^N, M)$, the sequence $(\tilde{X}^N, K^N) \to (\tilde{X}^N, K^N, M) \to (X^N, \tilde{X}^N, K^N)$ is a Markov chain, and

$$\begin{aligned} H(M|\tilde{X}^N, K^N) &= H(\tilde{X}^N, K^N, M|\tilde{X}^N, K^N) \\ &\leq H(X^N, \tilde{X}^N, K^N|\tilde{X}^N, K^N). \end{aligned}$$

Furthermore, by construction, $(M, K^N) \to (\tilde{X}^N, K^N, M) \to (X^N, \tilde{X}^N, K^N) \to (Y^N, K^N)$ is a Markov chain, so

$$\begin{aligned} H(M|Y^N, K^N) &= H(M, K^N|Y^N, K^N) \\ &\geq H(X^N, \tilde{X}^N, K^N|Y^N, K^N). \end{aligned}$$

Combining these inequalities yields

$$\begin{aligned} I(M;Y^N|K^N) &\leq H(X^N, \tilde{X}^N, K^N|\tilde{X}^N, K^N) - H(X^N, \tilde{X}^N, K^N|Y^N, K^N) \\ &= I(X^N, \tilde{X}^N, K^N;Y^N|K^N) - I(X^N, \tilde{X}^N, K^N;\tilde{X}^N|K^N) \\ &= [H(Y^N|K^N) - H(Y^N|X^N, \tilde{X}^N, K^N)] - H(\tilde{X}^N|K^N) \\ &= H(Y^N|K^N) - \sum_{i=1}^{N} H(Y_i|X_i) - \sum_{i=1}^{N} H(\tilde{X}_i|K_i) \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{i=1}^{N} \Big[H(Y_i|Y_1, \cdots, Y_{i-1}, K^N) - H(Y_i|X_i) - H(\tilde{X}_i|K_i) \Big] \\
&\overset{(b)}{\leq} \sum_{i=1}^{N} \Big[H(Y_i|K_i) - H(Y_i|X_i) - H(\tilde{X}_i|K_i) \Big] \\
&= \sum_{i=1}^{N} \Big[H(Y_i|K_i) - H(Y_i|X_i, \tilde{X}_i, K_i) - H(\tilde{X}_i|K_i) \Big] \\
&= \sum_{i=1}^{N} \Big[I(X_i, \tilde{X}_i, K_i; Y_i|K_i) - I(X_i, \tilde{X}_i, K_i; \tilde{X}_i|K_i) \Big]
\end{aligned}
$$

where equality (a) follows from the chain rule for conditional entropy [23], and inequality (b) is because conditioning reduces entropy. Now defining the N random variables $U_i = (X_i, \tilde{X}_i, K_i)$, we obtain

$$
\begin{aligned}
I(M; Y^N|K^N) &\leq \sum_{i=1}^{N} \Big[I(U_i; Y_i|K_i) - I(U_i; \tilde{X}_i|K_i) \Big] \\
&\leq N \max_{i} \Big[I(U_i; Y_i|K_i) - I(U_i; \tilde{X}_i|K_i) \Big] \\
&\leq N \max_{\tilde{Q}} \Big[I(U; Y|K) - I(U; \tilde{X}|K) \Big] .
\end{aligned}
$$

Hence the converse theorem follows.

C Proof of Proposition 3.5

Here we prove convexity properties of the payoff (3.4).

(i) Both sets $\tilde{\mathcal{Q}}$ and $\mathcal{Q}$ are defined by the linear inequalities (3.1) and (3.3), respectively.

(ii)—(v) These properties are immediate extensions of results in [29].

(vi) Fix $p(u|\tilde{x}, k)$. The distortion constraint (3.1) written as a function of $p(x|\tilde{x}, u, k)$ becomes

$$
\sum_{x, \tilde{x}, u, k} d(\tilde{x}, x) p(x|\tilde{x}, u, k) p(u, \tilde{x}, k) \leq D_1.
$$

Ignoring momentarily the distortion constraint, we see that the set $\mathcal{P}_0'$ of admissible $p(x|\tilde{x}, u, k)$ is the Cartesian product of $|\mathcal{X}||\mathcal{U}||\mathcal{K}|$ probability simplices in $I\!R^{|\mathcal{X}|}$. Hence $\mathcal{P}_0'$ is a closed, convex, polyhedral set. Its vertices (extremal points) are characterized by the following property: the functions $p(\cdot|u, \tilde{x}, k)$ are zero/one functions for all $(u, \tilde{x}, k)$. In the presence of distortion constraints, the admissible set $\mathcal{P}_0$ is the intersection of $\mathcal{P}_0'$ above with the half hyperplane specified by the distortion constraint. So $\mathcal{P}_0$ is a closed, convex, polyhedral set whose vertices lie on an edge between two vertices of $\mathcal{P}_0$. For compactness of notation, denote by $j(p)$ the payoff (3.4) viewed as a function of $p(x|\tilde{x}, u, k)$. From Part (iv) of the proposition, $j(p)$ is convex and hence its maximum over $\mathcal{P}_0$ is achieved on one of the vertices of $\mathcal{P}_0$. This proves the claim.

D Proof of Theorem 4.1

By application of Theorem 3.3, the payoff is $I(X;Y|\tilde{X})$. For Gaussian $p(\tilde{x})$, we shall verify that $p(x|\tilde{x})$ and $Q(y|x)$ given in the statement of the theorem form a saddlepoint of the payoff, $I(X;Y|\tilde{X})$. This is done in two steps. First we fix $Q(y|x)$, and verify that $p(x|\tilde{x})$ maximizes $I(X;Y|\tilde{X})$ subject to the distortion constraint (3.3). The second step consists of fixing $p(x|\tilde{x})$, and verifying that $Q(y|x)$ minimizes $I(X;Y|\tilde{X})$ subject to the constraint (3.3). Note that X has variance $\sigma^2 + D_1$. If $D_2 \geq \sigma^2 + D_1$, then the attack channel $Q(y|x) = \delta(y)$ is admissible, and $J = 0$ no matter what the covert channel is. So $C = 0$ in this case. In the following, we evaluate the alternative, $D_2 < \sigma^2 + D_1$.

