FUNDAMENTAL LIMITS OF COVERT COMMUNICATION

A Dissertation Presented

by

BOULAT A. BASH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2015

Computer Science

FUNDAMENTAL LIMITS OF COVERT COMMUNICATION

A Dissertation Presented

by

BOULAT A. BASH

Approved as to style and content by:

Donald F. Towsley, Chair

James F. Kurose, Member

Deepak K. Ganesan, Member

Dennis L. Goeckel, Member

Saikat Guha, Member

Lori A. Clarke, Chair
Computer Science

Моим родителям посвящается...

To my parents...

ACKNOWLEDGMENTS

I have had the good fortune of being advised on this thesis by a team of three brilliant scientists: Professors Don Towsley and Dennis Goeckel, and Dr. Saikat Guha. Prof. Towsley, my formal advisor at the UMass School of Computer Science and the chair of my dissertation committee, expertly guided me on the long journey to this thesis. First and foremost, I thank him for his patience. Prof. Towsley advises students by asking "big picture" questions, inspiring students to crystallize connections between what initially seem like disparate ideas. He encourages the pursuit of open problems in underexplored fields rather than incremental gains in existing research areas. While painful at times, this path is far more rewarding. This thesis is certainly a product of this approach.

Prof. Goeckel was not officially a co-chair of my dissertation committee because of the bureaucratic complications arising from him being on the faculty of the Electrical and Computer Engineering Department. However, he has effectively co-advised me with Prof. Towsley through almost the entirety of my UMass career. He taught me most of what I know about the mathematics of classical communication. I am grateful that he agreed to supervise the CMPSCI691WS seminar course on wireless network security in the Spring of 2011. I conceived the idea for that course when my graduate research stalled, and, in preparing to teach it, I discovered the topic of this thesis.

Dr. Guha has opened my eyes to the fascinating world of quantum optics and quantum information theory. I thank him for all the time he took answering my questions, as well as his patience with me as I worked through the complicated mathematics of quantum communication systems. During my internship under his supervision at Raytheon BBN Technologies in the Summer of 2013 not only was I able to explore

the theoretical aspects of covert optical communication but also to carry out the experimental evaluation that corroborated the theoretical results.

These experiments would not have been possible without tireless work of my fellow intern at BBN, Andrei Gheorghe. I grateful for the countless hours he spent building and debugging the testbed. I am also thankful to Drs. Jonathan Habif and Monika Patel of BBN for their help with the experiment, as well as the rest of the Quantum Information Processing Group for hosting me and Andrei.

The Computer Networks Research Group, where I spent the entirety of my career at UMass, has been a place of incredible intellectual stimulation. I would like to thank my past and present labmates: Yung-Chih Chen, Sookhyun Yang, Chang Liu, Daniel Sadoc Menasche, Fabricio Murai Ferreira, Antonio "Guto" de Aragão Rocha, Yu Gu, Mostafa Deghan, Dan Gyllstrom, and everyone else for the many great discussions, as well as putting up with the mess in my cubicle. I would like to extend special thanks to Bo Jiang and James Atwood for sharing their mathematical expertise with me, as well as Peter Desnoyers from the Laboratory for Advanced System Software for being a friend as well as a co-author on my first UMass paper. I also had the pleasure of working with the students at the Wireless Systems Laboratory in the Electrical and Computer Engineering Department, and am grateful to Tamara Sobers, Kyle Morrison, and Çağatay Çapar for many insightful discussions. I especially would like to acknowledge Ramin Soltani's work on extending the results of this thesis to networked settings, to thank him for his questions (which not only improved my explanatory skills but also provided me with a deeper understanding of the information theory), as well as to wish him the best in his research career.

I am grateful to my thesis committee members, Professors Jim Kurose and Deepak Ganesan, for their valuable feedback before and during my thesis defense. I would like to thank the present and former staff at the UMass School of Computer Science. Special thanks go to Laurie Connors for efficiently running the business aspects of

the Networks Group, to Sharon Mallory for handling the paperwork when I started graduate school and to Leeanne Leclerc for ensuring that everything was in order at the end of my studies. I am grateful to the staff at the Computer Science Computing Facility (CSCF) for not only keeping the computers running, but also being a friendly place to drop by for a chat, whether it's about the New England Patriots football (Gary Rehorka), heavy metal music and everything New England sports (Jamie Foster), golf (Terrie Kellogg), miscellaneous Windows tomfoolery (Andy Berkvist and David Korpiewski), and how to fix things (Rob Rice and Glenn Loud). I am especially grateful to Tyler Trafford, who not only fixed the many computer systems that I broke, but also became a great friend.

Indeed, spending so much time at UMass allowed me to meet many fascinating people and build many friendships. I will miss dropping by Prof. David Mix Barrington's office for fascinating discussions on mathematics, literature, sports, politics, theater, and all the combinations of these and other subjects. I will, however, continue to attend UMass sports events, and hope to see Brian Lynn at UMass basketball matches and John McColgan at UMass football and hockey games. I hope that someday I will run into Jeremy Wolf and Anna Curtis, as well as other fellow Graduate Employee Organization stewards. Outside of UMass I am grateful for the fellowship of my alumni brothers from the Dartmouth College chapter of the Sigma Nu fraternity, the connection to my alma-mater through the involvement on the executive board of the Dartmouth Club of Pioneer Valley, as well as the enduring friendship with Bob Hatcher. Bob, in this thesis I made a good use of the English writing skills that you have instilled in me.

I would not be where I am right now without the love, the support, and the inspiration from my family: my father Dr. Anvar Bash, my mother Mrs. Zoukhra Bash, and my sister Dina. I started on the road to a doctorate in Computer Science at age 8, when my mother (who is an amazing computer programmer) taught me the

`for`-loops in FOCAL, a programming language/operating system running my first computer, Electronika BK-0010. My dad, a professor of physics (now retired), has instilled in me the love of mental exercise through lifelong learning as well as the importance of physical exercise to sustain every pursuit. Even though over the years we butted heads as any siblings do, my sister was always there for me with words of advice and encouragement. My family has supported me when things were at their worst and shared the joy of my triumphs: for that I am eternally grateful.

I would not have emotionally survived graduate school were it not for my girlfriend and my partner Gail Sweeney. The gratitude I feel for her being in my life is difficult to describe with words alone. I would like to thank her for her love and support in this long journey, for keeping our relationship emotionally close while being geographically separated, for bearing with all of my idiosyncrasies as well as obsession with sports, for cooking meals for me in Amherst on the weekends and leaving them in the freezer (thus preventing me from starving during the week), for accompanying me on all the trips we took all over New England and beyond, and for finding fascinating things to do, places to visit, and foods to eat on these trips. Meeting Gail is the best thing to ever happen in my life, and I am grateful every day for having her alongside me.

Finally, I would like to acknowledge the National Science Foundation for funding my doctoral work.

ABSTRACT

FUNDAMENTAL LIMITS OF COVERT COMMUNICATION

FEBRUARY 2015

BOULAT A. BASH

A.B., DARTMOUTH COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Donald F. Towsley

Traditional security (e.g., encryption) prevents unauthorized access to message content; however, detection of the mere presence of a message can have significant negative impact on the privacy of the communicating parties. Unlike these standard methods, *covert* or low probability of detection (LPD) communication not only protects the information contained in a transmission from unauthorized decoding, but also prevents the detection of a transmission in the first place. In this thesis we investigate the fundamental laws of covert communication.

We first study covert communication over additive white Gaussian noise (AWGN) channels, a standard model for radio-frequency (RF) communication. We present a square root limit on the amount of information transmitted covertly and reliably over such channels. Specifically, we prove that if the transmitter has channels to the intended receiver and the warden that are both AWGN, then $\mathcal{O}(\sqrt{n})$ covert bits can be

reliably transmitted to the receiver in n uses of the channel. Conversely, attempting to transmit more than $\mathcal{O}(\sqrt{n})$ bits either results in detection by the warden with probability one or a non-zero probability of decoding error at the receiver as $n \to \infty$.

Next we study the impact of warden's ignorance of the communication attempt time. We prove that if the channels from the transmitter to the intended receiver and the warden are both AWGN, and if a single n-symbol period slot out of $T(n)$ such slots is selected secretly (forcing the warden to monitor all $T(n)$ slots), then $\mathcal{O}(\min\{\sqrt{n \log T(n)}, n\})$ covert bits can be transmitted reliably using this slot. Conversely, attempting to transmit more than $\mathcal{O}(\sqrt{n \log T(n)})$ bits either results in detection with probability one or a non-zero probability of decoding error at the receiver.

We then study covert optical communication and characterize the ultimate limit of covert communication that is secure against the most powerful physically-permissible adversary. We show that, although covert communication is impossible when a channel injects the minimum noise allowed by quantum mechanics, it is attainable in the presence of any noise excess of this minimum (such as the thermal background). In this case, $\mathcal{O}(\sqrt{n})$ covert bits can be transmitted reliably in n optical channel uses using standard optical communication equipment. The all-powerful adversary may intercept all transmitted photons not received by the intended receiver, and employ arbitrary quantum memory and measurements. Conversely, we show that this square root scaling cannot be circumvented. Finally, we corroborate our theory in a proof-of-concept experiment on an optical testbed.

CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Security and privacy are critical in modern-day wireless communications, with conventional cryptography [67, 80], information-theoretic secrecy [90, 22], and quantum cryptography [7] offering progressively higher levels of security against the unauthorized access to transmitted information. Widely-deployed conventional cryptography presents the adversary with a problem that he/she is assumed not to be able to solve because of computational constraints, while information-theoretic secrecy presents the adversary with a signal from which he/she cannot extract information about the message contained therein. Quantum key distribution (QKD) lets two distant parties generate a shared secret key that is secure from the most powerful adversary allowed by physics. This key, when used as a one-time pad [76], yields an unconditionally-secure cipher. However, while these approaches address security in many domains by protecting the content of the message, they do not mitigate the threat to users' privacy from the discovery of the very existence of the message itself.

Indeed, transmission attempts expose connections between the parties involved, and recent disclosures [6] of massive surveillance programs revealed that this "metadata" is widely collected and often used for nefarious means. Furthermore, the transmission of encrypted data can arouse suspicion, and many cryptographic schemes can be defeated by a determined adversary using non-computational means such as side-channel analysis. Anonymous communication tools [23] such as Tor resist metadata collection and traffic analysis by randomly directing encrypted traffic through a large network. While these tools conceal the identities of source and destination nodes in

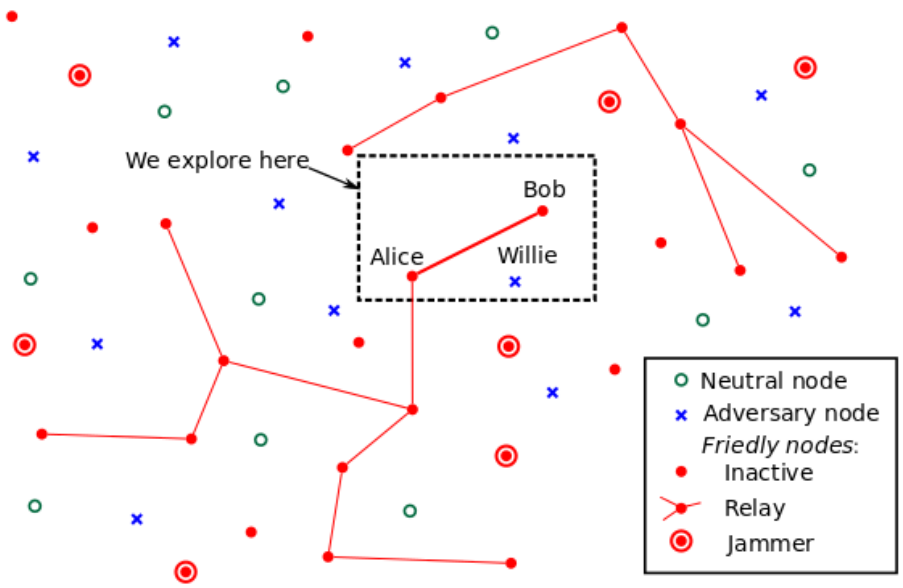


Figure 1.1: Our vision of a "shadow network" assembled from "friendly" nodes. Relays (each indicated by a filled red circle with lines expressing active connections to other relays) generate, transmit, receive, and consume data, while jammers (each indicated by a filled red circle with a concentric red circle) generate artificial noise that impairs the ability of wardens (indicated by the blue crosses) to detect the presence of communication. Inactive friendly nodes (indicated by filled red circles with neither connections to other nodes nor the concentric circles) are capable of receiving, transmitting and jamming, while the neutral nodes (indicated by empty green circles) produce background interference but do not participate in the shadow network nor assist the wardens. Most of this article focuses on the scenario involving only three nodes: transmitter Alice, receiver Bob, and warden Willie, as indicated in the diagram.

a "crowd" of relays, they are designed for the Internet and are not effective in wireless networks, which are typically orders of magnitude smaller. Moreover, such tools offer little protection to users whose communications are already being monitored by the adversaries. Thus, secure communication systems should also provide *covert*, or low probability of detection (LPD) communication. Such systems not only protect the information contained in the message from being decoded, but also prevent the adversary from detecting the transmission attempt in the first place and allow communication where it is prohibited.

While traditionally covert communication received relatively little attention, our recent work [4] on its fundamental limits has spurred a revival of interest. The overarching goal of covert communication research is the establishment of "shadow networks," an example of which is depicted in Figure 1.1. However, to create such networks, we must first learn how to connect its component nodes by stealthy communication links. Therefore, in this thesis we focus on the fundamental limits of such point-to-point links and address the following question: how much information can a sender Alice reliably transmit (if she chooses to transmit) to the intended recipient Bob while hiding it from the adversary, warden Willie?

