Information Theoretic Analysis of Watermarking Systems

by

Aaron Seth Cohen

S.B., Electrical Engineering and Computer Science
Massachusetts Institute of Technology (1997)

M.Eng., Electrical Engineering and Computer Science
Massachusetts Institute of Technology (1997)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2001

Author .....
Department of Electrical Engineering and Computer Science
August 31, 2001

Certified by .....
Amos Lapidoth
Associate Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Information Theoretic Analysis of Watermarking Systems

by

Aaron Seth Cohen

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Watermarking models a copyright protection mechanism where an original data sequence is modified before distribution to the public in order to embed some extra information. The embedding should be transparent (i.e., the modified data should be similar to the original data) and robust (i.e., the information should be recoverable even if the data is modified further). In this thesis, we describe the information-theoretic capacity of such a system as a function of the statistics of the data to be watermarked and the desired level of transparency and robustness. That is, we view watermarking from a communication perspective and describe the maximum bit-rate that can be reliably transmitted from encoder to decoder.

We make the conservative assumption that there is a malicious attacker who knows how the watermarking system works and who attempts to design a forgery that is similar to the original data but that does not contain the watermark. Conversely, the watermarking system must meet its performance criteria for any feasible attacker and would like to force the attacker to effectively destroy the data in order to remove the watermark. Watermarking can thus be viewed as a dynamic game between these two players who are trying to minimize and maximize, respectively, the amount of information that can be reliably embedded.

We compute the capacity for several scenarios, focusing largely on Gaussian data and a squared difference similarity measure. In contrast to many suggested watermarking techniques that view the original data as interference, we find that the capacity increases with the uncertainty in the original data. Indeed, we find that out of all distributions with the same variance, a Gaussian distribution on the original data results in the highest capacity. Furthermore, for Gaussian data, the capacity increases with its variance.

One surprising result is that with Gaussian data the capacity does not increase if the original data can be used to decode the watermark. This is reminiscent of a similar model, Costa's "writing on dirty paper", in which the attacker simply adds independent Gaussian noise. Unlike with a more sophisticated attacker, we show that the capacity does not change for Costa's model if the original data is not Gaussian.

Thesis Supervisor: Amos Lapidoth
Title: Associate Professor of Electrical Engineering

Acknowledgments

First, I would like to thank my advisor, Prof. Amos Lapidoth, for all of his advice and encouragement. Even from long distance, he has graciously helped me overcome the many hurdles I have encountered throughout the writing of this thesis.

I would also like to thank the members of my thesis committee, Prof. Bob Gallager and Prof. Greg Wornell, for their insightful comments. The many comments by Neri Merhav and Anelia Somekh-Baruch have also been extremely helpful.

My brief visit to the ISI in Zürich was quite pleasant, thanks mainly to the friendly and accommodating students there. I would specifically like to thank Renate Agotai for taking care of all of the important details and Ibrahim Abou Faycal for being a great traveling companion.

The intellectual environment created by the faculty, staff and students of LIDS has made it a wonderful place to pursue a graduate education. I would especially like to thank my many office-mates in 35-303 including Randy Berry, Anand Ganti, Hisham Kassab, Thierry Klein, Emre Koksal, Mike Neely, Edmund Yeh and Won Yoon; the office was always a good place for getting feedback about research or (more typically) discussing such important topics as sports or the stock market.

Of course, I could not have done any of this without my parents, Michael and Jackie, who have always encouraged me to vigorously pursue my goals. I would like to thank them and the rest of my family for the support they have given me over the years.

My final and most important acknowledgment goes to my lovely wife Dina Mayzlin. She has motivated and inspired me, and she has made the many years we have spent together at MIT worthwhile.

This research was supported in part by an NSF Graduate Fellowship.

Contents

List of Figures

Chapter 1

Introduction

Watermarking can model situations where source sequences need to be copyright-protected before distribution to the public. The copyright needs to be embedded in the distributed version so that no adversary with access to the distributed version will be able produce a forgery that resembles the original source sequence and yet does not contain the embedded message. The watermarking process should, of course, introduce limited distortion so as to guarantee that the distributed sequence closely resembles the original source sequence. The original source sequence can be any type of data such as still image, audio or video that can be modified slightly and still maintain its inherent qualities.

Watermarking research has exploded over the past several years. For example, see [KP00b, LSL00, PAK99, SKT98] and their extensive references. This interest has stemmed from the ease by which data can now be reproduced and transmitted around the world, which has increased the demand for copyright protection. Furthermore, ordinary encryption is not sufficient since, in order to be enjoyed by the public, the data must be accessed at some point. Thus, there is a need to embed information directly in the distributed data, which is precisely what a watermarking system does. Much of the work on watermarking has focused on designing ad-hoc systems and testing them in specific scenarios. Relatively little work has been done in assessing the fundamental performance trade-offs of watermarking systems. In this thesis, we seek to describe these performance trade-offs.

The main requirements for a watermarking system are *transparency* and *robustness*. For copyright protection, transparency means that the distributed data should be "similar" to the original data, while robustness means that the embedded information should be

recoverable from any forgery that is "similar" to the distributed data. Another way of thinking about robustness is that only by destroying the data could a pirate remove the copyright. We formalize these requirements by specifying a distortion measure and claiming that two data sequences are "similar" if the distortion between them is less than some threshold. The threshold for transparency is in general different from the threshold for robustness.

In this thesis, we view watermarking as a communication system and we seek to find the trade-off between transparency, robustness, and the amount of information that can be successfully embedded. In particular, we find the information-theoretic *capacity* of a watermarking system depending on the transparency and robustness thresholds and the statistical properties of the data to be watermarked.

For a general watermarking system, the data distributed to the public and the data that is used to recover the embedded information will be different. This difference might be caused by a variety of signal processing techniques, e.g., photocopying or cropping an image and recording or filtering an audio clip. Instead of making any assumptions on the type of signal processing that will take place, we make the conservative assumption that there is a malicious attacker whose sole intent is to disrupt the information flow. For example, a pirate might wish to remove copyright information in order to make illegal copies. In order to please his customers, this pirate would also want the illegal copies to be of the highest quality possible. Conversely, the watermarking system wishes to ensure that the act of removing the embedded information causes the data to be unusable to such a pirate. We can thus think of watermarking as a game between two players: a communication system (encoder and decoder) and an attacker. The players are trying to respectively maximize and minimize the amount of information that can be embedded.

We assume throughout the thesis that the attacker can only use one watermarked version of the original data sequence to create a forgery. That is, we assume one of following two things about the system. The first possible assumption is that only one watermark will be embedded in each original data sequence. Such an assumption is plausible if the watermark only contains copyright information. The second possible assumption is that even though many different versions exist, an attacker can only access one of them. This assumption is reasonable if the cost of obtaining multiple copies is prohibitively high. A system that has to deal with attackers with different watermarked copies is usually called a fingerprint-

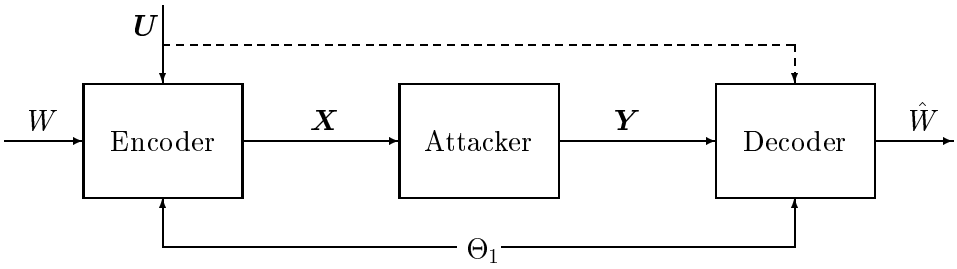

Figure 1-1: A diagram of watermarking. The dashed line is used in the private version, but not in the public version.

ing system, although the terms "watermarking" and "fingerprinting" are sometimes used interchangeably. See Section 2.3.3 for more about fingerprinting.

We consider two versions of watermarking. In the *private* version, the decoder can use both the forgery and the original data sequence recover the embedded information. In the *public* version, the decoder must recover the embedded information from the forgery alone. The private version is sometimes called non-oblivious or non-blind watermarking and the public version is sometimes called oblivious or blind watermarking. The private version is easier to analyze and is applicable, for example, when a copyright system is centralized. The public version is more difficult to analyze, but it is applicable in a much wider context. Surprisingly, we find that when the original data is Gaussian, then the capacity is the same for both versions. However, the capacity-achieving technique is more complex for the public version.

Our model of the watermarking game is illustrated in Figure 1-1 and can be described briefly as follows. A more thorough mathematical model is given below in Section 2.1. The first player consists of the encoder and decoder who share a secret key Θ_1 that allows them to implement a randomized strategy. The attacker is the second player and it is assumed to have full knowledge of the first player's strategy. We now discuss the functions of each of the encoder, the attacker and the decoder. The encoder takes the original data sequence $\boldsymbol{U}$ (which we will call the "covertext") and the watermark W and produces the "stegotext" $\boldsymbol{X}$ for distribution to the public. The encoder must ensure that the covertext and the stegotext are similar according to the given distortion measure. The attacker produces a forgery $\boldsymbol{Y}$ from the stegotext, and he must also ensure that the forgery and the stegotext are similar according to the given distortion measure. Finally, the decoder uses the forgery (in the

public version) or both the forgery and the covertext (in the private version) in order to produce an estimate of the watermark $\hat{W}$. Although the encoder, attacker and decoder act in that order, it is important to remember that the encoder and decoder are designed first and then the attacker is designed with knowledge of how the encoder and decoder work, but not with knowledge of the realizations of their inputs.

Although we have and will use copyright protection as the main watermarking application, a modified watermarking model can be used to describe several other scenarios. For example, the covertext could be a signal from an existing transmission technique (e.g., FM radio) and the watermark could be supplemental digital information [CS99]. The stegotext produced by the encoder would be the signal that is actually transmitted. Since the transmitted signal is required to be similar to the original signal, existing receivers will still work while newer (i.e., more expensive) receivers will be able to decode the supplemental information as well. For this example, instead of an active attacker that arbitrarily modifies the stegotext, it is more reasonable to say that the received signal is simply the transmitted signal plus independent ambient noise. This modified watermarking model can also be used to analyze a broadcast channel (i.e., one transmitter, many receivers) [CS01]. In this case, the transmitter can use its knowledge of the signal it is sending to one user to design the signal it is simultaneously sending to another user. The modified watermarking model with Gaussian covertext and Gaussian ambient noise is also known as "writing on dirty paper" [Cos83]; see Section 2.5.4 for more on this model including two extensions.

We conclude the introduction by considering an example watermarking system. Let's say that the rock band The LIzarDS has created a new hit song (this corresponds to our covertext $\boldsymbol{U}$). Instead of directly releasing the song to the public, the band submits it to a watermarking system. This system takes the original song and the watermark (e.g., song title, artist's name, etc.) and produces a version that will be distributed to the public (this is our stegotext $\boldsymbol{X}$). To respect the band's artistic integrity, the distributed version should be similar to the original version (hence, our transparency requirement). Whenever the song is played on the radio, the watermarking system could decode the watermark and ensure that the proper royalty is paid to the artist. The system could also block copying over the Internet based on the contents of the watermark, as the music industry would like to happen. Finally, the watermarking system would like to be able to recover the information in the watermark even if the distributed song has been altered, but is still essentially the

same (hence, our robustness requirement). Note that the watermarking system is primarily interested in the information embedded in the watermark and is only indirectly interested in the song itself through the transparency and robustness requirements. In other words, the part of the watermarking system that listens to the song is only required to extract the watermark and is not required to improve the quality of the song.

1.1 Outline of Thesis

This thesis is organized as follows. In Chapter 2, we give a precise definition of watermarking and give our results on the capacity of a watermarking system for several scenarios. We also compare our watermarking model to other watermarking research and to two well-studied information theoretic problems – communication with side information and the arbitrarily varying channel. We conclude this chapter with two extensions of a communication with side information problem, Costa's writing on dirty paper, which is similar to our watermarking model. In Chapter 3, we define and solve two mutual information games which are related to the private and public versions of watermarking. Chapters 4, 5 and 6 are devoted to proving the main watermarking capacity results. In Chapter 7, we give some conclusions and some directions for future work.

1.2 Notation

We use script letters, e.g., $\mathcal{U}$ and $\mathcal{X}$, to denote sets. The n-th Cartesian product of a set $\mathcal{U}$ (e.g., $\mathcal{U} \times \mathcal{U} \times \cdots \times \mathcal{U}$) is written $\mathcal{U}^n$. Random variables and random vectors are written in upper case, while their realizations are written in lower case. Unless otherwise stated, 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).

For real vectors, 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.

Chapter 2

Watermarking Model and Results

This chapter is organized as follows. In Section 2.1, we give precise definitions of watermarking and its information-theoretic capacity. In Section 2.2, we summarize our main results on the watermarking capacity for six different scenarios. In Section 2.3, we compare our model and results to prior work that has been done on watermarking. In Section 2.4, some of the assumptions we have made in our watermarking model are discussed, with an emphasis on which assumptions can be dropped and which need improvement. In Section 2.5, we show that watermarking can be thought of as a combination of two well-studied information-theoretic problems: communication with side information and the arbitrarily varying channel. In Section 2.5.4, we consider a specific communication with side information model – Costa's writing on dirty paper – and describe two extensions to this model.

2.1 Precise Definition of Watermarking

We now give a more detailed description of our watermarking model. Recall that this model is illustrated in Figure 1-1 above.

Prior to the use of the watermarking system, a *secret key* (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. 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.

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 nonnegative function. We require that the encoder satisfy

$$d_1(\boldsymbol{U}, \boldsymbol{X}) \leq D_1, \text{ a.s.}, \tag{2.1}$$

where $D_1 > 0$ is a given constant called the *encoder distortion level*, and a.s. stands for "almost surely", i.e., with probability 1. We will also consider an average distortion constraint on the encoder; see Section 2.2.3.

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.,

$$g_n : (\boldsymbol{x}, \theta_2) \mapsto \boldsymbol{y} \in \mathcal{Y}^n. \tag{2.2}$$

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., $\boldsymbol{Y} = g_n(\boldsymbol{X}, \Theta_2)$. 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 nonnegative function. The attacker is required to satisfy

$$d_2(\boldsymbol{X}, \boldsymbol{Y}) \leq D_2, \text{ a.s.}, \tag{2.3}$$

where $D_2 > 0$ is a given constant called the *attacker distortion level*. We will also consider an average distortion constraint on the attacker; see Section 2.2.3.

In the public version of watermarking, 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

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

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 watermarking 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.

The *coding capacity* of watermarking is the supremum of all achievable rates. It depends on five parameters: the encoder distortion function $d_1(\cdot,\cdot)$ and level D_1, the attacker distortion function $d_2(\cdot,\cdot)$ and level D_2, and the covertext distribution $\{P_{\boldsymbol{U}}\}$. The distortion functions will be made obvious from context, and thus we write the generic coding capacity of watermarking 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.

2.2 Capacity Results for Watermarking

In this section, we describe the capacity of watermarking under various assumptions on the covertext distribution, distortion constraints, and attacker capabilities. We find the capacity for the standard watermarking model of Section 2.1 when the covertext distribution is IID scalar Gaussian (Section 2.2.1), IID vector Gaussian (Section 2.2.4) and IID Bernoulli(1/2) (Section 2.2.6). We deviate from the standard model by considering an attacker that only

has to meet the distortion constraint in expectation (Section 2.2.3) and an attacker that can only inject additive noise (Section 2.2.2). Finally, we find a general formula for the capacity when no covertext is present (Section 2.2.5). The detailed proofs of all of these results can be found in later chapters; we present a proof sketch and a reference to the detailed proof following each result.

2.2.1 Scalar Gaussian Watermarking Game

We now consider a watermarking system where all of the alphabets are the real line (i.e., $\mathcal{U} = \mathcal{X} = \mathcal{Y} = \mathbb{R}$) and where the distortion measures for both the encoder and attacker will be squared error, i.e., $d_1(u, x) = (x - u)^2$ and $d_2(x, y) = (y - x)^2$. Of particular interest is when the covertext $\boldsymbol{U}$ is a sequence of independent and identically distributed (IID) random variables of law $\mathcal{N}(0, \sigma_u^2)$, i.e., zero-mean Gaussian. We refer to the *scalar Gaussian watermarking (SGWM) game* when the distortion constraints and covertext distribution are as specified above. Surprisingly, we find that the capacity of the SGWM game is the same for the private and public versions. Furthermore, we show that for all stationary and ergodic covertext distributions, the capacity of the watermarking game is upper bounded by the capacity of the SGWM game.

To state our results on the capacity of the SGWM game we need to define the interval

$$\mathcal{A}(D_1, D_2, \sigma_u^2) = \left\{ A : \max\left\{ D_2, \left(\sigma_u - \sqrt{D_1}\right)^2 \right\} \leq A \leq \left(\sigma_u + \sqrt{D_1}\right)^2 \right\}, \tag{2.4}$$

and the mappings

$$s(A; D_1, D_2, \sigma_u^2) = \frac{D_1}{D_2}\left(1 - \frac{D_2}{A}\right)\left(1 - \frac{(A - \sigma_u^2 - D_1)^2}{4 D_1 \sigma_u^2}\right), \tag{2.5}$$

and[1]

$$C^*(D_1, D_2, \sigma_u^2) = \begin{cases} \max_{A \in \mathcal{A}(D_1, D_2, \sigma_u^2)} \frac{1}{2} \log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr) & \text{if } \mathcal{A}(D_1, D_2, \sigma_u^2) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}. \tag{2.6}$$

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

Note that a closed-form solution for (2.6) 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; see Lemma A.1. Further note that $C^*(D_1, D_2, \sigma_u^2)$ is zero only if $D_2 \geq \sigma_u^2 + D_1 + 2\sigma_u\sqrt{D_1}$.

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 IID 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.

Theorem 2.1. *For the watermarking game with real alphabets and squared error distortion measures, if $\{P_{\boldsymbol{U}}\}$ defines an ergodic covertext $\boldsymbol{U}$ such that $E\left[U_k^4\right] < \infty$ and $E\left[U_k^2\right] \leq \sigma_u^2$, then*

$$\begin{aligned} C_{\text{pub}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C_{\text{priv}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) && (2.7) \\ &\leq C^*(D_1, D_2, \sigma_u^2). && (2.8) \end{aligned}$$

Equality is achieved in both (2.7) and (2.8) if $\boldsymbol{U}$ is an IID Gaussian sequence with mean zero and variance σ_u^2, i.e. if $P_{\boldsymbol{U}} = (\mathcal{N}(0, \sigma_u^2))^n$ for all n.

This theorem shows that, of all ergodic covertexts with a given power, the IID 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 2.5.2), this result differs from usual "Gaussian is the worst-case additive noise" idea, see e.g., [CT91, Lap96]. The basic reason that a Gaussian covertext is the best case 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) out of all distributions with the same second moment.

As an example, consider an IID covertext in which each sample U_k is either $-\sigma_u$ or $+\sigma_u$ with probability 1/2 each, 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, let us further consider an attacker that creates the forgery by quantizing each stegotext sample X_k to the nearest of $-\sigma_u$ or $+\sigma_u$. Even in the private version, the only way the encoder could send information is by changing U_k by at least σ_u, and the encoder can do this for only a small percentage of the samples since $D_1 \ll \sigma_u^2$.

Indeed, using the results of Section 2.2.6 on the binary watermarking game, we see that the largest achievable rate for this fixed attacker is[2] $H_b\left(D_1/\sigma_u^2\right)$ bits/symbol, which is smaller than 1/2 bits/symbol for $D_1/\sigma_u^2 < 0.11$, i.e., the regime of interest. Note that the capacity for this scenario is even smaller since we have only considered a known attacker.

We also find that the capacity of the SGWM game is *increasing* in σ_u^2; see Figure 2-1. Thus, again we see that the greater the uncertainty in the covertext the more bits the watermarking system can hide in it.

Another interesting aspect of this theorem is that, as in the "writing on dirty paper" model (see Section 2.5.4 below and [Cos83]), the capacity of the SGWM game is unaffected by the presence or absence of side-information (covertext) at the receiver. See [Cov99] for some comments on the role of receiver side-information, particularly in card games.

Moulin and O'Sullivan [MO99, MO00] give a capacity for the SGWM game that is strictly smaller than $C^*(D_1, D_2, \sigma_u^2)$. In particular, they claim that the capacity is given by $\frac{1}{2}\log(1+s(A; D_1, D_2, \sigma_u^2))$ when A is *fixed* to $\sigma_u^2+D_1$ instead of optimized over $\mathcal{A}(D_1, D_2, \sigma_u^2)$ as in (2.6), while the optimal A is strictly larger than $\sigma_u^2 + D_1$; see Lemma A.1. This difference is particularly noticeable when $\sigma_u^2 + D_1 < D_2 < \sigma_u^2 + D_1 + 2\sigma_u\sqrt{D_1}$, since $C^*(D_1, D_2, \sigma_u^2) > 0$ in this range while the capacity given in [MO99, MO00] is zero. An example of the two capacity expressions is plotted in Figure 2-1. Both capcity expressions are bounded above by $\frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right)$ and approach this bound as σ_u^2 increases. Note that the watermarking game here is defined differently than in [MO99, MO00], but we believe that the capacity of the SGWM game should be the same for both models. Indeed, the general capacity expression in [MO99, MO00] is similar to our mutual information game (see Chapter 3), and we find that the value of the mutual information game for a Gaussian covertext is also $C^*(D_1, D_2, \sigma_u^2)$ (see Theorem 3.1).

Theorem 2.1 is proved in Chapter 4 in two steps: a proof of achievability for Gaussian covertexts and a converse for general covertexts. Although achievability for the public version implies achievability for the private version, we give separate proofs for the private version (Section 4.2) and the public version (Section 4.3). We have chosen to include both proofs because the coding technique for the private version has a far lower complexity (than the coding technique for the public version) and may give some insight into the design of practical watermarking systems for such scenarios. We now provide a sketch of the proof.

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

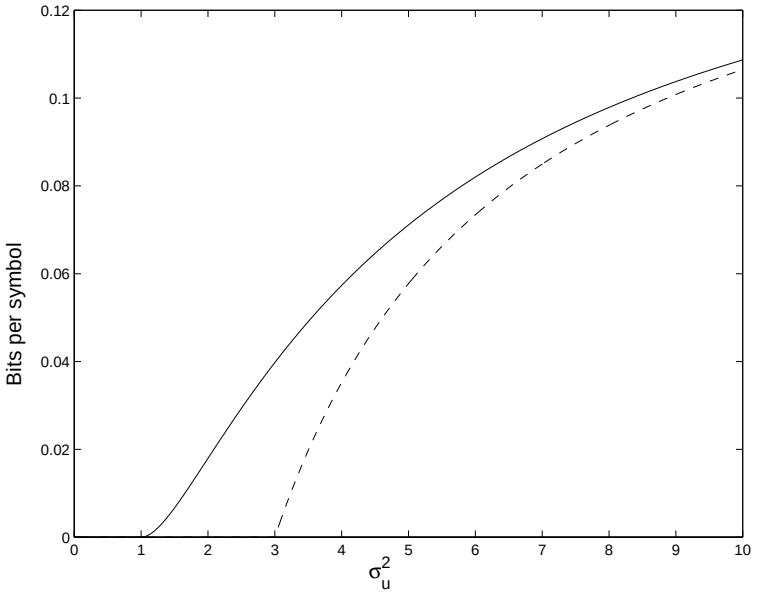

Figure 2-1: Scalar Gaussian watermarking capacity versus σ_u^2 with $D_1 = 1$ and $D_2 = 4$. The dashed line is the capacity expression from [MO99, MO00].

Achievability

We now argue that for a Gaussian covertext all rates less than $C^*(D_1, D_2, \sigma_u^2)$ are achievable in the public version; see Section 4.3 for a full proof. This will also demonstrate that all such rates are achievable in the private version as well. The parameter A corresponds to the desired power in the covertext, i.e., our coding strategy will have $n^{-1}\|\boldsymbol{X}\| \approx A$. We now describe a coding strategy that depends on A (and the given parameters D_1, D_2 and σ_u^2) and can achieve all rates up to $\frac{1}{2}\log(1 + s(A; D_1, D_2, \sigma_u^2))$. Hence, all rates less than $C^*(D_1, D_2, \sigma_u^2)$ are achievable with the appropriate choice of A. The coding strategy is motivated by the works of Marton [Mar79], Gel'fand and Pinsker [GP80], Heegard and El Gamal [HEG83], and Costa [Cos83]. The encoder/decoder pair use their common source of randomness to generate a codebook consisting of 2^{nR_1} IID codewords that are partitioned into 2^{nR} bins of size 2^{nR_0} each (hence, $R = R_1 - R_0$). Each codeword is uniformly distributed on an n-sphere with radius depending on A. Given the covertext $\boldsymbol{u}$ and the watermark w, the encoder finds the codeword in bin w that is closest (in Euclidean distance) to $\boldsymbol{u}$. Let $\boldsymbol{v}_w(\boldsymbol{u})$ be the chosen codeword. The encoder then forms the stegotext as a linear combination of

the chosen codeword and the covertext,

$$\boldsymbol{x} = \boldsymbol{v}_w(\boldsymbol{u}) + (1-\alpha)\boldsymbol{u},$$

where α is a constant that depends on A. The distortion constraint will be met with high probability if R_0 is large enough. The decoder finds the closest codeword (out of all 2^{nR_1} codewords) to the forgery, and estimates the watermark as the bin of this closest codeword. If R_1 is small enough, then the probability of error can be made arbitrarily small. The two constraints on R_0 and R_1 combine to give the desired bound on the overall rate R.

Converse

We now argue that no rates larger than $C^*(D_1, D_2, \sigma_u^2)$ are achievable in the private version for any ergodic covertext distribution with power at most σ_u^2; see Section 4.5 for a full proof. The main idea is to show using a Fano-type inequality that in order for the probability of error to tend to zero, a mutual information term must be greater than the watermarking rate. The mutual information term of interest is roughly $I(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{U})$, which is related to the capacity with side information at the encoder and decoder; see Section 2.5.2. A consequence of this proof is that these rates are not achievable even if the decoder knew the statistical properties of the attacker. The basic attacker that guarantees that the mutual information will be small is based on the Gaussian rate distortion forward channel. That is, such an attacker computes A (i.e., the power in the stegotext) and implements the channel that minimizes the mutual information between the stegotext and the forgery subject to a distortion constraint, assuming that the stegotext were an IID sequence of mean-zero variance-A Gaussian random variables. The method that the attacker uses to compute A is critical. If A is the average power of the stegotext (averaged over all sources of randomness), then the mutual information will be small but the attacker's a.s. distortion constraint might not be met. If A is the power of the realization of the stegotext, then the a.s. distortion constraint will be met but the encoder could potentially use A to transmit extra information. A strategy that avoids both of these problems is to compute A by quantizing the power of the realization of the stegotext to one of finitely many values. This attacker will both meet the distortion constraint (if the quantization points are dense enough) and prevent the encoder from transmitting extra information.

2.2.2 Additive Attack Watermarking Game

In this section, we describe a variation of the watermarking game for real alphabets and squared error distortions, 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 suboptimal for the attacker to produce the forgery by combining the stegotext with a power-limited jamming sequence generated independently of the stegotext. Similarly to Costa's writing on dirty paper result (see Section 2.5.4 and [Cos83]), we will show that if the covertext $\boldsymbol{U}$ is IID 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; see Section 2.5.4 for more discussion of this extension.

In the additive attack watermarking game the attacker is more restricted than in the general game. Rather than allowing general attacks of the form (2.2), we restrict the attacker to mappings that are of the form

$$g_n(\boldsymbol{x},\theta_2) = \boldsymbol{x} + \tilde{g}_n(\theta_2) \tag{2.9}$$

for some mapping $\tilde{g}_n$. In particular, the jamming sequence

$$\tilde{\boldsymbol{Y}} = \tilde{g}_n(\Theta_2) \tag{2.10}$$

is produced independently of the stegotext $\boldsymbol{X}$, and must satisfy the distortion constraint

$$\frac{1}{n}\left\|\tilde{\boldsymbol{Y}}\right\|^2 \le D_2, \text{ a.s..} \tag{2.11}$$

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

$$\begin{aligned} C_{\text{pub}}^{\text{AA}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C_{\text{priv}}^{\text{AA}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) && (2.12) \\ &= \frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right). && (2.13) \end{aligned}$$

Equality is achieved in (2.12) if $\boldsymbol{U}$ *is an IID Gaussian sequence.*

We first sketch the converse for both versions. An IID mean-zero, variance-D_2 Gaussian sequence $\tilde{\boldsymbol{Y}}$ does not satisfy (2.11). However, for any $\delta > 0$, an IID mean-zero, variance-$(D_2 - \delta)$ Gaussian sequence $\tilde{\boldsymbol{Y}}$ satisfies $n^{-1}\|\tilde{\boldsymbol{Y}}\| \leq D_2$ with arbitrarily large probability for sufficiently large blocklength n. Since the capacity here cannot exceed the capacity when $\boldsymbol{U}$ is absent, the capacity results on an additive white noise Gaussian channel imply that the capacity of either version is at most $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$.

We now argue that the capacity of the private version is as in the theorem. When the sequence $\boldsymbol{U}$ is known to the decoder, then the results of [Lap96] can be used to show that all rates less than $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$ are achievable using Gaussian codebooks and nearest neighbor decoding. This establishes the validity of (2.13).

To complete the proof of this theorem, we must show that $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$ is achievable in the public version of the game with IID Gaussian covertext. We present a coding strategy and demonstrate that all such rates are achievable in Chapter 4.3.

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 2.2, we see that for an IID Gaussian covertext this bound is loose. Thus, for such covertexts, it is suboptimal for the attacker in the general watermarking game to take the form (2.9). See Section 2.5.4 for more discussion on the additive attack watermarking game.

2.2.3 Average Distortion Constraints

In this section, we show that if the almost sure distortion constraints are replaced with average distortion constraint, then the capacity is typically zero. That is, we replace the a.s. constraints (2.1) and (2.3) on the encoder and attacker, respectively, with

$$E\left[d_1(\boldsymbol{U}, \boldsymbol{X})\right] \leq D_1, \tag{2.14}$$

and

$$E\left[d_2(\boldsymbol{X},\boldsymbol{Y})\right] \leq D_2, \tag{2.15}$$

where the expectations are with respect to all relevant random quantities. In particular, we have the following theorem.

Theorem 2.3. *For the watermarking game with real alphabets and squared error distortion, if the covertext* $\boldsymbol{U}$ *satisfies*

$$\liminf_{n\to\infty} E\left[\frac{1}{n}\|\boldsymbol{U}\|^2\right] < \infty, \tag{2.16}$$

and if the average distortion constraints (2.14), (2.15) are in effect instead of the a.s. distortion constraints (2.1), (2.3), 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) [HN87] and from the theory of general AVCs with constrained inputs and states [CN88a], where under average power constraints no positive rates are achievable[3].

The detailed proof of this theorem is given in Appendix B.1; the basic idea is as follows. The average power of the covertext is bounded and hence the average power of the stegotext is bounded as well. Thus, the attacker can set the forgery equal to the zero vector with some fixed probability and still meet the average distortion constraint. For this attacker, the probability of error is bounded away from zero for any positive rate. Hence, no positive rates are achievable when the attacker is only required to meet an average distortion constraint.

2.2.4 Vector Gaussian Watermarking Game

We now consider a generalization of the SGWM game, where the covertext consists of an IID sequence of zero-mean Gaussian vectors of a given covariance. This will be called the *vector Gaussian watermarking (VGWM) game*. Here, the alphabets are all the m-dimensional Euclidean space, i.e., $\mathcal{U} = \mathcal{X} = \mathcal{Y} = \mathbb{R}^m$, and the distortion measures are squared Euclidean distance, i.e., $d_1(\boldsymbol{u},\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{u}\|^2$ and $d_2(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{y}-\boldsymbol{x}\|^2$. Furthermore, the covertext is an IID sequence of m-vectors $\boldsymbol{U} = (\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n)$, where each $\boldsymbol{U}_i$ is a zero-mean Gaussian random vector with a given $m \times m$ covariance matrix $S_{\boldsymbol{u}}$. Note that the vector size m is

[3]The ϵ-capacity is, however, typically positive for $\epsilon > 0$

fixed, while the blocklength n is allowed to be arbitrarily large. We use $C_{\text{priv}}^{\text{VGWM}}(D_1, D_2, S_{\boldsymbol{u}})$ and $C_{\text{pub}}^{\text{VGWM}}(D_1, D_2, S_{\boldsymbol{u}})$ to denote the capacity of the VGWM game for the private and public versions, respectively.

Theorem 2.4. *For the vector Gaussian watermarking game,*

$$C_{\text{pub}}^{\text{VGWM}}(D_1, D_2, S_{\boldsymbol{u}}) = C_{\text{priv}}^{\text{VGWM}}(D_1, D_2, S_{\boldsymbol{u}}) \tag{2.17}$$

$$= \max_{\boldsymbol{D}_1 \geq 0 \,:\, \boldsymbol{e}^t \boldsymbol{D}_1 \leq D_1} \; \min_{\boldsymbol{D}_2 \geq 0 \,:\, \boldsymbol{e}^t \boldsymbol{D}_2 \leq D_2} \sum_{j=1}^{m} C^*(D_{1j}, D_{2j}, \sigma_j^2), \tag{2.18}$$

where $C^*(D_1, D_2, \sigma^2)$ is defined in (2.6), $(\sigma_1^2, \ldots, \sigma_m^2)$ are the eigenvalues of $S_{\boldsymbol{u}}$ and $\boldsymbol{e}$ is the m-vector containing all 1's.

This theorem is proved in detail in Chapter 5, but we now briefly describe the coding strategy that achieves the desired rates for the public version. The covariance matrix $K_{\boldsymbol{u}}$ can be diagonalized using an orthogonal transformation that does not affect the distortion. Thus, we can assume that $K_{\boldsymbol{u}}$ is diagonal so that $\boldsymbol{U}$ consists of m components, each a length-n sequence of IID zero-mean Gaussian random variables with respective variances $(\sigma_1^2, \ldots, \sigma_m^2) = \boldsymbol{\sigma}^2$. After choosing m-dimensional vectors $\boldsymbol{D}_1$, $\tilde{\boldsymbol{D}}_2$ and $\boldsymbol{A}$, the encoder encodes component j using the scalar encoder for the SGWM game (see the discussion after Theorem 2.1 and Chapter 4) based on $A = A_j$, $D_1 = D_{1j}$, $D_2 = \tilde{D}_{2j}$, and $\sigma_u^2 = \sigma_j^2$. Thus, the vector $\tilde{\boldsymbol{D}}_2$ acts as an estimate of the amount of distortion the attacker will place in each component. Every attacker is associated with a feasible $\boldsymbol{D}_2$ (not necessarily equal to $\tilde{\boldsymbol{D}}_2$), where D_{2j} describes the amount of distortion the attacker inflicts on component j. However, for the optimal choice of $\tilde{\boldsymbol{D}}_2$ by the encoder, the attacker will choose $\boldsymbol{D}_2 = \tilde{\boldsymbol{D}}_2$ in order to minimize the achievable rates. This allows us to describe the achievable rates using the simple form of (2.18).

We now discuss some aspects of this theorem, focusing on the differences and similarities between SGWM and VGWM. One major similarity is that in both cases the public and private versions have the same capacity. One major difference between the two games is that in the vector version an attacker based on the Gaussian rate distortion solution is no longer optimal, i.e., it does not necessarily prevent rates larger than capacity from being achievable. A rate-distortion based attacker calculates the second order statistics of the stegotext, and designs the attack to minimize (subject to an average distortion constraint)

the mutual information between the stegotext and the forgery, assuming that the stegotext was Gaussian. In the SGWM game, this attacker does not necessarily meet the almost sure distortion constraint, but it does prevent rates higher than capacity from being achievable. However, in the vector version, if such an attacker is used, then rates strictly larger than capacity can be achieved. See Section 5.7 for more detail. The difference is that an optimal attacker does not distribute his distortion to the different components of the stegotext using the familiar waterfilling algorithm (see e.g., [CT91]). However, having chosen the correct distortion distribution, a parallel concatenation of optimal attackers for the SGWM game (and hence a parallel concatenation of scalar Gaussian rate distortion solutions) does prevent rates larger than capacity from being achievable.

We also note that the order in which the watermarking game is played remains critical in the vector version. In particular, the max and min in (2.18) cannot be switched. We highlight the significance of this observation by restricting the encoder and attacker to parallel concatenations of optimal scalar strategies based on some vectors $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. There is no single vector $\boldsymbol{D}_2$ that the attacker could pick to ensure that no rates higher than the capacity are achieved. Instead, the attacker must use his advantage of playing second (i.e., his knowledge of the encoder's strategy) in order to accomplish this goal. This differs from the vector Gaussian arbitrarily varying channel [HN88] where the attacker (resp. encoder) can choose a distortion distribution to ensure that no rates more than (resp. all rates up to) the capacity can be achieved.

2.2.5 Discrete Alphabets, No Covertext

In this section, we examine an extreme watermarking scenario in which there is no covertext to hide the message in. In this situation, the attacker can directly modify (subject to a distortion constraint) the codeword produced by the encoder. This can be viewed as an extension of [CN88a], which found the random coding capacity of an arbitrarily varying channel (AVC) with constrained inputs and states (see Section 2.5.3 for more on the AVC). The primary difference is that in [CN88a] the inputs and states are chosen independently of each other, while here the states of the channel are chosen as a function of the input sequence.

Before stating the main result of this section, we first give our assumptions and some notation. We assume that the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. Since there is no covertext,

the distortion constraint (2.1) is replaced by $n^{-1}\sum_{i=1}^{n} d_1(X_i) \leq D_1$ a.s. for some function $d_1 : \mathcal{X} \mapsto \mathbb{R}_+$. The distortion constraint (2.3) imposed on the attacker remains the same. The lack of covertext also means that there is no distinction between the private and public versions, and thus we write the capacity for this scenario as $C^{\text{NoCov}}(D_1, D_2)$. For any distributions P_X and $P_{Y|X}$, we write $I_{P_X P_{Y|X}}(X;Y)$ to be the mutual information between random variables X and Y under joint distribution $P_X P_{Y|X}$.

Theorem 2.5. *When there is no covertext and discrete alphabets, the capacity of the watermarking game is given by*

$$C^{\text{NoCov}}(D_1, D_2) = \max_{P_X : E_{P_X}[d_1(X)] \leq D_1} \min_{P_{Y|X} : E_{P_X P_{Y|X}}[d_2(X,Y)] \leq D_2} I_{P_X P_{Y|X}}(X;Y). \tag{2.19}$$

The proof of this Theorem can be found in Section 6.1; we now briefly sketch the arguments behind the proof.

Achievability

For a fixed n, the encoder chooses a distribution P_X such that the constraint in (2.19) is satisfied and $n \cdot P_X(x)$ is an integer for every $x \in \mathcal{X}$. The encoder then generates 2^{nR} IID codewords $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{nR}}\}$, with each codeword uniformly distributed over all n-sequences whose empirical distribution is given by P_X. Given the codebook and the watermark w, the transmitted sequence is simply $\boldsymbol{x}_w$. Note that $n^{-1} d_1(\boldsymbol{x}_w) = E_{P_X}[d_1(X)] \leq D_1$, and thus the distortion constraint is satisfied. The decoder uses the maximum mutual information (MMI) decoding rule. That is, the estimate of the watermark is given by

$$\hat{w} = \arg\max_{1 \leq w' \leq 2^{nR}} I(\boldsymbol{x}_{w'} \wedge \boldsymbol{y}),$$

where $I(\boldsymbol{x} \wedge \boldsymbol{y})$ is the mutual information between random variables X and Y when they have the joint empirical distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$. The probability of error only depends on the attacker through the conditional empirical distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$. Using techniques from [CK81], we can show that the probability of error goes to zero as long as the rate R is less than $I(\boldsymbol{x}_w \wedge \boldsymbol{y})$ for the correct watermark w. Finally, the conditional empirical distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ must satisfy the constraint in (2.19) in order for the attacker to meet his distortion constraint, and thus the encoder can guarantee that the score of the

correct codeword $I(\boldsymbol{x}_w \wedge \boldsymbol{y})$ is arbitrarily close to $C^{\mathrm{NoCov}}(D_1, D_2)$ by making the blocklength n large enough.

Converse

The attacker finds the minimizing $P_{Y|X}$ in (2.19) for the empirical distribution of the transmitted sequence $\boldsymbol{x}$. He then implements a memoryless channel based on this $P_{Y|X}$. The distortion constraint will be met with high probability as long as any $\tilde{D}_2 < D_2$ is used instead of D_2 in (2.19). A Fano-type inequality can be used to show that no rates higher than $C^{\mathrm{NoCov}}(D_1, \tilde{D}_2)$ are achievable for this attacker. The converse follows by continuity of (2.19) in D_2.

2.2.6 Binary Watermarking Game

In this section, we consider the watermarking game binary alphabets, i.e., $\mathcal{U} = \mathcal{X} = \mathcal{Y} = \{0, 1\}$. Further, we assume that the covertext $\boldsymbol{U}$ is an IID sequence of Bernoulli(1/2) random variables, i.e., $\Pr(U_i = 0) = \Pr(U_i = 1) = 1/2$. We use Hamming distortion constraints for both encoder and decoder, i.e., $d_1(\boldsymbol{u}, \boldsymbol{x}) = n^{-1} w_h(\boldsymbol{u} \oplus \boldsymbol{x})$ and $d_2(\boldsymbol{x}, \boldsymbol{y}) = n^{-1} w_h(\boldsymbol{x} \oplus \boldsymbol{y})$. We write the capacity in this scenario as $C_{\mathrm{priv}}^{\mathrm{BinWM}}(D_1, D_2)$ and $C_{\mathrm{pub}}^{\mathrm{BinWM}}(D_1, D_2)$ for the private and public versions, respectively.

Theorem 2.6. *For the binary watermarking game with* $0 \le D_1 \le 1/2$ *and* $0 \le D_2 \le 1/2$,

$$C_{\mathrm{priv}}^{\mathrm{BinWM}}(D_1, D_2) = H_b(D_1 \otimes D_2) - H_b(D_2), \tag{2.20}$$

and

$$C_{\mathrm{pub}}^{\mathrm{BinWM}}(D_1, D_2) = \max_{2D_1 \le g \le 1} g \left(H_b\left(\frac{D_1}{g}\right) - H_b(D_2) \right), \tag{2.21}$$

where $D_1 \otimes D_2 = D_1(1 - D_2) + (1 - D_1)D_2$ *and* $H_b(\cdot)$ *is the binary entropy, i.e.,* $H_b(p) = -p \log p - (1 - p) \log(1 - p)$.

See figure 2-2 for an example plot of $C_{\mathrm{priv}}^{\mathrm{BinWM}}(D_1, D_2)$ and $C_{\mathrm{pub}}^{\mathrm{BinWM}}(D_1, D_2)$. Note that $C_{\mathrm{priv}}^{\mathrm{BinWM}}(D_1, D_2) > C_{\mathrm{pub}}^{\mathrm{BinWM}}(D_1, D_2)$ for $0 < D_1 < 1/2$ and $0 < D_2 < 1/2$. Thus, unlike the Gaussian watermarking games, the capacity of the private version can exceed the capacity

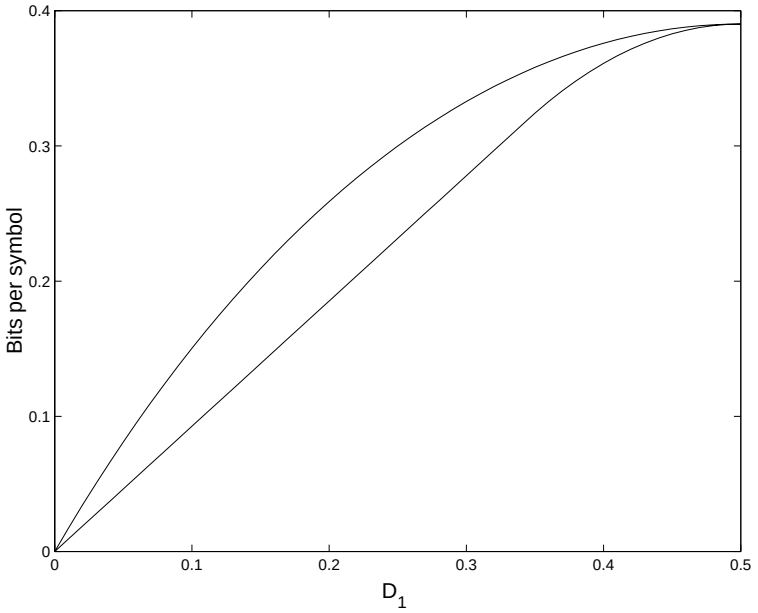

Figure 2-2: Binary watermarking capacity (private and public versions) versus D_1 with $D_2 = 0.15$.

of the public version. Also note that the maximizing g in (2.21) is given by

$$g^* = \begin{cases} \frac{D_1}{1-2^{-H_b(D_2)}} & \text{if } D_1 < 1 - 2^{-H_b(D_2)} \\ 1 & \text{otherwise} \end{cases}, \tag{2.22}$$

where $1 - 2^{-H_b(D_2)} \leq 1/2$ and thus $g^* \geq 2D_1$. Further, we can rewrite (2.21) as

$$C_{\text{pub}}^{\text{BinWM}}(D_1, D_2) = \begin{cases} D_1 \cdot \left(\frac{H_b(1-2^{-H_b(D_2)}) - H_b(D_2)}{1-2^{-H_b(D_2)}} \right) & \text{if } D_1 < 1 - 2^{-H_b(D_2)}, \\ H_b(D_1) - H_b(D_2) & \text{otherwise} \end{cases}. \tag{2.23}$$

Note that Barron, Chen and Wornell [BCW00] found identical expressions for the capacity when the attacker is *fixed* to be a binary symmetric channel with crossover probability D_2. Indeed, we prove the converse part of this theorem by fixing the attacker to be such a channel and computing the resulting capacity using an extension (Lemma 2.1) of Gel'fand and Pinsker's work [GP80] on channels with side information. The detailed proof of the converse and the achievability parts of the theorem can be found in Section 6.2. We give a brief sketch of the achievability proofs below.

