June 2001

UILU-ENG-01-2214
CRHC 2001-04

THE PARALLEL-GAUSSIAN WATERMARKING GAME

Pierre Moulin and M. Kivanc Mihcak

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1308 West Main Street, Urbana, IL 61801

SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB NO. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE MAY 2001	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE The Parallel-Gaussian Watermarking Game	5. FUNDING NUMBERS CCR 00-81268 CDA 96-24396
6. AUTHOR(S) Pierre Moulin and M. Kivanc Mihcak	
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES) Coordinated Science Laboratory 1308 W. Main Street Urbana, IL 61801	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Science Foundation	10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.	12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Rates of reliable transmission of hidden information are derived for watermarking problems involving parallel Gaussian sources, which are often used to model host images and audio signals. Constraints are imposed on the average squared-error distortion that can be introduced by the information hider and by the attacker. When distortions are measured with respect to the original host data, the optimal attack is the cascade of a bank of minimum-mean-squared-error estimators for the host data and a bank of Gaussian test channels. The solution to the watermarking game involves an optimal allocation of distortions by the information hider and by the attacker to the different channels. While the resulting maxmin optimization problem is nonconcave with respect to the maximizing variable, we present a reparameterization that maps the original problem into a convex programming problem with separable cost function and separable constraints. A fast algorithm is given for computing the optimal solution. For each channel we derive analytical expressions for two asymptotic regimes: weak and strong host signals. Finally we extend these results to the class of stationary Gaussian host signals with bounded, continuous spectral density. This analysis provides an upper bound on watermarking capacity for non-Gaussian host signals.

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

The Parallel-Gaussian Watermarking Game [*]

Pierre Moulin and M. Kıvanç Mıhçak
University of Illinois
Beckman Inst., Coord. Sci. Lab & ECE Dept.
405 N. Mathews Ave., Urbana, IL 61801
Email: *moulin@ifp.uiuc.edu*

December 16, 2000. Revised June 15, 2001

Abstract

Rates of reliable transmission of hidden information are derived for watermarking problems involving parallel Gaussian sources, which are often used to model host images and audio signals. Constraints are imposed on the average squared-error distortion that can be introduced by the information hider and by the attacker. When distortions are measured with respect to the original host data, the optimal attack is the cascade of a bank of minimum-mean-squared-error estimators for the host data and a bank of Gaussian test channels. The solution to the watermarking game involves an optimal allocation of distortions by the information hider and by the attacker to the different channels. While the resulting maxmin optimization problem is nonconcave with respect to the maximizing variable, we present a reparameterization that maps the original problem into a convex programming problem with separable cost function and separable constraints. A fast algorithm is given for computing the optimal solution. For each channel we derive analytical expressions for two asymptotic regimes: weak and strong host signals. Finally we extend these results to the class of stationary Gaussian host signals with bounded, continuous spectral density. This analysis provides an upper bound on watermarking capacity for non-Gaussian host signals.

Index terms: Watermarking, game theory, channel capacity, rate-distortion theory, parallel Gaussian channels, random processes.

[*] Work supported by NSF grants CCR 00-81268 and CDA 96-24396.

1 Introduction

The widespread dissemination of images, video, audio and text data on public communication networks raises intellectual-property and security issues that can be addressed using watermarking and data hiding techniques. Other applications of data hiding include close captioning and embedding of text and audio in images and video. These areas have seen the development of a plethora of algorithms in the last five years [2, 3], but an information-theoretic treatment of the problem is just emerging [4, 5, 6, 7, 8, 9]. In particular, a theory has recently been developed to establish the fundamental limits of the watermarking (data hiding) problem depicted in Fig. 1 [4, 5].

In this framework, a message M is to be embedded in a length-N sequence $S^N = (S_1, \cdots, S_N)$ termed *host data set*, typically data from an host image, video, or audio signal. The embedding is done using a cryptographic key. The resulting *watermarked data* $X^N = (X_1, \cdots, X_N)$ are subject to *attacks* that attempt to remove any trace of M from X^N. The output of the attack is a sequence $Y^N = (Y_1, \cdots, Y_N)$. The decoder has access to Y^N and the key and produces an estimate $\hat{M}$ for the message that was transmitted.

The watermarking system should satisfy two basic requirements. The first is usually referred to as *transparency*, or *unobtrusiveness*: the data set X^N should be similar to S^N, according to a suitable distortion measure. The second requirement is referred to as *robustness*: the hidden message should survive the application of any attack (within a certain class) to X^N. For instance, there is typically a limit on the amount of distortion that an attacker is willing to introduce. A watermarking system can be analyzed by defining a statistical model for S^N and the key, a distortion function, and specifying constraints on the admissible distortion levels for the information hider and the attacker. In particular, we seek the *maximum rate of reliable transmission for M*, over *any* possible watermarking strategy and *any* attack that satisfy the specified constraints.

Following a brief review of background in Sec. 2, three related problems are considered in this paper. First, in Sec. 3, capacity expressions are derived for Gaussian channels when all distortions are evaluated with respect to the host data. The solution serves as a building block for the second problem, in Secs. 4 and 5. There the source S can be decomposed as a parallel-Gaussian source, with K channels carrying independent and identically distributed (i.i.d.) host data. In typical applications, S would be a K-dimensional block of transform data (such as an 8×8 block of discrete cosine transform coefficients, or a subtree of wavelet coefficients) from an host image, video, or audio signal. In the third problem (Sec. 6), the source S is a stationary Gaussian random process with bounded and continuous spectral density. The paper concludes with a discussion in Sec. 7.

Notation. We use capital letters to denote random variables, small letters to denote their individual values, and a superscript N to denote length-N vectors. We let $I(X;Y)$ denote the mutual information between two random variables X and Y, and $I(X;Y|Z)$ denote the conditional mutual information between X and Y, conditioned on Z.

2 Background

2.1 Mathematical Model

The problems in Secs. 3—5 admit the following general description. The host-data source emits an i.i.d. sequence of K-dimensional Gaussian random vectors $S \sim \mathcal{N}(0, R)$, where R is a $K \times K$ correlation matrix. The distortion metric is the squared Euclidean distance $d(x, y) = \|x - y\|^2$, for $x, y \in \mathbb{R}^K$.

In Sec. 3, we have $K = 1$. In Sec. 4, we assume that the correlation matrix is diagonal with diagonal entries $\sigma_k^2, 1 \le k \le K$, and denote it by $\Sigma = diag\{\sigma_k^2\}_{k=1}^K$. Equivalently, S may be represented by means of K parallel Gaussian channels. The channel inputs are K independent sources $S_k, 1 \le k \le K$, each producing i.i.d. Gaussian random variables $\mathcal{N}(0, \sigma_k^2)$. If the host-signal correlation matrix is not diagonal, the problem can be reduced to the above case by means of the Karhunen-Loève transform, see Sec. 5.

The message M of interest is uniformly distributed over the message set $\mathcal{M}$, and is to be reliably transmitted to the decoder. M is independent of S.

The decoder has access to side information. It is assumed that randomized watermarking codes are used, and that the decoder knows the cryptographic key used to select the particular code used. If the decoder has access to no other side information, the problem is referred to as *blind watermarking.* If in addition the decoder has access to the original host signal, the problem is referred to as *private watermarking.*

Maximum distortion levels are specified for the information hider and the attacker. Let $d^N(x^N, y^N) = \frac{1}{N}\sum_{i=1}^N d(x_i, y_i)$. A *length-$N$ watermarking code subject to distortion D_1* is a triple $(\mathcal{M}, f_N, \phi_N)$, where: $\mathcal{M}$ is the message set of cardinality $|\mathcal{M}|$; $f_N : \mathbb{R}^{NK} \times \mathcal{M} \to \mathbb{R}^{NK}$ is the encoder mapping and is subject to the distortion constraint

$$\sum_{s^N \in \mathbb{R}^{NK}} \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{M}|} p(s^N)\, d^N(s^N, f_N(s^N, m)) \le D_1; \tag{2.1}$$

and $\phi_N : \mathbb{R}^{NK} \to \mathcal{M}$ (resp. $\phi_N : \mathbb{R}^{NK} \times \mathbb{R}^{NK} \to \mathcal{M}$) is the decoder mapping, producing the decoded message $\hat{m} = \phi_N(y^N)$ (resp. $\hat{m} = \phi_N(y^N, s^N)$) for blind watermarking (resp. private watermarking). The choice of f_N induces a conditional probability density function (p.d.f.) $p(x^N|s^N)$ on the watermarked data.

A *memoryless attack channel, subject to distortion D_2*, is a conditional p.d.f. $A(y|x)$, $x, y \in \mathbb{R}^K$, subject to linear distortion constraints. The length-N extension of this channel is defined as

$A^N(y^N|x^N) = \prod_{i=1}^N A(y_i|x_i)$. Two types of distortion constraints are considered, giving rise to two distinct classes of attack channels.

Type X. Constraint on $Ed(X^N, Y^N)$:

$$\int\int d^N(x^N, y^N) A^N(y^N|x^N)\, p(x^N)\, dx^N\, dy^N \le D_2. \tag{2.2}$$

Type S. Constraint on $Ed(S^N, Y^N)$:

$$\int\int\int d^N(s^N, y^N) A^N(y^N|x^N)\, p(x^N|s^N) p(s^N)\, ds^N\, dx^N\, dy^N \le D_2. \tag{2.3}$$

For Type-S constraints, we normally require $D_2 \ge D_1$, so that the set of attack channels includes $Y^N = X^N$ (no attack). The distortions for the information hider and the attacker are equal in this special case. The scenario $D_2 < D_1$ appears to have only limited practical interest; in fact, the set of attack channels that satisfy (2.3) is empty if D_2 is too small.

The rate of the watermarking code is $R = \frac{1}{N}\log|\mathcal{M}|$, and the average probability of error is $P_{e,N} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} P(\phi_N(Y^N) \ne m \mid M = m)$ and $P_{e,N} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} P(\phi_N(Y^N, S^N) \ne m \mid M = m)$ for blind and private watermarking, respectively. A rate R is said to be 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 admissible, memoryless attack. The watermarking capacity is the supremum of all achievable rates for distortions (D_1, D_2).

2.2 Watermarking Capacity

The paper [5] has shown that the watermarking capacity defined above is the value of a mutual-information game between the information hider and the attacker. This result is stated in Theorem 2.1 below. In order to minimize the payoff, the attacker designs an optimal memoryless attack channel $A(y|x)$ that satisfies the distortion constraint

$$\int\int d(x,y) A(y|x) p(x)\, dxdy \le D_2 \tag{2.4}$$

under type-X constraints, and

$$\int\int\int d(s,y) A(y|x) p(x|s) p(s)\, dsdxdy \le D_2 \tag{2.5}$$

under type-S constraints. In order to maximize the payoff, the information hider optimally designs a *covert channel* $Q(x, u|s)$, where U is an auxiliary $\mathbb{R}^K$-valued random variable. The covert channel satisfies the distortion constraint

$$\int\int\int d(s,x) Q(x,u|s) p(s)\, dxdsdu \le D_1. \tag{2.6}$$

We let $\mathcal{A}_X(D_2)$, $\mathcal{A}_S(D_2)$, and $\mathcal{Q}(D_1)$ be the set of channels that satisfy the constraints (2.4), (2.5), and (2.6), respectively. We omit the subscript X or S for results that apply to both $\mathcal{A}_X(D_2)$ and $\mathcal{A}_S(D_2)$.

For any arbitrarily complicated encoding scheme and memoryless attack, Theorem 2.1 upper bounds the rate of reliable transmission for the information hider, under the assumptions that the attacker knows f_N, and that the decoder knows both f_N and A.

