The Gaussian Watermarking Game*

Aaron S. Cohen† Amos Lapidoth‡

Dedicated to the memory of Aaron D. Wyner

אהרן דניאל זכרו לברכה

with gratitude for his mentoring and hospitality. A.L.

Abstract

Watermarking models a copyright protection mechanism where an original source sequence or "covertext" is modified before distribution to the public in order to embed some extra information. The embedding should be transparent (i.e., the modified data sequence or "stegotext" should be similar to the covertext) and robust (i.e., the extra information should be recoverable even if the stegotext is modified further, possibly by a malicious "attacker").

We compute the coding capacity of the watermarking game for a Gaussian covertext and squared-error distortions. Both the public version of the game (covertext known to neither attacker nor decoder) and the private version of the game (covertext unknown to attacker but known to decoder) are treated. While the capacity of the former cannot, of course, exceed the capacity of the latter, we show that the two are, in fact, identical. These capacities depend critically on whether the distortion constraints are required to be met in expectation or with probability one. In the former case the coding capacity is zero, whereas in the latter it coincides with the value of related zero-sum dynamic mutual informations games of complete and perfect information.

*Parts of this work were presented at the 2000 Conference on Information Sciences and Systems (CISS '00), Princeton University, Princeton, NJ, March 15–17, 2000, and at the 2000 IEEE International Symposium on Information Theory (ISIT '00), Sorrento, Italy, June 25–30, 2000.

†A. Cohen was with the Laboratory for Information and Decision Systems (LIDS) at the department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT). He is now with the Research Laboratory for Electronics at MIT, Cambridge, MA 02139 (email: acohen@mit.edu). His work was supported in part by an NSF Graduate Fellowship.

‡A. Lapidoth was also with LIDS. He is now with the Swiss Federal Institute of Technology, ETH Zentrum, CH-8092, Zurich, Switzerland (email: lapidoth@isi.ee.ethz.ch). The work of A. Lapidoth was supported in part by the NSF Faculty Early Career Development (CAREER) program.

We also compute the capacity when the attacker is restricted to additive attacks. This capacity turns out to be strictly larger than the watermarking capacity, thus demonstrating that additive attacks are sub-optimal. In fact, under the additive attack restriction, capacity turns out to coincide with the capacity of Costa's model for "writing on dirty paper", thus demonstrating that in Costa's model the i.i.d. Gaussian "noise" is the most malevolent power-limited "noise". Additionally, Costa's observation that in the presence of i.i.d. Gaussian "noise", an i.i.d. Gaussian "dirt" process that is non-causally known to the transmitter (but not receiver) does not reduce capacity, is extended here to general ergodic "dirt" and to stationary (but not necessarily white) Gaussian "noise".

Contrary to the average power limited jamming game, where the analysis of a modified (saddle-point achieving) white Gaussian jammer suffices to prove the converse, the watermarking game does not have a memoryless saddle-point, and our proof of the converse requires the analysis of an attacker that depends on the entire stegotext sequence (but not otherwise on the encoder). This dependence must be carefully controlled to guarantee that the choice of the attacking strategy will asymptotically not reveal any information about the embedded message.

The proof of the converse exploits only the ergodicity and the second-order properties of the covertext, thus allowing for the characterization of the memoryless Gaussian covertext as the covertext that has the highest watermarking capacity among all finite fourth-moment ergodic covertexts of a given second moment.

1 Introduction

The watermarking game can model a situation where a source sequence ("covertext") needs to be copyright-protected before it is distributed to the public. The copyright ("message") needs to be embedded in the distributed version ("stegotext") so that no "attacker" with access to the stegotext will be able produce a "forgery" that resembles the covertext and yet does not contain the embedded message. The watermarking process ("encoding") should, of course, introduce limited distortion so as to guarantee that the stegotext closely resembles the original covertext.

In the public version of the game we require that any party with access to a valid forgery (i.e., a forgery that introduces limited distortion) should be able to decode the message with a small probability of error. In the private version of the game the decoding is only required of parties with access to both the forgery and the original covertext.

The different messages may correspond to different possible owners of the covertext or to other relevant data, and it is thus of interest to study the number of distinct messages that can be embedded in the text, if each message is to be reliably decoded from any valid forgery. The highest exponential rate at which this number can grow in relation to the covertext size is defined as the coding capacity of the game.

In this paper we focus on memoryless Gaussian sources when the distortions introduced by

the encoder and by the attacker are measured using Euclidean distances. For such scenarios we compute the coding capacity of both the private and the public versions of the game.

The precise nature of the distortion constraints imposed on the encoder and the attacker greatly influences the resulting coding capacities. We focus on average distortion constraints and on almost-sure distortion constraints. We show that the former constraints typically lead to zero coding capacities, whereas the latter lead to capacities that are equal to the values of related mutual information games. These mutual information games also motivate optimal encoding and attacking strategies.

Some of the sources that need to be watermarked cannot be modeled as memoryless Gaussians. For such sources we show that the memoryless Gaussian model is optimistic. Thus, we show that among all finite fourth-moment ergodic[1] sources of a given second moment, the memoryless Gaussian covertext is the easiest to watermark (yielding the highest coding capacity). Intuitively, this follows since the encoder utilizes the uncertainty of the covertext when transmitting the message, and a Gaussian distribution has the most uncertainty (i.e., highest entropy) among all distributions of a given second moment. See the discussion before Theorem 2.1 for an example of a covertext distribution with a coding capacity strictly smaller than a Gaussian covertext distribution with the same second order statistics. See [1] for an analysis of Gaussian sources with memory.

Knowing the covertext at the decoder cannot hurt, because such information can always be ignored. Consequently, the coding capacity of the private version of the game cannot be lower than the coding capacity of the public version. For memoryless Gaussian covertext and Euclidean distance distortions, however, we show that the two capacities are identical. Thus, while the decoder's knowledge of the covertext may help to reduce the complexity of the watermarking process, it does not increase the watermarking capacity.

To guarantee that no rate above capacity is achievable, the attacker must be familiar with the details of the encoder (excluding, of course, the realization of the secret key). However, as we shall see, it need not know the structure of the decoder. Thus, the converse would continue to hold even if the decoder (but not encoder) were cognizant of the attacking strategy. Combined with the achievability theorems in which the decoder does not know the attacking strategy, this observation demonstrates that the ignorance of the decoder of the attack rule is irrelevant to the value of the game. The fact that the attacker may depend on the encoder, however, is critical.

[1] We shall use the term "ergodic" to also imply stationarity.

We also compute the coding capacity for a variant of the watermarking game — "the additive attack watermarking game" — where the attacker is restricted to be purely additive. This capacity turns out to be strictly larger than the watermarking capacity, thus demonstrating that additive attacks are sub-optimal. In fact, under the additive attack restriction, the capacity turns out to coincide with the capacity of Costa's model for "writing on dirty paper" [2], thus demonstrating that in Costa's model the independent and identically distributed (i.i.d.) Gaussian "noise" is the most malevolent power-limited "noise". We revisit Costa's result — that in the presence of i.i.d. Gaussian "noise", an i.i.d. Gaussian "dirt" process that is non-causally known to the transmitter (but not receiver) does not reduce capacity — and extend it to general ergodic "dirt" and to stationary (but not necessarily white) Gaussian "noise".

We finally consider two related mutual information games whose solutions provide motivation for both the coding schemes and the converse strategy used in the watermarking game. It should be noted, however, that the mutual information games in themselves do not suffice to prove the coding theorems or the converses. For example, the mutual information games do not address the difference between almost-sure and average distortion constraints — a distinction that greatly influences capacity.

Watermarking has attracted interest in recent years due to the ease by which data can now be reproduced and transmitted around the world, for example see [3, 4, 5, 6] and references therein. The information theoretic model of the watermarking game was introduced by O'Sullivan, Moulin and Ettinger [7]. They formulated private watermarking as a max-min game over conditional mutual information, and extended their approach in [8, 9]. For similar models (but with somewhat different distortion constraints) error exponents were studied in [10, 11] and identification capacities in [12]. Information rates were investigated in [13] for Gaussian covertexts and for the fixed (and as we shall see, sub-optimal) independent additive attack strategy. In [14], a coding strategy was introduced (distortion-compensated quantization index modulation or "DC-QIM"), which was shown to achieve capacity for several scenarios when the decoder knows the statistics of the attack channel. The capacity region of a joint watermarking and quantization technique was investigated in [15].

In almost all these studies the decoder is assumed to be cognizant of the attack strategy, thus allowing for maximum-likelihood decoding; the notable exception is [11]. This assumption is implicit in studies where some attack strategy is hypothesized (and not optimized), and is more explicit

in those studies that optimize over the attack strategy. In the present paper we shall avoid this assumption. Thus, we shall require that the encoder and decoder be designed so that the desired level of performance can be met regardless of the actual attacker used. In fact, we shall prove the coding theorems assuming that the decoder is ignorant of the attack strategy, and prove the converses under maximum-likelihood decoding conditions. The encoding will always be performed in ignorance of the attacking strategy.

After concluding this section with some notes on notation, we turn in Section 2 to formalize the models and to present our main results on their coding capacities. The remainder of the paper is devoted to proving and discussing these results. In Section 3, we present the solutions to the mutual information games, which provide motivation for the encoding and attacker strategies discussed later. In Section 4, we prove the achievability parts of the main theorems on the coding capacity of the watermarking game. That is, we describe coding strategies and then demonstrate that for the appropriate rates and any attack strategy, the probability of error tends to zero. In Section 5, we propose an attack strategy that proves that no ergodic source of finite fourth moment and of second moment σ_u^2 can be reliably watermarked — publicly or privately — at all appropriate rates (i.e., any rate larger than our proposed capacity). In Section 6, we show that no positive rate is achievable when average distortion constraints are imposed rather than the almost-sure constraints. Finally, in Section 7, we give some concluding remarks.

1.1 Notation and Definitions

All the alphabets used in this paper are the real line, but for clarity we denote them by separate letters $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$ for the covertext, stegotext, and forgery, respectively, which we define below. The n-th Cartesian products of these sets (e.g., $\mathcal{U}\times\mathcal{U}\times\cdots\times\mathcal{U}$) are written $\mathcal{U}^n$, $\mathcal{X}^n$, and $\mathcal{Y}^n$, respectively. Random variables and random vectors are written in upper case, while their realizations are written in lower case. The use of bold refers to a vector of length n, for example $\boldsymbol{U} = (U_1, \ldots, U_n)$ (random) or $\boldsymbol{u} = (u_1, \ldots, u_n)$ (deterministic).

We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and inner product, respectively. That is, for any $\boldsymbol{\mu}, \boldsymbol{\psi} \in \mathbb{R}^n$, $\langle\boldsymbol{\mu}, \boldsymbol{\psi}\rangle = \sum_{i=1}^n \mu_i\psi_i$, and $\|\boldsymbol{\mu}\| = \sqrt{\langle\boldsymbol{\mu}, \boldsymbol{\mu}\rangle}$. If $\langle\boldsymbol{\mu}, \boldsymbol{\psi}\rangle = 0$, then we say that $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are orthogonal and write $\boldsymbol{\mu} \perp \boldsymbol{\psi}$. We denote by $\boldsymbol{\psi}^\perp$ the linear sub-space of all vectors that are

orthogonal to $\boldsymbol{\psi}$. If $\boldsymbol{\psi} \neq 0$, then $\boldsymbol{\mu}|_{\boldsymbol{\psi}}$ denotes the projection of $\boldsymbol{\mu}$ onto $\boldsymbol{\psi}$, i.e.,

$$\boldsymbol{\mu}|_{\boldsymbol{\psi}} = \frac{\langle \boldsymbol{\mu}, \boldsymbol{\psi} \rangle}{\|\boldsymbol{\psi}\|^2} \boldsymbol{\psi}.$$

Similarly, $\boldsymbol{\mu}|_{\boldsymbol{\psi}^\perp}$ denotes the projection of $\boldsymbol{\mu}$ onto the subspace orthogonal to $\boldsymbol{\psi}$, i.e., $\boldsymbol{\mu}|_{\boldsymbol{\psi}^\perp} = \boldsymbol{\mu} - \boldsymbol{\mu}|_{\boldsymbol{\psi}}$.

We use P to denote a generic probability measure on the appropriate Borel σ-algebra. For example, $P_{\boldsymbol{U}}(\cdot)$ is the distribution of $\boldsymbol{U}$ on the Borel σ-algebra of subsets of $\mathcal{U}^n$. Similarly, $P_{\boldsymbol{X}|\boldsymbol{U}}$ denotes the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{U}$, and $f_{\boldsymbol{X}|\boldsymbol{U}}(\boldsymbol{x}|\boldsymbol{u})$ denotes the conditional density, when it exists.

Finally, we shall use the following definitions throughout the paper in order to describe both capacity expressions and optimal strategies. Let us first define the interval

$$\mathcal{A}(D_1, D_2, \sigma_u^2) = \left\{ A : \max\left\{ D_2, \left(\sigma_u - \sqrt{D_1}\right)^2 \right\} \leq A \leq \left(\sigma_u + \sqrt{D_1}\right)^2 \right\}, \tag{1}$$

where $\mathcal{A}(D_1, D_2, \sigma_u^2)$ is empty if $D_2 \geq \sigma_u^2 + D_1 + 2\sigma_u\sqrt{D_1}$. Let us also define the mappings

$$\rho(A; D_1, \sigma_u^2) = \frac{1}{2}(A - \sigma_u^2 - D_1), \tag{2}$$

$$b_1(A; D_1, \sigma_u^2) = 1 + \frac{\rho(A; D_1, \sigma_u^2)}{\sigma_u^2}, \tag{3}$$

$$b_2(A; D_1, \sigma_u^2) = D_1 - \frac{\rho^2(A; D_1, \sigma_u^2)}{\sigma_u^2}, \tag{4}$$

$$c(A; D_2) = 1 - \frac{D_2}{A}, \tag{5}$$

$$\alpha(A; D_1, D_2, \sigma_u^2) = 1 - \frac{b_1(A; D_1, \sigma_u^2) D_2}{D_2 + c(A; D_2) b_2(A; D_1, \sigma_u^2)}, \tag{6}$$

$$s(A; D_1, D_2, \sigma_u^2) = \frac{b_2(A; D_1, \sigma_u^2) c(A; D_2)}{D_2}, \tag{7}$$

and[2]

$$C^*(D_1, D_2, \sigma_u^2) = \begin{cases} \max_{A \in \mathcal{A}(D_1, D_2, \sigma_u^2)} \frac{1}{2} \log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr) & \text{if } \mathcal{A}(D_1, D_2, \sigma_u^2) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

[2]Unless otherwise specified, all logarithms in this paper are base-2 logarithms.

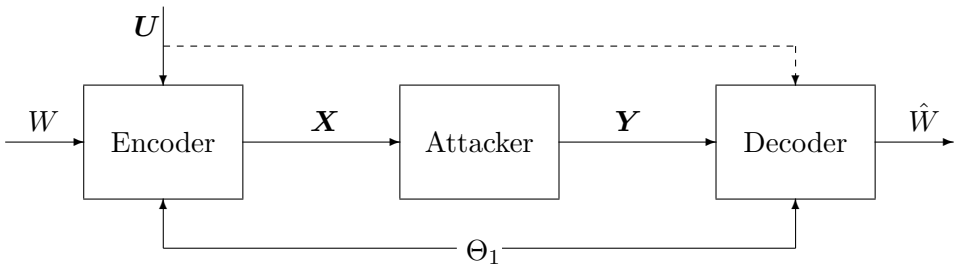

Figure 1: A diagram of the watermarking game. The dashed line is used in the private version of the game, but not in the public version.

We shall see that $C^*(D_1, D_2, \sigma_u^2)$ is the capacity for the Gaussian watermarking game (where the precise definitions of capacity and the watermarking game are given below in Section 2.1.1). Note that a closed-form solution for (8) can be found by setting the derivative with respect to A to zero. This yields a cubic equation in A that can be solved analytically. Further note that $C^*(D_1, D_2, \sigma_u^2)$ is zero only if $D_2 \geq \sigma_u^2 + D_1 + 2\sigma_u\sqrt{D_1}$.

2 Main Results

2.1 The Watermarking Game

2.1.1 The Game

The watermarking game is illustrated in Figure 1 and can be described as follows. Prior to the use of the watermarking system, a *secret key*[3] (random variable) Θ_1 is generated and revealed to the *encoder* and *decoder*. Independently of the secret key Θ_1, a source subsequently emits a blocklength-n *covertext* sequence $\boldsymbol{U} \in \mathcal{U}^n$ according to the law $P_{\boldsymbol{U}}$, where $\{P_{\boldsymbol{U}}\}$ is a collection of probability laws indexed by the blocklength n. We will be mostly interested in the case where $\boldsymbol{U}$ is a sequence of independent and identically distributed (i.i.d.) random variables of law P_U^G, where P_U^G denotes the Gaussian distribution of zero mean and variance $\sigma_u^2 > 0$. Independently of the covertext $\boldsymbol{U}$ and of the secret key Θ_1, a copyright *message* W is drawn uniformly over the set $\mathcal{W}_n = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$, where R is the *rate* of the system.

[3]We do not limit the amount of randomness provided by the secret key, but it must be independent of the message and the covertext.

Using the secret key, the encoder maps the covertext and message to the *stegotext* $\boldsymbol{X}$. For every blocklength n, the encoder thus consists of a measurable function f_n that maps realizations of the covertext $\boldsymbol{u}$, the message w, and the secret key θ_1 into the set $\mathcal{X}^n$, i.e.,

$$f_n : (\boldsymbol{u}, w, \theta_1) \mapsto \boldsymbol{x} \in \mathcal{X}^n.$$

The random vector $\boldsymbol{X}$ is the result of applying the encoder to the covertext $\boldsymbol{U}$, the message W, and the secret key Θ_1, i.e., $\boldsymbol{X} = f_n(\boldsymbol{U}, W, \Theta_1)$. The distortion introduced by the encoder is measured by

$$d_1(\boldsymbol{u}, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} d_1(u_i, x_i),$$

where $d_1 : \mathcal{U} \times \mathcal{X} \to \mathbb{R}^+$ is a given non-negative function. We assume throughout that $d_1(u, x) = (x - u)^2$. We require that the encoder satisfy

$$d_1(\boldsymbol{U}, \boldsymbol{X}) \leq D_1, \text{ a.s.}, \tag{9}$$

where $D_1 > 0$ is a given constant called the *encoder distortion level*, and a.s. stands for "almost surely". We will also consider an average distortion constraint on the encoder, i.e.,

$$E\left[d_1(\boldsymbol{U}, \boldsymbol{X})\right] \leq D_1, \tag{10}$$

where the expectation is with respect to the covertext, the message, and the secret key; a similar average distortion constraint was considered in [7, 9]. Still other types of constraints have been considered, e.g., $E\left[d_1(\boldsymbol{U}, \boldsymbol{X}) | \boldsymbol{U} = \boldsymbol{u}\right] \leq D_1$ for all $\boldsymbol{u}$ [10] and $\Pr\left\{d_1(\boldsymbol{U}, \boldsymbol{X}) > D_1 | \boldsymbol{U} = \boldsymbol{u}\right\} \leq \exp(-\nu n)$ for some ν and all $\boldsymbol{u}$ [11]. The latter constraint reduces to (9) for $\nu = \infty$. Unless otherwise stated, we shall focus on the a.s. distortion constraint (9). We feel that this constraint best represents the specification that every stegotext produced by the encoder should be within distortion D_1 of the covertext.

Independently of the covertext $\boldsymbol{U}$, the message W, and the secret key Θ_1 the *attacker* generates an *attack key* (random variable) Θ_2. For every $n > 0$, the attacker consists of a measurable function g_n that maps realizations of the stegotext $\boldsymbol{x}$ and the attack key θ_2 into the set $\mathcal{Y}^n$, i.e.,

$$g_n : (\boldsymbol{x}, \theta_2) \mapsto \boldsymbol{y} \in \mathcal{Y}^n. \tag{11}$$

The *forgery* $\boldsymbol{Y}$ is a random vector that is the result of applying the attacker to the stegotext $\boldsymbol{X}$ and the attacker's source of randomness Θ_2, i.e., $\boldsymbol{Y} = g_n(\boldsymbol{X}, \Theta_2)$. In other studies of watermarking, e.g., [7, 10], the attacker is formalized equivalently as a conditional distribution of the forgery given the stegotext. Here, we use a deterministic mapping to emphasize that, similarly to the encoder, the attacker directly produces the forgery from the stegotext, with some randomness used to assist. In fact, we argue in Section 4.1.2 that any rate achievable against a completely deterministic attacker is also achievable against a randomized attacker. The distortion introduced by the attacker is measured by

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} d_2(x_i, y_i),$$

where $d_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is a given non-negative function. We assume throughout that $d_2(x, y) = (y - x)^2$. The attacker is required to satisfy

$$d_2(\boldsymbol{X}, \boldsymbol{Y}) \leq D_2, \text{ a.s.}, \tag{12}$$

where $D_2 > 0$ is a given constant called the *attacker distortion level*. We will also consider an average distortion constraint on the attacker, i.e.,

$$E\left[d_2(\boldsymbol{X}, \boldsymbol{Y})\right] \leq D_2, \tag{13}$$

where expectation is over the covertext $\boldsymbol{U}$, the message W, the secret key Θ_1, and the attack key Θ_2. Again, unless otherwise stated, we shall focus on the a.s. distortion constraint (12).

In the public version of the watermarking game, the decoder attempts to recover the copyright message based only on realizations of the secret key θ_1 and the forgery $\boldsymbol{y}$. In this version the decoder is a measurable mapping

$$\phi_n : (\boldsymbol{y}, \theta_1) \mapsto \hat{w} \in \mathcal{W}_n \text{ (public version)}.$$

In the private version, however, the decoder also has access to the covertext. In this case the decoder is a measurable mapping

$$\phi_n : (\boldsymbol{y}, \boldsymbol{u}, \theta_1) \mapsto \hat{w} \in \mathcal{W}_n \text{ (private version)}.$$

The *estimate of the message* $\hat{W}$ is a random variable that is the result of applying the decoder to the forgery $\boldsymbol{Y}$, the covertext $\boldsymbol{U}$ (in the private version), and the same source of randomness used by the encoder Θ_1. That is, $\hat{W} = \phi_n(\boldsymbol{Y}, \boldsymbol{U}, \Theta_1)$ in the private version, and $\hat{W} = \phi_n(\boldsymbol{Y}, \Theta_1)$ in the public version.

The realizations of the covertext $\boldsymbol{u}$, message w, and sources of randomness (θ_1, θ_2) determine whether the decoder errs in decoding the copyright message, i.e., if the estimate of the message $\hat{w}$ differs from the original message w. We write this error indicator function (for the private version) as

$$e(\boldsymbol{u}, w, \theta_1, \theta_2, f_n, g_n, \phi_n) = \begin{cases} 1 & \text{if } w \neq \phi_n\Big(g_n\big(f_n(\boldsymbol{u}, w, \theta_1), \theta_2\big), \mathbf{u}, \theta_1\Big) \\ 0 & \text{otherwise} \end{cases},$$

where the expression for the public version is the same, except that the decoder mapping ϕ_n does not take the covertext $\boldsymbol{u}$ as an argument. We consider the probability of error averaged over the covertext, message and both sources of randomness as a functional of the mappings f_n, g_n, and ϕ_n. This is written as

$$\bar{P}_e(f_n, g_n, \phi_n) = E_{\boldsymbol{U}, W, \Theta_1, \Theta_2}[e(\boldsymbol{U}, W, \Theta_1, \Theta_2, f_n, g_n, \phi_n)] = \Pr\big(\hat{W} \neq W\big),$$

where the subscripts on the right hand side (RHS) of the first equality indicate that the expectation is taken with respect to the four random variables $\boldsymbol{U}$, W, Θ_1, and Θ_2.

We adopt a conservative approach to the watermarking game and assume that once the watermarking system is employed, its details — namely the encoder mapping f_n, the distributions (but not realizations) of the covertext $\boldsymbol{U}$ and of the secret key Θ_1, and the decoder mapping ϕ_n — are made public. The attacker can be malevolently designed accordingly. The watermarking game is thus played so that the encoder and decoder are designed prior to the design of the attacker. This, for example, precludes the decoder from using the maximum-likelihood decoding rule which requires knowledge of the law $P_{\boldsymbol{Y}|W}$ and thus, indirectly, knowledge of the attack strategy. We thus say that a rate R is *achievable* if there exists a sequence $\{(f_n, \phi_n)\}$ of allowable rate-R encoder and decoder pairs such that for any sequence $\{g_n\}$ of allowable attackers the average probability of error $\bar{P}_e(f_n, g_n, \phi_n)$ tends to zero as n tends to infinity.

2.1.2 The Coding Capacity of the Game

The *coding capacity* of the game is the supremum of all achievable rates. It depends on three parameters: the encoder distortion level D_1, the attacker distortion level D_2, and the covertext distribution $\{P_{\boldsymbol{U}}\}$. We thus write the coding capacity as $C_{\text{priv}}(D_1, D_2, \{P_{\boldsymbol{U}}\})$ and $C_{\text{pub}}(D_1, D_2, \{P_{\boldsymbol{U}}\})$ for the private and public version, respectively.