Achievability for Private Version

The encoder and the decoder can use their combined knowledge of the covertext $\boldsymbol{U}$ to provide secrecy about a transmitted sequence chosen independently of $\boldsymbol{U}$. To see this, let a codeword $\tilde{\boldsymbol{X}} = f_n(W, \Theta_1)$ be chosen independently of $\boldsymbol{U}$ (but depending on the watermark W and the secret key Θ_1). The encoder will form the stegotext as $\boldsymbol{X} = \boldsymbol{U} \oplus \tilde{\boldsymbol{X}}$, and thus the distortion constraint on the encoder becomes $n^{-1} w_h(\tilde{\boldsymbol{X}}) \leq D_1$ a.s.. Furthermore, $\boldsymbol{U}$ is an IID sequence of Bernoulli(1/2) random variables, and thus $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are independent. Thus, any rate achievable for the AVC with constrained inputs and states is achievable here; see Section 2.5.3 and [CN88a]. In particular, all rates less than $C_{\text{priv}}^{\text{BinWM}}(D_1, D_2)$ are achievable.

Achievability for Public Version

Let us first fix a parameter g as in (2.21). The encoder/decoder pair select ng indices uniformly out of all subsets of $\{1, \ldots, n\}$ of size ng. The encoder will use only these indices of the covertext to encode the watermark. We use a codebook similar to that used for the public version of the SGWM game. In particular, every watermark $w \in \{1, \ldots, 2^{nR}\}$ corresponds to a bin of 2^{nR_0} codewords. Each codeword is a length-ng IID sequence of Bernoulli(1/2) random variables. Given the watermark w and the covertext $\boldsymbol{u}$, the encoder finds the codeword in bin w that agrees with the covertext at the selected indices as closely as possible. The encoder then creates the stegotext by replacing the selected positions of the covertext with the closest codeword. The distortion constraint will be satisfied if

$$R_0 > g \cdot \left(1 - H_b\left(\frac{D_1}{g}\right)\right). \tag{2.24}$$

The decoder finds the codeword closest to the forgery at the selected indices, and estimates the watermark as the bin of this codeword. Let $\tilde{\boldsymbol{y}} = \boldsymbol{y} \oplus \boldsymbol{x}$ be the difference between the forgery and the stegotext. The probability of error only depends on the attacker through the Hamming weight of $\tilde{\boldsymbol{y}}$, which can be at most nD_2. With high probability, the Hamming weight of $\tilde{\boldsymbol{y}}$ at the selected positions will not greatly exceed ngD_2. This observation allows us to show that the probability of error tends to zero as long as

$$R + R_0 < g \cdot (1 - H_b(D_2)). \tag{2.25}$$

The combination of (2.24) and (2.25) completes the proof.

2.3 Prior Work on Watermarking

In this section, we discuss some of the related literature and compare it to the results presented above. We first briefly describe some techniques that have been proposed, and we then give an overview of the information-theoretic work that has been done.

2.3.1 Practical Approaches to Watermarking

The simplest watermarking systems convey information by modifying the least significant parts of the covertext data, e.g., changing the low-order bits in a digital representation of an image. These systems are transparent, but they are easily corrupted. For example, lossy compression will remove the least significant portions of the data or an attacker might replace low-order bits with random bits without greatly affecting the quality of the data. It was recognized [CKLS97] that in order to achieve robustness, information must be embedded in significant portions of the data. Thus, for a given desired watermarking rate, there is a non-trivial trade-off between robustness and transparency.

The most widely studied class of watermarking systems consist of "spread spectrum" techniques, introduced in [CKLS97]. In these systems, a noise-like sequence is added to the covertext at the encoder and a correlation detector is used at the decoder. The watermark (i.e., the added sequence) is often scaled to achieve the desired robustness or transparency requirement, but otherwise the watermark is independent of the covertext. The watermark can be added either directly or in transform domains like Discrete Cosine Transform (DCT) [CKLS97], Fourier-Mellon transform [ORP97] or wavelets [XA98]. One important feature of such systems is that when the covertext is not available at the decoder (i.e., the public version), then the covertext acts as interference in the decoding process. Thus, as the variability in the original data increases, the amount of information that can be embedded in this manner decreases. However, we have seen that the capacity for our watermarking model can increase as the variability of the covertext increases, e.g., for the SGWM game. Thus, forming the stegotext by linearly combining the covertext and a signal that only depends on the watermark is suboptimal.

One new watermarking method that does not suffer from the problem of interference

from the covertext is Quantization Index Modulation (QIM), introduced by Chen and Wornell [Che00, CW01]. In QIM, a quantizer is used for the covertext depending on the value of the watermark. By varying the number and coarseness of the quantizers, one can trade off between transparency, robustness and data rate. Some of the watermarking techniques described in this thesis are similar to distortion compensated QIM, in which the stegotext is a linear combination of the covertext and the quantized version of the covertext (where again the quantizer depends on the value of the watermark). For example, in the public version of the SGWM game, the stegotext is a linear combination of the covertext and a codeword selected from the bin associated with the watermark; see the discussion after Theorem 2.1 and Section 4.3. The process of selecting the codeword is similar to quantization since the chosen codeword is the one closest to the covertext. In [Che00, CW01], it was shown that distortion compensated QIM achieves the capacity for situations with a known attacker. Here, we show that a similar technique also achieves the capacity for an unknown and arbitrary attacker.

2.3.2 Information-Theoretic Watermarking

The basic information theoretic model of watermarking was introduced by O'Sullivan, Moulin and Ettinger [MO99, MO00, OME98]. They investigated the capacity of a model that is similar to that described above but with several important differences. First, they assume a maximum likelihood decoder, which requires the decoder to be cognizant of the attack strategy. In contrast, we require that one encoder/decoder pair be robust against any potential attack. Second, they focus exclusively on average distortion constraints, while we compare the average and almost sure constraints. In fact, we find that average distortion constraints typically result in a capacity of zero. Finally, despite our stricter requirements, we have seen that our capacity with a Gaussian covertext is larger than that given in [MO99, MO00]; see Figure 2-1 for a comparison of the two capacities. Mittelhozer [Mit99] independently introduced a similar model for watermarking. Still others [BBDRP99, BI99, LC01, LM00, RA98, SPR98, SC96] have investigated the capacity of watermarking systems, but only for specific encoding schemes or types of attacks.

The most similar model to ours has been recently proposed by Somekh-Baruch and Merhav [SBM01a, SBM01b]. In their model, the probability that the distortion introduced by the encoder or the attacker is greater than some threshold must decay to zero exponen-

tially, i.e., $\Pr\{d_2(\boldsymbol{X},\boldsymbol{Y}) > D_2|\boldsymbol{X}=\boldsymbol{x}\} \leq e^{-\lambda n}$ for some λ and for all $\boldsymbol{x} \in \mathcal{X}^n$, and similarly for the encoder. This type of constraint is equivalent to our a.s. constraints when $\lambda = \infty$. In [SBM01a], they find a general expression (that does not depend on λ) for the coding capacity of the private version for finite alphabets. This result supersedes our result of Theorem 2.5 on the capacity of the watermarking game with no covertext and finite alphabets. Their capacity expression is similar to the mutual information game of [MO99, OME98]. We also see that for a scalar Gaussian covertext, the capacity is the same as the value of a related mutual information game; compare Theorems 2.1 and 3.1.

Besides capacity, several other information theoretic quantities have begun to be addressed for watermarking. Merhav [Mer00] and Somekh-Baruch and Merhav [SBM01a, SBM01b] have studied error exponents (i.e., how the probability of error decreases to zero as the blocklength increases for rates less than capacity) for a similar watermarking model, but with slightly different distortion constraints; see above. Also, Steinberg and Merhav [SM01] have investigated the identification capacity of a watermarking system with a fixed attack channel. In identification, questions of the form "Was watermark w sent?" need to be answered reliably instead of the usual "Which watermark was sent?". This more lenient requirement results in a doubly exponential growth in the number of watermarks; see also [AD89]. Furthermore, questions of this form might be what needs to be answered in some copyright protection applications. Finally, Karakos and Papamarcou [KP00a] have studied the trade-off between quantization and watermarking rate for data that needs to be both watermarked and compressed.

2.3.3 Similar Models: Steganography and Fingerprinting

In this section, we consider some models that are similar to watermarking and that have also generated recent interest. In steganography, the objective is to embed information so that an adversary cannot decide whether or not information has been embedded. This differs from our watermarking model since we assume that the attacker knows that information has been embedded, but has only limited means to remove it. For more on steganography see e.g., [AP98, Cac98, KP00b]. In fingerprinting, the embedded information is used to identify one of many users as opposed to a single owner. That is, the same covertext is given to different users with different watermarks. Thus, *collusive* attacks are possible on a fingerprinting system, while they are not possible on a watermarking system. In a

collusive attack, many users contribute their distinct fingerprinted copies in order to create a better forgery. Several researchers [BBK01, BS98, CFNP00, CEZ00, SEG00] have studied the number of fingerprinted objects that a system can distribute under various conditions. The research on fingerprinting has focused largely on combinatorial lower bounds on the number of possible fingerprints, while there has been less work on information-theoretic upper bounds.

2.3.4 Communication Games

We have seen that watermarking can be viewed as a communication game. At a low level, the encoder and decoder are playing a game against the attacker in which they are trying to communicate over a channel where the encoder's input sequence can be changed arbitrarily (subject to a distortion constraint), while the attacker is trying to prevent such reliable communication. This is similar to the arbitrarily varying channel (AVC), in which the encoder and decoder have to be designed to reliably send a message over a channel with many possible states, in which the channel state can change arbitrarily (as opposed to stochastically). At a higher level, the encoder and decoder are trying to to maximize the set of achievable rates while the attacker tries to minimize this set. This is similar to many mutual information games, in which a communicator and a jammer try to maximize and minimize, respectively, a mutual information expression. The solution to a mutual information game can sometimes be used to describe the maximum achievable rate for a communication system. The AVC and a mutual information game are discussed in more detail in Section 2.5.3 and Chapter 3, respectively.

We now consider a sample of other communication games that have been investigated. In one game [Baş83, BW85], a power-constrained transmitter tries to send a sequence of Gaussian random variables to a receiver with minimum mean-squared error, while a jammer (with some knowledge of the transmitter's input) attempts to maximize the error. In another game [MSP00], a transmitter can choose which slots in a slotted communication channel to transmit and the jammer can choose which slots to jam. Both transmitter and jammer are constrained by a dissipative energy model so that if power P_n (which can be either zero or some fixed value) is used in slot n, then $\sum_{n=0}^{m-1} \delta^n P_{m-n} \leq P_{\max}$ for all m where δ and $P_{\max}$ are given constants. In a final game [GH99, SV00], a transmitter tries to use the timing of packets to send information over a network (as in "Bits through Queues" [AV96]),

while a jamming network provider attempts to minimize the information rate subject to a constraint that he must deliver the packets in a timely fashion.

2.4 Assumptions in Watermarking Model

In this section, we review some of the assumptions made in the watermarking model. In Section 2.4.1, we briefly discuss if capacity is a good measure for a watermarking system. We then discuss randomization, and in particular when it is not necessary, for the encoder/decoder (Section 2.4.2) and for the attacker (Section 2.4.3). In Section 2.4.4, we discuss the distortion constraints that we impose in the watermarking model. In Section 2.4.5, we discuss the covertext distributions that we have chosen to study.

2.4.1 Is Capacity Meaningful?

In Section 2.2, we described the watermarking capacity for many scenarios, but we have not addressed whether the capacity of a watermarking system is a meaningful concept; we now discuss this issue. In order for the asymptotic analysis in the definition of capacity to be meaningful, there should be effectively limitless covertext data and unending watermark information to embed. This might not always be the case for a copyright protection application, since there would usually be a fixed length covertext and one of a fixed number of messages to embed. However, in many instances the data to be watermarked is quite long (e.g., a movie or an album), and the asymptotic regime can be safely assumed. Furthermore, there are other applications, such as hybrid digital/analog transmission and closed captioning, in which the above assumptions are met more generally. In any case, we think that the capacity achieving scheme should shed light on how to design a good watermarking system even for a non-asymptotic situation.

2.4.2 Randomization for Encoder/Decoder

There is a difference between the randomized coding used here and Shannon's classical random coding argument (see, for example, [CT91, 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 water-

marking 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 [LN98] for more on this subject.

Nevertheless, in the private version of the watermarking game, common randomness is typically not needed between the encoder and the decoder and deterministic codes suffice. For example, consider an IID Gaussian covertext. Part of the covertext to which both the encoder and the decoder have access, can be used instead of the secret key Θ_1. Indeed, 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.

Even if the covertext does not have a density, a similar technique can be used to show that a secret key is not necessary in the private version, as long as the number of samples from the covertext used for randomization does not asymptotically affect the distortion. Indeed, Ahlswede [Ahl78] has shown that only[4] $O(n^2)$ codebooks are necessary to achieve randomization in many situations. Thus, only $O(\log n)$ random bits available to both the encoder and decoder are needed to specify which codebook to use. Thus, if the covertext is a discrete memoryless source, then $O(\log n)$ samples from the covertext (which is known to both the encoder and decoder in the private version) can be used to specify the codebook. In order to prevent the attacker from learning anything about the codebook, the encoder should make the stegotext samples independent of the covertext samples that are used to specify the codebook, which results in some extra distortion. However, if the distortion constraint is bounded, then the extra distortion that is needed to implement this scheme is $O\left(\frac{\log n}{n}\right)$, which can be made negligible by making the blocklength n large enough.

2.4.3 Randomization for Attacker - Deterministic is Sufficient

We allow the attacker to implement a randomized strategy. However, to prove achievability in the watermarking game, we can without loss of generality limit the attacker to determin-

[4]For any two functions $f(n)$ and $g(n)$, $f(n) = O(g(n))$ if $f(n)/g(n)$ is bounded for all n.

istic 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 attacker mappings

$$\boldsymbol{y} = g_n(\boldsymbol{x}) \tag{2.26}$$

instead of the more general $g_n(\boldsymbol{x}, \theta_2)$. With an attacker of this form, the distortion constraint (2.3) can be rewritten as $d_2(\boldsymbol{X}, g_n(\boldsymbol{X})) \leq D_2$, 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 (2.26) is small, then the average probability of error for any general attacker mapping of the form (2.2) is also small. This idea is similar to the argument (which we outlined about in Section 2.4.2) that deterministic codebooks are sufficient for a fixed channel.

2.4.4 Distortion Constraints

Admittedly, the technique we have used to decide whether two data sequences are "similar" has some flaws. However, the simplicity of our technique allows us to derive closed form solutions that hopefully will give some intuition for more realistic scenarios. To review, we say that data sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ are similar if $n^{-1} \sum_i d(x_i, y_i) \leq D$ for some non-negative function $d(\cdot, \cdot)$ and some threshold D. The first potential problem is that $\boldsymbol{y}$ could be a shifted or rotated version of $\boldsymbol{x}$ and thus very "similar" to $\boldsymbol{x}$. However, our distortion measure would not recognize the similarity. This will affect our watermarking performance since we only require decoding from forgeries that are similar according to our distortion measure. One way to overcome this problem is to watermark in a domain (e.g., Fourier) that is relatively robust to such transformations [LWB+01, ORP97]. Another way to overcome this problem is for the encoder to use some its available distortion to introduce a synchronization signal that the decoder can use to align the samples of the covertext and the forgery [PP99]. The second potential problem is that there might not be a pointwise function $d(\cdot, \cdot)$ so that our distortion measure corresponds to perceptual distortion. Much work has been devoted to developing models of human perception to design good data compression schemes; see e.g., [JJS93, MS74] and references therein. It is clear that the squared difference distortion measure that we have mainly used does not directly correspond to human perceptual distortion, but our

distortion measure is tractable and provides a decent first approximation. We would like to integrate some more knowledge of the human perceptual system into our watermarking model in the future.

We require that the attacker satisfy a distortion constraint between the stegotext $\boldsymbol{X}$ and the forgery $\boldsymbol{Y}$. This is plausible because the attacker observes the stegotext and thus he knows exactly what forgeries are allowed. Since one basic purpose of this constraint is to ensure that the forgery is similar to the covertext $\boldsymbol{U}$, some have suggested that the attacker's constraint be between $\boldsymbol{U}$ and $\boldsymbol{Y}$ [Mou01]. However, with this alternative constraint, if the amount of distortion the attacker can add is small (but non-zero), then the watermarking system can send unlimited information, which seems unreasonable. On the other hand, for the SGWM game (with our original constraint) we saw that the capacity is zero only if $D_2 > \sigma_u^2 + D_1 + \sqrt{\sigma_u^2 D_1}$, while if $D_2 > \sigma_u^2$, then the attacker could set the forgery to zero, resulting in no positive achievable rates and a distortion between $\boldsymbol{U}$ and $\boldsymbol{Y}$ of approximately $\sigma_u^2 < D_2$. Thus, the capacity under our constraint is potentially too large for large attacker distortion levels, while the capacity under the alternative constraint is potentially too large for small attacker distortion levels.

2.4.5 Statistics of Covertext

In our study of watermarking, we have largely focused on Gaussian covertext distributions. Such a distribution might arise in transform domains where each sample is a weighted average of many samples from the original domain, in which case one would expect the central limit theorem to play a role. Indeed, some studies [BBPR98, JF95, Mül93] have found that the discrete cosine transform (DCT) coefficients for natural images are well modeled as generalized Gaussian random variables, which include the standard Gaussian distribution as a special case[5]. While a Gaussian model is reasonable for many types of sources that might need to be watermarked, there are other sources that require watermarking that cannot be so modeled; examples include VLSI designs [Oli99] and road maps [KZ00].

A shortcoming of the Gaussian assumption is that the data we are interested in will be stored on a computer, and hence the distribution could only be a quantized approximation

[5]The generalized Gaussian density is defined by $f_X(x) = \frac{\nu\alpha(\nu)}{2\sigma\Gamma(1/\nu)} \exp(-(\alpha(\nu)|x/\sigma|)^\nu)$, where $\alpha(\nu) = \sqrt{\Gamma(3/\nu)/\Gamma(1/\nu)}$, $\Gamma(\cdot)$ is the usual gamma function and ν is the so-called shape parameter. The generalized Gaussian is equivalent to the standard Gaussian when $\nu = 2$.

of a Gaussian distribution. If the quantization is too coarse, then the Gaussian assumption would not be reasonable. For example, one-bit quantization of every covertext sample would lead to the binary watermarking game, which we have seen to have much smaller capacity than the scalar Gaussian watermarking game. However, we are more likely to be interested in high fidelity storage of the data, and the Gaussian approximation is more reasonable in this case.

2.5 Uncertainty in the Watermarking Model

In watermarking, an encoder/decoder pair has to deal with two sources of uncertainty, the covertext and the attacker. In our model, the covertext is generated stochastically from some known distribution while the attacker can take on any form subject to a distortion constraint.

In Section 2.5.1, we formalize the differences between these two types of uncertainty into *stochastically* generated states and *arbitrarily* generated states. We then consider two models: one that contains only stochastically generated states (communication with side information, Section 2.5.2) and one that contains only arbitrarily generated states (the arbitrarily varying channel, Section 2.5.3). In Section 2.5.4, we consider an instance of communication with side information, Costa's "writing on dirty paper" model [Cos83], and describe two extensions to this model.

2.5.1 Types of State Generators

In order to discuss the types of state generators, we consider a communication channel that has a transition probability that depends on a state s. That is, given the value of the current state s and the current input x, the output of the channel is a random variable Y with distribution $P_{Y|X,S}(\cdot|x,s)$, where we assume throughout that $P_{Y|X,S}$ is known. Furthermore, given the state sequence $\boldsymbol{s}$ and the input sequence $\boldsymbol{x}$, the output sequence $\boldsymbol{Y}$ is generated in a memoryless fashion, so that

$$P(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{s}) = \prod_{i=1}^{n} P_{Y|X,S}(Y_i|x_i,s_i). \tag{2.27}$$

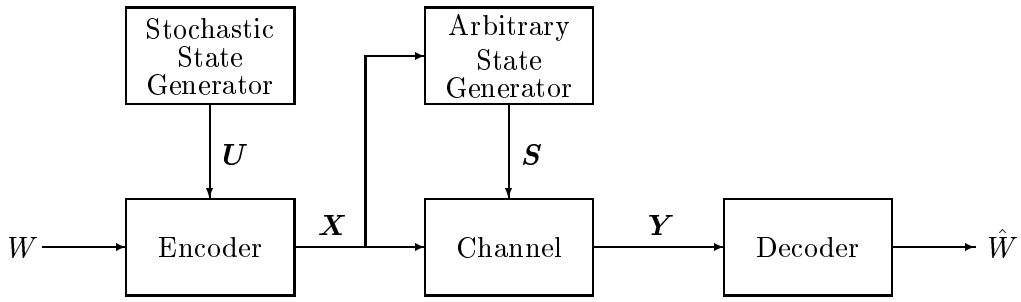

Figure 2-3: Watermarking model with state sequences.

We would like to describe the coding capacity for such a channel. That is, we would like to answer the usual question, "For rates R can we reliably communicate nR bits using the channel n times?". In general, reliable communication means that the probability of error can be as small as desired by making the blocklength n large enough. The definition of probability of error that we use affects the capacity and depends on how the state sequence is generated. Unless stated otherwise, we focus on probability of error averaged[6] (as opposed to maximized) over all possible bit sequences and sequence-wise probability of error (as opposed to bit-wise). We will also assume that the encoder and decoder share a source of randomness and that the probability of error is averaged over this source of randomness as well.

We now consider two possible methods for generating the state sequence:

1. The state sequence $\boldsymbol{S}$ could be generated *stochastically* from some known distribution $P_{\boldsymbol{S}}$ (usually independently of the other sources of randomness). In this case, we will be interested in the probability of error averaged over the possible values of the state sequence.

2. The state sequence $\boldsymbol{s}$ could be generated *arbitrarily*, possibly subject to some constraint. In this case, we will want to insure that the probability of error can be made small for every possible state sequence $\boldsymbol{s}$.

Restatement of Watermarking Model

We can think of our watermarking model as having two state sequences, one generated stochastically and one generated arbitrarily. This idea is depicted in Figure 2-3 (for the public version only). Here, the stochastically generated state $\boldsymbol{U}$ is the covertext and the arbitrarily generated state[7] $\boldsymbol{S}$ describes the mapping between the stegotext $\boldsymbol{X}$ and the forgery $\boldsymbol{Y}$. For example, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are real random vectors, then $\boldsymbol{S}$ could be the difference between $\boldsymbol{Y}$ and $\boldsymbol{X}$ and $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{S}}(\cdot|\boldsymbol{x},\boldsymbol{s})$ is the unit mass on $\boldsymbol{x}+\boldsymbol{s}$. This form is particularly useful when the attacker's distortion function can be written $d_2(x,y) = d_2(y-x)$. In this case, the attacker's distortion constraint becomes a constraint solely on the sequence $\boldsymbol{S}$. Note that the attacker knows the stegotext $\boldsymbol{X}$, and thus the arbitrary state sequence $\boldsymbol{S}$ is actually a mapping from $\mathcal{X}^n$ into $\mathcal{Y}^n$. Thus, the encoder/decoder pair wishes to make the average probability of error small for every possible attacker mapping, where the probability of error is averaged over all sources of randomness including the covertext. Both the encoder and the arbitrary state sequence $\boldsymbol{S}$ are subject to distortion constraints. Thus, although the stochastically generated state sequence $\boldsymbol{U}$ does not directly affect the channel, it does indirectly affect the channel through the constraint on the encoder's output.

2.5.2 Communication with Side Information

We now consider a model with only stochastically generated states, like the covertext in the watermarking game. When known at the encoder or decoder, the state sequence is called side information and thus this model is referred to as communication with side information. An example where the side information is known at the encoder only is depicted in Figure 2-4. All of the models in this section assume that the stochastic state sequence is generated in an IID manner according to a known distribution P_U.

Shannon [Sha58] first studied this problem under the assumption that the encoder must be causal with respect to the side information. That is, the ith channel input x_i can be a function of only the message and the channel states up to and including time i. Gel'fand and Pinsker [GP80] later found the capacity assuming (as we do in the watermarking game) that the encoder has non-causal access to the side information. That is, the channel input vector

[6] We make the usual assumption that all bit sequences are equally likely.

[7] This arbitrarily generated state is actually an arbitrary mapping $\boldsymbol{s}(\boldsymbol{X})$ that we write as the random vector $\boldsymbol{S}$.

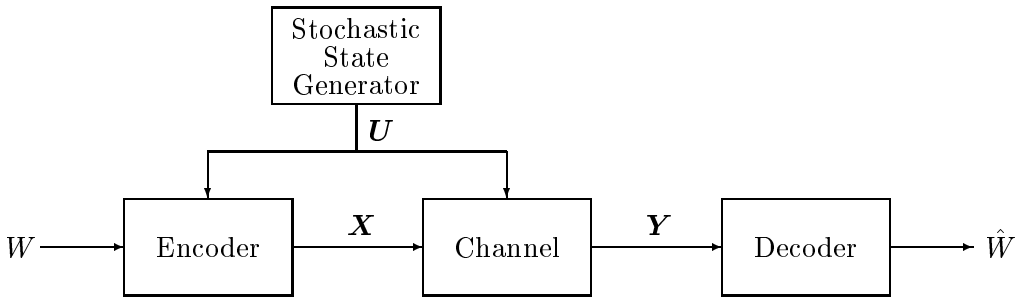

Figure 2-4: Communication with side information at the encoder.

$\boldsymbol{x} \in \mathcal{X}^n$ can be a function of the message and the channel state vector $\boldsymbol{u} \in \mathcal{U}^n$. A causal encoder makes practical sense in many real time applications, but a non-causal encoder also makes sense in other situations, such as watermarking or storing information on a partially defective hard drive. Heegard and El Gamal [HEG83] considered a generalization of [GP80] where the state sequence can be described non-causally to both the encoder and decoder, but only using rates R_e and R_d, respectively.

Capacity Results

We now give the capacity of the channel with side information in two scenarios: when the state sequence $\boldsymbol{U}$ is known non-causally to the encoder only, and when the state sequence $\boldsymbol{U}$ is known to non-causally to both the encoder and decoder. As in the watermarking game, we will refer to these scenarios as the public and private versions, respectively. Note that these results are proved assuming that the sets $\mathcal{X}, \mathcal{U}$ and $\mathcal{Y}$ are finite.

For the private version with non-causal side information, the capacity is given by [Wol78, HEG83]

$$C_{\text{priv}}^{\text{NCSI}} = \max_{P_{X|U}} I(X;Y|U), \tag{2.28}$$

where the mutual information is evaluated with respect to the joint distribution $P_{U,X,Y} = P_U P_{X|U} P_{Y|X,U}$. Recall that P_U and $P_{Y|X,U}$ are given.

For the public version with non-causal side information, the capacity is given by [GP80,

HEG83]

$$C_{\text{pub}}^{\text{NCSI}} = \max_{P_{V|U},\, f:\mathcal{V}\times\mathcal{U}\mapsto\mathcal{X}} I(V;Y) - I(V;U), \tag{2.29}$$

where V is an auxiliary random variable with alphabet $|\mathcal{V}| \leq |\mathcal{X}| + |\mathcal{U}| - 1$, and the mutual informations are evaluated with respect to the joint distribution

$$P_{U,V,X,Y}(u,v,x,y) = \begin{cases} P_U(u)P_{V|U}(v|u)P_{Y|X,U}(y|x,u) & \text{if } x = f(v,u) \\ 0 & \text{otherwise} \end{cases}. \tag{2.30}$$

The achievability of this capacity is proved using a random binning argument that we will also use to prove the watermarking capacity result. Note that the capacity with causal side information is similar [Sha58], except that $P_{V|U}$ is replaced by P_V in (2.29) and (2.30). The capacity with non-causal side information can be strictly greater than the capacity with causal side information. Thus, we would not expect the results on watermarking to directly carry over to a causal situation.

Fixed Attack Watermarking

One potential attack strategy in the watermarking game is a memoryless channel based on some conditional distribution $P_{Y|X}^{\text{attack}}$. Of course, the attacker should choose this distribution so that the distortion constraint is met either with high probability or in expectation. Assuming such an attack strategy is used and known to both the encoder and decoder, an extension of (2.28) or (2.29) can be used to describe the achievable rates for this scenario.

In the following lemma, we describe the capacity of the public version with non-causal side information when the encoder is required to meet a distortion constraint between the side information and the channel input, which can be used to describe the watermarking capacity with a fixed attack channel.

Lemma 2.1. *For the communication with side information model with finite alphabets, if the side information is available non-causally to the encoder only and the encoder is required to satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} d_1(u_i, x_i) \leq D_1, \ \ a.s., \tag{2.31}$$

for some non-negative function $d_1(\cdot,\cdot)$. Then, the capacity is given by

$$C_{\text{pub}}^{\text{NCSI}}(D_1) = \max_{\substack{P_{V|U},\, f:\mathcal{V}\times\mathcal{U}\mapsto\mathcal{X},\\ E[d_1(U,X)]\leq D_1}} I(V;Y) - I(V;U), \tag{2.32}$$

where V is an auxiliary random variable with finite alphabet, and the mutual informations are evaluated with respect to the joint distribution (2.30).

The proof of this lemma can be found in Appendix B.2. The achievability part and most of the converse part of the proof follow directly from the proof of Gel'fand and Pinsker [GP80]. One tricky part involves showing that the conditional distribution $P_{X|V,U}$ is deterministic (i.e., only takes values of 0 and 1). We will use this lemma to simplify the evaluation of the public version of the binary watermarking game; see Section 6.2.2.

An attacker in the watermarking game cannot implement a general channel based on both the input and the state since the attacker does not directly know the state sequence $\boldsymbol{U}$ (i.e., the covertext). However, this result can be used to analyze fixed attack watermarking by substituting $P_{Y|X,U}(y|x,u) = P_{Y|X}^{\text{attack}}(y|x)$ for all $u \in \mathcal{U}$.

This analysis inspires the mutual information games that we will describe in Chapter 3. In short, the mutual information game will further modify (2.32) and the analogous result for the private version by adding a minimization over feasible attack "channels" $P_{Y|X}^{\text{attack}}$, where feasible means that the distortion constraint is met in expectation. It is not clear that the solution to the mutual information game describes the capacity of the watermarking game. This is partly because a decoder for communication with side information uses knowledge about the channel's conditional distribution, while in the watermarking game, the attacker can choose any feasible attack channel *after* the decoder has been deployed.

2.5.3 Arbitrarily Varying Channels

We now turn our attention to states that can be generated *arbitrarily*. That is, there is no probability distribution on the state sequences, and any performance guarantees have to be valid for any possible state sequence. In the watermarking game, the attacker (under the a.s. distortion constraint) can produce an arbitrary sequence (subject to the distortion constraint) in its attempt to confuse the encoder and decoder.

The basic arbitrarily varying channel (AVC) was introduced in [BBT60] and has a single arbitrarily generated state sequence $\boldsymbol{s}$ that determines the conditional distribution

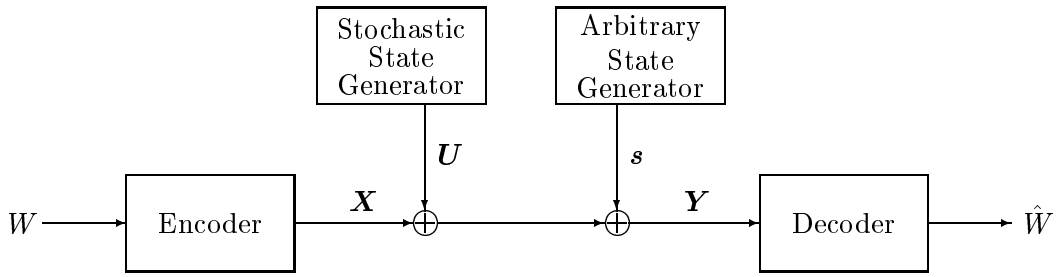

Figure 2-5: Gaussian arbitrarily varying channel: $\boldsymbol{U}$ is an IID Gaussian sequence and $\boldsymbol{s}$ is an arbitrary power constrained sequence.

of the channel as in (2.27). Unlike the usual communication scenario (e.g., a memoryless channel), the capacity depends on whether average or maximum probability of error is used and on whether there is a common source of randomness available to the encoder and decoder. Many variations of the AVC have been studied, see e.g., [CK81, LN98] for extensive references. Unlike the watermarking game, the state sequence and the input to the channel are usually assumed to be chosen independently. However, see [Ahl86] for analysis of the AVC when the state sequence is known to the encoder and [AW69] for analysis of the AVC when the input sequence is known to the state selector. Csiszár and Narayan [CN88a, CN88b] considered an instance of the AVC that has particular relevance to the watermarking game in which the input sequence $\boldsymbol{x}$ and the state sequence $\boldsymbol{s}$ must satisfy respective constraints. The capacity results depend on whether the constraints are enforced almost surely or in expectation, as is also the case for the watermarking game (compare Theorems 2.1 and 2.3).

The Gaussian Arbitrarily Varying Channel

The Gaussian arbitrarily varying channel (GAVC), introduced by Hughes and Narayan [HN87], is a particular AVC with constrained inputs and states that is related to the Gaussian watermarking game. In the GAVC (illustrated in Figure 2-5), the input and state sequences must both satisfy power constraints, and the channel is given by $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{s} + \boldsymbol{Z}$, where $\boldsymbol{Z}$ is an IID sequence of $\mathcal{N}(0, \sigma^2)$ random variables, $\boldsymbol{s}$ is an arbitrary sequence (subject to $n^{-1}\|\boldsymbol{s}\|^2 \leq D_2$), and the input $\boldsymbol{X}$ is similarly power limited to D_1.

Hughes and Narayan [HN87] found that the capacity of the GAVC (when a source of

common randomness is available to the encoder and decoder) is given by

$$C^{\mathrm{GAVC}}(D_1, D_2, \sigma^2) = \frac{1}{2}\log\left(1 + \frac{D_1}{D_2 + \sigma^2}\right). \tag{2.33}$$

Note that this is the same capacity that would result if $\boldsymbol{s}$ were replaced by an IID sequence of $\mathcal{N}(0, D_2)$ random variables. Further note that if the a.s. power constraints are replaced by expected power constraints then the capacity of the GAVC is zero, although the ϵ-capacity[8] is positive and increasing with ϵ.

Let us now consider an alternate description of the GAVC in order to highlight the similarities with the watermarking game with an IID Gaussian covertext. The GAVC can be obtained from the watermarking game by slightly modifying the capabilities of both the encoder and the attacker, as we now outline. First, the encoder must be of the form $\boldsymbol{X} = \boldsymbol{U} + \tilde{\boldsymbol{X}}$, where $\tilde{\boldsymbol{X}}$ is *independent* of $\boldsymbol{U}$ (but not independent of the watermark W). Second, the attacker must form the attack sequence $\boldsymbol{s}$ *independently* of $\boldsymbol{X}$. Thus, the overall channel is given by $\boldsymbol{Y} = \tilde{\boldsymbol{X}} + \boldsymbol{s} + \boldsymbol{U}$, where $\tilde{\boldsymbol{X}}$ is a power limited sequence depending on the message, $\boldsymbol{s}$ is a power limited arbitrary sequence, and $\boldsymbol{U}$ is an IID sequence of Gaussian random variables independent of $\tilde{\boldsymbol{X}}$ and $\boldsymbol{s}$. Although both the encoder and attacker are less powerful in the GAVC than in the watermarking game, the effect does not cancel out. Indeed, the capacity of the GAVC decreases with the variance of $\boldsymbol{U}$ while the watermarking capacity increases; compare (2.6) and (2.33).

Finally, note that the additive attack watermarking game of Section 2.2.2 with an IID Gaussian covertext is a combination of the GAVC and the scalar Gaussian watermarking game. In particular, this game uses the encoder from the watermarking game and the attacker from the GAVC. In this compromise between the two models, the capacity does not depend on the variance of $\boldsymbol{U}$; see Theorem 2.2.

2.5.4 Extended Writing on Dirty Paper

A special case of communication with side information (see Section 2.5.2) is Costa's writing on dirty paper [Cos83], which is depicted in Figure 2-6. In this model, all of the the sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ and $\mathcal{Z}$ are the real line. Further, the encoder knows the state sequence $\boldsymbol{U}$ non-causally and its output $\boldsymbol{X} = \boldsymbol{x}(W, \boldsymbol{U})$ must satisfy a power constraint, i.e., $n^{-1}\|\boldsymbol{X}\|^2 \leq D_1$

[8] The ϵ-capacity is the supremum of all rates such that the probability of error is at most ϵ.

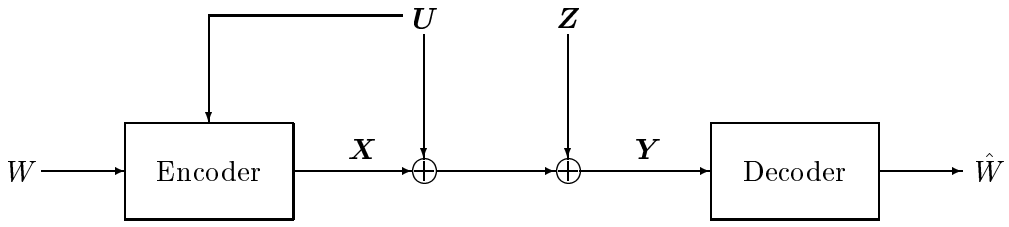

Figure 2-6: Writing on dirty paper. $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent IID Gaussian sequences.

a.s.. Finally, the output of the channel is given by

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{U} + \boldsymbol{Z} \tag{2.34}$$

where both $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent IID sequences of zero-mean Gaussian random variables of variances σ_u^2 and D_2, respectively. We will call $\boldsymbol{U}$ the covertext and $\boldsymbol{Z}$ the jamming sequence. Costa's main result is that the capacity is the same whether or not the covertext $\boldsymbol{U}$ is known to the decoder. When $\boldsymbol{U}$ is known to the decoder, the channel effectively become $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$, i.e. the classical power limited Gaussian channel. Thus, the capacity is given by $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$, which does not depend on σ_u^2. Others [Che00, YSJ$^+$01] have extended this result to when $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent non-white (i.e., colored) Gaussian processes. In this section, we describe two further extensions of Costa's result. First, when $\boldsymbol{U}$ has any (power limited) distribution and $\boldsymbol{Z}$ is an independent colored Gaussian process, we show that the capacity with non-causal side information (the random vector $\boldsymbol{U}$) at the encoder is the same as the capacity with side information at both the encoder and decoder. A similar result was given simultaneously by Erez, Shamai, and Zamir [ESZ00]. Second, we show that the additive attack watermarking game with Gaussian covertext (see Section 2.2.2) is an extension of Costa's result to arbitrarily varying noise.

Extension 1 : Any Distribution on Covertext, Colored Gaussian Jamming Sequence

We first generalize Costa's result to where the side information $\boldsymbol{U}$ is an IID sequence of random variables with some arbitrary (but known) distribution, while the noise sequence $\boldsymbol{Z}$ is still an IID sequence of mean-zero, variance-D_2 Gaussian random variables. We will

then be able to further generalize to the above assumptions.

Recall that the maximum over $I(V;Y) - I(U;V)$ (see (2.32)) is the capacity for a channel with non-causal side information at the encoder only. Although this result was only proved for finite alphabets, it is straightforward to extend the achievability part to infinite alphabets, which is all that we will need. Indeed, we will specify the joint distribution of an auxiliary random variable V, the input X and the side information U such that $I(V;Y) - I(U;V)$ equals the capacity when $\boldsymbol{U}$ is not present at all, which also acts as an upper bound on the capacity for writing on dirty paper.

We now specify the necessary joint distribution. Let X be a zero-mean Gaussian random variable of variance D_1, which is independent of U, which clearly satisfies $E[X^2] \leq D_1$. Also, let the auxiliary random variable $V = \alpha U + X$, where $\alpha = \frac{D_1}{D_1+D_2}$. (As in (2.32), we could have first generated V conditioned on U and then generated X as a function of V and U.) The preceding steps replicates Costa's original proof. At this point, he calculated $I(V;Y) - I(U;V)$ to be $\frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right)$, assuming that both U and Z are Gaussian random variables. This is sufficient to prove the original result since all rates less than this are achievable and the capacity cannot exceed the capacity without $\boldsymbol{U}$, which is also given by this expression. We shall assume that only Z is Gaussian, but we shall obtain the same result.

With our choice of the auxiliary random variable V,

$$V - \alpha(X + U + Z) = X - \alpha(X + Z), \tag{2.35}$$

and with our choice of α the random variables $X - \alpha(X + Z)$ and $X + Z$ are uncorrelated and hence, being zero-mean jointly Gaussian, also independent[9]. Furthermore, the random variables $X - \alpha(X + Z)$ and $X + U + Z$ are independent since U is independent of (X, Z).

[9] Another way to view the choice of α is that $\alpha(X + Z)$ is the minimum mean squared error (MMSE) estimate of X given $X + Z$.

Consequently,

$$\begin{aligned} h(V|X+U+Z) &= h\big(V-\alpha(X+U+Z)|X+U+Z\big) \\ &= h\big(X-\alpha(X+Z)\big) \\ &= h\big(X-\alpha(X+Z)|X+Z\big) \\ &= h(X|X+Z), \end{aligned} \tag{2.36}$$

where all of the differential entropies exist since X and Z are independent Gaussian random variables, and the second and third equalities follow by (2.35) and the above discussed independence. Also, the independence of U and X implies that

$$\begin{aligned} h(V|U) &= h(\alpha U + X|U) \\ &= h(X|U) \\ &= h(X). \end{aligned} \tag{2.37}$$

We can now compute that

$$\begin{aligned} I(V;X+U+Z) - I(V;U) &= h(V) - h(V|X+U+Z) - h(V) + h(V|U) \\ &= I(X;X+Z) \\ &= \frac{1}{2}\log\left(1+\frac{D_1}{D_2}\right), \end{aligned}$$

where the first equality follows by the definition of mutual information; the second equality follows from (2.36) and (2.37); and the last equality because X and Z are independent Gaussian random variables of variance D_1 and D_2, respectively.

Let us now consider general independent random processes $\boldsymbol{U}$ and $\boldsymbol{Z}$ as the known and unknown, respectively, additive noise components in the writing on dirty paper model. Also, let the random process $\boldsymbol{X}^*$ have the capacity achieving distribution for a channel with additive noise $\boldsymbol{Z}$ (i.e., $P_{\boldsymbol{X}^*} = \arg\max_{P_{\boldsymbol{X}}} I(\boldsymbol{X};\boldsymbol{X}+\boldsymbol{Z})$, where the maximum is over distributions that satisfy any required constraints). The preceding arguments can be repeated as long as there exists a linear[10] function $\alpha(\cdot)$ such that $\boldsymbol{X}^* - \alpha(\boldsymbol{X}^*+\boldsymbol{Z})$ is independent of $\boldsymbol{Z}$. (We also need $\boldsymbol{U}$ to be power limited so that all of the differential entropies are finite.)

[10]The linearity of $\alpha(\cdot)$ is needed in (2.35).

That is, for this random process $\boldsymbol{X}^*$ and this linear function $\alpha(\cdot)$, if $\boldsymbol{V} = \alpha(\boldsymbol{U}) + \boldsymbol{X}^*$, then $I(\boldsymbol{V};\boldsymbol{Y}) - I(\boldsymbol{U};\boldsymbol{V}) = I(\boldsymbol{X}^*;\boldsymbol{X}^* + \boldsymbol{Z})$, which is the capacity without $\boldsymbol{U}$ by our choice of $\boldsymbol{X}^*$.

We can thus show that for that Costa's result can be extended to any (power-limited) distribution on $\boldsymbol{U}$ and a colored Gaussian distribution on $\boldsymbol{Z}$. This follows since the capacity achieving $\boldsymbol{X}^*$ associated with $\boldsymbol{Z}$ is also Gaussian (with variances given by the waterfilling algorithm) [CT91]. Furthermore, for any two independent Gaussian (and hence jointly Gaussian) processes, we can find a linear function $\alpha(\cdot)$ that satisfies the above independence property.

We can also use an interleaving argument to show that if Costa's result holds for any power-limited IID law on $\boldsymbol{U}$, then it should also hold for any power-limited ergodic law. Furthermore, 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 IID [HM88, Lap96] it should be clear that it suffices to prove (as we have done above) this result for the case where $\boldsymbol{Z}$ is IID.