The contributions of this thesis are:

- The development of the fundamental theory of covert communication over additive white Gaussian noise (AWGN) channels (Chapters 4 and 5):
  - We show that Shannon capacity [75] does not apply to quantifying the limits of covert communication. Unlike standard secure communication (e.g., encryption and information-theoretic secrecy) that only protects the message content, covert communication is subject to the *square root law*: when both Bob and Willie have AWGN channels from Alice, she can reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits in n uses of her channel to Bob; attempting to transmit more information either results in detection with high probability or unreliable communication [4]. Since Shannon capacity is the number of bits that can be transmitted per channel use as the number of channel uses approaches infinity, our result implies that the Shannon capacity of covert communication over AWGN channel is zero.
  - We also demonstrate that ignorance on the part of Willie as to when Alice might transmit can be used to increase the covert communication throughput [5].

- The development of the fundamental theory of covert optical communication (Chapter 6): Since modern high-sensitivity optical components are primarily limited by noise of quantum-mechanical origin, we employ quantum information theory to derive the ultimate limit of covert communication that is secure against the most powerful adversary physically permissible. This is the same benchmark of security to which quantum cryptography adheres for encrypted communication. As in the AWGN case, the standard result on the limits of communication over a quantum channel (the Holevo capacity [43], a generalization of Shannon capacity to quantum channels) does not apply to quantum-noise limited covert optical communication. We demonstrate that a square root law similar to the one for AWGN channels holds for covert optical communication as long as there is a positive amount of non-adversarial noise (e.g., thermal background). We also show that non-adversarial noise is critical as covert communication is impossible in its absence, which also sharply contrasts the standard quantum cryptography results [3].

- Experimental validation of the square root law for covert optical communication (Chapter 7): We corroborate the theory developed in Chapter 6 in a proof-of-concept experiment. This is the first known implementation of a truly quantum-information-theoretically secure covert communication system that allows communication when all transmissions are prohibited [3].

This thesis is structured as follows: we begin with a high-level overview of covert communication in Chapter 2, including both the intuitive treatment of our results as well as the related work. We provide the mathematical prerequisites in Chapter 3, including the rigorous definition of covert communication. In Chapter 4 we study covert communication over the additive white Gaussian noise (AWGN) channels, establishing the fundamental square root limit. In Chapter 5 we relax the assumption that Willie knows *when* to monitor his AWGN channel from Alice for a possible transmis-

sion and show that this increases the covert communication throughput. We employ quantum information theory to analyze covert optical communication in Chapter 6. In Chapter 7 we describe our proof-of-concept experiment that corroborates the theory developed in Chapter 6. We conclude the thesis in Chapter 8 by summarizing our contributions and discussing both the ongoing research and the potential future work in covert communication.

CHAPTER 2

OVERVIEW OF COVERT COMMUNICATION

The objective of this chapter is to frame the thesis in the context of previous work on covert communication as well as to intuitively sketch the results of the technical chapters that follow. We limit the prerequisite mathematical knowledge for this chapter to the basic asymptotic notation (which is formally defined in Section 3.1). We begin by briefly reviewing the ancient field of steganography in Section 2.1. Steganography is the practice of hiding messages in innocuous objects. It is important because not only was the first covert communication system based on steganography, but it also was the subject of the first information-theoretic investigation of stealthy communication.

However, the use of steganography for covert communication requires the transmission of objects containing the hidden messages, which is challenging when all transmissions are prohibited. Thus, in Section 2.2, we discuss covert communication over noisy channels. The focus of the bulk of Section 2.2 are analog radio frequency (RF) channels, where the information is hidden in the channel artifacts such as additive white Gaussian noise (AWGN). After a brief overview of the classical spread spectrum methods, we introduce our work on the fundamental limits of covert communication over AWGN channels. Our presentation in Section 2.2 is largely intuitive and we defer the technical details to Chapters 4 and 5. At the end of Section 2.2 we discuss covert communication over digital communication channels, where the progress by other research groups was inspired by our work, as well as briefly touch upon the covert broadcast scenario.

Optical communication has intrinsically high resistance to detection, however, analysis of optical systems demands the use of quantum mechanics. We outline our results on covert optical communication in Section 2.3 while deferring the technical details to Chapter 6. At the end of Section 2.3 we touch on the experimental work from Chapter 7. We conclude this chapter in Section 2.4 by discussing the relationship of this thesis to other work in communication.

2.1 Steganography

The first known description of covert communication is by Herodotus circa 440 BCE in *The Histories* [42], an account of the Greco-Persian Wars: in Chapter 5 Paragraph 35, Histiaeus shaves the head of his slave, tattoos the message on his scalp, waits until the hair grows back, and then sends the slave to Aristagoras with instructions to shave the head and read the message that calls for an anti-Persian revolt in Ionia; in Chapter 7 Paragraph 239, Demaratus warns Sparta of an imminent Persian invasion by scraping the wax off a wax tablet, scribbling a message on the exposed wood, and concealing the message by covering the tablet with wax. This practice of hiding sensitive messages in innocuous objects is known as *steganography*.

Modern digital steganography conceals messages in finite-length, finite-alphabet *covertext* objects, such as images or software binary code. Embedding hidden messages in covertext produces *stegotext*, necessarily changing the properties of the covertext. The countermeasure for steganography, *steganalysis* (an analog of cryptanalysis for cryptography), looks for these changes. Covertext is usually unavailable for steganalysis (when it is, steganalysis consists of the trivial comparison between the covertext and the suspected stegotext). However, Willie is assumed to have a complete statistical model of the covertext. The amount of information that can be embedded without being discovered depends on whether Alice also has access to this model.

If she does, then *positive-rate steganography* is achievable: given an $\mathcal{O}(n)$-bit[1] secret "key" that is shared with Bob prior to the embedding, $\mathcal{O}(n)$ bits can be embedded in an n-symbol covertext without being detected by Willie [29, Chapter 13.1].

Recent work focuses on the more general scenario where the complete statistical model of the covertext is unavailable to Alice. Then, Alice can safely embed $\mathcal{O}(\sqrt{n}\log n)$ bits by modifying $\mathcal{O}(\sqrt{n})$ symbols out of n in the covertext, at the cost of pre-sharing $\mathcal{O}(\sqrt{n}\log n)$ secret bits with Bob. Note that this *square root law of digital steganography* yields *zero-rate steganography* since $\lim_{n\to\infty} \frac{\mathcal{O}(\sqrt{n}\log n)}{n} = 0$ bits/symbol. The square root law was first observed empirically in the experimental analysis of existing steganalysis systems and the proof is available in Chapter 13.2.1 of the review of pre-2009 work in digital steganography [29]. More recent work shows that an *empirical* model of covertext suffices to break the square root law and achieve positive-rate steganography [20]. Essentially, while embedding at a positive rate lets Willie obtain $\mathcal{O}(n)$ stegotext observations (enabling detection of Alice when statistics of covertext and stegotext differ), the increasing size n of the covertext allows Alice to improve her covertext model and produce statistically-matching stegotext.

Although it is an active research area, steganography has limited application for covert communication. First, analysis of the steganographic systems generally assumes that stegotext is not corrupted by a noisy channel. Second, the generalization of the results for steganographic systems is limited because of their finite-alphabet discrete nature. Finally, the most serious drawback of using steganography for covert communication is the necessity of transmitting the stegotext from Alice to Bob—a potentially unrealizable requirement when all communication is prohibited. We thus consider covert communication over noisy channels.

---

[1]The asymptotic notation is formally defined in Section 3.1.

2.2 Covert Communication over Noisy Channels

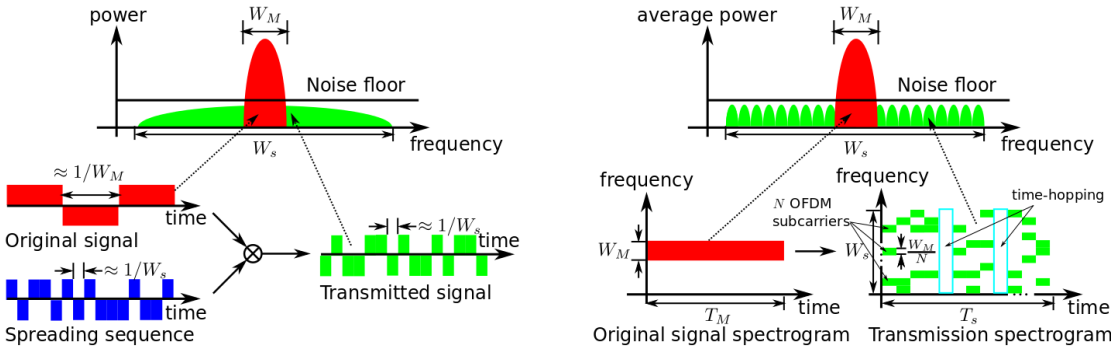
We begin the investigation of covert communication over noisy channels by considering RF wireless communication. Since its emergence in the early 20th century, protecting wireless RF communication from detection, jamming and eavesdropping has been of paramount concern. *Spread spectrum* techniques, devised between the two world wars to address this issue, have constituted the earliest and, arguably, the most enduring form of physical layer security.

2.2.1 Spread Spectrum Communication

Essentially, the spread spectrum approach involves transmitting a signal that requires a bandwidth W_M on a much wider bandwidth $W_s \gg W_M$, thereby suppressing the power spectral density of the transmission below the noise floor. Spread spectrum systems provide both a covert communication capability as well as resistance to jamming, fading, and other forms of interference. A comprehensive review of this field is available in [78, 81]. Typical spread spectrum techniques include *direct sequence* spread spectrum (DSSS), *frequency-hopping* spread spectrum (FHSS), and their combination.

When Alice uses DSSS, she multiplies the signal waveform by the *spreading sequence*—a randomly-generated binary waveform with a substantially higher bandwidth than the original signal. The resulting waveform is thus "spread" over a wider bandwidth, which reduces the power spectral density of the transmitted signal. Bob uses the same spreading sequence to de-spread the received waveform and obtain the original signal. The spreading sequence is exchanged by the communicating parties prior to transmission and is kept secret from Willie.[2] Outside of security applications,

[2]While an exchange of a secret prior to covert communication is similar to a key exchange in symmetric-key cryptography [67, Chapter 1.5] (e.g., one-time pad [76]), an important distinction is that public-key cryptography techniques [67, Chapter 1.8] cannot be used to exchange this secret on a channel monitored by Willie without revealing the intention to communicate.



(a) DSSS

(b) FHSS with OFDM and time-hopping.

Figure 2.1: Illustration of spread spectrum techniques. Direct sequence spread spectrum (DSSS) is described in (a): the signal with bandwidth W_M is multiplied by a spreading sequence with bandwidth $W_s \gg W_M$ prior to the transmission, reducing below the noise floor the power spectral density of the transmission. Frequency-hopping spread spectrum (FHSS) achieves the same by re-tuning the transmitter to a different carrier frequency within a wide frequency band. As illustrated in (b), orthogonal frequency-division multiplexing (OFDM) enables the use of multiple frequency bands at each transmission and time-hopping allows arbitrary varying of the transmission duty cycle. Note that the spreading sequence and the frequency/time hopping pattern are kept secret from the adversaries.

the use of *public* uncorrelated spreading sequences between transmitter/receiver pairs enables multiple access; DSSS thus forms the basis of code-division multiple access (CDMA) protocols used in cellular telephony [85]. The operation of DSSS is illustrated in Figure 2.1(a).

Unlike DSSS, FHSS re-tunes the carrier frequency for each transmitted symbol. However, like the spreading sequence in DSSS, the frequency-hopping pattern is also randomly generated and secretly shared between the communicating parties prior to the transmission. FHSS can be combined with orthogonal frequency-division multiplexing (OFDM), enabling the use of multiple carrier frequencies. To further reduce the average transmitted symbol power, FHSS can be used with *time-hopping* techniques that randomly vary the duty cycle (the time-hopping pattern is also secretly

pre-shared between the communicating parties prior to the transmission). The operation of FHSS with OFDM and time-hopping is illustrated in Figure 2.1(b).

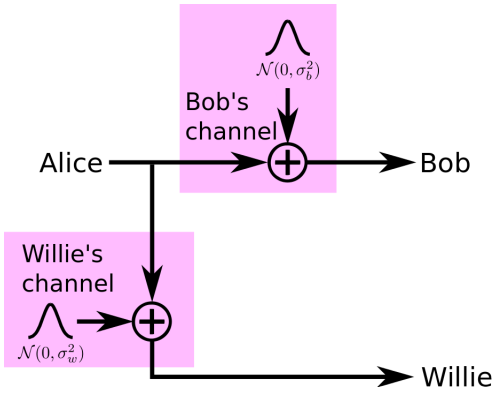
Although spread spectrum architectures for covert communication are well-developed, their fundamental information-theoretic limits have not been explored. Knowledge of the limits of communication systems is important, particularly since modern coding techniques (such as Turbo codes [8] and low-density parity check [31, 61] codes) allow 3G/4G cellular systems to operate near their theoretical *channel capacity*, the maximum rate of reliable communication without imposing any security requirement [75]. We thus discuss the fundamental limits of covert communication next.

2.2.2 Square Root Law for Covert Communication over AWGN Channels

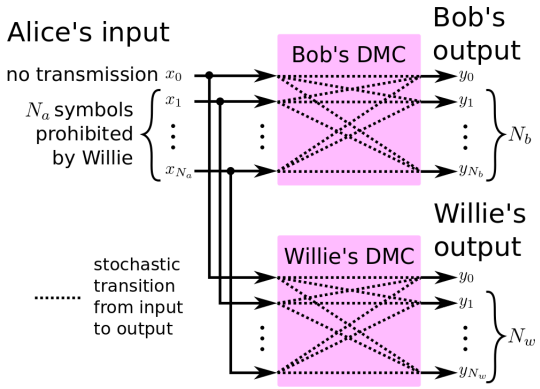
Spread spectrum systems allow communication where it is prohibited because spreading the signal power over a large time-frequency space substantially reduces the adversary's signal-to-noise ratio (SNR). This impairs his/her capability to discriminate between the noise and the information-carrying signal corrupted by noise. Here we determine just how small the power has to be for the communication to be fundamentally undetectable, and how much covert information can be transmitted reliably.