The following theorem demonstrates that if the covertext has power σ_u^2, then the coding capacity of the private and public watermarking games cannot exceed $C^*(D_1, D_2, \sigma_u^2)$. Furthermore, if the covertext $\boldsymbol{U}$ is an i.i.d. zero-mean Gaussian sequence with power σ_u^2, then the coding capacities of the private and public versions are equal, and they coincide with this upper bound.

We see from this theorem that, as in the "writing on dirty paper" model (see Section 2.1.3 below and [2]), the capacity of the Gaussian watermarking game is unaffected by the presence or absence of side-information (covertext) at the receiver. See [16] for some comments on the role of receiver side-information, particularly in card games.

This theorem also shows that, of all ergodic covertexts with a given power, the i.i.d. zero-mean Gaussian covertext has the largest watermarking capacity. Although the covertext can be thought of as additive noise in a communication with side information situation (see Section 2.1.3), this result differs from usual "Gaussian is the worst-case additive noise" idea, see e.g., [17, 18]. The reason that a Gaussian covertext is the best case (i.e., easiest to watermark) is that the encoder is able to transmit the watermark using the uncertainty of the covertext, and a Gaussian distribution has the most uncertainty (i.e., highest entropy) among all distributions of a given second moment.

As an example of this extremal property of the Gaussian distribution, consider an i.i.d. covertext in which each sample U_k takes on the values $\pm\sigma_u$ equiprobably, so that $E\left[U_k^2\right] = \sigma_u^2$. If $D_1 = D_2 \ll \sigma_u^2$, then $C^*(D_1, D_2, \sigma_u^2) \approx 1/2$ bits/symbol, but a watermarking system could not reliably transmit at nearly this rate with this covertext. To see this, consider further an attacker that creates the forgery by quantizing each stegotext sample X_k to the nearest of $\pm\sigma_u$. Even in the private version, the encoder can only send information by changing U_k by at least σ_u, which can be done for only a small percentage of the samples since $D_1 \ll \sigma_u^2$. Indeed, it can be shown (see e.g., [19, 20]) that the largest achievable rate for this fixed attacker is[4] $H_b\left(D_1/\sigma_u^2\right)$ bits/symbol, which is smaller than $1/2$ bits/symbol for $D_1/\sigma_u^2 < 0.11$. The capacity for this scenario could be even smaller since we

[4]We use $H_b(\cdot)$ to denote the binary entropy, i.e., $H_b(p) = -p\log p - (1-p)\log(1-p)$.

have only considered a known attacker.

Theorem 2.1. *If $\{P_{\boldsymbol{U}}\}$ defines an ergodic covertext $\boldsymbol{U}$ such that*

$$E\left[U_k^4\right] < \infty, \tag{14}$$

and

$$E\left[U_k^2\right] \leq \sigma_u^2, \tag{15}$$

then

$$\begin{aligned} C_{\text{pub}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C_{\text{priv}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) && (16)\\ &\leq C^*(D_1, D_2, \sigma_u^2). && (17) \end{aligned}$$

Equality is achieved in both (16) and (17) if $\boldsymbol{U}$ is an i.i.d. Gaussian sequence with mean zero and variance σ_u^2, i.e. if $P_{\boldsymbol{U}} = (P_U^G)^n$ for all n.

In Section 4, we prove the achievability result of this theorem for the public version, assuming an i.i.d. Gaussian covertext; we also outline the simpler coding strategy for the private version. We prove the converse result for general covertexts in Section 5. We now briefly describe the optimal strategies.

The optimal encoder for the public version with a Gaussian covertext uses random binning [21, 22, 23, 2] and can be described as follows. A value of $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$ is chosen and a codebook of 2^{nR} bins is generated with 2^{nR_0} codewords in each bin, where R and R_0 depend on A. Given the covertext $\boldsymbol{u}$ and the message w, the encoder forms the stegotext as $\boldsymbol{x} = \boldsymbol{v} + (1-\alpha)\boldsymbol{u}$ where $\boldsymbol{v}$ is a codeword from bin w chosen so that $n^{-1}\langle \boldsymbol{x} - \boldsymbol{u}, \boldsymbol{u}\rangle \approx \rho$. Our choice of α and ρ will ensure that the distortion constraint is met and that the norm of the stegotext is $n^{-1}\|\boldsymbol{x}\|^2 \approx A$. Given the forgery, the decoder finds the codeword (out of all $2^{n(R+R_0)}$ codewords) that is closest to the forgery and estimates the message as the bin containing this codeword. We will show that the target correlation will be met if R_0 is large enough and that the correct message will be recovered if $R + R_0$ is small enough. Combining these bounds will show that all rates $R < \frac{1}{2}\log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr)$ are achievable, which, by (8), will complete the achievability proof.

In order to guarantee that no rates larger than $C^*(D_1, D_2, \sigma_u^2)$ are achievable for any covertext satisfying (14) and (15), we consider an attacker that creates the forgery by attenuating the stego-

text by $c(\hat{A}; D_2)$ of (5) and adding independent Gaussian noise of variance $c(\hat{A}; D_2) \cdot D_2$, where $\hat{A}$ is a quantization of $n^{-1}\|\boldsymbol{x}\|^2$. This attacker is related to the Gaussian rate distortion forward channel, which for a variance-A Gaussian random variable and allowable distortion D_2 also multiplies by $c(A; D_2)$ and adds noise of variance $c(A; D_2) \cdot D_2$. An optimal encoder with an i.i.d. Gaussian covertext produces a stegotext that is approximately Gaussian, and thus, in this case, the optimal attacker is essentially performing optimal lossy compression. Note that optimal lossy compression is not the optimal attack for all covertexts. In fact, it is not even optimal for Gaussian covertexts, if they exhibit memory [1].

The next theorem, which is proved in Section 6, addresses the case where average distortion constraints (10), (13) rather than a.s. distortion constraints (9), (12) are in effect. In this case no positive rates are achievable.

Theorem 2.2. *If the covertext $\boldsymbol{U}$ satisfies*

$$\liminf_{n\to\infty} E\left[\frac{1}{n}\|\boldsymbol{U}\|^2\right] < \infty,$$

and if the average distortion constraints (10), (13) are in effect instead of the a.s. distortion constraints (9), (12), then no rate is achievable in either version of the game.

This result is reminiscent of results from the theory of Gaussian arbitrarily varying channels (AVCs) [24] and from the theory of general AVCs with constrained inputs and states [25], where under average power constraints no positive rates are achievable.

2.1.3 The Watermarking Game and the Jamming Game

By writing the forgery $\boldsymbol{Y}$ in the form

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}} + \boldsymbol{U} + \tilde{\boldsymbol{Y}},$$

where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{U}$, and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}$, we can interpret the watermarking game as a non-causal communication game with side information, see e.g. [26]. To this end we may think of $\tilde{\boldsymbol{X}}$ as a transmitted signal that is corrupted by an additive noise $\boldsymbol{U}$, which is non-causally known to the transmitter, and by an additive jammer signal $\tilde{\boldsymbol{Y}}$, which may depend non-causally on $\tilde{\boldsymbol{X}} + \boldsymbol{U}$.

If the transparency measure $d_1(u, x)$ is a difference measure, then the transparency constraint

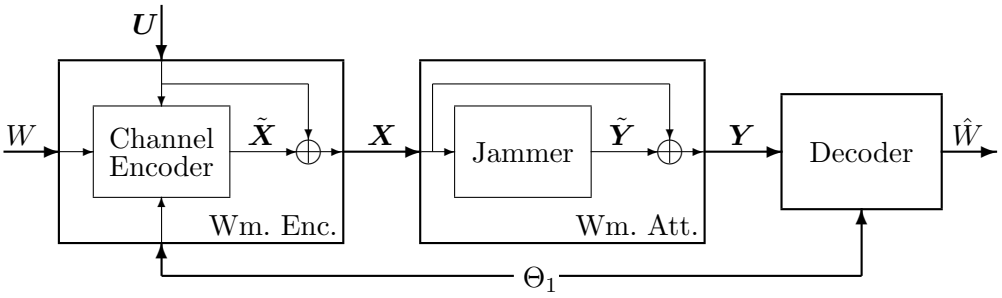

Figure 2: An alternative diagram of the public version of the watermarking game.

translates to a constraint on $\tilde{\boldsymbol{X}}$. In particular, in our case where $d_1(u,x) = (x-u)^2$, the transparency constraint translates to a power constraint on the transmitted signal $\tilde{\boldsymbol{X}}$. Similarly if $d_2(x,y) = (x-y)^2$ then the jamming signal $\tilde{\boldsymbol{Y}}$ becomes power limited. See figure 2 for a block diagram of this interpretation of the game.

The Gaussian watermarking game with $d_1(u,x) = (x-u)^2$ and $d_2(x,y) = (y-x)^2$ is thus reminiscent of Costa's "writing on dirty paper" model [2], which corresponds to the case where, as in the Gaussian watermarking game, the sequence $\boldsymbol{U}$ is memoryless and Gaussian, but where *the attacker ignores* $\tilde{\boldsymbol{X}} + \boldsymbol{U}$ *and fixes* $\tilde{\boldsymbol{Y}}$ *to be a sequence of i.i.d. Gaussian random variables, independent of* $\tilde{\boldsymbol{X}} + \boldsymbol{U}$.

In Costa's model, if the sequence $\boldsymbol{U}$ is known at the receiver then it can be subtracted from the received signal, and in this case we can conclude that $\boldsymbol{U}$ has no adverse effect on capacity. (Additionally, this simple strategy results in the highest capacity in this scenario.) As in the watermarking game, Costa has shown for his model that the capacity does not decrease when $\boldsymbol{U}$ is unknown to the decoder. In either case, this capacity is given by $\frac{1}{2}\log(1+\frac{D_1}{D_2})$. We extend this result to non-Gaussian attackers in Section 2.2 and to non-Gaussian covertexts in Section 2.4.

2.2 The Additive Attack Watermarking Game

In this section, we describe a variation of the watermarking game, which we call the *additive attack watermarking game*. (When it is necessary to distinguish the two models, we will refer to the original model of Section 2.1 as the general watermarking game.) The study of this model will show that it is sub-optimal for the attacker of Figure 2 to produce the jamming sequence $\tilde{\boldsymbol{Y}}$ independently of the stegotext $\boldsymbol{X}$. Also, similarly to Costa's writing on dirty paper result, we will show that if the

covertext $\boldsymbol{U}$ is i.i.d. Gaussian then the capacities of the private and public versions are the same and are given by $\frac{1}{2}\log(1+\frac{D_1}{D_2})$. This result can be thus viewed as an extension of Costa's result to arbitrarily varying noises.

In the additive attack watermarking game the attacker is more restricted than in the general game. Rather than allowing general attacks of the form (11), we restrict the attacker to mappings that are of the form

$$g_n(\boldsymbol{x}, \theta_2) = \boldsymbol{x} + \tilde{g}_n(\theta_2) \tag{18}$$

for some mapping $\tilde{g}_n$. In particular, the jamming sequence $\tilde{\boldsymbol{Y}} = \tilde{g}_n(\Theta_2)$ is produced independently of the stegotext $\boldsymbol{X}$, and must satisfy the distortion constraint

$$\frac{1}{n}\left\|\tilde{\boldsymbol{Y}}\right\|^2 \leq D_2, \text{ a.s..} \tag{19}$$

The capacity of the additive attack watermarking game is defined similarly to the capacity of the general game and is written as $C^{\mathrm{AA}}_{\mathrm{priv}}(D_1, D_2, \{P_{\boldsymbol{U}}\})$ and $C^{\mathrm{AA}}_{\mathrm{pub}}(D_1, D_2, \{P_{\boldsymbol{U}}\})$ for the private and public versions, respectively. Our main result in this section is to describe these capacities.

Theorem 2.3. *For any covertext distribution* $\{P_{\boldsymbol{U}}\}$*,*

$$\begin{aligned} C^{\mathrm{AA}}_{\mathrm{pub}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C^{\mathrm{AA}}_{\mathrm{priv}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) && (20)\\ &= \frac{1}{2}\log\left(1+\frac{D_1}{D_2}\right). && (21) \end{aligned}$$

Equality is achieved in (20) if $\boldsymbol{U}$ *is an i.i.d. Gaussian sequence.*

As with Theorem 2.1, we prove the achievability and converse parts of this theorem in Sections 4 and 5, respectively. Since any allowable additive attacker is also an allowable general attacker, the capacity of the additive attack watermarking game provides an upper bound to the capacity of the general watermarking game. However, comparing Theorems 2.1 and 2.3, we see that for an i.i.d. Gaussian covertext this bound is loose. Thus, for such covertexts, it is sub-optimal for the attacker in the general watermarking game to take the form (18).

When the covertext $\boldsymbol{U}$ is i.i.d. Gaussian, then the additive attack watermarking game differs from Costa's writing on dirty paper in only two respects. First, the jamming sequence distribution is arbitrary (subject to (19)) instead of being an i.i.d. Gaussian sequence. Second, the jamming sequence distribution is unknown to the encoder and decoder. Nevertheless, the two models give the

same capacity, thus demonstrating that the most malevolent additive attack for the watermarking game is an i.i.d. Gaussian one.

2.3 Mutual Information Games

In this section, we consider two mutual information games that are motivated by the results on the capacity of channels with states, when the states are known to both transmitter and receiver and when they are known to the transmitter only (i.e., private and public versions). The motivation is discussed in more detail in Section 2.3.1. In Section 2.3.2, we define the games precisely and give our main result on their value.

2.3.1 Motivation: Capacity with Side Information

Let us consider a communication channel of transition probability that depends on a state u. That is, given the value of the current state u and the current input x, the output of the channel is a random variable Y with distribution $P_{Y|X,U}(\cdot|x,u)$, where we assume throughout that $P_{Y|X,U}$ is known. Furthermore, given a state sequence $\boldsymbol{u}$ and an input sequence $\boldsymbol{x}$, the output sequence $\boldsymbol{Y}$ is generated in a memoryless fashion, so that

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{x},\boldsymbol{u}) = \prod_{i=1}^{n} P_{Y|X,U}(y_i|x_i,u_i). \tag{22}$$

Let us assume that the state sequence $\boldsymbol{U}$ is generated in an i.i.d. manner according to a known distribution P_U and let us also (temporarily) assume that the alphabets $\mathcal{U}$, $\mathcal{X}$ and $\mathcal{Y}$ are finite. As in the watermarking game, we are interested in the capacity of this channel when the encoder knows $\boldsymbol{u}$ (non-causally) and the decoder does (private) or does not (public) know $\boldsymbol{u}$. For the private version, Wolfowitz [27] showed that the capacity is given by

$$C_{\text{priv}}^{\text{CSI}} = \max_{P_{X|U}} I(X;Y|U) \tag{23}$$

where the mutual information is defined in the usual manner and is evaluated with respect to the joint distribution $P_{U,X,Y} = P_U P_{X|U} P_{Y|X,U}$. For the public version, Gel'fand and Pinsker [22]

showed the capacity is given by

$$C_{\text{pub}}^{\text{CSI}} = \max_{P_{X,V|U}} I(V;Y) - I(V;U), \tag{24}$$

where V is an auxiliary random variable with alphabet $|\mathcal{V}| \leq |\mathcal{X}| + |\mathcal{U}| - 1$, and where the optimal conditional distribution takes the form

$$P_{X,V|U}(x, v|u) = P_{V|U}(v|u) \cdot 1_{x=f(v,u)} \tag{25}$$

for some $P_{V|U}$ and some function $f : \mathcal{V} \times \mathcal{U} \mapsto \mathcal{X}$.

In the watermarking game, the above channel model corresponds to a fixed memoryless attacker, and the capacity of the watermarking game for such an attacker can be found by modifying (23) and (24) to include a distortion constraint. In the mutual information games we will further generalize these expressions to include a minimization over possible "attack channels." We shall see that the solutions to an instance of these games agree with the capacity of the Gaussian watermarking game and provide insight into how to approach it. Others [7, 8, 11] have shown that general capacity expressions for similar watermarking games are given by related mutual information games.

2.3.2 Definitions and Result

For a general covertext distribution $P_{\boldsymbol{U}}$, conditional law $P_{\boldsymbol{X}|\boldsymbol{U}}$ ("watermarking channel") and conditional law $P_{\boldsymbol{Y}|\boldsymbol{X}}$ ("attack channel") we can compute the conditional mutual information

$$I_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}) = D(P_{\boldsymbol{U},\boldsymbol{X},\boldsymbol{Y}} || P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{Y}|\boldsymbol{U}}),$$

where $D(\cdot||\cdot)$ is the Kullback-Leibler distance, defined for any probability measures P and Q as

$$D(P||Q) = \begin{cases} \int \log \frac{dP}{dQ} dP & \text{if } P \ll Q \\ \infty & \text{otherwise} \end{cases}.$$

Here, $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q, and $P \ll Q$ means that P is absolutely continuous with respect to Q. If P and Q have densities f_P and f_Q, then $D(P||Q) = E_P[\log \frac{f_P}{f_Q}]$. We can similarly compute other mutual information quantities.

Like the watermarking game, the *mutual information game* is a game played between two players in which the second player (attacker) has full knowledge of the strategy of the first player (encoder). The main difference between the two games is that the strategies in the mutual information game are conditional distributions instead of mappings, and the payoff function is mutual information, which may or may not have an operational significance in terms of achievable rates.

We first describe the *private mutual information game*, which is based on the capacity of a channel with state information known at both encoder and decoder (23). For every n, the encoder chooses a *watermarking channel* $P_{\boldsymbol{X}|\boldsymbol{U}}$ that satisfies the average distortion constraint (10), and the attacker then chooses an *attack channel* $P_{\boldsymbol{Y}|\boldsymbol{X}}$ that satisfies the average distortion constraint (13). The quantity that the encoder wishes to maximize and that the attacker wishes to minimize is

$$I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) = \frac{1}{n} I_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}). \tag{26}$$

The *value of the private mutual information game* is thus

$$C_{\text{priv}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) = \liminf_{n\to\infty} \sup_{P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}})} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}), \tag{27}$$

where

$$\mathcal{D}_1(D_1, P_{\boldsymbol{U}}) = \left\{ P_{\boldsymbol{X}|\boldsymbol{U}} : E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}}}[d_1(\boldsymbol{U}, \boldsymbol{X})] \le D_1 \right\}, \tag{28}$$

and

$$\mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}) = \left\{ P_{\boldsymbol{Y}|\boldsymbol{X}} : E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{Y}|\boldsymbol{X}}}[d_2(\boldsymbol{X}, \boldsymbol{Y})] \le D_2 \right\}. \tag{29}$$

Note that the choice of $P_{\boldsymbol{X}|\boldsymbol{U}}$ influences the set of distributions from which $P_{\boldsymbol{Y}|\boldsymbol{X}}$ can be chosen. Thus, this is not a standard static zero-sum game; it is better described as a dynamic two-stage zero-sum game of complete and perfect information. Also note that we take the limit in (27) since there is no a-priori guarantee that the attacker will use a memoryless strategy.

We next describe the *public mutual information game*, which is based on the capacity of a channel with state information known non-causally to the encoder (24). We first define an auxiliary random vector $\boldsymbol{V}$ that depends on the random vectors $\boldsymbol{U}$ and $\boldsymbol{X}$. The watermarking channel is

expanded to include not only the conditional distribution $P_{\boldsymbol{X}|\boldsymbol{U}}$ but also the conditional distribution $P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}$. Given the random vector $\boldsymbol{X}$, the random vector $\boldsymbol{Y}$ is independent of both $\boldsymbol{U}$ and $\boldsymbol{V}$, so that the joint distribution of the random vectors $\boldsymbol{U}$, $\boldsymbol{X}$, $\boldsymbol{V}$ and $\boldsymbol{Y}$ is the product of the laws $P_{\boldsymbol{U}}$, $P_{\boldsymbol{X}|\boldsymbol{U}}$, $P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}$, and $P_{\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{X},\boldsymbol{V}} = P_{\boldsymbol{Y}|\boldsymbol{X}}$. In the public version, the quantity of interest is $n^{-1}(I(\boldsymbol{V};\boldsymbol{Y}) - I(\boldsymbol{V};\boldsymbol{U}))$, which is written more explicitly as

$$I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) = \frac{1}{n}\left(I_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{V};\boldsymbol{Y}) - I_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}}(\boldsymbol{V};\boldsymbol{U})\right).$$

The *value of the public mutual information game* is thus

$$C_{\text{pub}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) = \liminf_{n\to\infty} \sup_{\substack{P_{\boldsymbol{X}|\boldsymbol{U}}\in\mathcal{D}_1(D_1,P_{\boldsymbol{U}}),\\ P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}}} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}\in\mathcal{D}_2(D_2,P_{\boldsymbol{U}},P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}). \tag{30}$$

Note that we do not restrict the overall watermarking channel $P_{\boldsymbol{V},\boldsymbol{X}|\boldsymbol{U}}$ to have $\boldsymbol{X}$ be a deterministic function of $\boldsymbol{V}$ and $\boldsymbol{U}$ as in (25) since the argument in [22] that (25) is optimal assumes finite alphabets. However, we shall see in Section 3.2 that the optimal strategy takes this form.

In the following theorem, which is proved in Section 3, we show that the upper bound of Theorem 2.1 on the coding capacity of the watermarking game is also an upper bound on the values of the mutual information games. Moreover, for i.i.d. Gaussian covertexts, this upper bound is tight.

Theorem 2.4. *For the covertext $\{P_{\boldsymbol{U}}\}$ and the distortions D_1 and D_2*

$$\begin{aligned} C_{\text{pub}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C_{\text{priv}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &(31)\\ &\leq C^*(D_1, D_2, \underline{\sigma}_u^2), &(32)\end{aligned}$$

where $\underline{\sigma}_u^2$ is defined by

$$\underline{\sigma}_u^2 = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E_{P_{\boldsymbol{U}}}[U_i^2]$$

and is assumed finite. Equality is achieved in both (31) and (32) if the covertext is zero-mean i.i.d. Gaussian.

Note that for a zero-mean variance-σ_u^2 i.i.d. Gaussian covertext, Theorem 2.4 differs from the initial results of Moulin and O'Sullivan [8, 9], but agrees with an updated version [28]. In our

notation, the results of [8, 9] can be expressed as $\frac{1}{2}\log(1+s(A;D_1,D_2,\sigma_u^2))$, where the parameter A is fixed to $\sigma_u^2+D_1$, rather than, as in (8), being optimized over $A\in\mathcal{A}(D_1,D_2,\sigma_u^2)$. (The value $\sigma_u^2+D_1$ is *not* the optimal choice for $A\in\mathcal{A}(D_1,D_2,\sigma_u^2)$.) See Section 3.4 for further discussion on this difference.

Also note that the mutual information games introduced here are somewhat different from the classical information theoretic games corresponding to communication in the presence of an unknown jammer [29] or the lossless encoding of an unknown source [30]. While all these games can be viewed as two-player zero-sum games (encoder vs. attacker, communicator vs. jammer, and compressor vs. nature) the watermarking games are *dynamic*, whereas the latter two games are *static*. Thus, in the latter two games the feasible actions of each player do not depend on the actions of the other player, and under the proper convexity/mixture conditions a saddle-point guarantees that the value of the game does not depend on who plays first. In contrast, in the watermarking mutual information games the feasible actions of the attacker depend on the action of the encoder. The watermarking games are thus two-stage games where in the first stage the first player (encoder) chooses a watermarking channel, and in doing so determines the set of attack channels from which the second player (attacker) can choose in the second stage. In the terminology of game theory, the watermarking mutual information games are dynamic (two-stage) zero-sum games of complete and perfect information [31]; see Section 3.4 for further game theoretic interpretation.

2.4 Another Extension of Costa [2] (Non-white "noise", non-Gaussian "dirt")

In Section 2.2, we showed that Costa's result on "writing on dirty paper" [2] can be extended to situations when the unknown noise is power limited but arbitrary. In this section, we use the capacity with side information results of Section 2.3.1 to extend Costa's result to more general channels as follows. Consider a channel whose output $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y}=\tilde{\boldsymbol{x}}+\boldsymbol{U}+\tilde{\boldsymbol{Y}},$$

where the input $\tilde{\boldsymbol{x}}$ is average-power limited to D_1; $\boldsymbol{U}$ is any power-limited ergodic process, which is non-causally known to the encoder but not to the decoder; and $\tilde{\boldsymbol{Y}}$ is a stationary Gaussian process, which is known to neither encoder nor decoder. Assume that $\boldsymbol{U}$ and $\tilde{\boldsymbol{Y}}$ are independent, and that their joint law does not depend on the input $\tilde{\boldsymbol{x}}$.

We next show that the capacity of this channel is the same as the capacity that would result if $\boldsymbol{U}$ were not only known to the encoder but also to the receiver, namely, the Gaussian water-filling capacity [17]. Costa's result then follows by considering the case where both $\boldsymbol{U}$ and $\tilde{\boldsymbol{Y}}$ are i.i.d. zero-mean Gaussian random processes. Similar extensions have been given recently for $\boldsymbol{U}$ and $\tilde{\boldsymbol{Y}}$ both Gaussian but not necessarily white [32, 33], and for $\tilde{\boldsymbol{Y}}$ i.i.d. Gaussian and $\boldsymbol{u}$ an arbitrary sequence [34, 35]. In light of this independently derived latter result, we will only sketch the proof of this theorem.