Extension 2 : IID Gaussian Covertext, Arbitrary Jamming Sequence

For the additive attack watermarking game with IID Gaussian covertext, we have shown that the capacity is the same for both the private and public versions; see Section 2.2.2. This provides an extension of Costa's writing on dirty paper result to when the jamming is an arbitrarily varying power-limited sequence. Note that the stegotext $\boldsymbol{X}$ in the watermarking game corresponds to $\boldsymbol{U} + \boldsymbol{X}$ here.

When the covertext $\boldsymbol{U}$ is IID Gaussian, then the additive attack watermarking game is similar to Costa's writing on dirty paper. In particular, the former model differs from the latter only in two respects. First, the jamming sequence distribution is arbitrary (subject to (2.11)) instead of being an IID 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 IID Gaussian one.

Chapter 3

Mutual Information Games

In this chapter, we consider two mutual information games that are motivated by the capacity results of Wolfowitz [Wol78] and Gel'fand and Pinsker [GP80] on communication with side information discussed in Section 2.5.2. We define the *private mutual information game* based on the capacity of a communication channel with side information non-causally available to both the encoder and decoder; see (2.28). Similarly, we define the *public mutual information game* based on the capacity of a communication channel with side information non-causally available to only the encoder; see (2.29). Mutual information games have been considered in the context of watermarking previously by Moulin and O'Sullivan [OME98, MO99, MO00]. We focus on squared error distortion and IID Gaussian sources, and the resulting solution provides insight into how to approach the scalar Gaussian watermarking (SGWM) game.

The remainder of this chapter is organized as follows. In Section 3.1, we precisely define our mutual information games and give our main result on the value of the games. In Section 3.2, we sketch the proof of the main result using three main lemmas; the proofs of these lemmas can be found in Appendix B. In Section 3.3, we give a game theoretic interpretation of the mutual information games. In Section 3.4, we discuss some other mutual information games that have been previously considered

3.1 Definition and Main Result

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

mation

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

where $D(\cdot||\cdot)$ is the Kullback-Leibler distance, which is 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*. For every n, the encoder chooses a *watermarking channel* $P_{\boldsymbol{X}|\boldsymbol{U}}$ that satisfies the average distortion constraint (2.14), and the attacker then chooses an *attack channel* $P_{\boldsymbol{Y}|\boldsymbol{X}}$ that satisfies the average distortion constraint (2.15). The quantity that the encoder wishes to maximize and that the attacker wishes to minimize is

$$I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) = \frac{1}{n} I_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U}), \tag{3.1}$$

which is the mutual information term in (2.28). The *value of the private mutual information game* is thus

$$C_{\text{priv}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) = \liminf_{n\to\infty} \sup_{P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}})} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}), \tag{3.2}$$

where

$$\mathcal{D}_1(D_1, P_{\boldsymbol{U}}) = \left\{ P_{\boldsymbol{X}|\boldsymbol{U}} : E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}}}[d_1(\boldsymbol{U}, \boldsymbol{X})] \leq D_1 \right\}, \tag{3.3}$$

and

$$\mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}) = \left\{ P_{\boldsymbol{Y}|\boldsymbol{X}} : E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{Y}|\boldsymbol{X}}}[d_2(\boldsymbol{X}, \boldsymbol{Y})] \leq D_2 \right\}. \tag{3.4}$$

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.

We next describe the *public mutual information game*. 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 mutual information term from (2.29) is $n^{-1}(I(\boldsymbol{V};\boldsymbol{Y}) - I(\boldsymbol{V};\boldsymbol{U}))$, which is written more explicitly as

$$I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) = \frac{1}{n}\left(I_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}} P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{V};\boldsymbol{Y}) - I_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}} P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}}(\boldsymbol{V};\boldsymbol{U}) \right), \tag{3.5}$$

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

$$C_{\text{pub}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) = \liminf_{n\to\infty} \sup_{\substack{P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}}) \\ P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}}} \inf_{P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}})} I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}}). \tag{3.6}$$

Note that the supremum is over a slightly more general set than (2.32), since we have not shown (as we did for finite alphabets in Lemma 2.1) that the maximizing joint distribution on the random vectors $\boldsymbol{U}$, $\boldsymbol{X}$ and $\boldsymbol{V}$ makes $\boldsymbol{X}$ a deterministic function of $\boldsymbol{U}$ and $\boldsymbol{V}$.

In the following theorem, which is proved in Section 3.2, we show that the capacity of

the SGWM game $C^*(D_1, D_2, \sigma_u^2)$ is an upper bound on the values of the mutual information games for real alphabets and squared error distortions. Moreover, for IID Gaussian covertexts, this upper bound is tight.

Theorem 3.1. *For real alphabets and squared error distortions*

$$\begin{aligned} C_{\text{pub}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) &\leq C_{\text{priv}}^{\text{MI}}(D_1, D_2, \{P_{\boldsymbol{U}}\}) && (3.7) \\ &\leq C^*(D_1, D_2, \underline{\sigma}_u^2), && (3.8) \end{aligned}$$

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

$$\underline{\sigma}_u^2 = \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E_{P_{\boldsymbol{U}}}[U_i^2] \tag{3.9}$$

and is assumed finite.

Equality is achieved in both (3.7) and (3.8) if the covertext is zero-mean IID Gaussian.

Recall the definition of $C^*(D_1, D_2, \sigma_u^2)$ given in (2.6). This definition and some other relevant definitions used in this chapter are summarized in Appendix A.

3.2 Proof of Mutual Information Game Result

In this section, we sketch a proof of Theorem 3.1. The upper bound on the values of the games is based on a family of attack channels that will be described in Section 3.2.1. The equality for IID zero-mean Gaussian covertexts is based on the watermarking channels that will be described in Section 3.2.2. In Section 3.2.3, we will show that the proposed attack channels prove the upper bound (3.8) and that for IID zero-mean Gaussian covertexts, the proposed watermarking channels guarantee the claimed equality.

3.2.1 Optimal Attack Channel

The attack channel we propose does not depend on the version of the game, and is described next. 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

$$A_n = \frac{1}{n} E_{P_{\boldsymbol{U}} P_{\boldsymbol{X}|\boldsymbol{U}}}[\|\boldsymbol{X}\|^2]. \tag{3.10}$$

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[1]

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

where $c(A;D_2) = 1 - \frac{D_2}{A}$ (also defined in (A.4)). 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 [CT91] 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

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

that is, $P_{\boldsymbol{Y}|\boldsymbol{X}}$ has a product form with marginal $P^{A_n}_{Y|X}$, where A_n is given in (3.10).

Notice that by (3.10) and the structure of the attack channel

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

Thus the attack channel $(P^{A_n}_{Y|X})^n$ satisfies the distortion constraint. Compare this attack channel with the attacker (defined in Section 4.5.1) used in the proof of the converse of the SGWM game.

3.2.2 Optimal Watermarking Channel

In this section we focus on IID zero-mean variance-σ_u^2 Gaussian covertexts and describe watermarking channels that will demonstrate that for such covertexts (3.7) and (3.8) both

[1] 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.

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 $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, where $\mathcal{A}(D_1, D_2, \sigma_u^2)$ is defined in (A.7). The choice of A 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 $P_{X|U}^A$ be Gaussian with mean $b_1(A; D_1, \sigma_u^2)U$ and variance $b_2(A; D_1, \sigma_u^2)$, i.e., have the density

$$f_{X|U}^A(x|u) = \mathcal{N}(x; b_1(A; D_1, \sigma_u^2)u, b_2(A; D_1, \sigma_u^2)),$$

where $b_1(A; D_1, \sigma_u^2) = \frac{A+\sigma_u^2-D_1}{2\sigma_u^2}$ and $b_2(A; D_1, \sigma_u^2) = D_1 - \frac{(A-\sigma_u^2-D_1)^2}{4\sigma_u^2}$ (also defined in (A.2) and (A.3)). Equivalently, under $P_{X|U}^A$ the random variable X is distributed as $b_1(A; D_1, \sigma_u^2)U + S_1$, where S_1 is a zero-mean variance-$b_2(A; D_1, \sigma_u^2)$ Gaussian random variable that is independent of U.

For IID zero-mean variance-σ_u^2 Gaussian covertexts we have

$$\begin{aligned} E_{P_{\boldsymbol{U}}(P_{X|U}^A)^n}\left[\frac{1}{n}\|\boldsymbol{X}-\boldsymbol{U}\|^2\right] &= \big(b_1(A; D_1, \sigma_u^2) - 1\big)^2\sigma_u^2 + b_2(A; D_1, \sigma_u^2) \\ &= D_1. \end{aligned}$$

Thus for this covertext distribution (and, in fact, for any covertext distribution with variance σ_u^2), the watermarking channel $(P_{X|U}^A)^n$ satisfies the distortion constraint. Furthermore,

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

which gives an interpretation of the parameter A as the power in the stegotext induced by the covertext and the watermarking channel. Compare this watermarking channel with the achievability scheme for the private SGWM game given in Section 4.2.1.

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

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

where $\alpha(A; D_1, D_2, \sigma_u^2) = 1 - \frac{b_1(A; D_1, \sigma_u^2)}{1+s(A; D_1, D_2, \sigma_u^2)}$ and $s(A; D_1, D_2, \sigma_u^2) = \frac{c(A; D_2) b_2(A; D_1, \sigma_u^2)}{D_2}$ (also defined in (A.6) and (A.5)). Finally, 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 SGWM game given in Section 4.3.1.

3.2.3 Analysis

In this section, we state three lemmas, which together prove Theorem 2.4. Lemma 3.1 (proved in Appendix B.3) 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 (proved in Appendix B.4) shows that, by using the attack channel proposed in Section 3.2.1, 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 (3.9). Lemma 3.3 (proved in Appendix B.5) shows that by watermarking an IID zero-mean variance-σ_u^2 Gaussian source using the channel proposed in Section 3.2.2 with the appropriate choice of A, the encoder can guarantee a value for the public mutual information game of at least $C^*(D_1, D_2, \sigma_u^2)$.

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 (3.7).

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* $D_2 > A_n$

$$\begin{aligned} I_{\text{priv}}\left(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, (P_{Y|X}^{A_n})^n\right) &\leq I_{\text{priv}}\left((P_U^G)^n, (P_{X|U}^{A_n})^n, (P_{Y|X}^{A_n})^n\right) \\ &= \frac{1}{2}\log\big(1 + s(A_n; D_{1,n}, D_2, \sigma_{u,n}^2)\big), \end{aligned} \tag{3.11}$$

where

$$\sigma_{u,n}^2 = E_{P_{\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{U}\|^2\right]; \tag{3.12}$$

$$D_{1,n} = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}[n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2]; \tag{3.13}$$

$$A_n = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{X}\|^2\right]; \tag{3.14}$$

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 3.2.2 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 3.2.1 for the parameters* D_2 *and* A_n.

This lemma proves (3.8). To see this note that for any $\epsilon > 0$ and any integer n_0 there exists some $n > n_0$ such that

$$\sigma_{u,n}^2 < \underline{\sigma}_u^2 + \epsilon, \tag{3.15}$$

where $\underline{\sigma}_u^2$ is defined in (3.9) and $\sigma_{u,n}^2$ is defined in (3.12). Also, since the watermarking channel must satisfy the distortion constraint (i.e. $P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1(D_1, P_{\boldsymbol{U}})$),

$$D_{1,n} \leq D_1, \tag{3.16}$$

where $D_{1,n}$ is defined in (3.13).

If A_n defined in (3.14) 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, (3.8) is satisfied in this case. We thus focus on the case when $A_n > D_2$. We also note that

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

by the triangle inequality so that $A_n \in \mathcal{A}(D_{1,n}, D_2, \sigma_{u,n}^2)$. By the definition of $C^*(\cdot,\cdot,\cdot)$ (A.8), it follows that the right hand side (RHS) of (3.11) 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 (3.15) and (3.16), because $C^*(D_1, D_2, \sigma_u^2)$ is non-decreasing in D_1 and σ_u^2 (see Appendix A). Finally, since $\epsilon > 0$ is arbitrary and $C^*(\cdot, \cdot, \cdot)$ is continuous in its arguments, 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 $C^*(D_1, D_2, \underline{\sigma}_u^2)$.

This lemma also shows that for an IID Gaussian covertext, if the memoryless attack channel $(P_{Y|X}^A)^n$ 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 $(P_{X|U}^A)^n$ of Section 3.2.2.

Lemma 3.3. *Consider an IID zero-mean variance-σ_u^2 Gaussian covertext (denoted $(P_U^G)^n$) and fixed distortions D_1 and D_2. If the attack channel $P_{\boldsymbol{Y}|\boldsymbol{X}}$ satisfies*

$$E_{(P_U^G P_{X|U}^A)^n P_{\boldsymbol{Y}|\boldsymbol{X}}}\left[n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\|\right] \leq D_2,$$

then for all $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$,

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

Here, $P_{X|U}^A$ and $P_{V|U,X}^A$ are the watermarking channels described in Section 3.2.2 for the parameters σ_u^2, D_1 and A and $P_{Y|X}^A$ is the attack channel described in Section 3.2.1 for the parameters D_2 and A.

This lemma implies that for a zero-mean variance-σ_u^2 IID 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 $(P_{X|U}^{A^*})^n$ and $(P_{V|U,X}^{A^*})^n$ where A^* achieves the maximum in the definition of C^*. Since for any covertext distribution (and in particular for an IID 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 IID Gaussian covertext, C^* is also a lower bound on the value of the private Gaussian mutual information game.

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

Lemma 3.3 also shows that when the covertext is zero-mean IID 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 $(P^A_{Y|X})^n$.

3.3 Game Theoretic Interpretation

In this section, we look at the the private mutual information game, defined in (3.2), with IID zero-mean variance-σ_u^2 Gaussian covertext from 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. [Gib92]), 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 will 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 pay-off to the first player. We will also illustrate a mistake that could be made when computing the value of the game.

We first rederive the value of the game using this game theoretic interpretation. 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 (3.3). However, the attacker plays second and thus his strategy space consists of all mappings of the form

$$\psi : P_{\boldsymbol{X}|\boldsymbol{U}} \mapsto P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2\left(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}\right), \quad \forall P_{\boldsymbol{X}|\boldsymbol{U}} \in \mathcal{D}_1\left(D_1, (P_U^G)^n\right), \tag{3.17}$$

where $\mathcal{D}_2\big(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}\big)$ is defined in (3.4). That is, for every possible strategy $P_{\boldsymbol{X}|\boldsymbol{U}}$ the encoder might use, the attacker must choose a feasible response $\psi(P_{\boldsymbol{X}|\boldsymbol{U}})$.

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

$$I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P_{\boldsymbol{X}|\boldsymbol{U}})\right) \leq I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right), \tag{3.18}$$

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

$$I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right) \leq I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right), \tag{3.19}$$

for every mapping $\psi(\cdot)$ of the form (3.17). 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 $\psi^*(\cdot)$.

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 one encoder strategy as in the regular Nash equilibrium. The value of the game is given by evaluating the mutual information at any subgame-perfect Nash equilibrium (there is not necessarily a unique equilibrium). The value of the game is thus $I_{\text{priv}}\left((P_U^G)^n, P^*_{\boldsymbol{X}|\boldsymbol{U}}, \psi^*(P^*_{\boldsymbol{X}|\boldsymbol{U}})\right)$.

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 3.2.2, A^* achieves the maximum in (A.8), and $\psi^*\left((P^A_{X|U})^n\right) = (P^A_{Y|X})^n$ for every $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, where $P^A_{Y|X}$ is defined in Section 3.2.1. The value of the game is thus $C^*(D_1, D_2, \sigma_u^2)$.

Using the above concepts, we now discuss the value of this game that was given in [MO99, MO00]. For $A_0 = \sigma_u^2 + D_1$,

$$I_{\text{priv}}\left((P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}}, (P^{A_0}_{Y|X})^n\right) \leq I_{\text{priv}}\left((P_U^G)^n, (P^{A_0}_{X|U})^n, (P^{A_0}_{Y|X})^n\right), \tag{3.20}$$

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

$$I_{\text{priv}}\left((P_U^G)^n, (P_{X|U}^{A_0})^n, (P_{Y|X}^{A_0})^n\right) \leq I_{\text{priv}}\left((P_U^G)^n, (P_{X|U}^{A_0})^n, P_{\boldsymbol{Y}|\boldsymbol{X}}\right), \tag{3.21}$$

for every $P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2\left(D_2, (P_U^G)^n, (P_{X|U}^{A_0})^n\right)$. Thus, it would seem that if $\psi_0(P_{\boldsymbol{X}|\boldsymbol{U}}) = (P_{Y|X}^{A_0})^n$ for all $P_{\boldsymbol{X}|\boldsymbol{U}}$, then the pair $\left((P_{X|U}^{A_0})^n, \psi_0(\cdot)\right)$ form a Nash equilibrium according to the definitions (3.18) and (3.19). The value of the game given in [MO99, MO00] is the mutual information evaluated with this pair. However, this attack strategy is not valid since $(P_{Y|X}^{A_0})^n \notin \mathcal{D}_2(D_2, (P_U^G)^n, P_{\boldsymbol{X}|\boldsymbol{U}})$ for some $P_{\boldsymbol{X}|\boldsymbol{U}}$, and in particular for any $P_{\boldsymbol{X}|\boldsymbol{U}}$ with $n^{-1}E\left[\|\boldsymbol{X}\|^2\right] > A_0$. Indeed, the optimal encoder strategy $(P_{X|U}^{A^*})^n$ has $n^{-1}E\left[\|\boldsymbol{X}\|^2\right] = A^*$ and $A^* > A_0$ (see Lemma A.1). Thus, the expression on the RHS of (3.20) is strictly less than $C^*(D_1, D_2, \sigma_u^2)$; see Figure 2-1 for a comparison of the two expressions.

3.4 Other Mutual Information Games

Zero-sum games in which one player tries to maximize some mutual information expression while the other player tries to minimize the same mutual information have also been investigated in [BMM85, SM88, Yan93]. As in the watermarking game, typically the first player is a communicator and the second player is a jammer. Assuming maximum-likelihood decoding, the mutual information between the input and output of a channel gives the rate at which reliable communication can take place. However, the decoder in the watermarking game is not necessarily performing maximum-likelihood decoding, and thus the mutual information games do not necessarily describe the capacity.

Most of the research in this area has focused on the channel $Y = X + Z$, where X is the input specified by the first player, Z is the noise specified by the second player, and X and Z are *independent*. For this game, the mutual information expression of interest is $I(X;Y)$. If X and Z are both power-constrained in expectation (i.e., $E[X^2] \leq P$ and $E[Z^2] \leq N$), then zero-mean Gaussian distributions for both X and Z form a saddlepoint in mutual information [Bla57]. That is, if $X^* \sim \mathcal{N}(0, P)$ and $Z^* \sim \mathcal{N}(0, N)$, then

$$I(X; X + Z^*) \leq I(X^*; X^* + Z^*) \leq I(X^*; X^* + Z), \tag{3.22}$$

for any feasible random variables X and Z. In our mutual information game with Gaussian

covertext and power-constraints, the optimal strategies are also (conditionally) Gaussian. However, the one-dimensional solution to our mutual information game does not form a saddlepoint. Another result that is reflected in our mutual information game is that even if a player is allowed to choose random vectors instead of random variables, then he will choose the random vector to consist of independent and identically distributed (IID) random variables [SM88]. Thus, it is sufficient to consider the one-dimensional mutual information game for the additive channel discussed above.

Chapter 4

The Scalar Gaussian Watermarking Game

This chapter is devoted to proving Theorem 2.1, which describes the capacity of the scalar Gaussian watermarking (SGWM) game and gives an upper bound on the capacity for a general ergodic covertext. In the SGWM game, the covertext is an IID sequence of zero-mean variance-σ_u^2 random variables and the distortion is measured using the squared difference. The proof of this theorem is divided into two main parts, achievability and converse.

The achievability part of the proof (Sections 4.2 and 4.3) consists of showing that all rates less than $C^*(D_1, D_2, \sigma_u^2)$ are achievable for the SGWM game for the private and public versions, respectively. In Section 4.3, we also show that all rates less than $\frac{1}{2}\log\left(1+\frac{D_1}{D_2}\right)$ are achievable for the public version of the additive attack watermarking game with Gaussian covertext, which completes the proof of Theorem 2.2. To assist in these arguments, we describe the allowable attacks in Section 4.1. We also show in Section 4.4 that it is sufficient to consider covertexts that are uniformly distribution on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$.

In the converse part in Section 4.5, we show that no rates higher than $C^*(D_1, D_2, \sigma_u^2)$ are achievable in the SGWM game. In fact, we show that no such rates are achievable for any ergodic covertext distribution with second moment at most σ_u^2.

In this chapter, we will use uniform distributions on the n-dimensional sphere as an approximation for an IID Gaussian distribution. We denote the n-dimensional sphere centered

at $\boldsymbol{\mu} \in \mathbb{R}^n$ with radius $r \geq 0$ by $\mathcal{S}^n(\boldsymbol{\mu}, r)$, i.e.,

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

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

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

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.

Note that many of the other definitions used in this chapter are summarized in Appendix A. Most importantly, recall that if $\mathcal{A}(D_1, D_2, \sigma_u^2)$ is non-empty, then

$$C^*(D_1, D_2, \sigma_u^2) = \max_{A \in \mathcal{A}(D_1, D_2, \sigma_u^2)} \frac{1}{2} \log\big(1 + s(A; D_1, D_2, \sigma_u^2)\big), \tag{4.1}$$

where $\mathcal{A}(D_1, D_2, \sigma_u^2)$ and $s(A; D_1, D_2, \sigma_u^2)$ are defined in Appendix A.

4.1 Deterministic Attacks

In Section 2.4.3, we argued that deterministic attacks are sufficient to analyze achievability for the watermarking game. In this section, we describe in more detail a deterministic additive attack (Section 4.1.1) and a deterministic general attack (Section 4.1.2).

4.1.1 Deterministic Additive Attack

For the additive attack watermarking game with real alphabets and squared error distortion, a deterministic attacker takes on a particularly simple form. Indeed, combining the forms (2.26) and (2.9), we see that the attacker can be written as

$$g_n(\boldsymbol{x}) = \boldsymbol{x} + \tilde{\boldsymbol{y}} \tag{4.2}$$

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

$$\frac{1}{n}\|\tilde{\boldsymbol{y}}\|^2 \leq D_2. \tag{4.3}$$

4.1.2 Deterministic General Attack

For the general watermarking game with real alphabets and squared error distortions, 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

$$g_n(\boldsymbol{x}) = \gamma_1(\boldsymbol{x})\boldsymbol{x} + \boldsymbol{\gamma}_2(\boldsymbol{x}), \tag{4.4}$$

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

$$\gamma_3(\boldsymbol{x}) = n^{-1}\|\boldsymbol{\gamma}_2(\boldsymbol{x})\|^2, \tag{4.5}$$

we can rewrite the attacker's distortion constraint (2.3) in terms of $\gamma_1(\boldsymbol{X})$, $\boldsymbol{X}$, and $\gamma_3(\boldsymbol{X})$ as

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

and consequently,

$$\frac{\gamma_3(\boldsymbol{x})}{\gamma_1^2(\boldsymbol{x})} \leq \frac{D_2}{c(n^{-1}\|\boldsymbol{x}\|^2; D_2)}, \tag{4.6}$$

for almost all $\boldsymbol{x}$ such that $n^{-1}\|\boldsymbol{x}\|^2 > D_2$, where $c(A; D_2) = 1 - D_2/A$ (also defined in (A.4)).

4.2 Achievability for Private Version

In this section, we show that for the private version of the watermarking game all rates up to $C^*(D_1, D_2, \sigma_u^2)$ are achievable when the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$. This result is extended to IID Gaussian covertexts in Section 4.4.

4.2.1 Coding Strategy

The coding strategy for the private version of the watermarking game is motivated by the solution to the corresponding mutual information game; see Theorem 3.1 and its proof in Section 3.2.

Constants

The encoder and decoder choose some $\delta > 0$ and a value of $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$, where the interval $\mathcal{A}(D_1, D_2, \sigma_u^2)$ is defined in (A.7). We assume throughout that the above interval is non-empty, because otherwise the claimed coding capacity is zero, and there is no need for a coding theorem.

Let the rate R of the coding strategy be

$$R = \frac{1}{2}\log\bigl(1 + s(A; D_1, D_2, \sigma_u^2)\bigr) - \delta, \tag{4.7}$$

where $s(A; D_1, D_2, \sigma_u^2)$ is defined in (A.5). Note that if the chosen A achieves the maximum in (4.1), then the RHS of (4.7) is $R = C^*(D_1, D_2, \sigma_u^2) - \delta$; we show in Lemma A.1 that such an A can be chosen. We will show that for any $\delta > 0$, and for $\boldsymbol{U}$ that is uniformly distributed over the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$, the rate R is achievable.

The encoder and decoder also compute the constants $\rho = \rho(A; D_1, \sigma_u^2)$, $b_1 = b_1(A; D_1, \sigma_u^2)$ and $b_2(A; D_1, \sigma_u^2)$, which are all defined in Appendix A. Recall that $\rho = (A - \sigma_u^2 - D_1)/2$, $b_1 = 1 + \rho/\sigma_u^2$ and $b_2 = D_1 - \rho^2/\sigma_u^2$.

Encoder and Decoder

The encoder and decoder use their common randomness Θ_1 to generate 2^{nR} independent random vectors $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_{2^{nR}}\}$, where each random vector $\boldsymbol{C}_i$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, 1)$.

Given a covertext $\boldsymbol{U} = \boldsymbol{u}$, a message $W = w$, and the vector $\boldsymbol{C}_w = \boldsymbol{c}_w$, let $\boldsymbol{c}_w(\boldsymbol{u})$ be the projection of $\boldsymbol{c}_w$ onto the subspace orthogonal to $\boldsymbol{u}$, but scaled so that

$$n^{-1}\|\boldsymbol{c}_w(\boldsymbol{u})\|^2 = b_2. \tag{4.8}$$

That is,

$$\boldsymbol{c}_w(\boldsymbol{u}) = \sqrt{nb_2}\frac{\boldsymbol{c}_w|_{\boldsymbol{u}^\perp}}{\left\|\boldsymbol{c}_w|_{\boldsymbol{u}^\perp}\right\|}. \tag{4.9}$$

Note that

$$\langle \boldsymbol{c}_w(\boldsymbol{u}), \boldsymbol{u}\rangle = 0. \tag{4.10}$$

Encoder: Using the covertext $\boldsymbol{u}$, the message w, and the source of common randomness θ_1 the encoder creates the stegotext $\boldsymbol{x}$ as

$$\boldsymbol{x} = f_n(\boldsymbol{u}, w, \theta_1) = b_1\boldsymbol{u} + \boldsymbol{c}_w(\boldsymbol{u}). \tag{4.11}$$

By (4.10) and the definitions of the constants b_1 and b_2 (A.2), (A.3), it follows that

$$n^{-1}\|\boldsymbol{x} - \boldsymbol{u}\|^2 = (b_1 - 1)^2\sigma_u^2 + b_2 = D_1,$$

thus demonstrating that the encoder satisfies the distortion constraint (2.1). We can further calculate that

$$n^{-1}\|\boldsymbol{x}\|^2 = A, \tag{4.12}$$

which demonstrates the operational significance of the constant A as the power of the stegotext.

Decoder: The decoder uses a modified nearest-neighbor decoding rule. 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 distance between $\boldsymbol{y}|_{\boldsymbol{u}^\perp}$ and $\boldsymbol{c}_{\tilde{w}}(\boldsymbol{u})$. The decoder's output $\hat{w} = \phi_n(\boldsymbol{y}, \boldsymbol{u}, \theta_1)$ is thus given as

$$\hat{w} = \phi_n(\boldsymbol{y}, \boldsymbol{u}, \theta_1) = \operatorname*{arg\,min}_{1\le\tilde{w}\le 2^{nR}} \left\| \boldsymbol{y}|_{\boldsymbol{u}^\perp} - \boldsymbol{c}_{\tilde{w}}(\boldsymbol{u})\right\|^2 \tag{4.13}$$

$$= \operatorname*{arg\,max}_{1\le\tilde{w}\le 2^{nR}} \langle \boldsymbol{y}|_{\boldsymbol{u}^\perp}, \boldsymbol{c}_{\tilde{w}}(\boldsymbol{u})\rangle, \tag{4.14}$$

where the last equality follows by noting that $n^{-1}\|\boldsymbol{c}_{\tilde{w}}(\boldsymbol{u})\|^2 = b_2$ irrespective of $\tilde{w}$; see (4.8).

If more than one message achieves the minimum in (4.13), then an error is declared. Note that $\hat{w}$ of (4.13) is with probability one unique.

4.2.2 Analysis of Probability of Error

We now proceed to analyze our proposed encoding and decoding scheme. To this end we shall find it convenient to define the random variables,

$$Z_1 = \frac{1}{n}\left\|\boldsymbol{Y}|_{\boldsymbol{U}^\perp}\right\|^2, \tag{4.15}$$

$$Z_2 = \frac{1}{n}\left\langle \boldsymbol{Y}|_{\boldsymbol{U}^\perp}, \boldsymbol{C}_W(\boldsymbol{U})\right\rangle, \tag{4.16}$$

and the mapping

$$\beta_1(z_1, z_2) = \frac{z_2}{\sqrt{b_2 z_1}},$$

which will be shown to capture the effect of the attacker on the decoder's performance. Note that $|\beta_1(Z_1, Z_2)| \leq 1$, which follows from (4.8), (4.15), and (4.16) using the Cauchy-Schwarz inequality.

By the definition of the decoder (4.14) and of the random variable Z_2 (4.16) it follows that a decoding error occurs if, and only if, there exists a message $w' \neq W$ such that

$$\begin{aligned} \frac{1}{n}\left\langle \boldsymbol{Y}|_{\boldsymbol{U}^\perp}, \boldsymbol{C}_{w'}(\boldsymbol{U})\right\rangle &\geq \frac{1}{n}\left\langle \boldsymbol{Y}|_{\boldsymbol{U}^\perp}, \boldsymbol{C}_W(\boldsymbol{U})\right\rangle \\ &= Z_2. \end{aligned}$$

Equivalently, an error occurs if, and only if, there exists some $w' \neq W$ such that

$$\begin{aligned} \left\langle \frac{\boldsymbol{Y}|_{\boldsymbol{U}^\perp}}{\sqrt{nZ_1}}, \frac{\boldsymbol{C}_{w'}(\boldsymbol{U})}{\sqrt{nb_2}}\right\rangle &\geq \frac{Z_2}{\sqrt{b_2 Z_1}} \\ &= \beta_1(Z_1, Z_2). \end{aligned} \tag{4.17}$$

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 (4.17).

Conditional on the covertext $\boldsymbol{U} = \boldsymbol{u}$ and for any message $w' \neq W$, the random vector $\boldsymbol{C}_{w'}(\boldsymbol{u})/\sqrt{nb_2}$ is uniformly distributed over $\mathcal{S}^n(0,1) \cap \boldsymbol{u}^\perp$ (i.e., all unit vectors that are orthogonal to $\boldsymbol{u}$) and is independent of the random vector $\boldsymbol{Y}|_{\boldsymbol{u}^\perp}/\sqrt{nZ_1}$, which also takes value on $\mathcal{S}^n(0,1) \cap \boldsymbol{u}^\perp$. Since $\mathcal{S}^n(0,1) \cap \boldsymbol{u}^\perp$ is isometric to $\mathcal{S}^{n-1}(0,1)$,[1] it follows from the above observation that the distribution of the random variable on the LHS of (4.17) does not depend on the distribution of $\boldsymbol{Y}|_{\boldsymbol{u}^\perp}/\sqrt{nZ_1}$. Consequently, for any $w' \neq W$,

$$\Pr\left(\left\langle \frac{\boldsymbol{Y}|_{\boldsymbol{U}^\perp}}{\sqrt{nZ_1}}, \frac{\boldsymbol{C}_{w'}(\boldsymbol{U})}{\sqrt{nb_2}} \right\rangle \geq \beta_1(z_1,z_2) \Big| Z_1 = z_1, Z_2 = z_2, \boldsymbol{U} = \boldsymbol{u}\right) = \frac{C_{n-1}\big(\arccos \beta_1(z_1,z_2)\big)}{C_{n-1}(\pi)}, \quad (4.18)$$

where recall that $C_{n-1}(\theta)$ is the surface area of a spherical cap of half-angle θ on an $(n-1)$-dimensional unit sphere.

To continue the analysis of the probability of a decoding error, we note that conditional on $\boldsymbol{U} = \boldsymbol{u}$, the random vectors $\{\boldsymbol{C}_{w'}(\boldsymbol{u}) : w' \neq W\}$ are independent of each other. Thus, the probability of correct decoding is given by the product of the probabilities that each of these $2^{nR} - 1$ vectors did not cause an error. Since the probability of error for each individual vector is given in (4.18), we can write the conditional probability of error for this coding strategy as

$$\Pr(\text{error}|Z_1 = z_1, Z_2 = z_2, \boldsymbol{U} = \boldsymbol{u}) = \Pr(\text{error}|Z_1 = z_1, Z_2 = z_2) = 1 - \left(1 - \frac{C_{n-1}\big(\arccos \beta_1(z_1,z_2)\big)}{C_{n-1}(\pi)}\right)^{2^{nR}-1}. \quad (4.19)$$

We now find an upper bound on the average of the RHS of (4.19) over the random variables Z_1 and Z_2. The function $\Pr(\text{error}|Z_1 = z_1, Z_2 = z_2)$ is a monotonically non-increasing function of $\beta_1(z_1,z_2)$ and is upper bounded by one. Consequently, for any real number Υ we have

$$\Pr(\text{error}) \leq \Pr(\text{error}|\beta_1(Z_1,Z_2) = \Upsilon) + \Pr\big(\beta_1(Z_1,Z_2) < \Upsilon\big). \quad (4.20)$$

[1] To see this, it is sufficient to consider $\boldsymbol{u} = (1,0,\ldots,0)$. In this case, $\boldsymbol{u}' \in \mathcal{S}^n(0,1) \cap \boldsymbol{u}^\perp$ if $u'_1 = 0$ and $\sum_{i=2}^n u'_i = 1$.

We will show that the RHS of (4.20) is small when $\Upsilon = \beta_1^* - \epsilon_1$, where

$$\beta_1^* = \sqrt{\frac{cb_2}{cb_2 + D_2}}, \tag{4.21}$$

$c = c(A; D_2) = 1 - \frac{D_2}{A}$ (see (A.4)) and ϵ_1 is a small positive number to be specified later. We analyze the first term on the RHS of (4.20) in Lemma 4.2 and the second term in Lemma 4.3. In order to do so, we recall that Shannon [Sha59] derived bounds on the ratio of the surface areas of spherical caps that asymptotically yield

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \frac{C_n(\arccos\tau)}{C_n(\pi)} &= \log\big(\sin(\arccos\tau)\big) \\ &= \log(1-\tau^2), \end{aligned} \tag{4.22}$$

for every $0 < \tau < 1$; see also [Wyn67]. We shall also need the following lemma.

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

$$-\eta_1 = \lim_{t\to\infty} \frac{1}{t} \log f(t) \tag{4.23}$$

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

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

Proof. First, recall the well known fact that

$$\lim_{t\to\infty} \left(1 - 2^{-t\eta_1}\right)^{2^{t\eta_2}} = \begin{cases} 1 & \text{if } \eta_1 > \eta_2 \\ e^{-1} & \text{if } \eta_1 = \eta_2 \\ 0 & \text{if } \eta_1 < \eta_2 \end{cases}. \tag{4.24}$$

Fix $\epsilon > 0$. Let us consider the case where $\eta_1 > \eta_2$. There exists a t_1 such that $t^{-1}\log(f(t)) - (-\eta_1) > \frac{\eta_2 - \eta_1}{2}$ for all $t > t_1$ since $\eta_1 > \eta_2$ and by (4.23). There also exists a t_2 such that

$$\left(1 - 2^{-t(\eta_1+\eta_2)/2}\right)^{2^{t\eta_2}} > 1 - \epsilon$$

for all $t > t_2$ since $(\eta_1+\eta_2)/2 > \eta_2$ and by (4.24). Thus, we can write that $(1-f(t))^{2^{t\eta_2}} > 1-\epsilon$ for all $t > \max\{t_1, t_2\}$. The claim follows in this case since $(1-f(t))^{2^{t\eta_2}} \leq 1$.

The claim follows in the case $\eta_1 < \eta_2$ by similar logic. □

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

$$1 - \left(1 - \frac{C_{n-1}\big(\arccos(\beta_1^* - \epsilon_1)\big)}{C_{n-1}(\pi)}\right)^{2^{nR}-1} < \epsilon.$$

Proof. With the definitions of β_1^* (4.21) and R (4.7) we have

$$\frac{1}{2}\log\left(\frac{1}{1-(\beta_1^*)^2}\right) = R + \delta,$$

and consequently there must exist some $\epsilon_1 > 0$ such that

$$R < \frac{1}{2}\log\left(\frac{1}{1-(\beta_1^* - \epsilon_1)^2}\right). \tag{4.25}$$

By the result on the asymptotic area of spherical caps (4.22) and by the inequality (4.25), it follows from Lemma 4.1 that there exists a positive integer n_1 such that for all $n > n_1$

$$\left(1 - \frac{C_{n-1}\big(\arccos(\beta_1^* - \epsilon_1)\big)}{C_{n-1}(\pi)}\right)^{2^{nR}} > 1 - \epsilon,$$

and the claim follows by noting that the LHS cannot decrease when the exponent 2^{nR} is replaced by $2^{nR} - 1$. □

Our achievability proof will thus be complete once we demonstrate that the second term on the RHS of (4.20) converges to zero for $\Upsilon = \beta_1^* - \epsilon_1$. This is demonstrated in the following lemma, which is proved in Appendix B.6 and which concludes the achievability proof for the private version of the SGWM game.

Lemma 4.3. *For any $\epsilon > 0$ and $\epsilon_1 > 0$, there exists an integer $n_2 > 0$ such that for all*

$n > n_2$

$$\Pr\big(\beta_1(Z_1, Z_2) < \beta_1^* - \epsilon_1\big) < \epsilon.$$

4.3 Achievability for Public Version

In this section, we show that all rates up to $C^*(D_1, D_2, \sigma_u^2)$ and $\frac{1}{2}\log(1+\frac{D_1}{D_2})$ are achievable for the public version of the general watermarking game and for the additive attack watermarking game, respectively, when the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$. We extend these results to IID Gaussian covertexts in Section 4.4.

4.3.1 Coding Strategy

The coding strategies for the public versions of both the additive attack and the general watermarking games are motivated by the works of Marton [Mar79], Gel'fand and Pinsker [GP80], Heegard and El Gamal [HEG83], and Costa [Cos83].

For both models, we fix a $\delta > 0$. In the following subsections, we define the set of constants $\{\alpha, \sigma_v^2, R_0, R_1, R\}$ separately for each model. Using these constants we then describe the encoder and decoder used for both models. Thus, while the constants have different values for the two models, in terms of these constants the proposed coding schemes are identical.

Constants for the Additive Attack Watermarking Game

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

$$\alpha = \frac{D_1}{D_1 + D_2}, \tag{4.26}$$

$$\sigma_v^2 = D_1 + \alpha^2\sigma_u^2, \tag{4.27}$$

$$R_0 = \frac{1}{2}\log\left(1 + \frac{D_1\sigma_u^2}{(D_1+D_2)^2}\right) + \delta, \tag{4.28}$$

$$R_1 = \frac{1}{2}\log\left(1 + \frac{D_1}{D_2} + \frac{D_1\sigma_u^2}{D_2(D_1+D_2)}\right) - \delta, \tag{4.29}$$

and

$$R = R_1 - R_0 = \frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right) - 2\delta. \tag{4.30}$$

Constants for the General Watermarking Game

The choice of the constants for the general watermarking game is inspired by the solution to the public Gaussian mutual information game; see Theorem 3.1 and its derivation in Section 3.2. 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 (A.7). 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.

First, let $\rho = \rho(A; D_1, \sigma_u^2)$, $b_1 = b_1(A; D_1, \sigma_u^2)$, $b_2 = b_2(A; D_1, \sigma_u^2)$, $c = c(A; D_2)$ and $\alpha = \alpha(A; D_1, D_2, \sigma_u^2)$ as defined in Appendix A. In particular, recall that $\alpha = 1 - \frac{b_1 D_2}{D_2 + cb_2}$. We can then define the other constants as

$$\sigma_v^2 = \alpha^2\sigma_u^2 + 2\alpha\rho + D_1, \tag{4.31}$$

$$R_0 = \frac{1}{2}\log\left(1 + \frac{(\alpha\sigma_u^2 + \rho)^2}{D_1\sigma_u^2 - \rho^2}\right) + \delta, \tag{4.32}$$

$$R_1 = \frac{1}{2}\log\left(1 + \frac{Acb_2}{D_2(D_2 + cb_2)}\right) - \delta, \tag{4.33}$$

and

$$R = R_1 - R_0 = \frac{1}{2}\log\big(1 + s(A; D_1, D_2, \sigma_u^2)\big) - 2\delta, \tag{4.34}$$

where $s(A; D_1, D_2, \sigma_u^2)$ is defined in (A.5). If A is chosen to maximize (4.34) as in (4.1), then $R = C^*(D_1, D_2, \sigma_u^2) - 2\delta$; we show in Lemma A.1 that such an A can be chosen.

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 $2^{nR_1} = 2^{n(R+R_0)}$ IID random vectors $\{\boldsymbol{V}_{j,k}\}$, where $1 \le j \le 2^{nR}$, $1 \le k \le 2^{nR_0}$, and each random vector $\boldsymbol{V}_{j,k}$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_v^2})$. Thus, the codebook consists of 2^{nR} bins (indexed by j), each containing 2^{nR_0} auxiliary codewords. In Figure 4-1, we give an example codebook with

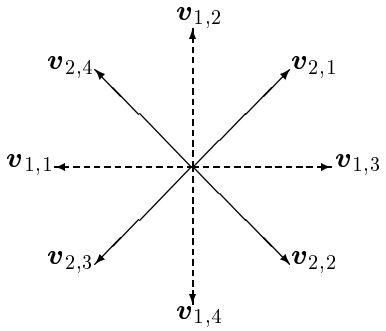

Figure 4-1: Example codebook for public version. Dashed vectors are in bin 1 and dotted vectors are in bin 2.

$n = 2$, $R_0 = 1$ and $R = 1/2$. Instead of being selected randomly, the codewords in this example have been placed regularly in the 2-sphere (i.e., circle).

Encoder: 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.

Mathematically, the encoder behaves 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

$$k^*(\boldsymbol{u}, w) = \underset{1 \le k \le 2^{nR_0}}{\arg\max} \langle \boldsymbol{u}, \boldsymbol{v}_{w,k} \rangle, \tag{4.35}$$

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

$$\boldsymbol{v}_w(\boldsymbol{u}) = \boldsymbol{v}_{w,k^*(\boldsymbol{u},w)}. \tag{4.36}$$

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

$$\boldsymbol{x} = \boldsymbol{v}_w(\boldsymbol{u}) + (1 - \alpha)\boldsymbol{u}. \tag{4.37}$$

The example of Figure 4-1 is continued in Figure 4-2, where the encoding procedure is illustrated.

Decoder: The decoder finds the auxiliary codeword that, among all the 2^{nR_1} sequences in

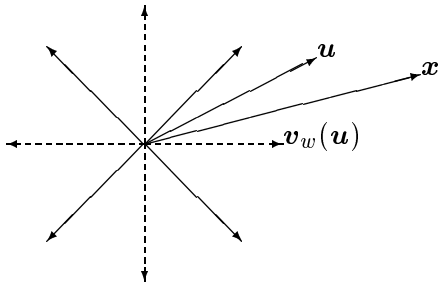

Figure 4-2: Example encoding for public version with $w = 1$ (bin with dashed vectors).

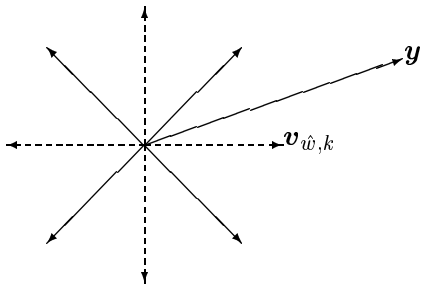

Figure 4-3: Example decoding for public version.

the codebook, is closest to the received sequence $\boldsymbol{y}$. He then declares the estimate of the message to be the bin to which this auxiliary codeword belongs. Mathematically, given the received sequence $\boldsymbol{y}$ and the codebook $\{\boldsymbol{v}_{j,k}\}$, the estimate is given by

$$\hat{w} = \underset{1\leq\tilde{w}\leq 2^{nR}}{\arg\min} \left(\min_{1\leq k\leq 2^{nR_0}} \|\boldsymbol{y} - \boldsymbol{v}_{\tilde{w},k}\|^2 \right) \tag{4.38}$$

$$= \underset{1\leq\tilde{w}\leq 2^{nR}}{\arg\max} \left(\max_{1\leq k\leq 2^{nR_0}} \langle \boldsymbol{y}, \boldsymbol{v}_{\tilde{w},k} \rangle \right), \tag{4.39}$$

where the last equality follows by noting that $n^{-1}\|\boldsymbol{v}_{\tilde{w},k}\|^2 = \sigma_v^2$ irrespective of $\tilde{w}$ and k. Note that $\hat{w}$ of (4.38) is with probability one unique. The example is completed in Figure 4-3 with an illustration of the decoding process. In this example, the decoder successfully recovered the value of the watermark.