Step 1. Fix $Q(y|x)$ as indicated in (4.3). We do not require $p(\tilde{x})$ to be Gaussian. The Gaussian test channel (4.3) takes the form shown in Fig. 6, where $W \sim \mathcal{N}(0, \beta D_2)$ is independent of X, $V = X + W$, and $Y = \beta^{-1}V$. We write

$$J \overset{(a)}{=} I(X;Y|\tilde{X}) \overset{(b)}{=} I(X;V|\tilde{X}) \overset{(c)}{=} h(V|\tilde{X}) - h(V|X,\tilde{X}) \overset{(d)}{=} h(V|\tilde{X}) - h(V|X), \tag{D.1}$$

where (a) follows from Corollary 3.4, (b) is because Y is a deterministic function of V, (c) follows from the definition of conditional mutual information, and (d) holds because $\tilde{X} \to X \to V$ forms a Markov chain. The second term in the right–hand side of (D.1) is given by $h(V|X) = \frac{1}{2}\log(2\pi e \beta D_2)$, and only the first term, $h(V|\tilde{X})$, depends on $p(x|\tilde{x})$. Let $Z = X - \tilde{X}$. Observe that Z is independent of W, because $Z \to X \to W$ forms a Markov chain, and X is independent of W. Also,

$$E|V - \tilde{X}|^2 = E|Z + W|^2 \overset{(a)}{=} E|Z|^2 + E|W|^2 \leq D_1 + \beta D_2$$

where equality (a) holds because Z and W are independent, and the inequality holds because of the distortion constraint (3.1). Hence

$$h(V|\tilde{X}) \leq \frac{1}{2}\log(2\pi e(D_1 + \beta D_2))$$

and $J \leq \frac{1}{2}\log\left(1 + \frac{D_1}{\beta D_2}\right)$, with equality if $Z \sim \mathcal{N}(0, D_1)$ is independent of $\tilde{X}$. This proves the claim (ii).

Step 2. Fix $p(x|\tilde{x})$ as indicated in the statement of the theorem. Assume that $\tilde{X} \sim \mathcal{N}(0, \sigma^2)$, so $X \sim \mathcal{N}(0, \sigma^2 + D_1)$. For any given $Q(y|x)$, let $a = \frac{E[XY]}{E[X^2]}$, $V = a^{-1}Y$, and $W = V - X$. It follows from this definition that $E[XW] = 0$. The distortion constraint (3.3) takes the form

$$D_2 \geq E|Y - X|^2 = E|(a-1)X + aW|^2 = (a-1)^2(\sigma^2 + D_1) + a^2 E|W|^2. \tag{D.2}$$

Now

$$\begin{aligned} J &= I(X;Y|\tilde{X}) = I(X;V|\tilde{X}) = I(X - \tilde{X}; V - \tilde{X}|\tilde{X}) = I(Z; Z + W|\tilde{X}) \\ &\overset{(a)}{=} h(Z) - h(Z|Z + W, \tilde{X}) \overset{(b)}{\geq} h(Z) - h(Z|Z + W) = I(Z; Z + W) \overset{(c)}{\geq} I(Z; Z + W^*) \end{aligned}$$

where W^* is a Gaussian random variable that has zero mean and the same variance as W. `discuss mean` Here equality (a) is because Z and $\tilde{X}$ are independent, inequality (b) follows from the fact that conditioning reduces entropy, and inequality (c) can be found in [41]. These inequalities are satisfied with equality if W is Gaussian and independent of $\tilde{X}$. Since X and W^* are independent and $Z \to X \to W^*$ forms a Markov chain, Z and W^* are also independent. Hence the lower bound above becomes

$$J \geq I(Z; Z + W^*) = \frac{1}{2} \log \left(1 + \frac{D_1}{E|W|^2} \right),$$

which is a decreasing function of $E|W|^2$. The lowest possible bound is obtained by maximizing $E|W|^2$ subject to the constraint (D.2), where $a \in I\!R$. After some algebra, we obtain $E|W|^2 = \beta D_2$ and $a = \beta^{-1}$, namely, $Q(y|x)$ is the Gaussian test channel. In this case, all the inequalities above are satisfied with equality. Hence the Gaussian test channel is the optimal attack.

E Proof of Theorem 4.2

In order to prove the saddlepoint property, the first step is to fix $Q(y|x)$ and show that the optimal $\tilde{Q}(x, u|\tilde{x})$ is the one specified in the statement of the theorem. We only briefly sketch this step, which is a simple variation on a result in [38]. The Gaussian test channel $Q(y|x)$ may be viewed as the cascade of an additive white Gaussian noise channel with variance βD_2 and a multiplier D_2/β, as illustrated in Fig. 6. The payoff is given by $I(U; Y) - I(U; \tilde{X}) = I(U; V) - I(U; \tilde{X})$. Consider the particular design $U = Z + \alpha\tilde{X}$. The rate of reliable transmission in this case is given by

$$R(\alpha) = \begin{cases} \frac{1}{2} \log \frac{D_1(D_1+\sigma^2+\beta D_2)}{D_1\sigma^2(1-\alpha)^2+\beta D_2(D_1+\alpha^2\sigma^2)} & : \text{ if } D_2 < \sigma^2 + D_1 \\ 0 & : \text{ else.} \end{cases} \tag{E.1}$$

The value of α that maximizes $R(\alpha)$ is $\alpha = \frac{D_1}{D_1+\beta D_2}$. The corresponding value of $R(\alpha)$ is equal to (4.2). Since capacity could not possibly be larger than in the case where $\tilde{X}$ is known at the decoder, the design $U = Z + \alpha\tilde{X}$ with $\alpha = \frac{D_1}{D_1+\beta D_2}$ is optimal.