Theorem 2.1 *[5] Assume the decoder knows the attack channel. A rate R is achievable subject to distortions (D_1, D_2) if and only if $R < C$, where*

$$C = \max_{Q(x,u|s)\in\mathcal{Q}(D_1)} \min_{A(y|x)\in\mathcal{A}(D_2)} J(Q,A) \tag{2.7}$$

where

$$J(Q,A) = \begin{cases} I(U;Y) - I(U;S) & : \textit{for blind watermarking} \\ I(U;Y|S) & : \textit{for private watermarking.} \end{cases} \tag{2.8}$$

For private watermarking, $U = X$ is optimal.

2.3 Gaussian Channels – Type-X Constraints

Assume that $K = 1$, $S \sim \mathcal{N}(0, \sigma^2)$, and $d(x,y) = (x-y)^2$ (squared-error distortion on the real line). The capacity-achieving distributions have been explicitly calculated under the type-X distortion constraints (2.2) [5]. The capacity is the same for both blind and private watermarking problems:

$$C = \begin{cases} \frac{1}{2}\log\left(1 + \frac{D_1}{\beta D_2}\right) & : \text{ if } D_2 < \sigma^2 + D_1, \\ 0 & : \text{ otherwise} \end{cases} \tag{2.9}$$

where

$$\beta = \left(1 - \frac{D_2}{\sigma^2 + D_1}\right)^{-1} \geq 1. \tag{2.10}$$

For small distortions ($\sigma^2 >> D_1, D_2$) we have $\beta \sim 1$, and so $C \sim \frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right)$, i.e., the capacity expression is *asymptotically independent* of σ^2. The optimal attack is the Gaussian test channel from rate-distortion theory [10], $A(y|x) = \mathcal{N}(\beta^{-1}x, \beta^{-1}D_2)$. The noise introduced by this optimal attack channel is independent of the channel input.

For blind watermarking, the optimal covert channel $Q(x,u|s)$ is given by

$$X = S + Z, \tag{2.11}$$

$$U = Z + \alpha S, \tag{2.12}$$

where

$$\alpha = \frac{D_1}{D_1 + \beta D_2}, \tag{2.13}$$

and $Z \sim \mathcal{N}(0, D_1)$ is independent of S. The optimal distribution $Q(x,u|s)$ the same optimal distribution that achieves capacity in a problem studied by Costa [11]. For private watermarking, the optimal $Q(x,u|s)$ is given by (2.11), with $U = Z$ or $U = X$.

3 Gaussian Channels – Type-S Constraints

For type-S distortion constraints, we show the optimal covert channel is similar to that given in Sec. 2.3, but the optimal attack is no longer the Gaussian test channel. The optimal attack is now the cascade of the minimum-mean-squared-error (MMSE) estimator of S followed by a Gaussian test channel introducing the maximum possible distortion, see Fig. 2. The solution is stated in Theorem 3.1.

Theorem 3.1 *Let $K = 1$ and $d(x, y) = (x - y)^2$ be the squared-error distortion measure. Let the distortion constraints for the information hider and the attacker be given by (2.6) and (2.5), respectively. Assume that $S \sim \mathcal{N}(0, \sigma^2)$ and that $D_2 \geq \frac{\sigma^2}{\sigma^2+D_1} D_1$.*

(i) If $D_2 \geq \sigma^2$, the optimal attack channel is given by $Y = 0$, and the watermarking capacity is $C = 0$.

(ii) For blind watermarking with $\frac{\sigma^2}{\sigma^2+D_1} D_1 \leq D_2 < \sigma^2$, the optimal covert channel is given by (2.11) (2.12), where $\alpha = \frac{D_1}{D_1+D}$, $D = -D_1 + \frac{\sigma^2}{\sigma^2-D_2} D_2$, and $Z \sim \mathcal{N}(0, D_1)$ is independent of S.

(iii) For private watermarking, the optimal covert channel is the same as with type-X constraints.

(iv) For both blind and private watermarking with $\frac{\sigma^2}{\sigma^2+D_1} D_1 \leq D_2 < \sigma^2$, the optimal attack channel A is given by $Y = (X + W)/\beta$, where $W \sim \mathcal{N}(0, D)$ is independent of X, and $\beta = \frac{\sigma^2}{\sigma^2-D_2}$. Equivalently, $A(y|x) = A^*(y|\hat{s})$, where $\hat{S} = \frac{\sigma^2}{\sigma^2+D_1} X$ is the MMSE estimator for S given X, and A^* is the Gaussian test channel with distortion equal to $D' = -\frac{\sigma^2}{\sigma^2+D_1} D_1 + D_2$.

(v) For both blind and private watermarking with $\frac{\sigma^2}{\sigma^2+D_1} D_1 \leq D_2 < \sigma^2$, the watermarking capacity is given by

$$C = \Gamma(\sigma^2, D_1, D_2) = \frac{1}{2} \log \left(1 + \frac{D_1}{D}\right) = \frac{1}{2} \log \left(1 - D_1 (\frac{1}{D_2} - \frac{1}{\sigma^2})\right)^{-1}. \tag{3.1}$$

Proof. See appendix.

Note #1. For $D_2 = \frac{\sigma^2}{\sigma^2+D_1} D_1$ (necessarily smaller than σ^2), the admissible set of attack channels reduces to one single element, namely, the MMSE estimator $Y = \frac{\sigma^2}{\sigma^2+D_1} X$. This attack is reversible, and $C = \infty$ in that case. For $D_2 < \frac{\sigma^2}{\sigma^2+D_1} D_1$, the admissible set of attack channels is empty.

Note #2. For small distortions ($\sigma^2 >> D_1, D_2$), we have $\beta \sim 1$, $D \sim D_2 - D_1$, and $C \sim \frac{1}{2} \log \left(1 + \frac{D_1}{D_2 - D_1}\right)$.

4 Parallel Gaussian Channels

In this section, we develop watermarking capacity expressions for parallel Gaussian channels, and specialize the results to sparse signal models, which have been used in recent literature.

4.1 Main Result

First we prove that optimal watermarking and attack strategies decouple the K channels and make use of Gaussian distributions in each channel, see Fig. 3. The power allocations for the information hider and the attacker are denoted by $d_1 = \{d_{1k}, 1 \le k \le K\}$ and $d_2 = \{d_{2k}, 1 \le k \le K\}$, respectively. Define the host-signal rates $r_k = 1/K$, $1 \le k \le K$. Later we shall see that channels with same variance can be combined, yielding nonuniform $\{r_k\}$.

Lemma 4.1 *The watermarking capacity for both blind and private parallel-Gaussian watermarking games subject to distortion constraints (D_1, D_2) is equal to*

$$C = \max_{d_1} \min_{d_2} \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k}), \tag{4.1}$$

where

$$\begin{aligned}\Gamma(\sigma_k^2, d_{1k}, d_{2k}) &= \frac{1}{2} \log \left(1 - d_{1k}\left(\frac{1}{d_{2k}} - \frac{1}{\sigma_k^2}\right)\right)^{-1} &(4.2)\\ &= \frac{1}{2} \log \left(\frac{\sigma_k^2 + d_{1k}}{\sigma_k^2} - \frac{d_{1k}}{d_{2k}}\right)^{-1}, &(4.3)\end{aligned}$$

and the maximization and minimization are subject to the overall distortion constraints

$$\begin{aligned}\sum_{k=1}^{K} r_k d_{1k} &\le D_1 &(4.4)\\ \sum_{k=1}^{K} r_k d_{2k} &\le D_2 &(4.5)\end{aligned}$$

and the inequality constraints (see Fig. 4)

$$\begin{aligned}0 &\le d_{1k} &(4.6)\\ \frac{\sigma_k^2}{\sigma_k^2 + d_{1k}} d_{1k} &\le d_{2k} &(4.7)\\ d_{2k} &\le \sigma_k^2, &(4.8)\end{aligned}$$

for $1 \le k \le K$. The capacity-achieving distributions are of the form $Q(x^K, u^K|s^K) = \prod_{k=1}^{K} Q_k(x_k, u_k|s_k)$ and $A(y^K|x^K) = \prod_{k=1}^{K} A_k(y_k|x_k)$, where Q_k and A_k are the capacity-achieving distributions for a single Gaussian channel with distortion levels d_{1k} and d_{2k}, respectively.

Proof. The host-data source may be viewed as a blockwise memoryless scalar source with block length K, where data within any given block $S^K = (S_1, S_2, \cdots, S_K)$ are independent (but not identically distributed). The class $\mathcal{A}(D_2)$ may be similarly viewed as a set of blockwise memoryless attack channels. Hence we can apply Prop. 8.2 in [5], which states that the optimal blockwise memoryless channel is memoryless: $A(y^K|x^K) = \prod_{k=1}^K A_k(y_k|x_k)$. Conversely, a straightforward extension of Prop. 4.2 in [5] shows that under the assumption of a memoryless attack channel, the optimal covert channel is also memoryless: $Q(x^K, u^K|s^K) = \prod_{k=1}^K Q_k(x_k, u_k|s_k)$. In other words, the optimal pair of information hiding and attack strategies leaves the K channels decoupled. Let $d_{1k} = E|S_k - X_k|^2$ and $d_{2k} = E|Y_k - X_k|^2$ be the distortion levels in channel k under these optimal strategies. The distortion levels $d_1 = \{d_{1k}\}$ and $d_2 = \{d_{2k}\}$ satisfy the constraints (4.4)—(4.8). The capacity-achieving distributions Q_k and A_k are obtained by applying Theorem 3.1 with distortion levels d_{1k} and d_{2k}. These distributions are Gaussian. The resulting capacity for channel k is given by (4.2), and the total capacity is given by (4.1). □

It remains to optimally allocate the powers $d_1 = \{d_{1k}\}$ and $d_2 = \{d_{2k}\}$ between the K channels. Theorem 4.3 below reduces this problem to a simple convex/concave programming problem and presents its solution. The proof of Theorem 4.3 uses the following lemma.

Lemma 4.2 *The function $\Gamma(\sigma_k^2, d_{1k}, d_{2k})$ in (4.2) is convex in d_{1k} and is convex in d_{2k}, over the set defined by the inequality constraints (4.6)-(4.8).*

Proof: see appendix.

The main ideas used in the proof of Theorem 4.3 are:

1. The payoff function $\sum_{k=1}^K r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k})$ is additive over k, and so are the distortion constraints (4.4) and (4.5). The other $3K$ constraints (4.6) (4.7), and (4.8) apply to each channel separately.

2. For any d_1, the constrained minimization problem is reformulated as the dual maximization problem $\max_{\lambda_2 \geq 0} q(d_1, \lambda_2)$, where the dual variable $\lambda_2 \geq 0$ corresponds to the distortion constraint (4.5).

3. A closed-form solution for each optimal d_{2k} is derived in terms of d_{1k} and λ_2.

4. The function $q(d_1, \lambda_2)$ is nonconcave with respect to d_1, but a reparameterization is found that makes the cost function concave.

5. The constraint set is still convex under the reparameterization above. The maximization problem is converted to a dual minimization problem $\min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2)$, where the dual variable $\lambda_1 \leq 0$ corresponds to the distortion constraint (4.4).

over d_1

6. We have $C = \max_{\lambda_2 \geq 0} \min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2)$ where r is strictly convex in λ_1. This maxmin problem is solved using a standard numerical algorithm.

A numerical optimization algorithm based on these properties is described in Appendix D. The dual variables λ_1 and λ_2 represent sensitivity parameters with respect to changes in distortion levels D_1 and D_2.