4.3.2 Probability of Error

In this section, we derive the conditional probability of error in the above coding strategy. We first define the random variables on which we will condition. Let the random variable Z be the maximum (normalized) inner product achieved in (4.35),

$$Z = \frac{1}{n}\langle \boldsymbol{U}, \boldsymbol{V}_W(\boldsymbol{U})\rangle. \tag{4.40}$$

Next, let the random variable Z_3 be the normalized power in the sequence $\boldsymbol{Y}$,

$$Z_3 = \frac{1}{n}\|\boldsymbol{Y}\|^2. \tag{4.41}$$

Next, let the random variable Z_4 be the normalized inner product between the sequence $\tilde{\boldsymbol{Y}}$, which is defined by

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}, \tag{4.42}$$

and the auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$,

$$Z_4 = \frac{1}{n}\langle \tilde{\boldsymbol{Y}}, \boldsymbol{V}_W(\boldsymbol{U})\rangle. \tag{4.43}$$

Finally, let us define a mapping $\beta_2(z, z_3, z_4)$ as

$$\beta_2(z, z_3, z_4) = \frac{\sigma_v^2 + (1-\alpha)z + z_4}{\sqrt{z_3\sigma_v^2}}. \tag{4.44}$$

By the definition of the decoder (4.39), it follows that a decoding error occurs if, and only if, there exists a message $w' \neq W$ and an index k' such that

$$\begin{aligned}
\frac{1}{n}\langle \boldsymbol{Y}, \boldsymbol{V}_{w',k'}\rangle &\geq \frac{1}{n}\langle \boldsymbol{Y}, \boldsymbol{V}_W(\boldsymbol{U})\rangle \\
&= \frac{1}{n}\langle \boldsymbol{X}, \boldsymbol{V}_W(\boldsymbol{U})\rangle + \frac{1}{n}\langle \tilde{\boldsymbol{Y}}, \boldsymbol{V}_W(\boldsymbol{U})\rangle \\
&= \sigma_v^2 + (1-\alpha)Z + Z_4,
\end{aligned}$$

where the first equality follows by the definition of $\tilde{\boldsymbol{Y}}$ (4.42) and the second equality follows by the definitions of the encoder (4.37) and the random variables Z and Z_4. 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

$$\begin{aligned} \left\langle \frac{\boldsymbol{Y}}{\sqrt{nZ_3}}, \frac{\boldsymbol{V}_{w',k'}}{\sqrt{n\sigma_v^2}} \right\rangle &\geq \frac{\sigma_v^2 + (1-\alpha)Z + Z_4}{\sqrt{Z_3\sigma_v^2}} \\ &= \beta_2(Z, Z_3, Z_4). \end{aligned} \tag{4.45}$$

The random vector $\boldsymbol{V}_{w',k'}/\sqrt{n\sigma_v^2}$ is uniformly distributed on the unit n-sphere $\mathcal{S}^n(0,1)$ and is independent of $\boldsymbol{Y}$, Z, Z_3, and Z_4. Indeed, the encoder does not examine the auxiliary codewords in bins other than in the one corresponding to the message W. The random vector $\boldsymbol{Y}/\sqrt{nZ_3}$ also takes value on the unit n-sphere $\mathcal{S}^n(0,1)$, and thus, by symmetry (see Section 4.2.2), the distribution of the LHS of (4.45) does not depend on the distribution of $\boldsymbol{Y}$. In particular, for any $w' \neq W$,

$$\Pr\left(\left\langle \frac{\boldsymbol{Y}}{\sqrt{nz_3}}, \frac{\boldsymbol{V}_{w',k'}}{\sqrt{n\sigma_v^2}} \right\rangle \geq \beta_2(z, z_3, z_4) \middle| Z = z, Z_3 = z_3, Z_4 = z_4 \right) = \frac{C_n\big(\arccos \beta_2(z, z_3, z_4)\big)}{C_n(\pi)}. \tag{4.46}$$

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 that there was *not* an error is given by the product of the probabilities that each of these $2^{nR_1} - 2^{nR_0}$ vectors did *not* cause an error. Since the probability of error for each individual vector is given in (4.46), we can write the conditional probability of error for this coding strategy as

$$\Pr(\text{error}|Z = z, Z_3 = z_3, Z_4 = z_4) = 1 - \left(1 - \frac{C_n\big(\arccos \beta_2(z, z_3, z_4)\big)}{C_n(\pi)}\right)^{2^{nR_1} - 2^{nR_0}}. \tag{4.47}$$

The expression $\Pr(\text{error}|Z = z, Z_3 = z_3, Z_4 = z_4)$ is a monotonically non-increasing function of $\beta_2(z, z_3, z_4)$ and is upper-bounded by 1. Consequently, as in Section 4.2.2,

$$\Pr(\text{error}) \leq \Pr\big(\text{error}|\beta_2(Z, Z_3, Z_4) = \Upsilon\big) + \Pr\big(\beta_2(Z, Z_3, Z_4) < \Upsilon\big), \tag{4.48}$$

for any real number Υ. For both games under consideration, we will show that, by choosing a sufficiently large blocklength n, the RHS of (4.48) can be made arbitrarily small when

$\Upsilon = \beta^*(R_1 + \delta) - \epsilon_2$. Here

$$\beta^*(R_1 + \delta) = \left(1 - 2^{-2(R_1+\delta)}\right)^{1/2}, \tag{4.49}$$

ϵ_2 is a small number to be specified later, and the constant R_1 is defined in (4.29) and (4.33) for the additive attack and general watermarking games, respectively.

We now analyze the first term on the RHS of (4.48) for both games simultaneously. The analysis of the second term is performed separately for the additive attack watermarking game in Lemma 4.8 and for the general watermarking game in Lemma 4.10.

Lemma 4.4. *For any $\epsilon > 0$, there exists some $\epsilon_2 > 0$ and some integer $n_1 > 0$ such that for all $n > n_1$*

$$1 - \left(1 - \frac{C_n\Big(\arccos\big(\beta^*(R_1 + \delta) - \epsilon_2\big)\Big)}{C_n(\pi)}\right)^{2^{nR_1} - 2^{nR_0}} < \epsilon,$$

where R_1 is defined according to either (4.29) or (4.33).

Proof. Rewriting (4.49) as

$$\frac{1}{2}\log\left(\frac{1}{1 - \big(\beta^*(R_1 + \delta)\big)^2}\right) = R_1 + \delta,$$

demonstrates the existence of some $\epsilon_2 > 0$ such that

$$\frac{1}{2}\log\left(\frac{1}{1 - \big(\beta^*(R_1 + \delta) - \epsilon_2\big)^2}\right) > R_1, \tag{4.50}$$

because in both (4.29) and (4.33) the rate R_1 satisfies $0 < \beta^*(R_1 + \delta) < 1$. By the result on the asymptotic area of spherical caps (4.22) and by the inequality (4.50), it follows by Lemma 4.1 that there exists a positive integer n_1 such that for all $n > n_1$

$$\left(1 - \frac{C_n\Big(\arccos\big(\beta^*(R_1 + \delta) - \epsilon_2\big)\Big)}{C_n(\pi)}\right)^{2^{nR_1}} > 1 - \epsilon,$$

and the claim follows by noting that the LHS cannot decrease when the exponent 2^{nR_1} is replaced by $2^{nR_1} - 2^{nR_0}$. □

4.3.3 Distribution of Chosen Auxiliary Codeword

To continue with the performance analysis, we shall need the distribution of the chosen auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$ (defined in (4.36)), both unconditionally and conditioned on the random vector $\boldsymbol{X}$ and the random variable Z (defined in (4.37) and (4.40), respectively).

Lemma 4.5. *The random vector $\boldsymbol{V}_W(\boldsymbol{U})$ defined in (4.36) is uniformly distributed over the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_v^2})$.*

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 that $W = 1$.

Since all the auxiliary random vectors $\{\boldsymbol{V}_{1,k}\}$ in bin 1 take value in the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_v^2})$, 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 4.6. *Given $\boldsymbol{X} = \boldsymbol{x}$ and $Z = z$, the random vector $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the set*

$$\mathcal{V}(\boldsymbol{x}, z) = \left\{ a_1 \boldsymbol{x} + \boldsymbol{v} : \boldsymbol{v} \in \mathcal{S}^n(0, \sqrt{na_2}) \cap \boldsymbol{x}^\perp \right\}, \tag{4.51}$$

where

$$a_1 = \frac{\sigma_v^2 + (1-\alpha)z}{n^{-1}\|\boldsymbol{x}\|^2},$$

and

$$a_2 = \frac{(1-\alpha)^2(\sigma_u^2\sigma_v^2 - z^2)}{n^{-1}\|\boldsymbol{x}\|^2}. \tag{4.52}$$

The proof of this lemma can be found in Appendix B.7.

4.3.4 Analysis for Additive Attack Watermarking Game

Using the encoder and decoder described above, we now show that for any positive δ the rate R defined in (4.30) is achievable for the additive attack watermarking game (when the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$). That is, the probabilities that the encoder does not meet the distortion constraint and that a decoding error occur can both be made arbitrarily small by choosing some finite blocklength n. In order to prove these facts, we first show in the following subsection that the random variable Z takes value close to $\alpha\sigma_u^2$ with high probability.

A Law of Large Numbers for Z

In this section, we state and prove a claim that describes the behavior of the random variable Z defined in (4.40). This claim will be used to show that encoder and decoder behave properly (i.e., meeting the distortion constraint and recovering the correct message) with arbitrarily high probability.

Lemma 4.7. *If the constants defined for the additive attack watermarking game are used to design the sequence of encoders of Section 4.3.1, then*

$$\lim_{n\to\infty} \Pr\big(Z \geq \alpha\sigma_u^2\big) = 1.$$

Proof. Let $\boldsymbol{V}$ be uniformly distributed on $\mathcal{S}^n(0, \sqrt{n\sigma_v^2})$ independent of $\boldsymbol{U}$. Then

$$\begin{aligned} \Pr(Z \geq \alpha\sigma_u^2) &= 1 - \Pr\left(\max_{1\leq k\leq 2^{nR_0}} n^{-1}\langle \boldsymbol{U}, \boldsymbol{V}_{W,k}\rangle < \alpha\sigma_u^2\right) \\ &= 1 - \left(1 - \Pr\left(n^{-1}\langle \boldsymbol{U}, \boldsymbol{V}\rangle \geq \alpha\sigma_u^2\right)\right)^{2^{nR_0}}, \end{aligned} \tag{4.53}$$

where the first equality follows by the definition of Z (4.40) and of $\boldsymbol{V}_{W,k}$, and the second equality follows because $\{\boldsymbol{V}_{W,k}\}_{k=1}^{2^{nR_0}}$ are IID and also independent of $\boldsymbol{U}$. The RHS of (4.53) can be further simplified using

$$\begin{aligned} \Pr\left(\frac{1}{n}\langle \boldsymbol{U}, \boldsymbol{V}\rangle \geq \alpha\sigma_u^2\right) &= \Pr\left(\left\langle \frac{\boldsymbol{U}}{\sqrt{n\sigma_u^2}}, \frac{\boldsymbol{V}}{\sqrt{n\sigma_v^2}}\right\rangle \geq \frac{\alpha\sigma_u}{\sigma_v}\right) \\ &= \frac{C_n\left(\arccos\left(\frac{\alpha\sigma_u}{\sigma_v}\right)\right)}{C_n(\pi)}, \end{aligned} \tag{4.54}$$

which follows since both normalized random vectors are uniformly distributed on $\mathcal{S}^n(0,1)$ and they are independent of each other. By (4.22) we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{C_n\left(\arccos\left(\frac{\alpha\sigma_u}{\sigma_v}\right)\right)}{C_n(\pi)} = \frac{1}{2}\log\left(1-\frac{\alpha^2\sigma_u^2}{\sigma_v^2}\right) \quad (4.55)$$

$$= -(R_0-\delta), \quad (4.56)$$

where the second equality follows by the definitions of α (4.26), σ_v^2 (4.27), and R_0 (4.28). Combining Lemma 4.1 with (4.53), (4.54), and (4.56) concludes the proof. □

The Encoding Distortion Constraint

We now show that the encoder's distortion constraint is met with arbitrarily high probability. By Lemma 4.7, it is sufficient to show that $Z \geq \alpha\sigma_u^2$ 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 (4.37) and (4.40)),

$$n^{-1}\|\boldsymbol{X}-\boldsymbol{U}\|^2 = \sigma_v^2 - 2\alpha Z + \alpha^2\sigma_u^2. \quad (4.57)$$

Since α is positive (4.26), the RHS of (4.57) is decreasing in Z. Consequently, the condition $Z \geq \alpha\sigma_u^2$ implies

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

where the last equality follows from (4.27).

The Decoding Error

We now show that the second term on the RHS of (4.48) is vanishing in n when $\Upsilon = \beta^*(R_1+\delta)-\epsilon_2$. Here R_1 and $\beta^*(R_1+\delta)$ are defined in (4.29) and (4.49) respectively, and $\epsilon_2 > 0$ is specified in Lemma 4.4. The combination of this fact with Lemma 4.4 will show that, as the blocklength n tends to infinity, the probability of decoding error approaches zero. The following lemma is proved in Appendix B.8.

Lemma 4.8. *If the constants defined for the additive attack watermarking game are used to design the sequence of encoders of Section 4.3.1, then for any $\epsilon > 0$ and $\epsilon_2 > 0$, there*

exists an integer $n_2 > 0$ such that for all $n > n_2$ and for all the deterministic attacks of Section 4.1.1

$$\Pr\big(\beta_2(Z, Z_3, Z_4) < \beta^*(R_1 + \delta) - \epsilon_2\big) < \epsilon.$$

4.3.5 Analysis for General Watermarking Game

We return to the public version of the general watermarking game to demonstrate that the encoder and decoder for the general watermarking game (defined in Section 4.3.1) guarantee that the rate R of (4.34) is achievable, for any $\delta > 0$. That is, we show that both the probability that the encoding distortion constraint is not met and the probability of a decoding error are vanishing in the blocklength n. We first show in the following subsection that the random variable Z concentrates around $\alpha\sigma_u^2 + \rho$.

A Law of Large Numbers for Z

In this section, we prove a law of large numbers for the random variable $Z = \frac{1}{n}\langle \boldsymbol{U}, \boldsymbol{V}_W(\boldsymbol{U})\rangle$, which is defined in (4.40), and which corresponds to the normalized inner product between the source sequence $\boldsymbol{U}$ and the chosen auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$. This law will be useful for the later analysis of the probability of exceeding the allowed encoder distortion and the probability of a decoding error.

Lemma 4.9. *For every $\delta > 0$ used to define the encoder for the general watermarking game (see equations (4.32), (4.33), (4.34) and Section 4.3.1), there exists $\epsilon(\delta) > 0$ such that*

$$\lim_{n\to\infty} \Pr\big(\alpha\sigma_u^2 + \rho \le Z \le \alpha\sigma_u^2 + \rho + \epsilon(\delta)\big) = 1,$$

and

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

Proof. The proof that $\Pr(Z \ge \alpha\sigma_u^2 + \rho) \to 1$ is almost identical to the proof of Lemma 4.7. One need only replace $\alpha\sigma_u^2$ with $\alpha\sigma_u^2 + \rho$ and use the definitions of the constants that are for the general watermarking game as opposed to the constants for the additive attack watermarking game; see Section 4.3.1.

To complete the proof of the present claim, we now choose $\epsilon(\delta) > 0$ such that

$$\log\left(1 - \left(\frac{\alpha\sigma_u^2 + \rho + \epsilon(\delta)}{\sigma_u\sigma_v}\right)^2\right) < -R_0. \tag{4.58}$$

This can be done because the LHS of (4.58) equates to $-(R_0+\delta)$ when $\epsilon(\delta)$ is set to zero (in analogy to the equality between (4.55) and (4.56)), and because $\log(1-x^2)$ is continuous and decreasing in x, for $0 < x < 1$. Using Lemma 4.1, we see that $\Pr(Z > \alpha\sigma_u^2 + \epsilon(\delta)) \to 0$. Finally, we can choose $\epsilon(\delta) \to 0$ as $\delta \to 0$ by the continuity of $\log(1-x^2)$. □

The Encoding Distortion Constraint

We now show that for an appropriate choice of n and δ, the distortion constraint is met with arbitrarily high probability. As in Section 4.3.4, if $\alpha \geq 0$, then (4.57) demonstrates that whenever $Z \geq \alpha\sigma_u^2 + \rho$ holds we also have $n^{-1}\|\boldsymbol{X} - \boldsymbol{U}\|^2 \leq D_1$. Thus, our claim follows from Lemma 4.9 if $\alpha \geq 0$.

However, contrary to the additive attack game, in the general watermarking game the constant α need not be non-negative. To address this case we note that for $\alpha < 0$, 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 4.9 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.

The Decoding Error

In this section, we show that the second term on the RHS of (4.48) is vanishing in n when $\Upsilon = \beta^*(R_1 + \delta) - \epsilon_2$. Here R_1 and $\beta^*(R_1 + \delta)$ are defined in (4.33) and (4.49) and ϵ_2 is specified in Lemma 4.4. The combination of this fact with Lemma 4.4 will show that the probability of decoding error approaches zero, as the blocklength n tends to infinity. We state the desired result in the following lemma, which is proved in Appendix B.9.

Lemma 4.10. *If the constants defined for the general watermarking game are used to design the sequence of encoders of Section 4.3.1, then for any $\epsilon > 0$ and $\epsilon_2 > 0$, there exists*

an integer $n_2 > 0$ such that for all $n > n_2$ and for all attackers of Section 4.1.2

$$\Pr\big(\beta_2(Z, Z_3, Z_4) < \beta^*(R_1 + \delta) - \epsilon_2\big) < \epsilon.$$

4.4 Spherically Uniform Covertext is Sufficient

We have shown in the early sections of this chapter that if the covertext $\boldsymbol{U}$ is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$, then the coding capacity of both the private and public versions of the watermarking games are lower bounded by $C^*(D_1, D_2, \sigma_u^2)$. We have also shown that for such covertexts, the coding capacity of the additive attack watermarking game is at least $\frac{1}{2}\log(1+\frac{D_1}{D_2})$. In this section, we extend these results to zero-mean variance-σ_u^2 IID Gaussian covertexts.

We first transform the IID Gaussian sequence $\boldsymbol{U}$ into a random vector $\boldsymbol{U}'$ that is uniformly distributed on the n-sphere $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$. 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 $\mathcal{S}^n(0, \sqrt{n\sigma_u^2})$.

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^* (2.6), 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 4.2.1 and 4.3.1. Let ϕ' be the corresponding decoder. Consider now

an encoder for the covertext $\boldsymbol{U}$ that 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 (2.1) 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 4.2.1 and 4.3.1 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

$$\lim_{n\to\infty} \Pr(\boldsymbol{X} = \boldsymbol{X}') = 1 \tag{4.59}$$

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 (4.59), and the first term approaches zero by the achievability results for covertexts that are uniformly distributed over the n-sphere.

To clarify the latter argument consider, for example, the public watermarking game with an additive attacker as in (4.2). We would then argue that

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

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

4.5 Converse for Squared Error Distortion

In this section, we prove the converse part of Theorem 2.1 for the watermarking game with real alphabets and squared error distortions. That is, we 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 $C^*(D_1, D_2, \sigma_u^2)$. In particular, for any fixed $R > C^*(D_1, D_2, \sigma_u^2)$ and any sequence of rate-R encoders that satisfy the distortion constraint (2.1), we will propose a sequence of attackers $\{g_n\}$ that satisfy the distortion constraint (2.3) and that guarantee that, irrespective of the decoding rule, the probability of error will be 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.

The remainder of this section is organized as follows. In Section 4.5.1, we describe the proposed sequence of attackers. In Section 4.5.2, we study the distortion they introduce, and in Section 4.5.3 we show that, for the appropriate rates, these attack strategies guarantee a probability of error that is bounded away from zero. We conclude with a discussion of the necessity of the ergodicity assumption in Section 4.5.4.

4.5.1 Attacker

Intuitive Definition

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 (3.1) — 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 3.2.1 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 3.2.1, 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 (2.1) (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 3.2.3 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 (2.3) 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 3.2.3 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 there are only finitely many intervals. These arguments will be made more precise in the upcoming sections.

Precise Definition

Let R be a fixed rate that 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

$$\tilde{D}_2 = D_2 - \tilde{\delta}, \tag{4.60}$$

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 $\mathcal{A}_k$,

$$\tilde{D}_2 \left(1 + \frac{\tilde{D}_2}{A}\left(\frac{A'}{A} - 1\right)\right) < \tilde{D}_2 + \frac{\tilde{\delta}}{2}, \ \forall A, A' \in \mathcal{A}_k. \tag{4.61}$$

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

$$k(\boldsymbol{x}) = \begin{cases} l & \text{if } n^{-1}\|\boldsymbol{x}\|^2 \in \mathcal{A}_l \\ 0 & \text{if no such } l \text{ exists} \end{cases}. \tag{4.62}$$

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

$$a_k = E\left[\frac{1}{n}\|\boldsymbol{X}\|^2 \middle| K = k\right], \ \forall\ 0 \leq k \leq m. \tag{4.63}$$

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

$$\mu_k = E\left[\frac{1}{n}\|\boldsymbol{U}\|^2 \middle| K = k\right], \ \forall\ 0 \leq k \leq m. \tag{4.64}$$

Using only the source of randomness Θ_2, the attacker generates a random vector $\boldsymbol{V}$ as a sequence of IID 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.

We now describe an attacker g_n^* that does not necessarily meet the distortion constraint. For this attacker, the stegotext is computed as

$$g_n^*(\boldsymbol{x}, \theta_2) = \begin{cases} c(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{x} + c^{1/2}(a_{k(\boldsymbol{x})}; \tilde{D}_2)\boldsymbol{v}(\theta_2) & \text{if } k(\boldsymbol{x}) > 0 \\ \left(\sqrt{nD_2} - \sqrt{n\tilde{D}_2}\right)\boldsymbol{v}(\theta_2)/\|\boldsymbol{v}(\theta_2)\| & \text{otherwise} \end{cases}, \tag{4.65}$$

where $c(A; D_2) = 1 - D_2/A$ (also see (A.4)). Conditionally on $\boldsymbol{X} = \boldsymbol{x}$ satisfying $k(\boldsymbol{x}) \geq 1$, 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 3.2.1.

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,

$$g_n(\boldsymbol{x}, \theta_2) = \begin{cases} g_n^*(\boldsymbol{x}, \theta_2) & \text{if } n^{-1}\|g_n^*(\boldsymbol{x}, \theta_2) - \boldsymbol{x}\|^2 \leq D_2 \\ \boldsymbol{x} & \text{otherwise} \end{cases}. \tag{4.66}$$

The attacker g_n thus satisfies the distortion constraint with probability one. Note that if instead of a_k being calculated as in (4.66) it was chosen arbitrarily from $\mathcal{A}_k$, then the upcoming proof would still be valid (provided that each A_k is small enough). The resulting attacker is independent of the encoder and decoder and guarantees that no rates greater than $C^*(D_1, D_2, \sigma_u^2)$ are achievable.

4.5.2 Analysis of Distortion

The attackers $\{g_n^*\}$ do not, in general, satisfy the distortion constraint (2.3). 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. In the terminology of (4.66) we shall thus show that

$$\lim_{n\to\infty} \Pr\big(g_n(\boldsymbol{X}, \Theta_2) = g_n^*(\boldsymbol{X}, \Theta_2)\big) = 1. \tag{4.67}$$

Once this is shown, for the purposes of proving the converse, it will suffice to show that, for the appropriate rates, the attackers $\{g_n^*\}$ guarantee a non-vanishing probability of error. 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 (4.60) so that $R > C^*(D_1, \tilde{D}_2, \sigma_u^2)$ and consider the attacker (4.65). Assume that we have managed to prove that the attackers $\{g_n^*\}$ of (4.65) guarantees a non-vanishing probability of error. In this case (4.67) 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.

We now turn to the proof of (4.67). In order to summarize the distortion introduced by the attacker, we define the following random variables,

$$\Delta_1(k) = c(a_k; \tilde{D}_2)\left(n^{-1}\|\boldsymbol{V}\|^2 - \tilde{D}_2\right), \quad k = 1, \ldots, m, \tag{4.68}$$

and

$$\Delta_2(k) = \left(c(a_k;\tilde{D}_2) - 1\right) c^{1/2}(a_k;\tilde{D}_2) n^{-1}\langle \boldsymbol{X}, \boldsymbol{V}\rangle, \quad k = 1, \ldots, m. \tag{4.69}$$

Note that for any $1 \leq k \leq m$, the random variables $\Delta_1(k)$ and $\Delta_2(k)$ converge to zero in probability, because $\boldsymbol{V}$ is a sequence of IID $\mathcal{N}(0, \tilde{D}_2)$ random variables independent of $\boldsymbol{X}$, and because $0 < c(a_k;\tilde{D}_2) < 1$ for all $1 \leq k \leq m$.

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 $n^{-1}\|\boldsymbol{X}\|^2 \leq \tilde{D}_2$ or $n^{-1}\|\boldsymbol{X}\|^2 > (2\sigma_u + \sqrt{D_1})^2$. In the former case,

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

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 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 (A.4), (4.68) and (4.69)), and the inequality follows by (4.61) 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.

4.5.3 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 (4.65) 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 (4.66).

This analysis is carried out in a series of claims. In Lemma 4.11 we use a Fano-type inequality to show that an achievable rate cannot exceed some limit of mutual informations. In Lemma 4.12, we upper bound these mutual informations by simpler expectations, and in Lemma 4.13 we finally show that, in the limit, these expectations do not exceed $C^*(D_1, D_2, \sigma_u^2)$.

Lemma 4.11. *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*

$$R \leq \liminf_{n\to\infty} \frac{1}{n} I_{P_{\boldsymbol{U}}P_{\Theta_1}P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1}P_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U},\Theta_1). \tag{4.70}$$

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 (4.70) 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

$$\lim_{n\to\infty} \frac{1}{n}\Big(I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U},\Theta_1) - I(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1)\Big) = 0. \tag{4.71}$$

To prove (4.71) 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 $\boldsymbol{X}$ (4.62). Thus, the mutual information on the RHS of (4.70) can be written as

$$I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U},\Theta_1) = I(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1) + I(K;\boldsymbol{Y}|\boldsymbol{U},\Theta_1). \tag{4.72}$$

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

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

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

$$\lim_{n\to\infty} \frac{1}{n} I(K;\boldsymbol{Y}|\boldsymbol{U},\Theta_1) = 0. \tag{4.73}$$

Equation (4.71) now follows from (4.73) and (4.72).

It now follows from Lemma 4.11 and from (4.71) that in order to prove that the rate R

is not achievable, it suffices to show that

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

We upper bound in the following Lemma, which is proved in Appendix B.10.

Lemma 4.12. *For any encoder with corresponding watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$ satisfying (2.1), if the attacker g_n^* of (4.65) with corresponding attack channel $P^*_{\boldsymbol{Y}|\boldsymbol{X}}$ is used, then*

$$\begin{aligned} \frac{1}{n} I_{P_{\boldsymbol{U}} P_{\Theta_1} P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1} P^*_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) &\leq \sum_{k=1}^{m} \Pr(K=k) \cdot \frac{1}{2} \log\big(1 + s(a_k; D_1, \tilde{D}_2, \mu_k)\big) \\ &\leq E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right]. \end{aligned} \tag{4.74}$$

To proceed with the proof of the converse we would now like to upper bound the RHS of (4.74). 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 B.11.

Lemma 4.13. *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, where*

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

P2. For every $\epsilon > 0$, $n > n_0(\epsilon)$, and event $\mathcal{E}$, 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 4.13 we can now upper bound the RHS of (4.74). 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 (4.65) is used, then

$$\limsup_{n\to\infty} E_K\left[C^*(D_1, \tilde{D}_2, \mu_K)\right] \leq C^*(D_1, \tilde{D}_2, \sigma_u^2). \tag{4.75}$$

To see this, let $\delta(\epsilon, n)$ and $n_0(\epsilon)$ be the mappings of Lemma 4.13 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 (4.64), it is clear that $E\left[n^{-1}\|\boldsymbol{U}\|^2 | K \in \mathcal{K}^*(\epsilon)\right] > \sigma_u^2 + 5\epsilon$. Thus, by the Lemma 4.13, $\Pr(K \in \mathcal{K}^*(\epsilon)) < \delta(\epsilon, n)$. Since $C^*(D_1, D_2, \sigma_u^2)$ is non-decreasing in σ_u^2 and is upper bounded by $\frac{1}{2}\log(1 + \frac{D_1}{D_2})$,

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

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,

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

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, (4.75) 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 (4.60) so that $R > C^*(D_1, \tilde{D}_2, \sigma_u^2)$ and consider the attacker g_n^* of (4.65). 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 4.11. Here the first inequality is justified by the choice of $\tilde{D}_2$ (4.60), the second inequality by (4.75), the third inequality by (4.74), and the final equality by (4.71).

4.5.4 Discussion: The Ergodicity Assumption

We have proved that the IID 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 $C^*\left(D_1, D_2, E\left[U_i^2\right]\right)$ 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 $n^{-1}\sum_{i=1}^{n} U_i^2$ approach zero, as n tends to infinity; see Appendix B.11.

This condition can sometimes be further relaxed if the process has an ergodic decomposition (see e.g. [Gra88]). 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 IID 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 described here is non-ergodic, it is still true that no rate above $C^*\left(D_1, D_2, E\left[U_i^2\right]\right)$ 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 (4.66) 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.

Chapter 5

The Vector Gaussian Watermarking Game

In this chapter, we prove Theorem 2.4 on the capacity of the vector Gaussian watermarking (VGWM) game. Recall that in the VGWM game, the covertext is a sequence of IID Gaussian random vectors with common $m \times m$ covariance matrix $S_{\boldsymbol{u}}$. Both distortion measures are squared Euclidean distance so that the distortion between the $m \times n$covertext matrix $\boldsymbol{u}$ and the $m \times n$ stegotext matrix $\boldsymbol{x}$ is given by

$$d_1(\boldsymbol{u}, \boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ji} - u_{ji})^2,$$

and similarly for the distortion between the stegotext and the forgery.

The remainder of the chapter is organized as follows. In Section 5.1, we show that it is sufficient to consider covariance matrices that are diagonal. In Section 5.2, we introduce some notation used throughout the rest of the chapter. In Section 5.3, we outline the proof of the theorem using several main lemmas. Sections 5.4-5.6 are devoted to proving these lemmas. Finally, in Section 5.7, we compare the optimal attack to optimal lossy compression of the stegotext.

5.1 Diagonal Covariance is Sufficient

Let us consider the eigenvalue-eigenvector decomposition of the $m \times m$ covariance matrix $S_{\boldsymbol{u}}$, which can be a general non-negative definite symmetric matrix. We can write this

matrix as

$$S_{\boldsymbol{u}} = \Theta \Sigma \Theta^t, \tag{5.1}$$

where Σ is diagonal with the eigenvalues $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_m^2)$ on the diagonal and Θ is a orthogonal matrix (i.e., $\Theta\Theta^t = I$) comprised of normalized eigenvectors. We will see that the capacity of the watermarking game depends only on the eigenvalues $\boldsymbol{\sigma}^2$, and thus we will make the simplifying assumption that $S_{\boldsymbol{u}}$ is diagonal.

Given a $m \times n$ covertext matrix $\boldsymbol{u}$, let $\boldsymbol{u}' = \Theta \boldsymbol{u}$ and let $\boldsymbol{x}'$ be a stegotext that satisfies the distortion constraint (2.1) with respect to $\boldsymbol{u}'$. Then, since Θ is an orthogonal transformation, $\boldsymbol{x} = \Theta^t \boldsymbol{x}'$ satisfies the distortion constraint with respect to $\boldsymbol{u}$. Similarly, if $\boldsymbol{y}$ satisfies the distortion constraint (2.3) with respect to $\boldsymbol{x}$, then $\boldsymbol{y}' = \Theta \boldsymbol{y}$ satisfies the distortion constraint with respect to $\boldsymbol{x}'$. Thus, a codebook for a general covariance covertext can be created from a codebook for a diagonal covariance covertext while retaining the same probability of error. Thus, it is sufficient to consider diagonal covariance covertexts.

Let $\boldsymbol{U}' = \Theta \boldsymbol{U} = (\Theta \boldsymbol{U}_1, \ldots, \Theta \boldsymbol{U}_n)$. While the columns of $\boldsymbol{U}$ are independent $\mathcal{N}(0, S_{\boldsymbol{u}})$ random vectors, the rows of $\boldsymbol{U}'$ are independent with row j containing a sequence of IID $\mathcal{N}(0, \sigma_j^2)$ random variables.

Throughout the sequel we will assume that the covariance matrix $S_{\boldsymbol{u}}$ is diagonal with diagonal elements $\boldsymbol{\sigma}^2$. Furthermore, we will write the covertext as a $n \times m$ random matrix

$$\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m), \tag{5.2}$$

where the columns $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m$ are independent and column j contains a length-n IID sequence of $\mathcal{N}(0, \sigma_j^2)$ random variables. We will also write the stegotext $\boldsymbol{x}$ and the forgery $\boldsymbol{y}$ as $n \times m$ matrices. We will refer to the columns of $\boldsymbol{U}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the *components* of the covertext, stegotext and forgery, respectively.

5.2 Definitions

In this section, we give some definitions that will be used throughout the rest of the chapter. Also see Appendix A where many definitions used here and other chapters are summarized.

In order to differentiate between distortion in a component and total distortion, we will

use Δ_1 and Δ_2 for the total allowed distortion for the encoder and attacker, respectively. We use vectors $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ to describe the amount of distortion placed in each component by the encoder and attacker respectively. In order for encoder and attacker to meet their respective distortion constraints, $\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)$ and $\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)$, where

$$\mathcal{D}_m(\Delta) = \left\{ \boldsymbol{D} \in \mathbb{R}_+^m \ : \ \sum_{j=1}^m D_j \leq \Delta \right\}. \tag{5.3}$$

We also use a vector $\boldsymbol{A}$ to describe the variance of each stegotext component. Given the vector of stegotext variances $\boldsymbol{\sigma}^2$ and the vector of encoder distortion levels $\boldsymbol{D}_1$, the vector $\boldsymbol{A}$ must belong to (by the triangle inequality)

$$\mathcal{A}(\boldsymbol{D}_1, \boldsymbol{\sigma}) = \left\{ \boldsymbol{A} : (\sigma_j - \sqrt{D_{1j}})^2 \leq A_j \leq (\sigma_j + \sqrt{D_{1j}})^2, \ 1 \leq j \leq m \right\}. \tag{5.4}$$

Conversely, given $\boldsymbol{A}$ and $\boldsymbol{\sigma}$, the vector $\boldsymbol{D}_1$ must belong to

$$\mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma}) = \left\{ \boldsymbol{D}_1 : (\sigma_j - \sqrt{A_j})^2 \leq D_{1j} \leq (\sigma_j + \sqrt{A_j})^2, \ 1 \leq j \leq m \right\}. \tag{5.5}$$

Given $\boldsymbol{\sigma}$, $\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)$, $\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)$, and $\boldsymbol{A} \in \mathcal{A}(\boldsymbol{D}_1, \boldsymbol{\sigma})$, let us define

$$r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) = \sum_{j=1}^m \frac{1}{2} \log \left(1 + \left| s(A_j; D_{1j}, D_{2j}, \sigma_j^2) \right|^+ \right), \tag{5.6}$$

where $s(A; D_1, D_2, \sigma^2)$ is defined in (A.5). For the scalar Gaussian watermarking (SGWM), we proved in Chapter 4 that $\frac{1}{2}\log(1 + s(A; D_1, D_2, \sigma^2))$ is the maximum guaranteed achievable rate when the distortion levels are D_1 and D_2 and the covertext and stegotext variance are σ^2 and A, respectively. We will see here that $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ describes the maximum guaranteed achievable rate for VGWM when the distortion levels in the components are $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ and the variances of the covertext and stegotext components are $\boldsymbol{\sigma}^2$ and $\boldsymbol{A}$, respectively.

The definitions we have introduced so far are sufficient to analyze the private version of the VGWM game. However, the public version requires some additional definitions since we will describe an encoder that explicitly estimates the amount of distortion that the attacker will use in each component. We will denote this estimate by the non-negative m-vector $\tilde{\boldsymbol{D}}_2$.

Let us further define

$$v(A, D_1, \tilde{D}_2, \sigma^2) = \alpha^2(A; D_1, \tilde{D}_2, \sigma^2)\sigma^2 + 2\alpha(A; D_1, \tilde{D}_2, \sigma^2)\rho(A; D_1, \sigma^2) + D_1, \tag{5.7}$$

where α and ρ are defined in Appendix A. The function v will be used to describe the variances of entries in the codebook; compare v with σ_v^2 of (4.31). Let us next define

$$r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}) = \sum_{j=1}^{m} \frac{1}{2} \log \left(\frac{v(A_j, D_{1j}, \tilde{D}_{2j}, \sigma_j^2)}{b_2(A_j; D_{1j}, \sigma_j^2)} \right), \tag{5.8}$$

where $b_2(\cdot;\cdot,\cdot)$ is defined in (A.3) and each summand is non-negative as long as $0 \leq \tilde{D}_{2j} \leq A_j$. Next, let

$$r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) = \sum_{j=1}^{m} \frac{1}{2} \log^+ \left(\frac{A_j + (\tilde{D}_{2j})^2 G(A_j, D_{1j}, \sigma_j^2)}{D_{2j} + (\tilde{D}_{2j})^2 G(A_j, D_{1j}, \sigma_j^2)} \right), \tag{5.9}$$

where

$$G(A, D_1, \sigma^2) = \frac{1}{b_2(A; D_1, \sigma^2)} - \frac{1}{A}. \tag{5.10}$$

Here, 2^{nr_0} will represent the minimum number of codewords needed in each bin and 2^{nr_1} will represent the maximum number of total codewords; compare to R_0 and R_1 of Section 4.3. The final definition necessary for the public version is

$$\tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) = r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) - r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}), \tag{5.11}$$

which will describe the maximum guaranteed achievable rate when the component variances and distortions are given by the arguments of $\tilde{r}$.

If $\tilde{\boldsymbol{D}}_2 = \boldsymbol{D}_2 < \boldsymbol{A}$, then we can relate (5.6) and (5.11) by

$$\tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) = r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.12}$$

Thus, if the encoder's estimate of the attacker's distortion distribution is accurate, then the achievable rates should be the same for both private and public versions. Indeed, we will see that this is the case.

5.3 Outline of Proof

We recall that our objective in this chapter is to show that the capacity for both versions of the VGWM game is given by

$$\max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \sum_{j=1}^{m} C^*(D_{1j}, D_{2j}, \sigma_j^2), \tag{5.13}$$

where $C^*(D_1, D_2, \sigma^2)$, defined in (A.8), is the capacity of both versions of the SGWM game. We show that (5.13) is the capacity for the VGWM game in the following steps. We give lower and upper bounds on the capacity for the private version in Lemmas 5.1 and 5.2, respectively. We then show that these bounds coincide and are equal to the right hand side (RHS) of (5.13) in Lemma 5.3. Finally, in Lemma 5.4, we give a lower bound on the capacity for the public version that is also an upper bound on the capacity of the for the private version. Since the capacity of the public version cannot exceed the capacity of the private version, this line of argument completes the proof of Theorem 2.4. Several of the proof techniques are borrowed from [HN88] on the vector Gaussian arbitrarily varying channel, which is to the Gaussian arbitrarily varying channel (see Section 2.5.3) what the VGWM game is to the SGWM game.

Lemma 5.1.

$$C_{\text{priv}}^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}) \geq \max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.14}$$

We prove this lemma in Section 5.4. The basic idea is that given any feasible $\boldsymbol{A}$ and $\boldsymbol{D}_1$, an encoder/decoder pair can be designed so that the message can be reliably recovered regardless of the attacker strategy (which can be described by some feasible $\boldsymbol{D}_2$) for any rate less than $\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$. The encoder/decoder pair that we use is essentially a parallel concatenation of encoder/decoder pairs for the private version of the SGWM game. Note that the choice of $\boldsymbol{A}$ is constrained in that $\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$ must be non-empty.

Our second lemma of this section gives an upper bound on the capacity of the private version.

Lemma 5.2.

$$C_{\text{priv}}^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}) \leq \max_{\boldsymbol{A}} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.15}$$

In order to prove this lemma, we consider an attacker that chooses its distortion distribution $\boldsymbol{D}_2$ based on the empirical variances of the stegotext components $\boldsymbol{A}$. This choice of $\boldsymbol{D}_2$ will correspond to the minimum (as a function of $\boldsymbol{A}$ and Δ_1) on the RHS of (5.15). Note that although the attacker knows the total distortion Δ_1 that the encoder can use, the attacker cannot calculate the distortion distribution $\boldsymbol{D}_1$ that has been used to produce the stegotext. The attacker implements an optimal attack for the SGWM game for each component based on the vector $\boldsymbol{D}_2$. Thus, even if the encoder knows how the attacker will choose $\boldsymbol{D}_2$ based on $\boldsymbol{A}$, the maximum achievable rate is described by the RHS of (5.15). A more detailed proof of this lemma is similar to the lengthy converse of the SGWM game in Section 4.5 and is omitted.

Our next lemma shows that the lower and upper bounds of Lemma 5.1 and Lemma 5.2 coincide and are equal to the proposed capacity.

Lemma 5.3. *The following expressions are equal:*

1. *(5.13)*
2. *The RHS of (5.14)*
3. *The RHS of (5.15)*

We prove this lemma in Section 5.6. The key steps in proving the equality of the three expressions are two applications of the Sion-Kakutani Minimax theorem. The combination of these three lemmas implies that the capacity of the private version is given by any of the three expressions, and in particular (5.13).

Our final lemma gives a lower bound on the capacity for the public version that is in turn an upper bound on the capacity of the private version.

Lemma 5.4.

$$\begin{aligned} & C_{\text{pub}}^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}) \\ & \quad \geq \max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})} \max_{0 \leq \tilde{\boldsymbol{D}}_2 \leq \boldsymbol{A}} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) && (5.16) \\ & \quad \geq C_{\text{priv}}^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}). && (5.17) \end{aligned}$$

We prove (5.16) in Section 5.5 and (5.17) in Section 5.6.2. We now briefly discuss the proof of the two inequalities. To prove the first inequality, an encoder/decoder pair is designed using any feasible $\boldsymbol{D}_1$, $\boldsymbol{A}$ and $\tilde{\boldsymbol{D}}_2$ so that the message can be reliably recovered regardless of the attacker strategy (which can be described by some feasible $\boldsymbol{D}_2$) for any rate less that $\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma})$. As in the proof of Lemma 5.1, we essentially use a parallel concatenation of encoder/decoder pairs from the SGWM game. The encoder for each component is designed using an estimate of the attacker's distortion in that component. The vector $\tilde{\boldsymbol{D}}_2$ contains these estimates for all of the components. This estimate is necessary because although the encoder knows the total distortion allowed to the attacker, the encoder does not know how the attacker will distribute that distortion to the components. In order to prove the second inequality (5.17), we choose $\tilde{\boldsymbol{D}}_2$ depending on $\boldsymbol{A}$, $\boldsymbol{D}_1$ and $\boldsymbol{\sigma}$ to be the minimizing $\boldsymbol{D}_2$ on the RHS of (5.14). The resulting minimizing $\boldsymbol{D}_2$ in (5.17) is equal to our $\tilde{\boldsymbol{D}}_2$, and we can use (5.12) to prove the desired inequality.

Since the capacity of the public version cannot exceed the capacity of the private version, this lemma implies that the capacities of the two versions are equal, and, combined with the earlier lemmas, completes the proof of Theorem 2.4.

5.4 Achievability for the Private Version

In this section, we will prove Lemma 5.1. That is, we will show that any rate less the RHS of (5.14) is achievable in the private version of the vector Gaussian watermarking game.

Codebook generation: The encoder and decoder first choose a vector of stegotext powers $\boldsymbol{A}$. Second, they choose a distortion distribution vector $\boldsymbol{D}_1$ that is in the interior[1] of $\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$ so that $\sum_{j=1}^m D_{1j} < \Delta_1$. They then choose a rate R and a positive constant ϵ. They next use their shared secret key Θ_1 to generate an IID sequence of codewords $(\boldsymbol{S}(1), \ldots, \boldsymbol{S}(2^{nR}))$, where for each w, $\boldsymbol{S}(w)$ is a $n \times m$ random matrix with independent elements so that $S_{ij}(w)$ is a zero-mean variance-b_{2j} Gaussian random variable, where $b_{2j} = b_2(A_j; D_{1j}, \sigma_j^2)$ as defined in (A.3).

[1] The interior of a set $\mathcal{S}$ is the union of all open sets contained in $\mathcal{S}$ and will be denoted by $\mathrm{Int}(\mathcal{S})$.