The second step in the proof is to fix $\tilde{Q}(x, u|\tilde{x})$ and show that the optimal $Q(y|x)$ is the Gaussian test channel. Notice that the payoff (3.4) depends on Q only through $I(U; Y)$. We shall in fact show that for any Gaussian–distributed pair (U, X), the distribution $Q(y|x)$ that minimizes $I(U; Y)$ over $\mathcal{Q}$ is the Gaussian test channel. Let Y^* be a Gaussian random variable such that (U, X, Y) and (U, X, Y^*) have the same second–order statistics. Then we use the classical inequality [41]

$$I(U; Y) \geq I(U; Y^*) = \frac{1}{2} \log \frac{r_{uu}r_{yy}}{r_{uu}r_{yy} - r_{uy}^2} \tag{E.2}$$

where we use the notation $r_{XY} = E[XY] - E[X]E[Y]$. Equality is achieved in (E.2) if (U, Y) is jointly Gaussian distributed with zero mean. We minimize the lower bound $I(U; Y^*)$ over all possible distributions that satisfy the distortion constraint (3.3) and show that the minimum is achieved for the Gaussian test channel. The argument is based on the geometric representation of second–order random variables. Let P be the joint probability distribution of the random variables $A = (U, X, Y)$. Define the Hilbert space $L_2(A)$ induced by the inner product $< a, b >= E[ab]$, where $a, b \in L_2(A)$, and the expectation is taken with respect to the distribution P. Then minimizing the right–hand side $-\frac{1}{2}\log(1 - \frac{r_{uy}^2}{r_{uu}r_{yy}})$ of (E.2) is equivalent to maximizing

$$\frac{r_{uu}r_{yy}}{r_{uy}^2} = \frac{\|U\|^2\|Y^*\|^2}{< U, Y^* >^2} = \cos^2\theta,$$

where $\|\cdot\|$ is the norm induced by the inner product above, and $\cos\theta$ is the angle between the vectors U and Y in $L_2(A)$. The constraint (3.3) takes the form

$$\|Y - X\|^2 \leq D_2, \tag{E.3}$$

which is the equation of a circle of radius $\sqrt{D_2}$ centered at X, see Fig. 10. The vector Y that minimizes $\cos^2\theta$ subject to the constraint (E.3) is tangent to the circle above. Hence the solution satisfies the orthogonality property $< Y, Y - X >= 0$, no matter what U is. This solution is the Gaussian test channel. □

F Proof of Theorem 5.1

For clarity, here we explicitly indicate the subscripts on the pdfs $p_{\tilde{X}}(\tilde{x})$, $p_X(x)$ and $p_Y(y)$. Let $\tilde{Q}^*(x, u|\tilde{x})$ and $Q^*(y|x)$ be the two distributions specified in the statement of the theorem. Hence

$$\tilde{Q}^*(x, u|\tilde{x}) = \frac{1}{\sqrt{D_1}}\phi\left(\frac{x - \tilde{x}}{\sqrt{D_1}}\right)\delta(x - u - (1-\alpha)\tilde{x}), \quad \text{and} \quad Q^*(y|x) = \frac{1}{\sqrt{D_2}}\phi\left(\frac{y - x}{\sqrt{D_2}}\right)$$

where $\phi(\cdot)$ denote the standard normal distribution $\mathcal{N}(0, 1)$, and $\delta(\cdot)$ is the Dirac distribution. Now, using elementary properties of the mutual information, we write the payoff (3.4) as

$$J(\tilde{Q}^*, Q) = \int\int p_{\tilde{X}}(\tilde{x})\psi(Q, \tilde{x})\, d\tilde{x}, \tag{F.1}$$

where

$$\psi(Q, \tilde{x}) = \int\int\int Q(y|x)\tilde{Q}^*(x, u|\tilde{x}) \log\frac{p(u|y)}{p(u|\tilde{x})}\, dydxdu. \tag{F.2}$$

Since $p_{\tilde{X}}(\tilde{x})$ is smooth, the minimizing $Q(y|x)$ is also smooth in both x and y; hence we conduct our analysis assuming that $Q(y|x)$ is smooth. We first prove the following claim.

Claim: For fixed Q and $\tilde{x}$, $\psi(Q, \tilde{x})$ does not depend on the distribution $p_{\tilde{X}}$ as $D_1, D_2 \to 0$.

Proof: We have $p(u|\tilde{x}) = \int \tilde{Q}^*(x, u|\tilde{x})\, dx$, so any dependency of $\psi(Q, \tilde{x})$ on $p_{\tilde{X}}$ is via the term

$$p(u|y) = \frac{p(y, u)}{p_Y(y)} = \frac{1}{p_Y(y)}\int\int Q(y|x)\tilde{Q}^*(x, u|\tilde{x})p_{\tilde{X}}(\tilde{x})\, dx\, d\tilde{x}. \tag{F.3}$$

Consider

$$p_Y(y) = \int Q(y|x)p_X(x)\, dx \tag{F.4}$$

where

$$\begin{aligned} p_X(x) &= \int \left[\int \tilde{Q}^*(x, u|\tilde{x})\, du \right] p_{\tilde{X}}(\tilde{x})\, d\tilde{x} \\ &= \int \frac{1}{\sqrt{D_1}}\, \phi\left(\frac{x - \tilde{x}}{\sqrt{D_1}} \right) p_{\tilde{X}}(\tilde{x})\, d\tilde{x}. \end{aligned} \tag{F.5}$$