Theorem 4.3 *The optimal power allocations d_1, d_2 for the watermarking game in Lemma 4.1 are as follows. If $\sum_{k=1}^{K} r_k \sigma_k^2 \leq D_2$, the optimal attack is $d_{2k} = \sigma_k^2 \ \forall k$; d_1 is arbitrary, and $C = 0$. If $\sum_{k=1}^{K} r_k \sigma_k^2 > D_2$, the watermarking game admits a unique solution. For each $1 \leq k \leq K$, the optimal d_{2k} is zero if $\sigma_k^2 = 0$; otherwise d_{2k} is the unique root of the fourth-order polynomial*

$$p_k(d_{2k}) = \frac{\lambda_2}{\sigma_k^4} {d_{2k}}^4 + \left(\frac{1}{\sigma_k^4} - \frac{2\lambda_2}{\sigma_k^2}\right) {d_{2k}}^3 + \left(\lambda_2 + \lambda_1 - \frac{5}{2\sigma_k^2}\right) {d_{2k}}^2 + \left(\frac{\lambda_1}{\lambda_2} + \frac{3}{2} - \frac{1}{2\lambda_2 \sigma_k^2}\right) d_{2k} + \frac{1}{2\lambda_2} \tag{4.9}$$

in the semi–open interval $(0, \sigma_k^2]$. The optimal d_{1k} is given by

$$d_{1k} = \left[1/(2\lambda_2 {d_{2k}}^2) + 1/d_{2k} - 1/\sigma_k^2\right]^{-1}. \tag{4.10}$$

The Lagrange multipliers $\lambda_1 \leq 0$ and $\lambda_2 \geq 0$ are such that the distortion constraints (4.4) and (4.5), respectively, are satisfied with equality.

Proof: see appendix.

4.2 Properties of Solution

1. (*Symmetry*). If $\sigma_k^2 = \sigma_l^2$ for some $k \neq l$, then $d_{1k} = d_{1l}$ and $d_{2k} = d_{2l}$. This follows directly from (4.9), as the polynomials p_k and p_l are identical: we have the same optimization problem, and hence the same solution, in channels k and l. Hence the problem can be reduced to one involving distinct channel variances $\{\sigma_k^2\}$ and nonuniform host-signal rates $\{r_k\}$, where r_k is the fraction of channels that have variance σ_k^2. Theorem 4.3 holds for arbitrary $\{r_k\}$.

2. (*Bounds on Optimal power allocation*). The right side of (4.10) is strictly increasing in d_{2k}. From (4.8) and (4.10), we have:

$$d_{1k} \leq 2\lambda_2 \sigma_k^4. \tag{4.11}$$

4.3 Asymptotics of Weak Channels

It is shown below that if $\sigma_k^2 \to 0$, the bounds (4.8) and (4.11) become tight.

Proposition 4.4 *If $\sigma_k^2 \to 0$ for some channel k, the optimal power allocations in that channel are*

$$\begin{aligned} d_{2k} &\sim \sigma_k^2 - 3\lambda_2\sigma_k^4, & (4.12)\\ d_{1k} &\sim 2\lambda_2\sigma_k^4, & (4.13)\end{aligned}$$

and the contribution of channel k to capacity is linear in σ_k^2:

$$\Gamma(\sigma_k^2, d_{1k}, d_{2k}) \sim 3\lambda_2^2\sigma_k^4. \tag{4.14}$$

Proof. When $\sigma_k^2 \to 0$, the polynomial (4.9) is asymptotic to

$$p_k(d_{2k}) \sim \frac{1}{\sigma_k^4}{d_{2k}}^3 - \frac{5}{2\sigma_k^2}{d_{2k}}^2 - \frac{1}{2\lambda_2\sigma_k^2}d_{2k} + \frac{1}{2\lambda_2}. \tag{4.15}$$

Due to (4.8), the only possible asymptotic balance that satisfies $p_k(d_{2k}) = 0$ is between the last two terms of (4.15). This yields $d_{2k} \sim \sigma_k^2$. The next order term will be needed to derive (4.14). Writing $d_{2k} \sim \sigma_k^2 + \alpha\sigma_k^4$ and substituting this expression for d_{2k} in (4.15), we obtain $\alpha = -3\lambda_2$; hence (4.12) follows. In (4.10), the first term in brackets dominates the sum; hence (4.13) follows. Finally, the contribution of channel k to capacity is obtained from (4.2), (4.12) and (4.13):

$$\begin{aligned}\Gamma(\sigma_k^2, d_{1k}, d_{2k}) &\sim -\frac{1}{2}\log(1 - 2\lambda_2\sigma_k^4(3\lambda_2))\\ &\sim 3\lambda_2^2\sigma_k^4.\end{aligned}$$

□

4.4 Asymptotics of Strong Channels

Proposition 4.5 *If $\sigma_k^2 \to \infty$ for some channel k, we have $\lambda_1 + \lambda_2 < 0$, and the optimal power allocations are given by*

$$\begin{aligned} d_{2k} &\sim -\frac{1}{2(\lambda_1+\lambda_2)} & (4.16)\\ d_{1k} &\sim \frac{\lambda_2}{2\lambda_1(\lambda_1+\lambda_2)}. & (4.17)\end{aligned}$$

The contribution of channel k to capacity is

$$\Gamma(\sigma_k^2, d_{1k}, d_{2k}) \sim \frac{1}{2}\log\frac{\lambda_1}{\lambda_1+\lambda_2}. \tag{4.18}$$

Proof. If $\sigma_k^2 \to \infty$, the polynomial equation (4.9) reduces to a quadratic equation in d_{2k} (the cubic and quartic terms vanish because d_{2k} is bounded from above):

$$p_k(d_{2k}) \sim (\lambda_1 + \lambda_2){d_{2k}}^2 + \left(\frac{3}{2} + \frac{\lambda_1}{\lambda_2}\right) d_{2k} + \frac{1}{2\lambda_2}. \tag{4.19}$$

The quadratic expression in the right side has two roots at $-\frac{1}{\lambda_2}$ and $-\frac{1}{2(\lambda_1+\lambda_2)}$. The first is negative, so the second must be positive. Hence the first part of the claim.

Now (4.10) yields

$${d_{1k}}^{-1} \sim \frac{1}{2\lambda_2 {d_{2k}}^2} + \frac{1}{d_{2k}},$$

whence (4.17). Using (4.2), (4.17) and (4.16), we obtain (4.18). □

4.5 Spike processes

Recently Weidman and Vetterli have introduced a simplified model for sparse signal compression [12]. The signal is decomposed into a set of significant components and a set of insignificant components. This model can also be used to derive approximate closed-form expressions for watermarking capacity. Assume there exists an integer $K^* \le K$ such that $\sigma_k^2 \to \infty$ for $1 \le k \le K^*$, and $\sigma_k^2 \to 0$ for $K^* < k \le K$. From Prop. 4.4, the optimal power allocations tend to zero in the weak channels ($K^* < k \le K$). From Prop. 4.5, the optimal power allocations d_{1k} and d_{2k} in the strong channels ($1 \le k \le K^*$) tend to a constant value, which is independent of k and must therefore be equal to D_1/a and D_2/a, respectively:

$$d_{1k} \to \left\{ \begin{array}{l} D_1/a \\ 0 \end{array} \right. \qquad d_{2k} \to \left\{ \begin{array}{ll} D_2/a & : \ 1 \le k \le K^* \\ 0 & : \ K^* < k \le K, \end{array} \right. \tag{4.20}$$

where $a = \sum_{k=1}^{K^*} r_k \in (0,1]$ is the fraction of significant samples of the host signal. The contribution (4.2) of a strong channel ($1 \le k \le K^*$) to capacity becomes

$$\Gamma(\sigma_k^2, d_{1k}, d_{2k}) \sim \frac{1}{2} \log \left(1 + \frac{D_1}{D_2 - D_1}\right). \tag{4.21}$$

From (4.20), we obtain

$$\begin{aligned} \frac{D_2}{a} &\sim d_{2k}, \quad 1 \le k \le K^* \\ &\sim \frac{1}{2\lambda_2} \frac{D_1}{D_2 - D_1} \end{aligned}$$

where the last line is due to the asymptotic equality of (4.18) and (4.21). Hence

$$\lambda_2 = \frac{a}{D_2}(2D_2/D_1 - 2)^{-1}. \tag{4.22}$$

Solving (4.16) (with $d_{2k} = D_2/a$) for λ_1, we obtain a similar expression:

$$-\lambda_1 = \frac{a}{D_1}(2D_2/D_1 - 1)^{-1}. \tag{4.23}$$

5 Correlated Gaussian Sources

So far we have assumed the host-data vector $S \in \mathbb{R}^K$ has diagonal correlation matrix. Assume now this correlation matrix is nondiagonal, and denote it by R. The solution to the watermarking game with squared-error distortion levels D_1 and D_2 in this case is simply obtained by diagonalizing S using the Karhunen-Loève transform, thereby converting the problem to one involving independent parallel Gaussian channels and the same distortion levels D_1 and D_2, see Fig. 5. Then by direct application of Theorem 4.3, we obtain

Proposition 5.1 *Let $d(x,y) = \|x-y\|^2$ be the squared Euclidean distortion measure in $\mathbb{R}^K$. Assume that $S \sim \mathcal{N}(0,R)$. Let $\Sigma = diag\{\sigma_k^2\}$ be a $K \times K$ diagonal matrix with the eigenvalues of R on the diagonal. The watermarking game subject to distortion constraints (D_1, D_2) is equivalent to a parallel Gaussian watermarking game with channel variances $\{\sigma_k^2\}$, and distortion constraints (D_1, D_2). The solution to this game is given in Theorem 4.3.*

If the Gaussian assumption on S is relaxed, we can obtain upper bounds on capacity. One key result in the proof is [5, Prop. 8.3], which can be extended to continuous alphabets under the regularity conditions of [5, Sec. 6].

Proposition 5.2 *Consider the blind and private watermarking problems with squared-error distortion $d(x,y) = \|x-y\|^2$, $x, y \in \mathbb{R}^K$. If S is non-Gaussian with correlation matrix R, the watermarking capacity is upper-bounded by the capacity given in Prop. 5.1.*

Proof. Assume without loss of generality that the components of the vector S are uncorrelated with variances $\{\sigma_k^2, 1 \le k \le K\}$. (If $\{S_k\}$ are correlated, apply the proposition to $\mathbf{T}S$, where $\mathbf{T}$ is the Karhunen-Loève transform). It is sufficient to prove the proposition for private watermarking, because capacity for blind watermarking can only be lower. Moreover, it is sufficient to prove the proposition for host-signal distributions of the product form $p(s) = \prod_{k=1}^K p(s_k)$, because any dependency between the components S_k would reduce capacity [5, Prop. 8.3].

By Prop. 8.3 in [5], the capacity-achieving p.d.f.'s under $p(s)$ are separable: $\pi(x|s) = \prod_{k=1}^K \pi_k(x_k|s_k)$, and $A(y|x) = \prod_{k=1}^K A_k(y_k|x_k)$. For any fixed power allocation d_1, d_2, one can seek optimal channels $\{\pi_k, A_k\}$ subject to distortion constraints $\{d_{1k}, d_{2k}\}$. We find it convenient to write the payoff function using these channels as $j(p_S, d_1, d_2)$; capacity is the solution to the optimal power allocation problem $C = \max_{d_1} \min_{d_2} j(p_S, d_1, d_2)$.

Let p_S^* be the Gaussian distribution with the same second-order statistics as p_S. The capacity under p_S^* is equal to $C^* = \max_{d_1} \min_{d_2} j(p_S^*, d_1, d_2)$; let d_1^*, d_2^* be the resulting optimal power allocations. Also let d_1, d_2 be the optimal power allocations under p_S. We have

$$C = j(p_S, d_1, d_2) \le j(p_S, d_1, d_2^*) \le j(p_S^*, d_1, d_2^*) \le j(p_S^*, d_1^*, d_2^*) = C^*$$

where the first inequality is because d_2 is optimal under p_S and d_1, the second is obtained by applying Theorem 5.1(ii) in [5] (more exactly a routine extension of that result to the case of Type-S constraints) to each channel, and the third is because d_1^*, d_2^* are optimal under p_S^*. □.

6 Stationary Gaussian Processes

Assume now that S is a d-dimensional stationary Gaussian process in $\mathbb{Z}^d$ with zero mean and continuous spectral density $\nu(f), f \in \Omega = [-\frac{1}{2}, \frac{1}{2}]^d$. It is assumed that ν is bounded away from zero and infinity: $\underline{\nu} = \min_{f \in \Omega} \nu(f) > 0$, and $\overline{\nu} = \max_{f \in \Omega} \nu(f) < \infty$. First we define the watermarking game with maximum distortion levels D_1 and D_2, and then derive its solution.