Consider an additive white Gaussian noise (AWGN) channel model where the signaling sequence is corrupted by the addition of a sequence of independent and identically distributed zero-mean Gaussian random variables with variance σ^2. This is the standard model for a free-space RF channel. Suppose that the channels from Alice to Bob and to Willie are subject to AWGN with respective variances $\sigma_b^2 > 0$ and $\sigma_w^2 > 0$,[3] as illustrated in Figure 2.2(a). Let *channel use* denote the unit of communication resource—a fixed time period that is used to transmit a fixed-

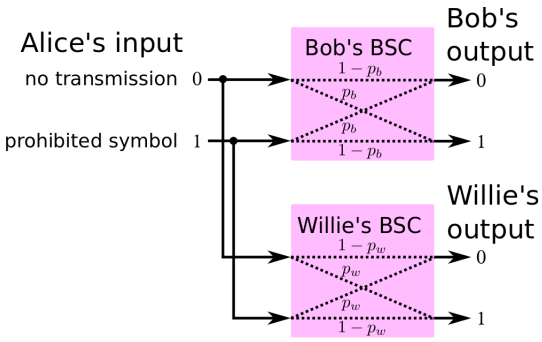
[3]If the channel from Alice to Bob is noiseless ($\sigma_b^2 = 0$) and the channel from Alice to Willie is noisy ($\sigma_w^2 > 0$), then Alice can transmit an infinite amount of information to Bob; if the channel from Alice to Willie is noiseless ($\sigma_w^2 = 0$), then covert communication is impossible.



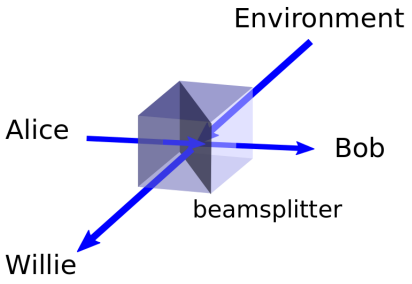
(a) Additive white Gaussian noise channel.



(b) Discrete memoryless channel.



(c) Binary symmetric channel.



(d) Lossy-noisy optical channel.

Figure 2.2: Channel models.

bandwidth signal—and let n be the total number of channel uses available to Alice and Bob (e.g., $n = W_s T_s$ in Figure 2.1(b)). Willie's ability to detect Alice's transmission depends on the amount of total power that she uses. Let's intuitively derive[4] Alice's power constraint assuming that Willie observes these n channel uses. When Alice is not transmitting, Willie observes AWGN with total power $\sigma_w^2 n$ over n channel observations on average. By standard statistical arguments, with high probability, observations of the total power lie within $\pm c\sigma_w^2\sqrt{n}$ of this average, where c is a

[4]The formal proof is in Chapter 4.

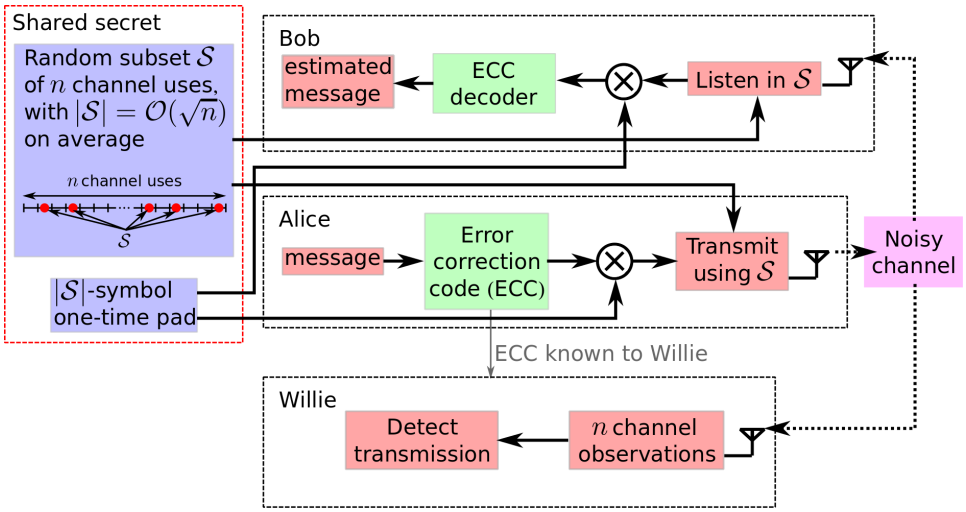


Figure 2.3: Design of a covert communication system that allows Alice and Bob to use any error-correction codes (including those known to Willie) to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using $\mathcal{O}(\sqrt{n}\log n)$ pre-shared secret bits. Subset $\mathcal{S}$ is effectively a frequency/time-hopping pattern, generated by flipping a biased random coin n times, with probability of heads $\mathcal{O}(1/\sqrt{n})$: the i^{th} channel use is selected for transmission if the i^{th} flip is heads. On average, the number of channel uses selected $|\mathcal{S}| = \mathcal{O}(\sqrt{n})$. Knowledge of $\mathcal{S}$ allows Bob to discard the observations of his channel from Alice that are not in $\mathcal{S}$ and decode her message; Willie observes mostly noise since he does not know $\mathcal{S}$. Furthermore, application of a one-time pad prevents Willie's exploitation of the error correction code's structure to detect Alice (rather than protects the message content). In Appendix A.4 a binary amplitude modulation is used while in Chapter 7 this scheme is implemented on an optical testbed using a Q-ary pulse position modulation.

constant. Since Willie observes Alice's signal power when she transmits in addition to the noise power, to prevent Willie from getting suspicious, the total amount of power that Alice can use is limited to $\mathcal{O}(\sqrt{n})$, or $P = \mathcal{O}(1/\sqrt{n})$ per symbol; otherwise her transmission will be detected. We show (Chapter 4) that this allows her to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits to Bob in n channel uses, but no more than that. The similarity of this *square root law for covert communications* to the steganographic square root law from Section 2.1 is attributable to the mathematics of statistical hypothesis testing (as discussed in Section 4.2.3.2). The additional $\log n$ factor in the steganographic square root law comes from the fact that the steganographic "channel" to Bob is noiseless.

As in steganography and spread spectrum communication, prior to communicating, Alice and Bob may share a secret signaling scheme. Figure 2.3 depicts a method used in Appendix A.4 that allows Alice and Bob to use any error-correction code (which can be known to Willie) on top of binary modulation to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using $\mathcal{O}(\sqrt{n}\log n)$ pre-shared secret bits; this method is trivially extended to higher-order modulation schemes (e.g., in Chapter 7 is is used with Q-ary pulse position modulation to implement optical communication on an optical testbed.) While the size of the key is asymptotically larger than the size of the transmitted message, there are many real-world scenarios where this is an acceptable trade-off to being detected. Furthermore, the recent extension of our work in Chapter 4 to digital covert communication that we describe next suggests that the pre-shared secret can be eliminated in some scenarios.

2.2.3 Digital Covert Communication

The *discrete memoryless channel* (DMC) model [19, Chapter 7] describing digital communication often sheds light on what is feasible in practical communication systems. Discrete input and output allow the DMC to be represented using a bipartite graph where the two sets of vertices correspond to input and output alphabets, and edges correspond to the stochastic transitions from input to output symbols. The memoryless nature of the DMC means that its output is statistically independent from any symbol other than the input at that time. We illustrate this model in Figure 2.2(b), which we augment by designating one of Alice's inputs as "no transmission"—a necessary default channel input permitted by Willie.[5]

We first consider the *binary symmetric channel* (BSC) illustrated in Figure 2.2(c), which restricts the DMC to binary input and output alphabet $\{0, 1\}$, and the probability of a crossover from zero at the input to one at the output being equal to that of

[5]For example, this could be the zero-signal in the AWGN channel scenario.

a crossover from one to zero. Denote by $p_b > 0$ and $p_w > 0$ the crossover probabilities on Bob's and Willie's BSCs, respectively. It has been shown that, while no more than $\mathcal{O}(\sqrt{n})$ covert bits can be reliably transmitted in n BSC uses, if $p_w > p_b$, then the pre-shared secret is unnecessary [15].

Channel *resolvability* can be employed to generalize the square root law in [15] to DMCs. Channel resolvability is the minimum input entropy[6] needed to generate a channel output distribution that is "close" (by some measure of closeness between probability distributions[7]) to the channel output distribution for a given input; resolvability has been used to obtain new, stronger results for the information-theoretic secrecy capacity [10]. If the channels from Alice to both Willie and Bob are DMCs, and Willie's channel capacity is smaller than Bob's, then techniques in [44] can be used to demonstrate the square root law without a pre-shared secret [54]. The results in [15] and [54] provide evidence that secret-less covert communication over the AWGN channel should be possible.

2.2.4 Willie's Ignorance of Transmission Time Helps Alice

When deriving the square root laws, we assume that Willie knows *when* the transmission takes place, if it does. However, in many scenarios Alice and Bob have a pre-arranged time for communication that is unknown to Willie (e.g., a certain time and day). The transmission might also be short relative to the total time during which it may take place (e.g., a few seconds out of the day). If Willie does not know when the message may be transmitted, he has to monitor a much longer time period

[6]Essentially, entropy measures "surprise" associated with a random variable, or its "uncertainty". For example, a binary random variable describing a flip of a fair coin with equal probabilities of heads and tails has higher entropy than the binary random variable describing a flip of a biased coin with probability of heads larger than tails. The output of the biased coin is more predictable, and less surprising, as one should observe more heads. Introductory texts on the information theory (such as [19]) provide the in-depth discussion of entropy and other information-theoretic concepts.

[7]Examples of measures of closeness are variational distance and relative entropy (see Section 4.2.1 and [19, Chapter 11]).

than the time required for the transmission. It turns out that Willie's ignorance of Alice's transmission time allows her to transmit additional information to Bob. Surprisingly, under some mild conditions on the relationship between the total available transmission time and the transmission duration, Alice and Bob do not even have to pre-arrange the communication time.

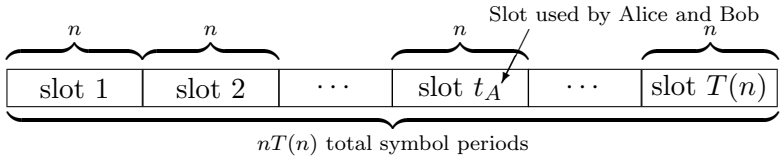


Figure 2.4: Slotted channel: each of the $T(n)$ slots contains n symbols. Alice and Bob use slot t_A to communicate.

Consider the scenario of Section 2.2.2 where the channels from Alice to Bob and to Willie are subject to AWGN. Suppose that time is slotted, with each of $T(n)$ slots containing n channel uses and $T(n)$ being an increasing function of n, as depicted in Figure 2.4. Clearly, if Alice used all $T(n)$ slots, by the square root law, she could reliably transmit $\mathcal{O}(\sqrt{nT(n)})$ covert bits in $nT(n)$ channel uses. However, suppose that she only employs a *single* time slot t_A, selected uniformly at random. While the naïve application of the square root law states that she can reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits, in fact, as we show in Chapter 5, Willie being subject to much more noise from having to monitor all $T(n)$ slots allows Alice to reliably transmit $\mathcal{O}(\min\{\sqrt{n \log T(n)}, n\})$ covert bits during the selected time slot, if she chooses to transmit. Furthermore, no additional bits of pre-shared secret are required if $T(n) < 2^{c_T n}$, where constant $c_T > 0$ depends on the relative power of AWGN on Bob's and Willie's channels. Conversely, no more than $\mathcal{O}(\sqrt{n \log T(n)})$ can be transmitted both reliably and covertly.

While it has been established that the absence of transmission timing knowledge by Willie allows Alice to transmit more covert bits, the proof in Chapter 5 is valid

only for an AWGN channel. The problem is open for DMCs; however, we suspect that the same scaling laws hold.

2.2.5 Positive-rate Covert Communication

The covert communication channels described above have zero rates, since the average number of bits that can be covertly transmitted per channel use tends to zero as the number of channel uses n gets large. Here we discuss the possibility of positive-rate covert communication, i.e. reliable transmission of $\mathcal{O}(n)$ covert bits in n channel uses. In general, the circumstances that allow Alice to covertly communicate with Bob at positive rates occur either when Willie *allows* Alice to transmit positive-entropy messages or when he is ignorant of the probabilistic structure of the noise on his channel (note that the applicability of the steganographic results [20] here is limited since estimation of the probabilistic structure of the noise on Willie's channel is insufficient unless Alice can "replace" this noise rather than add to it). When Willie allows transmissions, the covert capacity is the same as the information-theoretic secrecy capacity (see [44] for treatment of the DMCs). Incompleteness of Willie's noise model can also allow positive-rate covert communication: in the noisy digital channel setting, Willie's ignorance of the channel model is a special case of the scenario in [44]; while in the AWGN channel setting, random noise power fluctuations have been shown to yield positive-rate covert communication [57, 58]. The latter result holds even when the noise power can be bounded; a positive rate is achieved because Willie does not have a constant baseline of noise for comparison.

2.2.6 Covert Broadcast

Some of the results for the point-to-point covert communication in the presence of a single warden that are discussed in this section can easily be extended to scenarios with multiple receivers. For example, covert communication over an AWGN channel effectively imposes a power constraint on Alice. Since a pre-shared secret enables

covert communication in this setting, if each receiver obtains it prior to communication, Alice can use standard techniques [19, Chapter 15] to encode covert messages to multiple recipients. The extension to a multi-warden setting as well as other networked scenarios is the ongoing work discussed in Chapter 8.

2.3 Covert Optical Communication

Optical signaling enables such important cryptographic security techniques as QKD [7]. It is also ideal for covert communication because of the narrow beam spread of laser-based communication systems [30, 35] and the availability of time-domain reflectometry devices [2] for detecting taps in optical fiber. Therefore, in this section we outline the main results for covert optical communication from Chapters 6 and 7.

2.3.1 Optical channel: model and analysis

A lossy-noisy optical channel is typically modeled by a beamsplitter that takes inputs from Alice and the environment and outputs to Bob and Willie, as depicted in Figure 2.2(d). The analysis of Section 2.2.2 applies to an optical channel with a thermal environment (described later) where Alice uses a laser-light transmitter while both Bob and Willie use coherent-detection (i.e., homodyne or heterodyne) receivers. However, modern high-sensitivity optical components are limited by noise of quantum-mechanical origin. Thus, establishing the ultimate limit of covert communication that is secure against the most powerful adversary allowed by physics—the standard of security to which quantum cryptography adheres for encrypted communication—requires quantum information theory.