Encoding: Given the covertext $\boldsymbol{u}$, the message w, and the codebook, the encoder creates the stegotext $\boldsymbol{x}$ as

$$\boldsymbol{x}_j(w) = b_{1j}\boldsymbol{u}_j + \boldsymbol{s}_j(w), \tag{5.18}$$

for all $1 \leq j \leq m$, where $b_{1j} = b_1(A_j; D_{1j}, \sigma_j^2)$ as defined in (A.2). This is essentially a parallel concatenation of the encoders for the scalar Gaussian watermarking game. We show below that this encoder yields a small probability of error. Furthermore, the expected distortion induced by this encoding rule is given by $\sum_{j=1}^m D_{1j}$, and we have chosen $\boldsymbol{D}_1$ so that this quantity is strictly less than Δ_1. Thus, by choosing the blocklength n large enough, we can make the probability of excess distortion as small as desired. Thus, since we will show below that this encoder yields a small probability of error, we can create a modified encoder that both meets the almost sure distortion constraint (2.1) and ensures reliable decoding of the message.

Decoding: The decoder uses the scoring function

$$\pi(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{D}_2) = r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) + \frac{1}{n}\sum_{i=1}^n \sum_{j:A_j > D_{2j}} \left(\frac{(y_{ij} - c_j b_{1j} u_{ij})^2}{2c_j(c_j b_{2j} + D_{2j})} - \frac{(y_{ij} - c_j x_{ij})^2}{2c_j D_{2j}} \right), \tag{5.19}$$

where $c_j = c(A_j; D_{2j})$ as defined in (A.4), and only considers attacker distortion vectors $\boldsymbol{D}_2$ in the set

$$\mathcal{D}_m^{(n)}(\Delta_2) = \left\{ \boldsymbol{D}_2 \in \mathbb{R}^m : \boldsymbol{D}_2 > 0, \sum_{j=1}^m D_{2j} < \Delta_2 \left(1 + \frac{m}{n}\right), \frac{n\boldsymbol{D}_2}{\Delta_2} \in \mathbb{Z}^m \right\}. \tag{5.20}$$

Note that the cardinality of $\mathcal{D}_m^{(n)}(\Delta_2)$ is at most $(n+m)^m$. Given the covertext $\boldsymbol{u}$, the forgery $\boldsymbol{y}$, and the codebook, the decoder declares message w was sent if

$$\pi(\boldsymbol{u}, \boldsymbol{x}(w), \boldsymbol{y}, \boldsymbol{D}_2) > r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \epsilon, \text{ for some } \boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2). \tag{5.21}$$

If no w or more than one w satisfies (5.21), then an error is declared.

Probability of error: Let $\mathcal{E}_1$ be the event that some incorrect w' satisfies (5.21), and let $\mathcal{E}_2$ be the event where the correct w does not satisfy (5.21). An error occurs if and

only if $\mathcal{E}_1 \cup \mathcal{E}_2$ occurs, and thus the overall probability of error is at most the sum of the probabilities of the events $\mathcal{E}_1$ and $\mathcal{E}_2$. We analyze these probabilities using the following series of lemmas.

In the following lemma, we establish an upper bound on the rate R such that $\Pr(\mathcal{E}_1)$ can be made as small as desired with a proper choice of ϵ and the blocklength n.

Lemma 5.5. *For any $n \times m$ matrices $\boldsymbol{u}$ and $\boldsymbol{y}$ and length-m vector $\boldsymbol{D}_2 > 0$, if $\boldsymbol{X}_j = b_{1j}\boldsymbol{u}_j + \boldsymbol{S}_j(w)$, for all j and any w, then $\Pr\left(\pi(\boldsymbol{u}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{D}_2) > \eta\right) \leq 2^{-n\eta}$.*

Proof. The proof follows as in [HN88]. For simplicity, we assume that all logarithms and exponentials are with respect to e. Let $\xi_{ij} = y_{ij} - c_j b_{1j} u_{ij}$. We shall need the following expectation,

$$\begin{aligned} E\left[\exp\left(-\frac{(S_{ij} - \xi_{ij}/c_j)^2}{2D_{2j}/c_j}\right)\right] &= \frac{1}{\sqrt{1 + b_{2j}c_j/D_{2j}}} \exp\left(\frac{-(c_j/2D_j)(\xi_{ij}/c_j)^2}{1 + b_{2j}c_j/D_{2j}}\right) \\ &= \frac{1}{\sqrt{1 + s(A_j; D_{1j}, D_{2j}, \sigma_j^2)}} \exp\left(\frac{-\xi_{ij}^2}{2c_j(c_j b_{2j} + D_{2j})}\right), \end{aligned}$$

which follows since if Z is a Gaussian with mean μ and variance σ^2, then

$$E[\exp(\beta Z^2)] = \frac{\exp\bigl(\beta\mu^2/(1 - 2\beta\sigma^2)\bigr)}{\sqrt{1 - 2\beta\sigma^2}}, \tag{5.22}$$

for $\beta < (2\sigma^2)^{-1}$. Thus,

$$\begin{aligned} &\Pr\bigl(\pi(\boldsymbol{u}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{D}_2) > \eta\bigr) \\ &\quad\leq E\left[\exp\left(n(\pi(\boldsymbol{u}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{D}_2) - \eta)\right)\right] \\ &\quad= e^{(n(r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \eta))} \prod_{i=1}^{n} \prod_{j: A_j > D_{2j}} E\left[\exp\left(\frac{\xi_{ij}^2}{2c_j(c_j b_{2j} + D_{2j})} - \frac{(S_{ij} - \xi_{ij}/c_j)^2}{2D_{2j}/c_j}\right)\right] \\ &\quad= e^{(n(r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \eta))} \prod_{i=1}^{n} \prod_{j: A_j > D_{2j}} \frac{1}{\sqrt{1 + s(A_j; D_{1j}, D_{2j}, \sigma_j^2)}} \\ &\quad= e^{-n\eta}, \end{aligned}$$

where the inequality follows since the exponential is at least one when the condition is true and at least zero when the condition is false. The equalities follow by the above computation and by the definition of $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$; see (5.6). □

We can now upper bound the probability of the first error event. To do so, let B_1 be a $m \times m$ diagonal matrix with jth diagonal element equal to b_{1j}. Then,

$$\begin{aligned}\Pr(\mathcal{E}_1) &\leq \sum_{w' \neq W} \sum_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} \Pr\left\{\pi\big(\boldsymbol{U}, \boldsymbol{U}B_1 + \boldsymbol{S}(w'), \boldsymbol{Y}, \boldsymbol{D}_2\big) > r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \epsilon\right\} \\ &\leq (n+m)^m \exp\left(n\left(R - \min_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \epsilon\right)\right).\end{aligned}$$

Here, the first inequality follows by the definition of the decoder and by the union of events bound. The second inequality follows by Lemma 5.5 since for any incorrect message w', the codeword $\boldsymbol{S}(w')$ is independent of the covertext $\boldsymbol{U}$ and the forgery $\boldsymbol{Y}$. Also, recall that $\mathcal{D}_m^{(n)}(\Delta_2)$ has at most $(n+m)^m$ elements and that there are 2^{nR} total messages. Since ϵ can be chosen arbitrarily by the encoder, the probability the some incorrect w' satisfies (5.21) can be made to tend to zero as long as

$$\begin{aligned}R &< \limsup_{n\to\infty} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) &(5.23)\\ &= \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). &(5.24)\end{aligned}$$

To prove the equality, let the right-hand side (RHS) of (5.24) be denoted by $\bar{r}(\Delta_2)$, which is continuous in Δ_2. First, the RHS of (5.23) is at most $\limsup_{n\to\infty} \bar{r}(\Delta_2(n+m)/n)$, which equals $\bar{r}(\Delta_2)$ by the continuity of $\bar{r}$. Second, let $\boldsymbol{D}_2^* \in \mathcal{D}_m(\Delta_2)$ denote the minimizing vector in (5.24), i.e. $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{\sigma}) = \bar{r}(\Delta_2)$. There exists a sequence of vectors $\boldsymbol{D}_{2,1}, \ldots$ such that $\boldsymbol{D}_{2,n} \in \mathcal{D}_m^{(n)}(\Delta_2)$ and $\boldsymbol{D}_{2,n} \to \boldsymbol{D}_2^*$ pointwise. The RHS of (5.23) is at least $\limsup_{n\to\infty} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_{2,n}, \boldsymbol{\sigma})$, which also equals $\bar{r}(\Delta_2)$ by the continuity of r in $\boldsymbol{D}_2$.

We now show that the probability of the second error event, $\Pr(\mathcal{E}_2)$, tends to zero as the blocklength n tends to infinity regardless of the choice of ϵ and R. To do so, we establish conditions under which the scoring function is high (Lemma 5.6) and then show that these conditions are met with high probability (Lemmas 5.7 and 5.8). Lemmas 5.6 is proved in Appendix B.12, while we give a short proofs of Lemmas 5.7 and 5.8 below.

Lemma 5.6. *There exists a positive function $f(A, D_1, \sigma^2)$ such that if the n-vectors $\boldsymbol{u}$, $\boldsymbol{x}$,*

and $\boldsymbol{y}$*, and the scalars* D_2 *and* δ *satisfy*

$$\left|n^{-1}\|\boldsymbol{u}\|^2 - \sigma^2\right| < \delta, \tag{5.25}$$

$$\left|n^{-1}\langle \boldsymbol{u}, \boldsymbol{x} - b_1\boldsymbol{u}\rangle\right| < \delta, \tag{5.26}$$

$$\left|n^{-1}\|\boldsymbol{x} - b_1\boldsymbol{u}\|^2 - b_2\right| < \delta \tag{5.27}$$

$$\left|n^{-1}\langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{y}|_{\boldsymbol{x}}\rangle\right| < \delta, \tag{5.28}$$

$$n^{-1}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq D_2 < A, \tag{5.29}$$

$$\delta < \frac{A}{2(1+b_1)^2}, \tag{5.30}$$

then

$$\frac{n^{-1}\|\boldsymbol{y} - cb_1\boldsymbol{u}\|^2}{2c(cb_2 + D_2)} - \frac{n^{-1}\|\boldsymbol{y} - c\boldsymbol{x}\|^2}{2cD_2} > -\delta f(A, D_1, \sigma^2), \tag{5.31}$$

where $b_1 = b_1(A; D_1, \sigma^2)$*,* $b_2 = b_2(A; D_1, \sigma^2)$ *and* $c = c(A; D_2)$*.*

Lemma 5.7. *The random variables* $n^{-1}\langle \boldsymbol{U}_j, \boldsymbol{Y}_j - \boldsymbol{Y}_j|_{\boldsymbol{X}_j}\rangle$ *converge to zero in probability as* n *tends to infinity, for all* $1 \leq j \leq m$*.*

Proof. The jth component of the covertext $\boldsymbol{U}_j$ is an IID sequence of mean-zero variance-σ_j^2 Gaussian random variables. The jth component of the stegotext is generated from the covertext as

$$\boldsymbol{X}_j = b_{1j}\boldsymbol{U}_j + \boldsymbol{S}_j,$$

where $\boldsymbol{S}_j$ is an IID sequence of mean-zero variance-b_{2j} Gaussian random variables that is further independent of $\boldsymbol{U}_j$. Since $\boldsymbol{X}_j$ and $\boldsymbol{U}_j$ are jointly Gaussian random vectors, we can also write their relationship as

$$\boldsymbol{U}_j = \frac{b_{1j}\sigma_j^2}{A_j}\boldsymbol{X}_j + \boldsymbol{T}_j,$$

where $\boldsymbol{T}_j$ is an IID sequence of mean-zero variance-$b_{2j}\sigma_j^2/A_j$ Gaussian random variables that is further independent of $\boldsymbol{X}_j$. The random vector $\boldsymbol{T}_j$ is also independent of $\boldsymbol{Y}_j$ since $\boldsymbol{U} \multimap \boldsymbol{X} \multimap \boldsymbol{Y}$. For every realization of $\boldsymbol{x}_j$ and $\boldsymbol{y}_j$, the vectors $\boldsymbol{y}_j - \boldsymbol{y}_j|_{\boldsymbol{x}_j}$ and $\boldsymbol{x}_j$ are

perpendicular. Thus, in distribution,

$$n^{-1}\langle \boldsymbol{U}_j, \boldsymbol{Y}_j - \boldsymbol{Y}_j|_{\boldsymbol{X}_j}\rangle = n^{-1}\langle \boldsymbol{T}_j, \boldsymbol{Y}_j - \boldsymbol{Y}_j|_{\boldsymbol{X}_j}\rangle.$$

The proof is completed by recalling that $\boldsymbol{T}_j$ is independent of $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ and from the fact that $\boldsymbol{Y}_j - \boldsymbol{Y}_j|_{\boldsymbol{X}_j}$ has bounded norm. □

Lemma 5.8. *If $n \times m$ matrices $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy*

$$\sum_{j=1}^{m} \frac{1}{n}\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2 \le \Delta_2,$$

then there exists a $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$ such that for all $1 \le j \le m$,

$$\frac{1}{n}\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2 \le D_{2j}'. \tag{5.32}$$

Proof. For every j, let D_{2j}' be the smallest positive integer multiple of Δ_2/n that is at least $n^{-1}\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2$. This $\boldsymbol{D}_2'$ satisfies (5.32), and thus we only have to show that $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$, which is defined in (5.20). Our choice of $\boldsymbol{D}_2'$ is positive and satisfies $n\boldsymbol{D}_2'/\Delta_2 \in \mathbb{Z}^m$. Furthermore, it satisfies $D_{2j}' - \Delta_2/n \le n^{-1}\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2$, and thus

$$\begin{aligned}\sum_{j=1}^{m} D_{2j}' &\le \frac{m\Delta_2}{n} + \sum_{j=1}^{m}\frac{1}{n}\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2 \\ &\le \Delta_2\left(1 + \frac{m}{n}\right).\end{aligned}$$

Thus, $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$. □

The above lemmas allow us to analyze $\Pr(\mathcal{E}_2)$. To do so, let

$$\delta_j = \frac{\epsilon}{m f(A_j, D_{1j}, \sigma_j^2)}, \tag{5.33}$$

where $f(\cdot,\cdot,\cdot)$ is the function defined in Lemma 5.6. Given the covertext $\boldsymbol{U} = \boldsymbol{u}$ and the forgery $\boldsymbol{Y} = \boldsymbol{y}$, the correct message w will be selected by the decoder if (but not only if) there exists a $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$ such that $\boldsymbol{u}_j$, $\boldsymbol{x}_j(w)$ and $\boldsymbol{y}_j$ satisfy the requirements of Lemma 5.6 with $D_2 = D_{2j}'$ and $\delta = \delta_j$ (along with $A = A_j$, $D_1 = D_{1j}$ and $\sigma^2 = \sigma_j^2$) for every j such that $D_{2j}' < A_j$. This follows since in this case, $\pi(\boldsymbol{u}, \boldsymbol{x}(w), \boldsymbol{y}, \boldsymbol{D}_2') > r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2', \boldsymbol{\sigma}) - \epsilon$;

compare (5.19), (5.31) and (5.33).

We now show that the above claim holds with probability tending to one as the block-length n tends to infinity. We first note that by the definition of the encoder, all of (5.25), (5.26) and (5.27) will be satisfied with high probability for each component. Next, Lemma 5.7 demonstrates that (5.28) is also satisfied with high probability for each component. Finally, Lemma 5.8 shows that if the attacker satisfies the distortion constraint (2.3), which it is required to do with probability one, then there exists a $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$ that satisfies either (5.29) or $D_{2j}' > A_j$ for every component. Since the above condition is sufficient for reliable recovery of the message, we have shown that $\Pr(\mathcal{E}_2)$ can be made arbitrarily small for any positive R and ϵ.

Combining the analysis of $\mathcal{E}_1$ and $\mathcal{E}_2$, we see that the probability of error can be made arbitrarily small as long as

$$R < \max_{\boldsymbol{A}} \sup_{\boldsymbol{D}_1 \in \mathrm{Int}(\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma}))} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.34}$$

This follows from (5.24) since we can choose any $\boldsymbol{A}$ and any $\boldsymbol{D}_1 \in \mathrm{Int}(\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma}))$ to design the encoder and decoder. The proof of Lemma 5.1 is completed by noting that the RHS of (5.24) is continuous in $\boldsymbol{D}_1$.

5.5 Achievability for the Public Version

In this section, we will prove the first inequality (5.16) of Lemma 5.4. That is, we will show that any rate less than the RHS of (5.16) is achievable in the public version of the vector Gaussian watermarking game. This proof parallels the proof in Section 5.4 for the private version. We denote with a tilde ($\tilde{}$) many of the quantities used here that are different from the private version but play an analogous role.

5.5.1 Codebook Generation

The encoder and decoder jointly choose the parameters they will use. Namely, they choose a vector of stegotext powers $\boldsymbol{A}$, a distortion distribution vector $\boldsymbol{D}_1$, an estimate of the attacker's distortion distribution $\tilde{\boldsymbol{D}}_2$, two rates R and R_0, and two positive constants ϵ and ϵ_0. They next generate $2^{n(R+R_0)}$ independent random matrices as follows. For every

message $w \in \{1, \ldots, 2^{nR}\}$ and every index $w_0 \in \{1, \ldots, 2^{nR_0}\}$, they generate a $n \times m$ random matrix $\boldsymbol{S}(w, w_0)$ with independent elements such that $S_{ij}(w, w_0)$ is a mean-zero variance-v_j Gaussian random variable, where $v_j = v(A_j, D_{1j}, \tilde{D}_{2j}, \sigma_j^2)$; see (5.7). We can think of the codebook of consisting of 2^{nR} bins (indexed by w), where each bin contains 2^{nR_0} codewords (indexed by w_0).

5.5.2 Encoding

Given the covertext $\boldsymbol{u}$, the message w, and the codebook, the encoder searches for a codeword in bin w that is "jointly typical" with $\boldsymbol{u}$. That is, the encoder finds a w_0 such that

$$\left| n^{-1} \|\boldsymbol{s}_j(w, w_0)\|^2 - v_j \right| < \epsilon_0, \tag{5.35}$$

and

$$\left| n^{-1} \langle \boldsymbol{s}_j(w, w_0), \boldsymbol{u}_j \rangle - (\alpha_j - 1 + b_{1j})\sigma_j^2 \right| < \epsilon_0, \tag{5.36}$$

for all $1 \leq j \leq m$, where $\alpha_j = \alpha(A_j; D_{1j}, \tilde{D}_{2j}, \sigma_j^2)$ and $b_{1j} = b_1(A_j; D_{1j}, \sigma_j^2)$; see (A.6) and (A.3). If no such w_0 is found, then an encoding failure is declared. If there was not an encoding failure, then the encoder creates the stegotext as

$$\boldsymbol{x}_j = \boldsymbol{s}_j(w, w_0) + (1 - \alpha_j)\boldsymbol{u}_j, \tag{5.37}$$

for all $1 \leq j \leq m$. If there was an encoding failure, then the encoder simply sets the stegotext equal to the covertext. Let us further require that

$$\left| n^{-1} \|\boldsymbol{u}_j\|^2 - \sigma_j^2 \right| < \epsilon_0, \tag{5.38}$$

for all $1 \leq j \leq m$, which occurs with arbitrarily high probability for blocklength n large enough. Then, all of

$$\left| n^{-1} \|\boldsymbol{x}_j - \boldsymbol{u}_j\|^2 - D_{1j} \right| < \epsilon_0 (1 + |\alpha_j|)^2, \tag{5.39}$$

$$\left| n^{-1} \|\boldsymbol{x}_j\|^2 - A_j \right| < \epsilon_0 (2 - \alpha_j)^2, \tag{5.40}$$

and

$$\left|n^{-1}\langle \boldsymbol{s}_j(w,w_0),\boldsymbol{x}\rangle - (\alpha_j-1)b_{1j}\sigma_j^2 - A_j\right| < \epsilon_0(2-\alpha_j) \tag{5.41}$$

are true for all $1 \le j \le m$. Thus, from (5.39), we see that for any $\boldsymbol{D}_1$ that lies in the interior of $\mathcal{D}_m(\Delta_1)$, there exists an $\epsilon_0 > 0$ such that encoding success implies that the distortion constraint (2.1) is met. Also, (5.40) demonstrates that $\boldsymbol{A}$ describes the power of the covertext components.

5.5.3 Decoding

In order to describe the decoding procedure, we first define

$$\beta_1(A,D_1,\tilde{D}_2,D_2,\sigma^2) = \frac{A-D_2}{\sqrt{v(A,D_1,\tilde{D}_2,\sigma^2)\big(A+\tilde{D}_2^2 G(A,D_1,\sigma^2)\big)}}, \tag{5.42}$$

and

$$\beta_2(A,D_1,\tilde{D}_2,D_2,\sigma^2) = (A-D_2)\frac{D_2+\tilde{D}_2^2 G(A,D_1,\sigma^2)}{A+\tilde{D}_2^2 G(A,D_1,\sigma^2)}, \tag{5.43}$$

where v and G are defined in (5.7) and (5.10), respectively.

Given the forgery $\boldsymbol{y}$, the decoder evaluates a codeword matrix $\boldsymbol{s}$ and an attacker distortion distribution $\boldsymbol{D}_2$ using the scoring function

$$\tilde{\pi}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{D}_2) = r_1(\boldsymbol{A},\boldsymbol{D}_1,\tilde{\boldsymbol{D}}_2,\boldsymbol{D}_2,\boldsymbol{\sigma}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j:A_j>D_{2j}}\left(\frac{y_{ij}^2}{2(A_j-D_{2j})} - \frac{(y_{ij}-\beta_{1j}s_{ij})^2}{2\beta_{2j}}\right), \tag{5.44}$$

where $\beta_{1j} = \beta_1(A_j,D_{1j},\tilde{D}_{2j},D_{2j},\sigma_j^2)$, $\beta_{2j} = \beta_2(A_j,D_{1j},\tilde{D}_{2j},D_{2j},\sigma_j^2)$, and r_1 is defined in (5.9). Given the forgery $\boldsymbol{y}$ and the codebook, the decoder declares message w was sent if

$$\tilde{\pi}\big(\boldsymbol{s}(w,w_0),\boldsymbol{y},\boldsymbol{D}_2\big) > r_1(\boldsymbol{A},\boldsymbol{D}_1,\tilde{\boldsymbol{D}}_2,\boldsymbol{D}_2,\boldsymbol{\sigma}) - \epsilon,$$
$$\text{for some } \boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2) \text{ and } w_0 \in \{1,\ldots,2^{nR_0}\}, \tag{5.45}$$

where $\mathcal{D}_m^{(n)}(\Delta_2)$ is defined in (5.20). If no w or more than one w satisfies (5.45), then a decoding failure is declared.

5.5.4 Probability of Error

Let $\mathcal{E}_e$ be the event that an encoding failure occurs. Let $\mathcal{E}_{d0}$ be the event that a decoding failure occurs because no w satisfies (5.45). Let $\mathcal{E}_{d2}$ be the event that a decoding failure occurs because two or more w satisfy (5.45). The overall probability of error is at most the sum of the probabilities of these three events. We show below that the probability of each of these error events can be made to vanish as long as the rate R does not exceed the RHS of (5.16).

Error $\mathcal{E}_e$: Encoding Failure

We now establish a lower bound on R_0 such that a proper choice of ϵ_0 ensures that $\Pr(\mathcal{E}_e)$ tends to zero as the blocklength n tends to infinity.

Let $A_{\epsilon_0}^{(n)}$ be the set of all $(\boldsymbol{s}, \boldsymbol{u})$ pairs of $n \times m$ matrices that satisfy (5.35), (5.36) and (5.38) for all j, i.e., $A_{\epsilon_0}^{(n)}$ are the jointly typical pairs. Using the continuous joint AEP (see e.g. [CT91, Thms 8.6.1 & 9.2.2]), we can upper bound (for any w and w_0)

$$\Pr\Big\{\big(\boldsymbol{S}(w, w_0), \boldsymbol{U}\big) \in A_{\epsilon_0}^{(n)}\Big\} \geq (1-\epsilon_0)\exp\Big(-n\big(r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}) - 3\epsilon_0\big)\Big), \tag{5.46}$$

where r_0 is defined in (5.8). This follows since if S_{ij} and U_{ij} are zero-mean Gaussian random variables with variances v_j and σ_j^2, respectively, and with covariance $(\alpha_j - 1 + b_{1j})\sigma_j^2$, then

$$I(S_{i1}, \ldots, S_{im}; U_{i1}, \ldots, U_{im}) = r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma})$$

for all i. For any message w, the probability of an encoding failure can now be bounded as

$$\begin{aligned}
\Pr(\mathcal{E}_e) &= \Pr\Big(\big(\boldsymbol{S}(w, w_0), \boldsymbol{U}\big) \notin A_{\epsilon_0}^{(n)}, \text{ for all } w_0 \in \{1, \ldots, \exp(nR_0)\}\Big) \\
&\leq \left(1 - (1-\epsilon_0)\exp\Big(-n\big(r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}) - 3\epsilon_0\big)\Big)\right)^{\exp(nR_0)},
\end{aligned}$$

where the inequality follows by (5.46). By choosing ϵ_0 small enough, the encoder can make the probability of an encoding failure as small as desired as long as

$$R_0 > r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}). \tag{5.47}$$

Error $\mathcal{E}_{d2}$: Incorrect Message Satisfying (5.45)

Using a similar argument to Lemma 5.5, we establish an upper bound on $R + R_0$ such that a proper choice of ϵ ensures that $\Pr(\mathcal{E}_{d2})$ tends to zero as the blocklength n tends to infinity.

Lemma 5.9. *For any $n \times m$ matrix $\boldsymbol{y}$, any m-vector $\boldsymbol{D}_2 > 0$, any message w, and any index w_0,*

$$\Pr\left\{\tilde{\pi}\left(\boldsymbol{S}(w, w_0), \boldsymbol{y}, \boldsymbol{D}_2\right) > \eta\right\} \leq 2^{-n\eta}. \tag{5.48}$$

Proof. The proof follows as in Lemma 5.5, and is thus omitted. □

We can now upper bound the probability of $\mathcal{E}_{d2}$.

$$\begin{aligned}
\Pr(\mathcal{E}_{d2}) &\leq \sum_{w' \neq W} \sum_{w_0} \sum_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} \Pr\left\{\tilde{\pi}\left(\boldsymbol{S}(w', w_0), \boldsymbol{Y}, \boldsymbol{D}_2\right) > r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \epsilon\right\} \\
&\leq (n+m)^m \exp\left(n\left(R + R_0 - \min_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) - \epsilon\right)\right),
\end{aligned}$$

where the two inequalities follow by a similar argument to the one following the inequalities after Lemma 5.5. Since ϵ can be chosen arbitrarily by the encoder, the probability of the error event $\mathcal{E}_{d2}$ can be made to tend to zero as long as

$$\begin{aligned}
R + R_0 &< \limsup_{n \to \infty} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m^{(n)}(\Delta_2)} r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) \\
&= \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}),
\end{aligned} \tag{5.49}$$

where the equality follows by a similar argument to the one following (5.24).

Error $\mathcal{E}_{d0}$: Correct Message not Satisfying (5.45)

Using similar arguments to Lemmas 5.6 and 5.7, we show that the correct message will satisfy (5.45) with arbitrarily high probability as the blocklength tends to infinity. The following lemma is proved in Appendix B.13.

Lemma 5.10. *There exists a positive function $\tilde{f}(A, D_1, \tilde{D}_2, \sigma^2)$ such that if the n-vectors*

$\boldsymbol{s}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, and the scalars D_2 and δ satisfy

$$\left|n^{-1}\|\boldsymbol{s}\|^2 - v\right| < \delta, \tag{5.50}$$

$$\left|n^{-1}\|\boldsymbol{x}\|^2 - A\right| < \delta, \tag{5.51}$$

$$\left|n^{-1}\langle \boldsymbol{s}, \boldsymbol{x}\rangle - (\alpha - 1)b_1\sigma^2 - A\right| < \delta, \tag{5.52}$$

$$\left|n^{-1}\langle \boldsymbol{s}, \boldsymbol{y} - \boldsymbol{y}|_{\boldsymbol{x}}\rangle\right| < \delta, \tag{5.53}$$

$$n^{-1}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq D_2 < A, \tag{5.54}$$

$$\delta < \frac{A}{2}, \tag{5.55}$$

then

$$\frac{n^{-1}\|\boldsymbol{y}\|^2}{2(A - D_2)} - \frac{n^{-1}\|\boldsymbol{y} - \beta_1\boldsymbol{s}\|^2}{2\beta_2} > -\delta\tilde{f}(A, D_1, \tilde{D}_2, \sigma^2), \tag{5.56}$$

where all of the parameters are computed with respect to A, D_1, $\tilde{D}_2$, D_2 and σ^2, i.e., $\alpha = \alpha(A; D_1, \tilde{D}_2, \sigma^2)$, $b_1 = b_1(A; D_1, \sigma^2)$, $v = v(A, D_1, \tilde{D}_2, \sigma^2)$, $\beta_1 = \beta_1(A, D_1, \tilde{D}_2, D_2, \sigma^2)$, and $\beta_2 = \beta_2(A, D_1, \tilde{D}_2, D_2, \sigma^2)$.

The above lemma allows us to analyze $\Pr(\mathcal{E}_{d0})$. To do so, let

$$\tilde{\delta}_j = \frac{\epsilon}{m\tilde{f}(A_j, D_{1j}, \tilde{D}_{2j}, \sigma_j^2)}, \tag{5.57}$$

where $\tilde{f}(\cdot,\cdot,\cdot,\cdot)$ is the function defined in Lemma 5.10.

Given the forgery $\boldsymbol{Y} = \boldsymbol{y}$, the correct message w will be selected by the decoder if there exists a $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$, an index w_0, and a $n \times m$ matrix $\boldsymbol{x}$ such that $\boldsymbol{s}_j(w, w_0)$, $\boldsymbol{x}_j$, and $\boldsymbol{y}_j$ satisfy the requirements of Lemma 5.10 for all j (such that $A_j > D_{2j}$) with $D_2 = D_{2j}'$ and $\delta = \tilde{\delta}_j$ (along with $A = A_j$, $D_1 = D_{1j}$, $\tilde{D}_2 = \tilde{D}_{2j}$ and $\sigma^2 = \sigma_j^2$). This follows since in this case $\tilde{\pi}(\boldsymbol{s}(w, w_0), \boldsymbol{y}, \boldsymbol{D}_2') > r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2', \boldsymbol{\sigma}) - \epsilon$; compare (5.44), (5.56) and (5.57).

We now show that the above claim (with the actual stegotext $\boldsymbol{x}$ and the actual index w_0) holds with probability tending to one as the blocklength n tends to infinity. We first note that if there was not an encoding failure, then all of (5.50), (5.51) and (5.52) will be satisfied for all j for small enough ϵ_0; see (5.35), (5.39) and (5.41). Next, an analogous result to Lemma 5.7 (which we do not prove here) demonstrates that (5.53) is satisfied with high probability. Finally, Lemma 5.8 shows that if the attacker satisfies the distortion constraint

(2.3), which it is required to do with probability one, then there exists a $\boldsymbol{D}_2' \in \mathcal{D}_m^{(n)}(\Delta_2)$ that satisfies either (5.54) or $D_{2j}' > A_j$ for every component. Since the above condition is sufficient for reliable recovery of the message, we have shown that $\Pr(\mathcal{E}_2)$ can be made arbitrarily small for any positive R, R_0 and ϵ and small enough ϵ_0.

Overall Rate Restriction

We thus find that a rate is achievable if

$$\begin{aligned} R &< \min_{0 \le \boldsymbol{D}_2 \le \boldsymbol{A} \,:\, \boldsymbol{e}^t \boldsymbol{D}_2 \le D_2} r_1(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) - r_0(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{\sigma}) && (5.58) \\ &= \min_{0 \le \boldsymbol{D}_2 \le \boldsymbol{A} \,:\, \boldsymbol{e}^t \boldsymbol{D}_2 \le D_2} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}), && (5.59) \end{aligned}$$

where the inequality follows by (5.47) and (5.49) and the equality follows by the definition of $\tilde{r}$ (5.11). Thus, the following rates are achievable,

$$R < \max_{\boldsymbol{A}} \sup_{\boldsymbol{D}_1 \in \mathrm{Int}(\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma}))} \max_{0 \le \tilde{\boldsymbol{D}}_2 \le \boldsymbol{A}} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}), \qquad (5.60)$$

since the encoder is free to choose any feasible $\boldsymbol{A}$, $\boldsymbol{D}_1$ and $\tilde{\boldsymbol{D}}_2$. The proof of (5.16) is completed by noting that the RHS of (5.60) after the supremum is continuous in $\boldsymbol{D}_1$.

5.6 Optimization Results

In this section, we will prove Lemma 5.3 and the second inequality (5.17) of Lemma 5.4.

5.6.1 Proof of Lemma 5.3

In this section, we will use the Sion-Kakutani Minimax Theorem (see e.g. [SW70, Theorem 6.3.7]) to show the equivalence of (5.13) and the RHSs of (5.14) and (5.15). Recall that the Minimax Theorem states that if $\mathcal{S}_1$ and $\mathcal{S}_2$ are compact convex sets and $\Phi : \mathcal{S}_1 \times \mathcal{S}_2 \mapsto \mathbb{R}$ is a continuous function such that $\Phi(x_1, x_2)$ is concave in x_1 (for x_2 fixed) and convex in x_2 (for x_1 fixed), then

$$\max_{x_1 \in \mathcal{S}_1} \min_{x_2 \in \mathcal{S}_2} \Phi(x_1, x_2) = \min_{x_2 \in \mathcal{S}_2} \max_{x_1 \in \mathcal{S}_1} \Phi(x_1, x_2)$$

The equivalence of the RHSs of (5.14) and (5.15) follows directly from the Minimax

Theorem since $\mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$ and $\mathcal{D}_m(\Delta_2)$ are compact convex sets and $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is concave in $\boldsymbol{D}_1$ and convex in $\boldsymbol{D}_2$. The fact that $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is concave in $\boldsymbol{D}_1$ relies critically on the fact that $\boldsymbol{D}_1 \in \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$ since in this set $\frac{1}{2}\log(1 + s(A_j; D_{1j}, D_{2j}, \sigma_j^2))$ is concave for every j.

We now show the equality of (5.13) and the RHS of (5.14). By the definitions of C^* (A.8) and r (5.6), we can rewrite (5.13) as

$$\max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \max_{\boldsymbol{A} \in \mathcal{A}(\boldsymbol{D}_1, \boldsymbol{\sigma})} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.61}$$

We can also trivially rewrite the RHS of (5.14) as

$$\max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)} \max_{\boldsymbol{A} \in \mathcal{A}(\boldsymbol{D}_1, \boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.62}$$

Note that we cannot directly apply the Minimax Theorem to show the equivalence of (5.61) and (5.62) since $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is not concave in $\boldsymbol{A}$. (It is not even quasi-concave in $\boldsymbol{A}$ despite being quasi-concave in each A_j.) We require some manipulations before we can apply the Minimax Theorem. We can replace $\mathcal{A}(\boldsymbol{D}_1, \boldsymbol{\sigma})$ in both (5.61) and (5.62) with $\mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})$, where

$$\mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma}) = \left\{ \boldsymbol{A} : \sigma_j^2 + D_{1j} \le A_j \le (\sigma_j + \sqrt{D_{1j}})^2,\ 1 \le j \le m \right\};$$

compare with (5.4). See the proof of Lemma A.1 for why this is true. Thus, the following lemma demonstrates that (5.13) and the RHS of (5.14) are equal.

Lemma 5.11.

$$\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) = \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). \tag{5.63}$$

Proof. Clearly, the left hand side (LHS) of (5.63) is at least as large as the RHS of (5.63). Thus, we only need to show the opposite inequality. To do so, let us define

$$r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) = \sum_{j=1}^{m} \frac{1}{2} \log\left(1 + s(A_j; D_{1j}, D_{2j}, \sigma_j^2)\right), \tag{5.64}$$

which differs from the definition of $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ (see (5.6)) only in that the positive part

of s is not taken here. Note that $r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is well-defined since $s(A; D_1, D_2, \sigma^2) > -1$ for $\sigma^2 + D_1 \leq A \leq (\sigma + \sqrt{D_1})^2$. Also note that, unlike $r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$, the function $r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is concave in $\boldsymbol{A}$. We can thus compute that

$$\begin{aligned}
&\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) \\
&= \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) && (5.65) \\
&= \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) && (5.66) \\
&\leq \max_{\boldsymbol{A} \in \mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}). && (5.67)
\end{aligned}$$

Here, (5.65) follows since each A_j is maximized separately and the contribution from component j to r is at least as great as the contribution to r' with equality if the contributions are positive; (5.66) follows from the Minimax Theorem since $\mathcal{A}'(\boldsymbol{D}_1, \boldsymbol{\sigma})$ and $\mathcal{D}_m(\Delta_2)$ are compact convex sets and $r'(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})$ is a continuous function that is concave in $\boldsymbol{A}$ and convex in $\boldsymbol{D}_1$; and (5.67) follows since $r' \leq r$, compare (5.6) and (5.64). This completes the proof of the Lemma. □

5.6.2 Proof of (5.17)

In this section, we prove the second inequality (5.17) of Lemma 5.4. To do so, we specify a $\tilde{\boldsymbol{D}}_2$ so that, with this choice of $\tilde{\boldsymbol{D}}_2$, the RHS of (5.16) equals the RHS of (5.14), which by Lemma 5.3 is the capacity of the private version. Since we are maximizing over $\tilde{\boldsymbol{D}}_2$ on the RHS of (5.16), this demonstrates the desired inequality.

The following lemma, describes the vector $\boldsymbol{D}_2$ that achieves the minimum on the RHS of (5.14).

Lemma 5.12. *For fixed $\boldsymbol{\sigma}$, $\boldsymbol{A} > 0$ and $\boldsymbol{D}_1 \in \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$,*

$$\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma}) = r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{\sigma}), \tag{5.68}$$

where if $\Delta_2 < \sum_j A_j$, then

$$D_{2j}^* = \begin{cases} \Gamma(G_j, \lambda) & \text{if } \Gamma(G_j, \lambda) < A_j \\ A_j & \text{otherwise} \end{cases}, \tag{5.69}$$

where

$$\Gamma(G,\lambda) = \frac{-1+\sqrt{1+\frac{4G}{\lambda}}}{2G}, \tag{5.70}$$

and $G_j = G(A_j, D_{1j}, \sigma_j^2)$*; see (5.10). Furthermore,* $\lambda > 0$ *is chosen such that* $\sum_j D_{2j}^* = \Delta_2$*. If* $\Delta_2 \geq \sum_j A_j$*, then* $\boldsymbol{D}_2^* = \boldsymbol{A}$ *and the minimum in (5.68) is zero.*

Proof. The case where $\Delta_2 \geq \sum_j A_j$ is straightforward and thus we assume that $\Delta_2 < \sum_j A_j$. We can further restrict $\boldsymbol{D}_2 \leq \boldsymbol{A}$ (pointwise) since the contribution to r from component j is zero if $D_{2j} \geq A_j$. With this further restriction, the LHS of (5.68) is a convex program with differentiable objective function and constraints. Furthermore, the Slater constraint qualification is met, i.e., there exists a $\boldsymbol{D}_2$ such that $0 < \boldsymbol{D}_2 < \boldsymbol{A}$ and $\sum_{j=1}^m D_{2j} < \Delta_2$. Thus, the Kuhn-Tucker conditions are both necessary and sufficient for a vector to achieve the minimum (see e.g. [SW70, Theorem 6.6.5]). To that end, let

$$L(\boldsymbol{D}_2, \mu_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \frac{2r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{\sigma})}{\log e} + \mu_0(\boldsymbol{e}^t\boldsymbol{D}_2 - \Delta_2) + \boldsymbol{\mu}_1^t(\boldsymbol{D}_2 - \boldsymbol{A}) - \boldsymbol{\mu}_2^t\boldsymbol{D}_2, \tag{5.71}$$

where $\boldsymbol{e} = (1, \ldots, 1)$. We will show that there exists a $\boldsymbol{\mu}_1^* \geq 0$ with $\boldsymbol{\mu}_{1j}^* > 0$ only if $D_{2j}^* = A_j$ such that

$$\frac{\partial}{\partial D_{2j}} L(\boldsymbol{D}_2^*, \lambda, \boldsymbol{\mu}_1^*, 0) = 0, \tag{5.72}$$

for $1 \leq j \leq m$. Since the Lagrange multipliers are positive only for the tight constraints, the Kuhn-Tucker conditions are satisfied and $\boldsymbol{D}_2^*$ is the unique minimum. Indeed, for $D_{2j} \leq A_j$,

$$\begin{aligned}
\frac{\partial}{\partial D_{2j}} L(\boldsymbol{D}_2, \mu_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2) &= \frac{\frac{\partial}{\partial D_{2j}} s(A_j; D_{1j}, D_{2j}, \sigma_j^2)}{1 + s(A_j; D_{1j}, D_{2j}, \sigma_j^2)} + \mu_0 + \mu_{1j} - \mu_{2j} \\
&= \frac{-\frac{b_{2j}}{D_{2j}^2}}{1 + \frac{c_j b_{2j}}{D_{2j}}} + \mu_0 + \mu_{1j} - \mu_{2j} \\
&= \frac{-1}{D_{2j}^2 G_j + D_{2j}} + \mu_0 + \mu_{1j} - \mu_{2j}.
\end{aligned} \tag{5.73}$$

Observe that $(\Gamma(G_j,\lambda))^2 G_j + \Gamma(G_j,\lambda) = 1/\lambda$. Thus, when $\Gamma(G_j,\lambda) < A_j$, setting $\mu_{1j}^* = 0$

satisfies (5.72). Further, if $\Gamma(G_j, \lambda) > A_j$, set

$$\mu_{1j}^* = \frac{1}{A_j^2 G_j + A_j} - \lambda > 0, \tag{5.74}$$

which satisfies (5.72) (the inequality follows since the first term in (5.73) is decreasing in D_{2j}). We have thus verified (5.72), which completes the proof of the lemma. □

Our next lemma, shows that if $\tilde{\boldsymbol{D}}_2$ is chosen to be $\boldsymbol{D}_2^*$ of the previous lemma, then the minimizing $\boldsymbol{D}_2$ on the RHS of (5.17) is also $\boldsymbol{D}_2^*$.

Lemma 5.13. *For fixed $\boldsymbol{\sigma}$, $\boldsymbol{A} > 0$ and $\boldsymbol{D}_1 \in \mathcal{D}(\boldsymbol{A}, \boldsymbol{\sigma})$, let the vector $\boldsymbol{D}_2^*$ be as described in Lemma 5.12. Then,*

$$\min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{D}_2, \boldsymbol{\sigma}) = \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{D}_2^*, \boldsymbol{\sigma}). \tag{5.75}$$

Proof. The case where $\Delta_2 \geq \sum_j A_j$ is straightforward and thus we assume that $\Delta_2 < \sum_j A_j$. As in Lemma 5.12, this is a convex program where the Kuhn-Tucker conditions are both necessary and sufficient. Thus, let

$$\tilde{L}(\boldsymbol{D}_2, \mu_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \tilde{\boldsymbol{D}}_2) = \frac{2\tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma})}{\log e} + \mu_0(\boldsymbol{e}^t \boldsymbol{D}_2 - \Delta_2) + \boldsymbol{\mu}_1^t(\boldsymbol{D}_2 - \boldsymbol{A}) - \boldsymbol{\mu}_2^t \boldsymbol{D}_2. \tag{5.76}$$

We will show that for the $\lambda > 0$ and $\boldsymbol{\mu}_1^* \geq 0$ specified in Lemma 5.12,

$$\frac{\partial}{\partial D_{2j}} \tilde{L}(\boldsymbol{D}_2^*, \lambda, \boldsymbol{\mu}_1^*, 0; \boldsymbol{D}_2^*) = 0, \tag{5.77}$$

for all $1 \leq j \leq m$. This will complete the proof of the current lemma. Indeed, if $D_{2j} \leq A_j$, then

$$\frac{\partial}{\partial D_{2j}} \tilde{L}(\boldsymbol{D}_2, \mu_0, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \tilde{\boldsymbol{D}}_2) = \frac{-1}{\tilde{D}_{2j}^2 G_j + D_{2j}} + \mu_0 + \mu_{1j} - \mu_{2j}, \tag{5.78}$$

where $G_j = G(A_j, D_{1j}, \sigma_j^2)$ defined in (5.10). Compare (5.78) with (5.73). By substituting $\boldsymbol{D}_2 = \tilde{\boldsymbol{D}}_2 = \boldsymbol{D}_2^*$, $\mu_0 = \lambda$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1^*$, and $\boldsymbol{\mu}_2 = 0$ into (5.78) we verify that (5.77) is true. □

The combination of these lemmas completes the proof of (5.17) as follows,

$$\begin{aligned}
&\max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A},\boldsymbol{\sigma})} \max_{0 \le \tilde{\boldsymbol{D}}_2 \le \boldsymbol{A}} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \tilde{\boldsymbol{D}}_2, \boldsymbol{D}_2, \boldsymbol{\sigma}) \\
&\quad \ge \max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A},\boldsymbol{\sigma})} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(\Delta_2)} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{D}_2, \boldsymbol{\sigma}) && (5.79) \\
&\quad = \max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A},\boldsymbol{\sigma})} \tilde{r}(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{D}_2^*, \boldsymbol{\sigma}) && (5.80) \\
&\quad = \max_{\boldsymbol{A}} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1) \cap \mathcal{D}(\boldsymbol{A},\boldsymbol{\sigma})} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}_2^*, \boldsymbol{\sigma}) && (5.81) \\
&\quad = C_{\text{priv}}^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}). && (5.82)
\end{aligned}$$

Here, (5.79) follows since we are using a particular choice of $\tilde{\boldsymbol{D}}_2$ (namely, $\boldsymbol{D}_2^*$), (5.80) follows by Lemma 5.13, (5.81) follows by (5.12), and (5.82) follows by Lemmas 5.12 and 5.3.