6.1 Blockwise Memoryless Approximation

The spectral representation theorem states that any stationary process can be represented as an integral of independent processes indexed by $f \in \Omega$. Moreover, given a stationary Gaussian process with bounded, continuous spectral density, one may construct a sequence of blockwise memoryless approximations (indexed by block length K) to the original stationary process. This approximation can be made arbitrarily accurate in a relative-entropy sense, as stated in Lemma 6.1 below. This allows us to relate the current problem to that studied in Sec. 5.

Let R_n be the correlation matrix of the vector S^n for $n \geq 1$. For any $N \geq 1$, let $R_{K,N}$ be the block-diagonal matrix with the first $\lfloor N/K \rfloor$ blocks equal to R_K, and the last block equal to $R_{N \bmod K}$, if $N \bmod K \neq 0$. Let $\hat{P}_K^N$ denote the Gaussian distribution with zero mean and correlation matrices $R_{K,N}$, for all $N \geq 1$.

Lemma 6.1 *The relative entropy between the actual distribution P^N of S^N and its blockwise-memoryless approximation $\hat{P}_K^N$ tends to zero in the following sense:*

$$\lim_{K \to \infty} \lim_{N/K \to \infty} \frac{1}{N} D(P^N || \hat{P}_K^N) = 0. \tag{6.1}$$

Proof. See appendix.

6.2 The Stationary-Gaussian Watermarking Game

Prop. 5.1 gives the solution of the watermarking game for blockwise-memoryless Gaussian processes:

$$C^{(K)} = \max_{d_1} \min_{d_2} \frac{1}{K} \sum_{k=1}^{K} \Gamma(\sigma_k^2, d_{1k}, d_{2k}) \tag{6.2}$$

where K is the blocklength, $\sigma_k^2, 1 \le k \le K$ are the eigenvalues of the correlation matrix R_K, and d_1 and d_2 are K-vectors that satisfy the constraints (4.4) – (4.8).

We show that $\lim_{K\to\infty} C^{(K)}$ is the value of a game in which the payoff is a functional of two power-allocation functions $d_1 = \{d_1(f), f \in \Omega\}$ and $d_2 = \{d_2(f), f \in \Omega\}$.

Theorem 6.2 *As $K \to \infty$, the sequence $C^{(K)}$ converges to the following limit:*

$$C_c = \max_{d_1} \min_{d_2} \int_\Omega \Gamma(\nu(f), d_1(f), d_2(f))\, df \tag{6.3}$$

where the subscript on C stands for "continuous", and d_1 and d_2 satisfy the two distortion constraints

$$\begin{aligned} \int_\Omega d_1(f)\, df &\le D_1, & (6.4)\\ \int_\Omega d_2(f)\, df &\le D_2, & (6.5)\end{aligned}$$

and the infinite set of constraints

$$\begin{aligned} 0 &\le d_1(f), & (6.6)\\ \frac{\nu(f)d_1(f)}{\nu(f)+d_1(f)} &\le d_2(f), & (6.7)\\ d_2(f) &\le \nu(f), & (6.8)\end{aligned}$$

for all $f \in \Omega$.

Proof: see appendix.

Note. Helly's theorem in game theory [19, Ch. 2] would be a standard tool for establishing a correspondence between the game (6.3) and the sequence of finite-dimensional approximations (6.2). However, this theorem is not applicable here, because the Helly metric $\rho(d_2, d_2') = \sup_{d_1} |H(d_1, d_2) - H(d_1, d_2')|$ (where $H(d_1, d_2)$ denotes the payoff function in (6.3)) is unbounded everywhere [1].

The payoff function in (6.3) is continuously Fréchet differentiable within the feasible set defined by the constraints (6.4)—(6.8) on d_1 and d_2. The solution to this game is given by the following proposition, whose proof parallels that of Theorem 4.3, using Fréchet derivatives instead of ordinary derivatives [13].

Proposition 6.3 *The game (6.3) subject to the constraints (6.4) – (6.8) admits the following solution. If $\int_\Omega \nu(f)\, df \le D_2$, the optimal attack is $d_2 = \nu$, d_1 is arbitrary, and $C = 0$. If*

[1] For any given d_2, one can construct a subset Ω_ϵ of Ω with positive measure and a feasible d_1 such that the difference between the left and right sides of the inequality constraint (6.7) is arbitrarily small for all $f \in \Omega_\epsilon$, thereby making $H(d_1, d_2)$ arbitrarily large.

$\int_\Omega \nu(f)\,df > D_2$, *the game admits a unique solution. For each* $f \in \Omega$, *the optimal* $d_2(f)$ *is zero if* $\nu(f) = 0$; *otherwise* $d_2(f)$ *is the unique root of the fourth-order polynomial*

$$p_f(v) = \frac{\lambda_2}{\nu^2(f)} v^4 + \left(\frac{1}{\nu^2(f)} - \frac{2\lambda_2}{\nu(f)}\right) v^3 + \left(\lambda_2 + \lambda_1 - \frac{5}{2\nu(f)}\right) v^2 + \left(\frac{\lambda_1}{\lambda_2} + \frac{3}{2} - \frac{1}{2\lambda_2 \nu(f)}\right) v + \frac{1}{2\lambda_2} \tag{6.9}$$

in the semi-open interval $(0, \nu(f)]$. *The optimal* $d_1(f)$ *is given by*

$$d_1(f) = \left[1/(2\lambda_2 d_2^2(f)) + 1/d_2(f) - 1/\nu(f)\right]^{-1}, \quad f \in \Omega. \tag{6.10}$$

The Lagrange multipliers $\lambda_1 \leq 0$ *and* $\lambda_2 \geq 0$ *are such that the distortion constraints (6.4) and (6.5), respectively, are satisfied with equality.*

The solution satisfies a symmetry property analogous to Property 1 in Sec. 4.2:

Property 6.4 *If there exist* $f, f' \in \Omega$ *such that* $\nu(f) = \nu(f')$, *then* $d_i(f) = d_i(f')$ *for* $i = 1, 2$.

6.3 Example: AR(1) Process

We present a numerical solution to the watermarking game for one-dimensional, first-order, autoregressive models with zero mean, unit variance and correlation coefficient $\rho \in [0, 1)$. The method of proof of Theorem 6.2 suggests the following approach. A uniform discretization of the range of $\log \nu$ into $K = 256$ channels is used to compute capacity using the algorithm in Appendix D. Fig. 6 shows capacity as a function of ρ when $D_1 = 0.1$ and $D_2 = 0.2$. Observe the monotonic reduction in capacity as ρ tends to 1, due to the fact that the spectral representation of the process is increasingly sparse. Fig. 7 shows the optimal power allocations and the contribution of each channel to capacity for three examples: $\rho = 0.05$, $\rho = 0.5$, and $\rho = 0.95$. The gradual decay of the power allocations as $\nu(f)$ decreases (f increases) is consistent with the weak-channel asymptotics in Sec. 4.3. The saturation of the power allocations as $\nu(f)$ becomes large ($f \to 0$) relative to D_1 and D_2 is consistent with the strong-channel asymptotics in Sec. 4.4. Fig. 8 shows the reduction in capacity as a function of D_2.

Other examples arising in image watermarking are presented in the paper [14].

7 Discussion

Optimal power allocation is a classical information-theoretic topic, arising in areas such as channel capacity and rate-distortion theory. For parallel Gaussian sources, the solution to the the optimization problems arising in the two above-mentioned areas is fairly simple and is given by the famous waterfilling and reverse-waterfilling formulas, respectively [10]. A significant challenge in watermarking for parallel-Gaussian sources comes from the game-theoretic formulation of the watermarking problem. The resulting optimization problem is maxmin rather than a simple maximization or minimization problem. The solution is more involved but still consistent with the notion, originally formulated by Cox *et al.* [1], that "watermarks should be hidden in perceptually significant signal components."

In our initial investigation of this problem, we sought an expression for capacity of parallel Gaussian channels under the type-X constraints (2.4) for the attacker. This can be done using an approach similar to the one outlined above Theorem 4.3. In this case however, we have been unable to find a reparameterization similar to the one used in Step 4, that would reduce the problem to a convex programming problem. Numerical implementation of the solution shows that significant power may be allocated by both the information hider and the attacker to weak channels (unlike the results in Sec. 4.3). This result may seem counterintuitive but makes sense from a game-theoretic point of view: the allocation of resources by the information hider to weak channels does force the attacker to use a similar strategy under type-X constraints and hence "waste" valuable power that might be better invested otherwise. This suggests that type-X constraints are less natural than type-S constraints, as no such conterintuitive solution exists in the problems studied in this paper.

Acknowledgements. The authors are grateful to T. Başar for helpful discussions regarding game theory in infinite-dimensional spaces.

A Proof of Theorem 3.1

(a) For $D_2 \geq \sigma^2$, the attack $Y = 0$ is admissible, in which case the value of the payoff function (2.8) is $J = 0$, which clearly achieves the required minimum; so $C = 0$.

(b) Blind Watermarking. Assume now that $D_2 < \sigma^2$. The proof parallels that of Theorem 5.2 in [5]. If the attack channel is as specified in (iv), it follows immediately from [11, 5, 7] that the covert channel that maximizes (2.8) is the one given in (ii). Conversely, if the covert channel is as specified above, the optimal attack channel A is the one that minimizes $I(U;Y)$. We have $I(U;Y) \geq I(U;Y^*)$ where (U, X, Y^*) is Gaussian with the same second-order statistics as (U, X, Y). Now minimize $I(U;Y)$ over all Gaussian distributions that satisfy the distortion constraint $E(S-Y)^2 \leq D_2$.

Because (X, Y) is jointly Gaussian, there exist two positive constants β and D and a random variable $W \sim \mathcal{N}(0, D)$ independent of X, such that $Y = \beta^{-1}(X+W) \stackrel{\Delta}{=} \beta^{-1}V$. Since $I(U;Y) = I(U;V)$, we have

$$J \geq \frac{1}{2}\log\left(1 + \frac{D_1}{D}\right). \tag{A.1}$$

For the attack channel $Y = \beta^{-1}V$ to be admissible, we need the following condition on D:

$$\begin{aligned} D_2 &\geq E(S-Y)^2 \\ &= E(S - V/\beta)^2 \\ &\stackrel{(a)}{=} E(S(1-1/\beta) - Z/\beta - W/\beta) \\ &\stackrel{(b)}{=} \sigma^2(1-1/\beta)^2 + (D_1 + D)/\beta^2 \end{aligned}$$

where (a) is because $V = S + Z + W$, and (b) follows from the independence of S, Z and W. We then obtain

$$D \leq -D_1 + \beta^2 D_2 - (\beta - 1)^2\sigma^2. \tag{A.2}$$

The right hand side of (A.1) is minimized over D by choosing β that maximizes $\beta^2 D_2 - (\beta-1)^2\sigma^2$, and D that achieves equality in (A.2). Hence

$$\beta = \frac{\sigma^2}{\sigma^2 - D_2}. \tag{A.3}$$

Substituting in (A.2), we obtain the largest admissible value of D as

$$D = -D_1 + \frac{\sigma^2}{\sigma^2 - D_2} D_2. \tag{A.4}$$

Hence the payoff for all admissible attacks channels is lower bounded by (A.1) where D is given by (A.4). The attack channel $Y = \beta^{-1}(X+W)$ with β and D given by (A.3) and (A.4) achieves that bound and hence is optimal. The value of C follows immediately.