Using an interleaving argument, it should be intuitively clear that if the result holds for any power-limited i.i.d. law on $\boldsymbol{U}$, then it should also hold for any power-limited ergodic law. Also, by diagonalizing the problem and reducing it to a set of parallel scalar channels whose noise component (the component that is known to neither encoder nor decoder) is i.i.d. [18, 36] it should be clear that it suffices to prove this result for the case where $\tilde{\boldsymbol{Y}}$ is i.i.d. .

Consider then the case where $\boldsymbol{U}$ and $\tilde{\boldsymbol{Y}}$ are i.i.d. sequences of random variables, and where U_k has an arbitrary distribution (with finite second moment) and $\tilde{Y}_k$ has a Gaussian distribution with mean zero and variance D_2. We will specify a joint distribution on U_k, the input $\tilde{X}_k$, and an auxiliary random variable V_k such that $E\left[\tilde{X}_k^2\right] \leq D_1$ and

$$I(V_k; \tilde{X}_k + U_k + \tilde{Y}_k) - I(V_k; U_k) = \frac{1}{2} \log \left(1 + \frac{D_1}{D_2}\right). \tag{33}$$

Note that the RHS of (33) coincides with the capacity if $\boldsymbol{U}$ were also known at the decoder. The desired result follows from the achievability part of (24), which does not depend strongly the alphabets being finite.

For our joint distribution, let $\tilde{X}_k$ be a zero-mean Gaussian of variance D_1 independent of U_k and $\tilde{Y}_k$. Also, let the auxiliary random variable $V_k = \alpha U_k + \tilde{X}_k$, where $\alpha = D_1/(D_1 + D_2)$. Notice that with this choice of V_k,

$$V_k - \alpha(\tilde{X}_k + U_k + \tilde{Y}_k) = \tilde{X}_k - \alpha(\tilde{X}_k + \tilde{Y}_k), \tag{34}$$

and that with this choice of α the random variables $\tilde{X}_k - \alpha(\tilde{X}_k + \tilde{Y}_k)$ and $\tilde{X}_k + \tilde{Y}_k$ are uncorrelated and hence, being zero-mean jointly Gaussian, also independent. Furthermore, the random variables

$\tilde{X}_k - \alpha(\tilde{X}_k + \tilde{Y}_k)$ and $\tilde{X}_k + U_k + \tilde{Y}_k$ are independent since U_k is independent of $(\tilde{X}_k, \tilde{Y}_k)$. Consequently,

$$\begin{aligned} h(V_k|\tilde{X}_k + U_k + \tilde{Y}_k) &= h\big(V_k - \alpha(\tilde{X}_k + U_k + \tilde{Y}_k)|\tilde{X}_k + U_k + \tilde{Y}_k\big) \\ &= h\big(\tilde{X}_k - \alpha(\tilde{X}_k + \tilde{Y}_k)\big) \\ &= h\big(\tilde{X}_k - \alpha(\tilde{X}_k + \tilde{Y}_k)|\tilde{X}_k + \tilde{Y}_k\big) \\ &= h(\tilde{X}_k|\tilde{X}_k + \tilde{Y}_k), \end{aligned} \tag{35}$$

where all of the differential entropies exist since $\tilde{X}_k$ and $\tilde{Y}_k$ are independent Gaussians, and the second and third equalities follow by (34) and the above discussed independence. Also, the independence of U_k and $\tilde{X}_k$ implies that

$$h(V_k|U_k) = h(\alpha U_k + \tilde{X}_k|U_k) = h(\tilde{X}_k|U_k) = h(\tilde{X}_k). \tag{36}$$

Combining the definition of mutual information for random variables with densities with (35) and (36) we see that

$$\begin{aligned} I(V_k; \tilde{X}_k + U_k + \tilde{Y}_k) - I(V_k; U_k) &= h(V_k) - h(V_k|\tilde{X}_k + U_k + \tilde{Y}_k) - h(V_k) + h(V_k|U_k) \\ &= I(\tilde{X}_k; \tilde{X}_k + \tilde{Y}_k). \end{aligned} \tag{37}$$

The proof is completed by noting that the RHS of (33) equals the RHS of (37).

3 Values of the Mutual Information Games

In this section, we study the mutual information games (27), (30) and prove Theorem 2.4. The upper bound on the values of the games is based on a family of attack channels that will be described in Section 3.1. The equality for i.i.d. zero-mean Gaussian covertexts is based on the watermarking channels that will be described in Section 3.2. In Section 3.3, we give a series of lemmas that demonstrate that the proposed channels are optimal. In Section 3.4, we provide a game theoretic interpretation of these results.

3.1 Optimal Attack Channel

The attack channel that we describe here does not depend on the version of the game. Since the attacker is assumed to be cognizant of the covertext distribution $P_{\boldsymbol{U}}$ and of the watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$, it can compute

$$A_n = \frac{1}{n} E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}}}[\|\boldsymbol{X}\|^2]. \tag{38}$$

It then bases its attack channel on A_n and on its allowed distortion D_2 as follows. If $A_n \leq D_2$ then the attacker can guarantee zero mutual information by setting the forgery $\boldsymbol{Y}$ deterministically to zero without violating the distortion constraint. We shall thus focus on the case $A_n > D_2$. For this case the proposed attack channel is memoryless, and we proceed to describe its marginal. For any $A > D_2$, let the conditional distribution $P^A_{Y|X}$ have the density[5]

$$f^A_{Y|X}(y|x) = \mathcal{N}\big(y\,;\, c(A;D_2)\cdot x\,,\, c(A;D_2)\cdot D_2\big),$$

where $c(\cdot;\cdot)$ is defined in (5), and where our notation $f^A_{Y|X}(y|x)$ makes D_2 implicit. Equivalently, under $P^A_{Y|X}$ the random variable Y is distributed as $c(A;D_2)X + S_2$, where S_2 is a zero-mean variance-$c(A;D_2)D_2$ Gaussian random variable independent of X. The conditional distribution $P^A_{Y|X}$ is thus equivalent to the Gaussian rate distortion forward channel [17] for a variance-A Gaussian source and an allowable distortion D_2.

For blocklength n and $A_n > D_2$, the proposed attacker $P_{\boldsymbol{Y}|\boldsymbol{X}}$ is

$$P_{\boldsymbol{Y}|\boldsymbol{X}} = \Big(P^{A_n}_{Y|X}\Big)^n,$$

that is, $P_{\boldsymbol{Y}|\boldsymbol{X}}$ has a product form with marginal $P^{A_n}_{Y|X}$, where A_n is given in (38). Notice that by (38) and the structure of the attack channel

$$E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} (P^{A_n}_{Y|X})^n}\left[\frac{1}{n}\|\boldsymbol{Y}-\boldsymbol{X}\|^2\right] = \big(c(A_n;D_2)-1\big)^2 A_n + c(A_n;D_2)\, D_2 = D_2.$$

Thus the attack channel $(P^{A_n}_{Y|X})^n$ satisfies the distortion constraint. Compare this attack channel with the attacker (defined in Section 5.2) used in the proof of the converse of the watermarking game.

[5] We use $\mathcal{N}(x;\mu,\sigma^2)$ to denote the density at x of a Gaussian distribution of mean μ and variance σ^2.

3.2 Optimal Watermarking Channel

In this section we focus on i.i.d. zero-mean variance-σ_u^2 Gaussian covertexts and describe watermarking channels that will demonstrate that for such covertexts (31) and (32) both hold with equality. The watermarking channels are memoryless, and it thus suffices to describe their marginals. The proposed watermarking channels depend on the version of the game, on $(\sigma_u^2,\ D_1,\ D_2)$, and on a parameter $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, whose choice is at the watermarker's discretion. Later, of course, we shall optimize over this choice.

Private Version: For any $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, let the conditional distribution $P^A_{X|U}$ be Gaussian with mean $b_1 U$ and variance b_2, i.e., have the density

$$f^A_{X|U}(x|u) = \mathcal{N}(x; b_1 u, b_2), \tag{39}$$

where $b_1 = b_1(A; D_1, \sigma_u^2)$ and $b_2 = b_2(A; D_1, \sigma_u^2)$ as in (3) and (4), and where our notation $f^A_{X|U}(x|u)$ makes D_1 and σ_u^2 implicit. Equivalently, under $P^A_{X|U}$ the random variable X is distributed as $b_1 U + S_1$, where S_1 is a zero mean Gaussian random variable of variance b_2 which is independent of U. For i.i.d. zero-mean variance-σ_u^2 Gaussian covertexts we have

$$E_{P_{\boldsymbol{U}}(P^A_{X|U})^n}\left[\frac{1}{n}\|\boldsymbol{X} - \boldsymbol{U}\|^2\right] = (b_1 - 1)^2\sigma_u^2 + b_2 = D_1.$$

Thus for this covertext distribution (and, in fact, for any covertext distribution of power σ_u^2), the watermarking channel $(P^A_{X|U})^n$ satisfies the distortion constraint. Furthermore,

$$E_{P_{\boldsymbol{U}}(P^A_{X|U})^n}\left[\frac{1}{n}\|\boldsymbol{X}\|^2\right] = A,$$

which gives an interpretation of the parameter A as the power in the stegotext induced by the covertext and the watermarking channel. This watermarking channel can be used as a basis for an achievability scheme for the private Gaussian watermarking game; see Section 4.2.

Public Version: For the public game, the conditional distribution of the random vector $\boldsymbol{V}$ given the random vectors $\boldsymbol{U}$ and $\boldsymbol{X}$ is also needed. The optimal such distribution turns out to be deterministic and memoryless. In particular, for A as above, let the distribution $P^A_{V|U,X}$ be described by

$$V = (\alpha(A; D_1, D_2, \sigma_u^2) - 1)U + X,$$

where $\alpha(A; D_1, D_2, \sigma_u^2)$ is defined in (68), and let

$$P^A_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}} = (P^A_{V|U,X})^n.$$

Compare this expanded watermarking channel with the achievability scheme for the public Gaussian watermarking game given in Section 4.3.

3.3 Analysis

In this section, we state three lemmas, which together prove Theorem 2.4. Lemma 3.1 demonstrates the intuitive fact that the value of the public version of the mutual information game cannot exceed the value of the private version. Lemma 3.2 shows that, by using the attack channel proposed in Section 3.1, the attacker can guarantee that the value of the private mutual information game not exceed $C^*(D_1, D_2, \underline{\sigma}_u^2)$, where $\underline{\sigma}_u^2$ is defined in (2.4). Lemma 3.3 shows that by watermarking an i.i.d. zero-mean variance-σ_u^2 Gaussian source using the channel proposed in Section 3.2 with the appropriate choice of A, the encoder can guarantee a value for the public mutual information game of at least $C^*(D_1, D_2, \sigma_u^2)$. The proofs of the following three lemmas are given in Appendices A.1, A.2 and A.3, respectively.

Lemma 3.1. *For any $n > 0$ and any covertext distribution $P_{\boldsymbol{U}}$,*

$$\sup_{\substack{P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}}) \\ P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}}} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) \leq$$

$$\sup_{P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}})} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}).$$

Since this lemma holds for every n, it implies (31).

Lemma 3.2. *For any $n > 0$, any covertext distribution $P_{\boldsymbol{U}}$, any watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$, and any fixed distortion $D_2 > A_n$*

$$\begin{aligned} I_{\text{priv}}\left(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, (P^{A_n}_{Y|X})^n\right) &\leq I_{\text{priv}}\left((P^G_U)^n, (P^{A_n}_{X|U})^n, (P^{A_n}_{Y|X})^n\right) \\ &= \frac{1}{2}\log\bigl(1 + s(A_n; D_{1,n}, D_2, \sigma^2_{u,n})\bigr), \end{aligned} \tag{40}$$

where

$$\sigma_{u,n}^2 = E_{P_{\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{U}\|^2\right]; \tag{41}$$

$$D_{1,n} = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}[n^{-1}\|\boldsymbol{X}-\boldsymbol{U}\|^2]; \tag{42}$$

$$A_n = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{X}\|^2\right]; \tag{43}$$

P_U^G denotes a zero-mean Gaussian distribution of variance $\sigma_{u,n}^2$; $P_{X|U}^{A_n}$ is the watermarking channel described in Section 3.2 for the parameters $\sigma_{u,n}^2$, $D_{1,n}$ and A_n; and $P_{Y|X}^{A_n}$ is the attack channel described in Section 3.1 for the parameters D_2 and A_n.

This lemma proves (32). To see this note that, by the definition of $\underline{\sigma}_u^2$, for any $\epsilon > 0$ and any integer n_0 there exists some $n > n_0$ such that

$$\sigma_{u,n}^2 < \underline{\sigma}_u^2 + \epsilon, \tag{44}$$

where $\sigma_{u,n}^2$ is defined in (41). Also, by the distortion constraint (i.e. $P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}})$),

$$D_{1,n} \leq D_1, \tag{45}$$

where $D_{1,n}$ is defined in (42).

If A_n defined in (43) is less than D_2, then the attack channel that sets the forgery deterministically to zero is allowable and the resulting mutual information is zero. Thus, (32) is satisfied in this case. We thus focus on the case when $A_n > D_2$. We also note that

$$\left(\sigma_{u,n} - \sqrt{D_{1,n}}\right)^2 \leq A_n \leq \left(\sigma_{u,n} + \sqrt{D_{1,n}}\right)^2$$

by the triangle inequality so that $A_n \in \mathcal{A}(D_{1,n}, D_2, \sigma_{u,n}^2)$. By definition of $C^*(\cdot,\cdot,\cdot)$ (8), it follows that the right hand side (RHS) of (40) is at most $C^*(D_{1,n}, D_2, \sigma_{u,n}^2)$. This in turn is upper bounded by $C^*(D_1, D_2, \underline{\sigma}_u^2 + \epsilon)$ in view of (44) and (45), because $C^*(D_1, D_2, \sigma_u^2)$ is non-decreasing in D_1 and σ_u^2.[6] Finally, since $\epsilon > 0$ is arbitrary and $C^*(\cdot,\cdot,\cdot)$ is continuous, it follows that the attacker $P_{\boldsymbol{Y}|\boldsymbol{X}}^{A_n}$ guarantees that $C_{\text{priv}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\})$ is upper bounded by $C^*(D_1, D_2, \underline{\sigma}_u^2)$.

[6]This follows since $s(D_1, D_2, \sigma_u^2)$ is increasing in D_1 and σ_u^2 and since if A^* optimizes (8) for D_1 and σ_u^2 then $A^* \in \mathcal{A}(D_1 + \epsilon, D_2, \sigma_u^2 + \epsilon)$ for some $\epsilon > 0$.

This lemma also shows that for an i.i.d. Gaussian covertext, if the memoryless attack channel $(P^A_{Y|X})^n$ is used, then, of all watermarking channels that satisfy $E\left[n^{-1}\|\boldsymbol{X}\|^2\right] = A$, mutual information is maximized by the memoryless watermarking channel $(P^A_{X|U})^n$ of Section 3.2.

Lemma 3.3. *Consider an i.i.d. zero-mean variance-σ_u^2 Gaussian covertext (denoted $(P^G_U)^n$) and fixed distortions D_1 and D_2. If the attack channel $P_{\boldsymbol{Y}|\boldsymbol{X}}$ satisfies $E_{(P^G_U P^A_{X|U})^n P_{\boldsymbol{Y}|\boldsymbol{X}}}\left[n^{-1}\|\boldsymbol{Y}-\boldsymbol{X}\|\right] \leq D_2$, then for all $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$,*

$$\begin{aligned} I_{\text{pub}}\left((P^G_U)^n, (P^A_{X|U})^n, (P^A_{V|U,X})^n, P_{\boldsymbol{Y}|\boldsymbol{X}}\right) &\geq I_{\text{pub}}\left((P^G_U)^n, (P^A_{X|U})^n, (P^A_{V|U,X})^n, (P^A_{Y|X})^n\right) \\ &= \frac{1}{2}\log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr). \end{aligned}$$

Here, $P^A_{X|U}$ and $P^A_{V|U,X}$ are the watermarking channels described in Section 3.2 for the parameters σ_u^2, D_1 and A and $P^A_{Y|X}$ is the attack channel described in Section 3.1 for the parameters D_2 and A.

This lemma implies that for a zero-mean variance-σ_u^2 i.i.d. Gaussian covertext, the value of the public mutual information game is lower bounded by $C^*(D_1, D_2, \sigma_u^2)$. Indeed, the encoder can use the watermarking channels defined by $(P^{A^*}_{X|U})^n$ and $(P^{A^*}_{V|U,X})^n$ where A^* achieves the maximum in the definition of C^*. Since for any covertext distribution (and in particular for an i.i.d. Gaussian covertext) the value of the private version is at least as high as the value of the public version (Lemma 3.1), it follows from the above that, for an i.i.d. Gaussian covertext, C^* is also a lower bound on the value of the private Gaussian mutual information game. This lemma also shows that when the covertext is zero-mean i.i.d. Gaussian and the memoryless watermarking channels $(P^A_{X|U})^n$ and $(P^A_{V|U,X})^n$ are used, then to minimize the mutual information the attacker should use the memoryless attack channel $(P^A_{Y|X})^n$.

The combination of Lemmas 3.1, 3.2 and 3.3 shows that for a zero-mean i.i.d. Gaussian covertext of variance σ_u^2, the value of both the private and public Gaussian mutual information games is exactly $C^*(D_1, D_2, \sigma_u^2)$.

3.4 Game Theoretic Interpretation

We have seen that, for an i.i.d. Gaussian covertext, memoryless encoders and attackers are optimal. However, there does not exist one memoryless attacker that is both valid for any memoryless encoder

and guarantees that the mutual information is less than $C^*(D_1, D_2, \sigma_u^2)$. That is, a memoryless saddlepoint does not exist.

In this section, we more carefully examine the private version of this mutual information game from a game theoretic perspective. Recall that the encoder is trying to maximize I_{priv} and the attacker is trying to minimize I_{priv}. In game theoretic terminology (see e.g., [31]), this is a zero-sum game with I_{priv} as the pay-off to the first player (encoder) and $-I_{\text{priv}}$ as the pay-off to the second player (attacker). Specifically, this mutual information game is a dynamic zero-sum game of complete and perfect information. In particular, the game is not static, and thus we need to consider an attacker strategy of lists of responses to every possible watermarking channel. We show that a subgame-perfect Nash equilibrium gives the value of the game, where we use the term "value of the game" to denote the highest possible guaranteed pay-off to the first player. We also illustrate the difference between the value of the game given here and the value of a similar game given in [8, 9].

For a dynamic game, a strategy space for each player is specified by listing a feasible action for each possible contingency in the game. Since the encoder plays first, his strategy space is simply the set of feasible watermarking channels, i.e., $\mathcal{D}_1\big(D_1, (P_U^G)^n\big)$ defined in (28). However, the attacker plays second and thus his strategy space consists of all mappings of the form

$$\psi : P_{\boldsymbol{X}|\boldsymbol{U}} \mapsto P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2\left(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}\right), \ \ \forall P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1\left(D_1, (P_U^G)^n\right), \tag{46}$$

where $\mathcal{D}_2\big(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}\big)$ is defined in (29). That is, for every possible strategy $P_{\boldsymbol{X}|\boldsymbol{U}}$ the encoder might use, the attacker must choose a feasible response $\psi(P_{\boldsymbol{X}|\boldsymbol{U}})$.

For this game, an encoder strategy $P^*_{\boldsymbol{X}|\boldsymbol{U}}$ and an attacker strategy $\psi^*(\cdot)$ form a *Nash equilibrium* if

$$I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P_{\boldsymbol{X}|\boldsymbol{U}})\right) \le I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right), \tag{47}$$

for every $P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, (P_U^G)^n)$, and if

$$I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right) \le I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right), \tag{48}$$

for every mapping $\psi(\cdot)$ of the form (46). That is, given that the attacker will use $\psi^*(\cdot)$, the encoder maximizes its pay-off by using $P^*_{\boldsymbol{X}|\boldsymbol{U}}$. Conversely, given that the encoder will use $P^*_{\boldsymbol{X}|\boldsymbol{U}}$, the attacker

maximizes its pay-off (minimizes the encoder's pay-off) by using $\psi^*(\cdot)$.

For this game, an encoder strategy $P^*_{\boldsymbol{X}|\boldsymbol{U}}$ and an attacker strategy $\psi^*(\cdot)$ form a *subgame-perfect Nash equilibrium* if they form a Nash equilibrium and if additionally

$$I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P_{\boldsymbol{X}|\boldsymbol{U}})\right) \leq I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}\right)$$

for all $P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, (P_U^G)^n)$ and for all $P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}})$. That is, the attacker must choose the best response to *any* possible encoder strategy, and not just the maximized encoder strategy as in the regular Nash equilibrium. The value of the game is given by evaluating the mutual information I_{priv} at any subgame-perfect Nash equilibrium

Using this terminology we see that Lemma 3.2 and Lemma 3.3 imply that there exists a subgame-perfect Nash equilibrium of the form $\left(\left(P^{A^*}_{X|U}\right)^n, \psi^*(\cdot)\right)$ where $P^A_{X|U}$ is defined above in Section 3.2, A^* achieves the maximum in (8), and $\psi^*\big((P^A_{X|U})^n\big) = (P^A_{Y|X})^n$ for every $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, where $P^A_{Y|X}$ is defined in Section 3.1. The value of the game is thus $C^*(D_1, D_2, \sigma_u^2)$, as we have demonstrated above.

Using the above concepts, we now discuss the value of this game that was initially given in [8, 9] and later revised in [28]. For $A_0 = \sigma_u^2 + D_1$,

$$I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, (P^{A_0}_{Y|X})^n\right) \leq I_{\text{priv}}\left((P_U^G)^n, (P^{A_0}_{X|U})^n, (P^{A_0}_{Y|X})^n\right), \tag{49}$$

for every $P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, (P_U^G)^n)$, and

$$I_{\text{priv}}\left((P_U^G)^n, (P^{A_0}_{X|U})^n, (P^{A_0}_{Y|X})^n\right) \leq I_{\text{priv}}\left((P_U^G)^n, (P^{A_0}_{X|U})^n, P_{\boldsymbol{Y}|\boldsymbol{X}}\right), \tag{50}$$

for every $P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2\big(D_2, (P_U^G)^n, (P^{A_0}_{X|U})^n\big)$. Thus, it would seem that if we were to define

$$\psi_0(P_{\boldsymbol{X}|\boldsymbol{U}}) = (P^{A_0}_{Y|X})^n, \ \ \forall P_{\boldsymbol{X}|\boldsymbol{U}}, \tag{51}$$

then the pair $\big((P^{A_0}_{X|U})^n, \psi_0(\cdot)\big)$ would form a Nash equilibrium according to the definitions (47) and (48). It is indeed the value for this pair that is given in [8, 9]. Note, however, that (51) *is not a feasible strategy* for the attacker since for some watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$ (in particular, for any $P_{\boldsymbol{X}|\boldsymbol{U}}$ with $n^{-1}E\left[\|\boldsymbol{X}\|^2\right] > A_0$), the attacker's response of (51) is not feasible, i.e., $(P^{A_0}_{Y|X})^n \notin \mathcal{D}_2(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}})$. This observation is critical since we have found that for some $A > A_0$ (in

particular, the maximizing A in (8)), the watermarking channel $(P^A_{X|U})^n$, which has $n^{-1}E\left[\|\boldsymbol{X}\|^2\right] = A$, guarantees a mutual information that is strictly larger than the value of the game reported in [8, 9].

4 Achievability for a.s. Constraints

In this section, we prove the watermarking achievability results of Theorem 2.1 (general attack) and Theorem 2.3 (additive attack). The achievability in the private versions will be discussed in Section 4.2 only briefly, because, as we will show, these rates are achievable even in the public versions. The public-version achievability results are based on a coding strategy that is described in Section 4.3 and analyzed in Section 4.4. We begin in Section 4.1 with some preliminary results that we will use throughout the section.

4.1 Preliminaries

4.1.1 Spherical caps

We now state some asymptotic properties of the surface area of a spherical cap on a unit n-sphere. We denote the n-dimensional sphere centered at $\boldsymbol{\mu} \in \mathbb{R}^n$ with radius $r \geq 0$ by $S^n(\boldsymbol{\mu}, r)$, i.e.,

$$S^n(\boldsymbol{\mu}, r) = \{\boldsymbol{\xi} \in \mathbb{R}^n : \|\boldsymbol{\xi} - \boldsymbol{\mu}\| = r\}.$$

For any vector $\boldsymbol{\mu} \in S^n(0,1)$ and any angle $0 \leq \theta \leq \pi$, we let $\mathcal{C}(\boldsymbol{\mu}, \theta) \subset S^n(0,1)$ denote the *spherical cap* centered at $\boldsymbol{\mu}$ with half-angle θ,

$$\mathcal{C}(\boldsymbol{\mu}, \theta) = \{\boldsymbol{\xi} \in S^n(0,1) : \langle \boldsymbol{\mu}, \boldsymbol{\xi} \rangle > \cos\theta\}. \tag{52}$$

The surface area of this spherical cap in $\mathbb{R}^n$ depends only on the angle θ, and is denoted by $C_n(\theta)$. Note that $C_n(\pi)$ is the surface area of the unit n-sphere.