2.3.2 Covert communication is impossible over pure-loss channels

Consider a *pure loss* optical channel, i.e., one with a "vacuum" environment, which corresponds to the minimum noise the channel must inject to preserve the Heisen-

berg inequality of quantum mechanics. If Willie has such a channel from Alice and is limited only by the laws of physics in his choice of a detector, regardless of the resources available to Alice and Bob, he can prevent *any* reliable covert communication by using an *ideal single photon detector* (SPD). An ideal SPD registers detection events only when one or more photons impinge on its receiver aperture, and, thus, Willie does not experience false alarms. A detection of a single photon gives away Alice's transmission attempt regardless of her signaling scheme. This restricts Alice to transmissions that are nearly indistinguishable from vacuum, rendering unreliable any communication designed to be covert and vice-versa.

2.3.3 Square root law for covert optical communication

The analysis of quantum cryptography schemes such as QKD assumes a nearly-omnipotent adversary that is in control of the channel noise and is capable of employing ideal detectors. While providing the highest level of security, such assumptions are unreasonably strong for covert communication: a positive amount of noise in addition to the quantum minimum is unavoidable in any practical setting. This *excess noise* is not controlled by the adversary and originates either from the thermal environment or the detector itself. First, consider the *thermal noise channel*, where the noise source is the thermal environment. The thermal environment, such as the background radiation from a 300K thermal blackbody, is modeled by a mixture of zero-mean complex Gaussian-distributed coherent states, where a coherent state is a quantum-mechanical description of ideal laser light. The thermal noise channel allows covert communication when Alice transmits $\mathcal{O}(1/\sqrt{n})$ photons per optical mode averaged over n available modes[8]. Signaling photons then blend in with the noise photons, resulting in the *square root law for covert optical communication*: Alice can reliably transmit

[8]Here an optical channel *mode* is a communication resource unit, equivalent to the channel use in the previous section. A more formal description is provided in the footnote on page 65 and in Appendix B.4.

$\mathcal{O}(\sqrt{n})$ covert bits to Bob using n modes. For the thermal noise channel the square root law holds even when Willie has access to all of Alice's signaling photons not captured by Bob and arbitrary quantum measurement and computation resources. Alice and Bob pre-share a secret codebook prior to communication and Bob needs only a suboptimal homodyne-detection receiver to decode Alice's transmissions. We note that these results are also relevant to the RF covert communication systems because of the recent advances in quantum-limited microwave-frequency coherent detectors and amplifiers [1].

However, assuming single-mode detection for Willie, thermal noise is negligible at optical frequencies. It is easy to show that the ideal photon number resolving (PNR) detector is asymptotically optimal in this (hypothetical) pure-loss channel scenario. PNR detectors count the number of photons observed in a mode. However, their practical implementations suffer from various excess noise sources, with *dark counts* being most prevalent. Dark counts are false photon detection events triggered by the internal spontaneous emission process, rather than photons impinging on the detector's active surface (they also plague practical implementations of SPDs). The square root law holds in this scenario, and Alice can use laser pulses to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits in n modes. A noisy PNR detector can also be used to prove the converse of the square root law, i.e., more than $\mathcal{O}(\sqrt{n})$ covert bits cannot be transmitted using n modes without either being detected or suffering from uncorrectable decoding errors. In fact, a noisy SPD suffices to prove the converse in all practical scenarios when codewords with bounded photon number per mode. In Chapter 7 we use noisy SPDs, laser-light pulse-position modulation and Reed-Solomon error-correction coding (implementing the covert communication system depicted in Figure 2.3) to experimentally demonstrate the square root law for covert optical communication on a testbed.

2.4 Relationship to Previous Work in Communications

Here we relate our thesis to other work in communication.

2.4.1 Information-theoretic secrecy

There exists a rich body of literature on the information-theoretic secrecy resulting from the legitimate receiver having a better channel to the transmitter than the adversary. Wyner was the first to show that if the adversary only has access to a noisy version of the signal received by the legitimate receiver (using a *wire-tap channel*), then the legitimate receiver can achieve a positive secure communication rate to the sender without the use of a shared one-time pad [90]. Cheong and Hellman extended this result to Gaussian channels [60], and Csiszár and Körner generalized it to broadcast channels [21]. Our approach considers the adversary's ability to detect rather than decode the transmissions, and it does not rely on the channel to the legitimate receiver being better than the channel to the adversary. However, as discussed in Section 2.2.3, the legitimate receiver having a better channel than the adversary may allow achievability of covert communication without a pre-shared secret.

2.4.2 Anonymous communication

Our problem is related to that of anonymous communication [23], specifically the task of defeating the network traffic timing analysis. While the objective is fundamentally the same, the setting and approaches are vastly different. The network traffic analysis involves the adversary inferring network properties (such as source-relay pairs) by correlating properties (such as the inter-packet timing) of two or more encrypted packet flows. Protecting against this kind of analysis is costly, as one needs to make flows look statistically independent by randomizing the timing of the packets, inserting dummy packets, or dropping a portion of the data packets. Recent work thus addressed the amount of common information that can be embedded into two flows that are generated by independent renewal processes [65]. However, in our

scenario Willie cannot perform traffic analysis (or any kind of network layer analysis), as Alice prevents him (with high probability) from detecting her transmission in the first place.

2.4.3 Cognitive Radio

The covert communication problem is also related to that of establishing a cognitive radio (CR) network [91]. Aspects of covert communication can be cast in the CR context by considering a problem of secondary users communicating while minimizing the interference from their transmissions to the primary users of the network. To do so, secondary users must monitor the channel for primary users and back off if their transmissions are detected. This task is identical to that of Willie in the covert communication scenario. In fact, the work showing that positive-rate covert communication is possible when Willie uses a power detector and has uncertainty about his noise variance was inspired by the primary user detection problem in CR networks [57, 58].

CHAPTER 3

INFORMATION-THEORETICALLY COVERT COMMUNICATION

Quantum and classical information-theoretic analyses of covert communication consider the *reliability* and *detectability* of a transmission. We introduce these concepts after a brief overview of the asymptotic notation used in this thesis. We conclude this chapter by presenting the general mathematical methodology of the proofs that follow.

3.1 Asymptotic Notation

We use asymptotic notation [18, Ch. 3.1] where:

- $f(n) = \mathcal{O}(g(n))$ denotes an asymptotic upper bound on $f(n)$ (i.e., there exist constants $m, n_0 > 0$ such that $0 \leq f(n) \leq mg(n)$ for all $n \geq n_0$),
- $f(n) = o(g(n))$ denotes an upper bound on $f(n)$ that is not asymptotically tight (i.e., for any constant $m > 0$, there exists constant $n_0 > 0$ such that $0 \leq f(n) < mg(n)$ for all $n \geq n_0$),
- $f(n) = \Omega(g(n))$ denotes an asymptotic lower bound on $f(n)$ (i.e., there exist constants $m, n_0 > 0$ such that $0 \leq mg(n) \leq f(n)$ for all $n \geq n_0$),
- $f(n) = \omega(g(n))$ denotes a lower bound on $f(n)$ that is not asymptotically tight (i.e., for any constant $m > 0$, there exists constant $n_0 > 0$ such that $0 \leq mg(n) < f(n)$ for all $n \geq n_0$), and

- $f(n) = \Theta(g(n))$ denotes an asymptotically tight bound on $f(n)$ (i.e., there exist constants $m_1, m_2, n_0 > 0$ such that $0 \leq m_1 g(n) \leq f(n) \leq m_2 g(n)$ for all $n \geq n_0$). $f(n) = \Theta(g(n))$ implies that $f(n) = \Omega(g(n))$ and $f(n) = \mathcal{O}(g(n))$.

3.2 Reliability

We consider a scenario where Alice attempts to transmit M bits to Bob over n uses of the channel while Willie attempts to detect her transmission attempt. A channel use corresponds to a signaling interval carrying one fixed-bandwidth modulation symbol. Each of the 2^M possible M-bit messages maps to an n-symbol *codeword*, and their collection forms a *codebook*. Desirable codebooks ensure that the codewords, when corrupted by the channel, are distinguishable from one another. This provides *reliability*: a guarantee that the probability of Bob's error in decoding Alice's message $\mathbb{P}_e^{(b)} < \delta$ with arbitrarily small $\delta > 0$ for n large enough. In practice, error-correction codes (ECCs) are used to enable reliability.

3.3 Detectability

Willie's detector reduces to a binary hypothesis test of Alice's transmission state given his observations of the channel, where the null hypothesis H_0 corresponds to the hypothesis that Alice does not transmit and the alternate hypothesis H_1 corresponds to the hypothesis that Alice transmits. Denote by $\mathbb{P}_{\text{FA}}$ the probability that Willie raises a false alarm when Alice does not transmit, and by $\mathbb{P}_{\text{MD}}$ the probability that Willie misses the detection of Alice's transmission. We assume equal prior probabilities for hypotheses H_0 and H_1, i.e., $\mathbb{P}(H_0 \text{ true}) = \mathbb{P}(H_1 \text{ true}) = \frac{1}{2}$, which corresponds to Willie's complete ignorance of Alice's transmission state. We examine the impact of using unequal prior probabilities (which corresponds to Willie possessing some information about the likelihood of Alice transmitting) in Appendix A.1 for classical hypothesis testing and in Appendix B.2 for quantum hypothesis testing,

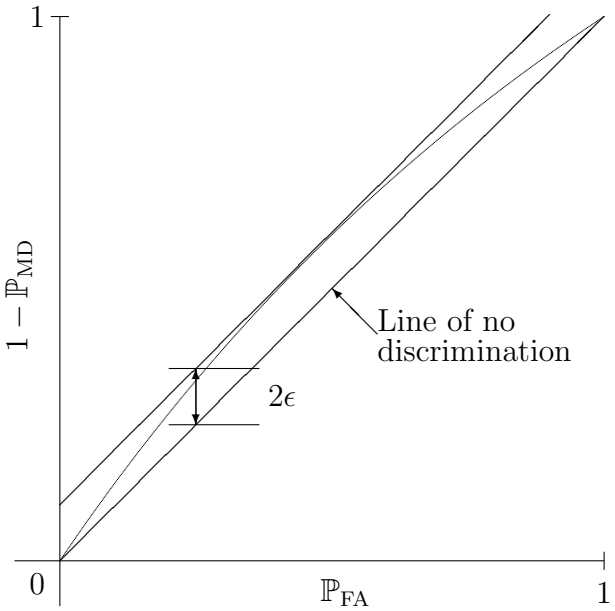


Figure 3.1: Illustration of Willie's ROC curve when Alice's maintains $\mathbb{P}_e^{(w)} \geq 1/2 - \epsilon$. The ROC for a detector that makes random decisions is the diagonal *line of no discrimination*. Since $\mathbb{P}_{\mathrm{FA}} \leq 1 - \mathbb{P}_{\mathrm{MD}}$, $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ implies $\mathbb{P}_{\mathrm{FA}} \leq 1 - \mathbb{P}_{\mathrm{MD}} \leq \mathbb{P}_{\mathrm{FA}} + 2\epsilon$. Thus, Willie's ROC curve is confined to a narrow region near the line of no discrimination.

and find that, unless one of the prior probabilities is unity, the asymptotic results that follow are unaffected. Under the assumption of equal prior probabilities, Willie's *detection error probability*, $\mathbb{P}_e^{(w)} = \frac{\mathbb{P}_{\mathrm{FA}} + \mathbb{P}_{\mathrm{MD}}}{2}$. Since $\mathbb{P}_e^{(w)} = \frac{1}{2}$ for a detector that guesses Alice's transmission state, $\mathbb{P}_e^{(w)} \leq \frac{1}{2}$. We call a signaling scheme *covert* if it ensures $\mathbb{P}_e^{(w)} \geq 1/2 - \epsilon$ for an arbitrarily small $\epsilon > 0$ regardless of Willie's detector choice. This has a natural signal processing interpretation via the receiver operating characteristic (ROC) curve [83, Ch. 2.2.2], which plots the probability of true detection $1 - \mathbb{P}_{\mathrm{MD}}$ versus the probability of false detection $\mathbb{P}_{\mathrm{FA}}$. Since $\mathbb{P}_{\mathrm{FA}} \leq 1 - \mathbb{P}_{\mathrm{MD}}$ and $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ imply that $\mathbb{P}_{\mathrm{FA}} \leq 1 - \mathbb{P}_{\mathrm{MD}} \leq \mathbb{P}_{\mathrm{FA}} + 2\epsilon$, when ϵ is small, the ROC curve lies very close to the line of no discrimination (the diagonal line where $1 - \mathbb{P}_{\mathrm{MD}} = \mathbb{P}_{\mathrm{FA}}$), as illustrated in Figure 3.1. Since the line of no discrimination corresponds to a detec-

tor that guesses Alice's transmission state randomly, a small ϵ implies that the best detector available to Willie does only slightly better than a random guess.

By decreasing her transmission power, Alice can decrease the effectiveness of Willie's hypothesis test at the expense of the reliability of Bob's decoding. *Information-theoretically secure covert communication* is both reliable and covert. To achieve it, prior to transmission, Alice and Bob share a secret, the cost of which we assume to be substantially less than that of being detected by Willie. This secret allows Alice to encode the message in such a way that it is reliably decoded Bob, but not distinguished from the noise by Willie; in fact, it is a codebook secretly shared between Alice and Bob prior to communication in some of the achievability proofs that follow. This follows "best practices" in security system design as the security of the covert communication system depends only on the shared secret [52, 67]. Secret-sharing is also consistent with other information-hiding systems [47, 29, 27, 50, 51, 77]; however, as evidenced by the recent results for a restricted class of channels [16, 46], certain scenarios (e.g., Willie's channel from Alice being worse than Bob's) may allow secret-less covert communication [54].