Next we show that this optimal attack channel is the cascade of the MMSE estimator and a Gaussian test channel. We briefly outline the proof. Let $\hat{S} = \frac{\sigma^2}{\sigma^2+D_1}X$ be the MMSE estimator of S given X, and apply the Gaussian test channel with distortion $E(\hat{S}-Y)^2 = D'$ to $\hat{S}$. This Gaussian test channel can be viewed as the cascade of an AWGN channel with distortion level $\gamma D'$ and a multiplicative constant γ^{-1}, where

$$\gamma^{-1} = 1 - D'/\sigma_{\hat{S}}^2 = 1 - D'\frac{\sigma^2 + D_1}{\sigma^4}. \tag{A.5}$$

For this system and the optimal system $Y = (X+W)/\beta$ to be equivalent, γ and D' must satisfy the two following conditions:

$$\beta^{-1} = \gamma^{-1}\frac{\sigma^2}{\sigma^2 + D_1} \tag{A.6}$$

$$\beta^{-2}D = \gamma^{-1}D' \tag{A.7}$$

corresponding to equality of the constants that multiplies X and of the additive noise powers, respectively. Substituting the value of β from (A.3) into (A.6), we obtain

$$\gamma = \frac{\sigma^4}{(\sigma^2 - D_2)(\sigma^2 + D_1)}. \tag{A.8}$$

Substituting the values of β, D and γ from (A.3), (A.4) and (A.8) into (A.7), we obtain

$$D' = -\frac{\sigma^2}{\sigma^2 + D_1}D_1 + D_2. \tag{A.9}$$

It can then be verified by substitution that the values of γ and D' in (A.8) and (A.9) satisfy the condition (A.5).

(c) Private Watermarking. The proof parallels that of Theorem 5.1 in [5]. If the attack channel is as specified in (iv), it follows immediately from Step 1 of that theorem that the covert channel that maximizes (2.8) is the one given in (ii). Conversely, if the covert channel is given by (ii), consider any given attack channel $A(y|x)$. 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 (2.4) takes the form

$$D_2 \geq E(Y-S)^2 = E[(a-1)S + aZ + aW]^2 = (a-1)^2\sigma^2 + a^2D_1 + a^2E(W^2). \tag{A.10}$$

Proceeding as in [5], we obtain

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

where W^* is a Gaussian random variable that has zero mean and the same variance as W. The inequality above is satisfied with equality if W is Gaussian and independent of X. The lowest possible bound is obtained by maximizing $E(W^2)$ subject to the constraint (A.10), where $a \in \mathbb{R}$. After some algebra, we obtain $E(W^2) = D$ and $a = \beta$ specified in (iv). The attack channel in (iv) satisfies the lower bound with equality. □

B Proof of Lemma 4.2

The proof of convexity with respect to d_{1k} is immediate from (4.2). In order to prove convexity with respect to d_{2k}, we verify the sign of second derivatives of (4.3). Let

$$a_k \stackrel{\triangle}{=} \frac{\sigma_k^2}{\sigma_k^2 + d_{1k}} d_{1k}. \tag{B.1}$$

The admissibility conditions (4.7) (4.8) can be written as $a_k \leq d_{2k} \leq \sigma_k^2$. From (4.3) we obtain

$$\begin{aligned} C &= -\frac{1}{2}\sum_{k=1}^{K} r_k \log\left(\frac{\sigma_k^2 + d_{1k}}{d_{1k}}(1 - \frac{a_k}{d_{2k}})\right), \\ \frac{\partial C}{\partial d_{2k}} &= -\frac{1}{2} r_k \frac{a_k}{d_{2k}(d_{2k} - a_k)}, \\ \frac{\partial^2 C}{\partial {d_{2k}}^2} &= \frac{r_k a_k}{2} \frac{2d_{2k} - a_k}{{d_{2k}}^2 (d_{2k} - a_k)^2}, \end{aligned}$$

which is strictly positive for all $d_{2k} > \frac{a_k}{2}$. □

C Proof of Theorem 4.3

Let $\mathcal{D}_1$ denote the feasible set for $d_1 = \{d_{1k}\}$, i.e., the set of d_1 that satisfy (4.4) and (4.6). Given d_1, let $\mathcal{D}_2(d_1)$ denote the feasible set for $d_2 = \{d_{2k}\}$, i.e., the set of d_2 that satisfy the $2K + 1$ linear constraints (4.5), (4.7) and (4.8). Because $\mathcal{D}_2(d_1)$ depends on d_1, the feasible set for the pair (d_1, d_2) is said to be nonrectangular. We follow the general approach developed by Shimizu and Aiyoshi [22] for solving maxmin problems with nonrectangular feasible sets.

The proof of the first statement of the theorem (case $\sum_{k=1}^{K} r_k \sigma_k^2 \leq D_2$) is immediate. The proof of the more interesting case $\sum_{k=1}^{K} r_k \sigma_k^2 > D_2$ is as follows.

Step 1. First we fix the power allocation d_1 for the information hider and derive the optimal power allocation d_2 for the attacker. These parameters minimize the convex [2] cost function (4.1) subject to the convex constraint $d_2 \in \mathcal{D}_2(d_1)$. According to the strong duality theorem [21], the solution to the constrained minimization problem is given by

$$\min_{d_2 \in \mathcal{D}_2(d_1)} \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k}) = \max_{\lambda_2 \geq 0} q(d_1, \lambda_2) \tag{C.1}$$

where λ_2 is the dual variable,

$$q(d_1, \lambda_2) \stackrel{\triangle}{=} \min_{d_2} \mathcal{L}_2(d_1, d_2, \lambda_2) \tag{C.2}$$

[2] per Lemma 4.2

is the dual function, and

$$\begin{aligned}\mathcal{L}_2(d_1, d_2, \lambda_2) &= \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k}) + \lambda_2 \left(\sum_{k=1}^{K} r_k d_{2k} - D_2 \right) + \sum_{k=1}^{K} \mu_k d_{2k} \\ &= -\frac{1}{2} \sum_{k=1}^{K} r_k \log \left(\frac{\sigma_k^2 + d_{1k}}{\sigma_k^2} - \frac{d_{1k}}{d_{2k}} \right) + \lambda_2 \left(\sum_{k=1}^{K} r_k d_{2k} - D_2 \right) + \sum_{k=1}^{K} \mu_k d_{2k} \\ &= constant - \frac{1}{2} \sum_{k=1}^{K} r_k \log \left(1 - \frac{a_k}{d_{2k}} \right) + \lambda_2 \left(\sum_{k=1}^{K} r_k d_{2k} - D_2 \right) + \sum_{k=1}^{K} \mu_k d_{2k} \quad \text{(C.3)}\end{aligned}$$

is the Lagrangian. Here a_k is given in (B.1), and $\{\mu_k\}$ are Lagrange multipliers representing the constraints (4.7) and (4.8). If $\lambda_2 = 0$, $\mathcal{L}_2$ is monotonically decreasing in d_{2k}, and so the minimizing d_{2k} is equal to σ_k^2 for all k. ($\Gamma(\sigma_k^2, d_{1k}, d_{2k}) \equiv 0$ in that case). But then $\sum_k r_k d_{2k} = \sum_k r_k \sigma_k^2 > D_2$. This would violate the distortion constraint (4.5), so we must have $\lambda_2 > 0$: the distortion constraint (4.5) must be satisfied with equality.

The partial derivative of $\mathcal{L}_2$ with respect to d_{2k} is given by

$$\frac{\partial \mathcal{L}_2}{\partial d_{2k}} = -\frac{1}{2} r_k \frac{a_k}{d_{2k}(d_{2k} - a_k)} + \lambda_2 r_k + \mu_k,$$

and $\partial \mathcal{L}_2 / \partial d_{2k} = 0$ at the solution. If the constraints $d_{2k} \le \sigma_k^2$ and $d_{2k} \ge a_k$ are inactive, then $\mu_k = 0$, and the condition $\partial \mathcal{L}_2 / \partial d_{2k} = 0$ implies that d_{2k} satisfies the quadratic equation

$$d_{2k}{}^2 - a_k d_{2k} - \frac{a_k}{2\lambda_2} = 0. \tag{C.4}$$

The only positive (and hence potentially admissible) solution of (C.4) is

$$d_{2k} = \frac{a_k}{2} + \sqrt{\frac{a_k^2}{4} + \frac{a_k}{2\lambda_2}}, \tag{C.5}$$

which satisfies the constraint (4.7). If the root (C.5) is admissible, it coincides with the solution, due to the convexity of $\mathcal{L}_2$ with respect to d_{2k}. If the root is nonadmissible, then the solution is $d_{2k} = \sigma_k^2$. To summarize, denote the minimizer of (C.3) by

$$\hat{d}_{2k}(d_{1k}, \lambda_2) = \min \left\{ \frac{a_k}{2} + \sqrt{\frac{a_k^2}{4} + \frac{a_k}{2\lambda_2}},\ \sigma_k^2 \right\}, \quad 1 \le k \le K. \tag{C.6}$$

Then

$$q(d_1, \lambda_2) = \mathcal{L}_2(d_1, \hat{d}_2(d_1, \lambda_2), \lambda_2). \tag{C.7}$$

If $\hat{d}_{2k} < \sigma_k^2$, equations (C.4) and (B.1) respectively yield

$$\frac{1}{a_k} = \frac{d_{2k}{}^{-2}}{2\lambda_2} + d_{2k}{}^{-1} \tag{C.8}$$

and

$$d_{1k} = (1/a_k - 1/\sigma_k^2)^{-1}. \tag{C.9}$$

Substituting $1/a_k$ from (C.8) into (C.9), we obtain (4.10), restated below for convenience:

$$d_{1k} = \left[1/(2\lambda_2 \hat{d}_{2k}^2) + 1/\hat{d}_{2k} - 1/\sigma_k^2\right]^{-1}.$$

The condition $\hat{d}_{2k} < \sigma_k^2$ is equivalent to $d_{1k} < 2\lambda_2\sigma_k^4$. We can thus rewrite (C.6) as

$$\hat{d}_{2k}(d_{1k}, \lambda_2) = \begin{cases} \frac{a_k}{2} + \sqrt{\frac{a_k^2}{4} + \frac{a_k}{2\lambda_2}} & : \ d_{1k} < 2\lambda_2\sigma_k^4 \\ \sigma_k^2 & : \ \text{else.} \end{cases} \tag{C.10}$$

Step 2. Next we derive an expression for d_1 that achieves the maximum in (4.1), subject to the constraint $d_1 \in \mathcal{D}_1$. Write

$$\begin{aligned} C &= \max_{d_1 \in \mathcal{D}_1} \min_{d_2 \in \mathcal{D}_2(d_1)} \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}, d_{2k}) \\ &= \max_{d_1 \in \mathcal{D}_1} \max_{\lambda_2 \geq 0} q(d_1, \lambda_2) \\ &= \max_{\lambda_2 \geq 0} \max_{d_1 \in \mathcal{D}_1} q(d_1, \lambda_2) \end{aligned} \tag{C.11}$$

where the second equality uses (C.1). We have $\Gamma(\sigma_k^2, d_{1k}, \hat{d}_{2k}(d_{1k}, \lambda_2)) = 0$ for $d_{1k} \geq 2\lambda_2\sigma_k^4$ ($\hat{d}_{2k} = \sigma_k^2$). So without loss of optimality, we restrict our search for the optimal d_{1k} to the interval $[0, 2\lambda_2\sigma_k^4]$.

Step 2a. The cost function $q(d_1, \lambda_2)$ is nonconcave with respect to d_1. However, define the reparameterization

$$\tilde{d}_{1k}(d_{1k}, \lambda_2) = \hat{d}_{2k}(d_{1k}, \lambda_2), \quad 0 \leq d_{1k} \leq 2\lambda_2\sigma_k^4, \ \lambda_2 \geq 0 \tag{C.12}$$

which is a strictly increasing and hence invertible (given λ_2) mapping. Let

$$\begin{aligned} \tilde{q}(\tilde{d}_1, \lambda_2) &\stackrel{\triangle}{=} q(d_1(\tilde{d}_1, \lambda_2), \lambda_2) \\ &= \mathcal{L}_2(d_1(\tilde{d}_1, \lambda_2), \tilde{d}_1, \lambda_2) \\ &= \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}(\tilde{d}_{1k}, \lambda_2), \tilde{d}_{1k}) + \lambda_2 \left(\sum_{k=1}^{K} r_k \tilde{d}_{1k} - D_2 \right). \end{aligned} \tag{C.13}$$

The second equality follows from (C.7) and (C.12). Note that while $q(d_1, \lambda_2)$ is concave in λ_2, $\tilde{q}(\tilde{d}_1, \lambda_2)$ does not necessarily have this property.