If a random vector $\boldsymbol{\Psi}$ is uniformly distributed over the unit n-sphere $S^n(0,1)$ and if $\boldsymbol{\mu}$ is any vector in $S^n(0,1)$, then for any $0 \leq \tau \leq 1$,

$$\Pr(\langle \boldsymbol{\Psi}, \boldsymbol{\mu} \rangle > \tau) = \frac{C_n(\arccos\tau)}{C_n(\pi)}. \tag{53}$$

Indeed, the vectors $\boldsymbol{\psi} \in S^n(0,1)$ that have the specified inner product with the vector $\boldsymbol{\mu}$ are precisely the elements of the set $\mathcal{C}(\boldsymbol{\mu}, \arccos\tau)$; see (52). To analyze a probability of error expression such as (84), we shall need some asymptotic properties of the ratio on the RHS of (53). In [37], Shannon derived bounds that asymptotically yield

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \frac{C_n(\arccos\tau)}{C_n(\pi)} &= \log\big(\sin(\arccos\tau)\big) \\ &= \log(1-\tau^2), \end{aligned} \tag{54}$$

for every $0 < \tau < 1$. See also [38]. To complete our asymptotic analysis, we shall also need the following technical lemma.

Lemma 4.1. *Let $f : \mathbb{R} \mapsto (0,1]$ be such that the limit*

$$-\eta_1 = \lim_{t\to\infty} \frac{1}{t} \log f(t)$$

exists and is negative so that $\eta_1 > 0$. Then

$$\lim_{t\to\infty} \big(1 - f(t)\big)^{2^{t\eta_2}} = \begin{cases} 1 & \text{if } \eta_1 > \eta_2 \\ 0 & \text{if } \eta_1 < \eta_2 \end{cases}. \tag{55}$$

The proof of this lemma is based on the fact that (55) is true for $f(t) = 2^{-t\eta_1}$; the details are omitted.

4.1.2 Deterministic Attackers are Sufficient

To prove achievability in the watermarking game, we can without loss of generality limit the attacker to deterministic attacks. That is, it is sufficient to show that the average probability of error (averaged over the side information, secret key and message) is small for all deterministic attacker mappings

$$\boldsymbol{y} = g_n(\boldsymbol{x}) \tag{56}$$

instead of the more general $g_n(\boldsymbol{x}, \theta_2)$. With an attacker of this form, the distortion constraint (12) can be rewritten as $n^{-1}\|g_n(\boldsymbol{X}) - \boldsymbol{X}\|^2 \le D_2$, almost surely.

Indeed, we can evaluate the average probability of error (averaged over everything including

the attack key Θ_2) by first conditioning on the attack key Θ_2. Thus, if the average probability of error given every attacker mapping of the form (56) is small, then the average probability of error for any general attacker mapping of the form (11) is also small.

In the additive attack watermarking game, a deterministic attacker takes on a particularly simple form. Indeed, combining the forms (56) and (18), we see that the attacker can be written as

$$g_n(\boldsymbol{x}) = \boldsymbol{x} + \tilde{\boldsymbol{y}} \tag{57}$$

for some sequence $\tilde{\boldsymbol{y}}$ that satisfies

$$\frac{1}{n}\|\tilde{\boldsymbol{y}}\|^2 \leq D_2. \tag{58}$$

In the general watermarking game, a deterministic attack $g_n(\boldsymbol{x})$ can be decomposed into its projection onto the stegotext $\boldsymbol{x}$ and its projection onto $\boldsymbol{x}^{\perp}$. That is, we can write

$$g_n(\boldsymbol{x}) = \gamma_1(\boldsymbol{x})\boldsymbol{x} + \boldsymbol{\gamma}_2(\boldsymbol{x}), \tag{59}$$

for some $\gamma_1 : \mathbb{R}^n \mapsto \mathbb{R}$ and some $\boldsymbol{\gamma}_2 : \mathbb{R}^n \mapsto \mathbb{R}^n$, where $\langle \boldsymbol{\gamma}_2(\boldsymbol{x}), \boldsymbol{x} \rangle = 0$. Defining

$$\gamma_3(\boldsymbol{x}) = n^{-1}\|\boldsymbol{\gamma}_2(\boldsymbol{x})\|^2, \tag{60}$$

we can rewrite the attacker's distortion constraint (12) in terms of $\gamma_1(\boldsymbol{X})$, $\boldsymbol{X}$, and $\gamma_3(\boldsymbol{X})$ as

$$\big(\gamma_1(\boldsymbol{X}) - 1\big)^2 n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X}) \leq D_2, \text{ a.s..} \tag{61}$$

4.2 Achievability: Private Versions

4.2.1 Additive Attack

For the private version of the watermarking game with an additive attack, all rates less than $\frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right)$ are achievable, *regardless* of the statistics of the covertext. Indeed, the encoder and decoder can both subtract off the covertext $\boldsymbol{U}$, resulting in a channel of output $\boldsymbol{Z} = \tilde{\boldsymbol{X}} + \tilde{\boldsymbol{Y}}$, where $\tilde{\boldsymbol{X}}$ is a function of the message and must satisfy $n^{-1}\|\tilde{\boldsymbol{X}}\|^2 \leq D_1$, a.s., and where $\tilde{\boldsymbol{Y}}$ is independent of $\tilde{\boldsymbol{X}}$ and must satisfy $n^{-1}\|\tilde{\boldsymbol{Y}}\|^2 \leq D_2$, a.s.. The required achievability thus follows from [18].

4.2.2 General Attack

For the private version of the watermarking game with a general attack, we argue in this section that all rates less than $C^*(D_1, D_2, \sigma_u^2)$ are achievable for an i.i.d. Gaussian covertext. We describe the coding and decoding strategy that achieves these rates. We will not analyze its performance in detail, since we do so in Section 4.4 for the more difficult public version; see [19] for a detailed analysis. For this brief analysis, we assume that the covertext, rather than being i.i.d. Gaussian, is instead uniformly distributed on the n-sphere $S^n(0, \sqrt{n\sigma_u^2})$. We show in Section 4.4.3 that any rates achievable for either of these related covertext distributions is also achievable for the other.

The optimal coding strategy is similar to the optimal watermarking channel described in Section 3.2 for the private version of the mutual information game. We first choose a parameter A and compute $b_1 = b_1(A; D_1, \sigma_u^2)$ and $b_2 = b_2(A; D_1, \sigma_u^2)$ as in (3) and (4). A codebook is then generated consisting of 2^{nR} i.i.d. codeword vectors, each uniformly distributed over the n-sphere $S^n(0, \sqrt{nb_2})$. Given the message w and the covertext $\boldsymbol{u}$, the stegotext is generated as a sum of $b_1\boldsymbol{u}$ and the w^{th} codeword, i.e.,

$$\boldsymbol{x} = b_1\boldsymbol{u} + \boldsymbol{c}_w(\boldsymbol{u}), \tag{62}$$

where $\boldsymbol{c}_w(\boldsymbol{u})$ is the w^{th} codeword projected onto $\boldsymbol{u}^\perp$ and renormalized to have norm nb_2. Compare (62) and (39) to see the similarity between this strategy and the strategy for the mutual information game. Note that for every covertext $\boldsymbol{u}$ and corresponding stegotext $\boldsymbol{x}$ produced by this encoding strategy, $n^{-1}\|\boldsymbol{x} - \boldsymbol{u}\|^2 = D_1$ and $n^{-1}\|\boldsymbol{x}\|^2 = A$.

The decoder uses a modified nearest-neighbor decoding rule to find its estimate $\hat{w}$ of the message. It projects the forgery $\boldsymbol{y}$ onto $\boldsymbol{u}^\perp$ to create $\boldsymbol{y}|_{\boldsymbol{u}^\perp}$ and produces the message $\hat{w}$ that, among all messages $\tilde{w}$, minimizes the Euclidean distance between $\boldsymbol{y}|_{\boldsymbol{u}^\perp}$ and $\boldsymbol{c}_{\tilde{w}}(\boldsymbol{u})$.

For this encoder and decoder, it can be shown that the probability of error tends to zero for every possible attacker as long as

$$R < \frac{1}{2}\log\left(1 + s(A; D_1, D_2, \sigma_u^2)\right),$$

where $s(\cdot;\cdot,\cdot,\cdot)$ is defined in (7). Since the encoder and decoder are free to choose A, we see from the definition (8) that all rates less than $C^*(D_1, D_2, \sigma_u^2)$ are achievable, which is the desired result.

4.3 Coding Strategies for Public Versions

The coding strategies for the public versions of both the additive attack and the general watermarking games are motivated by the works of Marton [21], Gel'fand and Pinsker [22], Heegard and El Gamal [23], and Costa [2].

For both models, we fix a $\delta > 0$. In Sections 4.3.1 and 4.3.2, we define the set of constants $\{\alpha_{\text{type}}, \rho_{\text{type}}, R_{0,\text{type}}, R_{1,\text{type}}, R_{\text{type}}\}$ for type equal to "add" (for additive attack) or "gen" (for general attack). We shall drop the subscripts for these constants in the sequel when we are referring to both cases. Using these constants we then describe the encoder and decoder used for both models. Briefly, we create a codebook consisting of 2^{nR} bins with 2^{nR_0} codewords in each bin for a total of 2^{nR_1} codewords (hence $R = R_1 - R_0$). Furthermore, the stegotext is formed as the sum of $(1-\alpha)$ times the covertext and a selected codeword, and ρ describes the target correlation between the covertext and the difference between the stegotext and the covertext. While the constants have different values for the two models, in terms of these constants the proposed coding schemes are identical.

4.3.1 Additive Attack Constants

For the additive attack watermarking game, we define the set of constants as

$$\alpha_{\text{add}} = \frac{D_1}{D_1 + D_2}, \tag{63}$$

$$\rho_{\text{add}} = 0, \tag{64}$$

$$R_{0,\text{add}} = \frac{1}{2}\log\left(1 + \frac{D_1\sigma_u^2}{(D_1+D_2)^2}\right) + \delta, \tag{65}$$

$$R_{1,\text{add}} = \frac{1}{2}\log\left(1 + \frac{D_1}{D_2} + \frac{D_1\sigma_u^2}{D_2(D_1+D_2)}\right) - \delta, \tag{66}$$

and

$$R_{\text{add}} = R_1 - R_0 = \frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right) - 2\delta. \tag{67}$$

4.3.2 General Attack Constants

The choice of the constants for the general watermarking game is inspired by the solution to the public Gaussian mutual information game; see Theorem 2.4 and its derivation in Section 3. The encoder and decoder choose a free parameter $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, where the interval $\mathcal{A}(D_1, D_2, \sigma_u^2)$

is defined in (1). We assume throughout that the above interval is non-empty, because otherwise the coding capacity is zero, and there is no need for a coding theorem.

We first let $b_1 = b_1(A; D_1, \sigma_u^2)$, $b_2 = b_2(A; D_1, \sigma_u^2)$, and $c = c(A; D_2)$ as in (3), (4), and (5). We define the main set of constants for the general watermarking game as

$$\alpha_{\text{gen}} = \alpha(A; D_1, D_2, \sigma_u^2), \tag{68}$$

$$\rho_{\text{gen}} = \rho(A; D_1, \sigma_u^2), \tag{69}$$

$$R_{0,\text{gen}} = \frac{1}{2} \log \left(1 + \frac{(\alpha_{\text{gen}} \sigma_u^2 + \rho_{\text{gen}})^2}{D_1 \sigma_u^2 - \rho_{\text{gen}}^2} \right) + \delta, \tag{70}$$

$$R_{1,\text{gen}} = \frac{1}{2} \log \left(1 + \frac{Acb_2}{D_2(D_2 + cb_2)} \right) - \delta, \tag{71}$$

and

$$R_{\text{gen}} = R_{1,\text{gen}} - R_{0,\text{gen}} = \frac{1}{2} \log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr) - 2\delta, \tag{72}$$

where $\alpha(A; D_1, D_2, \sigma_u^2)$, $\rho(A; D_1, \sigma_u^2)$ and $s(A; D_1, D_2, \sigma_u^2)$ are defined in (6), (2), and (7), respectively. If A is chosen to maximize (72) as in (8), then $R_{\text{gen}} = C^*(D_1, D_2, \sigma_u^2) - 2\delta$.

4.3.3 Encoder and Decoder

The encoder and decoder use their source of common randomness Θ_1 to create a codebook of auxiliary codewords as follows. They generate $2^{nR_{1,\text{type}}} = 2^{n(R_{\text{type}} + R_{0,\text{type}})}$ i.i.d. random vectors $\{\boldsymbol{V}_{j,k}\}$, where $1 \le j \le 2^{nR_{\text{type}}}$, $1 \le k \le 2^{nR_{0,\text{type}}}$, where each random vector $\boldsymbol{V}_{j,k}$ is uniformly distributed on the n-sphere $S^n(0, \sqrt{n\sigma_{v,\text{type}}^2})$, and where

$$\sigma_{v,\text{type}}^2 = \alpha_{\text{type}}^2 \sigma_u^2 + 2\alpha_{\text{type}} \rho_{\text{type}} + D_1. \tag{73}$$

Thus, the codebook consists of $2^{nR_{\text{type}}}$ bins (indexed by j), each containing $2^{nR_{0,\text{type}}}$ auxiliary codewords.

Given the message w and the covertext $\boldsymbol{u}$, the encoder looks in bin w and chooses the auxiliary codeword closest (in Euclidean distance) to the covertext. The output of the encoder $\boldsymbol{x}$ is then created as a linear combination of the covertext and the chosen auxiliary codeword. This can be written as follows. Given the message w, the covertext $\boldsymbol{u}$, and the codebook $\{\boldsymbol{v}_{j,k}\}$, let the chosen

index for message w be

$$k^*(\boldsymbol{u}, w) = \underset{1 \le k \le 2^{nR_{0,\text{type}}}}{\arg\max} \langle \boldsymbol{u}, \boldsymbol{v}_{w,k} \rangle, \tag{74}$$

which is unique with probability one. Further, let the chosen auxiliary codeword for message w be

$$\boldsymbol{v}_w(\boldsymbol{u}) = \boldsymbol{v}_{w,k^*(\boldsymbol{u},w)}. \tag{75}$$

The encoder creates its output $\boldsymbol{x}$ as

$$\boldsymbol{x} = \boldsymbol{v}_w(\boldsymbol{u}) + (1 - \alpha_{\text{type}})\boldsymbol{u}. \tag{76}$$

The decoder finds the auxiliary codeword that, among all the $2^{nR_{1,\text{type}}}$ sequences in the codebook, is closest to the received sequence $\boldsymbol{y}$. It then declares the estimate of the message to be the bin to which this auxiliary codeword belongs. Given the received sequence $\boldsymbol{y}$ and the codebook $\{\boldsymbol{v}_{j,k}\}$, the decoder's estimate is thus given by

$$\hat{w} = \underset{1 \le \tilde{w} \le 2^{nR_{\text{type}}}}{\arg\min} \left(\min_{1 \le k \le 2^{nR_{0,\text{type}}}} \|\boldsymbol{y} - \boldsymbol{v}_{\tilde{w},k}\|^2 \right) \tag{77}$$

$$= \underset{1 \le \tilde{w} \le 2^{nR_{\text{type}}}}{\arg\max} \left(\max_{1 \le k \le 2^{nR_{0,\text{type}}}} \langle \boldsymbol{y}, \boldsymbol{v}_{\tilde{w},k} \rangle \right), \tag{78}$$

where the last equality follows by noting that $n^{-1}\|\boldsymbol{v}_{\tilde{w},k}\|^2 = \sigma^2_{v,\text{type}}$ irrespective of $\tilde{w}$ and k. Note that $\hat{w}$ of (77) is with probability one unique.

4.4 Analysis for Public Version

We now show that for the sequence of encoders and decoders from the previous section, the probability of error tends to zero and the distortion constraint is met. We prove these two facts in Sections 4.4.1 and 4.4.2, respectively, assuming in both sections that the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $S^n\left(0, \sqrt{n\sigma_u^2}\right)$. We then extend these results to i.i.d. Gaussian covertexts in Section 4.4.3.

4.4.1 Probability of Error

In this section, we derive the conditional probability of error in the above coding strategy. Let us first define the random variables on which we will condition as

$$Z = \frac{1}{n}\langle \boldsymbol{U}, \boldsymbol{V}_W(\boldsymbol{U})\rangle, \quad Z_1 = \frac{1}{n}\|\boldsymbol{Y}\|^2, \quad Z_2 = \frac{1}{n}\langle \tilde{\boldsymbol{Y}}, \boldsymbol{V}_W(\boldsymbol{U})\rangle, \tag{79}$$

where, as in Section 2.1.3,

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}, \tag{80}$$

Let us also define a mapping $\beta_{\text{type}}(z, z_1, z_2)$ for both types as

$$\beta_{\text{type}}(z, z_1, z_2) = \frac{\sigma^2_{v,\text{type}} + (1-\alpha_{\text{type}})z + z_2}{\sqrt{z_1 \sigma^2_{v,\text{type}}}}. \tag{81}$$

By the definition of the decoder (78), it follows that a decoding error occurs if, and only if, there exists a message $w' \neq W$ and an index k' such that

$$\begin{aligned} \frac{1}{n}\langle \boldsymbol{Y}, \boldsymbol{V}_{w',k'}\rangle &\geq \frac{1}{n}\langle \boldsymbol{Y}, \boldsymbol{V}_W(\boldsymbol{U})\rangle \\ &= \frac{1}{n}\langle \boldsymbol{X}, \boldsymbol{V}_W(\boldsymbol{U})\rangle + \frac{1}{n}\langle \tilde{\boldsymbol{Y}}, \boldsymbol{V}_W(\boldsymbol{U})\rangle \\ &= \sigma_v^2 + (1-\alpha)Z + Z_2, \end{aligned}$$

where the first equality follows by the definition of $\tilde{\boldsymbol{Y}}$ (80) and the second equality follows by the definitions of the encoder (76) and the random variables Z and Z_2. Note that we do not need to consider the case where the decoder makes a mistake in the same bin since this does not result in an error. Equivalently, an error occurs if, and only if, there exists a message $w' \neq W$ and an index k' such that

$$\begin{aligned} \left\langle \frac{\boldsymbol{Y}}{\sqrt{nZ_1}}, \frac{\boldsymbol{V}_{w',k'}}{\sqrt{n\sigma_v^2}} \right\rangle &\geq \frac{\sigma_v^2 + (1-\alpha)Z + Z_2}{\sqrt{Z_1\sigma_v^2}} \\ &= \beta(Z, Z_1, Z_2). \end{aligned} \tag{82}$$

If a random vector $\boldsymbol{S}$ is uniformly distributed on an n-dimensional sphere, and if another vector $\boldsymbol{T}$ is independent of it and also takes value in that n-sphere, then, by symmetry, the inner

product $\langle \boldsymbol{S}, \boldsymbol{T} \rangle$ has a distribution that does not depend on the distribution of $\boldsymbol{T}$. We next use this observation to analyze the left hand side (LHS) of (82).

The random vector $\boldsymbol{V}_{w',k'}/\sqrt{n\sigma_v^2}$ is uniformly distributed on the unit n-sphere $S^n(0,1)$ and is independent of $\boldsymbol{Y}$, Z, Z_1, and Z_2. Indeed, the encoder does not examine the auxiliary codewords in bins other than in the one corresponding to the message W. The random vector $\boldsymbol{Y}/\sqrt{nZ_1}$ also takes value on the unit n-sphere $S^n(0,1)$, and thus, by the argument above, the distribution of the LHS of (82) does not depend on the distribution of $\boldsymbol{Y}$. In particular, for any $w' \neq W$,

$$\Pr\left(\left\langle \frac{\boldsymbol{Y}}{\sqrt{nz_1}}, \frac{\boldsymbol{V}_{w',k'}}{\sqrt{n\sigma_v^2}} \right\rangle \geq \beta(z,z_1,z_2) \Big| Z=z, Z_1=z_1, Z_2=z_2\right) = \frac{C_n\big(\arccos\beta(z,z_1,z_2)\big)}{C_n(\pi)}. \tag{83}$$

Furthermore, the random vectors $\{\boldsymbol{V}_{w',k'} : w' \neq W, 1 \leq k' \leq 2^{nR_0}\}$ are independent of each other. Thus, the probability of *no* error is given by the product of the probabilities that each of these $2^{nR_1} - 2^{nR_0}$ vectors does *not* cause an error. Since the probability of error for each individual vector is given in (83), we can write the conditional probability of error for this coding strategy as

$$\Pr(\text{error}|Z=z, Z_1=z_1, Z_2=z_2) = 1 - \left(1 - \frac{C_n\big(\arccos\beta(z,z_1,z_2)\big)}{C_n(\pi)}\right)^{2^{nR_1}-2^{nR_0}}. \tag{84}$$

The expression $\Pr(\text{error}|Z=z, Z_1=z_1, Z_2=z_2)$ is a monotonically non-increasing function of $\beta(z,z_1,z_2)$ and is upper-bounded by 1. Consequently,

$$\Pr(\text{error}) \leq \Pr\big(\text{error}|\beta(Z,Z_1,Z_2)=\Upsilon\big) + \Pr\big(\beta(Z,Z_1,Z_2)<\Upsilon\big), \tag{85}$$

for any real number Υ. For both games under consideration, we will show that, by choosing a sufficiently large blocklength n, the RHS of (85) can be made arbitrarily small when $\Upsilon = \beta^*(R_{1,\text{type}}+\delta) - \epsilon_1$. Here

$$\beta^*(R_1+\delta) = \left(1 - 2^{-2(R_1+\delta)}\right)^{1/2}, \tag{86}$$

ϵ_1 is a small number to be specified later, and $R_{1,\text{type}}$ is either $R_{1,\text{add}}$ of (66) or $R_{1,\text{gen}}$ of (71) depending on whether we are considering an additive attack or a general attack.

We now analyze the terms on the RHS of (85) using a series of lemmas. In Lemma 4.2, we show that the first term on the RHS of (85) can be made arbitrarily small. In Lemma 4.4, we perform similar analysis for the second term.

Lemma 4.2. *For any $\epsilon > 0$, there exists some $\epsilon_1 > 0$ and some integer $n_1 > 0$ such that for all $n > n_1$*

$$1 - \left(1 - \frac{C_n\Big(\arccos\big(\beta^*(R_{1,\text{type}} + \delta) - \epsilon_1\big)\Big)}{C_n(\pi)}\right)^{2^{nR_{1,\text{type}}} - 2^{nR_{0,\text{type}}}} < \epsilon,$$

where type *is either* add *or* gen.

Proof. There exists some $\epsilon_1 > 0$ such that

$$\frac{1}{2}\log\left(\frac{1}{1 - \big(\beta^*(R_{1,\text{type}} + \delta) - \epsilon_1\big)^2}\right) > R_{1,\text{type}}. \tag{87}$$

This follows since the LHS of (87) equals $R_{1,\text{type}} + \delta$ when $\epsilon_1 = 0$ (see (86)) and since in both (66) and (71) the rate $R_{1,\text{type}}$ satisfies $0 < \beta^*(R_{1,\text{type}} + \delta) < 1$. By the result on the asymptotic area of spherical caps (54) and by the inequality (87), it follows by Lemma 4.1 that there exists a positive integer n_1 such that for all $n > n_1$

$$\left(1 - \frac{C_n\Big(\arccos\big(\beta^*(R_{1,\text{type}} + \delta) - \epsilon_1\big)\Big)}{C_n(\pi)}\right)^{2^{nR_{1,\text{type}}}} > 1 - \epsilon,$$

and the lemma follows by noting that the LHS cannot decrease when the exponent $2^{nR_{1,\text{type}}}$ is replaced by $2^{nR_{1,\text{type}}} - 2^{nR_{0,\text{type}}}$. □

In order to analyze the second term on the RHS of (85) we need a lemma that describes the behavior of the random variable Z defined in (79). This lemma will also be used to show that the encoder meets the distortion constraint with arbitrarily high probability; see Section 4.4.2.

Lemma 4.3. *For every $\delta > 0$ used to define the encoder, there exists $\epsilon(\delta) > 0$ such that*

$$\lim_{n\to\infty} \Pr\big(\alpha_{\text{type}}\sigma_u^2 + \rho_{\text{type}} \le Z \le \alpha_{\text{type}}\sigma_u^2 + \rho_{\text{type}} + \epsilon(\delta)\big) = 1,$$

and

$$\lim_{\delta \downarrow 0} \epsilon(\delta) = 0,$$

where type *is either* add *or* gen.