5.7 The Optimal Attack and Lossy Compression

In this section, we compare the optimal attacker for the VGWM game to an attacker who implements optimal lossy compression based only on the statistics of the stegotext[2]. This comparison is of interest for two reasons. First, in the SGWM game (see Chapter 4.5), we found the two attackers to be essentially the same. Second, many watermarking systems are designed to be robust against compression attacks [FKK01]. Thus, we would like to see if our intuition for the SGWM game carries over to the VGWM game and if watermarking systems designed in the above manner are the best possible. We find that the answer to both of these questions is no.

5.7.1 Compression Attack

Optimal lossy compression with allowed distortion Δ_2 of a vector Gaussian stegotext with component variances given by $\boldsymbol{A}$ can be described as follows (see also [CT91]). The distortion is distributed to the components using the reverse waterfilling vector $\boldsymbol{D}^{\text{wf}}$, where

$$D_j^{\text{wf}} = \begin{cases} \lambda_0 & \text{if } \lambda_0 < A_j \\ A_j & \text{otherwise} \end{cases}, \qquad (5.83)$$

[2]For the purposes of this discussion, we assume that the statistics of the stegotext are Gaussian, which is approximately true for the optimal encoders described in Sections 5.5 and 5.4.

and where λ_0 is chosen such that $\sum_j D_j^{\text{wf}} = \Delta_2$. Then, optimal lossy compression is performed on each of the components using distortion D_j^{wf} for component j. We will optimal lossy compression of the stegotext the *compression attack*.

Similarly to the compression attack, the optimal attack in the VGWM game chooses a distortion distribution vector $\boldsymbol{D}_2$ and uses the optimal attack for the SGWM game on component j with distortion D_{2j}. Since the optimal attack for the SGWM game is essentially optimal lossy compression, the difference between the optimal attack and the compression attack lies in their distribution of their allowable distortion, i.e., $\boldsymbol{D}_2^*$ of (5.69) versus $\boldsymbol{D}^{\text{wf}}$ of (5.83). We see that the vector are different for an arbitrary choice of $\boldsymbol{A}$ and $\boldsymbol{D}_1$ by the encoder. In fact, the following example will demonstrate that $\boldsymbol{D}_2^* \neq \boldsymbol{D}^{\text{wf}}$ even for the optimal choice of $\boldsymbol{A}$ and $\boldsymbol{D}_1$.

We now consider an example where $m = 5$, $\boldsymbol{\sigma}^2 = [10 \;\; 8 \;\; 6 \;\; 4 \;\; 2]$, $\Delta_1 = 5$ and $\Delta_2 = 40$. For these parameters, the vectors for the optimal encoder and attack (i.e., the ones that solve the RHS of (5.14)) are given by

$$\begin{aligned}
\boldsymbol{D}_1^* &\approx [1.50 \;\; 1.38 \;\; 1.03 \;\; 0.70 \;\; 0.39]; \\
\boldsymbol{A}^* &\approx [15.50 \;\; 13.27 \;\; 10.29 \;\; 7.09 \;\; 3.89]; \\
\boldsymbol{D}_2^* &\approx [11.44 \;\; 10.36 \;\; 8.48 \;\; 6.11 \;\; 3.61].
\end{aligned}$$

For this example the capacity of the VGWM game, $C^{\text{VGWM}}(\Delta_1, \Delta_2, \boldsymbol{\sigma}) = r(\boldsymbol{A}^*, \boldsymbol{D}_1^*, \boldsymbol{D}_2^*, \boldsymbol{\sigma})$, is approximately 0.048 bits/vector. We see that

$$\boldsymbol{D}^{\text{wf}}(\boldsymbol{A}^*) \approx [9.673 \;\; 9.673 \;\; 9.673 \;\; 7.09 \;\; 3.89] \neq \boldsymbol{D}_2^*,$$

and indeed the compression attack is not the same as the optimal attack. The maximum achievable rate for the encoder defined by $\boldsymbol{D}_1^*$ and $\boldsymbol{A}^*$ and the corresponding compression attack defined by $\boldsymbol{D}^{\text{wf}}$ is given by $r(\boldsymbol{A}^*, \boldsymbol{D}_1^*, \boldsymbol{D}^{\text{wf}}, \boldsymbol{\sigma}) \approx 0.051$ bits/vector. Given that the encoder also has to be robust against a general attack, the effect of using the suboptimal compression attack is significant, but not very large. However, we will see in the next section that if the watermarking system is designed for the compression attack, then the gain in achievable rate is quite large.

5.7.2 Designing for a Compression Attack

Since the compression attack is not optimal, the watermarking system can send more bits if it knows it only has to protect against such an attack. In fact, when a system is only required to be robust to compression, it can send many more bits using a qualitatively different strategy. Before we return to the example of the previous section, note that all rates less than

$$\max_{\boldsymbol{D}_1 \in \mathcal{D}_m(\Delta_1)} \max_{\boldsymbol{A} \in \mathcal{A}(\boldsymbol{D}_1, \sigma)} r(\boldsymbol{A}, \boldsymbol{D}_1, \boldsymbol{D}^{\mathrm{wf}}(\boldsymbol{A}), \boldsymbol{\sigma}) \tag{5.84}$$

are achievable against a compression attack. This follows since such an attack uses distortion D_j^{wf} in component j and the encoder can thus reliably send at rates less than $\frac{1}{2}\log(1 + s(A_j; D_{1j}, D_j^{\mathrm{wf}}, \sigma_j^2))$ in component j alone. We denote the maximizing vectors in (5.84) by $\boldsymbol{D}_1^{**}$ and $\boldsymbol{A}^{**}$. For the example in the previous section

$$\boldsymbol{D}_1^{**} \approx [3.395 \;\; 0.035 \;\; 0.32 \;\; 0.83 \;\; 0.42],$$
$$\boldsymbol{A}^{**} \approx [17.28 \;\; 9.09 \;\; 9.09 \;\; 8.47 \;\; 4.25],$$

and the corresponding waterfilling vector is given by

$$\boldsymbol{D}^{\mathrm{wf}} \approx [9.10 \;\; 9.09 \;\; 9.09 \;\; 8.47 \;\; 4.25].$$

The maximum achievable rate in this scenario, (5.84), is approximately 0.105 bits/vector, which is more than double the capacity of the VGWM game. In Figure 5-1, we compare the parameters for the optimal attack and for the compression attack.

We now consider some of the qualitative differences between the two attacks. For the optimal attack, the encoder uses all of the components to transmit information, although the number of bits in a component is correlated with the variance of the components. The encoder that is designed for the compression attack is very different. This encoder places some distortion in components 2–5, but only to boost the variance of the stegotext and not to send any information. The waterfilling attacker is wasting distortion on these components (in fact, almost 80%), and thus this attacker is far from optimal for this particular encoder. Furthermore, the encoder designed for the compression attack only transmits information

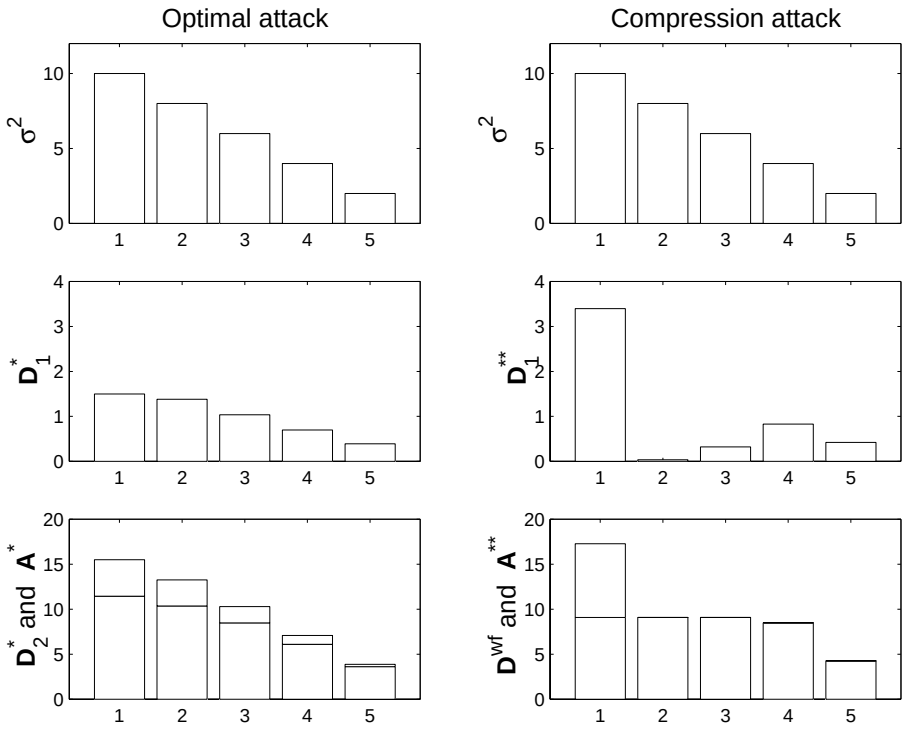

Figure 5-1: Comparison of watermarking system parameters for the optimal attack and the compression attack.

in the component with the highest variance, but it is able to send at a relatively high rate in this component since it devotes most of its power to this component and the attacker does not have much distortion left for this component. A more clever attacker would put all of its distortion into component 1, and no positive rates would be achievable. This example is somewhat extreme in that the ratio of Δ_2 to Δ_1 is quite high. However, in almost all regimes, the encoder designed for the compression attack will use some components as decoys and the rest for actually transmitting information.

Chapter 6

Watermarking with Discrete Alphabets

In this chapter, we consider two examples of the watermarking game when the alphabets (i.e., the sets $\mathcal{U}$, $\mathcal{X}$ and $\mathcal{Y}$) are finite. This contrasts with most of the rest of the thesis, where we have been assuming that all the alphabets are the real line. With finite alphabets, we are able to use combinatorial techniques that are not applicable in our other scenarios.

In Section 6.1, we consider a general watermarking game with the major exception that there is no covertext in which to hide the message. In Section 6.2, we assume that all of the alphabets are binary and that the covertext is an IID sequence of Bernoulli(1/2) random variables. These simple examples should provide some insight into how to approach a more general watermarking model with discrete alphabets.

Note that Somekh-Baruch and Merhav [SBM01a] have recently described the capacity of the private version of the watermarking game for finite alphabets and a general discrete memoryless source, which is more general than our proof in Section 6.1 on watermarking with no covertext. Also, Barron, Chen and Wornell [BCW00] have recently shown that our proposed capacity expressions for the binary watermarking game are also the capacity expressions for related fixed attack binary watermarking problems, which we could use as a converse in Section 6.2. Nevertheless, we are including the full proofs of our results due to their simplicity and their illustrative nature.

6.1 No Covertext

In this section, we give a proof of Theorem 2.5. Recall that this theorem states that the capacity of the watermarking game when there is no covertext is given by $C^{\mathrm{NoCov}}(D_1, D_2)$, which is defined in (2.19) and recalled below in (6.1). Further, recall that since there is no covertext, the distortion constraint on the encoder (2.1) is replaced by $n^{-1}\sum_{i=1}^{n} d_1(X_i) \leq D_1$ a.s. for some function $d_1 : \mathcal{X} \mapsto \mathbb{R}_+$. Other than this exception, the watermarking game with no covertext is exactly as described in Section 2.1. The remainder of this section is organized as follows. We first provide some relevant definitions in Section 6.1.1. We then show achievability in Section 6.1.2 and finally show a converse in Section 6.1.3.

6.1.1 Definitions

Let $\mathcal{P}(\mathcal{X})$ and $\mathcal{W}(\mathcal{Y}|\mathcal{X})$ denote the set of all distributions on $\mathcal{X}$ and the set of all conditional distributions on $\mathcal{Y}$ given $\mathcal{X}$. For particular $P \in \mathcal{P}(\mathcal{X})$ and $W \in \mathcal{W}(\mathcal{Y}|\mathcal{X})$, let $PW \in \mathcal{P}(\mathcal{Y})$ denote the resulting marginal distribution on $\mathcal{Y}$ and let $P \circ W$ denote the joint distribution[1]. We write $I(P, W)$ to denote the mutual information between random variables X and Y when they have joint distribution $P \circ W$. Similarly, we write $H(P)$, $H(PW)$ and $(W|P)$ to denote the entropy of X, the entropy of Y, and the conditional entropy of Y given X, respectively. Thus, we can simplify the mutual information as $I(P, W) = H(PW) - H(W|P)$.

Let $\mathcal{P}_n(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ denote the set of all distributions on $\mathcal{X}$ such that $nP(a)$ is an integer for all $P \in \mathcal{P}_n(\mathcal{X})$ and $a \in \mathcal{X}$. A distribution $P \in \mathcal{P}_n(\mathcal{X})$ is also referred to as a type of length n. For any $\boldsymbol{x} \in \mathcal{X}^n$, let $P_{\boldsymbol{x}} \in \mathcal{P}_n(\mathcal{X})$ denote the empirical distribution of $\boldsymbol{x}$, i.e., $P_{\boldsymbol{x}}(a) = N(a|\boldsymbol{x})$ for all $a \in \mathcal{X}$. Similarly, for $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$, let $P_{\boldsymbol{y}|\boldsymbol{x}}$ denote the empirical conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$. For a distribution $P \in \mathcal{P}_n(\boldsymbol{X})$, the set of sequences of type P is written $T_P = \{\boldsymbol{x} : P_{\boldsymbol{x}} = P\}$. Similarly, for $\boldsymbol{x} \in \mathcal{X}^n$ and $W \in \mathcal{W}(\mathcal{Y}|\mathcal{X})$, the W-shell of $\boldsymbol{x}$ is written $T_W(\boldsymbol{x}) = \{\boldsymbol{y} : P_{\boldsymbol{y}|\boldsymbol{x}} = W\}$. The empirical mutual information between two sequences $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ is the mutual information given by their empirical distributions. That is, $I(\boldsymbol{x} \wedge \boldsymbol{y}) = I(P_{\boldsymbol{x}}, P_{\boldsymbol{y}|\boldsymbol{x}}) = I(P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}})$.

[1] In this section, we use W to denote a conditional distribution instead of the message, as we do in the rest of the thesis. Furthermore, we use M to denote the watermark message.

Finally, we slightly rewrite the definition (2.19) as

$$C^{\mathrm{NoCov}}(D_1, D_2) = \max_{P \in \mathcal{P}(\mathcal{X}): E_P[d_1(X)] \le D_1} \; \min_{W \in \mathcal{W}(\mathcal{Y}|\mathcal{X}): E_{P \circ W}[d_2(X,Y)] \le D_2} I(P, W). \tag{6.1}$$

6.1.2 Achievability

We now show that all rates less than $C^{\mathrm{NoCov}}(D_1, D_2)$ are achievable. To do so, we use a codebook of IID vectors with each vector chosen uniformly over a type along with a maximum mutual information (MMI) decoder to yield the desired result.

First, fix n and choose a $Q_n \in \mathcal{P}_n(\mathcal{X})$ and $\delta > 0$. Let

$$R = I\big(Q_n, W^*(Q_n, D_2)\big) - \delta, \tag{6.2}$$

where

$$W^*(P, D_2) = \operatorname*{arg\,min}_{W \in \mathcal{W}(\mathcal{Y}|\mathcal{X}): E_{P \circ W}[d_2(X,Y)] \le D_2} I(P, W). \tag{6.3}$$

The codebook consists of 2^{nR} length-n IID vectors $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{nR}}\}$, where each $\boldsymbol{X}_j$ is uniformly distributed on T_{Q_n}. Note that $\boldsymbol{X}_m \in T_{Q_n}$ and hence $d_1(\boldsymbol{X}_m) = E_{Q_n}[d_1(X)]$ almost surely for all m.

Given the codebook and the forgery $\boldsymbol{y}$, an estimate of the message is found using an MMI decoder. That is,

$$\hat{m} = \operatorname*{arg\,max}_{1 \le m' \le 2^{nR}} I(\boldsymbol{x}_{m'} \wedge \boldsymbol{y}),$$

with ties decided arbitrarily. Without loss of generality, we assume that the correct message $M = 1$. Thus, an error occurs only if there exists a $m \neq 1$ such that $I(\boldsymbol{x}_m \wedge \boldsymbol{y}) \ge I(\boldsymbol{x}_1 \wedge \boldsymbol{y})$.

Given the stegotext $\boldsymbol{x}_1$, the attacker can choose as a forgery any $\boldsymbol{y}$ such that $d_2(\boldsymbol{x}_1, \boldsymbol{y}) \le D_2$. Note that $I(\boldsymbol{x}_1 \wedge \boldsymbol{y}) = I(Q_n, P_{\boldsymbol{y}|\boldsymbol{x}_1}) \ge I(Q_n, W^*(Q_n, D_2))$ since $P_{\boldsymbol{y}|\boldsymbol{x}_1}$ must satisfy the condition in (6.3). Further note that it is sufficient to choose a deterministic attacker to prove the achievability of the proposed rate; see Section 2.4.3.

Given the stegotext $\boldsymbol{x}_1$ and the forgery $\boldsymbol{y}$, let $\mathcal{E}(\boldsymbol{x}_1, \boldsymbol{y})$ be the set of all $\boldsymbol{x} \in T_{Q_n}$ that

could cause an error. That is,

$$\begin{aligned}\mathcal{E}(\boldsymbol{x}_1,\boldsymbol{y}) &= \{\boldsymbol{x}\in T_{Q_n} : I(\boldsymbol{x}\wedge\boldsymbol{y}) \geq I(\boldsymbol{x}_1\wedge\boldsymbol{y})\} \\ &= \bigcup_{\substack{V\in\mathcal{W}(\mathcal{X}|\mathcal{Y}):\\ T_{Q_n}\cap T_V(\boldsymbol{y})\neq\emptyset, H(V|P_{\boldsymbol{y}})\leq H(P_{\boldsymbol{x}_1|\boldsymbol{y}}|P_{\boldsymbol{y}})}} T_V(\boldsymbol{y}), \end{aligned} \tag{6.4}$$

where the inequality in the union follows since if $T_{Q_n}\cap T_V(\boldsymbol{y})\neq\emptyset$ implies that $P_{\boldsymbol{y}}V = Q_n$ and $I(\boldsymbol{x}\wedge\boldsymbol{y}) = H(P_{\boldsymbol{y}}V) - H(V|P_{\boldsymbol{y}})$ for $\boldsymbol{x}\in T_V(\boldsymbol{y})$. There are at most $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ elements in the above union and $|T_V(\boldsymbol{y})|\leq 2^{nH(V|P_{\boldsymbol{y}})}$ for every such V; see e.g. [CK81]. Thus,

$$\begin{aligned}|\mathcal{E}(\boldsymbol{x}_1,\boldsymbol{y})| &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}2^{nH(P_{\boldsymbol{x}_1|\boldsymbol{y}}|P_{\boldsymbol{y}})} \\ &= (n+1)^{|\mathcal{X}||\mathcal{Y}|}2^{n\left(H(Q_n)-I(Q_n,P_{\boldsymbol{y}|\boldsymbol{x}_1})\right)}, \end{aligned} \tag{6.5}$$

where the inequality follows by the above reasoning and (6.4) and the equality follows by the definition of mutual information since $P_{\boldsymbol{x}_1} = Q_n$.

Since each $\boldsymbol{X}_m$ for $m\neq 1$ is uniformly distributed on T_{Q_n} (and independent of $\boldsymbol{X}_1$ and $\boldsymbol{Y}$), the probability of error can be upper bounded using the union bound as

$$\begin{aligned}&\Pr(\text{error}|\boldsymbol{X}_1=\boldsymbol{x}_1,\boldsymbol{Y}=\boldsymbol{y}) \\ &\leq 2^{nR}\frac{|\mathcal{E}(\boldsymbol{x}_1,\boldsymbol{y})|}{|T_{Q_n}|} \\ &\leq 2^{nR}\frac{(n+1)^{|\mathcal{X}||\mathcal{Y}|}2^{n\left(H(Q_n)-I(Q_n,P_{\boldsymbol{y}|\boldsymbol{x}_1})\right)}}{(n+1)^{-|\mathcal{X}|}2^{nH(Q_n)}} \\ &= (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)}2^{n\left(R-I(Q_n,P_{\boldsymbol{y}|\boldsymbol{x}_1})\right)} \\ &\leq (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)}2^{-n\delta}, \end{aligned} \tag{6.6}$$

where the second inequality follows by (6.5) and since T_Q is non-empty then $|T_Q|\geq (n+1)^{-1}2^{nH(Q)}$ [CK81, Lemma 1.2.3], and the last inequality follows since $I(Q_n,P_{\boldsymbol{y}|\boldsymbol{x}_1})\geq I(Q_n,W^*(Q_n,D_2))$ and by (6.2). Note that this bound does not depend on $\boldsymbol{x}_1$ or $\boldsymbol{y}$, and thus the overall probability of error can also be bounded by the right hand side (RHS) of (6.6), which tends to zero.

The above reasoning demonstrates that if

$$R < \limsup_{n\to\infty} \max_{Q_n \in \mathcal{P}_n(\mathcal{X}): E_{Q_n}[d_1(X)] \le D_1} \min_{W \in \mathcal{W}(\mathcal{Y}|\mathcal{X}): E_{Q_n \circ W}[d_2(X,Y)] \le D_2} I(Q_n, W), \tag{6.7}$$

then a sequence of allowable encoder/decoder pairs can be constructed such that the probability of error goes to zero for all sequences of allowable attackers. The proof of achievability will be complete once we demonstrate that the RHS of (6.7) is at least $C^{\mathrm{NoCov}}(D_1, D_2)$; see (6.1). To see that this is true, note that any distribution $P \in \mathcal{P}(\mathcal{X})$ can be approached uniformly by a sequence of types, $\{Q_n \in \mathcal{P}_n(\mathcal{X})\}$.

6.1.3 Converse

We now show that no rates larger than $C^{\mathrm{NoCov}}(D_1, D_2)$ are achievable for discrete alphabets when there is no covertext. To do so, we describe an attacker that satisfies this requirement, even if the encoder and decoder are designed with full knowledge of this attacker.

The attacker's basic strategy can be described as follows. He first chooses a constant $\tilde{D}_2 < D_2$. Given the stegotext $\boldsymbol{x}$, the attacker computes the best response $W^*(P_{\boldsymbol{x}}, \tilde{D}_2)$, where W^* is defined in (6.3).

The attacker then creates the forgery by using the stegotext as an input to a memoryless channel with the conditional distribution $W^*(P_{\boldsymbol{x}}, \tilde{D}_2)$. If the attacker does not satisfy the distortion constraint (i.e., $n^{-1}d_2(\boldsymbol{x}, \boldsymbol{y}) > D_2$) after the application of this memoryless channel, then the attacker arbitrarily changes some components of the forgery so that the distortion constraint is satisfied.

We first note that the probability that the attacker must change the forgery tends to zero as the blocklength n tends to infinity. This follows since given the stegotext $\boldsymbol{x}$, the expected normalized distortion between the stegotext and the covertext $\boldsymbol{Y}$ is at most $\tilde{D}_2$ that is in turn smaller than D_2. We will thus analyze the probability assuming that that the attacker never needs to change the forgery.

We now show that for constant composition codebooks the probability of error cannot tend to zero unless the rate R is at most $C^{\mathrm{NoCov}}(D_1, \tilde{D}_2)$, where a constant composition codebook is one in which all of the codewords are of the same type. It is well known that if there are 2^{nR} codewords of type P and a memoryless channel W, then the probability of error is bounded away from zero for $R > I(P, W)$. The type P of the codewords must

satisfy $E_P[d_1(X)] \leq D_1$, and thus

$$I\big(P, W^*(P, \tilde{D}_2)\big) \leq C(D_1, \tilde{D}_2);$$

compare (6.3) and (6.1). Combining these two facts yields the preliminary result.

We can extend the above result on constant composition codebooks to general codebooks. To do so, we break up the codebook of 2^{nR} codewords into types that are represented by fewer than $(n+1)^{-2|\mathcal{X}|}2^{nR}$ codewords and types that are represented by at least $(n+1)^{-2|\mathcal{X}|}2^{nR}$ codewords. For codewords in the former category, we can trivially lower bound the probability of error by zero. Since there are at most $(n+1)^{|\mathcal{X}|}$ types of length n, the fraction of codewords in this category is at most $(n+1)^{-|\mathcal{X}|}$, which tends to zero. For codewords in the latter category we will use the result on constant composition codebooks to analyze the probability of error. In each of the constant composition sub-codebooks under consideration, the number of codewords is at least

$$(n+1)^{-2|\mathcal{X}|}2^{nR} = 2^{n\left(R - \frac{2|\mathcal{X}|\log(n+1)}{n}\right)}.$$

Since the exponent on the RHS is asymptotically equal to R, even a decoder that knows the type of the codeword will not be able to reliably decode the message if $R > C^{\mathrm{NoCov}}(D_1, \tilde{D}_2)$. Finally, the fraction of codewords for which the probability of error is bounded away from zero approaches one, and thus the average probability of error is also bounded away from zero.

We finally note that $C^{\mathrm{NoCov}}(\cdot, \cdot)$ is continuous in its arguments. Thus, since $\tilde{D}_2 < D_2$ is chosen arbitrarily, no rate higher than $C(D_1, D_2)$ is achievable.

6.2 Binary Covertext

In this section, we give a proof of Theorem 2.6. That is, we will describe the capacity for the watermarking game when all of the alphabets are binary and the covertext is an IID sequence of Bernoulli(1/2) random variables and the distortion is measured using Hamming distance.

6.2.1 Private Version

We first show that all rates less than $H_b(D_1 \otimes D_2) - H_b(D_2)$ are achievable in the private version. To do so, we show that any rate that is achievable using random coding for the AVC with constrained inputs and states (see [CN88a] and Section 2.5.3) is also achievable here. The fact that the covertext $\boldsymbol{U}$ is known to both the encoder and the decoder provides perfect secrecy about the transmitted sequence. To see this, let the encoder generate a "codeword" $\tilde{\boldsymbol{X}} = f_n(W, \Theta_1)$ independently of $\boldsymbol{U}$ (but depending on the watermark W and the secret key Θ_1), where the codeword must satisfy $n^{-1}w_h(\tilde{\boldsymbol{X}}) \leq D_1$ a.s.[2]. The encoder forms the stegotext as $\boldsymbol{X} = \tilde{\boldsymbol{X}} \oplus \boldsymbol{U}$, which is independent of the codeword $\tilde{\boldsymbol{X}}$, due to the distribution of $\boldsymbol{U}$. Since the forgery $\boldsymbol{Y}$ depends only on the stegotext, it follows that the sequence $\tilde{\boldsymbol{Y}} = \boldsymbol{X} \oplus \boldsymbol{Y}$ (which must satisfy $n^{-1}w_h(\tilde{\boldsymbol{Y}}) \leq D_2$ a.s.) is also independent of the codeword $\tilde{\boldsymbol{X}}$. The decoder knows the covertext, and thus he can base his estimate of the watermark on $\boldsymbol{Y} \oplus \boldsymbol{U} = \tilde{\boldsymbol{X}} \oplus \tilde{\boldsymbol{Y}}$, where again $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are independent. As in Section 2.4.3, it is sufficient to show that the probability of error tends to zero for every deterministic sequence $\tilde{\boldsymbol{y}}$ such that $n^{-1}w_h(\tilde{\boldsymbol{y}}) \leq D_2$. That is, it is sufficient to find a sequence of feasible rate-R encoders $\{f_n\}$ and decoders $\{\phi_n\}$ such that

$$\bar{P}_e(\tilde{\boldsymbol{y}}) = \Pr\Big(\phi_n\big(f_n(W, \Theta_1) \oplus \tilde{\boldsymbol{y}}\big) \neq W\Big) \tag{6.8}$$

vanishes for every feasible $\tilde{\boldsymbol{y}}$. Here, feasible means that $n^{-1}w_h(f_n(W, \Theta_1)) \leq D_1$ a.s. and $n^{-1}w_h(\tilde{\boldsymbol{y}}) \leq D_2$. This is an instance of randomized coding on an AVC with constrained inputs and states [CN88a, CN88b]. In [CN88b, Sect. IV], it was shown that this capacity is given by $H_b(D_1 \otimes D_2) - H_b(D_2)$.

We now show the converse for the private version, i.e., that no rates higher than $H_b(D_1 \otimes D_2) - H_b(D_2)$ are achievable. Let us fix the attacker to be a binary symmetric channel with crossover probability $\tilde{D}_2 < D_2$. By the weak law of large numbers, $n^{-1}w_h(\boldsymbol{X} \oplus \boldsymbol{Y})$ will be smaller than D_2 with arbitrarily high probability for blocklength n large enough, and thus a trivial modification of this attacker will satisfy the distortion constraint. With this fixed attack channel, it is straightforward to show that the capacity is given by $H_b(D_1 \otimes \tilde{D}_2) - H_b(\tilde{D}_2)$. The converse follows since this expression is continuous in $\tilde{D}_2$ and since we can

[2]Recall that the Hamming weight of a vector $\boldsymbol{x} \in \{0,1\}^n$ is the number of ones contained in $\boldsymbol{x}$ and is written $w_h(\boldsymbol{x})$.

make $\tilde{D}_2$ arbitrarily close to D_2.

6.2.2 Public Version

In this section, we show that the capacity of the public version of the binary watermarking game is given by

$$C_{\text{pub}}^{\text{BinWM}}(D_1, D_2) = \max_{2D_1 \le g \le 1} g \cdot \left(H_b\left(\frac{D_1}{g}\right) - H_b(D_2) \right).$$

Achievability

We now prove that all rates less than $C_{\text{pub}}^{\text{BinWM}}(D_1, D_2)$ are achievable. To do so, we describe a randomized encoding and decoding strategy. We then show that this strategy satisfies the distortion constraint. We finally show that the probability of error tends to zero for any attacker.

To describe the encoder and decoder, we fix $2D_1 \le g \le 1$ and $\epsilon > 0$, and let

$$R_0 = g \cdot \left(1 - H_b\left(\frac{D_1}{g}\right) + \epsilon \right), \tag{6.9}$$

and

$$R = g \cdot \left(H_b\left(\frac{D_1}{g}\right) - H_b\left(D_2\right) - 2\epsilon \right). \tag{6.10}$$

The encoder/decoder pair generates $2^{n(R+R_0)}$ IID vectors, each a length-ng IID sequence of Bernoulli(1/2) random variables. This codebook thus consists of $2^{n(R+R_0)}$ random vectors $\{\boldsymbol{V}(w,k),\ 1 \le w \le 2^{nR},\ 1 \le k \le 2^{nR_0}\}$. The encoder/decoder pair also selects ng indices uniformly out of all subsets of $\{1, \ldots, n\}$ of size ng, say $\boldsymbol{P} = \{P(1), \ldots, P(ng)\}$. For a length-$n$ vector $\boldsymbol{u}$ and a size-ng position set $\boldsymbol{p}$, we write $\boldsymbol{u}|_{\boldsymbol{p}}$ to mean the length-ng vector at the points $\boldsymbol{p}$, i.e., $\boldsymbol{u}|_{\boldsymbol{p}} = (u_{p(1)}, \ldots, u_{p(ng)})$. Given the covertext $\boldsymbol{u}$, the watermark w, the codebook $\{\boldsymbol{v}\}$ and the indices $\boldsymbol{p}$, the encoder selects the value

$$k^* = \operatorname*{arg\,min}_{1 \le k \le 2^{nR_0}} w_h\left(\boldsymbol{u}|_{\boldsymbol{p}} \oplus \boldsymbol{v}(w,k)\right). \tag{6.11}$$

The encoder then creates the stegotext as

$$\boldsymbol{x}_i = \begin{cases} \boldsymbol{v}_j(w, k^*) & \text{if } i \in \boldsymbol{p} \text{ and } i = p(j) \\ \boldsymbol{u}_i & \text{otherwise} \end{cases}. \tag{6.12}$$

In other words, the encoder finds the codeword that best matches the covertext at the selected points and then replaces the covertext with the codeword at those points. At the other end, the decoder finds the codeword closest to the forgery $\boldsymbol{y}$ at the selected points. That is, he estimates the watermark as

$$\hat{w} = \arg\min_{1 \le w' \le 2^{nR}} \min_{1 \le k \le 2^{nR_0}} w_h\left(\boldsymbol{y}|_{\boldsymbol{p}} \oplus \boldsymbol{v}(w', k)\right). \tag{6.13}$$

The main fact that we will use for the remainder of the proof is the following. For $2^{mR'}$ IID random vectors $\{\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_{2^{mR'}}\}$ where each $\boldsymbol{X}'_i$ is an IID sequence of m Bernoulli(1/2) random variables, let

$$P^{\text{bin}}(m, D, R') = \Pr\left(\min_{1 \le i \le 2^{mR'}} m^{-1} w_h(\boldsymbol{X}'_i) \le D\right). \tag{6.14}$$

Then, for any $0 \le D \le 1/2$,

$$\lim_{m \to \infty} P^{\text{bin}}(m, D, R') = \begin{cases} 1 & \text{if } R' > 1 - H_b(D) \\ 0 & \text{if } R' < 1 - H_b(D) \end{cases}. \tag{6.15}$$

To show that the encoder satisfies $n^{-1} w_h(\boldsymbol{U} \oplus \boldsymbol{X}) \le D_1$ with arbitrarily high probability, we apply (6.15) with $m = ng$, $D = D_1/g$, and $R' = R_0/g$. To see this, note that $w_h(\boldsymbol{U} \oplus \boldsymbol{X}) = w_h\left(\boldsymbol{U}|_{\boldsymbol{P}} \oplus \boldsymbol{V}(W, k)\right)$ since $\boldsymbol{U}$ and $\boldsymbol{X}$ can only differ at points in $\boldsymbol{P}$. We can now write that

$$\begin{aligned} &\Pr\left(\min_{1 \le k \le 2^{nR_0}} n^{-1} w_h\left(\boldsymbol{U}|_{\boldsymbol{P}} \oplus \boldsymbol{V}(W, k)\right) \le D_1\right) \\ &\quad = \Pr\left(\min_{1 \le k \le 2^{ng(R_0/g)}} (ng)^{-1} w_h\left(\boldsymbol{V}(W, k)\right) \le \frac{D_1}{g}\right) \\ &\quad = P^{\text{bin}}(ng, D_1/g, R_0/g), \end{aligned} \tag{6.16}$$

where the first equality follows since $\boldsymbol{U}|_{\boldsymbol{P}} \oplus \boldsymbol{V}(W, k)$ is independent of $\boldsymbol{U}$ and thus the

distribution of the Hamming weight of $\boldsymbol{U}|_{\boldsymbol{P}} \oplus \boldsymbol{V}(W,k)$ does not depend on the realization of $\boldsymbol{u}$ (in particular, it is the same when $\boldsymbol{u} = 0$). Finally, the RHS of (6.16) goes to zero by the definition R_0 (6.9) and by (6.15).

To analyze the probability of an incorrect decision being made by the decoder, we first note that the probability of error depends on the attack only in the amount of distortion that is introduced. This follows from the randomized construction of the encoder and decoder. Thus, it is sufficient to analyze the probability of error caused by an attacker of the form $\boldsymbol{Y} = \boldsymbol{X} \oplus \tilde{\boldsymbol{y}}$ for some deterministic sequence $\tilde{\boldsymbol{y}}$ with $w_h(\tilde{\boldsymbol{y}}) = \lfloor nD_2 \rfloor$. For example, we could let $\tilde{\boldsymbol{y}}_i$ be 1 for $1 \leq i \leq \lfloor nD_2 \rfloor$ and 0 otherwise. Thus, we can claim that $\Pr\left\{n^{-1} w_h(\boldsymbol{Y}|_{\boldsymbol{P}} \oplus \boldsymbol{V}(W,k^*)) \leq g(D_2+\delta)\right\}$ tends to one for any $\delta > 0$ and for the correct watermark W. Conditioning on this event, the probability that an incorrect watermark will be selected by the decoder is given by $P^{\text{bin}}(ng, D_2+\delta, (R+R_0)/g)$, which tends to zero for δ sufficiently small by the definitions of R_0 (6.9) and R (6.10) and by (6.15). Thus, the overall probability of error can be made as small as desired by choosing a large enough blocklength.

To conclude the achievability proof, we note that $2D_1 \leq g \leq 1$ and $\epsilon > 0$ can be arbitrarily chosen. Thus, any $R < C_{\text{pub}}^{\text{BinWM}}(D_1, D_2)$ is achievable.

Converse

In this section, we prove that no rates higher than $C_{\text{pub}}^{\text{BinWM}}(D_1, D_2)$ are achievable for the binary watermarking game. We do so, as in the private version, by fixing the attacker to be a binary symmetric channel with crossover probability $\tilde{D}_2 < D_2$. For this attacker, the distortion constraint will be met with arbitrarily high probability for blocklength n large enough. We will further show, using the results of Lemma 2.29, that the capacity with this fixed attacker is given by $C_{\text{pub}}^{\text{BinWM}}(D_1, \tilde{D}_2)$. The converse is completed by noting that this expression is continuous in $\tilde{D}_2$.

The remainder of this section will be devoted to evaluating the capacity of the following channel with side information. The side information vector $\boldsymbol{U}$ is a sequence of IID Bernoulli(1/2) random variables. The channel is given by

$$P_{Y|X,U}(y|x,u) = \begin{cases} 1-\tilde{D}_2 & \text{if } y = x \\ \tilde{D}_2 & \text{if } y \neq x \end{cases}.$$

Note that the channel does not depend on the side information. Instead, the side information restricts the possible inputs since the input sequence must be within distance D_1 of the side information, i.e., $n^{-1}w_h(\boldsymbol{U}, \boldsymbol{x}(W, \boldsymbol{U})) \leq D_1$ a.s.. We have shown in Lemma 2.29 that the capacity of this channel is given by

$$C(D_1) = \max_{\substack{P_{V|U},\, f:\mathcal{V}\times\mathcal{U}\mapsto\mathcal{X}, \\ E[w_h(U,X)]\leq D_1}} I(V;Y) - I(V;U), \tag{6.17}$$

where V is an auxiliary random variable with finite alphabet, and the mutual informations are evaluated with respect to the joint distribution

$$P_{U,V,X,Y}(u,v,x,y) = \begin{cases} P_U(u)P_{V|U}(v|u)P_{Y|X,U}(y|x,u) & \text{if } x = f(v,u) \\ 0 & \text{otherwise} \end{cases}.$$

In order to evaluate (6.17), let us set $\mathcal{V} = \{v_0, v_1, v_2\}$, which we will see to be sufficient. Recall that $\mathcal{U} = \mathcal{X} = \mathcal{Y} = \{0,1\}$. Without loss of generality, we can fix the function f to be

$$f(v,u) = \begin{cases} 0 & \text{if } v = v_0 \\ 1 & \text{if } v = v_1 \\ u & \text{if } v = v_2 \end{cases}.$$

The only other possibility for f would be to set $f(v,u) = u \oplus 1$ for some v, which turns out to be suboptimal. We now only need to optimize over $P_{V|U}$ in order to evaluate $C(D_1)$. The distortion constraint requires that $(P_{V|U}(v_0|1) + P_{V|U}(v_1|0))/2 = D_1$, since these are the only cases where u and $x = f(u,v)$ differ. In order to simplify the optimization, we also require that $P_V(v_2) = 1 - g$ for some $2D_1 \leq g \leq 1$. We later choose the best g as in the definition of $C_{\text{pub}}^{\text{BinWM}}(D_1, D_2)$. We now claim that under these constraints, the optimal distribution is given by

$$P^*_{V|U}(v|u) = \begin{cases} D_1 & \text{if } (u,v) = (0,v_1) \text{ or } (1,v_0) \\ g - D_1 & \text{if } (u,v) = (0,v_0) \text{ or } (1,v_1) \\ 1 - g & \text{if } v = v_2 \end{cases}.$$

Under this distribution, $I(V;Y) - I(V;U) = g \cdot (H_b(D_1/g) - H_b(\tilde{D}_2))$. Thus, the establishment of this claim will complete the proof of the converse.

In order to bound $I(V;Y) - I(V;U)$ for a generic distribution that satisfies the above constraints, we will use the following calculation

$$\begin{aligned}
& P_V(v_0)H(U|V=v_0) + P_V(v_1)H(U|V=v_1) \\
&= g\left(\frac{P_V(v_0)}{g} H_b\left(P_{U|V}(1|v_0)\right) + \frac{P_V(v_1)}{g} H_b\left(P_{U|V}(0|v_1)\right)\right) \\
&\leq g H_b\left(\frac{P_V(v_0)P_{U|V}(1|v_0) + P_V(v_1)P_{U|V}(0|v_1)}{g}\right) \\
&= g H_b\left(\frac{D_1}{g}\right), \qquad (6.18)
\end{aligned}$$

where recall that $g = P_V(v_0) + P_V(v_1)$ and the inequality follows by the concavity of entropy. We can thus bound

$$\begin{aligned}
I(U;V) &= H(U) - H(U|V) \\
&\geq 1 - g H_b\left(\frac{D_1}{g}\right) - (1-g)H_b(\xi), \qquad (6.19)
\end{aligned}$$

where $\xi = P_{U|V}(0|v_2)$ and the inequality follows by (6.18). We can also bound

$$\begin{aligned}
I(V;Y) &= H(Y) - H(Y|V) \\
&\leq 1 - g H_b(D_2) - (1-g)H_b\big(1 - D_2 + \xi(2D_2 - 1)\big), \qquad (6.20)
\end{aligned}$$

where the inequality follows since Y is a binary random variable. Combining (6.19) and (6.20), we see that

$$\begin{aligned}
& I(V;Y) - I(U;V) \\
&\leq g \cdot \left(H_b\left(\frac{D_1}{g}\right) - H_b\left(D_2\right)\right) + (1-g) \cdot \Big(H_b(\xi) - H_b\big(1 - D_2 + \xi(2D_2 - 1)\big)\Big) \\
&\leq g \cdot \left(H_b\left(\frac{D_1}{g}\right) - H_b\left(D_2\right)\right), \qquad (6.21)
\end{aligned}$$

where the second inequality follows by maximizing over ξ (the maximum is achieved at $\xi = 1/2$). The bound (6.21) is achieved with equality when $P^*_{V|U}$ is used. This establishes that $C(D_1) = C^{\text{BinWM}}_{\text{pub}}(D_1, \tilde{D}_2)$, which completes the proof of the converse.

Chapter 7

Conclusions

In this thesis, we have defined the information theoretic capacity of a watermarking system, and we have found this capacity in many scenarios. We now comment on some of their interesting aspects of these findings. We conclude in Section 7.1.1 with some ideas for future research.

We have formalized a watermarking model in which a malicious attacker attempts to prevent reliable transmission of the watermark. We assume that this attacker knows the workings of both the encoder and decoder (but not a secret key shared by the encoder and decoder). We also assume that any forgery created by the attacker is only useful to him if the distortion between the forgery and stegotext is less than some threshold. Thus, we only consider attackers that meet this distortion constraint with probability one; we show that the capacity is zero when the constraint is enforced in expectation (see Section 2.2.3). These assumptions require the watermarking system (both encoder and decoder) to be designed first so that they are robust against any feasible attacker.

When the covertext has a Gaussian distribution, we have shown that the capacity is the same in the private and public versions; see Theorems 2.1 and 2.4. This surprising result demonstrates that the capacity does not increase if the decoder can use the covertext. Costa's "writing on dirty paper" [Cos83] has this same feature; we gave two extensions of his result in Section 2.5.4. Although the capacity is the same for both versions, the capacity achieving coding scheme for the public version is much more complex than the scheme for the private version; compare Sections 4.2 and 4.3 for the SGWM game and Sections 5.4 and 5.5 for the VGWM game. As one might expect, the two versions of watermarking

do not always yield the same capacity. For example, in the binary watermarking game of Section 2.2.6, there is a difference between the two versions.

When the covertext is an IID sequence of Gaussian random variables, we have shown that an optimal lossy compression attack prevents rates greater than capacity from being achievable. This property would allow designers to test the robustness of their watermarking systems against existing technology. Unfortunately, this property does not hold in general. Indeed, for an IID vector Gaussian, the compression attack is not optimal, and designing for robustness against such an attack yields a qualitatively different watermarking system; see Section 5.7 for more discussion.

We have seen that the watermarking capacity increases with the uncertainty in the covertext. Indeed, for the SGWM game, the capacity is increasing in the variance of the covertext; see Figure 2-1. Furthermore, with squared error distortion measures and a fixed covertext variance, the covertext distribution with the largest capacity is the Gaussian distribution, which also yields the highest entropy for out of all distributions with the same 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 can only add an arbitrary sequence (see Section 2.2.2 on the additive attack watermarking game) or an independent random sequence (see Section 2.5.4 on extended 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.

One technical result that might be of general interest is Lemma B.7. There, we consider the mutual information between a Gaussian random variable and some other random variable, with the second order moments fixed. We show that this mutual information is maximized if the other random variable is jointly Gaussian with the first one.

7.1 Future Research

In this section, we offer some directions for future research that expand on the themes we have presented in the thesis.