Under the reparameterization (C.12), the feasible set $\mathcal{D}_1$ is mapped to the set

$$\tilde{\mathcal{D}}_1 = \left\{ \tilde{d}_1 \ : \ \sum_{k=1}^{K} r_k d_{1k}(\tilde{d}_{1k}, \lambda_2) \leq D_1 \right\}. \tag{C.14}$$

We have

$$\max_{d_1 \in \mathcal{D}_1} q(d_1, \lambda_2) = \max_{\tilde{d}_1 \in \tilde{\mathcal{D}}_1} \tilde{q}(\tilde{d}_1, \lambda_2). \tag{C.15}$$

<u>*Step 2b*</u>. Here we show that the function $\tilde{q}(\tilde{d}_1, \lambda_2)$ in (C.13) is concave in $\tilde{d}_1$, and that $\tilde{\mathcal{D}}_1$ is a convex set. To show the first part, it is sufficient to show that the function

$$f(\tilde{d}_{1k}) \triangleq \Gamma(\sigma_k^2, d_{1k}(\tilde{d}_{1k}, \lambda_2), \tilde{d}_{1k}) \tag{C.16}$$

is concave. Applying successively (4.2) and (4.10) twice, we obtain

$$\begin{aligned} f(\tilde{d}_{1k}) &= -\frac{1}{2}\log\left(d_{1k}({d_{1k}}^{-1} - \hat{d}_{2k}^{-1} + 1/\sigma_k^2)\right) \\ &= \frac{1}{2}\log\left(2\lambda_2 \hat{d}_{2k}^2 {d_{1k}}^{-1}\right) \\ &= \frac{1}{2}\log\left(1 + 2\lambda_2 \tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)\right). \end{aligned} \tag{C.17}$$

where $d_{1k} = d_{1k}(\tilde{d}_{1k}, \lambda_2)$. Note that $f(\tilde{d}_{1k}) = f(\sigma_k^2 - \tilde{d}_{1k})$. The first and second derivatives of f are given by

$$\begin{aligned} f'(\tilde{d}_{1k}) &= \frac{1}{2}\frac{2\lambda_2(1 - 2\tilde{d}_{1k}/\sigma_k^2)}{1 + 2\lambda_2\tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)}, \\ f''(\tilde{d}_{1k}) &= \frac{1}{2}\frac{\frac{-4\lambda_2}{\sigma_k^2}[1 + 2\lambda_2\tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)] - [2\lambda_2(1 - 2\tilde{d}_{1k}/\sigma_k^2)]^2}{[1 + 2\lambda_2\tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)]^2} \\ &< 0. \end{aligned} \tag{C.18}$$

Hence f is strictly concave. Moreover, $f(0) = f(\sigma_k^2) = 0$, and f has a unique maximum at $\tilde{d}_{1k} = \frac{1}{2}\sigma_k^2$.

For short write $g(\tilde{d}_{1k}) = d_{1k}(\tilde{d}_{1k}, \lambda_2)$. The function g is monotonically increasing, with $g(0) = 0$ and $g(\sigma_k^2) = 2\lambda_2\sigma_k^4$. We now show that g is strictly convex, which implies that the feasible set $\tilde{\mathcal{D}}_1 = \{\tilde{d}_1 \ : \ \sum_{k=1}^K r_k g(\tilde{d}_{1k}) \leq D_1\}$ is strictly convex. Indeed we have

$$g(\tilde{d}_{1k}) = \frac{\tilde{d}_{1k}^2}{1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2}, \tag{C.19}$$

$$\begin{aligned} g'(\tilde{d}_{1k}) &= \frac{2\tilde{d}_{1k}[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2] - \tilde{d}_{1k}^2(1 - 2\tilde{d}_{1k}/\sigma_k^2)}{[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^2} \\ &= \frac{\tilde{d}_{1k}^2 + \tilde{d}_{1k}/\lambda_2}{[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^2}, \end{aligned} \tag{C.20}$$

$$\begin{aligned} g''(\tilde{d}_{1k}) &= \Big[(2\tilde{d}_{1k} + 1/\lambda_2)[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^2 - 2(\tilde{d}_{1k}^2 + \tilde{d}_{1k}/\lambda_2)[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2] \\ &\qquad \times(1 - 2\tilde{d}_{1k}/\sigma_k^2)\Big] \Big/ [1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^4 \\ &= \frac{(2/\sigma_k^2)\tilde{d}_{1k}^3 + 1/(2\lambda_2^2)}{[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^3} \\ &> 0. \end{aligned}$$

Step 2c. Now define the second dual function

$$r(\lambda_1, \lambda_2) = \max_{\tilde{d}_1} \mathcal{L}_1(\tilde{d}_1, \lambda_1, \lambda_2) \tag{C.21}$$

which is convex in λ_1 (but nonconcave in λ_2.) The function $\mathcal{L}_1$ in (C.21) is the Lagrangian

$$\begin{aligned} \mathcal{L}_1(\tilde{d}_1, \lambda_1, \lambda_2) &\triangleq \tilde{q}(\tilde{d}_1, \lambda_2) + \lambda_1 \left(\sum_{k=1}^{K} r_k d_{1k}(\tilde{d}_{1k}, \lambda_2) - D_1 \right) \\ &= \sum_{k=1}^{K} r_k \Gamma(\sigma_k^2, d_{1k}(\tilde{d}_{1k}, \lambda_2), \tilde{d}_{1k}) + \lambda_2 \left(\sum_{k=1}^{K} r_k \tilde{d}_{1k} - D_2 \right) + \lambda_1 \left(\sum_{k=1}^{K} r_k d_{1k}(\tilde{d}_{1k}, \lambda_2) - D_1 \right). \end{aligned} \tag{C.22}$$

From the strong duality theorem, we have

$$\max_{\tilde{d}_1 \in \tilde{\mathcal{D}}_1} \tilde{q}(\tilde{d}_1, \lambda_2) = \min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2). \tag{C.23}$$

From (C.11), (C.15) and (C.23), we have

$$C = \max_{\lambda_2 \geq 0} \min_{\lambda_1 \leq 0} r(\lambda_1, \lambda_2). \tag{C.24}$$

Step 2d. We now evaluate (C.21). The Lagrangian (C.22) is to be maximized with respect to $\tilde{d}_1$.

First we show by contrapositive that the optimal $\lambda_1 < 0$. If $\lambda_1 = 0$, then $\tilde{d}_1$ that maximizes (C.22) is given by

$$\left. \frac{\partial \tilde{\mathcal{L}}_1}{\partial \tilde{d}_{1k}} \right|_{\lambda_1 = 0} = r_k \left[f'(\tilde{d}_{1k}) + \lambda_2 \right] = 0, \quad 1 \leq k \leq K,$$

where $f(\cdot)$ and $g(\cdot)$ are specified in (C.16) and (C.19). Using (C.18) in the equation above, we get

$$\begin{aligned} 0 &= \frac{1}{2} \frac{2\lambda_2(1 - 2\tilde{d}_{1k}/\sigma_k^2)}{1 + 2\lambda_2 \tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)} + \lambda_2, \\ 0 &= 1 - \frac{2\tilde{d}_{1k}}{\sigma_k^2} + 1 + 2\lambda_2 \tilde{d}_{1k} \left(1 - \frac{\tilde{d}_{1k}}{\sigma_k^2} \right), \\ &= 1 + \tilde{d}_{1k} \left(\lambda_2 - \frac{1}{\sigma_k^2} \right) - \frac{\lambda_2}{\sigma_k^2} \tilde{d}_{1k}^2, \\ &= \left(1 + \lambda_2 \tilde{d}_{1k} \right) \left(1 - \frac{\tilde{d}_{1k}}{\sigma_k^2} \right), \quad 1 \leq k \leq K, \end{aligned}$$

The only admissible root of this equation is $\tilde{d}_{1k} = \sigma_k^2$. But this means $\hat{d}_{2k} = \sigma_k^2$, $\forall k$, and thus $D_2 = \sum_{k=1}^{K} r_k \hat{d}_{2k} = \sum_{k=1}^{K} r_k \sigma_k^2$, which violates our initial assumption above Step 1. So $\lambda_1 < 0$, and the distortion constraint (4.4) is active.

Therefore, we necessarily have $\lambda_1 < 0$, We now show that the choice $\tilde{d}_{1k} = 0$ cannot be a maximizer of (C.22). Write (C.22) as

$$\tilde{\mathcal{L}}_1(\tilde{d}_1, \lambda_1, \lambda_2) = \sum_{k=1}^{K} r_k f(\tilde{d}_{1k}) + \lambda_2 \left(\sum_{k=1}^{K} r_k \tilde{d}_{1k} - D_2 \right) + \lambda_1 \left(\sum_{k=1}^{K} r_k g(\tilde{d}_{1k}) - D_1 \right). \tag{C.25}$$

From (C.25), (C.18) and (C.20), we have

$$\left. \frac{\partial \tilde{\mathcal{L}}_1}{\partial \tilde{d}_{1k}} \right|_{\tilde{d}_{1k}=0} = r_k f'(0) + r_k \lambda_2 + r_k \lambda_1 g'(0) = 2 r_k \lambda_2 > 0.$$

Hence the maximizing $\tilde{d}_{1k}$ must be strictly positive. For this reason we have not included a Lagrangian term corresponding to the inequality constraints (4.6), as those constraints are necessarily inactive.

Step 2e. Next we explicitly identify the optimal $\tilde{d}_1$. We have

$$\begin{aligned} \frac{\partial \tilde{\mathcal{L}}_1}{\partial \tilde{d}_{1k}} &= r_k [f'(\tilde{d}_{1k}) + \lambda_2 + \lambda_1 g'(\tilde{d}_{1k})] \\ &\overset{(a)}{=} r_k \left[\frac{1}{2} \frac{1 - 2\tilde{d}_{1k}/\sigma_k^2}{1/(2\lambda_2) + \tilde{d}_{1k}(1 - \tilde{d}_{1k}/\sigma_k^2)} + \lambda_2 + \lambda_1 \frac{\tilde{d}_{1k}^2 + \tilde{d}_{1k}/\sigma_k^2}{[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^2} \right] \\ &\overset{(b)}{=} \frac{r_k}{[1/(2\lambda_2) + \tilde{d}_{1k} - \tilde{d}_{1k}^2/\sigma_k^2]^2} \, p_k(\tilde{d}_{1k}) \end{aligned}$$

where (a) follows from (C.18) and (C.20), and $p_k(\cdot)$ in (b) is the fourth-order polynomial defined in (4.9). By the strict concavity of the Lagrangian, the gradient of $\tilde{\mathcal{L}}_1$ with respect to $\tilde{d}_1$ has a unique root. This implies that $\tilde{d}_{1k}$ is the unique root of $p_k(\cdot)$ in the interval $(0, \sigma_k^2)$. □

D Algorithm for Computing Capacity

In this section, we present an algorithm to solve the maxmin problem (C.24). The method is based on iterative one-dimensional optimization techniques and applies to the nontrivial case $\sum_{k=1}^{K} r_k \sigma_k^2 > D_2$.

Consider the Lagrangian (C.22). For a given (λ_1, λ_2) pair, let $\hat{d}_1(\lambda_1, \lambda_2), \hat{d}_2(\lambda_1, \lambda_2)$ be the solution to the game

$$\max_{d_1} \min_{d_2} \mathcal{L}(d_1, d_2; \lambda_1, \lambda_2) \tag{D.1}$$

which is explicitly stated in Theorem 4.3: for $1 \le k \le K$, $\hat{d}_{2k}(\lambda_1, \lambda_2)$ is the unique root of the fourth-order polynomial (4.9) in the interval $(0, \sigma_k^2)$, and $\hat{d}_{1k}(\lambda_1, \lambda_2)$ is given by (4.10).