Proof. In the proof, we drop the type subscripts unless they are relevant. We first show that $\Pr(Z \geq \alpha\sigma_u^2 + \rho) \to 1$. Let $\boldsymbol{V}$ be uniformly distributed on $S^n(0, \sqrt{n\sigma_v^2})$ independent of $\boldsymbol{U}$. Then

$$\begin{aligned}\Pr(Z \geq \alpha\sigma_u^2 + \rho) &= 1 - \Pr\left(\max_{1 \leq k \leq 2^{nR_0}} n^{-1}\langle \boldsymbol{U}, \boldsymbol{V}_{W,k}\rangle < \alpha\sigma_u^2 + \rho\right) \\ &= 1 - \left(1 - \Pr\left(n^{-1}\langle \boldsymbol{U}, \boldsymbol{V}\rangle \geq \alpha\sigma_u^2 + \rho\right)\right)^{2^{nR_0}}, \end{aligned} \tag{88}$$

where the first equality follows by the definition of Z (79) and of $\boldsymbol{V}_{W,k}$, and the second equality follows because $\{\boldsymbol{V}_{W,k}\}_{k=1}^{2^{nR_0}}$ are i.i.d. and also independent of $\boldsymbol{U}$. The RHS of (88) can be further simplified using

$$\begin{aligned}\Pr\left(\frac{1}{n}\langle \boldsymbol{U}, \boldsymbol{V}\rangle \geq \alpha\sigma_u^2 + \rho\right) &= \Pr\left(\left\langle \frac{\boldsymbol{U}}{\sqrt{n\sigma_u^2}}, \frac{\boldsymbol{V}}{\sqrt{n\sigma_v^2}}\right\rangle \geq \frac{\alpha\sigma_u^2 + \rho}{\sigma_u\sigma_v}\right) \\ &= \frac{C_n\left(\arccos\left(\frac{\alpha\sigma_u^2+\rho}{\sigma_u\sigma_v}\right)\right)}{C_n(\pi)}, \end{aligned} \tag{89}$$

which follows since both normalized random vectors are uniformly distributed on $S^n(0,1)$ and they are independent of each other. By (54) we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{C_n\left(\arccos\left(\frac{\alpha\sigma_u^2+\rho}{\sigma_u\sigma_v}\right)\right)}{C_n(\pi)} = \frac{1}{2}\log\left(1 - \frac{(\alpha\sigma_u^2 + \rho)^2}{\sigma_u^2\sigma_v^2}\right), \tag{90}$$

where for both types,

$$\frac{1}{2}\log\left(1 - \frac{(\alpha_{\text{type}}\sigma_u^2 + \rho_{\text{type}})^2}{\sigma_u^2\sigma_{v,\text{type}}^2}\right) = -(R_{0,\text{type}} - \delta). \tag{91}$$

To verify the final equality, see the relevant definitions for each type in Sections 4.3.1 and 4.3.2. Combining Lemma 4.1 with (88), (89), (90) and (91) demonstrates that $\Pr(Z \geq \alpha\sigma_u^2 + \rho) \to 1$.

To complete the proof, we find an appropriate choice of $\epsilon(\delta)$ such that $\Pr(Z > \alpha\sigma_u^2 + \rho + \epsilon(\delta)) \to 0$. We choose $\epsilon(\delta) > 0$ such that

$$\frac{1}{2}\log\left(1 - \left(\frac{\alpha\sigma_u^2 + \rho + \epsilon(\delta)}{\sigma_u\sigma_v}\right)^2\right) < -R_0. \tag{92}$$

This can be done because the LHS of (92) equates to $-(R_0 + \delta)$ when $\epsilon(\delta)$ is set to zero (as we

have seen in (91)), and because $\log(1-x^2)$ is continuous and decreasing in x, for $0 < x < 1$. Using Lemma 4.1, we see that $\Pr(Z > \alpha\sigma_u^2 + \rho + \epsilon(\delta)) \to 0$. Finally, we can choose $\epsilon(\delta) \to 0$ as $\delta \to 0$ by the continuity of $\log(1-x^2)$. □

We now proceed to show that the second term on the RHS of (85) is vanishing in n when $\Upsilon = \beta^*(R_{1,\text{type}} + \delta) - \epsilon_1$. Here, $R_{1,\text{type}}$ is defined in (66) and (71), $\beta^*(\cdot)$ is defined in (86), and $\epsilon_1 > 0$ is specified in Lemma 4.2. The combination of this fact with Lemma 4.2 will show that, as the blocklength n tends to infinity, the probability of decoding error approaches zero. The proof of this lemma can be found in Appendix A.4.

Lemma 4.4. *For any $\epsilon > 0$ and $\epsilon_1 > 0$, there exists an integer $n_2 > 0$ such that for the sequence of encoders of Section 4.3.3 and for all deterministic attacks of Section 4.1.2,*

$$\Pr\big(\beta_{\text{type}}(Z, Z_1, Z_2) < \beta^*(R_{1,\text{type}} + \delta) - \epsilon_1\big) < \epsilon, \quad \textit{for all } n > n_2$$

where type *is either* add *or* gen.

4.4.2 The Encoding Distortion Constraint

We now show that the encoder's distortion constraint is met with arbitrarily high probability. We claim that this is sufficient since the encoder can modify its behavior on an event of vanishing probability to meet the a.s. distortion constraint without significantly affecting the probability of error.

Let us first assume that $\alpha_{\text{type}} \geq 0$, which is always true for α_{add} and usually true for α_{gen}. By Lemma 4.3, it is sufficient to show that $Z \geq \alpha\sigma_u^2 + \rho$ implies $n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 \leq D_1$, which we proceed to prove. By the definitions of $\boldsymbol{X}$ and Z (see (76) and (79)),

$$n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 = \sigma_v^2 - 2\alpha Z + \alpha^2\sigma_u^2. \tag{93}$$

Since α is assumed to be positive, the RHS of (93) is decreasing in Z. Consequently, the condition $Z \geq \alpha\sigma_u^2 + \rho$ implies

$$n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 \leq \sigma_v^2 - \alpha^2\sigma_u^2 - 2\alpha\rho = D_1,$$

where the equality follows from (73).

Let us now address the case when $\alpha < 0$, which can only occur for the encoder designed for a general attacker. Note that whenever the inequality $\alpha\sigma_u^2 + \rho \leq Z \leq \alpha\sigma_u^2 + \rho + \epsilon(\delta)$ holds we also have $n^{-1}\|\boldsymbol{X}-\boldsymbol{U}\|^2 \leq D_1 - 2\alpha\epsilon(\delta)$. Thus, if we design our system for some $\tilde{D}_1 < D_1$ instead of D_1 as the encoder's distortion constraint, then by choosing δ sufficiently enough and n sufficiently large, Lemma 4.3 will guarantee that the encoder will meet the D_1 distortion constraint with arbitrarily high probability. The desired achievability result can be demonstrated by letting $\tilde{D}_1$ approach D_1, because $C^*(D_1, D_2, \sigma_u^2)$ is continuous in D_1.

4.4.3 Extension to i.i.d. Gaussian Covertext

We have shown that if the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $S^n(0, \sqrt{n\sigma_u^2})$, then for the public version, the coding capacity of the general watermarking game is lower bounded by $C^*(D_1, D_2, \sigma_u^2)$ and the coding capacity of the additive attack watermarking game is lower bounded by $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$. In this section, we extend these results to zero-mean variance-σ_u^2 i.i.d. Gaussian covertexts.

We first transform the i.i.d. Gaussian sequence $\boldsymbol{U}$ into a random vector $\boldsymbol{U}'$ which is uniformly distributed on the n-sphere $S^n(0, \sqrt{n\sigma_u^2})$. To this end we set $S_{\boldsymbol{U}} = n^{-1}\|\boldsymbol{U}\|^2$, which converges to σ_u^2 in probability, and let $\boldsymbol{U}' = \sqrt{\frac{\sigma_u^2}{S_{\boldsymbol{U}}}}\,\boldsymbol{U}$, which is well defined with probability 1, and which is uniformly distributed on $S^n(0, \sqrt{n\sigma_u^2})$. We will consider all the models simultaneously, but we will state our assumptions on the rate of each of the models separately:

General watermarking Assume that $0 < R < C^*(D_1, D_2, \sigma_u^2)$. By the definition of C^* (8), there exists some $A' \in \mathcal{A}(D_1, D_2, \sigma_u^2)$ such that $R < \frac{1}{2}\log(1 + s(A'; D_1, D_2, \sigma_u^2))$. Since $s(A'; D_1, D_2, \sigma_u^2)$ is continuous in D_1, there exists some $D_1' < D_1$ such that $R < \frac{1}{2}\log(1 + s(A'; D_1', D_2, \sigma_u^2))$.

Additive attack watermarking Assume that $0 < R < \frac{1}{2}\log(1 + \frac{D_1}{D_2})$. Then, there exists a $D_1' < D_1$ such that $R < \frac{1}{2}\log(1 + \frac{D_1'}{D_2})$.

Let $\boldsymbol{X}'$ be the output of the encoders as designed for the covertext $\boldsymbol{U}'$ and the parameters A' and D_1' in Sections 4.3.3. Let ϕ' be the corresponding decoder. Consider now an encoder for the

covertext $\boldsymbol{U}$ which produces the stegotext $\boldsymbol{X}$ according to the rule

$$\boldsymbol{x} = \begin{cases} \boldsymbol{x}' & \text{if } n^{-1}\|\boldsymbol{x}' - \boldsymbol{u}\|^2 \leq D_1 \\ \boldsymbol{u} & \text{otherwise} \end{cases}.$$

With this choice of $\boldsymbol{x}$, the distortion between $\boldsymbol{u}$ and $\boldsymbol{x}$ is less than D_1 almost surely, so that the encoding distortion constraint (9) is met.

We next claim that for a sufficiently large blocklength, $\boldsymbol{X} = \boldsymbol{X'}$ with arbitrarily high probability. Indeed, the distortion between the random vectors $\boldsymbol{X'}$ and $\boldsymbol{U}$ is given by

$$\begin{aligned} \frac{1}{n}\|\boldsymbol{X'} - \boldsymbol{U}\|^2 &= \frac{1}{n}\|\boldsymbol{X'} - \boldsymbol{U'} + \boldsymbol{U'} - \boldsymbol{U}\|^2 \\ &\leq \frac{1}{n}\|\boldsymbol{X'} - \boldsymbol{U'}\|^2 + \frac{1}{n}\|\boldsymbol{U'} - \boldsymbol{U}\|^2 + \frac{2}{n}\|\boldsymbol{X'} - \boldsymbol{U'}\| \cdot \|\boldsymbol{U'} - \boldsymbol{U}\| \\ &\leq D_1' + \frac{1}{n}\|\boldsymbol{U'} - \boldsymbol{U}\|^2 + \sqrt{D_1'}\frac{2}{n}\|\boldsymbol{U'} - \boldsymbol{U}\|, \end{aligned}$$

and $\frac{1}{n}\|\boldsymbol{U'} - \boldsymbol{U}\|^2 = \left(\sqrt{S_{\boldsymbol{U}}} - \sqrt{\sigma_u^2}\right)^2$ approaches, by the weak law of large numbers, zero in probability. In the above, the first inequality follows from the triangle inequality, and the second because the encoders of Sections 4.3.3 satisfy the encoder distortion constraint $n^{-1}\|\boldsymbol{X'} - \boldsymbol{U'}\|^2 \leq D_1'$ almost surely. Since $D_1' < D_1$, our claim that

$$\lim_{n\to\infty} \Pr(\boldsymbol{X} = \boldsymbol{X'}) = 1 \tag{94}$$

is proved.

Let $\hat{W}$ be the output of the decoder ϕ', and consider now any fixed deterministic attack. The probability of error can be written as

$$\begin{aligned} \Pr(\hat{W} \neq W) &= \Pr(\hat{W} \neq W, \boldsymbol{X} = \boldsymbol{X'}) + \Pr(\hat{W} \neq W, \boldsymbol{X} \neq \boldsymbol{X'}) \\ &\leq \Pr(\hat{W} \neq W, \boldsymbol{X} = \boldsymbol{X'}) + \Pr(\boldsymbol{X} \neq \boldsymbol{X'}), \end{aligned}$$

where the second term on the RHS of the above converges to zero (uniformly over all the deterministic attackers) by (94), and the first term approaches zero by the achievability results for covertexts that are uniformly distributed over the n-sphere. To clarify the final argument consider, for example, the public watermarking game with an additive attacker as in (57). We would then

argue that

$$\begin{aligned}\Pr(\hat{W} \neq W, \boldsymbol{X} = \boldsymbol{X}') &= \Pr\bigl(\phi'(\Theta_1, \boldsymbol{X} + \tilde{\boldsymbol{y}}) \neq W, \boldsymbol{X} = \boldsymbol{X}'\bigr) \\ &= \Pr\bigl(\phi'(\Theta_1, \boldsymbol{X}' + \tilde{\boldsymbol{y}}) \neq W, \boldsymbol{X} = \boldsymbol{X}'\bigr) \\ &\leq \Pr\bigl(\phi'(\Theta_1, \boldsymbol{X}' + \tilde{\boldsymbol{y}}) \neq W\bigr),\end{aligned}$$

which converges to zero by the achievability result on covertexts that are uniformly distributed on the n-sphere.

4.5 Discussion: Common Randomness

There is a difference between the randomized coding used here and Shannon's classical random coding argument (see, for example, [17, Chap. 8.7]). In the latter, codebooks are chosen from an ensemble according to some probability law, and it is shown that the ensemble-averaged probability of error is small, thus demonstrating the existence of at least one codebook from the ensemble for which the probability of error is small. For the watermarking game, on the other hand, randomization is not a proof technique that shows the existence of a good codebook, but a defining feature of the encoding. For example, the randomization at the encoder prevents the attacker from knowing the particular mapping used for each message; the attacker only knows the strategy used for generating the codewords. See [29] for more on this subject.

Nevertheless, in the private version of the i.i.d. Gaussian watermarking game, common randomness is not needed between the encoder and the decoder and deterministic codes suffice. Indeed, part of the covertext, to which both the encoder and the decoder have access, can be used instead of the secret key Θ_1. For example, the encoder could set $x_1 = 0$, and use the random variable U_1 as the common random experiment. The extra distortion incurred by this policy can be made arbitrarily small by making n sufficiently large. Since U_1 is a real-valued random variable with a density, it is sufficient to provide the necessary randomization.

5 Converses for a.s. Constraints

In this section, we prove the converse parts of Theorem 2.3 (additive attack) and Theorem 2.1 (general attack). The former is fairly straightforward and is described in Section 5.1. For the

latter we need to show that if the covertext distribution $\{P_{\boldsymbol{U}}\}$ is ergodic with finite fourth moment and $E\left[U_k^2\right] \leq \sigma_u^2$, then the capacity of the private version of the watermarking game is at most $C^*(D_1, D_2, \sigma_u^2)$. For any fixed $R > C^*(D_1, D_2, \sigma_u^2)$ and any sequence of rate-R encoders that satisfy the distortion constraint (9), a sequence of attackers $\{g_n\}$ is proposed in Section 5.2.2 that is shown in Section 5.2.3 to satisfy the distortion constraint (12) and is shown in Section 5.2.4 to guarantee that, irrespective of the decoding rule, the probability of error is bounded away from zero. Thus, even if the sequence of decoders were designed with full knowledge of this sequence of attackers, no rate above $C^*(D_1, D_2, \sigma_u^2)$ would be achievable. Although we argued that deterministic attacks are sufficient in Section 4.1.2, we prove the converses using a randomized attacks in much the same way that randomized strategies are used to prove coding theorems. We conclude in Section 5.3 with a discussion of the necessity of the ergodicity assumption.

5.1 Additive Attack

In this section, we describe an additive attacker that demonstrates that no rates greater than $\frac{1}{2}\log\left(1+\frac{D_1}{D_2}\right)$ are achievable in either the public or private versions of the additive-attack watermarking game. This will prove the converse part of Theorem 2.3.

The attacker first chooses a $\delta > 0$. The attacker initially generates its additive attack $\tilde{\boldsymbol{Y}}$ as a sequence of i.i.d. mean-zero, variance $(D_2 - \delta)$ Gaussian random variables. This sequence satisfies the distortion constraint (19) with arbitrarily high probability; when it does not, the attacker sets $\tilde{\boldsymbol{Y}} = 0$. If for the initial definition of $\tilde{\boldsymbol{Y}}$, the probability of error is bounded away from zero and is furthermore greater than the probability of modification, then the final probability of error is also bounded away from zero. Choosing the blocklength n large enough (depending on δ), the above necessary condition is satisfied by the law of large numbers. With the initial definition of $\tilde{\boldsymbol{Y}}$, Costa [2] showed that no rates larger than $\frac{1}{2}\log\left(1+\frac{D_1}{D_2-\delta}\right)$ are achievable for the public version (and hence for the private version). Since rates that are not achievable for the initial definition of $\tilde{\boldsymbol{Y}}$ are also not achievable for the final definition and since the attacker can choose δ arbitrarily small, no rates larger than $\frac{1}{2}\log\left(1+\frac{D_1}{D_2}\right)$ are achievable.

5.2 General Attack

5.2.1 Intuitive Definition of Attacker

We seek to provide some motivation for the proposed attack strategy by first describing two simple attacks that fail to give the desired converse. We then combine aspects of these simple strategies to form the attack strategy that we will use to prove the converse.

The upcoming discussion will utilize the correspondence between the encoder and attacker (mappings) (f_n, g_n) and the watermarking and attack channels (conditional laws) ($P_{\boldsymbol{X}|\boldsymbol{U}}$, $P_{\boldsymbol{Y}|\boldsymbol{X}}$) that they induce for given fixed laws on W, $\{P_{\boldsymbol{U}}\}$, Θ_1, and Θ_2. One way to prove the converse is to show using a Fano-type inequality that in order for the probability of error to tend to zero, a mutual information term similar to I_{priv} of (26) — evaluated with respect to the induced channels — must be greater than the watermarking rate. Thus, one would expect that the optimal attack channels of Section 3.1 for the mutual information games could be used to design good attacker mappings for the watermarking game.

The first simple attack strategy corresponds to the optimal attack channel $(P^A_{Y|X})^n$ of Section 3.1, where A is the average power in the stegotext based on the encoder, i.e., $A = E\left[n^{-1}\|\boldsymbol{X}\|^2\right]$. Since the encoder must satisfy the distortion constraint (9) (and thus the corresponding watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$ must be in $\mathcal{D}_1(D_1, P_{\boldsymbol{U}})$), the results of Section 3.3 show that this attacker guarantees that the mutual information is at most $C^*(D_1, D_2, \sigma_u^2)$. The problem with this attack strategy is that since it is based on the average power in the stegotext, there is no guarantee that the attacker's distortion constraint (12) will be met with probability one.

The second simple attack strategy corresponds to the optimal attack channel $(P^a_{Y|X})^n$, where a is the power in the *realization* (sample-path) of the stegotext, i.e., $a = n^{-1}\|\boldsymbol{x}\|^2$. The results of Section 3.3 again give the appropriate upper bound on the mutual information conditioned on the value of a. Furthermore, if a distortion level $\tilde{D}_2$ slightly smaller than the actual distortion level D_2 is used to design this attacker, then the distortion constraint will be met with high probability. The problem with this attack strategy is that the decoder can fairly accurately determine the value of a from the forgery. Thus, the encoder and decoder could potentially use the power of the stegotext to send extra information, so that the total rate might be higher than $C^*(D_1, D_2, \sigma_u^2)$.

The attack strategy that we use to prove the converse combines aspects of the two simple strategies described above. To form this attacker, we partition the possible values of $a = n^{-1}\|\boldsymbol{x}\|^2$

into a finite number of intervals, $\mathcal{A}_1, \ldots, \mathcal{A}_m$, and compute the average power in the stegotext conditioned on each interval, i.e., $a_k = E\left[n^{-1}\|\boldsymbol{X}\|^2 \,\middle|\, n^{-1}\|\boldsymbol{X}\|^2 \in \mathcal{A}_k\right]$. We then use the optimal attack channel $(P_{Y|X}^{a_k})^n$ whenever the actual power of the stegotext lies in the interval $\mathcal{A}_k$. Unlike the first simple strategy, the distortion constraint can be guaranteed by making the intervals small enough. Unlike the second simple strategy, the encoder and decoder cannot use the power of the stegotext to transmit extra information because the number of intervals is finite and does not grow with the blocklength n. These arguments will be made more precise in the upcoming sections.

5.2.2 Precise Definition of Attacker

Let R be a fixed rate which is strictly larger than $C^*(D_1, D_2, \sigma_u^2)$. For any rate-$R$ sequence of encoders and decoders, the attacker described below will guarantee some non-vanishing probability of error.

By the continuity of $C^*(D_1, D_2, \sigma_u^2)$ in D_2, it follows that there exists some $0 < \tilde{\delta} < D_2$ such that $R > C^*(D_1, D_2 - \tilde{\delta}, \sigma_u^2)$. Let

$$\tilde{D}_2 = D_2 - \tilde{\delta}, \tag{95}$$

for some such $\tilde{\delta}$. The attacker partitions the interval $\left(\tilde{D}_2, (2\sigma_u + \sqrt{D_1})^2\right)$ sufficiently finely into m sub-intervals $\mathcal{A}_1, \ldots, \mathcal{A}_m$, so that for each sub-interval $\mathcal{A}_k$,

$$\tilde{D}_2 \left(1 + \frac{\tilde{D}_2}{A}\left(\frac{A'}{A} - 1\right)\right) < \tilde{D}_2 + \frac{\tilde{\delta}}{2}, \ \forall A, A' \in \mathcal{A}_k. \tag{96}$$

Such a partition exists because this interval is finite, it does not include zero ($\tilde{D}_2 > 0$), and because the constant $\tilde{\delta}$ is positive.

We define the mapping k from $\mathbb{R}^n$ to $\{0, \ldots, m\}$ as

$$k(\boldsymbol{x}) = \begin{cases} l & \text{if } n^{-1}\|\boldsymbol{x}\|^2 \in \mathcal{A}_l \\ 0 & \text{if no such } l \text{ exists} \end{cases}. \tag{97}$$

This mapping will determine how the stegotext $\boldsymbol{x}$ will be attacked. Notice that it takes on a finite number of values. We also define the random variable $K = k(\boldsymbol{X})$. Using his knowledge of the

distribution of the covertext and the encoder mapping, the attacker computes

$$a_k = E\left[\frac{1}{n}\|\boldsymbol{X}\|^2 \Big| K = k\right],\ \forall\, 0 \le k \le m. \tag{98}$$

Note that $a_k \in \mathcal{A}_k$ for $k \neq 0$ since $\mathcal{A}_k$ is an interval (and hence convex) and since the event $K = k$ corresponds to the event $n^{-1}\|\boldsymbol{X}\|^2 \in \mathcal{A}_k$. The attacker also computes

$$\mu_k = E\left[\frac{1}{n}\|\boldsymbol{U}\|^2 \Big| K = k\right],\ \forall\, 0 \le k \le m. \tag{99}$$

Using only the source of randomness Θ_2, the attacker generates a random vector $\boldsymbol{V}$ as a sequence of i.i.d. zero-mean variance-$\tilde{D}_2$ Gaussian random variables. Recall that we assume that the random variable Θ_2 and the random vector $\boldsymbol{X}$ are independent, and thus the random vectors $\boldsymbol{V}$ and $\boldsymbol{X}$ are also independent.

Let us now consider an attacker g_n^* in which the forgery is computed as

$$g_n^*(\boldsymbol{x}, \theta_2) = \begin{cases} c(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{x} + c^{1/2}(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{v}(\theta_2) & \text{if } k(\boldsymbol{x}) > 0 \\ \left(\sqrt{nD_2} - \sqrt{n\tilde{D}_2}\right)\boldsymbol{v}(\theta_2)/\|\boldsymbol{v}(\theta_2)\| & \text{otherwise} \end{cases}, \tag{100}$$

where $c(A; D_2)$ is defined in (5). Conditionally on $\boldsymbol{X} = \boldsymbol{x}$ satisfying $k(\boldsymbol{x}) \ge 1$, the random vector $\boldsymbol{Y} = g_n^*(\boldsymbol{x}, \Theta_2)$ under this attacker is thus distributed as $c(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{x} + c^{1/2}(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{V}$. Note that if $K = k > 0$, the resulting conditional distribution $P_{\boldsymbol{Y}|\boldsymbol{X}}$ is the same as the optimal attack channel of the mutual information game corresponding to a_k and $\tilde{D}_2$; see Section 3.1.

Finally, our proposed attacker uses g_n^* if the distortion constraint is met and sets $\boldsymbol{y} = \boldsymbol{x}$ if the distortion constraint is not met. That is,

$$g_n(\boldsymbol{x}, \theta_2) = \begin{cases} g_n^*(\boldsymbol{x}, \theta_2) & \text{if } n^{-1}\|g_n^*(\boldsymbol{x}, \theta_2) - \boldsymbol{x}\|^2 \le D_2 \\ \boldsymbol{x} & \text{otherwise} \end{cases}. \tag{101}$$

The attacker g_n thus satisfies the distortion constraint with probability one.

5.2.3 Analysis of Distortion

The attackers $\{g_n^*\}$ do not, in general, satisfy the distortion constraint (12). But in this section we show that, as the blocklength tends to infinity, the probability that the distortion they introduce exceeds D_2 tends to zero, i.e., that

$$\lim_{n\to\infty} \Pr\big(g_n(\boldsymbol{X},\Theta_2) = g_n^*(\boldsymbol{X},\Theta_2)\big) = 1. \tag{102}$$

Using this property, we will be able to complete the converse in the next section.