Since $p_{\tilde{X}}$ is bounded and continuous, (F.5) implies that p_X converges uniformly to $p_{\tilde{X}}$ as $D_1 \to 0$. A similar but slightly weaker argument can be applied to (F.4): since $Q(y|x)$ satisfies the distortion constraint (3.3), the Chebyshev–like inequality

$$\int p(x) \int_{x-\delta}^{x+\delta} Q(y|x)\, dy dx \geq 1 - \frac{D_2}{\delta^2}, \quad \forall D_2, \delta > 0,$$

ensures convergence of p_Y to p_X (and hence to $p_{\tilde{X}}$). This convergence is nonuniform unless $p_X(x)$ is bounded away from zero. We now write (F.3) as

$$\begin{aligned} p(u|y) &= \frac{1}{p_Y(y)} \int \int Q(y|x) \frac{1}{\sqrt{D_1}}\, \phi\left(\frac{x - \tilde{x}}{\sqrt{D_1}} \right) \delta(x - u - (1-\alpha)\tilde{x})\, p_{\tilde{X}}(\tilde{x})\, dx\, d\tilde{x} \\ &= \int Q(y|u + (1-\alpha)\tilde{x}) \frac{1}{\sqrt{D_1}}\, \phi\left(\frac{u - \alpha\tilde{x}}{\sqrt{D_1}} \right) \frac{p_{\tilde{X}}(\tilde{x})}{p_Y(y)}\, d\tilde{x}. \end{aligned}$$

Define the family of sets $\mathcal{X}(\eta) = \{\tilde{x} \;:\; p_{\tilde{X}}(\tilde{x}) > \eta\}$. The union of these sets is dense in the support set of $p_{\tilde{X}}$. For any $\delta, \eta > 0$ and $y \in \mathcal{X}(\eta)$, there exist D_1, D_2 small enough so that (1) the integral above can be approximated with arbitrary accuracy by an integral from $y - \delta$ to $y + \delta$, and (2) the ratio $\frac{p_{\tilde{X}}(\tilde{x})}{p_Y(y)}$ is arbitrarily close to 1, for any $\tilde{x} \in [y - \delta, y + \delta]$. Hence $p(u|y)$ does not depend on $p_{\tilde{X}}$ as $D_1, D_2 \to 0$. We conclude that $\psi(Q, \tilde{x})$ in (F.2) does not depend on $p_{\tilde{X}}$ as $D_1, D_2 \to 0$. This proves the claim.

The developments above imply that the function $\psi(Q, \tilde{x})$ essentially involves local values of $Q(y|.)$ in a small neighborhood of $\tilde{x}$. Since p_X converges uniformly to $p_{\tilde{X}}$, we have

$$J(\tilde{Q}^*, Q) = \int p_{\tilde{X}}(\tilde{x}) \psi(Q, \tilde{x})\, d\tilde{x} \sim \int p_X(x) \psi(Q, x)\, dx \quad \text{as } D_1, D_2 \to 0.$$

We now minimize the asymptotic form of $J(\tilde{Q}^*, Q)$ above with respect to Q, subject to the distortion constraint (3.3). The inequality (3.3) must be satisfied with equality. We define the Lagrange functional

$$\begin{aligned} \mathcal{L}(Q, \lambda) &= J(\tilde{Q}^*, Q) + \lambda \int \int p_X(x) Q(y|x) |y - x|^2\, dx dy \\ &= \int p_X(x)\, \phi(Q, \lambda, x)\, dx \end{aligned} \tag{F.6}$$

where

$$\phi(Q, \lambda, x) = \psi(Q, x) + \lambda \int Q(y|x) |y - x|^2\, dy \tag{F.7}$$

satisfies the same asymptotic locality properties as $\psi(Q, x)$, and the Lagrange multiplier $\lambda \geq 0$ is selected so as to satisfy the constraint (3.3). Let Q_λ^{**} be the minimum of $\mathcal{L}(Q, \lambda)$ with respect to

Q. Assume momentarily that $p_{\tilde{X}}$ is Gaussian. Then according to Theorem 4.2, Q_λ^{**} is the Gaussian test channel Q_λ^* corresponding to a particular distortion level $D_2(\lambda)$. Due to the asymptotic locality property of $\phi(Q, \lambda, x)$, Q_λ^* also minimizes $\phi(Q, \lambda, x)$, as $D_1, D_2 \to 0$. Also, the term that multiplies λ in (F.7) is equal to $D_2(\lambda)$ for all x when $Q = Q_\lambda^*$, so the distortion integral $\int\int p_X(x) Q_\lambda^*(y|x)|y - x|^2 \, dxdy = D_2(\lambda)$ for all p_X.

The second term in the right–hand side of (F.7) is thus independent of p_X, and the first term is asymptotically independent of p_X. Hence the minimum of $\phi(Q, \lambda, x)$ is achieved by some $Q_{\lambda x}$ that is independent of p_X. But Q_λ^* minimizes $\phi(Q, \lambda, x)$ under Gaussian $p_{\tilde{X}}$, and therefore does so under any other $p_{\tilde{X}}$. Hence Q_λ^* minimizes $\mathcal{L}(Q, \lambda)$ under any $p_{\tilde{X}}$, for any $\lambda \geq 0$. Choosing λ so as to satisfy the distortion constraint (3.3) with equality, we conclude that Q^* minimizes $J(\tilde{Q}^*, Q)$ subject to the constraint (3.3). Moreover, by Theorem 4.1(i), $\min_Q J(\tilde{Q}^*, Q) \sim \frac{1}{2} \log\left(1 + \frac{D_1}{D_2}\right)$, even if $p_{\tilde{X}}$ is non–Gaussian. Theorem 4.1(ii) then implies that no $\tilde{Q}$ could be better than $\tilde{Q}^*$, hence $\tilde{Q}^*, Q^*$ are asymptotically capacity–achieving distributions, as $D_1, D_2 \to 0$.

G Proof of Proposition 6.4

Here we show that the optimal attack for a time–invariant, memoryless covert channel is time–invariant and memoryless.