Now, using the result above, our goal is to find optimal pair (λ_1, λ_2) pair in the sense of (C.24). Let $\hat{\lambda}_1(\lambda_2)$ be the solution to (C.23) for a given λ_2. The algorithm consists of an inner loop and an outer loop. The inner loop solves the inner convex minimization problem (C.23), i.e., finds $\hat{\lambda}_1(\lambda_2)$ for a given λ_2. The outer loop solves the problem

$$\max_{\lambda_2 \geq 0} r\left(\hat{\lambda}_1(\lambda_2), \lambda_2\right),$$

in an iterative fashion. For both loops, we employ the "Golden Section" method to solve the one-dimensional optimization problem. This method assumes that the cost function is strictly unimodal in a pre-defined finite interval. The Golden Section method also uses function values only for the iteration (unlike some other methods that require derivatives), thereby avoiding errors due to numerical derivation. For more details on the algorithm, we refer the reader to [23].

E Proof of Lemma 6.1

Let $r_n, n \in \mathbb{Z}$ be the correlation sequence for the process S. By our assumptions on the spectral density ν, ν^{-1} is continuous, and $1/\overline{\nu} \leq \nu^{-1} \leq 1/\underline{\nu}$. Let $t_n = \int_\Omega \nu^{-1}(f) e^{j2\pi n f}\, df$, $n \in \mathbb{Z}$ denote the inverse Fourier transform of ν^{-1}, and T_N be the $N \times N$ Toeplitz matrix with entries $t_{n-m}, 0 \leq n, m < N$. The sequences of matrices T_N and R_N^{-1} are asymptotically equivalent [17, 18], i.e., they are uniformly bounded (by $1/\underline{\nu}$) in strong norm, and $\lim_{N\to\infty} ||R_N - R_{K,N}||_{HS} = 0$, where $||A||_{HS} = Tr(A^T A)^{1/2}$ denotes the Hilbert-Schmidt (weak) norm of a matrix.

The relative entropy between the Gaussian distributions P^N and $\hat{P}_K^N$ is given by [16, 17]

$$\begin{aligned} D(P^N || \hat{P}_K^N) &= \frac{1}{2}\ln\det R_{K,N} R_N^{-1} - \frac{1}{2} + \frac{1}{2} Tr(R_{K,N} R_N^{-1}) \\ &= \frac{1}{2}\ln\det R_{K,N} - \frac{1}{2}\ln\det R_N - \frac{1}{2} + \frac{1}{2} Tr(R_{K,N} R_N^{-1}) \end{aligned}$$

where

$$\begin{aligned} \frac{1}{N}\ln\det R_{K,N} &= \frac{1}{K}\ln\det R_K + O(1/N), \\ \frac{1}{N}\ln\det R_N &\sim \int_\Omega \ln\nu(f)\, df, \\ Tr(R_{K,N} R_N^{-1}) &\sim Tr(R_{K,N} T_N), \quad \text{as } N \to \infty. \end{aligned}$$

Now

$$\alpha_l \stackrel{\triangle}{=} (R_{K,N} T_N)_{ll}, \quad 0 \leq l < N$$

is a periodic sequence with period equal to K:

$$\alpha_l = \sum_{n=-l}^{K-l-1} r_n t_n, \quad 0 \leq l < K.$$

Hence

$$\frac{1}{N}Tr(R_{K,N}R_N^{-1}) \sim \frac{1}{K}\sum_{l=0}^{K-1}\alpha_l = \sum_{n\in\mathbb{Z}} w_{K,n}r_n t_n$$

where

$$w_{K,n} \triangleq \begin{cases} 1-\frac{|n|}{K} & : \ |n| < K \\ 0 & : \ \text{else} \end{cases}$$

is a triangular window function. Defining $W_K(f) = \sum_n w_{K,n}e^{-j2\pi nf}$ and $\nu_K(f) = (\nu \star W_K)(f)$ for $f \in \Omega$, and applying Parseval's identity, we obtain

$$\frac{1}{N}Tr(R_{K,N}R_N^{-1}) \sim \sum_{n\in\mathbb{Z}}(w_{K,n}r_n)t_n = \int_\Omega \frac{\nu_K(f)}{\nu(f)}\,df$$

and so

$$\lim_{N\to\infty}\frac{1}{N}D(P^N\|\hat{P}_K^N) = \frac{1}{2}\left[-\int_\Omega \ln\nu(f)\,df + \frac{1}{K}\ln\det R_K\right] - \frac{1}{2} + \frac{1}{2}\int_\Omega \frac{\nu_K(f)}{\nu(f)}\,df. \tag{E.1}$$

Now $\lim_{K\to\infty}\frac{1}{K}\ln\det R_K = \int_\Omega \ln\nu(f)\,df$ [16]. Also $\lim_{K\to\infty}\int_\Omega \frac{\nu_K(f)}{\nu(f)}\,df = 1$, because ν_K converges uniformly to ν over Ω (by continuity of ν), and $\nu(f)$ is bounded away from zero. Hence (6.1) follows. □

F Proof of Theorem 6.2

Given two K-dimensional vectors $\underline{r} = \{r_k\}_{k=1}^K$ and $\underline{s} = \{s_k\}_{k=1}^K$, let

$$C^{(K)}(\underline{r},\underline{s}) = \max_{d_1}\min_{d_2}\sum_k r_k\Gamma(s_k,d_{1k},d_{2k}). \tag{F.1}$$

View the value of the game (6.2) as a function of the length-K vector $\Sigma = \{\sigma_k^2, 1 \le k \le K\}$ and write it as $C^{(K)}(\frac{1}{K}\underline{1},\Sigma)$. This function is continuous:

$$\forall\Sigma,\Sigma',\,\forall\epsilon>0,\,\exists\delta_K>0\ :\ \max_{1\le k\le K}|\Sigma_k-\Sigma'_k|<\delta_K \quad\Rightarrow\quad |C^{(K)}(K^{-1}\underline{1},\Sigma)-C^{(K)}(K^{-1}\underline{1},\Sigma')|<\epsilon. \tag{F.2}$$

Also note the following result from Toeplitz theory [16]: $\underline{\nu} \le \sigma_k^2 \le \overline{\nu}$ for $1 \le k \le K$.

Likewise, view the value of the game (6.3) as a functional of ν and write it as $C(\nu)$. This functional is absolutely continuous with respect to the sup norm on ν:

$$\forall\nu,\nu',\,\forall\epsilon>0,\,\exists\delta_c>0\ :\ \sup_{f\in\Omega}|\nu(f)-\nu'(f)|<\delta_c \quad\Rightarrow\quad |C(\nu)-C(\nu')|<\epsilon. \tag{F.3}$$

Choose ϵ arbitrarily small and let $\delta = \min(\delta_c,\delta_K)$, where δ_c and δ_K are given by (F.3) and (F.2), respectively. Define a partition $\{P_i, 1 \le i \le I\}$ of $[\underline{\nu},\overline{\nu}]$, where $\max_i |P_i| < \delta$. The collection of sets $\Delta_i = \{f \ : \ \nu(f) \in P_i\}$, and $\tilde{\Delta}_i = \{k \ : \ \sigma_k^2 \in P_i\}, 1 \leq i \leq I$ are partitions of the sets Ω and $\{1, 2, \cdots, K\}$, respectively. For each $1 \leq i \leq I$, let s_i, f_i, and k_i be arbitrary elements of P_i, Δ_i, and $\tilde{\Delta}_i$, respectively. Define the vectors $\underline{s}$, $\underline{r}$ and $\underline{\tilde{r}}$ with components $\{s_i, 1 \leq i \leq I\}$, $\{|\Delta_i|, 1 \leq i \leq I\}$ and $\{|\tilde{\Delta}_i|, 1 \leq i \leq I\}$, respectively.

Construct the following piecewise-constant approximations to ν and Σ:

$$\begin{aligned} \nu_\Delta &= \left\{ \sum_{1 \leq i \leq I} s_i \, \chi_{\Omega_i}(f), \quad f \in \Omega \right\}, \\ \Sigma_\Delta &= \left\{ \sum_{1 \leq i \leq I} s_i \, \chi_{\tilde{\Delta}_i}(k), \quad 1 \leq k \leq K \right\}, \end{aligned}$$

where $\chi_{\mathcal{D}}$ denotes the characteristic function of a set $\mathcal{D}$. By construction, we have

$$|C(\nu) - C(\nu_\Delta)| < \epsilon \quad \text{and} \quad |C^{(K)}(K^{-1}\underline{1}, \Sigma) - C^{(K)}(K^{-1}\underline{1}, \Sigma_\Delta)| < \epsilon. \tag{F.4}$$

Let d_1 and d_2 denote the optimal power-allocation functions under ν_Δ. Property 6.4 implies that the optimal d_1, d_2 is piecewise-constant over $\{\Delta_i\}$:

$$\begin{aligned} d_1 &= \sum_i d_1(f_i) \, \chi_{\Delta_i}, \\ d_2 &= \sum_i d_2(f_i) \, \chi_{\Delta_i}. \end{aligned}$$

We obtain

$$\begin{aligned} C(\nu_\Delta) &= \int_\Omega \Gamma(\nu_\Delta(f), d_1(f), d_2(f)) \, df \\ &= \sum_{1 \leq i \leq I} |\Delta_i| \, \Gamma(s_i, d_1(f_i), d_2(f_i)) \\ &= C^{(I)}(\underline{r}, \underline{s}). \end{aligned} \tag{F.5}$$

Similarly to (F.5), we have

$$C^{(K)}(\frac{1}{K}\underline{1}, \Sigma_\Delta) = C^{(I)}(\underline{\tilde{r}}, \underline{s}). \tag{F.6}$$

Next we use (the multidimensional extension of) the Kac-Murdock-Szegö theorem, which states that the asymptotic distribution of the eigenvalues σ_k^2 satisfies the property [15] [16, p. 64]:

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} h(\sigma_k^2) = \int_\Omega h(\nu(f)) \, df \tag{F.7}$$

for any continuous function h defined over the interval $[\underline{\nu}, \overline{\nu}]$. Applying this result [3] to $h = \chi_{P_i}, 1 \leq i \leq I$, we obtain $\lim_{K \to \infty} \underline{\tilde{r}} = \underline{r}$. By continuity of $C^{(I)}$ with respect to its first argument, we

[3] While the indicator function of a set is not continuous everywhere, the Kac-Murdock-Szegö theorem is in fact applied to an elementary modification of this function, where the variation is made arbitrarily small [17].

have $\lim_{K\to\infty} C^{(I)}(\underline{\tilde{r}}, \underline{s}) = C^{(I)}(\underline{r}, \underline{s})$. Combining this result with (F.4), (F.5), and (F.6), proves the claim. □

References

[1] 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.

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

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

[4] 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.

[5] P. Moulin and J. A. O'Sullivan, "Information–Theoretic Analysis of Information Hiding," *submitted to IEEE Trans. Information Theory*, Oct. 1999; revised, Nov. 2000. Available from `www.ifp.uiuc.edu/~moulin/Papers/ IThiding99r.ps.gz`. Short version in *IEEE Int. Symp. on Info. Thy*, p. 45, Sorrento, Italy, June 2000.

[6] N. Merhav, "On Random Coding Error Exponents of Watermarking Codes," *IEEE Trans. Info Thy*, Vol. 46, No. 2, pp. 420—430, Mar. 2000.

[7] B. Chen, "Preprocessed and Postprocessed Quantization Index Modulation Methods for Digital Watermarking, *Proc. SPIE*, Vol. 3971, San Jose, CA, Jan. 2000.

[8] B. Chen and G. W. Wornell, "Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding," *IEEE Trans. Info. Thy*, Vol. 47, No. 4, pp. 1423—1443, May 2001.