We now turn to the proof of (102). In order to summarize the distortion introduced by the attacker, we define the following random variables,

$$\Delta_1(k) = c(a_k;\tilde{D}_2)\left(n^{-1}\|\boldsymbol{V}\|^2 - \tilde{D}_2\right), \quad k = 1,\ldots,m, \tag{103}$$

and

$$\Delta_2(k) = \Big(c(a_k;\tilde{D}_2) - 1\Big)\, c^{1/2}(a_k;\tilde{D}_2) n^{-1}\langle \boldsymbol{X},\boldsymbol{V}\rangle, \quad k = 1,\ldots,m. \tag{104}$$

Note that for any $1 \le k \le m$, the random variables $\Delta_1(k)$ and $\Delta_2(k)$ converge to zero in probability, because $\boldsymbol{V}$ is a sequence of i.i.d. $\mathcal{N}(0,\tilde{D}_2)$ random variables independent of $\boldsymbol{X}$, and because $0 < c(a_k;\tilde{D}_2) < 1$ for all $1 \le k \le m$. The probability of exceeding the allowed distortion can be written as

$$\Pr\left(\frac{1}{n}\|g_n^*(\boldsymbol{X},\Theta_2) - \boldsymbol{X}\|^2 > D_2\right) = \sum_{l=0}^{m} \Pr\left(\frac{1}{n}\|g_n^*(\boldsymbol{X},\Theta_2) - \boldsymbol{X}\|^2 > D_2, K = l\right).$$

We shall next show that each of the terms in the above sum converges to zero in probability. We begin with the first term, namely $l = 0$. The event $K = 0$ corresponds to either $n^{-1}\|\boldsymbol{X}\|^2 \le \tilde{D}_2$ or $n^{-1}\|\boldsymbol{X}\|^2 > (2\sigma_u + \sqrt{D_1})^2$. In the former case,

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\|^2 &= \frac{1}{n}\left\|\left(\sqrt{nD_2} - \sqrt{n\tilde{D}_2}\right)\boldsymbol{V}/\|\boldsymbol{V}\| - \boldsymbol{X}\right\|^2 \\
&\le \left(\sqrt{D_2} - \sqrt{\tilde{D}_2}\right)^2 + 2\left(\sqrt{D_2} - \sqrt{\tilde{D}_2}\right)\sqrt{\tilde{D}_2} + \tilde{D}_2 \\
&= D_2,
\end{aligned}$$

where the inequality follows by the triangle inequality and since $n^{-1}\|\boldsymbol{X}\|^2 \leq \tilde{D}_2$ here. Thus,

$$\begin{aligned}\Pr\left(\frac{1}{n}\|g_n^*(\boldsymbol{X},\Theta_2)-\boldsymbol{X}\|^2 > D_2, K=0\right) &= \Pr\left(n^{-1}\|\boldsymbol{X}\|^2 > (2\sigma_u+\sqrt{D_1})^2\right)\\ &\leq \Pr\left(n^{-1}\|\boldsymbol{U}\|^2 > 4\sigma_u^2\right),\end{aligned}$$

which converges in probability to zero by the ergodicity of the covertext. To study the limiting behavior of the rest of the terms, fix some $1 \leq l \leq m$. If $k(\boldsymbol{x}) = l$ then

$$\begin{aligned}\frac{1}{n}\|g_n^*(\boldsymbol{x},\theta_2)-\boldsymbol{x}\|^2 &= \frac{1}{n}\left\|\left(c(a_l;\tilde{D}_2)-1\right)\boldsymbol{x} + c^{1/2}(a_l;\tilde{D}_2)\boldsymbol{V}\right\|^2\\ &= \tilde{D}_2\left(1+\frac{\tilde{D}_2}{a_l}\left(\frac{n^{-1}\|\boldsymbol{x}\|^2}{a_l}-1\right)\right)+\Delta_1(l)+\Delta_2(l)\\ &\leq D_2 - \frac{\tilde{\delta}}{2}+\Delta_1(l)+\Delta_2(l),\end{aligned}$$

where the second equality follows by the definitions of c, $\Delta_1(l)$, and $\Delta_2(l)$ (see (5), (103) and (104)), and the inequality follows by (96) since both $n^{-1}\|\boldsymbol{x}\|^2$ and a_l are in the set $\mathcal{A}_l$. Thus,

$$\begin{aligned}\Pr\left(\frac{1}{n}\|g_n^*(\boldsymbol{X},\Theta_2)-\boldsymbol{X}\|^2 > D_2,\ K=l\right) &\leq \Pr\left(\Delta_1(l)+\Delta_2(l) \geq \tilde{\delta}/2,\ K=l\right)\\ &\leq \Pr\left(\Delta_1(l)+\Delta_2(l) \geq \tilde{\delta}/2\right),\end{aligned}$$

which converges to zero because both $\Delta_1(l)$ and $\Delta_2(l)$ converge to zero in probability.

5.2.4 Analysis of Probability of Error

In this section, we show that whenever the watermarking rate R exceeds $C^*(D_1, D_2, \sigma_u^2)$, the sequence of attackers $\{g_n^*\}$ defined in (100) prevents the probability of error from decaying to zero. In the previous section, we have shown that for blocklength n large enough $g_n(\boldsymbol{X},\Theta_2) = g_n^*(\boldsymbol{X},\Theta_2)$ with arbitrarily high probability. The combination of these two facts will show that the probability of error is also prevented from decaying to zero by the sequence of attackers $\{g_n\}$ defined in (101). To see this, fix any $R > C^*(D_1, D_2, \sigma_u^2)$ and fix some encoder sequence $\{f_n\}$ and a corresponding decoder sequence $\{\phi_n\}$. Let $\tilde{D}_2$ be chosen as in (95) so that $R > C^*(D_1, \tilde{D}_2, \sigma_u^2)$ and consider the attacker (100). Assume that we have managed to prove that the attackers $\{g_n^*\}$ of (100) guarantee a non-vanishing probability of error. In this case (102) will guarantee that the probability of error

must also be bounded away from zero in the presence of the attacker g_n. Since $\{g_n\}$ do satisfy the distortion constraint, this will conclude the proof of the converse.

The probability of error analysis for $\{g_n^*\}$ is carried out in a series of lemmas. In Lemma 5.1 we use a Fano-type inequality to show that an achievable rate cannot exceed some limit of mutual informations. In Lemma 5.2, we upper bound these mutual informations by simpler expectations, and in Lemma 5.3 we finally show that, in the limit, these expectations do not exceed $C^*(D_1, D_2, \sigma_u^2)$.

Lemma 5.1. *For any sequence of encoders, attackers, and decoders $\{(f_n, g_n, \phi_n)\}$ with corresponding sequence of conditional distributions $\{(P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1}, P_{\boldsymbol{Y}|\boldsymbol{X}})\}$, if $\bar{P}_e(f_n, g_n, \phi_n) \to 0$ as $n \to \infty$, then*

$$R \leq \liminf_{n\to\infty} \frac{1}{n} I_{P_{\boldsymbol{U}} P_{\Theta_1} P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1} P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{U}, \Theta_1). \tag{105}$$

Proof. Utilizing Fano's inequality and the data processing theorem,

$$\begin{aligned} nR &= H(W \mid \boldsymbol{U}, \Theta_1) \\ &= H(W \mid \boldsymbol{U}, \Theta_1, \boldsymbol{Y}) + I(W; \boldsymbol{Y} \mid \boldsymbol{U}, \Theta_1) \\ &\leq 1 + nR\bar{P}_e(f_n, g_n, \phi_n) + I(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{U}, \Theta_1), \end{aligned}$$

where the first equality follows since W is independent of $(\boldsymbol{U}, \Theta_1)$ and uniformly distributed over $\{1, \ldots, 2^{nR}\}$, and the inequality follows by the data processing theorem and by Fano's inequality. Dividing by n and taking the lim inf, yields the desired result. □

The mutual information term in the RHS of (105) is a little cumbersome to manipulate, and we next exploit the fact that K takes on at most $m+1$ possible values to prove that $n^{-1}I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}, \Theta_1)$ has the same limiting behavior as $n^{-1}I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1)$, i.e., that

$$\lim_{n\to\infty} \frac{1}{n}\Big(I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}, \Theta_1) - I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1)\Big) = 0. \tag{106}$$

To prove (106) write

$$\begin{aligned} I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) &= h(\boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, K, \boldsymbol{U}, \Theta_1) \\ &= h(\boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{U}, \Theta_1) \\ &= I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}, \Theta_1) - I(K; \boldsymbol{Y}|\boldsymbol{U}, \Theta_1), \end{aligned}$$

where all differential entropies exist for the attacker g_n^*, and the second equality follows since K is a function of $\boldsymbol{X}$ (97). Thus, the mutual information on the RHS of (105) can be written as

$$I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U},\Theta_1) = I(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1) + I(K;\boldsymbol{Y}|\boldsymbol{U},\Theta_1). \tag{107}$$

Since K takes on at most $m+1$ different values, it follows that

$$0 \leq I(K;\boldsymbol{Y}|\boldsymbol{U},\Theta_1) \leq H(K) \leq \log(m+1),$$

and thus, since m is fixed and does not grow with the blocklength,

$$\lim_{n\to\infty} \frac{1}{n} I(K;\boldsymbol{Y}|\boldsymbol{U},\Theta_1) = 0. \tag{108}$$

Equation (106) now follows from (108) and (107). It now follows from Lemma 5.1 and from (106) that in order to prove that the rate R is not achievable, it suffices to show that

$$R > \liminf_{n\to\infty} \frac{1}{n} I(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1). \tag{109}$$

We upper bound the RHS of (109) in the following Lemma, which is proved in Appendix A.5.

Lemma 5.2. *For any encoder with corresponding watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$ satisfying (9), if the attacker g_n^* of (100) with corresponding attack channel $P^*_{\boldsymbol{Y}|\boldsymbol{X}}$ is used, then*

$$\frac{1}{n} I_{P_{\boldsymbol{U}} P_{\Theta_1} P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1} P^*_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1) \leq E_K\left[C^*(D_1,\tilde{D}_2,\mu_K)\right]. \tag{110}$$

To proceed with the proof of the converse we would now like to upper bound the RHS of (110). Since the function $C^*(D_1,D_2,\sigma_u^2)$ is not necessarily concave in σ_u^2, we cannot use Jensen's inequality. However, $C^*(D_1,D_2,\sigma_u^2)$ is increasing in σ_u^2 and is upper bounded by $1/2\log(1+D_1/D_2)$ for all σ_u^2. Thus, we will complete the proof by showing in the following lemma that if μ_k is larger than σ_u^2, albeit by a small constant, then $\Pr(K=k)$ must be vanishing. The proof of this lemma can be found in Appendix A.6.

Lemma 5.3. *For any ergodic covertext distribution $P_{\boldsymbol{U}}$ with $E\left[U_k^4\right] < \infty$ and $E\left[U_k^2\right] \leq \sigma_u^2$, there exists mappings $\delta(\epsilon,n)$ and $n_0(\epsilon)$ such that both the properties P1 and P2 stated below hold:*

P1. For every $\epsilon > 0$, $\lim_{n\to\infty} \delta(\epsilon, n) = 0$.

P2. For every $\epsilon > 0$, $n > n_0(\epsilon)$, *and event* $\mathcal{E}$ *of non-zero probability, if* $E\left[n^{-1}\|\boldsymbol{U}\|^2|\mathcal{E}\right] > \sigma_u^2 + 5\epsilon$, *then* $\Pr(\mathcal{E}) < \delta(\epsilon, n)$.

With the aid of Lemma 5.3 we can now upper bound the RHS of (110). Specifically, we next show that for any ergodic stegotext $\{P_{\boldsymbol{U}}\}$ of finite fourth moment and of second moment σ_u^2, if $R > C^*(D_1, \tilde{D}_2, \sigma_u^2)$ and the attacker g_n^* of (100) is used, then

$$\limsup_{n\to\infty} E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right] \le C^*(D_1, \tilde{D}_2, \sigma_u^2). \tag{111}$$

To see this, let $\delta(\epsilon, n)$ and $n_0(\epsilon)$ be the mappings of Lemma 5.3 corresponding to the stegotext $\{P_{\boldsymbol{U}}\}$. For any $\epsilon > 0$, let us define the set

$$\mathcal{K}^*(\epsilon) = \{k : \mu_k > \sigma_u^2 + 5\epsilon\}.$$

By the definition of μ_k (99), it is clear that $E\left[n^{-1}\|\boldsymbol{U}\|^2|K \in \mathcal{K}^*(\epsilon)\right] > \sigma_u^2 + 5\epsilon$. Thus, by Lemma 5.3, $\Pr(K \in \mathcal{K}^*(\epsilon)) < \delta(\epsilon, n)$. Since $C^*(D_1, D_2, \sigma_u^2)$ is non-decreasing in σ_u^2 and is upper bounded by $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$,

$$\begin{aligned}
&E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right] \\
&\quad = \Pr\left(K \notin \mathcal{K}^*(\epsilon)\right) E\left[C_K^* \middle| K \notin \mathcal{K}^*(\epsilon)\right] + \Pr\left(K \in \mathcal{K}^*(\epsilon)\right) E\left[C_K^* \middle| K \in \mathcal{K}^*(\epsilon)\right] \\
&\quad \le C^*(D_1, \tilde{D}_2, \sigma_u^2 + 5\epsilon) + \delta(\epsilon, n) \cdot \frac{1}{2}\log\left(1 + \frac{D_1}{\tilde{D}_2}\right),
\end{aligned}$$

where $C_K^* = C^*(D_1, \tilde{D}_2, \mu_K)$. Since this is true for every sufficiently large n and since $\delta(\epsilon, n)$ approaches zero as n tends to infinity,

$$\limsup_{n\to\infty} E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right] \le C^*(D_1, \tilde{D}_2, \sigma_u^2 + 5\epsilon).$$

Furthermore, since this is true for every $\epsilon > 0$ and since $C^*(D_1, D_2, \sigma_u^2)$ is continuous in σ_u^2, (111) follows.

We now have all of the necessary ingredients to prove that if the rate R exceeds $C^*(D_1, D_2, \sigma_u^2)$, then the sequence of attackers $\{g_n^*\}$ prevents the probability of error from decaying to zero. Indeed,

let $\tilde{D}_2$ be chosen as in (95) so that $R > C^*(D_1, \tilde{D}_2, \sigma_u^2)$ and consider the attacker g_n^* of (100). Then

$$\begin{aligned} R &> C^*(D_1, \tilde{D}_2, \sigma_u^2) \\ &\geq \limsup_{n\to\infty} E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right] \\ &\geq \limsup_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) \\ &= \limsup_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}, \Theta_1), \end{aligned}$$

and the probability of error must be bounded away from zero by Lemma 5.1. Here the first inequality is justified by the choice of $\tilde{D}_2$ (95), the second inequality by (111), the third inequality by (110), and the final equality by (106).

5.3 Discussion: The Ergodicity Assumption

We have proved that the i.i.d. zero-mean Gaussian covertext is easiest to watermark among all ergodic covertexts of finite fourth moment and of a given second moment. That is, we have shown that for any covertext satisfying these conditions, no rate above $C^*\big(D_1, D_2, E\left[U_i^2\right]\big)$ is achievable.

An inspection of the proof, however, reveals that full ergodicity is not required, and it suffices that the covertext law $\{P_{\boldsymbol{U}}\}$ be stationary and satisfy a second-moment ergodicity assumption, i.e., that the variance of $S_{U^2,n}$ of (153) approach zero, as n tends to infinity.

This condition can sometimes be further relaxed if the process has an ergodic decomposition (see e.g. [39]). We illustrate this point with a simple example of a covertext that has two ergodic modes. Let Z take on the values zero and one equiprobably, and assume that conditional on Z the covertext $\{U_i\}$ is i.i.d. zero-mean Gaussian with variance $\sigma_{u,0}^2$, if $Z = 0$, and with variance $\sigma_{u,1}^2$, if $Z = 1$. Assume that $\sigma_{u,0}^2 < \sigma_{u,1}^2$. The covertext is thus not ergodic, but it is stationary with $E\left[U_k^2\right] = (\sigma_{u,0}^2 + \sigma_{u,1}^2)/2$. Even though the covertext is non-ergodic, it is still true that no rate above $C^*(D_1, D_2, E\left[U_i^2\right])$ is achievable. In fact, no rate above $C^*(D_1, D_2, \sigma_{u,0}^2)$ can be achieved, as an attacker of the form (101) designed for the parameters $(D_1, D_2, \sigma_{u,0}^2)$ demonstrates. This type of argument naturally extends to any covertext with a finite number of ergodic modes, and in fact, with the proper modifications, to more general covertexts too.

6 Converse for Average Constraints

In this section, we prove Theorem 2.2. That is, we show that if the covertext distribution $\{P_{\boldsymbol{U}}\}$ satisfies

$$\sigma^2 = \liminf_{n\to\infty} E\left[\frac{1}{n}\|\boldsymbol{U}\|^2\right] < \infty, \tag{112}$$

and if *average* distortion constraints rather than *almost sure* distortion constraints are in effect, then the capacity of the public and private watermarking games is zero. We shall prove the theorem by exhibiting an attacker that satisfies the average distortion constraint (but not the a.s. constraint) and that guarantees that no positive rate is achievable.

For a given covertext $\{\boldsymbol{U}\}$ and for a given encoder sequence $\{f_n\}$, let the average power in the stegotext $\boldsymbol{X} = f_n(\boldsymbol{U}, W, \Theta_1)$ be given by

$$\tilde{a}_n = \frac{1}{n} E\left[\|\boldsymbol{X}\|^2\right] \tag{113}$$

Note that the encoder average distortion constraint $E\left[n^{-1}\|\boldsymbol{X}-\boldsymbol{U}\|^2\right] \le D_1$ and the triangle inequality $\|\boldsymbol{X}\| \le \|\boldsymbol{X}-\boldsymbol{U}\| + \|\boldsymbol{U}\|$ guarantee that $\tilde{a}_n \le (\sqrt{D_1} + \sqrt{E\left[\|\boldsymbol{U}\|^2\right]/n})^2$. Consequently, it follows by (112) that for any $\epsilon > 0$ and any integer $n_0 > 0$ there exists some $n^* > n_0$ such that

$$\tilde{a}_{n^*} \le (\sigma + \epsilon + \sqrt{D_1})^2. \tag{114}$$

Let the attack key Θ_2 take on the value 0 with probability p, and take on the value 1 with probability $1-p$, where

$$p = \min\left\{\frac{D_2}{(\sigma+\epsilon+\sqrt{D_1})^2}, 1\right\}. \tag{115}$$

For the blocklength n^* consider now the attacker

$$\tilde{g}_{n^*}(\boldsymbol{x}, \theta_2) = \theta_2 \boldsymbol{x} \tag{116}$$

that with probability p produces the all-zero forgery, and with probability $(1-p)$ does not alter the stegotext at all. Irrespective of the rate (as long as $\lfloor 2^{nR} \rfloor > 1$) and of the version of the game, this attacker guarantees a probability of error of at least $p/2$. It remains to check that $\tilde{g}_{n^*}(\boldsymbol{x}, \theta_2)$ satisfies the average distortion constraint. Indeed, the average distortion introduced by $\tilde{g}_{n^*}$ is given

by

$$\begin{aligned}\frac{1}{n^*}E\left[\|\boldsymbol{X}-\tilde{g}_{n^*}(\boldsymbol{X},\Theta_2)\|^2\right] &= p\cdot\frac{1}{n^*}E\left[\|\boldsymbol{X}\|^2\right]\\ &= p\cdot\tilde{a}_{n^*}\\ &\le p\cdot(\sigma+\epsilon+\sqrt{D_1})^2\\ &\le D_2,\end{aligned}$$

where the first equality follows from (116), the second by (113), the subsequent inequality by (114), and the last inequality by (115).

7 Conclusions

In Section 2.1.3, we showed that the watermarking game is equivalent to a communication system where the encoder and the jammer have non-causal access to different side informations. The covertext $\boldsymbol{U}$ in the watermarking game corresponds to the additive noise sequence that is known non-causally to the transmitter. It is interesting to note that, in terms of the communication system, the encoders that achieve capacity use part of their allowed power to *enhance* the noise. Indeed, in the private version of the watermarking game, the stegotext $\boldsymbol{X}$ is created by adding an independent (except for being orthogonal to $\boldsymbol{U}$) random vector to the covertext $\boldsymbol{U}$ scaled by $b_1(A; D_1, \sigma_u^2)$. Since for the optimal choice of A (i.e., the one that achieves the maximum in (8)) the constant b_1 is greater than one, this corresponds to enhancing the covertext; see Section 4.2. Similarly, in the public version of the game, if $Z \approx \alpha\sigma_u^2+\rho$ (which is very likely by Lemma 4.3), then the projection of the stegotext $\boldsymbol{X}$ onto the covertext $\boldsymbol{U}$ is also produced by roughly multiplying by the constant b_1, i.e., $\boldsymbol{X}|_{\boldsymbol{U}} \approx b_1\boldsymbol{U}$.

We have seen that the watermarking capacity increases with the uncertainty in the covertext in the following sense. First, for an i.i.d. Gaussian covertext, the capacity of the watermarking game is increasing in the variance of the covertext. Second, with squared error distortion measures and a fixed covertext variance, the covertext distribution with the largest watermarking capacity is the Gaussian distribution, which has the highest entropy among all distributions of a given variance. Intuitively, if the uncertainty in the covertext is large, then the encoder can hide more information in the stegotext since the attacker learns little about the covertext from observing the stegotext.

If the attacker does not take advantage of its knowledge of the stegotext, then this property is not as strong. For example, if the attacker adds an arbitrary sequence as in the additive attack watermarking game or if the attacker adds a Gaussian process as in the extension of writing on dirty paper, then the amount of uncertainty in the covertext has little bearing on the capacity. In all cases, the watermarking system's knowledge of the covertext should be used to its advantage. It is suboptimal to ignore the encoder's knowledge of the covertext, as some systems do by forming the stegotext by adding the covertext and a sequence that depends only on the watermark.

A Proofs of Technical Lemmas

A.1 Proof of Lemma 3.1

We first show following Chen [32] that for arbitrary distributions $P_{\boldsymbol{U}}$, $P_{\boldsymbol{X}|\boldsymbol{U}}$, $P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}$, and $P_{\boldsymbol{Y}|\boldsymbol{X}}$ the mutual information terms satisfy $I_{\text{priv}} \geq I_{\text{pub}}$. All of the below mutual information terms are evaluated in terms of these distributions. We will assume that I_{priv} is finite, since otherwise the lemma is trivial. We can write that

$$\begin{aligned}
I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) &= n^{-1} I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U}) \\
&\geq n^{-1} I(\boldsymbol{V}; \boldsymbol{Y}|\boldsymbol{U}) \qquad (117) \\
&= n^{-1}\big(I(\boldsymbol{V}; \boldsymbol{U}, \boldsymbol{Y}) - I(\boldsymbol{V}; \boldsymbol{U})\big) \qquad (118) \\
&\geq n^{-1}\big(I(\boldsymbol{V}; \boldsymbol{Y}) - I(\boldsymbol{V}; \boldsymbol{U})\big) \\
&= I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}})
\end{aligned}$$

where (117) follows by the data processing inequality (see e.g. [17]) because $\boldsymbol{V}$ and $\boldsymbol{Y}$ are conditionally independent given $(\boldsymbol{X}, \boldsymbol{U})$; and where (118) follows by the chain rule for mutual informations. We next show that the values of the mutual information games also behave as desired. Fix n and $\epsilon > 0$ and let $P^*_{\boldsymbol{X}|\boldsymbol{U}}$ and $P^*_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}$ be distributions that are within ϵ of the supremum in (30). Thus,

$$\begin{aligned}
\sup_{P_{\boldsymbol{X}|\boldsymbol{U}}} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}} I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) &\geq \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}} I_{\text{priv}}(P_{\boldsymbol{U}}, P^*_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) \\
&\geq \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}} I_{\text{pub}}(P_{\boldsymbol{U}}, P^*_{\boldsymbol{X}|\boldsymbol{U}}, P^*_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) \\
&\geq \sup_{P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}}} I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) - \epsilon,
\end{aligned}$$

where the second inequality follows by the preceding paragraph and the final inequality follows by our choice of $P^*_{\boldsymbol{X}|\boldsymbol{U}}$ and $P^*_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}$. The lemma follows since $\epsilon > 0$ can be chosen as small as desired. □

A.2 Proof of Lemma 3.2

The proof is organized as follows. In Lemma A.1, we show that a Gaussian covertext distribution and a jointly Gaussian watermarking channel maximize the mutual information term of interest. Using this result and some basic mutual information manipulations, we then complete the proof.

Lemma A.1. *Let $P_{U,X}$ be an arbitrary distribution with covariance matrix K_{UX}, and let $P^*_{U,X}$ be a jointly Gaussian distribution of covariance matrix $K^*_{UX} = K_{UX}$. Then,*

$$I_{\text{priv}}(P_U, P_{X|U}, P^A_{Y|X}) \leq I_{\text{priv}}(P^*_U, P^*_{X|U}, P^A_{Y|X}),$$

where $P^A_{Y|X}$ is defined in Section 3.1 and $A > D_2$ is arbitrary.