(i) (Memoryless). By definition of the mutual information, we have

$$I(U^N; Y^N|K^N) - I(U^N; \tilde{X}^N|K^N) = H(U^N|\tilde{X}^N, K^N) - H(U^N|Y^N, K^N)$$

The attacker has no control over the first conditional entropy on the right side but tries to maximize the second one. Now

$$H(U^N|Y^N, K^N) = \sum_{i=1}^N H(U_i|U^{i-1}, Y^N, K^N) \leq \sum_{i=1}^N H(U_i|Y_i, K_i) \qquad \text{(G.1)}$$

where the equality follows from the chain rule for entropy [23, p. 21], and the first inequality follows from the fact that conditioning reduces entropy. Since K_i are mutually independent, equality is achieved in this inequality if and only if U_i is independent of $Y_j, \forall j \neq i$. Equality is achieved for a memoryless attack channel. Hence $J^N(\tilde{Q}^N, Q^N) \geq J^N(\tilde{Q}^N, \overline{Q}^N)$.

The final step in the proof is to observe that if the attack channel $Q^N(y^N|x^N)$ lies in the admissible set $\mathcal{Q}^N$, then so does the product of its marginals, $\overline{Q}^N(y^N|x^N)$, as the expected distortions under Q^N and $\overline{Q}^N$ are equal.

(ii) (Time–invariant). From part (i), the cost function to be minimized by the attacker takes the form $J^N(\tilde{Q}^N, Q^N) = \sum_{i=1}^N J(\tilde{Q}, Q_i)$. The attack is subject to linear constraints $\sum_{i=1}^N E_{Q_i} d(x, y) \leq D_2$ and $\sum_y Q_i(y|x) = 1$, for all $x \in \mathcal{X}$. The cost function is strictly convex in $\{Q_i\}$, and both the cost function and the constraints are invariant to permutations of the indices i. Hence Q_i must be independent of i to achieve the minimum.

References

[1] *IEEE Journal on Selected Areas in Communications*, Special Issue on Copyright and Privacy Protection, Vol. 16, No. 4, May 1998.

[2] *Proceedings IEEE*, Special Issue on Identification and Protection of Multimedia Information, Vol. 87, No. 7, July 1999.

[3] 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.

[4] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual Watermarks for Digital Images and Video," pp. 1108—1126 in [2].

[5] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," pp. 1079—1107 in [2].

[6] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information Hiding – A Survey," pp. 1062—1078 in [2].

[7] I. J. Cox, M. L. Miller and A. L. McKellips, "Watermarking as Communications with Side Information," pp. 1127—1141 in [2].

[8] N. Merhav, "On Random Coding Error Exponents of Watermarking Codes," *preprint*, Sep. 1998.

[9] B. Chen, *seminar at the University of Illinois*, April 1999.

[10] J. R. Hernández and F. Pérez-González, "Statistical Analysis of Watermarking Schemes for Copyright Protection of Images," pp. 1142—1166 in [2].

[11] I. J. Cox, J. Killian, F. T. Leighton and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," *IEEE Trans. Image Proc.*, Vol. 6, No. 12, pp. 1673—1687, Dec. 1997.

[12] D. Boneh and J. Shaw, "Collusion–Secure Fingerprinting for Digital Data," in *Advances in Cryptology: Proc. CRYPTO'95*, Springer–Verlag, New York, 1995.

[13] B. Macq and J.-J. Quisquater, "Cryptology for Digital TV Broadcasting," *Proc. IEEE*, Vol. 83, No. 6, pp. 944—957, June 1995.

[14] C. Cachin, "An Information–Theoretic Model for Steganography," *Proc. 1998 Workshop on Information Hiding*, Portland, Oregon, Lecture Notes in Computer Sciences, Springer–Verlag, 1998.

[15] L. Marvel, C. G. Boncelet, Jr., and C. T. Retter, "Spread–Spectrum Image Steganography," *IEEE Trans. Image Proc.*, Vol. 8, No. 8, pp. 1075—1083, Aug. 1999.

[16] B. Chen and G. W. Wornell, "An Information–Theoretic Approach to the Design of Robust Digital Watermarking Systems," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, AZ, March 1999.

[17] J. M. Ettinger, "Steganalysis and Game Equilibria," *Proc. 1998 Workshop on Information Hiding*, Portland, Oregon, Lecture Notes in Computer Sciences, Springer–Verlag, 1998.

[18] 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.

[19] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell Syst. Tech. J.*, Vol. 28, No. 4, pp. 656-715, 1949.

[20] F. A. P. Petitcolas and M. G. Kuhn: StirMark software, available from `www.cl.cam.ac.uk/~fapp2/watermarking/image_watermarking/stirmark/`.

[21] A. Piva, M. Barni, F. Bartolini and V. Cappellini, “DCT-Based Watermark Recovering without Resorting to the Uncorrupted Original Image,” *Proc. IEEE Int. Conf. on Image Proc.*, Santa Barbara, CA, Vol. I, pp. 520—523, 1997.

[22] F. Hartung and B. Girod, ”Digital Watermarking of Uncompressed and Compressed Video,” *Signal Processing*, Vol. 66, pp. 283—301, 1998.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[24] I. Csiszár and J. Körner, *Information Theory: Coding Theory for Discrete Memoryless Systems*, Academic Press, NY, 1981.

[25] M. V. Hegde, W. E. Stark, and D. Teneketzis, “On the Capacity of Channels with Unknown Interference,” *IEEE Trans. Info. Thy*, Vol. 35, No. 4, pp. 770—783, July 1989.

[26] P. W. Wong, “A Public Key Watermark for Image Verification and Authentication,” *Proc. Int. Conf. Im. Proc.*, Vol. I, pp. 455—459, 1998.

[27] W. Bender, D. Gruhl and N. Morimoto, “Techniques for Data Hiding,” *Proc. SPIE*, Vol. 2420, p. 40, Feb. 1995.

[28] C. Busch, W. Funk and S. Wolthusen,“Digital Watermarking: From Concepts to Real—Time Video Applications,” *IEEE Computer Graphics and Applications*, pp. 25–35, Jan/Feb. 1999.

[29] 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.