3.4 Covert Communication Proof Methodology

Each theorem presented in this thesis can be classified as either an "achievability" or a "converse". Achievability theorems (4.2.1, 4.2.2, 5.2.1, 6.3.1, 6.4.1, and 7.1.1) establish the lower limit on the amount of information that can be covertly transmitted from Alice to Bob, while the converse theorems (4.3.1, 5.3.1, 6.2.1 and 6.5.1) demonstrate the upper limit. In essence, the achievability results are obtained by

1. fixing Alice's and Bob's communication system and revealing its construction in entirety (except the shared secret) to Willie;

2. showing that, even with such information, Willie's optimal detector (that also satisfies the constraints of a particular scenario discussed) is ineffective at discriminating Alice's transmission state; and

3. demonstrating that the transmission can be reliably decoded by Bob using the shared secret.

On the other hand, converses are established by

1. fixing Willie's detection scheme (and revealing it to Alice and Bob); and

2. demonstrating that no amount of resources allows Alice to both remain undetected by Willie and exceed the upper limit on the amount of information that is reliably transmitted to Bob.

CHAPTER 4

COVERT COMMUNICATION OVER AWGN CHANNELS

In this chapter we develop the fundamental bounds on covert communication over channels that are subject to additive white Gaussian noise (AWGN). AWGN channel model is standard for many practical communication systems, including wireless devices operating on radio frequencies (RF). In our scenario, Alice has an AWGN channel to Bob, while passive warden Willie attempts to detect transmissions on this channel. The channel between Alice and Willie is also AWGN. Willie is passive in that he only observes and does not actively jam Alice's channel. Willie attempts to classify his observations as either noise on his channel from Alice or Alice's signals to Bob. If he detects communication, Willie can potentially shut the channel down or otherwise punish Alice (however, punishing innocent Alice is costly). If the noise on the channel between Willie and Alice has non-zero power, Alice can communicate with Bob while tolerating a certain probability of detection, which she can drive down by transmitting with low enough power. Thus, Alice potentially transmits non-zero mutual information covertly to Bob in n uses of the channel.

The main result of this chapter is the following theorem:

Theorem 4.1 (Square root law for covert communication over AWGN channels)**.** *Suppose the channels between Alice and each of Bob and Willie experience additive white Gaussian noise (AWGN) with powers $\sigma_b^2 > 0$ and $\sigma_w^2 > 0$, respectively, where σ_b^2 and σ_w^2 are constants. Then, provided that Alice and Bob have a shared secret of sufficient length, for any $\epsilon > 0$ and unknown σ_w^2, Alice can reliably (i.e., with Bob's decoding error probability $\mathbb{P}_e^{(b)} \leq \delta$ for arbitrary $\delta > 0$) transmit $o(\sqrt{n})$ information*

bits to Bob in n channel uses while lower-bounding Willie's detection error probability $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$. Moreover, if Alice knows a lower bound $\hat{\sigma}_w^2 > 0$ to the power of the AWGN on Willie's channel σ_w^2 (i.e. $\sigma_w^2 \geq \hat{\sigma}_w^2$), she can transmit $\mathcal{O}(\sqrt{n})$ bits in n channel uses while maintaining the lower bound $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$. Conversely, if Alice attempts to transmit $\omega(\sqrt{n})$ bits in n channel uses, then, as $n \to \infty$, either Willie detects her with arbitrarily low probability of error or Bob cannot decode her message reliably, regardless of the length of the shared secret.

We note that, since covert communication allows transmission of $\mathcal{O}(\sqrt{n})$ bits in n channel uses and, considering $\lim_{n\to\infty} \frac{\mathcal{O}(\sqrt{n})}{n} = 0$, the information-theoretic capacity of the covert channel is zero, unlike many other communications settings where it is a positive constant. However, a significant amount of information can still be transmitted using this channel. We are thus concerned with the number of information bits transmitted in n channel uses, as opposed to the number of bits per channel use.

After introducing our channel model in Section 4.1, we prove the achievability of the square root law in Section 4.2. We then prove the converse in Section 4.3.

4.1 Channel Model

The discrete-time AWGN channel model with real-valued symbols is a standard mathematical description of the free-space radio-frequency (RF) communication system, as well as of certain optical communications systems (see the introduction to Chapter 6). We defer discussion of the mapping to a continuous-time channel to Appendix A.2. Our formal system framework is depicted in Figure 4.1. Alice transmits a vector of n real-valued symbols $\mathbf{f} = \{f_i\}_{i=1}^n$. Bob receives vector $\mathbf{y}_b = \{y_i^{(b)}\}_{i=1}^n$ where $y_i^{(b)} = f_i + z_i^{(b)}$ with an independent and identically distributed (i.i.d.) sequence $\{z_i^{(b)}\}_{i=1}^n$ of zero-mean Gaussian random variables with variance σ_b^2 (i.e., $z_i^{(b)} \sim \mathcal{N}(0, \sigma_b^2)$). Willie observes vector $\mathbf{y}_w = \{y_i^{(w)}\}_{i=1}^n$ where $y_i^{(w)} = f_i + z_i^{(w)}$, with an i.i.d. sequence $\{z_i^{(w)}\}_{i=1}^n$ of zero-mean Gaussian random variables with variance σ_w^2

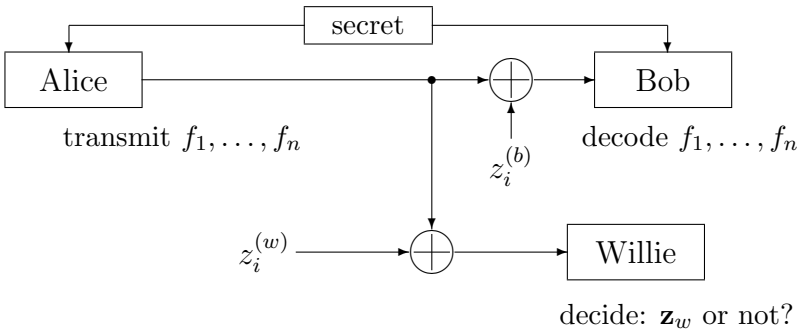


Figure 4.1: System framework: Alice and Bob share a secret before the transmission. Alice encodes information into a vector of real symbols $\mathbf{f} = \{f_i\}_{i=1}^n$ and transmits it on an AWGN channel to Bob, while Willie attempts to classify his vector of observations of the channel from Alice $\mathbf{y}_w$ as either an AWGN vector $\mathbf{z}_w = \{z_i^{(w)}\}_{i=1}^n$ or a vector $\{f_i + z_i^{(w)}\}_{i=1}^n$ of transmissions corrupted by AWGN.

(i.e., $z_i^{(w)} \sim \mathcal{N}(0, \sigma_w^2)$). Thus, when Alice does not transmit (i.e., the null hypothesis H_0 is true), samples are i.i.d. with $y_i^{(w)} \sim \mathcal{N}(0, \sigma_w^2)$; and when Alice transmits (i.e., the alternate hypothesis H_1 is true), samples $y_i^{(w)}$ come from a different distribution.

4.2 Achievability

4.2.1 Information-theoretic analysis of classical hypothesis testing

Willie's objective is to determine whether Alice transmits given the vector of observations $\mathbf{y}_w$ of his channel from Alice. For converse results, we demonstrate existence of a detector that allows Willie to upper-bound $\mathbb{P}_e^{(w)}$ arbitrarily close to zero. The necessary upper bounds are typically derived using probability concentration inequalities such as Chebyshev's and Chernoff's. On the other hand, achievability proofs require analyzing the performance of an arbitrary detector. Here we provide the mathematical machinery for such analysis.

Denote the probability distribution of Willie's channel observations when Alice does not transmit (i.e. when H_0 is true) as $\mathbb{P}_0$, and the probability distribution of the observations when Alice transmits (i.e. when H_1 is true) as $\mathbb{P}_1$. To strengthen

the achievability results, we assume that Alice's channel input distribution, as well as the statistics of the AWGN on the channel between Alice and Willie, are known to Willie. Then $\mathbb{P}_0$ and $\mathbb{P}_1$ are known to Willie, and he can construct an optimal statistical hypothesis test (such as the Neyman–Pearson test) that minimizes the detection error probability $\mathbb{P}_e^{(w)}$ [59, Ch. 13]. The following holds for such a test:

Lemma 4.1 (Theorem 13.1.1 in [59])**.** *For the optimal test,*

$$\mathbb{P}_e^{(w)} = \frac{1}{2} - \frac{1}{2}V(\mathbb{P}_0, \mathbb{P}_1)$$

where $V(\mathbb{P}_0, \mathbb{P}_1)$ is the variational distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ defined as follows:

Definition 4.1 (Variational distance [59])**.** *The* variational distance *(also known as the* total variation distance*) between two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$ is*

$$V(\mathbb{P}_0, \mathbb{P}_1) = \frac{1}{2}\|p_0(x) - p_1(x)\|_1 \tag{4.1}$$

where $p_0(x)$ and $p_1(x)$ are the density functions of $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively, and $\|a-b\|_1$ is the $\mathcal{L}_1$ norm.

Proof of Lemma A.1 in Appendix A.1 includes Lemma 4.1 as a special case, since Lemma A.1 is a generalization of Lemma 4.1 to unequal prior probabilities of hypotheses H_0 and H_1.

Since variational distance lower-bounds the error of all hypothesis tests Willie can use, a clever choice of $\mathbf{f}$ allows Alice to limit Willie's detector performance. Unfortunately, the variational distance is unwieldy for products of probability measures, which are used in the analysis of the vectors of observations. We thus use Pinsker's inequality:

Lemma 4.2 (Pinsker's inequality (Lemma 11.6.1 in [19]))**.**

$$V(\mathbb{P}_0, \mathbb{P}_1) \leq \sqrt{\frac{1}{2} D(\mathbb{P}_0 \| \mathbb{P}_1)}$$

where relative entropy $D(\mathbb{P}_0 \| \mathbb{P}_1)$ is defined as follows:

Definition 4.2. *The* relative entropy *(also known as* Kullback–Leibler divergence*) between two continuous probability measures* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *is:*

$$D(\mathbb{P}_0 \| \mathbb{P}_1) = \int_{\mathcal{X}} p_0(x) \ln \frac{p_0(x)}{p_1(x)} \mathrm{d}x, \tag{4.2}$$

while if the probability measures $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *are discrete, the relative entropy is:*

$$D(\mathbb{P}_0 \| \mathbb{P}_1) = \sum_{x \in \mathcal{X}} p_0(x) \ln \frac{p_0(x)}{p_1(x)}, \tag{4.3}$$

where $\mathcal{X}$ *is the support of* $p_1(x)$*.*

If $\mathbb{P}^n$ is the distribution of a sequence $\{X_i\}_{i=1}^n$ where each $X_i \sim \mathbb{P}$ is i.i.d., then:

Lemma 4.3 (Relative entropy product)**.** *From the chain rule for relative entropy [19, Eq. (2.67)]:*

$$D(\mathbb{P}_0^n \| \mathbb{P}_1^n) = n D(\mathbb{P}_0 \| \mathbb{P}_1)$$

Relative entropy is directly related to Neyman–Pearson hypothesis testing via the Chernoff–Stein Lemma [19, Ch. 11.8]: for a given $\mathbb{P}_{\mathrm{FA}} < \nu$ with $0 < \nu < \frac{1}{2}$, $\lim_{\nu \to 0} \lim_{n \to \infty} \frac{1}{n} \ln \mathbb{P}_{\mathrm{MD}}^* = -D(\mathbb{P}_0 \| \mathbb{P}_1)$ where $\mathbb{P}_{\mathrm{MD}}^* = \min \mathbb{P}_{\mathrm{MD}}$. Thus, upper-bounding the relative entropy limits the performance of the Neyman–Pearson hypothesis test. Indeed, the steganography community often concludes their proofs by

showing an upper bound on the relative entropy [12, 29]. However, we take the extra step of lower-bounding $\mathbb{P}_e^{(w)}$ since it has a natural signal processing interpretation, as described in Section 3.3. Next we show that this results in the square root limit on the amount of information that can be covertly transmitted between Alice and Bob when Bob and Willie both have AWGN channels from Alice.

4.2.2 Achievability of the square root law for covert communication over AWGN channels

We use Taylor's theorem with the Lagrange form of the remainder to upper-bound the relative entropy, and here we restate it as a lemma.

Lemma 4.4 (Taylor's theorem with the remainder)**.** *If $f(x)$ is a function with $n+1$ continuous derivatives on the interval $[u, v]$, then*

$$f(v) = f(u) + f'(u)(v-u) + \ldots + \frac{f^{(n)}(u)}{n!}(v-u)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(v-u)^{n+1},$$

where $f^{(n)}(x)$ denotes the n^{th} derivative of $f(x)$, and ξ satisfies $u \leq \xi \leq v$.

The proof is available in [56, Ch. V.3]. Note that if the remainder term is negative on $[u, v]$, then the sum of the zeroth through n^{th} order terms yields an upper bound on $f(v)$.

We now state the achievability theorem under an average power constraint:

Theorem 4.2.1 (Achievability under an average power constraint)**.** *Suppose Willie's channel is subject to AWGN with average power $\sigma_w^2 > 0$ and suppose that Alice and Bob share a secret of sufficient length. Then Alice can maintain Willie's detection error probability $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ for any $\epsilon > 0$ while reliably transmitting $o(\sqrt{n})$ bits to Bob over n uses of an AWGN channel when σ_w^2 is unknown and $\mathcal{O}(\sqrt{n})$ bits over n channel uses if she knows a lower bound $\sigma_w^2 \geq \hat{\sigma}_w^2$ for some $\hat{\sigma}_w^2 > 0$.*

Proof. **Construction:** Alice's channel encoder takes as inputs blocks of length M bits and encodes them into codewords of length n. We employ random coding arguments and independently generate 2^M codewords $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^M\}$ from $\mathbb{R}^n$ for messages $\{W_k\}_{k=1}^{2^M}$, each according to $p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i)$, where $X \sim \mathcal{N}(0, P_f)$ and P_f is defined later. The codebook is used only to send a single message and is the secret not revealed to Willie, though he knows how it is constructed, including the value of P_f. The size of this secret is discussed following the proof of Theorem 4.2.2.