7.1.1 Gaussian Sources with Memory

We would like to find the capacity of the watermarking game for squared error distortion and a stationary Gaussian covertext. That is, let us assume that the covertext $\boldsymbol{U}$ is a stationary Gaussian process with covariance $E[U_j U_k] = t_{|j-k|}$. We also assume that the covertext has a finite memory m_0 so that $t_m = 0$ for $m > m_0$.

We believe that we can use the results on the vector Gaussian watermarking game (see Section 2.2.4) to describe the capacity for this covertext distribution. Indeed, for any m, the vectors $\boldsymbol{U}'_j = (U_{j(m+m_0)+1}, \ldots, U_{j(m+m_0)+m})$, for $j = 0, 1, \ldots$ form an IID sequence of Gaussian random vectors with covariance matrix $T^{(m)}$. Here, $T^{(m)}$ is the $m \times m$ matrix with $T^{(m)}_{jk} = t_{|j-k|}$. We will write the set of eigenvalues of $T^{(m)}$ as $\{\lambda^{(m)}_k,\ 1 \leq k \leq m\}$. The encoder/decoder could use the coding strategy for the vector Gaussian source with the additional restriction of making the stegotext independent of the covertext at the indices not used in forming $\{\boldsymbol{U}'\}$. For example, the encoder could set $x_{j(m+m_0)+k} = 0$ for $m < k \leq m + m_0$. This restriction is needed so that the attacker cannot gain any knowledge of the covertext samples used for encoding the watermark. This restriction uses some of the encoder's available distortion, but this extra distortion can be made negligible by taking m large enough. Thus, we conjecture that any rate less than the following limit should be achievable:

$$\lim_{m\to\infty} \max_{\boldsymbol{D}_1 \in \mathcal{D}_m(m\Delta_1)} \min_{\boldsymbol{D}_2 \in \mathcal{D}_m(m\Delta_2)} C^* \left(D_{1k}, D_{2k}, \lambda^{(m)}_k \right), \tag{7.1}$$

where the term inside the limit is the normalized capacity of the vector Gaussian watermarking game with covariance $T^{(m)}$, encoder distortion level mD_1 and attacker distortion level mD_2. Furthermore, we believe that there exist attackers that guarantee that no rates larger than (7.1) are achievable. Such an attacker would assume that the covertext is a blockwise IID sequence of Gaussian random vectors.

We would also like to simplify the limit (7.1) into a more meaningful expression. We can use the fact that the covariance matrices $T^{(m)}$ are Toeplitz matrices, and thus we can

describe the limiting behavior of their eigenvalues (see e.g., [Gra00, GS58]). This is similar to the approach that Gallager [Gal68, Sec. 8.5] takes in describing the capacity of an additive Gaussian noise channel.

7.1.2 Discrete Memoryless Covertext

We would like to study the capacity of the public version of a watermarking game with a general discrete memoryless covertext and general distortion constraints. One conjecture is that the general capacity is given by the related mutual information games. In the private version, this is the solution that Moulin and O'Sullivan [MO99, OME98] derive with a maximum-likelihood decoder and average distortion constraints. Furthermore, Somekh-Baruch and Merhav [SBM01a] have recently shown that for the private version with finite alphabets, the private mutual information game also gives the capacity for a fixed decoder and almost sure distortion constraints. In the public version, all of the watermarking capacities that we described in Section 2.2 have coincided with values of the related mutual information games. However, no one has yet given a proof for the general public version.

7.1.3 Deterministic Code Capacity for Public Version

We would like to find the capacity when no secret key is available to the encoder and decoder. We have addressed this for the private version (see Sections 2.4.2 and 6.2), where we have found that the capacity without a secret key is typically the same as with a secret key. However, this result hinges on the fact that the encoder and decoder both have access to the covertext and they essentially use part its randomness as a secret key. Thus, these arguments do not work in the public version, i.e., when the decoder does not know the covertext. We call the capacity when no secret key is available — and thus the attacker knows the exact encoding and decoding mappings — the *deterministic code capacity*.

We first show that the deterministic code capacity is in general smaller than the random code capacity. For squared error distortion, if $D_2 > 4D_1$, then the attacker can make the forgery into any possible output from the encoder. This implies that the attacker can make the decoder have any possible output as well. Thus, no positive rate is achievable in this regime. Recall, however, that the the capacity of the Gaussian watermarking game with randomized codes is positive in this regime for σ_u^2 large enough. Thus, the deterministic code capacity does not equal the randomized code capacity for the public version.

Additive Attack Watermarking

We now discuss the deterministic code capacity for the additive attack watermarking game with IID Gaussian covertext and squared error distortions. This scenario is similar to the Gaussian arbitrarily varying channel (GAVC), except here the encoder can base his transmitted sequence on the entire Gaussian noise sequence. See Section 2.5.3 for more on the GAVC. Csiszár and Narayan [CN91] studied the deterministic code capacity of the GAVC and found that it is given by

$$C^{\mathrm{DetGAVC}}(D_1, D_2, \sigma^2) = \begin{cases} \frac{1}{2}\log\left(1 + \frac{D_1}{D_2+\sigma^2}\right) & \text{if } D_1 > D_2 \\ 0 & \text{otherwise} \end{cases}. \tag{7.2}$$

In other words, the capacity is either the random code capacity or zero, depending on the allowed distortion levels. In particular, the capacity is not continuous in the parameters. We believe that, unlike the GAVC, the deterministic code capacity for the additive attack watermarking game is continuous in the parameters. Further, we believe that there exists values of the parameters such that the deterministic code capacity is non-zero yet strictly less than the random code capacity. While this is not possible for AVCs without constraints, Csiszár and Narayan [CN88b] showed that this is possible for AVCs with input and state constraints.

Our argument for the above claims is briefly as follows. For D_2 small enough, we believe that we can construct deterministic codes which achieve the random code capacity for the additive attack watermarking game, namely $\frac{1}{2}\log\left(1 + \frac{D_1}{D_2}\right)$. Such a code would be similar to the random code of Section 4.3.1, and could be analyzed using techniques from [CN91]. One difference from [CN91] is that we would have to guarantee that each bin has a codeword that has a small inner product with the covertext. For any coding strategy of this form, the critical distortion level D_2 will be determined by the energy in the covertext which is in the direction of the correct codeword. We believe that by increasing the number of codewords in each bin, we can increase this energy at the expense of overall rate. Thus, the achievable rates for this coding strategy should continuously decrease to zero as D_2 increases instead of a non-continuously as in the GAVC. Besides analyzing such a coding strategy, we also need a converse to show that no higher rates are achievable.

7.1.4 Multiple Rate Requirements

We now consider a watermarking model where the amount of information that can be recovered depends on the distortion introduced by the attacker. For example, let there be two independent watermarks W_h of R_h bits and W_l of R_l bits and two attacker distortion levels $D_{2,h} > D_{2,l}$. The encoder will produce the stegotext as a function of the covertext and both watermarks such that the distortion between the stegotext and covertext is less than D_1. If the distortion between the forgery and stegotext is $D_{2,h}$, then the decoder should be able to recover W_h. However, if the distortion between the forgery and stegotext is $D_{2,l}$, then the decoder should be able to recover both W_h and W_l. The main question is what rates pairs are achievable for given values of $D_{2,h}$ and $D_{2,l}$. (Or conversely, what distortion pairs are allowable for given values of R_h and R_l.) This problem can be thought of as a broadcast channel with degraded message set, see e.g., [Cov75, KM77]. However, the broadcast channel is arbitrarily varying as in [Jah81].

Let us consider this example for an IID Gaussian covertext (zero-mean, variance-σ^2) and squared error distortion. Using the results of Theorem 2.1, we can say that both $(R_h, R_l) = (C^*(D_1, D_{2,h}, \sigma^2), 0)$ and $(R_h, R_l) = (0, C^*(D_1, D_{2,l}, \sigma^2))$ are achievable. However, it is not immediately clear that any linear combination of these rate pairs are achievable using the usual time-sharing argument. Indeed, it seems that in order to effectively time-share against the attacker, both codes will have to use the same stegotext power (i.e., the same value of A). The optimal value of A depends on the attacker's distortion level and hence for any common value of A that the two codes choose, at least one of the codes will be transmitting below capacity. On the positive side, for any value of A and any $0 \leq \lambda \leq 1$, the following rate pairs are achievable

$$(R_h, R_l) = \left(\lambda \cdot \frac{1}{2} \log(1 + s(A; D_1, D_{2,h}, \sigma^2)), (1 - \lambda) \cdot \frac{1}{2} \log(1 + s(A; D_1, D_{2,l}, \sigma^2)) \right). \tag{7.3}$$

This follows since the encoder and decoder can randomly decide on λn locations for the high distortion code, with the remaining positions for the low distortion code. Since the two codes are using the same stegotext power, the attacker cannot focus his distortion on either code. The question remains as to how much better, if any, can we do than this simple time sharing strategy.

Appendix A

Definitions for Gaussian covertext

In this section, we present many of the definitions that are used with Gaussian covertexts (i.e., Chapters 3, 4, and 5). We also discuss some of the basic properties of some of the mappings.

We now summarize the definitions that are used for all of the chapters with Gaussian covertexts. Recall that

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

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

$$b_2 = b_2(A; D_1, \sigma^2) = D_1 - \frac{\rho^2(A; D_1, \sigma^2)}{\sigma^2}, \tag{A.3}$$

$$c(A; D_2) = 1 - \frac{D_2}{A}, \tag{A.4}$$

$$s(A; D_1, D_2, \sigma^2) = \frac{c(A; D_2) b_2(A; D_1, \sigma^2)}{D_2}, \tag{A.5}$$

$$\alpha(A; D_1, D_2, \sigma^2) = 1 - \frac{b_1(A; D_1, \sigma^2)}{1 + s(A; D_1, D_2, \sigma^2)}, \tag{A.6}$$

$$\mathcal{A}(D_1, D_2, \sigma^2) = \left\{ A : \max\left\{ D_2, \left(\sigma - \sqrt{D_1}\right)^2 \right\} \leq A \leq \left(\sigma + \sqrt{D_1}\right)^2 \right\}, \tag{A.7}$$

and, finally,

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

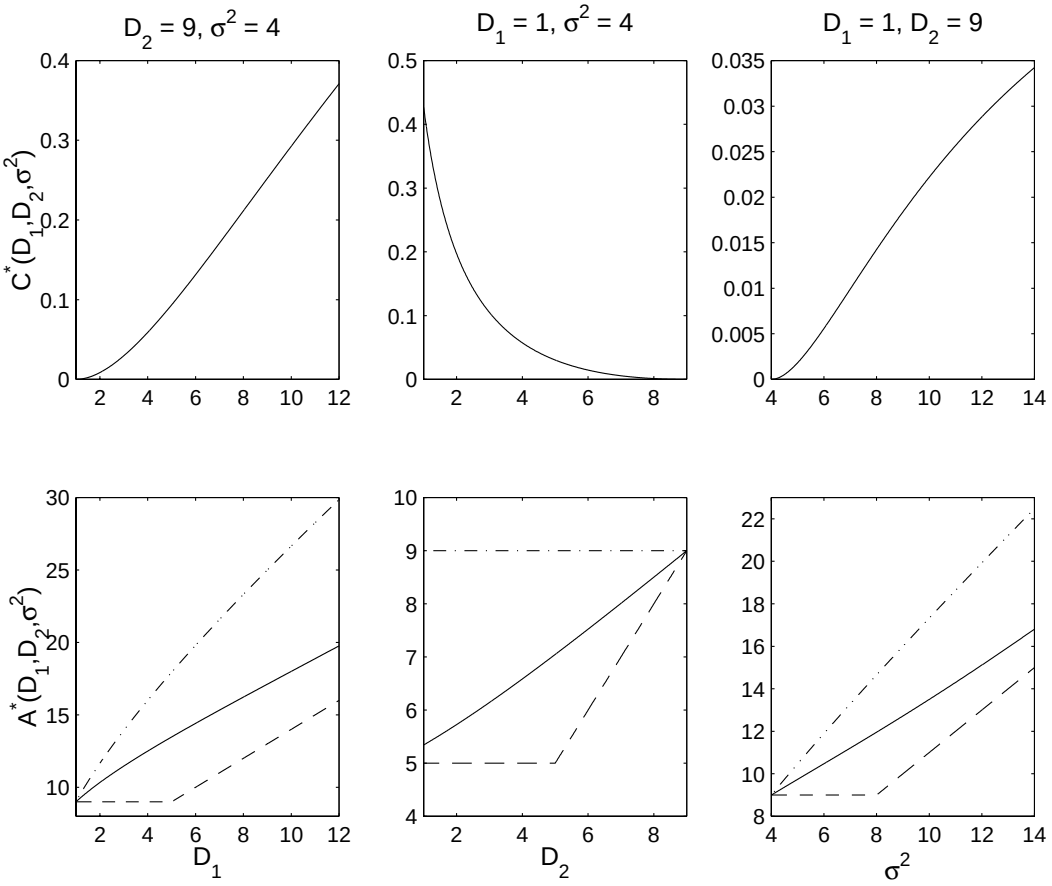

Figure A-1: Example plots of $C^*(D_1, D_2, \sigma^2)$ and $A^*(D_1, D_2, \sigma^2)$ for different parameter values.

The function $C^*(D_1, D_2, \sigma^2)$ is the capacity of the scalar Gaussian watermarking game (Theorem 2.1) and the value of the Gaussian mutual information game (Theorem 3.1); it also plays a critical role in the capacity of the vector Gaussian watermarking game (Theorem 2.4). In Figure A-1, we have plotted $C^*(D_1, D_2, \sigma^2)$ against each of its three arguments. We have also plotted the maximizing A in (A.8) along with the lower and upper limits in the definition of $\mathcal{A}(D_1, D_2, \sigma^2)$. We see that $C^*(D_1, D_2, \sigma^2)$ is non-decreasing in D_1 and σ^2 and non-increasing in D_2. Further, note that $C^*(D_1, D_2, \sigma^2)$ is neither convex nor concave in D_1 and σ^2 (there are points of inflection at $D_1 \approx 8.2$ and $\sigma^2 \approx 6.9$ in the first and third plots of Figure A-1). However, $C^*(D_1, D_2, \sigma^2)$ is convex in D_2 (this follows since $\frac{1}{2}\log(1 + s(A; D_1, D_2, \sigma^2))$ is convex in D_2).

In the following lemma, we describe the A^* that achieves the maximum in the definition

of $C^*(D_1, D_2, \sigma^2)$.

Lemma A.1. *If $\mathcal{A}(D_1, D_2, \sigma^2)$ is non-empty, then the maximizing A in (A.8) is achieved by the unique $\max\{\sigma^2 + D_1, D_2\} < A^* < \left(\sigma + \sqrt{D_1}\right)^2$ such that $p(A^*; D_1, D_2, \sigma^2) = 0$, where*

$$p(A; D_1, D_2, \sigma^2) = A^3 - \left(D_1 + \sigma^2 + \frac{D_2}{2}\right) A^2 + \frac{D_2}{2}(\sigma^2 - D_1)^2.$$

Proof. First note that, independent of D_2, it is sufficient to consider only A such that $\sigma^2 + D_1 \leq A \leq \left(\sigma + \sqrt{D_1}\right)^2$. This follows since for any $\left(\sigma - \sqrt{D_1}\right)^2 \leq A \leq \sigma^2 + D_1$, there exists a $\sigma^2 + D_1 \leq A' \leq \left(\sigma + \sqrt{D_1}\right)^2$ such that $\left|s(A'; D_1, D_2, \sigma^2)\right|^+ \geq \left|s(A; D_1, D_2, \sigma^2)\right|^+$, independently of D_2. Namely, let $A' = 2(\sigma^2 + D_1) - A$.

Since $\log(x)$ is monotonically increasing in x, the maximizing A in (A.8) is also the $A \in \mathcal{A}(D_1, D_2, \sigma^2)$ that maximizes the product $c(A; D_2) b_2(A; D_1, \sigma^2)$. We can calculate that

$$\frac{\partial}{\partial A} c(A; D_2) b_2(A; D_1, \sigma^2) = \frac{-p(A; D_1, D_2, \sigma^2)}{2A^2\sigma^2},$$

and

$$\frac{\partial^2}{\partial A^2} c(A; D_2) b_2(A; D_1, \sigma^2) = \frac{-1}{2\sigma^2}\left(1 - \frac{D_2(\sigma^2 - D_1)^2}{A^3}\right). \tag{A.9}$$

Since $A \geq D_2$ and $A \geq \sigma^2 + D_1 > \sqrt{|\sigma^2 - D_1|}$, the RHS of (A.9) is negative and hence $c(A; D_2) b_2(A; D_1, \sigma^2)$ is strictly concave in A. Thus, there can be at most one local extremum in $\mathcal{A}(D_1, D_2, \sigma^2)$, which would also be the maximizing value. There is exactly one local extremum since there exists an $A^* \in \mathcal{A}(D_1, D_2, \sigma^2)$ such that $p(A^*; D_1, D_2, \sigma^2) = 0$. This follows since $p(A; D_1, D_2, \sigma^2)$ is continuous in A; since

$$p(\sigma^2 + D_1; D_1, D_2, \sigma^2) = -2D_1 D_2 \sigma^2 < 0;$$

since if $\sigma^2 + D_1 < D_2 < \left(\sigma + \sqrt{D_1}\right)^2$, then

$$p(D_2; D_1, D_2, \sigma^2) = \frac{D_2}{2}\left(\left(D_2 - (\sigma^2 + D_1)\right)^2 - 4\sigma^2 D_1\right) < 0,$$

where the inequality follows by the above assumption; and since if $D_2 < (\sigma + \sqrt{D_1})^2$, then

$$\begin{aligned} p\left(\left(\sigma+\sqrt{D_1}\right)^2; D_1, D_2, \sigma^2\right) &= 2\sqrt{\sigma^2 D_1}\left(\left(\sigma+\sqrt{D_1}\right)^4 - \frac{D_2}{2}\left(\sigma^2 + D_1 + 4\sqrt{\sigma^2 D_1}\right)\right) \\ &> \sqrt{\sigma^2 D_1}\left(\sigma+\sqrt{D_1}\right)^2\left(\sigma-\sqrt{D_1}\right)^2 \\ &> 0, \end{aligned}$$

where the first inequality follows from the above assumption. □

Appendix B

Technical Proofs

In this appendix, we prove many of the technical claims that have been given throughout the thesis. In each section, we repeat the statement of the theorem or lemma to be proved followed by the proof. A reference to the theorem or lemma being proved is given in the title of each section as well as in parenthesis at the beginning of each restatement.

B.1 Proof of Theorem 2.3

Theorem B.1 (2.3). *For the watermarking game with real alphabets and squared error distortion, 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 (2.14), (2.15) are in effect instead of the a.s. distortion constraints (2.1), (2.3), then no rate is achievable in either version of the game.*

Proof. For a given covertext $\{\boldsymbol{U}\}$ and for a given encoder sequence $\{f_n\}$, let the average power in the stegotext be given by

$$\tilde{a}_n = E\left[\|\boldsymbol{X}\|^2\right]/n = E\left[\|f_n(\boldsymbol{U}, W, \Theta_1)\|^2\right]/n. \tag{B.1}$$

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 (2.16) that for any $\epsilon > 0$ and any integer $n_0 > 0$ there exists some $n^* > n_0$ such that

$$\tilde{a}_{n^*} \le A_{\max}, \tag{B.2}$$

where $A_{\max} = (\sigma + \epsilon + \sqrt{D_1})^2$. 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

$$p = \min\left\{\frac{D_2}{A_{\max}}, 1\right\}. \tag{B.3}$$

For the blocklength n^* consider now the attacker

$$\tilde{g}_{n^*}(\boldsymbol{x}, \theta_2) = \theta_2 \boldsymbol{x} \tag{B.4}$$

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

$$\begin{aligned}\frac{1}{n^*} E\left[\|\boldsymbol{X} - \tilde{g}_{n^*}(\boldsymbol{X}, \Theta_2)\|^2\right] &= p \cdot \frac{1}{n^*} E\left[\|\boldsymbol{X}\|^2\right] \\ &\leq p \cdot A_{\max} \\ &\leq D_2,\end{aligned}$$

where the equality follows from (B.4), the subsequent inequality by (B.1) and (B.2), and the last inequality by (B.3). □

B.2 Proof of Lemma 2.1

Lemma B.1 (2.1). *For the communication with side information model with finite alphabets, if the side information is available non-causally to the encoder only and the encoder is required to satisfy* $\frac{1}{n}\sum_{i=1}^{n} d_1(u_i, x_i) \leq D_1$, *a.s., for some non-negative function* $d_1(\cdot,\cdot)$. *Then, the capacity is given by*

$$C_{\text{pub}}^{\text{NCSI}}(D_1) = \max_{\substack{P_{V|U},\, f:\mathcal{V}\times\mathcal{U}\mapsto\mathcal{X}, \\ E[d_1(U,X)]\leq D_1}} I(V;Y) - I(V;U), \tag{B.5}$$

where V *is an auxiliary random variable with finite alphabet, and the mutual informations are evaluated with respect to the joint distribution (2.30).*

Proof. The achievability part follows directly from the proof of Gel'fand and Pinsker [GP80]. We simply choose a $P_{V|U}$ and a function $f : \mathcal{V} \times \mathcal{U} \mapsto \mathcal{X}$ such that $E[d_1(U, X)] \leq \tilde{D}_1 < D_1$. We then use the same coding strategy as in [GP80]. The distortion between the side information and the transmitted sequence will be approximately $\tilde{D}_1$. By choosing n large enough, we can ensure that this distortion exceed D_1 with arbitrarily small probability. The achievability proof is completed by noting that $C_{\text{pub}}^{\text{NCSI}}(D_1)$ is continuous in D_1. Furthermore, it is non-decreasing and convex in D_1; see [BCW00]. Combining the converse of Gel'fand and Pinsker [GP80] with the usual converse for channels with input constraints (see e.g., [CT91][Sect. 10.2]), we can show that no rates greater than

$$\max_{\substack{P_{V|U},\, P_{X|V,U} \\ E[d_1(U,X)] \leq D_1}} I(V;Y) - I(V;U) \tag{B.6}$$

are achievable. Thus, we only need to show that (B.6) is equal to the RHS of (B.5), the proposed capacity expression. Gel'fand and Pinsker [GP80] showed this equivalence without the distortion constraint. However, their proof does not carry through to this case. Their basic idea is that $I(V;Y) - I(V;U)$ is convex in $P_{X|V,U}$ for all other distributions fixed. Thus, a general $P_{X|V,U}$, which can be written as a convex combination of deterministic $P_{X|V,U}$'s, will always be dominated by a deterministic $P_{X|V,U}$. However, a general $P_{X|V,U}$ that satisfies the distortion constraint might not be a convex combination of deterministic $P_{X|V,U}$'s that also satisfy the distortion constraint.

We now prove that (B.6) is equal to the RHS of (B.5). We make the assumption that $\mathcal{X}$, $\mathcal{U}$ and $\mathcal{Y}$ are finite. We also assume that $\mathcal{V}$ is finite, which Gel'fand and Pinsker show to be sufficient (this does not change with the distortion constraint).

Without loss of generality, there exists $v_0 \in \mathcal{V}$ such that $0 < P_{X|V,U}(x|v_0, u) < 1$ for some x and s. If no such v_0 existed, then there would be nothing to prove. Let us choose functions $f_1, \ldots, f_n : \mathcal{U} \mapsto \mathcal{X}$ and positive constants $c_1, \cdots, c_n$ with $\sum_i c_i = 1$ such that

$$P_{X|V,U}(x|v_0, u) = \sum_{i=1}^{n} c_i 1_{\{x = f_i(u)\}}, \quad \forall x \in \mathcal{X},\ s \in \mathcal{U}. \tag{B.7}$$

Let $\mathcal{V}' = \{v_1', \ldots, v_n'\}$ and $\tilde{\mathcal{V}} = \mathcal{V}' \cup \mathcal{V} \backslash \{v_0\}$. Let the random variable $\tilde{V}$ take values in $\tilde{\mathcal{V}}$

and have joint distributions

$$P_{X|\tilde{V},U}(x|\tilde{v},u) = \begin{cases} P_{X|V,U}(x|\tilde{v},u) & \text{if } \tilde{v} \in \mathcal{V}\backslash\{v_0\} \\ 1_{\{x=f_i(u)\}} & \text{if } \tilde{v} = v_i' \in \mathcal{V}' \end{cases}, \tag{B.8}$$

and

$$P_{\tilde{V}|U}(\tilde{v}|u) = \begin{cases} P_{V|U}(\tilde{v}|u) & \text{if } \tilde{v} \in \mathcal{V}\backslash\{v_0\} \\ c_i P_{V|U}(v_0|u) & \text{if } \tilde{v} = v_i' \in \mathcal{V}' \end{cases}. \tag{B.9}$$

We now compare the original joint distribution between V, U, X and Y with the new joint distribution $\tilde{V}$, U, X and Y. We claim that $I(\tilde{V};Y) - I(\tilde{V};U) \geq I(V;Y) - I(V;U)$ and the expected distortion is the same under both distributions, which will complete the proof of the claim. This follows since we can repeat this process until there is no such v_0.

We first note that the joint distribution on U, X, and Y is the same under both distributions. That is,

$$\sum_{\tilde{v}\in\tilde{\mathcal{V}}} P_{Y|X,U}(y|x,u) P_{X|\tilde{V},U}(x|\tilde{v},u) P_{\tilde{V}|U}(\tilde{v}|u) P_U(u)$$
$$= \sum_{v\in\mathcal{V}} P_{Y|X,U}(y|x,u) P_{X|V,U}(x|v,u) P_{V|U}(v|u) P_U(u).$$

In particular, both $H(Y)$ and $E[d(U,X)]$ are unaffected by the change in distribution. Second, we consider a joint distribution between V, $\tilde{V}$ and U defined by

$$P_{\tilde{V}|V,U}(\tilde{v}|v,u) = \begin{cases} 1_{\{\tilde{v}=v\}} & \text{if } \tilde{v} \in \mathcal{V}\backslash\{v_0\} \\ c_i 1_{\{v=v_0\}} & \text{if } \tilde{v} = v_i' \in \mathcal{V}' \end{cases},$$

which is consistent with both joint distributions. Under this distribution, the random variables U, V and $\tilde{V}$ form a Markov chain. Thus, by the data processing inequality, $I(V;U) \geq I(\tilde{V};U)$. We finally note that

$$P_{Y|V}(y|v_0) = \sum_{i=1}^{n} c_i P_{Y|\tilde{V}}(y|v_i'),$$

which follows by the definitions (B.7), (B.8) and (B.9). We can thus show, using the

concavity of entropy, that $H(Y|\tilde{V}) \leq H(Y|V)$ and thus $I(\tilde{V};Y) \geq I(V;Y)$ (since $H(Y)$ is the same for both distributions). These three observations finish the proof of the claim. □

B.3 Proof of Lemma 3.1

Lemma B.2 (3.1). *For any $n > 0$ and any covertext distribution $P_{\boldsymbol{U}}$,*

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

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

Proof. We first show following Chen [Che00] 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 claim is trivial. We can write that

$$\begin{aligned}
I_{\text{priv}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{Y}|\boldsymbol{X}}) &= n^{-1} I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{U}) \\
&\geq n^{-1} I(\boldsymbol{V};\boldsymbol{Y}|\boldsymbol{U}) && \text{(B.10)} \\
&= n^{-1}\big(I(\boldsymbol{V};\boldsymbol{U},\boldsymbol{Y}) - I(\boldsymbol{V};\boldsymbol{U})\big) && \text{(B.11)} \\
&\geq n^{-1}\big(I(\boldsymbol{V};\boldsymbol{Y}) - I(\boldsymbol{V};\boldsymbol{U})\big) \\
&= I_{\text{pub}}(P_{\boldsymbol{U}}, P_{\boldsymbol{X}|\boldsymbol{U}}, P_{\boldsymbol{V}|\boldsymbol{U},\boldsymbol{X}}, P_{\boldsymbol{Y}|\boldsymbol{X}})
\end{aligned}$$

where (B.10) follows by the data processing inequality (see e.g. [CT91]) because $\boldsymbol{V}$ and $\boldsymbol{Y}$ are conditionally independent given $(\boldsymbol{X},\boldsymbol{U})$; and where (B.11) 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 (3.6). 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. □

B.4 Proof of Lemma 3.2

Lemma B.3 (3.2). *For any $n > 0$, any covertext distribution $P_{\boldsymbol{U}}$, any watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$, and any fixed distortion $D_2 > A_n$*

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

where $\sigma^2_{u,n} = E_{P_{\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{U}\|^2\right]$; $D_{1,n} = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}[n^{-1}\|\boldsymbol{X}-\boldsymbol{U}\|^2]$; $A_n = E_{P_{\boldsymbol{U}}P_{\boldsymbol{X}|\boldsymbol{U}}}\left[n^{-1}\|\boldsymbol{X}\|^2\right]$; P^G_U denotes a zero-mean Gaussian distribution of variance $\sigma^2_{u,n}$; $P^{A_n}_{X|U}$ is the watermarking channel described in Section 3.2.2 for the parameters $\sigma^2_{u,n}$, $D_{1,n}$ and A_n; and $P^{A_n}_{Y|X}$ is the attack channel described in Section 3.2.1 for the parameters D_2 and A_n.

Proof. The proof is organized as follows. In Lemma B.4, 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 B.4. *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 3.2.1 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 $c = c(A; D_2)$ (defined in (A.4)) and S_2 is mean-zero variance-cD_2 Gaussian random variable independent of X. Thus,

$$h_{P_U P_{X|U} P^A_{Y|X}}(Y|X) = h(S_2) = h_{P^*_U P^*_{X|U} P^A_{Y|X}}(Y|X), \tag{B.12}$$

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,

$$\begin{aligned}
h_{P_U P_{X|U} P^A_{Y|X}}(Y|U) &= h_{P_U P_{X|U} P^A_{Y|X}}(Y - \beta U|U) \\
&\leq h_{P_U P_{X|U} P^A_{Y|X}}(Y - \beta U) \\
&\leq \frac{1}{2}\log\left(2\pi e E_{P_U P_{X|U} P^A_{Y|X}}\left[(Y - \beta U)^2\right]\right) \\
&= \frac{1}{2}\log\left(2\pi e E_{P^*_U P^*_{X|U} P^A_{Y|X}}\left[(Y - \beta U)^2\right]\right) \\
&= h_{P^*_U P^*_{X|U} P^A_{Y|X}}(Y|U),
\end{aligned} \tag{B.13}$$

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 (B.13) 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 (B.12) and (B.13) with the definition of I_{priv} (see (3.1)) completes the proof of Lemma B.4. □

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

$$I_{\text{priv}}(P^*_U, P^*_{X|U}, P^A_{Y|X}) = \frac{1}{2}\log\left(1 + \frac{c(A; D_2)\, b_2\left(E\left[X^2\right]; E\left[(X-U)^2\right], E\left[U^2\right]\right)}{D_2}\right), \tag{B.14}$$

where $b_2(\cdot;\cdot,\cdot)$ is defined in (A.3) and $A > D_2$. Note that b_2 and hence the whole expression (B.14) is concave in the triple $(E\left[U^2\right], E\left[(X-U)^2\right], E\left[X^2\right])$, as can be verified by

checking that the Hessian is non-negative definite. We can now compute that

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

where the first inequality follows by the chain rule and since conditioning reduces entropy, the second inequality follows by Lemma B.4 and by (B.14), the third inequality follows by the above discussed concavity of (B.14), and the final equality follows by the definition of $s(\cdot;\cdot,\cdot,\cdot)$ (A.5). 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. □

B.5 Proof of Lemma 3.3

Lemma B.5 (3.3). *Consider an IID zero-mean variance-σ_u^2 Gaussian covertext (denoted $(P_U^G)^n$) and fixed distortions D_1 and D_2. If $P_{\boldsymbol{Y}|\boldsymbol{X}}$ satisfies $E_{(P_U^G P_{X|U}^A)^n P_{\boldsymbol{Y}|\boldsymbol{X}}}\left[n^{-1}\|\boldsymbol{Y}-\boldsymbol{X}\|\right] \leq D_2$, then for all $A \in \mathcal{A}(D_1, D_2, \sigma_u^2)$,*

$$\begin{aligned}
I_{\text{pub}}\left((P_U^G)^n, (P_{X|U}^A)^n, (P_{V|U,X}^A)^n, P_{\boldsymbol{Y}|\boldsymbol{X}}\right) &\geq I_{\text{pub}}\left((P_U^G)^n, (P_{X|U}^A)^n, (P_{V|U,X}^A)^n, (P_{Y|X}^A)^n\right) \\
&= \frac{1}{2}\log\bigl(1+s(A; D_1, D_2, \sigma_u^2)\bigr). \qquad \text{(B.15)}
\end{aligned}$$

Here, $P_{X|U}^A$ and $P_{V|U,X}^A$ are the watermarking channels described in Section 3.2.2 for the parameters σ_u^2, D_1 and A and $P_{Y|X}^A$ is the attack channel described in Section 3.2.1 for the parameters D_2 and A.

Proof. 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 3.2.2

$$M(D_2, A) = \inf_{P_{Y|X} \in \mathcal{D}_2(D_2, P_U^G, P_{X|U}^A)} I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}\right). \tag{B.16}$$

In Lemma B.6, 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 (B.15). In Lemma B.7, 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 claim to compute $M(D_2, A)$.

Lemma B.6. *For a fixed A, the function $M(D_2, A)$ defined in (B.16) 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 the random variables V, X and Y form a Markov chain.

Given the parameters A, D_r, D_s, and $\epsilon > 0$, let the watermarking channels $P_{Y|X}^r \in \mathcal{D}_2(D_r, P_U^G, P_{X|U}^A)$ and $P_{Y|X}^s \in \mathcal{D}_2(D_s, P_U^G, P_{X|U}^A)$ be such that

$$I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^r\right) \leq M(D_r, A) + \epsilon, \tag{B.17}$$

and

$$I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^s\right) \leq M(D_s, A) + \epsilon. \tag{B.18}$$

For any $0 \leq \lambda \leq 1$, let $P_{Y|X}^\lambda = \lambda P_{Y|X}^r + \bar{\lambda} P_{Y|X}^s$, where $\bar{\lambda} = (1 - \lambda)$. We complete the proof

with

$$\begin{aligned} M(\lambda D_r + \bar{\lambda} D_s, A) &\leq I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^{\lambda}\right) \\ &\leq \lambda I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^{r}\right) + \bar{\lambda} I_{\text{pub}}\left(P_U^G, P_{X|U}^A, P_{V|U,X}^A, P_{Y|X}^{s}\right) \\ &\leq \lambda M(D_r, A) + \bar{\lambda} M(D_s, A) + \epsilon, \end{aligned}$$

where the first inequality follows since $E_{P_U P_{X|U}^A P_{Y|X}^{\lambda}}[(X-Y)^2] \leq \lambda D_r + \bar{\lambda} D_s$, 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 (B.17) and (B.18). The claim follows since ϵ is an arbitrary positive number. □

We continue with the proof of Lemma 3.3 by demonstrating that $M(D_2, A)$ is a lower bound on the LHS of (B.15). Indeed, if

$$P_{\boldsymbol{Y}|\boldsymbol{X}} \in \mathcal{D}_2(D_2, (P_U^G)^n, (P_{X|U}^A)^n), \tag{B.19}$$

then

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

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 B.6 and by (B.19) 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 $P_{Y|X}^A$ of Section 3.2.1. To do so, we first show in Lemma B.7 that we only need to consider conditional distributions $P_{Y|X}$ under which V and Y are jointly Gaussian. A similar lemma was given in a preliminary version of [SVZ98] and in [MS01], but neither proof is as general as the one below.

Lemma B.7. *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 $I(V;Y) \geq I(V;Z)$.*

Proof. It suffices to prove the claim 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

$$I(V;Y) < \infty. \tag{B.20}$$

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

$$\begin{aligned}
I(V;Y) &= h(V) - h(V|Y) && \text{(B.21)}\\
&\geq h(V) - h(V - \beta Y) && \text{(B.22)}\\
&\geq h(V) - \frac{1}{2}\log 2\pi e E[(V - \beta Y)^2] && \text{(B.23)}\\
&= h(V) - \frac{1}{2}\log 2\pi e E[(V - \beta Z)^2] && \text{(B.24)}\\
&= I(V;Z) && \text{(B.25)}\\
&= \frac{1}{2}\log\left(\frac{E[V^2]E[Z^2]}{|K_{VZ}|}\right) && \text{(B.26)}
\end{aligned}$$

and the claim is proved. Here, (B.21) follows since we have assumed that a conditional density exists; (B.22) follows since conditioning reduces entropy; (B.23) follows since a Gaussian maximizes differential entropy subject to a second moment constraint; (B.24) follows since $K_{VY} = K_{VZ}$ and hence all second order moments are the same; (B.25) follows since $V - \beta Z$ is both Gaussian and independent of Z; and (B.26) follows since V and Z are zero-mean jointly Gaussian random variables.

By (B.20) the conditional density $f_{V|Y}$ exists if Y takes on a countable number of values. This follows since (B.20) 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 poof of the claim, 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 $q_\Delta : \mathbb{R} \mapsto \{\dots, -2\Delta, -\Delta, 0, \Delta, 2\Delta, \dots\}$ 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,

$$I(V;Y) \geq I(V;Y_\Delta). \tag{B.27}$$

The random variable Y_Δ takes on a countable number of values and by (B.20) and (B.27), $I(V;Y_\Delta) < \infty$. Thus, the conditional density $f_{V|Y_\Delta}$ exists and

$$I(V;Y_\Delta) \geq \frac{1}{2} \log \left(\frac{E[V^2]E[Y_\Delta^2]}{|K_{VY_\Delta}|} \right). \tag{B.28}$$

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 (B.27) and (B.28) hold for all $\Delta > 0$, the claim follows by letting Δ approach zero. $\square$

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 claim 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 $Y = c_1 X + S_1$, 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 $c_1 X + S_1$, where S_1 is a random variable independent of X, which is Gaussian of zero mean and variance c_2. Under this distribution,

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

We require that $P_{Y|X}^{c_1,c_2} \in \mathcal{D}_2(D_2, P_U^G, P_{X|U}^A)$, and thus

$$\frac{c_2}{c_1^2} \leq \frac{D_2}{c(A; D_2)}, \tag{B.29}$$

where equality is achieved by $c_1 = c(A; D_2)$ and $c_2 = c(A; D_2)D_2$, and where $c(\cdot;\cdot)$ is defined in (A.4). Thus, if $\alpha = \alpha(A; D_1, D_2, \sigma_u^2)$, $\rho = \rho(A; D_1, \sigma_u^2)$ and $b_1 = b_1(A; D_1, \sigma_u^2)$ (see Appendix A), 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 (B.29). 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. Indeed, the choice of the memoryless watermarking channels $(P_{X|U}^A)^n$ and $(P_{V|U,X}^A)^n$ guarantees a mutual information of at least $\frac{1}{2}\log(1 + s(A; D_1, D_2, \sigma_u^2))$. Furthermore, when these watermarking channels are used, the memoryless attack channel $(P_{Y|X}^A)^n$ is optimal. □

B.6 Proof of Lemma 4.3

Lemma B.8 (4.3). *For any $\epsilon > 0$ and $\epsilon_1 > 0$, there exists an integer $n_2 > 0$ such that for all $n > n_2$, $\Pr\left(\beta_1(Z_1, Z_2) < \beta_1^* - \epsilon_1\right) < \epsilon$.*

Proof. Recall that the attacker has the form given in Section 4.1.2 and that the random

variables Z_1 and Z_2 are defined in (4.15) and (4.16). Thus,

$$\begin{aligned} Z_1 &= \frac{1}{n}\left\|\boldsymbol{Y}|_{\boldsymbol{U}^\perp}\right\|^2 \\ &= \frac{1}{n}\left\|\big(\gamma_1(\boldsymbol{X})\boldsymbol{X}+\boldsymbol{\gamma}_2(\boldsymbol{X})\big)\big|_{\boldsymbol{U}^\perp}\right\|^2 \\ &= \frac{1}{n}\left\|\gamma_1(\boldsymbol{X})\boldsymbol{C}_W(\boldsymbol{U})+\boldsymbol{\gamma}_2(\boldsymbol{X})|_{\boldsymbol{U}^\perp}\right\|^2 \\ &\leq \gamma_1^2(\boldsymbol{X})b_2+\gamma_3(\boldsymbol{X})+2\gamma_1(\boldsymbol{X})n^{-1}\big\langle\boldsymbol{\gamma}_2(\boldsymbol{X})|_{\boldsymbol{U}^\perp},\boldsymbol{C}_W(\boldsymbol{U})\big\rangle \\ &= \gamma_1^2(\boldsymbol{X})b_2+\gamma_3(\boldsymbol{X})+2\gamma_1(\boldsymbol{X})n^{-1}\big\langle\boldsymbol{\gamma}_2(\boldsymbol{X}),\boldsymbol{C}_W(\boldsymbol{U})\big\rangle, \end{aligned} \tag{B.30}$$

where the first equality follows from the definition of Z_1 (4.15); the second equality from the representation of the forgery as in (4.4); the third equality from the structure of the encoder (4.11); the subsequent inequality from (4.8), the bound $\left\|\boldsymbol{\gamma}_2(\boldsymbol{X})|_{\boldsymbol{U}^\perp}\right\|^2 \leq \left\|\boldsymbol{\gamma}_2(\boldsymbol{X})\right\|^2$ and (4.5); and the final equality because $\boldsymbol{C}_W(\boldsymbol{U}) \in \boldsymbol{U}^\perp$ (4.9).

Similarly, we can show that

$$Z_2 = \gamma_1(\boldsymbol{X})b_2 + n^{-1}\langle\boldsymbol{\gamma}_2(\boldsymbol{X}),\boldsymbol{C}_W(\boldsymbol{U})\rangle. \tag{B.31}$$

We now argue that the sequence of random variables $n^{-1}\langle\boldsymbol{\gamma}_2(\boldsymbol{X}),\boldsymbol{C}_W(\boldsymbol{U})\rangle$ approaches, as n tends to infinity, zero in probability uniformly over all attackers. First, note that given the stegotext $\boldsymbol{X}=\boldsymbol{x}$, the random vector $\boldsymbol{C}_W(\boldsymbol{U})$ is distributed like $b_2\boldsymbol{x}/A+\boldsymbol{C}$, where $\boldsymbol{C}$ is uniformly distributed on $S^n(0,\sqrt{nb_3})$ and $b_3 = b_2(A-b_2)/A$. Consequently, for any $0<\zeta<\sqrt{D_2b_3}$,

$$\begin{aligned} &\Pr\left(\left|n^{-1}\left\langle\boldsymbol{\gamma}_2(\boldsymbol{X}),\boldsymbol{C}_W(\boldsymbol{U})\right\rangle\right|>\zeta\,\middle|\,\boldsymbol{X}=\boldsymbol{x}\right) \\ &\quad= \Pr\left(\left|\left\langle\boldsymbol{\gamma}_2(\boldsymbol{x})/\sqrt{n\gamma_3(\boldsymbol{x})},\boldsymbol{C}/\sqrt{nb_3}\right\rangle\right|>\zeta/\sqrt{\gamma_3(\boldsymbol{x})b_3}\right) \\ &\quad\leq \Pr\left(\left|\left\langle\boldsymbol{\gamma}_2(\boldsymbol{x})/\sqrt{n\gamma_3(\boldsymbol{x})},\boldsymbol{C}/\sqrt{nb_3}\right\rangle\right|>\zeta/\sqrt{D_2b_3}\right) \\ &\quad= \frac{2C_{n-1}\left(\arccos\left(\zeta/\sqrt{D_2b_3}\right)\right)}{C_{n-1}(\pi)}. \end{aligned}$$

Here the first equality follows by the conditional distribution of $\boldsymbol{C}_W(\boldsymbol{U})$ and the fact that $\langle\boldsymbol{\gamma}_2(\boldsymbol{x}),\boldsymbol{x}\rangle = 0$, the subsequent inequality follows from $\gamma_3(\boldsymbol{x}) \leq D_2$ (see (4.5)), and the final equality follows since $\boldsymbol{C}/\sqrt{nb_3}$ is uniformly distributed on $S^n(0,1)\cap\boldsymbol{x}^\perp$ 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 $\boldsymbol{x}$, it must also hold for the unconditional probability.

Combining this fact with (B.30) and (B.31), we see that for any $\epsilon_2 > 0$ there exists some n_2 such that

$$\Pr\big(Z_1 \leq \gamma_1^2(\boldsymbol{X})b_2 + \gamma_3(\boldsymbol{X}) + \epsilon_2 \text{ and } Z_2 \geq \gamma_1(\boldsymbol{X})b_2 - \epsilon_2\big) \geq 1 - \epsilon \tag{B.32}$$

for all $n > n_2$.