Proof. Recall that under the attack channel $P^A_{Y|X}$, the random variables Y and X are related by $Y = cX + S_2$, where $c = c(A; D_2)$ and S_2 is mean-zero variance-cD_2 Gaussian random variable independent of X. Thus,

$$h_{P_U P_{X|U} P^A_{Y|X}}(Y|X) = h(S_2) = h_{P^*_U P^*_{X|U} P^A_{Y|X}}(Y|X), \tag{119}$$

where these and the below differential entropies exist by the structure of the attack channel under consideration. Let βU be the linear minimum mean squared-error estimator of Y given U. Note that β depends on second-order statistics only, so that its value under $P^*_{U,X}$ is the same as under $P_{U,X}$. Thus,

$$\begin{aligned}
h_{P_U P_{X|U} P^A_{Y|X}}(Y|U) &= h_{P_U P_{X|U} P^A_{Y|X}}(Y - \beta U|U) \\
&\leq h_{P_U P_{X|U} P^A_{Y|X}}(Y - \beta U) \\
&\leq \frac{1}{2}\log\left(2\pi e E_{P_U P_{X|U} P^A_{Y|X}}\left[(Y-\beta U)^2\right]\right) \\
&= \frac{1}{2}\log\left(2\pi e E_{P^*_U P^*_{X|U} P^A_{Y|X}}\left[(Y-\beta U)^2\right]\right) \\
&= h_{P^*_U P^*_{X|U} P^A_{Y|X}}(Y|U),
\end{aligned} \tag{120}$$

where the first inequality follows since conditioning reduces entropy, the second inequality follows since a Gaussian distribution maximizes entropy subject to a second moment constraint, and (120) follows since under P^*_U, $P^*_{X|U}$ and $P^A_{Y|X}$ the random variables U and Y are jointly Gaussian and hence $Y - \beta U$ is Gaussian and independent of U. Combining (119) and (120) with the definition of I_{priv} (see (26)) completes the proof of Lemma A.1. □

To continue with the proof of Lemma 3.2, if under P^*_U and $P^*_{X|U}$ the random variables U and X are zero-mean and jointly Gaussian, then

$$I_{\text{priv}}(P^*_U, P^*_{X|U}, P^A_{Y|X}) = \frac{1}{2}\log\left(1 + \frac{c(A; D_2)\, b_2\left(E\left[X^2\right]; E\left[(X-U)^2\right], E\left[U^2\right]\right)}{D_2}\right), \tag{121}$$

where $b_2(\cdot;\cdot,\cdot)$ is defined in (4) and $A > D_2$. Note that b_2 and hence the whole expression (121)

is concave in the triple $\left(E\left[U^2\right], E\left[(X-U)^2\right], E\left[X^2\right]\right)$, as can be verified by checking that the Hessian is non-negative definite. We can now compute that

$$\begin{aligned} I_{\text{priv}}\left(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, (P_{Y|X}^{A_n})^n\right) &\leq \frac{1}{n}\sum_{i=1}^{n} I_{\text{priv}}\left(P_{U_i}, P_{X_i|U_i}, P_{Y|X}^{A_n}\right) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\left(1+\frac{c(A_n;D_2)\, b_2\left(E\left[X_i^2\right]; E\left[(X_i-U_i)^2\right], E\left[U_i^2\right]\right)}{D_2}\right) \\ &\leq \frac{1}{2}\log\left(1+\frac{c(A_n;D_2)\, b_2\left(A_n; D_{1,n}, \sigma_{u,n}^2\right)}{D_2}\right) \\ &= \frac{1}{2}\log\big(1+s(A_n; D_{1,n}, D_2, \sigma_{u,n}^2)\big), \end{aligned}$$

where the first inequality follows by the chain rule and since conditioning reduces entropy, the second inequality follows by Lemma A.1 and by (121), the third inequality follows by the above discussed concavity of (121), and the final equality follows by the definition of $s(\cdot;\cdot,\cdot,\cdot)$ (7). We obtain equality in each of the above inequalities when $P_{\boldsymbol{U}} = (P_U^G)^n$ and $P_{\boldsymbol{X}|\boldsymbol{U}} = (P_{X|U}^{A_n})^n$. This completes the proof of Lemma 3.2. □

A.3 Proof of Lemma 3.3

For every $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, consider the one-dimensional optimization based on the watermarking channel described in Section 3.2

$$M(D_2, A) = \inf_{P_{Y|X}\in\mathcal{D}_2(D_2, P_U^G, P_{X|U}^A)} I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}\right). \tag{122}$$

In Lemma A.2, we derive some properties of $M(D_2, A)$, which we subsequently use to show that $M(D_2, A)$ is a lower bound on the LHS of (46). In Lemma A.3, we show that when computing $M(D_2, A)$ we only need to consider attack channels that make the random variables Y and V jointly Gaussian. We finally use this lemma to compute $M(D_2, A)$.

Lemma A.2. *For a fixed A, the function $M(D_2, A)$ defined in (122) is convex and non-increasing in D_2.*

Proof. The function $M(D_2, A)$ is non-increasing in D_2 since increasing D_2 only enlarges the feasible set $\mathcal{D}_2(D_2, P_U^G, P_{X|U}^A)$. To show that $M(\cdot, A)$ is convex in D_2, we first note that

$$I_{\text{pub}}\left(P_U, P_{X|U}, P_{V|U,X}, P_{Y|X}\right) = I(V;Y) - I(V;U)$$

is convex in $P_{Y|X}$. Indeed, $I(V;U)$ does not depend on $P_{Y|X}$ and $I(V;Y)$ is convex in $P_{Y|V}$ and hence also convex in $P_{Y|X}$ since[7] $V \multimap X \multimap Y$. Given parameters A, D_r, D_s, and $\epsilon > 0$, let

[7]The notation $V \multimap X \multimap Y$ will mean that the random variables V, X, and Y form a Markov chain.

watermarking channels $P^r_{Y|X} \in \mathcal{D}_2(D_r, P^G_U, P^A_{X|U})$ and $P^s_{Y|X} \in \mathcal{D}_2(D_s, P^G_U, P^A_{X|U})$ be such that

$$I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P^r_{Y|X}\right) \leq M(D_r, A) + \epsilon, \tag{123}$$

and

$$I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P^s_{Y|X}\right) \leq M(D_s, A) + \epsilon. \tag{124}$$

For any $0 \leq \lambda \leq 1$, let $P^\lambda_{Y|X} = \lambda P^r_{Y|X} + \bar{\lambda} P^s_{Y|X}$, where $\bar{\lambda} = (1-\lambda)$. We complete the proof with

$$\begin{aligned} M(\lambda D_r + \bar{\lambda} D_s, A) &\leq I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P^\lambda_{Y|X}\right) \\ &\leq \lambda I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P^r_{Y|X}\right) + \bar{\lambda} I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P^s_{Y|X}\right) \\ &\leq \lambda M(D_r, A) + \bar{\lambda} M(D_s, A) + \epsilon, \end{aligned}$$

where the first inequality follows since $E_{P_U P^A_{X|U} P^\lambda_{Y|X}}[(X-Y)^2] \leq \lambda D_r + \bar{\lambda} D_s$, the second inequality follows by the convexity of $I_{\text{pub}}(P_U, P_{X|U}, P_{V|U,X}, \cdot)$, and the final inequality follows by (123) and (124). The lemma follows since ϵ is an arbitrary positive number. □

We continue by showing that $M(D_2, A)$ is a lower bound on the LHS of (46). Indeed, if

$$P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, (P^G_U)^n, (P^A_{X|U})^n), \tag{125}$$

then

$$\begin{aligned} I_{\text{pub}}\left((P^G_U)^n, (P^A_{X|U})^n, (P^A_{V|U,X})^n, P_{\boldsymbol{Y}|\boldsymbol{X}}\right) &\geq \frac{1}{n}\sum_{i=1}^n I_{\text{pub}}\left(P^G_U, P^A_{X|U}, P^A_{V|U,X}, P_{Y_i|X_i}\right) \\ &\geq \frac{1}{n}\sum_{i=1}^n M\left(E_{P^G_U P^A_{X|U} P_{Y_i|X_i}}\left[(Y_i - X_i)^2\right], A\right) \\ &\geq M\left(E_{(P^G_U P^A_{X|U})^n P_{\boldsymbol{Y}|\boldsymbol{X}}}\left[n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\|^2\right], A\right) \\ &\geq M(D_2, A), \end{aligned}$$

where the first inequality follows since the watermarking channel is memoryless, by the chain rule, and by the fact that conditioning reduces entropy; the second inequality follows by the definition of $M(\cdot,\cdot)$; and the final two inequalities follow by Lemma A.2 and by (125) so that the expected distortion is less than D_2.

To complete the proof of Lemma 3.3, we show that a minimum in the definition of $M(D_2, A)$ is achieved by the distribution $P^A_{Y|X}$ of Section 3.1. We now show that we only need to consider conditional distributions $P_{Y|X}$ under which V and Y are jointly Gaussian.

Lemma A.3. *Let V and Z be jointly Gaussian random variables with covariance matrix K_{VZ}. Let Y be another (not necessarily Gaussian) random variable related to V via the covariance matrix K_{VY}. If $K_{VY} = K_{VZ}$, then $I(V;Y) \geq I(V;Z)$.*

Remark: Similar lemmas have been given in a preliminary version of [40] and in [41], assuming that V and Y have a joint density.

Proof. It suffices to prove the lemma when all random variables are zero mean. If $I(V;Y)$ is infinite then there is nothing to prove. Thus, we only consider the case where

$$I(V;Y) < \infty. \tag{126}$$

For the fixed covariance matrix $K = K_{VY} = K_{VZ}$, let the linear minimum mean squared-error estimator of V given Y be βY. Note that the constant β is determined uniquely by the correlation matrix K and thus βZ is also the linear minimum mean squared-error estimator of V given Z. Since the random variables V and Z are jointly Gaussian, this is also the minimum mean squared-error estimator, and furthermore $V - \beta Z$ is independent of Z. If the conditional density $f_{V|Y}$ exists, then

$$\begin{aligned}
I(V;Y) &= h(V) - h(V|Y) && (127)\\
&\geq h(V) - h(V - \beta Y) && (128)\\
&\geq h(V) - \frac{1}{2}\log 2\pi e E[(V-\beta Y)^2] && (129)\\
&= h(V) - \frac{1}{2}\log 2\pi e E[(V-\beta Z)^2] && (130)\\
&= I(V;Z) && (131)\\
&= \frac{1}{2}\log\left(\frac{E[V^2]E[Z^2]}{|K_{VZ}|}\right) && (132)
\end{aligned}$$

and the lemma is proved. Here, (127) follows since we have assumed that a conditional density exists; (128) follows since conditioning reduces entropy; (129) follows since a Gaussian maximizes differential entropy subject to a second moment constraint; (130) follows since $K_{VY} = K_{VZ}$ and hence all second order moments are the same; (131) follows since $V - \beta Z$ is both Gaussian and independent of Z; and (132) follows since V and Z are zero-mean jointly Gaussian random variables.

By (126) the conditional density $f_{V|Y}$ exists if Y takes on a countable number of values. This follows since (126) implies $P_{V,Y} \ll P_V P_Y$, i.e., the joint distribution is absolutely continuous with respect to the product of the marginals. In particular, $P_{V|Y}(\cdot|y) \ll P_V$ for every y such that $P_Y(y) > 0$. Furthermore, V is Gaussian and hence $P_V \ll \lambda$, where λ is the Lebesgue measure. Thus, $P_{V|Y}(\cdot|y) \ll \lambda$ for every y such that $P_Y(y) > 0$ and hence the conditional density exists.

To conclude the proof of the lemma, we now consider the case where Y does not necessarily take on a countable number of values and $I(V;Y) < \infty$. This case follows using an approximation argument. For any $\Delta > 0$, let $q_\Delta : \mathbb{R} \mapsto \{\ldots, -2\Delta, -\Delta, 0, \Delta, 2\Delta, \ldots\}$ be a uniform quantizer with cell size Δ, i.e., $q_\Delta(x)$ maps x to the closest integer multiple of Δ. Let

$$Y_\Delta = q_\Delta(Y).$$

By the data processing inequality,

$$I(V;Y) \geq I(V;Y_\Delta). \tag{133}$$

The random variable Y_Δ takes on a countable number of values and by (126) and (133), $I(V;Y_\Delta) < \infty$. Thus, the conditional density $f_{V|Y_\Delta}$ exists and

$$I(V;Y_\Delta) \geq \frac{1}{2}\log\left(\frac{E[V^2]E[Y_\Delta^2]}{|K_{VY_\Delta}|}\right). \tag{134}$$

Since $|Y - Y_\Delta| \leq \Delta/2$, it follows that $E[Y_\Delta^2] \to E[Y^2]$ and $|K_{VY_\Delta}| \to |K_{VY}|$ as $\Delta \downarrow 0$. Since (133) and (134) hold for all $\Delta > 0$, the lemma follows by letting Δ approach zero. □

To continue with the evaluation of $M(D_2, A)$, we note that since under the distributions P_U^G, $P_{X|U}^A$, and $P_{V|U,X}^A$, the random variable V has a Gaussian distribution, the above lemma allows us to assume throughout the rest of the proof that the attack channel $P_{Y|X}$ makes the random variables V and Y jointly Gaussian. Recall that the random variables V, X, and Y form a Markov chain. Thus, if we let $Y = c_1X + S_1$, where S_1 is Gaussian random variable independent of X with variance $c_2 \geq 0$, then we can generate all possible correlation matrices K_{VY} by varying the parameters c_1 and c_2. Since the mutual information $I(V;Y)$ only depends on the correlation matrix K_{VY}, we can compute the quantity $M(D_2, A)$ by only considering such attack channels.

Let $P_{Y|X}^{c_1,c_2}$ be the attack channel such that the random variable Y is distributed as $c_1X + S_1$, where S_1 is a random variable independent of X which is Gaussian of zero mean and variance c_2. Under this distribution,

$$E_{P_U P_{X|U}^A P_{Y|X}^{c_1,c_2}}[(X-Y)^2] = (1-c_1)^2 A + c_2 \tag{135}$$

so that the condition $P_{Y|X}^{c_1,c_2} \in \mathcal{D}_2(D_2, P_U^G, P_{X|U}^A)$ is thus equivalent to the requirement:

$$(1-c_1)^2 A + c_2 \leq D_2. \tag{136}$$

We next note that for $A > D_2$ (and $c_2 \geq 0$) any pair (c_1, c_2) satisfying (136) must also satisfy

$$\frac{c_2}{c_1^2} \leq \frac{D_2}{c(A;D_2)}, \tag{137}$$

where $c(\cdot;\cdot)$ is defined in (5). Conversely, for any $0 \leq \kappa \leq D_2/c(A;D_2)$ there exists such a pair (c_1, c_2) for which $c_2/c_1^2 = \kappa$. Indeed, by (136) we have

$$\begin{aligned} P_{Y|X}^{c_1,c_2} \in \mathcal{D}_2(D_2, P_U^G, P_{X|U}^A) &\iff c_2 \leq D_2 - (1-c_1)^2 A \\ &\iff \frac{c_2}{c_1^2} \leq \frac{D_2 - (1-c_1)^2 A}{c^2} \leq \frac{D_2}{c(A;D_2)} \end{aligned}$$

where the last inequality is achieved with equality if $c_1 = c(A;D_2)$. The converse statement follows by solving $c_2/c_1^2 = \kappa$ subject to (136) holding with equality.

Thus, if $\alpha = \alpha(A; D_1, D_2, \sigma_u^2)$, $\rho = \rho(A; D_1, \sigma_u^2)$ and $b_1 = b_1(A; D_1, \sigma_u^2)$, then

$$\begin{aligned} I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^{c_1,c_2}\right) &= \frac{1}{2}\log\left(\frac{\alpha^2\sigma_u^2 + 2\alpha\rho + D_1 - (\alpha + b_1 - 1)^2\sigma_u^2}{\alpha^2\sigma_u^2 + 2\alpha\rho + D_1 - ((\alpha-1)b_1\sigma_u^2 + A)^2/(A + \frac{c_2}{c_1^2})}\right) \\ &\geq \frac{1}{2}\log\left(1 + s(A; D_1, D_2, \sigma_u^2)\right), \end{aligned}$$

where the equality follows by evaluating I_{pub} with the given distributions and the inequality follows by the relevant definitions and by (137). Equality is achieved when $c_1 = c(A; D_2)$ and $c_2 = c(A; D_2)D_2$. The combination of all of the above arguments shows that Lemma 3.3 is valid. □

A.4 Proof of Lemma 4.4

A.4.1 Distribution of Chosen Auxiliary Codeword

In order to prove Lemma 4.4, we shall need the distribution of the chosen auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$ (defined in (75)), both unconditionally and conditioned on the random vector $\boldsymbol{X}$ and the random variable Z (defined in (76) and (79), respectively). We present these distributions in the following two lemmas.

Lemma A.4. *The random vector $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the n-sphere $S^n(0, \sqrt{n\sigma_{v,\text{type}}^2})$, where* type *is* add *or* gen *as appropriate.*

Proof. By the symmetry of the encoding process it is apparent that $\boldsymbol{V}_W(\boldsymbol{U})$ is independent of the message W. Assume then without loss of generality (w.l.g.) that $W = 1$. Since all the auxiliary random vectors $\{\boldsymbol{V}_{1,k}\}$ in bin 1 take value in the n-sphere $S^n(0, \sqrt{n\sigma_v^2})$, it follows that the chosen auxiliary codeword must take value in the same n-sphere. Finally, since the joint distribution of $\{\boldsymbol{V}_{1,k}\}$ is invariant under any unitary transformation as is the distribution of $\boldsymbol{U}$, and since $\boldsymbol{U}$ and $\{\boldsymbol{V}_{1,k}\}$ are independent, it follows that the unconditional distribution of $\boldsymbol{V}_W(\boldsymbol{U})$ is as stated above. In other words, the fact that $\boldsymbol{V}_W(\boldsymbol{U})$ achieves the maximum inner product with $\boldsymbol{U}$ does not tell us anything about the direction of $\boldsymbol{V}_W(\boldsymbol{U})$. □

Lemma A.5. *Given $\boldsymbol{X} = \boldsymbol{x}$ and $Z = z$, the random vector $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the set*[8]

$$\mathcal{V}(\boldsymbol{x}, z) = \left\{a_1\boldsymbol{x} + \boldsymbol{v} : \boldsymbol{v} \in S^n(0, \sqrt{na_2})\boldsymbol{x}^\perp\right\}, \tag{138}$$

where

$$a_1 = \frac{\sigma_{v,\text{type}}^2 + (1 - \alpha_{\text{type}})z}{n^{-1}\|\boldsymbol{x}\|^2}, \qquad a_2 = \frac{(1 - \alpha_{\text{type}})^2(\sigma_u^2\sigma_{v,\text{type}}^2 - z^2)}{n^{-1}\|\boldsymbol{x}\|^2}, \tag{139}$$

and type *is* add *or* gen *as appropriate.*

Proof. We drop the subscripts since the proof is the same for both types. Conditional on $\boldsymbol{U} = \boldsymbol{u}$ and on $Z = z$, the auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the set

$$\mathcal{V}'(\boldsymbol{u}, z) = \left\{\boldsymbol{v} : n^{-1}\|\boldsymbol{v}\|^2 = \sigma_v^2 \text{ and } n^{-1}\langle\boldsymbol{v}, \boldsymbol{u}\rangle = z\right\},$$

[8]Recall that we use $\boldsymbol{x}^\perp$ to denote the linear sub-space of vectors that are orthogonal to $\boldsymbol{x}$.

as follows by the definition of Z (79) and the distribution of the codebook $\{\boldsymbol{V}_{j,k}\}$. Using the deterministic relation (76) we can now relate the appropriate conditional densities as

$$f_{\boldsymbol{V}_W(\boldsymbol{U})|\boldsymbol{X},Z}(\boldsymbol{v}|\boldsymbol{X}=\boldsymbol{x},Z=z)=f_{\boldsymbol{V}_W(\boldsymbol{U})|\boldsymbol{U},Z}\left(\boldsymbol{v}\Big|\boldsymbol{U}=\frac{\boldsymbol{x}-\boldsymbol{v}}{1-\alpha},Z=z\right).$$

The proof will be concluded once we demonstrate that, irrespective of z, it holds that $\boldsymbol{v}\in\mathcal{V}(\boldsymbol{x},z)$ if, and only if, $\boldsymbol{v}\in\mathcal{V}'\big((\boldsymbol{x}-\boldsymbol{v})/(1-\alpha),z\big)$. Indeed, if $\boldsymbol{v}\in\mathcal{V}(\boldsymbol{x},z)$, then we can calculate that $n^{-1}\|\boldsymbol{v}\|^2=a_1^2n^{-1}\|\boldsymbol{x}\|^2+a_2=\sigma_v^2$ using the fact that

$$n^{-1}\|\boldsymbol{x}\|^2=\sigma_v^2+2(1-\alpha)z+(1-\alpha)^2\sigma_u^2. \tag{140}$$

Furthermore,

$$\frac{1}{n}\left\langle\boldsymbol{v},\frac{\boldsymbol{x}-\boldsymbol{v}}{1-\alpha}\right\rangle=\frac{\sigma_v^2+(1-\alpha)z-\sigma_v^2}{1-\alpha}=z,$$

and thus $\boldsymbol{v}\in\mathcal{V}'\big((\boldsymbol{x}-\boldsymbol{v})/(1-\alpha),z\big)$. Conversely, if $\boldsymbol{v}\in\mathcal{V}'\big((\boldsymbol{x}-\boldsymbol{v})/(1-\alpha),z\big)$, then

$$\frac{1}{n}\left\langle\boldsymbol{v},\frac{\boldsymbol{x}-\boldsymbol{v}}{1-\alpha}\right\rangle=\frac{n^{-1}\langle\boldsymbol{v},\boldsymbol{x}\rangle-\sigma_v^2}{1-\alpha}=z,$$

and hence $\boldsymbol{v}|_{\boldsymbol{x}}=a_1\boldsymbol{x}$. Furthermore,

$$\frac{1}{n}\big\|\boldsymbol{v}|_{\boldsymbol{x}^\perp}\big\|^2=\frac{1}{n}\|\boldsymbol{v}\|^2-\frac{1}{n}\big\|\boldsymbol{v}|_{\boldsymbol{x}}\big\|^2=\sigma_v^2-\frac{a_1^2\|\boldsymbol{x}\|^2}{n}=a_2,$$

where we have again used (140), and thus $\boldsymbol{v}\in\mathcal{V}(\boldsymbol{x},z)$. □

A.4.2 Case I: Additive Attacker

Recall that a deterministic additive attacker described in Section 4.1.2 is specified by a vector $\tilde{\boldsymbol{y}}$ satisfying (58). Fix some $\epsilon_3>0$ (to be chosen later) and choose n_2 large enough to ensure

$$\Pr(\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3)\geq 1-\epsilon,\quad\forall n>n_2, \tag{141}$$

where the events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ are defined by

$$\mathcal{E}_1=\left\{\left|2n^{-1}\langle\boldsymbol{X},\tilde{\boldsymbol{y}}\rangle\right|\leq\epsilon_3\right\},\quad\mathcal{E}_2=\left\{\left|n^{-1}\langle\boldsymbol{V}_W(\boldsymbol{U}),\tilde{\boldsymbol{y}}\rangle\right|\leq\epsilon_3\right\},\quad\mathcal{E}_3=\{Z\geq\alpha\sigma_u^2\}.$$

Note that whenever $\epsilon_3>0$, such an n_2 can always be found by the union of events bound, because the probability of the complement of each of the events is vanishing uniformly in $\tilde{\boldsymbol{y}}$, for all $\tilde{\boldsymbol{y}}$ satisfying (58). Indeed, $\mathcal{E}_1^c$ and $\mathcal{E}_2^c$ have vanishing probabilities because both $\boldsymbol{U}$ and $\boldsymbol{V}_W(\boldsymbol{U})$ are uniformly distributed on n-spheres (see Lemma A.4) and since $\boldsymbol{X}=\boldsymbol{V}+(1-\alpha)\boldsymbol{U}$, and $\mathcal{E}_3^c$ has

vanishing probability by Lemma 4.3. Event $\mathcal{E}_1$ guarantees that

$$\begin{aligned} Z_1 &= \frac{1}{n}\|\boldsymbol{X}\|^2 + \frac{2}{n}\langle \boldsymbol{X}, \tilde{\boldsymbol{y}}\rangle + \frac{1}{n}\|\tilde{\boldsymbol{y}}\|^2 \\ &\leq \sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2, \end{aligned} \tag{142}$$

where the equality follows by the definition of Z_1 (79) and the form of the additive attacker given in Section 4.1.2, and where the inequality follows by (140), (58), and the inequality defining $\mathcal{E}_1$. From the definition of Z_2 (79) it follows that $\mathcal{E}_2$ guarantees that $Z_2 \geq -\epsilon_3$. Consequently, the intersection $\mathcal{E}_1\mathcal{E}_2$ guarantees that

$$\beta(Z, Z_1, Z_2) \geq \frac{\sigma_v^2 + (1-\alpha)Z - \epsilon_3}{\sqrt{\sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2}}. \tag{143}$$

For any $\epsilon_3 > 0$, the RHS of (143) is monotonically increasing in Z, so that the intersection $\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3$ implies

$$\beta(Z, Z_1, Z_2) \geq \frac{\sigma_v^2 + (1-\alpha)\alpha\sigma_u^2 - \epsilon_3}{\sqrt{\sigma_v^2 + 2(1-\alpha)\alpha\sigma_u^2 + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2}}. \tag{144}$$

Recalling the definitions in Section 4.3.1 and the definition of $\beta^*(R_1+\delta)$ (86), one can show using some algebra that for $\epsilon_3 = 0$, the RHS of (144) equals $\beta^*(R_1+\delta)$. Since the RHS of (144) is continuous in ϵ_3, we can choose some $\epsilon_3 > 0$ small enough (and the resulting n_2 large enough) so that the intersection $\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3$ will guarantee that

$$\beta(Z, Z_1, Z_2) \geq \beta^*(R_1+\delta) - \epsilon_1.$$

In the case of an additive attacker, the lemma thus follows from (141).