The channel between Alice and Willie is corrupted by AWGN with power σ_w^2. Willie applies statistical hypothesis testing on a vector of n channel readings $\mathbf{y}_w$ to decide whether Alice transmits. Next we show how Alice can limit the performance of Willie's methods.

Analysis: Consider the case when Alice transmits codeword $\mathbf{c}(W_k)$. Suppose that Willie employs a detector that implements an optimal hypothesis test on his n channel readings. His null hypothesis H_0 is that Alice does not transmit and that he observes noise on his channel. His alternate hypothesis H_1 is that Alice transmits and that he observes Alice's codeword corrupted by noise. By Fact 4.1, Willie's detection error probability is expressed by $\mathbb{P}_e^{(w)} = \frac{1}{2} - \frac{1}{2}V(\mathbb{P}_0, \mathbb{P}_1)$, where the variational distance is between the distribution $\mathbb{P}_0$ of n noise readings that Willie expects to observe under his null hypothesis and the distribution $\mathbb{P}_1$ of the codeword transmitted by Alice corrupted by noise. Alice can lower-bound $\mathbb{P}_e^{(w)}$ by upper-bounding the variational distance: $V(\mathbb{P}_0, \mathbb{P}_1) \leq 2\epsilon$.

The realizations of noise $z_i^{(w)}$ in vector $\mathbf{z}_w$ are zero-mean i.i.d. Gaussian random variables with variance σ_w^2, and, thus, $\mathbb{P}_0 = \mathbb{P}_w^n$ where $\mathbb{P}_w = \mathcal{N}(0, \sigma_w^2)$. Recall that Willie does not know the codebook and that noise is independent of the transmitted symbols. Therefore, when Alice transmits, Willie observes vector $\mathbf{y}_w$, where $y_i^{(w)} \sim \mathcal{N}(0, P_f + \sigma_w^2) = \mathbb{P}_s$ is i.i.d., and thus, $\mathbb{P}_1 = \mathbb{P}_s^n$. By Facts 4.2 and 4.3:

$$V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \sqrt{\frac{1}{2} D(\mathbb{P}_w^n \| \mathbb{P}_s^n)} = \sqrt{\frac{n}{2} D(\mathbb{P}_w \| \mathbb{P}_s)}.$$

In our case the relative entropy is:

$$D(\mathbb{P}_w \| \mathbb{P}_s) = \frac{1}{2} \left[\ln\left(1 + \frac{P_f}{\sigma_w^2}\right) - \left(1 + \left(\frac{P_f}{\sigma_w^2}\right)^{-1}\right)^{-1} \right].$$

Since the first three derivatives of $D(\mathbb{P}_w \| \mathbb{P}_s)$ with respect to P_f are continuous, we can apply Lemma 4.4. The zeroth and first order terms of the Taylor series expansion with respect to P_f around $P_f = 0$ are zero. However, the second order term is:

$$\frac{P_f^2}{2!} \times \left. \frac{\partial^2 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial P_f^2} \right|_{P_f = 0} = \frac{P_f^2}{4\sigma_w^4}.$$

That relative entropy is locally quadratic is well-known; in fact $\left. \frac{\partial^2 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial P_f^2} \right|_{P_f=0} = \frac{1}{2\sigma_w^4}$ is the Fisher information that an observation of noise carries about its power [55, Ch. 2.6]. Now, the remainder term is:

$$\frac{P_f^3}{3!} \times \left. \frac{\partial^3 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial P_f^3} \right|_{P_f = \xi} = \frac{P_f^3}{3!} \times \frac{\xi - 2\sigma_w^2}{(\xi + \sigma_w^2)^4},$$

where ξ satisfies $0 \leq \xi \leq P_f$. Suppose Alice sets her average symbol power $P_f \leq \frac{cf(n)}{\sqrt{n}}$, where $c = 4\epsilon\sqrt{2}$ and $f(n) = \mathcal{O}(1)$ is a function defined later. Since the remainder is negative when $P_f < 2\sigma_w^2$, for n large enough, we can upper-bound relative entropy with the second order term as follows:

$$V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \frac{P_f}{2\sigma_w^2} \sqrt{\frac{n}{2}} \leq \frac{\epsilon f(n)}{\sigma_w^2}. \tag{4.4}$$

In most practical scenarios Alice knows a lower bound $\sigma_w^2 \geq \hat{\sigma}_w^2$ and can set $f(n) = \hat{\sigma}_w^2$ (a conservative lower bound is the thermal noise power of the best currently available receiver). If σ_w^2 is unknown, Alice can set $f(n)$ such that $f(n) = o(1)$

and $f(n) = \omega(1/\sqrt{n})$ (the latter condition is needed to bound Bob's decoding error probability). In either case, Alice upper-bounds $V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq 2\epsilon$, limiting the performance of Willie's detector.

Next we examine the probability $\mathbb{P}_e^{(b)}$ of Bob's decoding error averaged over all possible codebooks. Since Alice's symbol power P_f is a decreasing function of the codeword length n, the standard channel coding results for constant power (and constant rate) do not directly apply. Let Bob employ a maximum-likelihood (ML) decoder (i.e. minimum distance decoder) to process the received vector $\mathbf{y}_b$ when $\mathbf{c}(W_k)$ was sent. The decoder suffers an error event $E_i(\mathbf{c}(W_k))$ when $\mathbf{y}_b$ is closer to another codeword $\mathbf{c}(W_i)$, $i \neq k$. The decoding error probability, averaged over all codebooks, is then:

$$\begin{aligned}
\mathbb{P}_e^{(b)} &= \mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(\cup_{i=1,i\neq k}^{2^M} E_i(\mathbf{c}(W_k))\right)\right] \\
&\leq \mathbb{E}_{\mathbf{c}(W_k)}\left[\sum_{i=1,i\neq k}^{2^M} \mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right] \qquad (4.5) \\
&= \sum_{i=1,i\neq k}^{2^M} \mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right], \qquad (4.6)
\end{aligned}$$

where $\mathbb{E}_X[\cdot]$ denotes the expectation operator over random variable X and (4.5) follows from the union bound. Let $\mathbf{d} = \mathbf{c}(W_k) - \mathbf{c}(W_i)$. Then $\|\mathbf{d}\|_2$ is the distance between two codewords, where $\|\cdot\|_2$ is the $\mathcal{L}_2$ norm. Since codewords are independent and Gaussian, $d_j \sim \mathcal{N}(0, 2P_f)$ for $j = 1, 2, \ldots, n$ and $\|\mathbf{d}\|_2^2 = 2P_f U$, where $U \sim \chi_n^2$, with χ_n^2 denoting the chi-squared distribution with n degrees of freedom. Therefore, by [62, Eq. (3.44)]:

$$\mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right] = \mathbb{E}_U\left[Q\left(\sqrt{\frac{P_f U}{2\sigma_b^2}}\right)\right],$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$. Since $Q(x) \leq \frac{1}{2}e^{-x^2/2}$ [17, Eq. (5)]:

$$\mathbb{E}_U\left[Q\left(\sqrt{\frac{P_f U}{2\sigma_b^2}}\right)\right] \leq \mathbb{E}_U\left[\exp\left(-\frac{P_f U}{4\sigma_b^2}\right)\right]$$

$$= \int_0^\infty \frac{e^{-\frac{P_f u}{4\sigma_b^2}-\frac{u}{2}} 2^{-\frac{n}{2}} u^{\frac{n}{2}-1}}{\Gamma(n/2)} du \tag{4.7}$$

$$= 2^{-n/2}\left(\frac{1}{2}+\frac{P_f}{4\sigma_b^2}\right)^{-n/2}, \tag{4.8}$$

where (4.8) follows from the substitution $v = u\left(\frac{1}{2}+\frac{P_f}{4\sigma_b^2}\right)$ in (4.7) and the definition of the Gamma function $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}\mathrm{d}x$. Since $\frac{1}{2}+\frac{P_f}{4\sigma_b^2} = 2^{\log_2\left(\frac{1}{2}+\frac{P_f}{4\sigma_b^2}\right)}$:

$$\mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right] \leq 2^{-\frac{n}{2}\log_2\left(1+\frac{P_f}{2\sigma_b^2}\right)}$$

for all i, and (4.6) becomes:

$$\mathbb{P}_e^{(b)} \leq 2^{M-\frac{n}{2}\log_2\left(1+\frac{P_f}{2\sigma_b^2}\right)}. \tag{4.9}$$

Since $P_f = \frac{cf(n)}{\sqrt{n}}$ with $f(n) = \omega(1/\sqrt{n})$, if, for some constant $\gamma < 1$, Alice attempts to transmit $M = \frac{n\gamma}{2}\log_2\left(1+\frac{cf(n)}{2\sqrt{n}\sigma_b^2}\right)$ bits, as n increases, the probability of Bob's decoding error averaged over all codebooks decays exponentially to zero. Since $\log_2(1+x) \geq x$ when $x \in [0,1]$, $M \geq \frac{\sqrt{n}\gamma cf(n)}{4\sigma_b^2}$ for large n. Thus, Bob receives $o(\sqrt{n})$ bits in n channel uses, and $\mathcal{O}(\sqrt{n})$ bits in n channel uses if $f(n) = \hat{\sigma}_w^2$. □

Unlike Shannon's coding theorem for AWGN channels [19, Theorem 9.1.1, p. 268], we cannot purge codewords from our codebook to lower the maximal decoding error probability, as that would violate the i.i.d. condition for the codeword construction that is needed to limit Willie's detection ability in our proof. However, it is reasonable that users in sensitive situations attempting to hide their communications would prefer uniform rather than average decoding error performance, in essence demanding that the specific codebook they use be effective. In such a scenario, the

construction of Theorem 4.2.2 can be used with the modification given by the remark following its proof. This construction also satisfies both the peak and the average power constraints, as demonstrated below.

Theorem 4.2.2 (Achievability under a peak power constraint)**.** *Suppose Alice's transmitter is subject to the peak power constraint b, $0 < b < \infty$, and Willie's channel is subject to AWGN with power $\sigma_w^2 > 0$. Also suppose that Alice and Bob share a secret of sufficient length. Then Alice can maintain Willie's detection error probability $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ for any $\epsilon > 0$ while reliably transmitting $o(\sqrt{n})$ bits to Bob over n uses of an AWGN channel when σ_w^2 is unknown and $\mathcal{O}(\sqrt{n})$ bits in n channel uses if she knows a lower bound $\sigma_w^2 \geq \hat{\sigma}_w^2$ for some $\hat{\sigma}_w^2 > 0$.*

To prove Theorem 4.2.2, we introduce a variant of the Leibniz integral rule as a lemma:

Lemma 4.5 (Leibniz integral rule)**.** *Suppose that $f(x, a)$ is defined for $x \geq x_0$ and $a \in [u, v], u < v$, and satisfies the following properties:*

1. *$f(x, a)$ is continuous on $[u, v]$ for $x \geq x_0$;*
2. *$\frac{\partial f(x,a)}{\partial a}$ is continuous on $[u, v]$ for $x \geq x_0$;*
3. *There is a function $g(x)$ such that $|f(x, a)| \leq g(x)$ and $\int_{x_0}^{\infty} g(x)\mathrm{d}x < \infty$;*
4. *There is a function $h(x)$ such that $|\frac{\partial f(x,a)}{\partial a}| \leq h(x)$ and $\int_{x_0}^{\infty} h(x)\mathrm{d}x < \infty$.*

Then $\frac{\partial}{\partial a} \int_{x_0}^{\infty} f(x, a)\mathrm{d}x = \int_{x_0}^{\infty} \frac{\partial f(x,a)}{\partial a}\mathrm{d}x$.

The proof is available in [56, Ch. XIII.3]. We now prove Theorem 4.2.2.

Proof (Theorem 4.2.2). **Construction:** Alice encodes the input in blocks of length M bits into codewords of length n with the symbols drawn from alphabet $\{-a, a\}$, where a satisfies the peak power constraint $a^2 < b$ and is defined later. We independently generate 2^M codewords $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^M\}$ for messages $\{W_k\}$ from

$\{-a, a\}^n$ according to $p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i)$, where $p_X(-a) = p_X(a) = \frac{1}{2}$. As in the proof of Theorem 4.2.1, this single-use codebook is not revealed to Willie, though he knows how it is constructed, including the value of a. While the entire codebook is secretly shared between Alice and Bob, in the remark following the proof we discuss how to reduce the amount of shared secret information.

Analysis: When Alice transmits a symbol during the i^{th} symbol period, she transmits $-a$ or a equiprobably by construction and Willie observes the symbol corrupted by AWGN. Therefore, $\mathbb{P}_s = \frac{1}{2}\left(\mathcal{N}(-a, \sigma_w^2) + \mathcal{N}(a, \sigma_w^2)\right)$, and, with $\mathbb{P}_w = \mathcal{N}(0, \sigma_w^2)$, we have

$$D(\mathbb{P}_w \| \mathbb{P}_s) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\frac{1}{2}\left(e^{-\frac{(x+a)^2}{2\sigma_w^2}} + e^{-\frac{(x-a)^2}{2\sigma_w^2}}\right)} \mathrm{d}x. \tag{4.10}$$

Since (4.10) is an even function, we assume $a \geq 0$.