[30] C. Heegard and A. A. El Gamal, “On the Capacity of Computer Memory with Defects,” *IEEE Trans. Info. Thy*, Vol. 29, No. 5, pp. 731—739, Sep. 1983.

[31] K. Marton, “ A Coding Theorem for the Discrete Memoryless Broadcast Channel,” *IEEE Trans. Info. Thy*, Vol. 25, No. 25, pp. 306—311, May 1979.

[32] D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, 1995.

[33] A. Lapidoth and P. Narayan, “Reliable Communication Under Channel Uncertainty,” *IEEE Trans. Info. Thy*, Vol. 44, No. 6, pp. 2148—2177, Oct. 1998.

[34] T. Basar and G.J. Olsder, *Dynamic Noncooperative Game Theory*, SIAM Classics in Applied Mathematics, 1999.

[35] T. Basar, “The Gaussian Channel with an Intelligent Jammer,” *IEEE Trans. Info. Thy*, Vol. 29, No. 1, pp. 152—157, Jan. 1983.

[36] J. M. Borden, D. M. Mason, and R. J. McEliece, “Some Information Theoretic Saddlepoints,” *SIAM J. Control and Optimization*, Vol. 23, No. 1, pp. 129—143, Jan. 1985.

[37] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

[38] M. Costa, “Writing on Dirty Paper,” *IEEE Trans. Info. Thy*, Vol. 29, No. 3, pp. 439—441, May 1983.

[39] G.-I. Lin, M. S. thesis, U. of Illinois at Urbana–Champaign, Dept. of Electrical and Computer Engineering, Sep. 1999.

[40] A. M. Gerrish and P. M. Schulteiss, “Information Rates of Non–Gaussian Processes,” *IEEE Trans. Info. Thy*, Vol. X, No. Y, pp. 265—271, Oct. 1964.

[41] M. Pinsker, “Gaussian Sources,” *Probl. Inform. Transmission*, Vol. 14, pp. 59—100, 1963.

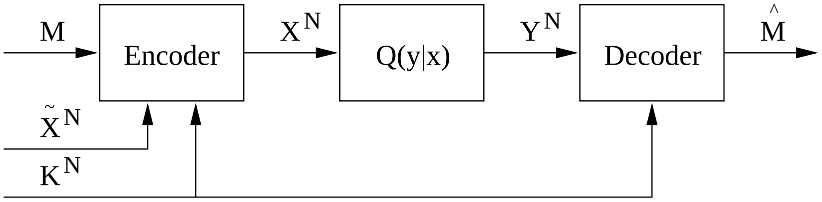

Figure 1: Formulation of information hiding as a communication problem. Here M is the message to be embedded in the host data set $\tilde{X}^N$ and transmitted to the decoder. The composite data set X^N is subject to attacks embodied by the channel $Q(y|x)$. The encoder and the decoder share a common cryptographic key K^N.

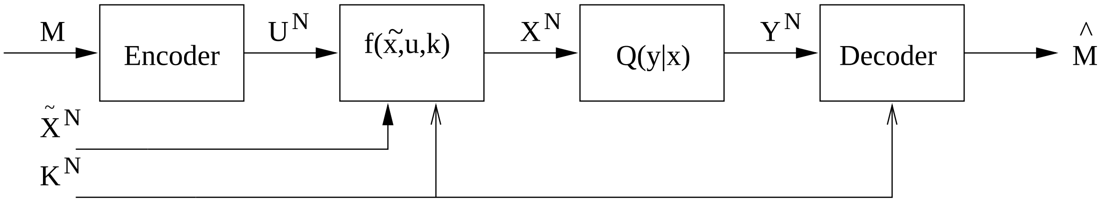

Figure 2: Constrained design of information–hiding system, used in current watermarking literature.

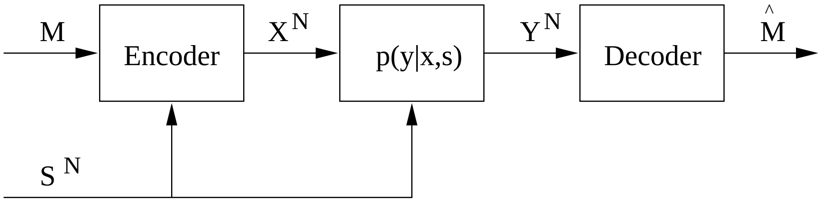

Figure 3: Channel with random parameters s that are known at the encoder but not at the decoder.

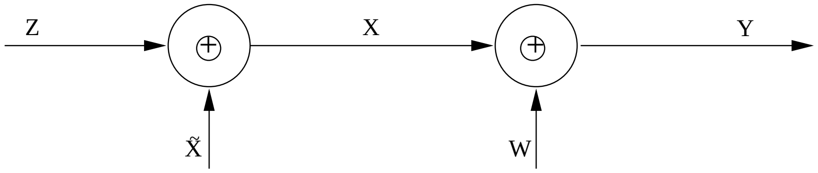

Figure 4: Binary channel with optimal information–hiding and attack strategies. Here $\tilde{X} \sim$ Bernoulli($\frac{1}{2}$), $U \sim$ Bernoulli(D_1), and $W \sim$ Bernoulli(D_2) are mutually independent random variables.