A.4.3 Case II: General Attacker

In order to prove the desired result for a general attacker, we need the following lemma.

Lemma A.6. *As n tends to infinity, the sequence of random variables $n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle$ approaches zero in probability uniformly over all the general attackers of Section 4.1.2.*

Proof. Conditional on $\boldsymbol{X} = \boldsymbol{x}$ and $Z = z$, the random vector $\boldsymbol{V}_W(\boldsymbol{U})$ is by Lemma A.5 distributed like $a_1\boldsymbol{x} + \boldsymbol{V}$, where $\boldsymbol{V}$ is uniformly distributed on $S^n(0, \sqrt{na_2})\boldsymbol{x}^\perp$, and a_2 defined in (139) depends on z. Consequently for any $0 < \zeta < \sqrt{D_2\sigma_v^2}$,

$$\begin{aligned} &\Pr\left(\left|n^{-1}\left\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\right\rangle\right| > \zeta \middle| \boldsymbol{X} = \boldsymbol{x}, Z = z\right) \\ &\quad= \Pr\left(\left|\left\langle \boldsymbol{\gamma}_2(\boldsymbol{x})/\sqrt{n\gamma_3(\boldsymbol{x})}, \boldsymbol{V}/\sqrt{na_2}\right\rangle\right| > \zeta/\sqrt{\gamma_3(\boldsymbol{x})a_2}\right) \\ &\quad\leq \Pr\left(\left|\left\langle \boldsymbol{\gamma}_2(\boldsymbol{x})/\sqrt{n\gamma_3(\boldsymbol{x})}, \boldsymbol{V}/\sqrt{na_2}\right\rangle\right| > \zeta/\sqrt{D_2\sigma_v^2}\right) \\ &\quad= \frac{2C_{n-1}\left(\arccos\left(\zeta/\sqrt{D_2\sigma_v^2}\right)\right)}{C_{n-1}(\pi)}. \end{aligned}$$

Here, the first equality follows by Lemma A.5 and the fact that $\boldsymbol{\gamma}_2(\boldsymbol{x}) \in \boldsymbol{x}^\perp$, the subsequent inequality follows from $\gamma_3(\boldsymbol{x}) \leq D_2$ and $a_2 \leq \sigma_v^2$ (see (60) and (139)), and the final equality follows since $\boldsymbol{V}/\sqrt{na_2}$ is uniformly distributed on $S^n(0,1)\boldsymbol{x}^\perp$ and since $\boldsymbol{\gamma}_2(\boldsymbol{x})/\sqrt{n\gamma_3(\boldsymbol{x})}$ also takes value in this set. Since the resulting upper bound, which tends to zero, does not depend on $\boldsymbol{x}$ or z, it must also hold for the unconditional probability. □

We now proceed to prove Lemma 4.4 for a general attacker. Choose n_2 large enough to ensure that

$$\Pr(\mathcal{E}_4\mathcal{E}_5) \geq 1-\epsilon, \quad \forall n > n_2,$$

where

$$\mathcal{E}_4 = \{Z \geq \alpha\sigma_u^2 + \rho\}, \qquad \mathcal{E}_5 = \left\{n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle \geq -\epsilon_1\sigma_v\big(\sqrt{A}-\sqrt{D_2}\big)\right\}.$$

Such an n_2 can be found by the union of events bound since both $\mathcal{E}_4^c$ and $\mathcal{E}_5^c$ have vanishing probabilities by Lemmas 4.3 and A.6, respectively. For the deterministic general attacker of Section 4.1.2, we can express the random variables Z_1 and Z_2 of (79) as

$$Z_1 = \gamma_1^2(\boldsymbol{X})n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X}),$$

and

$$Z_2 = (\gamma_1(\boldsymbol{X})-1)(\sigma_v^2 + (1-\alpha)Z) + n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle.$$

Substituting these expressions in $\beta(Z, Z_1, Z_2)$ of (81) yields

$$\begin{aligned}\beta(Z,Z_1,Z_2) &= \frac{\sigma_v^2 + (1-\alpha)Z + (\gamma_1(\boldsymbol{X})-1)(\sigma_v^2+(1-\alpha)Z) + n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle}{\sqrt{(\gamma_1^2 n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X}))\sigma_v^2}} \\ &= \frac{\sigma_v^2 + (1-\alpha)Z}{\sqrt{\left(n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X})\right)\sigma_v^2}} + \frac{n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle}{\sqrt{Z_1\sigma_v^2}}. \end{aligned} \tag{145}$$

We conclude the proof by showing that the intersection $\mathcal{E}_4\mathcal{E}_5$ implies that (145) exceeds $\beta^*(R_1+\delta)-\epsilon_1$. Using the expression (140) and the definitions of Section 4.3.2, we see that event $\mathcal{E}_4$ implies that $n^{-1}\|\boldsymbol{X}\|^2$ is at least A. When this is true, then the distortion constraint (12) and the triangle inequality imply that Z_1 is at least $(\sqrt{A}-\sqrt{D_2})^2$. Thus, the intersection $\mathcal{E}_4\mathcal{E}_5$ guarantees that the second term of (145) is at least $-\epsilon_1$. We now turn to the first term on the RHS of (145), which using (140) can be rewritten as

$$\frac{\sigma_v^2 + (1-\alpha)Z}{\sqrt{\left(\sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X})\right)\sigma_v^2}}. \tag{146}$$

Note that $\mathcal{E}_4$ implies $n^{-1}\|\boldsymbol{X}\|^2$ is at least A and that in this case $\gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X}) \leq D_2/c$ (this follows

using (61) as in the argument after (137)). Thus, (146) can be lower bounded by

$$\frac{\sigma_v^2 + (1-\alpha)Z}{\sqrt{\left(\sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \frac{D_2}{c}\right)\sigma_v^2}}.$$

Since $\alpha < 1$ (68), the above term is increasing in Z. Substituting $Z = \alpha\sigma_u^2 + \rho$ into this term yields $\beta^*(R_1 + \delta)$, as can be verified using the definitions of R_1 (71) and $\beta^*(\cdot)$ (86), which yields

$$\beta^*(R_1 + \delta) = \left(\sigma_v^2 + (1-\alpha)(\alpha\sigma_u^2 + \rho)\right)\sqrt{\frac{c}{A\sigma_v^2}}.$$

The event $\mathcal{E}_4$ thus implies that the first term on the RHS of (145) is at least $\beta^*(R_1 + \delta)$. □

A.5 Proof of Lemma 5.2

To simplify the proof of this lemma, we shall use the following notation:

$$c^{(k)} = c(a_k; \tilde{D}_2), \quad b_1^{(k)} = b_1(a_k; D_1, \mu_k), \quad b_2^{(k)} = b_2(a_k; D_1, \mu_k), \tag{147}$$

where the functions $c(\cdot;\cdot)$, $b_1(\cdot;\cdot,\cdot)$, and $b_2(\cdot;\cdot,\cdot)$ are defined in Section 1.1. We shall also need the following technical claim.

Lemma A.7. *If the encoder satisfies the a.s. distortion constraint (9), then*

$$E\left[\frac{1}{n}\left\|g_n^*(\boldsymbol{X}, \Theta_2) - b_1^{(k)}c^{(k)}\boldsymbol{U}\right\|^2 \middle| K = k\right] \le c^{(k)}\left(c^{(k)}b_2^{(k)} + \tilde{D}_2\right),$$

for all $k \ge 1$ such that $\Pr(K = k) > 0$.

Proof. Recall that the attacker g_n^* defined in (100) produces an i.i.d. sequence of $\mathcal{N}(0, \tilde{D}_2)$ random variables $\boldsymbol{V}$ that is independent of $(\boldsymbol{X}, \boldsymbol{U})$. Furthermore, since K is a function of $\boldsymbol{X}$, the random vector $\boldsymbol{V}$ is also independent of $\boldsymbol{X}$ and $\boldsymbol{U}$ given K. Thus, for all $k \ge 1$ with $\Pr(K = k) > 0$,

$$\begin{aligned}
&E\left[n^{-1}\left\|g_n^*(\boldsymbol{X}, \Theta_2) - b_1^{(k)}c^{(k)}\boldsymbol{U}\right\|^2 \middle| K = k\right] \\
&= E\left[n^{-1}\left\|c^{(k)}\left(\boldsymbol{X} - b_1^{(k)}\boldsymbol{U}\right) + \sqrt{c^{(k)}}\boldsymbol{V}\right\|^2 \middle| K = k\right] \\
&= (c^{(k)})^2 E\left[n^{-1}\left\|\boldsymbol{X} - b_1^{(k)}\boldsymbol{U}\right\|^2 \middle| K = k\right] + c^{(k)}E\left[n^{-1}\|\boldsymbol{V}\|^2 \middle| K = k\right] \\
&= (c^{(k)})^2 E\left[n^{-1}\left(\|\boldsymbol{X}\|^2 - b_1^{(k)}2\langle\boldsymbol{X}, \boldsymbol{U}\rangle + (b_1^{(k)})^2\|\boldsymbol{U}\|^2\right) \middle| K = k\right] + c^{(k)}\tilde{D}_2 \\
&= (c^{(k)})^2\left(a_k - b_1^{(k)}E\left[2n^{-1}\langle\boldsymbol{X}, \boldsymbol{U}\rangle | K = k\right] + (b_1^{(k)})^2\mu_k\right) + c^{(k)}\tilde{D}_2,
\end{aligned}$$

where the final equality follows by the definitions of a_k and μ_k (see (98) and (99)). The proof will

be concluded once we show

$$n^{-1}E\left[\langle \boldsymbol{X}, \boldsymbol{U}\rangle \mid K = k\right] \geq \frac{1}{2}(a_k + \mu_k - D_1), \tag{148}$$

because

$$a_k - b_1^{(k)}(a_k + \mu_k - D_1) + (b_1^{(k)})^2\mu_k = b_2^{(k)},$$

by (147). We verify (148) by noting that for every $k \geq 1$ such that $\Pr(K = k) > 0$,

$$\begin{aligned} D_1 &\geq E\left[n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 \big| K = k\right] \\ &= E\left[n^{-1}\|\boldsymbol{X}\|^2 - 2n^{-1}\langle \boldsymbol{X}, \boldsymbol{U}\rangle + n^{-1}\|\boldsymbol{U}\|^2 \big| K = k\right] \\ &= a_k - E\left[2n^{-1}\langle \boldsymbol{X}, \boldsymbol{U}\rangle \big| K = k\right] + \mu_k, \end{aligned}$$

where the inequality follows since $n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 \leq D_1$ almost-surely so that the expectation given any event with positive probability must also be at most D_1. □

We can now write the mutual information term of interest as

$$\begin{aligned} & I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) \\ &= \sum_{k=0}^{m} \Pr(K = k) \cdot I(\boldsymbol{X}; \boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) \\ &= \sum_{k=1}^{m} \Pr(K = k) \cdot \big(h(\boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, K = k, \boldsymbol{U}, \Theta_1)\big), \end{aligned} \tag{149}$$

since by the structure of the attack channel all of the above differential entropies exist for all $k \geq 1$, and since when $k = 0$ the above mutual information is zero. To continue our proof, we shall next verify that

$$I(\boldsymbol{X}; \boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) = h(\boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, K = k, \boldsymbol{U}, \Theta_1) \tag{150}$$

is upper bounded by $\frac{1}{2}\log(1 + s(a_k; D_1, \tilde{D}_2, \mu_k))$, for all $k \geq 1$ satisfying $\Pr(K = k) > 0$. We can upper bound the first term on the RHS of (150) as

$$\begin{aligned} h(\boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) &= h\left(g_n^*(\boldsymbol{X}, \Theta_2)\big|K = k, \boldsymbol{U}, \Theta_1\right) \\ &= h\left(g_n^*(\boldsymbol{X}, \Theta_2) - c^{(k)}b_1^{(k)}\boldsymbol{U}\big|K = k, \boldsymbol{U}, \Theta_1\right) \\ &\leq h\left(g_n^*(\boldsymbol{X}, \Theta_2) - c^{(k)}b_1^{(k)}\boldsymbol{U}\big|K = k\right) \\ &\leq \frac{n}{2}\log\left(2\pi e E\left[\frac{1}{n}\left\|g_n^*(\boldsymbol{X}, \Theta_2) - c^{(k)}b_1^{(k)}\boldsymbol{U}\right\|^2 \Big| K = k\right]\right) \\ &\leq \frac{n}{2}\log\left(2\pi e\left((c^{(k)})^2 b_2^{(k)} + c^{(k)}\tilde{D}_2\right)\right), \end{aligned} \tag{151}$$

where the first inequality follows since conditioning reduces entropy, the second inequality follows since a Gaussian has the highest entropy subject to a second moment constraint, and (151) follows

by Lemma A.7. We can write the second term on the RHS of (150) as

$$\begin{aligned} h(\boldsymbol{Y}|\boldsymbol{X}, K=k, \boldsymbol{U}, \Theta_1) &= h\left(\sqrt{c^{(k)}}\boldsymbol{V}\middle|K=k\right) \\ &= \frac{n}{2}\log\left(2\pi e c^{(k)}\tilde{D}_2\right), \end{aligned} \tag{152}$$

for all $k \geq 1$, where (152) follows since $\boldsymbol{V}$ is an i.i.d. sequence of $\mathcal{N}(0, \tilde{D}_2)$ random variables independent of $(\boldsymbol{X}, U, \Theta_1)$ and hence independent of K.

Combining (149), (151), and (152) and observing that $s(a_k; D_1, \tilde{D}_2, \mu_k) = c^{(k)} b_2^{(k)}/\tilde{D}_2$, we see that the LHS of (110) is at most $E_K\left[\frac{1}{2}\log\big(1 + s(a_K; D_1, \tilde{D}_2, \mu_K)\big)\right]$. To complete the proof, we note that this expression is upper bounded by the RHS of (110) by the definition of $C^*(D_1, D_2, \sigma_u^2)$ (8); this bound is not necessarily tight since a_k does not necessarily achieve the maximum in (8) for D_1, $\tilde{D}_2$ and μ_k. □

A.6 Proof of Lemma 5.3

First, note that the contrapositive (and hence equivalent) statement of property P2 is:

P2a. For every $\epsilon > 0$, $n > n_0(\epsilon)$, and event $\mathcal{E}$, if $\Pr(\mathcal{E}) \geq \delta(\epsilon, n)$, then $E\left[n^{-1}\|\boldsymbol{U}\|^2|\mathcal{E}\right] \leq \sigma_u^2 + 5\epsilon$.

Let us define

$$S_{U^2,n} = \frac{1}{n}\sum_{i=1}^{n} U_i^2, \tag{153}$$

and

$$m_{U^2} = E\left[U_i^2\right].$$

Since $\boldsymbol{U}$ is stationary, m_{U^2} does not depend on i and $E\left[S_{U^2,n}\right] = m_{U^2}$ for all n. Further recall the assumption that $m_{U^2} \leq \sigma_u^2$. We first prove the claim assuming that $S_{U^2,n}$ has a density for all n, and return later to the case when it does not. Fix $\epsilon > 0$, and choose $n_0(\epsilon)$ such that

$$\mathrm{Var}(S_{U^2,n}) \leq \epsilon^2/2, \;\; \forall n > n_0(\epsilon). \tag{154}$$

This can be done since $\boldsymbol{U}$ is ergodic with finite fourth moment, and hence $S_{U^2,n}$ is converging in mean square to m_{U^2}. Next, choose $\{s_n\}$ such that for all $n > n_0(\epsilon)$

$$\Pr(S_{U^2,n} \geq s_n) = \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon^2}, \tag{155}$$

and

$$m_{U^2} - \epsilon \leq s_n \leq m_{U^2} + \epsilon. \tag{156}$$

Such an s_n exists for all appropriate n by the intermediate value theorem of calculus because our assumption that $S_{U^2,n}$ has a density guarantees that $\Pr(S_{U^2,n} \geq \xi)$ is continuous in ξ, and because

$$\Pr\left(S_{U^2,n} \geq m_{U^2} + \epsilon\right) \leq \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon^2},$$

and

$$\begin{aligned}\Pr\left(S_{U^2,n} \geq m_{U^2} - \epsilon\right) &\geq 1 - \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon^2} \\ &\geq \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon^2},\end{aligned}$$

which follow from Chebyshev's inequality and (154).

From (155) it follows that the choice

$$\delta(\epsilon, n) = \Pr(S_{U^2,n} \geq s_n), \tag{157}$$

guarantees Property P1, because $\mathrm{Var}(S_{U^2,n})$ approaches zero. We now show that with this choice of $\delta(\epsilon, n)$, Property P2a is also satisfied. Let the event $\mathcal{E}$ satisfy $\Pr(\mathcal{E}) \geq \delta(\epsilon, n)$ so that by (157),

$$\Pr(\mathcal{E}) \geq \Pr(S_{U^2,n} \geq s_n). \tag{158}$$

Then,

$$\begin{aligned}E\left[S_{U^2,n}|\mathcal{E}\right] &= \int_0^\infty \Pr(S_{U^2,n} \geq t|\mathcal{E})\, dt \\ &= \frac{1}{\Pr(\mathcal{E})}\left(\int_0^{s_n} \Pr(S_{U^2,n} \geq t, \mathcal{E})\, dt + \int_{s_n}^\infty \Pr(S_{U^2,n} \geq t, \mathcal{E})\, dt\right) \\ &\leq \frac{1}{\Pr(\mathcal{E})}\left(\int_0^{s_n} \Pr(\mathcal{E})\, dt + \int_{s_n}^\infty \Pr(S_{U^2,n} \geq t)\, dt\right) \\ &\leq s_n + \frac{1}{\Pr(S_{U^2,n} \geq s_n)} \int_{s_n}^\infty \Pr(S_{U^2,n} \geq t)\, dt,\end{aligned}$$

where the first equality follows since $S_{U^2,n}$ is a non-negative random variable and the final inequality follows by (158). Furthermore, for $n > n_0(\epsilon)$,

$$\begin{aligned}\int_{s_n}^\infty \Pr(S_{U^2,n} \geq t)\, dt &= \int_{s_n}^{s_n+2\epsilon} \Pr(S_{U^2,n} \geq t)\, dt + \int_{s_n+2\epsilon}^\infty \Pr(S_{U^2,n} \geq t)\, dt \\ &\leq 2\epsilon \Pr(S_{U^2,n} \geq s_n) + \int_{s_n+2\epsilon}^\infty \frac{\mathrm{Var}(S_{U^2,n})}{(t - m_{U^2})^2}\, dt \\ &= 2\epsilon \Pr(S_{U^2,n} \geq s_n) + \frac{\mathrm{Var}(S_{U^2,n})}{s_n + 2\epsilon - m_{U^2}} \\ &\leq 2\epsilon \Pr(S_{U^2,n} \geq s_n) + \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon},\end{aligned}$$

where the first inequality follows since $\Pr(S_{U^2,n} \geq t)$ is non-increasing in t and by Chebyshev's inequality, and the final inequality is valid by (156). Therefore,

$$E\left[S_{U^2,n}|\mathcal{E}\right] \leq s_n + 2\epsilon + \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon \Pr(S_{U^2,n} \geq s_n)} \leq m_{U^2} + 4\epsilon,$$

where the final inequality follows by (155) and (156). This concludes the proof in the case where $S_{U^2,n}$ has a density.

We now return to the case when $S_{U^2,n}$ does not necessarily have a density. Fix $\epsilon > 0$, and let $Z_k = U_k^2 + \Xi_k$, for all $k \geq 1$, where $\Xi_1, \Xi_2, \ldots$ is an i.i.d. sequence of exponential random variables with mean ϵ independent of $\boldsymbol{U}$. Since $\boldsymbol{U}$ is ergodic, $\boldsymbol{Z}$ is also ergodic. Furthermore, $S_{Z,n} = n^{-1} \sum_{k=1}^{n} Z_k$ has a density, and thus the above results hold for $S_{Z,n}$. In particular, we can choose $\{s_n\}$ and $n_0(\epsilon)$ such that $\Pr(S_{Z,n} \geq s_n) \to 0$ and such that $\Pr(\mathcal{E}) \geq \Pr(S_{Z,n} \geq s_n)$ and $n > n_0(\epsilon)$ imply that

$$\begin{aligned} E\left[S_{Z,n}|\mathcal{E}\right] &\leq m_Z + 4\epsilon \\ &= m_{U^2} + 5\epsilon. \end{aligned}$$

We complete the proof by noting that $S_{U^2,n} \leq S_{Z,n}$ a.s. and thus $E\left[S_{U^2,n}|\mathcal{E}\right] \leq E\left[S_{Z,n}|\mathcal{E}\right]$ for any event $\mathcal{E}$ with non-zero probability. □

Acknowledgments

The authors would like to thank N. Merhav and A. Baruch for fruitful discussions and helpful comments.

References

[1] A. S. Cohen and A. Lapidoth, "The capacity of the vector Gaussian watermarking game," in *Proc. of ISIT*, (Washington, DC), p. 5, 2001.

[2] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, pp. 439–441, May 1983.

[3] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk, "Watermarking digital image and video data: A state-of-the-art overview," *IEEE Signal Processing Magazine*, vol. 17, pp. 20–46, Sept. 2000.

[4] S. Katzenbeisser and F. A. P. Petitcolas, eds., *Information Hiding Techniques for Steganography and Digital Watermarking*. Computer Security Series, Boston: Arthouse Tech, 2000.

[5] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding – a survey," *Proceedings of the IEEE*, vol. 87, pp. 1062–1078, July 1999.

[6] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technology," *Proceedings of the IEEE*, vol. 86, pp. 1064–1087, June 1998.

[7] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information theoretic analysis of steganography," in *Proc. of ISIT*, (Cambridge, MA), p. 297, 1998.

[8] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding." Preprint, available at `http://www.ifp.uiuc.edu/~moulin/paper.html`, 1999.

[9] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," in *Proc. of ISIT*, (Sorrento, Italy), p. 19, 2000.

[10] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, pp. 420–430, Mar. 2000.

[11] A. Somekh-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," in *Proc. of ISIT*, (Washington, DC), p. 7, 2001.

[12] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1410–1422, May 2001.

[13] S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, "Capacity issues in digital image watermarking," in *Proc. of the Inter. Conf. on Image Processing*, 1998.

[14] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.

[15] D. Karakos and A. Papamarcou, "Relationship between quantization and distribution rates of digitally watermarked data," in *Proc. of ISIT*, (Sorrento, Italy), p. 47, 2000.

[16] T. M. Cover, "Conflict between state information and intended information," in *Information Theory Workshop*, (Mestovo, Greece), p. 21, 1999.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

[18] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1529, Sept. 1996.

[19] A. S. Cohen, *Information Theoretic Analysis of Watermarking Systems*. PhD thesis, MIT, Cambridge, MA, 2001.

[20] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications." Preprint, Jan. 2000.

[21] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inform. Theory*, vol. 25, pp. 306–311, May 1979.

[22] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[23] C. Heegard and A. A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. 29, pp. 731–739, Sept. 1983.

[24] B. Hughes and P. Narayan, "Gaussian arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 33, pp. 267–284, Mar. 1987.

[25] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Trans. Inform. Theory*, vol. 34, pp. 27–34, Jan. 1988.

[26] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, vol. 87, pp. 1127–1141, July 1999.

[27] J. Wolfowitz, *Coding Theorems in Information Theory*. Spring-Verlag, third ed., 1978.

[28] J. A. O'Sullivan, "Some properties of optimal information hiding and information attacks," in *Proc. of the Allerton Conf. on Comm., Control and Computing*, 2001.

[29] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.

[30] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714–722, May 1995.

[31] R. Gibbons, *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press, 1992.

[32] B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems*. PhD thesis, MIT, Cambridge, MA, 2000.

[33] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, "Writing on colored paper," in *Proc. of ISIT*, (Washington, DC), 2001.

[34] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice-strategies for cancelling known interference," in *Proc. of the Cornell Summer Workshop on Inform. Theory*, Aug. 2000.

[35] R. Zamir, S. Shamai, and U. Erez, "Nested codes: An algebraic binning scheme for noisy multiterminal networks." Submitted to *IEEE Trans. Info. Theory*, Oct. 2001.

[36] W. Hirt and J. L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inform. Theory*, vol. 34, no. 3, pp. 380–388, 1988.

[37] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *The Bell System Technical Journal*, vol. 38, pp. 611–656, May 1959.

[38] A. D. Wyner, "Random packings and coverings of the unit n-sphere," *The Bell System Technical Journal*, vol. 46, pp. 2111–2118, Nov. 1967.

[39] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Spring-Verlag, 1988.

[40] S. Shamai, S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 564–579, Mar. 1998.

[41] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fibre communications," *Nature*, vol. 411, pp. 1027–1030, June 2001.