While there is no closed-form expression for (4.10), its integrand is well-behaved, allowing the application of Lemma 4.4 to (4.10). The Taylor series expansion with respect to a around $a = 0$ can be performed using Lemma 4.5. We demonstrate that the conditions for Lemmas 4.4 and 4.5 hold in Appendix A.3. The zeroth through third order terms of the Taylor series expansion of (4.10) are zero, as is the fifth term. The fourth order term is:

$$\frac{a^4}{4!} \times \left.\frac{\partial^4 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial a^4}\right|_{a=0} = \frac{a^4}{4\sigma_w^4}.$$

Suppose Alice sets $a^2 \leq \frac{cf(n)}{\sqrt{n}}$, where c and $f(n)$ are defined as in Theorem 4.2.1. The sixth derivative of (4.10) with respect to a is:

$$\frac{\partial^6 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial a^6} = -\int_{-\infty}^{\infty} \frac{8x^6 e^{-\frac{x^2}{2\sigma_w^2}}}{\sigma_w^{12}\sqrt{2\pi}\sigma_w} \left[15\operatorname{sech}^6 \frac{ax}{\sigma_w^2} - 15\operatorname{sech}^4 \frac{ax}{\sigma_w^2} + 2\operatorname{sech}^2 \frac{ax}{\sigma_w^2}\right] \mathrm{d}x, \tag{4.11}$$

where $\operatorname{sech} x = \frac{2}{e^x + e^{-x}}$ is the hyperbolic secant function. Evaluated at zero, the sixth derivative is $\left.\frac{\partial^6 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial a^6}\right|_{a=0} = -\frac{240}{\sigma_w^6}$. Since (4.11) is continuous (see Appendix A.3), there exists a neighborhood $[0, \mu]$ such that, for all $\xi \in [0, \mu]$, the remainder term $\frac{a^6}{6!} \times \left.\frac{\partial^6 D(\mathbb{P}_w \| \mathbb{P}_s)}{\partial a^6}\right|_{a=\xi} \leq 0$. Then, for n large enough, we can apply Lemma 4.4 to upper-bound relative entropy with the fourth order term as follows:

$$V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \frac{a^2}{2\sigma_w^2}\sqrt{\frac{n}{2}} \leq \frac{2\epsilon f(n)}{\sigma_w^2}. \tag{4.12}$$

Since the power of Alice's symbol is $a^2 = P_f$, (4.12) is identical to (4.4) and Alice obtains the upper bound $V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq 2\epsilon$, limiting the performance of Willie's detector.

Next let's examine the probability $\mathbb{P}_e^{(b)}$ of Bob's decoding error averaged over all possible codebooks. As in Theorem 4.2.1, we cannot directly apply the standard constant-power channel coding results to our system where the symbol power is a decreasing function of the codeword length. We upper-bound Bob's decoding error probability by analyzing a suboptimal decoding scheme. Suppose Bob uses a hard-decision device on each received symbol $y_i^{(b)} = f_i + z_i^{(b)}$ via the rule $\hat{f}_i = \left\{a \text{ if } y_i^{(b)} \geq 0; -a \text{ otherwise}\right\}$, and applies an ML decoder on its output. The effective channel for the encoder/decoder pair is a binary symmetric channel with cross-over probability $p_e = Q(a/\sigma_b)$ and the probability of the decoding error averaged over all possible codebooks is [32, Theorem 5.6.2]:

$$\mathbb{P}_e^{(b)} \leq e^{M - nE_0(p_e)}, \tag{4.13}$$

where $E_0(p_e) = \ln(2) - 2\ln\left(\sqrt{1-p_e} + \sqrt{p_e}\right)$ is the error exponent for a BSC from [32, Theorem 5.6.2] with parameter $\rho = 1$. We expand the analysis in [63, Section I.2.1] to characterize the rate R. We use Lemma 4.4 to upper-bound

$$p_e \leq \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\left(\frac{a}{\sigma_b} - \frac{a^3}{6\sigma_b^3}\right) \triangleq p_e^{(UB)},$$

where $p_e^{(UB)}$ denotes the sum of the zeroth through second terms of the Taylor series expansion of $Q(a/\sigma_b)$ around $a = 0$. The remainder term is non-positive for a/σ_b satisfying $\frac{8a^6}{\sigma_b^6} - \frac{60a^4}{\sigma_b^4} + \frac{90a^2}{\sigma_b^2} - 15 \leq 0$, and, since $a^2 = \frac{cf(n)}{\sqrt{n}}$, the upper bound thus holds for large enough n. Since $E_0(p)$ is a monotonically increasing function on the interval $\left[0, \frac{1}{2}\right]$, $E_0(p_e) \leq E_0(p_e^{(UB)})$. The Taylor series expansion of $E_0(p_e^{(UB)})$ with respect to a around $a = 0$ yields $E_0(p_e^{(UB)}) = \frac{a^2}{2\pi\sigma_b^2} + \mathcal{O}(a^4)$. Substituting $a^2 = \frac{cf(n)}{\sqrt{n}}$, we obtain $\mathbb{P}_e^{(b)} \leq e^{M - \frac{\sqrt{n}cf(n)}{2\pi\sigma_b^2} + \mathcal{O}(1)}$. Since $f(n) = \omega(1/\sqrt{n})$, if Alice attempts to transmit $M = \frac{\gamma c f(n)\sqrt{n}}{2\pi\sigma_b^2 \ln 2}$ bits with a constant $\gamma < 1$, the probability of Bob's decoding error averaged over all codebooks decays exponentially to zero as n increases and Bob obtains $M = o(\sqrt{n})$ bits in n channel uses, and $\mathcal{O}(\sqrt{n})$ bits in n channel uses if $f(n) = \hat{\sigma}_w^2$. □

4.2.3 Remarks

4.2.3.1 Employing the best codebook

The proof of Theorem 4.2.2 guarantees Bob's decoding error performance averaged over all binary codebooks. Following the standard coding arguments [19, p. 204], there must be at least one binary alphabet codebook that has at least average probability of error. Thus, to guarantee uniform performance, Alice and Bob must select "good" codebooks for communications. However, choosing specific codebooks would violate the i.i.d. condition for the codeword construction that is needed to limit Willie's detection capability in our proof.

Consider a codebook that has at least average probability of error, but now assume that it is public (i.e. known to Willie). Theorem 4.2.2 shows that Alice can use it to transmit $\mathcal{O}(\sqrt{n})$ bits to Bob in n channel uses with exponentially-decaying probability of error. However, since the codebook is public, unless Alice and Bob

take steps to protect their communication, Willie can use this codebook to detect Alice's transmissions by performing the same decoding as Bob. Here we demonstrate that to use a public codebook it suffices for Alice and Bob to share a secret random binary vector and note that this resembles the one-time pad scheme from traditional cryptography [76], but employed here for a very different application.

Suppose that, prior to communication, Alice and Bob generate and share binary vector $\mathbf{k}$ where $p_{\mathbf{K}}(\mathbf{k}) = \prod_{i=1}^{n} p_K(k_i)$ with $p_K(0) = p_K(1) = \frac{1}{2}$. Alice XORs $\mathbf{k}$ and the binary representation of the codeword $\mathbf{c}(W_k)$, resulting in an equiprobable transmission of $-a$ and a when Alice transmits a symbol during the i^{th} symbol period. Provided $\mathbf{k}$ is never re-used and is kept secret from Willie, the i.i.d. assumption for the vector $\mathbf{y}_w$ in Theorem 4.2.2 holds without the need to exchange an entire secret codebook between Alice and Bob. Bob decodes by XORing $\mathbf{k}$ with the output of the hard-decision device prior to applying the ML decoder. While the square root law implies that the shared $\mathcal{O}(n)$-bit secret here is quadratic in the length $M = \mathcal{O}(\sqrt{n})$ of a message, we offer a coding scheme that, on average, requires an $\mathcal{O}(\sqrt{n}\log n)$-bit secret in Appendix A.4. The development of covert communication with a shared secret either linear or sublinear in the message size is a subject of the ongoing research.

4.2.3.2 Relationship with Square Root Law in Steganography

The covert communication problem is related to steganography. A comprehensive review of steganography is available in a book by Fridrich [29]. In finite-alphabet imperfect steganographic systems at most $\mathcal{O}(\sqrt{n})$ symbols in the original covertext of length n may safely be modified to hide a steganographic message of length $\mathcal{O}(\sqrt{n}\log n)$ bits [29, Ch. 13] [48]. This result was extended to Markov covertext [27] and was shown to either require a key linear in the size of the message [50] or encryption of the message prior to embedding [51].

The square root law in steganography has the same form as our square root law because both laws follow from the property that relative entropy is locally quadratic [55, Ch. 2.6]: $D(\mathbb{P}_0\|\mathbb{P}_1) = \frac{\delta^2}{2}J(\theta) + \mathcal{O}(\delta^3)$, where $J(\theta) = \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 f(x;\theta)\mathrm{d}x$ is the Fisher information associated with parameter θ, and $\mathbb{P}_0$ and $\mathbb{P}_1$ are probability measures with density functions from the same family over the support $\mathcal{X}$, but with parameters differing by δ: $p_0(x) = f(x;\theta)$ and $p_1(x) = f(x;\theta+\delta)$. Fisher information is thus used as a metric for steganographic security [26, 49].

In a typical steganography scenario with a passive warden, coding techniques similar to Hamming codes allow embedding of $\log(n)$ bits per changed symbol [29, Ch. 8], which make hiding $\mathcal{O}(\sqrt{n}\log n)$ bits in n symbols possible. However, because of the noise on the channel between Alice and Bob, and the resultant need for error correction, no more than $\mathcal{O}(\sqrt{n})$ bits can be transmitted both reliably and covertly in n channel uses, as we prove in the following section.

4.3 Converse

Here, as in the proof of achievability, the channel between Alice and Bob is AWGN with power σ_b^2. Alice desires to transmit one of 2^M (equally likely) M-bit messages to Bob in n channel uses, where $M = \omega(\sqrt{n})$, with arbitrarily small probability of decoding error as n gets large, while limiting Willie's ability to detect her transmission. To this end, Alice encodes each message arbitrarily into n symbols.

Willie observes all n of Alice's channel uses, but he is oblivious to her signal properties and employs only a simple power detector. Nevertheless, we prove that, even if Willie only has these limited capabilities, Alice cannot transmit a message with $\omega(\sqrt{n})$ bits of information in n channel uses without either being detected by Willie or having Bob suffer a non-zero decoding error probability.

Theorem 4.3.1. *If over n channel uses, Alice attempts to transmit a message to Bob that is $\omega(\sqrt{n})$ bits long, then, as $n \to \infty$, either Willie can detect her with high probability, or Bob cannot decode with arbitrarily low probability of error.*

Proof. Suppose Alice employs an arbitrary codebook $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^M\}$. Detection of Alice's transmissions entails Willie deciding between the following hypotheses:

$$\begin{aligned} H_0 : &\quad y_i^{(w)} = z_i^{(w)},\ i = 1, \ldots, n \\ H_1 : &\quad y_i^{(w)} = f_i + z_i^{(w)},\ i = 1, \ldots, n \end{aligned}$$

Suppose Willie uses a power detector to perform the hypothesis test as follows: first, he collects a row vector of n independent readings $\mathbf{y}_w$ from his channel to Alice. Then he generates the test statistic $S = \frac{\mathbf{y}_w \mathbf{y}_w^T}{n}$ where $\mathbf{x}^T$ denotes the transpose of vector $\mathbf{x}$, and rejects or accepts the null hypothesis based on a comparison of S to a threshold that we discuss later. We first show how Willie can bound the error probabilities $\mathbb{P}_{\text{FA}}$ and $\mathbb{P}_{\text{MD}}$ of the power detector as a function of Alice's signal parameters. Then we show that if Alice's codebook prevents Willie's test from detecting her, Bob cannot decode her transmissions without error.

If the null hypothesis H_0 is true, Alice does not transmit and Willie observes AWGN on his channel. Thus, $y_i^{(w)} \sim \mathcal{N}(0, \sigma_w^2)$, and the mean and the variance of S when H_0 is true are:

$$\mathbb{E}[S] = \sigma_w^2 \tag{4.14}$$

$$\text{Var}[S] = \frac{2\sigma_w^4}{n} \tag{4.15}$$

Suppose Alice transmits codeword $\mathbf{c}(W_k) = \{f_i^{(k)}\}_{i=1}^n$. Then Willie's vector of observations $\mathbf{y}_{w,k} = \{y_i^{(w,k)}\}_{i=1}^n$ contains readings of mean-shifted noise $y_i^{(w,k)} \sim$

$\mathcal{N}(f_i^{(k)}, \sigma_w^2)$. The mean of each squared observation is $\mathbb{E}[y_i^2] = \sigma_w^2 + \left(f_i^{(k)}\right)^2$ and the variance is $\mathrm{Var}[y_i^2] = \mathbb{E}[y_i^4] - (\mathbb{E}[y_i^2])^2 = 4\left(f_i^{(k)}\right)^2 \sigma_w^2 + 2\sigma_w^4$. Denote the average symbol power of codeword $\mathbf{c}(W_k)$ by $P_k = \frac{\mathbf{c}(W_k)\mathbf{c}^T(W_k)}{n}$. Then the mean and variance of S when Alice transmits codeword $\mathbf{c}(W_k)$ are:

$$\mathbb{E}[S] = \sigma_w^2 + P_k \tag{4.16}$$

$$\mathrm{Var}[S] = \frac{4P_k\sigma_w^2 + 2\sigma_w^4}{n} \tag{4.17}$$

The variance of Willie's test statistic (4.17) is computed by adding the variances conditioned on $\mathbf{c}(W_k)$ of the squared individual observations $\mathrm{Var}[y_i^2]$ (and dividing by n^2) since the noise on the individual observations is independent.

The probability distribution for the vector of Willie's observations depends on which hypothesis is true. Denote by $\mathbb{P}_0$ the distribution when H_0 holds, and $\mathbb{P}_1^{(k)}$ when H_1 holds with Alice transmitting message W_k. While $\mathbb{P}_1^{(k)}$ is conditioned on Alice's codeword, we show that the average symbol power $P_k = \frac{\mathbf{c}(W_k)\mathbf{c}^T(W_k)}{n}$ of codeword $\mathbf{c}(W_k)$ determines its detectability by this detector, and that our result applies to all codewords with power of the same order.