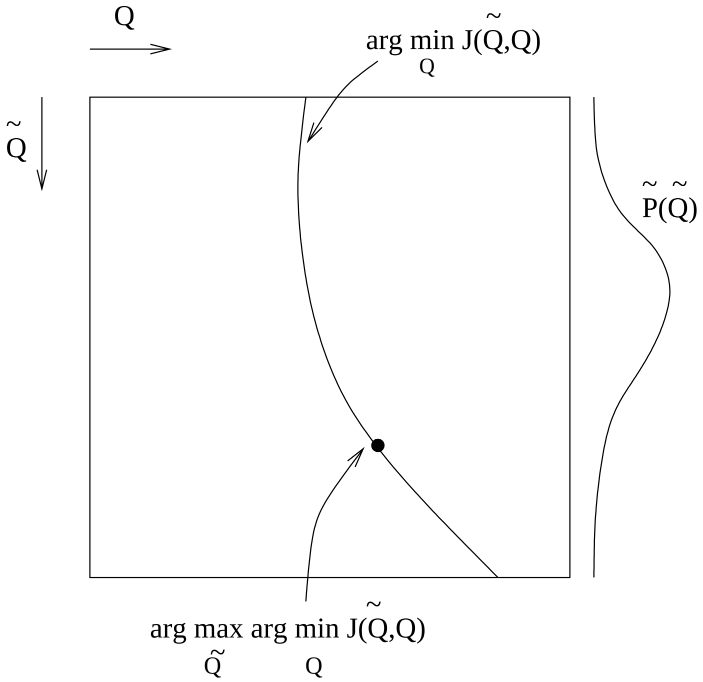

Figure 5: The information–hiding game. Any choice of $\tilde{Q}$ and Q by the information hider and the attacker constitutes a strategy. A Nash equilibrium of the game is attained when the information hider uses an optimal mixed strategy (chooses $\tilde{Q}$ according to some optimal distribution $\tilde{P}$), and the attacker uses a pure strategy.

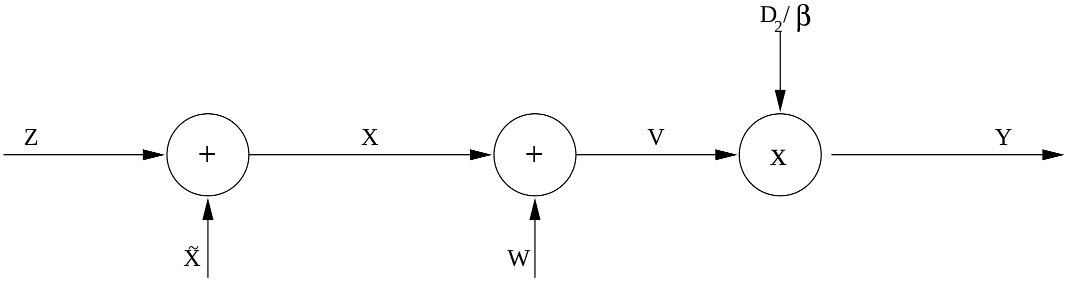

Figure 6: Optimal information–hiding and attack strategies when host data $\tilde{X} \sim \mathcal{N}(0, \sigma^2)$ are available at the decoder. Here $Z \sim \mathcal{N}(0, D_1)$ and $W \sim \mathcal{N}(0, \beta D_2)$ are mutually independent random variables, and $\beta = \left(1 - \frac{D_2}{\sigma^2 + D_1}\right)^{-1}$. The optimal attack is the Gaussian test channel with distortion level D_2.