[9] A. Cohen and A. Lapidoth, "The Gaussian Watermarking Game," *Proc. CISS*, Princeton, NJ, pp. TA4/21—26, Mar. 2000.

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

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

[12] C. Weidmann and M. Vetterli, "Rate–Distortion Analysis of Spike Processes," *Proc. Data Compression Conf.*, Snowbird, UT, March 1999.

[13] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd Ed., Addison-Wesley, 1984.

[14] P. Moulin, M. K. Mıhçak and G.-I. Lin, "An Information–Theoretic Model for Image Watermarking," *Proc. Int. Conf. on Image Proc.*, Vancouver, B.C., Sep. 2000.

[15] M. Kac, W. L. Murdock and G. Szegö, "On the Eigenvalues of Certain Hermitian Forms," *J. Rat. Mech. and Anal.*, Vol. 2, pp. 767—800, 1953.

[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, 2nd Ed., Chelsea Publishing Co., New York, 1981.

[17] R. M. Gray, "On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices," *IEEE Trans. Info. Thy*, Vol. 18, No. 6, pp. 725—730, Nov. 1972.

[18] R. M. Gray, *Toeplitz Matrices*, Stanford University Technical Report, 1990.

[19] N. N. Vorob'ev, *Game Theory*, Springer Verlag, New York, 1977.

[20] R. T. Rockafellar, *Convex Analysis*, Princeton U. Press, Princeton, NJ, 1970.

[21] D. P. Bertzekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[22] K. Shimizu and E. Aiyoshi, "Necessary Conditions for Min-Max Problems and Algorithms by a Relaxation Procedure," *IEEE Trans. on Automatic Control*, Vol. 25, No. 1, pp. 62—66, 1980.

[23] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1995.

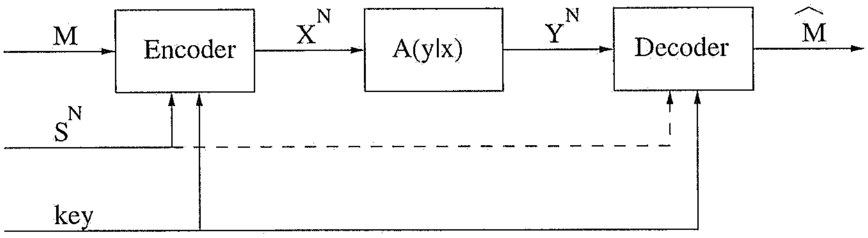


Figure 1: The watermark communication problem. In blind watermarking applications, the host data S^N are not available at the decoder; in private watermarking applications, they are.

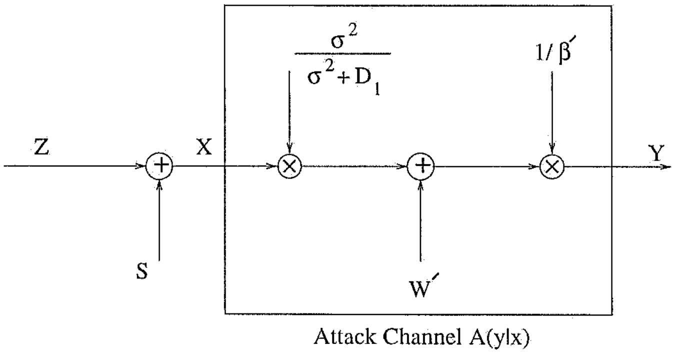


Figure 2: Optimal watermarking and attack strategies for i.i.d. Gaussian host data $S \sim \mathcal{N}(0, \sigma^2)$ under type-S distortion constraints. The optimal covert channel is given in Theorem 3.1; the optimal attack channel is the cascade of the MMSE estimator of S given X and a Gaussian test channel.

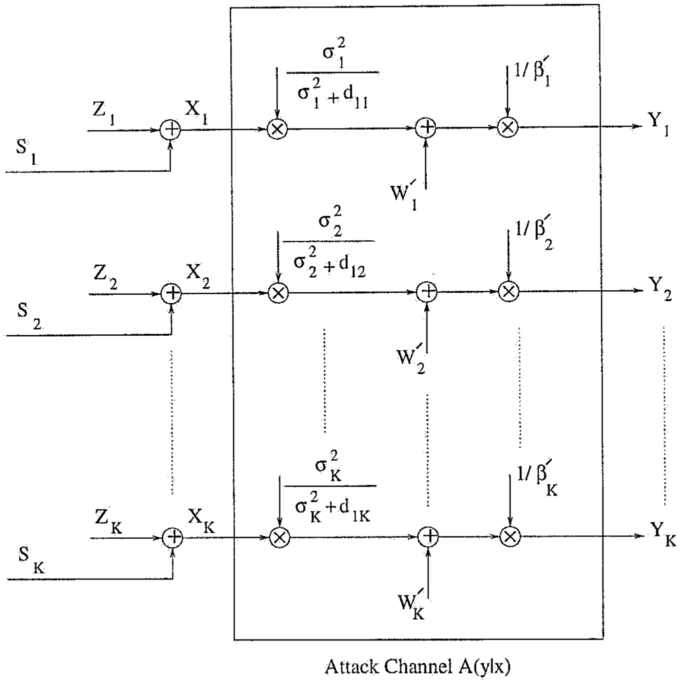


Figure 3: Optimal watermarking and attack strategies for parallel Gaussian channels $S_k \sim \mathcal{N}(0, \sigma_k^2), 1 \leq k \leq K$. The channels are decoupled, with optimal embedding and attack strategies in each channel as in Fig. 2. The optimal power allocation between channels is given by Theorem 4.3.

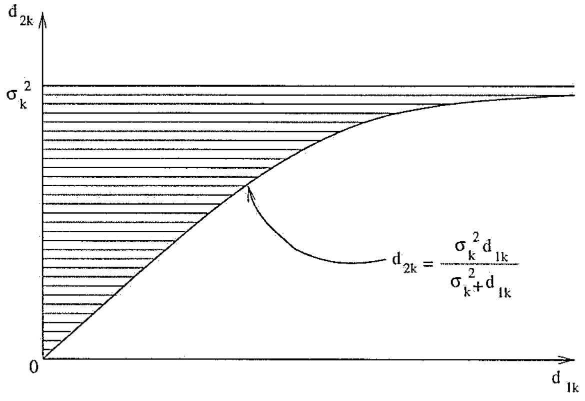


Figure 4: Nonconvex, shaded region represents values for d_{1k} and d_{2k} that satisfy the constraints (4.6) (4.7) (4.8). Distortions constraints are not represented in this figure.

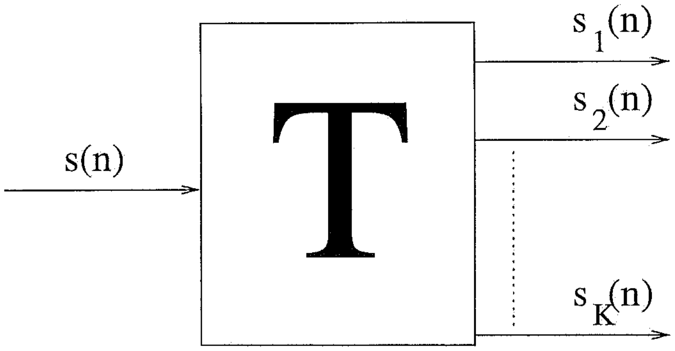


Figure 5: Decomposition of host signal S into K independent channels, using the Karhunen-Loève transform T.

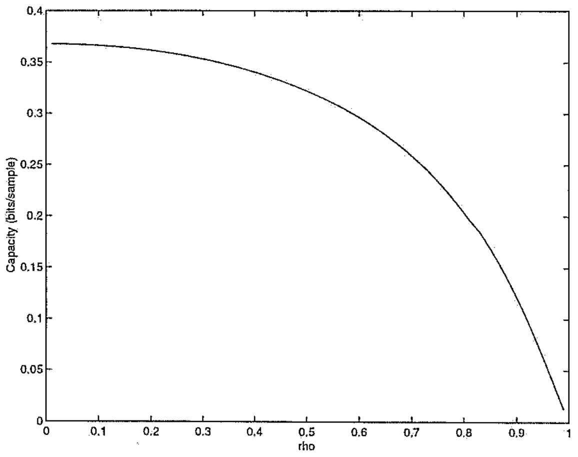
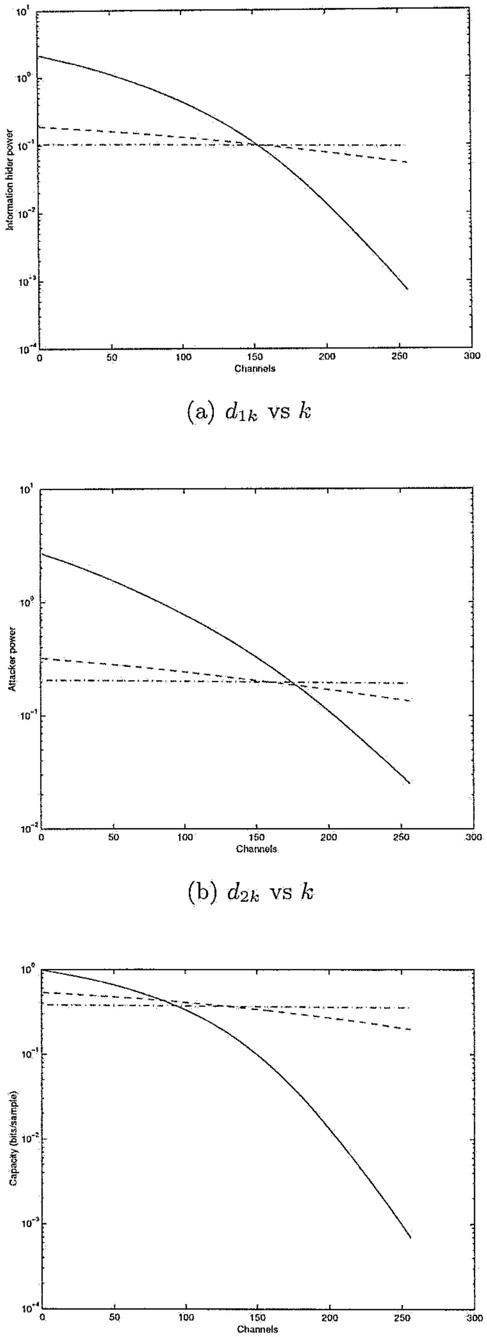


Figure 6: Watermarking capacity for AR(1) processes with unit variance, correlation ρ, and distortion levels $D_1 = 0.1$ and $D_2 = 0.2$.



(a) d_{1k} vs k

(b) d_{2k} vs k

(c) C_k vs k

Figure 7: Three AR(1) process with unit variance, correlations $\rho = 0.05$, 0.5 and 0.95, and distortion levels $D_1 = 0.1$ and $D_2 = 0.2$. (a), (b) Optimal power allocations for the information hider and the attacker; (c) resulting capacities as a function of frequency.

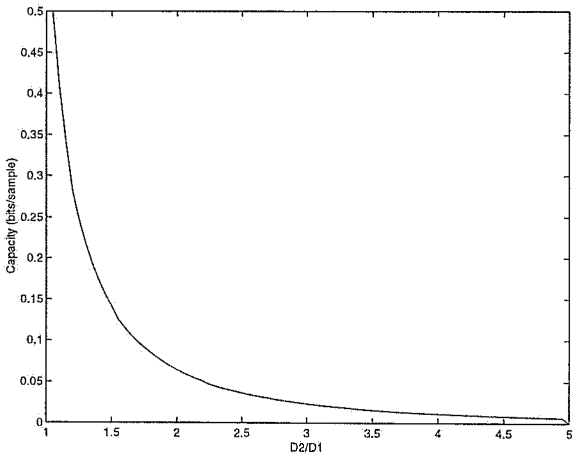


Figure 8: Watermarking capacity for AR(1) processes with unit variance, correlation $\rho = 0.95$, $D_1 = 0.1$, and D_2 ranging from D_1 to $5D_1$.