Since $\beta_1(z_1, z_2)$ is non-increasing in z_1 and non-decreasing in z_2, it follows that (B.32) implies

$$\Pr\left(\beta_1(Z_1, Z_2) \geq \frac{\gamma_1(\boldsymbol{X})b_2 - \epsilon_2}{\sqrt{b_2(\gamma_1^2(\boldsymbol{X})b_2 + \gamma_3(\boldsymbol{X}) + \epsilon_2)}}\right) \geq 1 - \epsilon \tag{B.33}$$

for all $n > n_2$. Since $n^{-1}\|\boldsymbol{X}\|^2 = A$ (4.12), it follows from (4.6) that with probability one $\gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X}) \leq D_2/c$ so that

$$\frac{\gamma_1(\boldsymbol{X})b_2}{\sqrt{b_2(\gamma_1^2(\boldsymbol{X})b_2 + \gamma_3(\boldsymbol{X}))}} = \sqrt{\frac{b_2}{b_2 + \gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X})}} \geq \beta_1^*. \tag{B.34}$$

Thus, we can choose ϵ_2 small enough (and the corresponding n_2 large enough) so that (B.32) will imply via (B.33) and (B.34) that $\Pr\left(\beta_1(Z_1, Z_2) \geq \beta_1^* - \epsilon_1\right) \geq 1 - \epsilon$, for all $n > n_2$. □

B.7 Proof of Lemma 4.6

Lemma B.9 (4.6). *Given $\boldsymbol{X} = \boldsymbol{x}$ and $Z = z$, the random vector $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the set $\mathcal{V}(\boldsymbol{x}, z) = \left\{a_1\boldsymbol{x} + \boldsymbol{v} : \boldsymbol{v} \in \mathcal{S}^n(0, \sqrt{na_2}) \cap \boldsymbol{x}^\perp\right\}$, where $a_1 = \frac{\sigma_v^2 + (1-\alpha)z}{n^{-1}\|\boldsymbol{x}\|^2}$, and $a_2 = \frac{(1-\alpha)^2(\sigma_u^2\sigma_v^2 - z^2)}{n^{-1}\|\boldsymbol{x}\|^2}$.*

Proof. Conditional on the covertext $\boldsymbol{U} = \boldsymbol{u}$ and on $Z = z$, the auxiliary codeword $\boldsymbol{V}_W(\boldsymbol{U})$ is uniformly distributed over the set

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

as follows by the definition of Z (4.40) and the distribution of the codebook $\{\boldsymbol{V}_{j,k}\}$. Using

the deterministic relation (4.37) we can now relate the appropriate conditional densities as

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

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

$$n^{-1}\|\boldsymbol{x}\|^2=\sigma_v^2+2(1-\alpha)z+(1-\alpha)^2\sigma_u^2. \tag{B.35}$$

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 (B.35), and thus $\boldsymbol{v}\in\mathcal{V}(\boldsymbol{x},z)$. □

B.8 Proof of Lemma 4.8

Lemma B.10 (4.8). *If the constants defined for the additive attack watermarking game are used to design the sequence of encoders of Section 4.3.1, then for any $\epsilon>0$ and $\epsilon_2>0$, there exists an integer $n_2>0$ such that for all $n>n_2$ and for all the deterministic attacks of Section 4.1.1,* $\Pr\big(\beta_2(Z,Z_3,Z_4)<\beta^*(R_1+\delta)-\epsilon_2\big)<\epsilon$.

Proof. Recall that a deterministic attacker of Section 4.1.1 is specified by a vector $\tilde{\boldsymbol{y}}$ satis-

fying (4.3). Fix some $\epsilon_3 > 0$ (to be chosen later) and choose n_2 large enough to ensure

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \epsilon, \quad \forall n > n_2, \tag{B.36}$$

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

$$\mathcal{E}_1 = \left\{ \left| 2n^{-1} \langle \boldsymbol{X}, \tilde{\boldsymbol{y}} \rangle \right| \leq \epsilon_3 \right\},$$

$$\mathcal{E}_2 = \left\{ \left| n^{-1} \langle \boldsymbol{V}_W(\boldsymbol{U}), \tilde{\boldsymbol{y}} \rangle \right| \leq \epsilon_3 \right\},$$

and

$$\mathcal{E}_3 = \{ Z \geq \alpha \sigma_u^2 \}.$$

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 (4.3). 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 4.5) and since $\boldsymbol{X} = \boldsymbol{V} + (1-\alpha)\boldsymbol{U}$, and $\mathcal{E}_3^c$ has vanishing probability by Lemma 4.7.

Event $\mathcal{E}_1$ guarantees that

$$\begin{aligned} Z_3 &= \frac{1}{n}\|\boldsymbol{X}\|^2 + \frac{2}{n}\langle \boldsymbol{X}, \tilde{\boldsymbol{y}} \rangle + \frac{1}{n}\|\tilde{\boldsymbol{y}}\|^2 \\ &\leq \sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2, \end{aligned} \tag{B.37}$$

where the equality follows by the definition of Z_3 (4.41) and the form of the attacker given in Section 4.1.1, and where the inequality follows by (B.35), (4.3), and the inequality defining $\mathcal{E}_1$.

From the definition of Z_4 (4.43) it follows that $\mathcal{E}_2$ guarantees that $Z_4 \geq -\epsilon_3$. Consequently, the intersection $\mathcal{E}_1 \cap \mathcal{E}_2$ guarantees that

$$\beta_2(Z, Z_3, Z_4) \geq \frac{\sigma_v^2 + (1-\alpha)Z - \epsilon_3}{\sqrt{\sigma_v^2 + 2(1-\alpha)Z + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2}}. \tag{B.38}$$

For any $\epsilon_3 > 0$, the RHS of (B.38) is monotonically increasing in Z, so that the inter-

section $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ implies

$$\beta_2(Z, Z_3, Z_4) \geq \frac{\sigma_v^2 + (1-\alpha)\alpha\sigma_u^2 - \epsilon_3}{\sqrt{\sigma_v^2 + 2(1-\alpha)\alpha\sigma_u^2 + (1-\alpha)^2\sigma_u^2 + \epsilon_3 + D_2}}. \tag{B.39}$$

Recalling the definitions in Section 4.3.1 and the definition of $\beta^*(R_1 + \delta)$ (4.49), one can show using some algebra that for $\epsilon_3 = 0$, the RHS of (B.39) equals $\beta^*(R_1 + \delta)$. Since the RHS of (B.39) 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 $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ will guarantee that

$$\beta_2(Z, Z_3, Z_4) \geq \beta^*(R_1 + \delta) - \epsilon_2.$$

The claim thus follows from (B.36). □

B.9 Proof of Lemma 4.10

Lemma B.11 (4.10). *If the constants defined for the general watermarking game are used to design the sequence of encoders of Section 4.3.1, then for any $\epsilon > 0$ and $\epsilon_2 > 0$, there exists an integer $n_2 > 0$ such that for all $n > n_2$ and for all attackers of Section 4.1.2,* $\Pr\big(\beta_2(Z, Z_3, Z_4) < \beta^*(R_1 + \delta) - \epsilon_2\big) < \epsilon$.

Proof. In order to prove the desired result, we need the following technical claim.

Lemma B.12. *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 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 4.6 distributed like $a_1\boldsymbol{x} + \boldsymbol{V}$, where $\boldsymbol{V}$ is uniformly distributed on $\mathcal{S}^n(0, \sqrt{na_2}) \cap \boldsymbol{x}^\perp$, and a_2 defined in (4.52) 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 4.6 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 (4.5) and (4.52)), and the final equality follows since $\boldsymbol{V}/\sqrt{na_2}$ is uniformly distributed on $\mathcal{S}^n(0,1) \cap \boldsymbol{x}^\perp$ 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 4.10. Choose n_2 large enough to ensure that

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

where $\mathcal{E}_4 = \{Z \geq \alpha\sigma_u^2 + \rho\}$ and $\mathcal{E}_5 = \left\{n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle \geq -\epsilon_2\sigma_v\left(\sqrt{A} - \sqrt{D_2}\right)\right\}$. 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 4.9 and B.12, respectively.

For the deterministic attacker of Section 4.1.2, we can express the random variables Z_3 and Z_4 of (4.41) and (4.43) as

$$Z_3 = \gamma_1^2(\boldsymbol{X})n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X}),$$

and

$$Z_4 = (\gamma_1(\boldsymbol{X}) - 1)(\sigma_v^2 + (1-\alpha)Z) + n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle.$$

Substituting these expressions in $\beta_2(Z, Z_3, Z_4)$ of (4.44) yields

$$\begin{aligned}
&\beta_2(Z, Z_3, Z_4) \\
&\quad = \frac{\sigma_v^2 + (1-\alpha)Z + (\gamma_1(\boldsymbol{X}) - 1)(\sigma_v^2 + (1-\alpha)Z) + n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle}{\sqrt{(\gamma_1^2 n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X}))\sigma_v^2}} \\
&\quad = \frac{\sigma_v^2 + (1-\alpha)Z}{\sqrt{\left(n^{-1}\|\boldsymbol{X}\|^2 + \gamma_3(\boldsymbol{X})/\gamma_1^2(\boldsymbol{X})\right)\sigma_v^2}} + \frac{n^{-1}\langle \boldsymbol{\gamma}_2(\boldsymbol{X}), \boldsymbol{V}_W(\boldsymbol{U})\rangle}{\sqrt{Z_3\sigma_v^2}}.
\end{aligned} \tag{B.40}$$

We conclude the proof by showing that the intersection $\mathcal{E}_4 \cap \mathcal{E}_5$ implies that (B.40) exceeds $\beta^*(R_1 + \delta) - \epsilon_2$.

We first focus on the second term on the RHS of (B.40). Using the expression (B.35) and the definitions of Section 4.3.1, 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 (2.3) and the triangle inequality imply

that Z_3 is at least $(\sqrt{A}-\sqrt{D_2})^2$. Thus, the intersection $\mathcal{E}_4 \cap \mathcal{E}_5$ guarantees that the second term of (B.40) is at least $-\epsilon_2$.

We now turn to the first term on the RHS of (B.40), which using (B.35) can be rewritten as

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

and which, using (4.6) and the fact that $\mathcal{E}_4$ implies $n^{-1}\|\boldsymbol{X}\|^2$ is at least A, 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 $\alpha < 1$ (see (A.6)), 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 (4.33) and $\beta^*(\cdot)$ (4.49), which yield

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

The event $\mathcal{E}_4$ thus implies that the first term on the RHS of (B.40) is at least $\beta^*(R_1+\delta)$. □

B.10 Proof of Lemma 4.12

Lemma B.13 (4.12). *For any encoder with corresponding watermarking channel $P_{\boldsymbol{X}|\boldsymbol{U}}$ satisfying (2.1), if the attacker g_n^* of (4.65) with corresponding attack channel $P^*_{\boldsymbol{Y}|\boldsymbol{X}}$ is used, then*

$$\begin{aligned}
&\frac{1}{n} I_{P_{\boldsymbol{U}} P_{\Theta_1} P_{\boldsymbol{X}|\boldsymbol{U},\Theta_1} P^*_{\boldsymbol{Y}|\boldsymbol{X}}}(\boldsymbol{X};\boldsymbol{Y}|K,\boldsymbol{U},\Theta_1) \\
&\quad\leq \sum_{k=1}^{m} \Pr(K=k)\cdot\frac{1}{2}\log\left(1+s(a_k;D_1,\tilde{D}_2,\mu_k)\right) &\text{(B.41)}\\
&\quad\leq E_K\left[C^*(D_1,\tilde{D}_2,\mu_K)\right]. &\text{(B.42)}
\end{aligned}$$

Proof. To simplify the proof of this lemma, we will use the following notation:

$$c^{(k)} = c(a_k; \tilde{D}_2), \tag{B.43}$$

$$b_1^{(k)} = b_1(a_k; D_1, \mu_k), \tag{B.44}$$

and

$$b_2^{(k)} = b_2(a_k; D_1, \mu_k), \tag{B.45}$$

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

Lemma B.14. *If the encoder satisfies the a.s. distortion constraint (2.1), 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 (4.65) produces an IID 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 (4.63) and (4.64)). The proof will be concluded once we show

$$n^{-1}E\left[\langle\boldsymbol{X},\boldsymbol{U}\rangle \mid K=k\right] \ge \frac{1}{2}(a_k + \mu_k - D_1), \tag{B.46}$$

because

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

by (B.44) and (B.45).

We verify (B.46) by noting that for every $k \geq 1$ such that $\Pr(K = k) > 0$,

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

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

$$\begin{aligned} & I(\boldsymbol{X}; \boldsymbol{Y}|K, \boldsymbol{U}, \Theta_1) \\ &= \sum_{k=0}^{m} \Pr(K = k) \cdot I(\boldsymbol{X}; \boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) \\ &= \sum_{k=1}^{m} \Pr(K = k) \cdot \big(h(\boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, K = k, \boldsymbol{U}, \Theta_1)\big), \end{aligned} \tag{B.47}$$

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 prove (B.41) we shall next verify that

$$I(\boldsymbol{X}; \boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) = h(\boldsymbol{Y}|K = k, \boldsymbol{U}, \Theta_1) - h(\boldsymbol{Y}|\boldsymbol{X}, K = k, \boldsymbol{U}, \Theta_1) \tag{B.48}$$

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 (B.48) as

$$\begin{aligned} h(\boldsymbol{Y}|K=k,\boldsymbol{U},\Theta_1) &= h\left(g_n^*(\boldsymbol{X},\Theta_2)\middle|K=k,\boldsymbol{U},\Theta_1\right) \\ &= h\left(g_n^*(\boldsymbol{X},\Theta_2)-c^{(k)}b_1^{(k)}\boldsymbol{U}\middle|K=k,\boldsymbol{U},\Theta_1\right) \\ &\le h\left(g_n^*(\boldsymbol{X},\Theta_2)-c^{(k)}b_1^{(k)}\boldsymbol{U}\middle|K=k\right) \\ &\le \frac{n}{2}\log\left(2\pi e E\left[\frac{1}{n}\left\|g_n^*(\boldsymbol{X},\Theta_2)-c^{(k)}b_1^{(k)}\boldsymbol{U}\right\|^2\middle|K=k\right]\right) \\ &\le \frac{n}{2}\log\left(2\pi e\left((c^{(k)})^2 b_2^{(k)}+c^{(k)}\tilde{D}_2\right)\right), \end{aligned} \tag{B.49}$$

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 (B.49) follows by Lemma B.14.

We can write the second term on the RHS of (B.48) as

$$\begin{aligned} h(\boldsymbol{Y}|\boldsymbol{X},K=k,\boldsymbol{U},\Theta_1) &= h\left(\sqrt{c^{(k)}}\boldsymbol{V}\middle|K\right) \\ &= \frac{n}{2}\log\left(2\pi e c^{(k)}\tilde{D}_2\right), \end{aligned} \tag{B.50}$$

for all $k \ge 1$, where (B.50) follows since $\boldsymbol{V}$ is an IID sequence of $\mathcal{N}(0,\tilde{D}_2)$ random variables independent of $(\boldsymbol{X},U,\Theta_1)$ and hence independent of K.

Combining (B.47), (B.49), and (B.50) and observing that $s(a_k;D_1,\tilde{D}_2,\mu_k) = c^{(k)}b_2^{(k)}/\tilde{D}_2$, proves (B.41). Finally, (B.42) follows from (B.41) by the definition of $C^*(D_1,D_2,\sigma_u^2)$ (A.8). □

B.11 Proof of Lemma 4.13

Lemma B.15 (4.13). *For any ergodic covertext distribution $P_{\boldsymbol{U}}$ with $E\left[U_k^4\right] < \infty$ and $E\left[U_k^2\right] \le \sigma_u^2$, there exists mappings $\delta(\epsilon,n)$ and $n_0(\epsilon)$ such that both the properties P1 and P2 stated below hold, where P1 is "For every $\epsilon > 0$, $\lim_{n\to\infty}\delta(\epsilon,n) = 0$." and P2 is "For every $\epsilon > 0$, $n > n_0(\epsilon)$, and event $\mathcal{E}$, if $E\left[n^{-1}\|\boldsymbol{U}\|^2|\mathcal{E}\right] > \sigma_u^2 + 5\epsilon$, then $\Pr(\mathcal{E}) < \delta(\epsilon,n)$."*

Proof. 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 $\Pr(\mathcal{E}) \ge \delta(\epsilon,n)$, then $E\left[n^{-1}\|\boldsymbol{U}\|^2|\mathcal{E}\right] \le$

$\sigma_u^2 + 5\epsilon$.

Let us define

$$S_{U^2,n} = \frac{1}{n}\sum_{i=1}^{n} U_i^2, \tag{B.51}$$

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

$$\mathrm{Var}(S_{U^2,n}) \leq \epsilon^2/2, \quad \forall n > n_0(\epsilon). \tag{B.52}$$

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 $n > n_0(\epsilon)$

$$\Pr(S_{U^2,n} \geq s_n) = \frac{\mathrm{Var}(S_{U^2,n})}{\epsilon^2}, \tag{B.53}$$

and

$$m_{U^2} - \epsilon \leq s_n \leq m_{U^2} + \epsilon. \tag{B.54}$$

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 (B.52).

From (B.53) it follows that the choice

$$\delta(\epsilon, n) = \Pr(S_{U^2,n} \geq s_n), \tag{B.55}$$

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 (B.55),

$$\Pr(\mathcal{E}) \geq \Pr(S_{U^2,n} \geq s_n) \tag{B.56}$$

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 (B.56). 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 (B.54). 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 (B.53) and (B.54). 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 IID 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. □

B.12 Proof of Lemma 5.6

Lemma B.16 (5.6). *There exists a positive function $f(A, D_1, \sigma^2)$ such that if the n-vectors $\boldsymbol{u}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, and the scalars D_2 and δ satisfy $\left|n^{-1}\|\boldsymbol{u}\|^2 - \sigma^2\right| < \delta$, $\left|n^{-1}\langle \boldsymbol{u}, \boldsymbol{x} - b_1\boldsymbol{u}\rangle\right| < \delta$, $\left|n^{-1}\|\boldsymbol{x} - b_1\boldsymbol{u}\|^2 - b_2\right| < \delta$, $\left|n^{-1}\langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{y}|_{\boldsymbol{x}}\rangle\right| < \delta$, $n^{-1}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq D_2 < A$, and $\delta < \frac{A}{2(1+b_1)^2}$, then*

$$\frac{n^{-1}\|\boldsymbol{y} - cb_1\boldsymbol{u}\|^2}{2c(cb_2 + D_2)} - \frac{n^{-1}\|\boldsymbol{y} - c\boldsymbol{x}\|^2}{2cD_2} > -\delta f(A, D_1, \sigma^2),$$

where $b_1 = b_1(A; D_1, \sigma^2)$, $b_2 = b_2(A; D_1, \sigma^2)$ and $c = c(A; D_2)$.

Proof. Consider the following chain of equalities and inequalities.

$$\begin{aligned} &2(cb_2 + D_2)\left(\frac{n^{-1}\|\boldsymbol{y} - cb_1\boldsymbol{u}\|^2}{2c(cb_2 + D_2)} - \frac{n^{-1}\|\boldsymbol{y} - c\boldsymbol{x}\|^2}{2cD_2}\right) \\ &\quad = -\frac{b_2}{D_2}n^{-1}\|\boldsymbol{y}\|^2 - c\left(1 + \frac{cb_2}{D_2}\right)n^{-1}\|\boldsymbol{x}\|^2 + cb_1^2 n^{-1}\|\boldsymbol{u}\|^2 + 2n^{-1}\left\langle \boldsymbol{y}, \left(1 + \frac{cb_2}{D_2}\right)\boldsymbol{x} - b_1\boldsymbol{u}\right\rangle \\ &\quad = -\frac{b_2}{D_2}n^{-1}\|\boldsymbol{y}\|^2 - c\left(1 + \frac{cb_2}{D_2}\right)n^{-1}\|\boldsymbol{x}\|^2 + cb_1^2 n^{-1}\|\boldsymbol{u}\|^2 \\ &\qquad + 2\left(1 + \frac{cb_2}{D_2} - \frac{b_1 n^{-1}\langle \boldsymbol{u}, \boldsymbol{x}\rangle}{n^{-1}\|\boldsymbol{x}\|^2}\right)n^{-1}\langle \boldsymbol{y}, \boldsymbol{x}\rangle - 2b_1 n^{-1}\langle \boldsymbol{y}|_{\boldsymbol{x}^\perp}, \boldsymbol{u}\rangle \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{b_2(1+c)}{A} - c\right) n^{-1}\|\boldsymbol{x}\|^2 + cb_1^2 n^{-1}\|\boldsymbol{u}\|^2 + 2\left(1 - \frac{b_2}{A} - \frac{b_1 n^{-1}\langle \boldsymbol{u}, \boldsymbol{x}\rangle}{n^{-1}\|\boldsymbol{x}\|^2}\right) n^{-1}\langle \boldsymbol{y}, \boldsymbol{x}\rangle \\
&\quad - \frac{b_2 n^{-1}\|\boldsymbol{y} - \boldsymbol{x}\|^2}{D_2} - 2b_1 n^{-1}\langle \boldsymbol{y}|_{\boldsymbol{x}^\perp}, \boldsymbol{u}\rangle \\
&> -\delta\left(\left|\frac{b_2(1+c)}{A} - c\right|(1+b_1)^2 + cb_1^2 + 2b_1\right) + 2\left(1 - \frac{b_2}{A} - \frac{b_1 n^{-1}\langle \boldsymbol{u}, \boldsymbol{x}\rangle}{n^{-1}\|\boldsymbol{x}\|^2}\right) n^{-1}\langle \boldsymbol{y}, \boldsymbol{x}\rangle \\
&> -\delta\left(\left|\frac{b_2(1+c)}{A} - c\right|(1+b_1)^2 + cb_1^2 + 2b_1\right. \\
&\qquad \left. + \frac{2\big((1-b_2/A)(1+b_1)^2 + b_1(1+b_1+b_1^2)\big)}{A - \delta(1+b_1)^2} n^{-1}\langle \boldsymbol{y}, \boldsymbol{x}\rangle\right),
\end{aligned}
$$

and thus,

$$
\begin{aligned}
&\frac{n^{-1}\|\boldsymbol{y} - cb_1\boldsymbol{u}\|^2}{2c(cb_2 + D_2)} - \frac{n^{-1}\|\boldsymbol{y} - c\boldsymbol{x}\|^2}{2cD_2} \\
&\quad > -\delta\left\{\left(\frac{1}{A} - \frac{1}{2b_2}\right)(1+b_1)^2 + \frac{b_1^2 + 2b_1}{2b_2} + \frac{6}{b_2}\left(\left(1 - \frac{b_2}{A}\right)(1+b_1)^2 + b_1(1+b_1+b_1^2)\right)\right\}.
\end{aligned}
$$

The first equality is simply an expansion of the terms of interest. The second equality uses the relations $\boldsymbol{y} = \boldsymbol{y}|_{\boldsymbol{x}} + \boldsymbol{y}|_{\boldsymbol{x}^\perp}$ and $\boldsymbol{y}|_{\boldsymbol{x}} = (\langle \boldsymbol{u}, \boldsymbol{x}\rangle / \|\boldsymbol{x}\|^2)\boldsymbol{x}$. The third equality uses the definition $c = 1 - D_2/A$ and the relation $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = (\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2)/2$. The first inequality uses (5.29), (5.28), (5.25), the fact that

$$
|n^{-1}\|\boldsymbol{x}\|^2 - A| < \delta(1+b_1)^2 \tag{B.57}
$$

(derived from (5.25), (5.26) and (5.27)), and the relation

$$
\left(\frac{b_2(1+c)}{A} - c\right) A + cb_1^2\sigma^2 - b_2 = 0.
$$

The second inequality uses (B.57), the fact that $|n^{-1}\langle \boldsymbol{u}, \boldsymbol{x}\rangle - (A + \sigma^2 - D_1)/2| < \delta(1 + b_1 + b_1^2)$ (derived from (5.25), (5.26) and (5.27)), and the relation

$$
1 - \frac{b_2}{A} - \frac{b_1(A + \sigma^2 - D_1)}{2A} = 0.
$$

The final inequality uses (5.30), the fact that $|n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle| < 3A$ (derived from (5.29), (5.30), (B.57) using Cauchy-Schwartz), and the facts that $c \leq 1$ and $cb_2 + D_2 \geq b_2$. □

B.13 Proof of Lemma 5.10

Lemma B.17 (5.10). *There exists a positive function $\tilde{f}(A, D_1, \tilde{D}_2, \sigma^2)$ such that if the n-vectors $\boldsymbol{s}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, and the scalars D_2 and δ satisfy $\left|n^{-1}\|\boldsymbol{s}\|^2 - v\right| < \delta$, $\left|n^{-1}\|\boldsymbol{x}\|^2 - A\right| < \delta$, $\left|n^{-1}\langle \boldsymbol{s}, \boldsymbol{x}\rangle - (\alpha-1)b_1\sigma^2 - A\right| < \delta$, $\left|n^{-1}\langle \boldsymbol{s}, \boldsymbol{y} - \boldsymbol{y}|_{\boldsymbol{x}}\rangle\right| < \delta$, $n^{-1}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \leq D_2 < A$, and $\delta < \frac{A}{2}$, then*

$$\frac{n^{-1}\|\boldsymbol{y}\|^2}{2(A - D_2)} - \frac{n^{-1}\|\boldsymbol{y} - \beta_1 \boldsymbol{s}\|^2}{2\beta_2} > -\delta \tilde{f}(A, D_1, \tilde{D}_2, \sigma^2),$$

where all of the parameters are computed with respect to A, D_1, $\tilde{D}_2$, D_2 and σ^2, i.e., $\alpha = \alpha(A; D_1, \tilde{D}_2, \sigma^2)$, $b_1 = b_1(A; D_1, \sigma^2)$, $v = v(A, D_1, \tilde{D}_2, \sigma^2)$, $\beta_1 = \beta_1(A, D_1, \tilde{D}_2, D_2, \sigma^2)$, and $\beta_2 = \beta_2(A, D_1, \tilde{D}_2, D_2, \sigma^2)$.

Proof. First, we compute that

$$\begin{aligned} n^{-1}\|\boldsymbol{y}\|^2 &= n^{-1}\|\boldsymbol{x} - \boldsymbol{y}\|^2 - n^{-1}\|\boldsymbol{x}\|^2 + 2n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \\ &< D_2 - A + \delta A + 2n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle, \end{aligned} \tag{B.58}$$

which follows by (5.54) and (5.51). Second, we compute that

$$\begin{aligned} n^{-1}\langle \boldsymbol{y}, \boldsymbol{s}\rangle &= n^{-1}\langle \boldsymbol{y}|_{\boldsymbol{x}}, \boldsymbol{s}\rangle + n^{-1}\langle \boldsymbol{y}|_{\boldsymbol{x}^\perp}, \boldsymbol{s}\rangle \\ &> \frac{n^{-1}\langle \boldsymbol{x}, \boldsymbol{s}\rangle}{n^{-1}\|\boldsymbol{x}\|^2} n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle - \delta, \end{aligned} \tag{B.59}$$

which follows by (5.53). Next, we compute that

$$\begin{aligned} n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \left((A - D_2)\frac{n^{-1}\langle \boldsymbol{x}, \boldsymbol{s}\rangle}{n^{-1}\|\boldsymbol{x}\|^2} - \beta_1 v\right) &> n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \left(\beta_1 v \left(\frac{A}{n^{-1}\|\boldsymbol{x}\|^2} - 1\right) - \delta \frac{A - D_2}{n^{-1}\|\boldsymbol{x}\|^2}\right) \\ &> -\delta n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle \left(\frac{2(\beta_1 v + A - D_2)}{A}\right) \qquad \text{(B.60)} \\ &> -6\delta(\beta_1 v + A - D_2), \qquad \text{(B.61)} \end{aligned}$$

where the first inequality follows by (5.52) and the relevant definitions; the second inequality follows by (5.51) and (5.55); and the final inequality follows since $n^{-1}\langle \boldsymbol{x}, \boldsymbol{y}\rangle < 3A$ using

Cauchy-Schwartz along with (5.54), (5.51) and (5.55). Thus,

$$
\begin{aligned}
&2(A-D_2)\beta_2\left(\frac{n^{-1}\|\boldsymbol{y}\|^2}{2(A-D_2)}-\frac{n^{-1}\|\boldsymbol{y}-\beta_1\boldsymbol{s}\|^2}{2\beta_2}\right)\\
&\quad=-\beta_1^2 v n^{-1}\|\boldsymbol{y}\|^2+2\beta_1(A-D_2)n^{-1}\langle\boldsymbol{y},\boldsymbol{s}\rangle-\beta_1^2(A-D_2)n^{-1}\|\boldsymbol{s}\|^2\\
&\quad>-\beta_1^2 v\left(D_2-A+\delta A+2n^{-1}\langle\boldsymbol{x},\boldsymbol{y}\rangle\right)+2\beta_1(A-D_2)\left(\frac{n^{-1}\langle\boldsymbol{x},\boldsymbol{s}\rangle}{n^{-1}\|\boldsymbol{x}\|^2}n^{-1}\langle\boldsymbol{x},\boldsymbol{y}\rangle-\delta\right)\\
&\qquad-\beta_1^2(A-D_2)\left(v+\delta\right)\\
&\quad=2\beta_1 n^{-1}\langle\boldsymbol{x},\boldsymbol{y}\rangle\left((A-D_2)\frac{n^{-1}\langle\boldsymbol{x},\boldsymbol{s}\rangle}{n^{-1}\|\boldsymbol{x}\|^2}-\beta_1 v\right)-\delta\left(\beta_1^2 vA+2\beta_1(A-D_2)+\beta_1^2(A-D_2)\right)\\
&\quad>-\delta\left\{12\beta_1(\beta_1 v+A-D_2)+\beta_1^2 vA+2\beta_1(A-D_2)+\beta_1^2(A-D_2)\right\},
\end{aligned}
\tag{B.62}
$$

where the first inequality follows by (5.50), (B.58) and (B.59) and the second inequality follows by (B.61). Dividing (B.62) by $2(A-D_2)\beta_2$ gives the desired result since β_1 and β_2 essentially only depend on D_2 through a $A-D_2$ term; see (5.42) and (5.43). □

Bibliography

[AD89] Rudolf Ahlswede and Gunter Dueck. Identification via channels. *IEEE Trans. Inform. Theory*, 35(1):15–29, January 1989.

[Ahl78] Rudolf Ahlswede. Elimination of correlation in random codes for arbitrarily varying channels. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 44:159–175, 1978.

[Ahl86] Rudolf Ahlswede. Arbitrarily varying channels with states sequence known to the sender. *IEEE Trans. Inform. Theory*, 32(5):621–629, September 1986.

[AP98] Ross J. Anderson and Fabien A. Petitcolas. On the limits of steganography. *IEEE Jour. on Sel. Areas in Comm.*, 16(4):463–473, May 1998.

[AV96] Venkat Anantharam and Sergio Verdú. Bits through queues. *IEEE Trans. Inform. Theory*, 42(1):4–18, January 1996.

[AW69] Rudolf Ahlswede and Jacob Wolfowitz. Correlated decoding for channels with arbitrarily varying channel probability functions. *Information and Control*, 14(5):457–473, May 1969.

[Baş83] Tamer Başar. The Gaussian test channel with an intelligent jammer. *IEEE Trans. Inform. Theory*, 29(1):152–157, January 1983.

[BBDRP99] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. Capacity of the watermark channel: How many bits can be hidden within a digital image. *Proc. SPIE Security and Watermarking of Multimedia Contents*, (3657):437–448, 1999.

[BBK01] Alexander Barg, G. R. Blakley, and G. Kabatiansky. Good digital fingerprint-

ing codes. In *Proc. of the Inter. Symposium on Info. Theory*, Washington, DC, 2001.

[BBPR98] M. Barni, F. Bartolini, A. Piva, and F. Rigacci. Statistical modelling of full-frame DCT coefficients. In *Proceedings of European Signal Processing Conference (EUSIPCO 98)*, volume 3, pages 1513–1516, Rhodes, Greece, 1998.

[BBT60] David Blackwell, Leo Breiman, and A. J. Thomasian. The capacity of certain channel classes under random coding. *Annals of Mathematical Statistics*, 31(3):558–567, September 1960.

[BCW00] Richard J. Barron, Brian Chen, and Gregory W. Wornell. The duality between information embedding and source coding with side information and some applications. Preprint, January 2000.

[BI99] Markus Breitbach and Hideki Imai. On channel capacity and modulation of watermarks in digital still images. In *Financial Cryptography*, number 1648 in Lecture Notes in Computer Science, pages 125–139, 1999.

[Bla57] Nelson M. Blachman. Communication as a game. In *IRE WESCON Convention Record*, number 2, pages 61–66, San Francisco, CA, 1957.

[BMM85] J. Martin Borden, David M. Mason, and Robert J. McEliece. Some information theoretic saddlepoints. *SIAM Journal on Control and Optimization*, 23(1):129–143, January 1985.

[BS98] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. Inform. Theory*, 44(5):1897–1905, September 1998.

[BW85] Tamer Başar and Ying-Wah Wu. A complete characterization of minimax and maximin encoder-decoder policies for communication channels with incomplete statistical description. *IEEE Trans. Inform. Theory*, 31(4):482–489, July 1985.

[Cac98] Christian Cachin. An information-theoretic model for steganography. In *Proc. of the Inter. Workshop on Info. Hiding*, number 1525 in Lecture Notes in Computer Science, pages 306–318, 1998.

[CEZ00] Gérard Cohen, Sylvia Encheva, and Gilles Zémor. Copyright protection for digital data. *IEEE Communication Letters*, 4(5):158–160, May 2000.

[CFNP00] Benny Chor, Amos Fiat, Moni Naor, and Benny Pinkas. Tracing traitors. *IEEE Trans. Inform. Theory*, 46(3):893–910, May 2000.

[Che00] Brian Chen. *Design and Analysis of Digital Watermarking, Information Embedding, and Data Hiding Systems*. PhD thesis, MIT, Cambridge, MA, 2000.

[CK81] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, Budapest, 1981.

[CKLS97] Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Proc.*, 6(12):1673–1687, December 1997.

[CN88a] Imre Csiszár and Prakash Narayan. Arbitrarily varying channels with constrained inputs and states. *IEEE Trans. Inform. Theory*, 34(1):27–34, January 1988.

[CN88b] Imre Csiszár and Prakash Narayan. The capacity of the arbitrarily varying channel revisited: Positivity, constraints. *IEEE Trans. Inform. Theory*, 34(2):181–193, March 1988.

[CN91] Imre Csiszár and Prakash Narayan. Capacity of the Gaussian arbitrarily varying channel. *IEEE Trans. Inform. Theory*, 37(1):18–26, January 1991.

[Cos83] Max H. M. Costa. Writing on dirty paper. *IEEE Trans. Inform. Theory*, 29(3):439–441, May 1983.

[Cov75] Thomas M. Cover. An achievable rate region for the broadcast channel. *IEEE Trans. Inform. Theory*, 21(4):399–404, July 1975.

[Cov99] Thomas M. Cover. Conflict between state information and intended information. In *Information Theory Workshop*, page 21, Mestovo, Greece, 1999.

[CS99] Brian Chen and Carl-Erik W. Sundberg. Broadcasting data in the FM band by means of adaptive contiguous band insertion and precancelling techniques.

In *Proceeding of the International Conference on Communications*, pages 823–827, 1999.

[CS01] Giuseppe Caire and Shlomo Shamai. On achievable rates in a multi-access Gaussian broadcast channel. In *Proc. of the Inter. Symposium on Info. Theory*, page 147, Washington, DC, 2001.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[CW01] Brian Chen and Gregory W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inform. Theory*, 47(4):1423–1443, May 2001.

[ESZ00] Uri Erez, Shlomo Shamai, and Ram Zamir. Capacity and lattice-strategies for cancelling known interference. In *Proceedings of the Cornell Summer Workshop on Information Theory*, August 2000.

[FKK01] Chuhong Fei, Deepa Kundu, and Raymond Kwong. The choice of watermark domain in the presence of compression. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, pages 79–84, 2001.

[Gal68] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, 1968.

[GH99] James R. Giles and Bruce Hajek. The jamming game for timing channels. In *Information Theory Workshop*, page 35, Mestovo, Greece, 1999.

[Gib92] Robert Gibbons. *Game Theory for Applied Economists*. Princeton University Press, Princeton, NJ, 1992.

[GP80] S. I. Gel'fand and M. S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

[Gra88] Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*. Spring-Verlag, New York, 1988.

[Gra00] Robert M. Gray. Toeplitz and circulant matrices: A review, 2000. Available at `http://ee.stanford.edu/~gray/toeplitz.html`.

[GS58] Ulf Grenander and Gabor Szegö. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

[HEG83] Chris Heegard and Abbas A. El Gamal. On the capacity of computer memory with defects. *IEEE Trans. Inform. Theory*, 29(5):731–739, September 1983.

[HM88] Walter Hirt and James L. Massey. Capacity of the discrete-time Gaussian channel with intersymbol interference. *IEEE Trans. Inform. Theory*, 34(3):380–388, 1988.

[HN87] Brian Hughes and Prakash Narayan. Gaussian arbitrarily varying channels. *IEEE Trans. Inform. Theory*, 33(2):267–284, March 1987.

[HN88] Brian Hughes and Prakash Narayan. The capacity of a vector Gaussian arbitrarily varying channel. *IEEE Trans. Inform. Theory*, 34(5):995–1003, September 1988.

[Jah81] Johann-Heinrich Jahn. Coding of arbitrarily varying multiuser channels. *IEEE Trans. Inform. Theory*, 27(2):212–226, March 1981.

[JF95] Rajan L. Joshi and Thomas R. Fischer. Comparison of generalized Gaussian and Laplacian modeling in DCT image coding. *IEEE Signal Processing Letters*, 2(5):81–82, May 1995.

[JJS93] Nikil Jayant, James Johnston, and Robert Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10):1385–1422, October 1993.

[KM77] János Körner and Katalin Marton. Images of a set via two channels and their role in multi-user communication. *IEEE Trans. Inform. Theory*, 23(6):751–761, November 1977.

[KP00a] Damianos Karakos and Adrian Papamarcou. Relationship between quantization and distribution rates of digitally watermarked data. In *Proc. of the Inter. Symposium on Info. Theory*, page 47, Sorrento, Italy, 2000.

[KP00b] Stefan Katzenbeisser and Fabien A. P. Petitcolas, editors. *Information Hiding Techniques for Steganography and Digital Watermarking*. Computer Security Series. Arthouse Tech, Boston, 2000.

[KZ00] Sanjeev Khanna and Francis Zane. Watermarking maps: Hiding information in structured data. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, CA, 2000.

[Lap96] Amos Lapidoth. Nearest neighbor decoding for additive non-Gaussian noise channels. *IEEE Trans. Inform. Theory*, 42(5):1520–1529, September 1996.

[LC01] Ching-Yung Lin and Shih-Fu Chang. Zero-error information hiding capacity of digital images. In *Proc. of the Inter. Conf. on Image Processing*, Thesaloniki, Greece, October 2001.

[LM00] Steven H. Low and Nicholas F. Maxemchuk. Capacity of text marking channel. *IEEE Signal Processing Letters*, 7(12):345–347, December 2000.

[LN98] Amos Lapidoth and Prakash Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, 44(6):2148–2177, October 1998.

[LSL00] Gerhard C. Langelaar, Iwan Setyawan, and Reginald L. Lagendijk. Watermarking digital image and video data: A state-of-the-art overview. *IEEE Signal Processing Magazine*, 17(5):20–46, September 2000.

[LWB+01] Ching-Yung Lin, Min Wu, Jeffrey A. Bloom, Matt L. Miller, Ingemar Cox, and Yui Man Lui. Rotation, scale, and translation resilient public watermarking for images. *IEEE Trans. Image Proc.*, 10(5):767–782, May 2001.

[Mar79] Katalin Marton. A coding theorem for the discrete memoryless broadcast channel. *IEEE Trans. Inform. Theory*, 25(3):306–311, May 1979.

[Mer00] Neri Merhav. On random coding error exponents of watermarking systems. *IEEE Trans. Inform. Theory*, 46(2):420–430, March 2000.

[Mit99] Thomas Mittelholzer. An information-theoretic approach to steganography and watermarking. In *Proc. of the Inter. Workshop on Info. Hiding*, number 1768 in Lecture Notes in Computer Science, 1999.

[MO99] Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. Preprint, available at `http://www.ifp.uiuc.edu/~moulin/paper.html`, 1999.

[MO00] Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. In *Proc. of the Inter. Symposium on Info. Theory*, page 19, Sorrento, Italy, 2000.

[Mou01] Pierre Moulin. The role of information theory in watermarking and its aplication to image watermarking. *Signal Processing*, 81(6):1121–1139, June 2001.

[MS74] James L. Mannos and David J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Trans. Inform. Theory*, 20(4):525–536, July 1974.

[MS01] Partha P. Mitra and Jason B. Stark. Nonlinear limits to the information capacity of optical fibre communications. *Nature*, 411:1027–1030, June 2001.

[MSP00] Ranjan K. Mallik, Robert A. Scholtz, and George P. Papavassilopoulos. Analysis of an on-off jamming situation as a dynamic game. *IEEE Trans. Comm.*, 48(8):1360–1373, August 2000.

[Mül93] F. Müller. Distribution of two-dimensional DCT coefficients of natural images. *Electronics Letters*, 29(22):1935–1936, October 1993.

[Oli99] Arlindo L. Oliveira. Robust techniques for watermarking sequential circuit designs. In *Proceeding of the Design Automation Conference*, pages 837–842, New Orleans, LA, 1999.

[OME98] Joseph A. O'Sullivan, Pierre Moulin, and J. Mark Ettinger. Information theoretic analysis of steganography. In *Proc. of the Inter. Symposium on Info. Theory*, page 297, Cambridge, MA, 1998.

[ORP97] Joseph J. K. Ó Ruanaidh and Thierry Pun. Rotation, scale and translation invariant digital image watermarking. In *Proc. of the Inter. Conf. on Image Processing*, pages 536–539, 1997.

[PAK99] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding – a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999.

[PP99] Shelby Pereira and Thierry Pun. Fast robust template matching for affine resistant image watermarks. In *Proc. of the Inter. Workshop on Info. Hiding*, number 1768 in Lecture Notes in Computer Science, pages 199–210, 1999.

[RA98] Mahalingam Ramkumar and Ali N. Akansu. Theoretical capacity measures for data hiding in compressed images. In *Proceedings of SPIE, Symposium on Voice, Video and Data Communication*, volume 3528, pages 482–492, Boston, MA, November 1998.

[SBM01a] Anelia Somekh-Baruch and Neri Merhav. On the error exponent and capacity games of private watermarking systems. Preprint, available at `http://tiger.technion.ac.il/users/merhav/`, 2001.

[SBM01b] Anelia Somekh-Baruch and Neri Merhav. On the error exponent and capacity games of private watermarking systems. In *Proc. of the Inter. Symposium on Info. Theory*, Washington, DC, 2001.

[SC96] Joshua R. Smith and Barrett O. Comiskey. Modulation and information hiding in images. In *Proc. of the Inter. Workshop on Info. Hiding*, number 1174 in Lecture Notes in Computer Science, pages 207–226, 1996.

[SEG00] Jonathan K. Su, Joachim J. Eggers, and Bernd Girod. Capacity of digital watermarks subjected to an optimal collusion attack. In *Proceedings of the European Signal Processing Conference*, 2000.

[Sha58] Claude E. Shannon. Channels with side information at the transmitter. *IBM Journal of Research and Development*, 2:289–293, October 1958.

[Sha59] Claude E. Shannon. Probability of error for optimal codes in a Gaussian channel. *The Bell System Technical Journal*, 38(3):611–656, May 1959.

[SKT98] Mitchell D. Swanson, Mei Kobayashi, and Ahmed H. Tewfik. Multimedia data-embedding and watermarking technology. *Proceedings of the IEEE*, 86(6):1064–1087, June 1998.

[SM88] Wayne E Stark and Robert J. McEliece. On the capacity of channels with block memory. *IEEE Trans. Inform. Theory*, 34(2):322–324, March 1988.

[SM01] Yossef Steinberg and Neri Merhav. Identification in the presence of side information with application to watermarking. *IEEE Trans. Inform. Theory*, 47(4):1410–1422, May 2001.

[SPR98] Sergio D. Servetto, Christine I. Podilchuk, and Kannan Ramchandran. Capacity issues in digital image watermarking. In *Proc. of the Inter. Conf. on Image Processing*, 1998.

[SV00] Rajesh Sundaresan and Sergio Verdú. Robust decoding for timing channels. *IEEE Trans. Inform. Theory*, 46(2):405–419, March 2000.

[SVZ98] Shlomo Shamai, Segio Verdú, and Ram Zamir. Systematic lossy source/channel coding. *IEEE Trans. Inform. Theory*, 44(2):564–579, March 1998.

[SW70] Josef Stoer and Christoph Witzgall. *Convexity and Optimization in Finite Dimensions I.* Springer-Verlag, 1970.

[Wol78] Jacob Wolfowitz. *Coding Theorems in Information Theory.* Spring-Verlag, third edition, 1978.

[Wyn67] Aaron D. Wyner. Random packings and coverings of the unit n-sphere. *The Bell System Technical Journal*, 46(9):2111–2118, November 1967.

[XA98] Liehua Xie and Gonzalo R. Arce. Joint wavelet compression and authentication watermarking. In *Proc. of the Inter. Conf. on Image Processing*, pages 427–431, 1998.

[Yan93] Kenjiro Yanagi. Optimal mutual information for coders and jammers in mismatched communication channels. *SIAM Journal of Control and Optimization*, 31(1):41–51, January 1993.

[YSJ+01] Wei Yu, Arak Sutivong, David Julian, Thomas M. Cover, and Mung Chiang. Writing on colored paper. In *Proc. of the Inter. Symposium on Info. Theory*, Washington, DC, 2001.