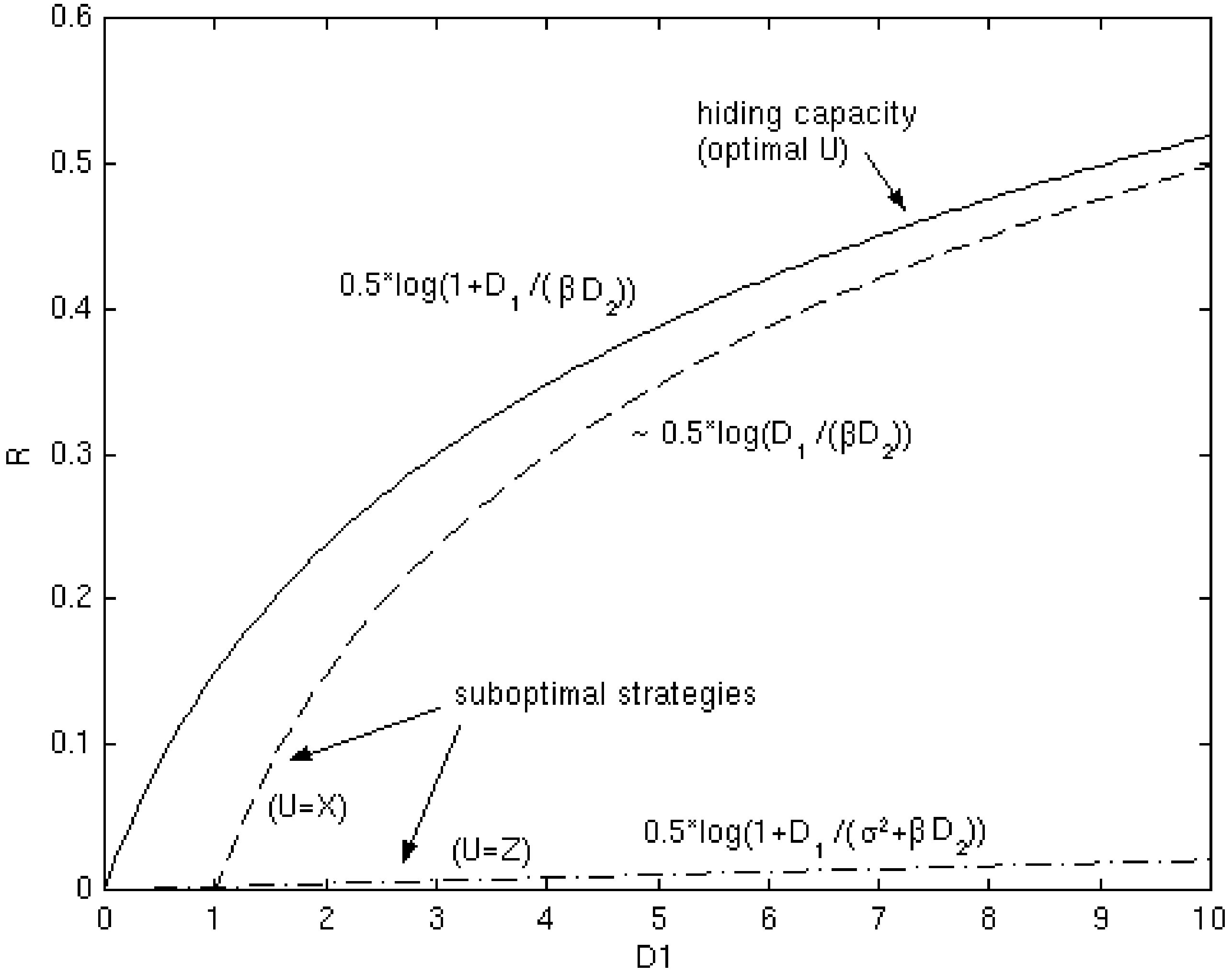

Figure 7: Rates of reliable transmission for Gaussian channels, using different information–hiding strategies, assuming $D_2 = 1$ and $\sigma = 10$. Current designs in the literature often operate far below capacity, as indicated on the graph and discussed in the text.

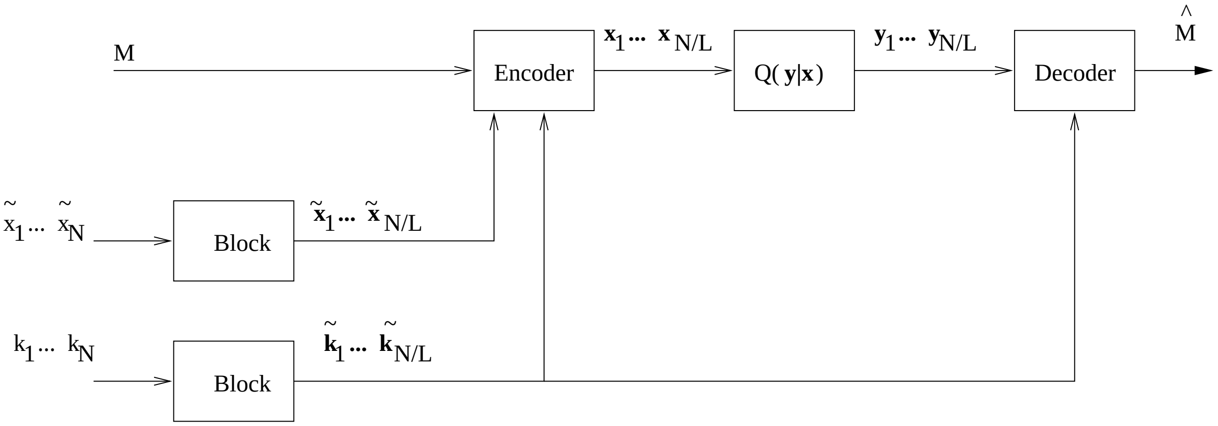

Figure 8: Block-diagram for blockwise-i.i.d. sources and blockwise memoryless attack channels.

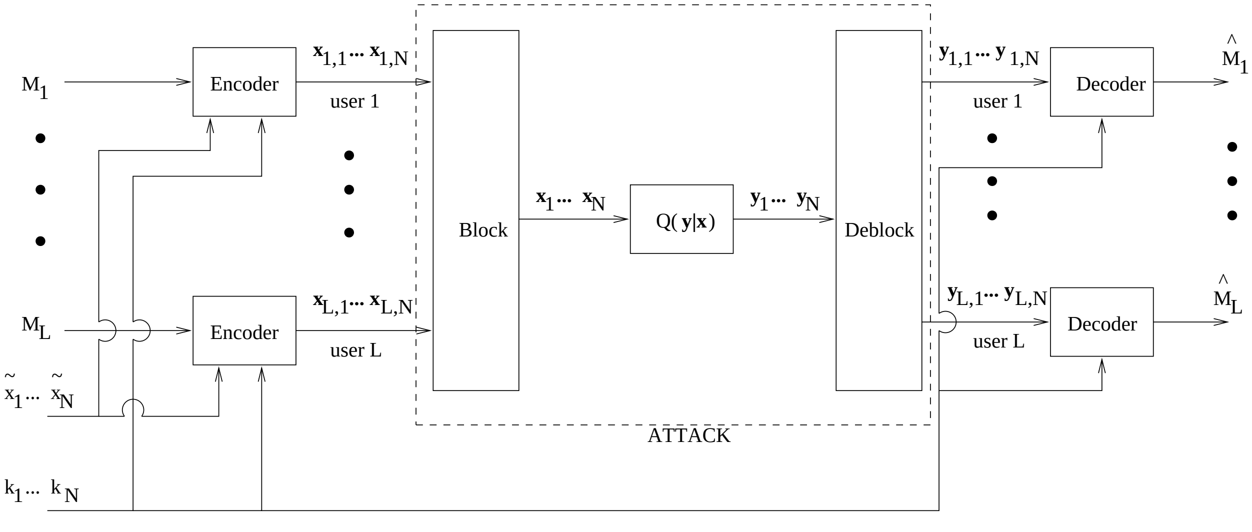

Figure 9: Block-diagram representation of the fingerprinting problem.

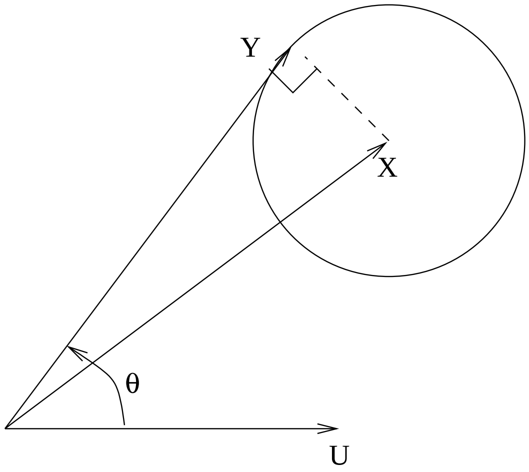

Figure 10: Minimization of $I(U;Y)$.