If H_0 is true, then S should be close to (4.14). Willie picks a threshold t and compares the value of S to $\sigma_w^2 + t$. He accepts H_0 if $S < \sigma_w^2 + t$ and rejects it otherwise. Suppose that he desires false positive probability $\mathbb{P}_{\mathrm{FA}}^*$, which is the probability that $S \geq \sigma_w^2 + t$ when H_0 is true. We bound it using (4.14) and (4.15) with Chebyshev's Inequality [19, Eq. (3.32)]:

$$\begin{aligned}
\mathbb{P}_{\mathrm{FA}} &= \mathbb{P}_0\left(S \geq \sigma_w^2 + t\right) \\
&\leq \mathbb{P}_0\left(|S - \sigma_w^2| \geq t\right) \\
&\leq \frac{2\sigma_w^4}{nt^2}
\end{aligned}$$

Thus, to obtain $\mathbb{P}^*_{\text{FA}}$, Willie sets $t = \frac{d}{\sqrt{n}}$, where $d = \frac{\sqrt{2}\sigma_w^2}{\sqrt{\mathbb{P}^*_{\text{FA}}}}$ is a constant. As n increases, t decreases, which is consistent with Willie gaining greater confidence with more observations.

Suppose Alice transmits codeword $\mathbf{c}(W_k)$. Then the probability of a miss $\mathbb{P}^{(k)}_{\text{MD}}$ is the probability that $S < \sigma_w^2 + t$, where $t = \frac{d}{\sqrt{n}}$. We bound $\mathbb{P}^{(k)}_{\text{MD}}$ using (4.16) and (4.17) with Chebyshev's Inequality:

$$\begin{aligned} \mathbb{P}^{(k)}_{\text{MD}} &= \mathbb{P}^{(k)}_1\left(S < \sigma_w^2 + t\right) \\ &\leq \mathbb{P}^{(k)}_1\left(\left|S - \sigma_w^2 - P_k\right| \geq P_k - t\right) \\ &\leq \frac{4P_k\sigma_w^2 + 2\sigma_w^4}{(\sqrt{n}P_k - d)^2} \end{aligned} \tag{4.18}$$

If the average symbol power $P_k = \omega(1/\sqrt{n})$, $\lim_{n\to\infty} \mathbb{P}^{(k)}_{\text{MD}} = 0$. Thus, with enough observations, Willie can detect with arbitrarily low error probability Alice's codewords with the average symbol power $P_k = \frac{\mathbf{c}(W_k)\mathbf{c}^T(W_k)}{n} = \omega(1/\sqrt{n})$. Note that Willie's detector is oblivious to any details of Alice's codebook construction.

On the other hand, if the transmitted codeword has the average symbol power $P_{\mathcal{U}} = \mathcal{O}(1/\sqrt{n})$, then (4.18) does not upper-bound the probability of a missed detection arbitrarily close to zero regardless of the number of observations. Thus, if Alice desires to lower-bound Willie's detection error probability by $\mathbb{P}^{(w)}_e \geq \zeta > 0$, her codebook must contain a positive fraction γ of such low-power codewords. Let's denote this subset of codewords with the average symbol power $P_{\mathcal{U}} = \mathcal{O}(1/\sqrt{n})$ as $\mathcal{U}$ and examine the probability of Bob's decoding error $\mathbb{P}^{(b)}_e$. The probability that a message from set $\mathcal{U}$ is sent is $\mathbb{P}(\mathcal{U}) = \gamma$, as all messages are equiprobable. We bound $\mathbb{P}^{(b)}_e = \mathbb{P}_e(\mathcal{U})\,\mathbb{P}(\mathcal{U}) + \mathbb{P}_e\left(\overline{\mathcal{U}}\right)\mathbb{P}\left(\overline{\mathcal{U}}\right) \geq \gamma\mathbb{P}_e(\mathcal{U})$, where $\overline{\mathcal{U}}$ is the complement of $\mathcal{U}$ and $\mathbb{P}_e(\mathcal{U})$ is the probability of decoding error when a message from $\mathcal{U}$ is sent:

$$\mathbb{P}_e(\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{W\in\mathcal{U}} \mathbb{P}_e(\mathbf{c}(W) \text{ sent}) \tag{4.19}$$

where $\mathbb{P}_e\left(\mathbf{c}(W) \text{ sent}\right)$ is the probability of error when codeword $\mathbf{c}(W)$ is transmitted, $|\cdot|$ denotes the set cardinality operator, and (4.19) holds because all messages are equiprobable.

When Bob uses the optimal decoder, $\mathbb{P}_e\left(\mathbf{c}(W) \text{ sent}\right)$ is the probability that Bob decodes the received signal as $\hat{W} \neq W$. This is the probability of a union of events E_j, where E_j is the event that sent message W is decoded as some other message $W_j \neq W$:

$$\begin{aligned} \mathbb{P}_e\left(\mathbf{c}(W) \text{ sent}\right) &= \mathbb{P}\left(\cup_{j=1, W_j \neq W}^{2^M} E_j\right) \\ &\geq \mathbb{P}\left(\cup_{W_j \in \mathcal{U} \setminus \{W\}} E_j\right) \triangleq \mathbb{P}_e^{(\mathcal{U})} \end{aligned} \tag{4.20}$$

Here the inequality in (4.20) is from the observation that the sets in the second union are contained in the first. From the decoder perspective, this is because the decoding error probability decreases when Bob knows that the message is from $\mathcal{U}$ (the set of messages on which the decoder can err is reduced).

Our analysis of $\mathbb{P}_e^{(\mathcal{U})}$ uses Cover's simplification of Fano's inequality similar to the proof of the converse to the coding theorem for Gaussian channels in [19, Ch. 9.2]. Since we are interested in $\mathbb{P}_e^{(\mathcal{U})}$, we do not absorb it into ϵ_n as done in (9.37) of [19]. Rather, we explicitly use:

$$H(W|\hat{W}) \leq 1 + (\log_2 |\mathcal{U}|)\mathbb{P}_e^{(\mathcal{U})} \tag{4.21}$$

where $H(W|\hat{W})$ denotes the entropy of message W conditioned on Bob's decoding $\hat{W}$ of W.

Noting that the size of the set $\mathcal{U}$ from which the messages are drawn is $\gamma 2^M$ and that, since each message is equiprobable, the entropy of a message W from $\mathcal{U}$ is

$H(W) = \log_2 |\mathcal{U}| = \log_2 \gamma + M$, we utilize (4.21) and carry out steps (9.38)–(9.53) in [19] to obtain:

$$\mathbb{P}_e^{(\mathcal{U})} \geq 1 - \frac{P_{\mathcal{U}}/2\sigma_b^2 + 1/n}{\frac{\log_2 \gamma}{n} + \frac{M}{n}} \tag{4.22}$$

Since Alice transmits $M = \omega(\sqrt{n})$ bits in n channel uses, $\frac{M}{n} = \omega(1/\sqrt{n})$. However, $P_{\mathcal{U}} = O(1/\sqrt{n})$, and, as $n \to \infty$, $\mathbb{P}_e^{(\mathcal{U})}$ is bounded away from zero. Since $\gamma > 0$, $\mathbb{P}_e^{(b)}$ is bounded away from zero if Alice tries to transmit $\omega(\sqrt{n})$ bits reliably while beating Willie's simple power detector. □

Goodput of Alice's Communication

Define the goodput $G(n)$ of Alice's communication as the average number of bits that Bob can receive from Alice over n channel uses with non-zero probability of a message being undetected as $n \to \infty$. Since only $\mathcal{U}$ contains such messages, by (4.22), the probability of her message being successfully decoded by Bob is $\mathbb{P}_s^{(\mathcal{U})} = 1 - \mathbb{P}_e^{(\mathcal{U})} = \mathcal{O}\left(\frac{\sqrt{n}}{M}\right)$ and the goodput is $G(n) = \gamma \mathbb{P}_s^{(\mathcal{U})} M = \mathcal{O}(\sqrt{n})$. Thus, Alice cannot break the square root law using an arbitrarily high transmission rate and retransmissions while keeping the power below Willie's detection threshold.

CHAPTER 5

WARDEN'S IGNORANCE OF TRANSMISSION TIME INCREASES COVERT THROUGHPUT

In the previous chapter we assume that Willie knows when Alice starts transmitting (if she transmits). However, there are many practical scenarios where this assumption can be relaxed and Alice's time of communication is unknown to Willie. Alice's message may also be much shorter than the total time available to transmit it (e.g., a few seconds out of the day when both Alice and Bob are available). Thus, since Willie does not know when Alice transmits, he has to monitor a much longer time period than the duration of Alice's transmission. In this chapter we show how Alice can leverage Willie's ignorance of her transmission time to transmit significant additional information to Bob.

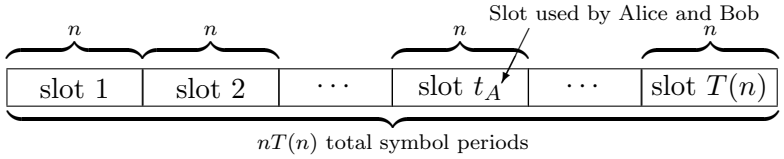


Figure 5.1: Slotted channel: each of the $T(n)$ slots contains n symbol periods. Alice and Bob use slot t_A to communicate (reprint of Figure 2.4).

In our scenario, Alice communicates with Bob over an additive white Gaussian noise (AWGN) channel. Willie also has an AWGN channel from Alice. Unlike the setting in Chapter 4, the channel is slotted, as described in Figure 5.1. Each of $T(n)$ slots contains n symbol periods, where $T(n)$ is an increasing function of n. If Alice used all $nT(n)$ symbol periods for transmission, then, by the square root law in Chapter 4, she could reliably transmit $\mathcal{O}(\sqrt{nT(n)})$ covert bits to Bob. However,

Alice uses only a single slot t_A, which she keeps secret from Willie, who is thus forced to monitor all $T(n)$ slots. A naïve application of the square root law from Chapter 4 allows Alice to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits in this scenario. We demonstrate that Alice can transmit $\mathcal{O}\left(\min\{\sqrt{n\log T(n)}, n\}\right)$ bits reliably on this channel while maintaining arbitrarily low probability of detection by Willie. Conversely, we show that the transmission of $\omega(\sqrt{n\log T(n)})$ bits either results in Alice being detected with high probability or unreliable communication.

The cost of covert communication on the AWGN channel is the secret that Alice and Bob share before the transmission. Remarkably, we demonstrate that the *multiplicative* increase (by a factor of $\sqrt{\log T(n)}$) in the number of covert bits that Alice can transmit reliably to Bob comes without any increase in the size of the pre-shared secret if $T(n) < 2^{c_T n}$, where c_T is a constant; to realize the $\sqrt{\log T(n)}$ gain when $T(n) \geq 2^{c_T n}$ only an additive expense of an extra $\log T(n)$ secret bits is needed to indicate to Bob the slot employed by Alice. Timing is thus a very efficient resource for covert communication. It also necessitates a vastly different analysis than that of the power alone in Chapter 4. Specifically, the relative entropy based bounds on the probability of detection error employed in Chapter 4 are too loose to yield our achievability results, and we therefore have to apply other techniques from mathematical statistics.

The main result of this chapter is the following theorem:

Theorem 5.1. *Suppose the channel between Alice and each of Bob and Willie experiences independent additive white Gaussian noise (AWGN) with constant power $\sigma_b^2 > 0$ and $\sigma_w^2 > 0$, respectively. Also suppose that, if Alice chooses to transmit, she uses one of the $T(n)$ slots chosen randomly. Each slot contains n symbol periods, where $T(n) = \omega(1)$. Then, for any $\epsilon > 0$, there exists n_0 such that, for all $n \geq n_0$, Alice can reliably transmit $\mathcal{O}\left(\min\{\sqrt{n\log T(n)}, n\}\right)$ bits to Bob in a selected slot while maintaining a probability of detection error by Willie greater than $\frac{1}{2} - \epsilon$. Conversely,*

if Alice tries to transmit $\omega(\sqrt{n \log T(n)})$ bits using n consecutive symbol periods, either Willie detects with arbitrarily low probability of error or Bob cannot decode her message with arbitrary low probability of decoding error.

After introducing our slotted channel model in Section 5.1, we prove the achievability and the converse in Sections 5.2 and 5.3, respectively. We conclude this chapter by discussing its relationship to steganography in Section 5.4.

5.1 Channel Model

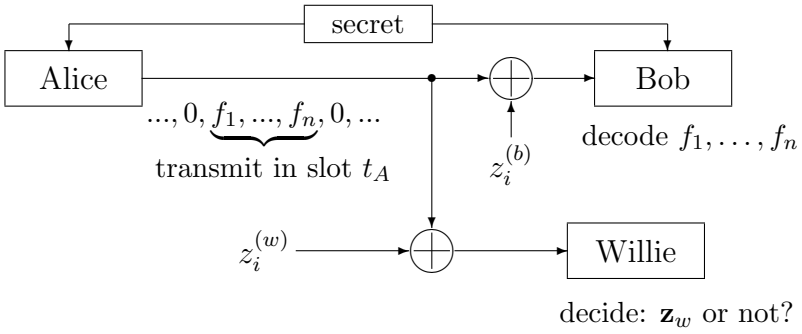


Figure 5.2: System framework: Alice and Bob share a secret before transmission. If Alice chooses to transmit, she encodes information into a vector of real symbols $\mathbf{f} = \{f_i\}_{i=1}^{n}$ and uses random slot t_A to send it on an AWGN channel to Bob (to ensure reliable decoding t_A is secretly shared with Bob before the transmission if $T(n) \geq 2^{c_T n}$, where c_T is a constant). Upon observing the channel from Alice, Willie has to classify his vector of readings $\mathbf{y}_w$ as either an AWGN vector $\mathbf{z}_w = \{z_i^{(w)}\}_{i=1}^{nT(n)}$ or an AWGN vector that contains a slot with transmissions corrupted by AWGN.

We use the discrete-time slotted AWGN channel model with real-valued symbols depicted in Figures 5.1 and 5.2. The channel has $T(n)$ slots, each containing n symbol periods. Alice selects slot t_A uniformly at random prior to transmission; it is shared secretly with Bob before the transmission if $T(n) \geq 2^{c_T n}$, where c_T is a constant that we determine later. If Alice chooses to transmit, she uses t_A to send a vector of n real-valued symbols $\mathbf{f} = \{f_i\}_{i=1}^{n}$. Bob receives a vector $\mathbf{y}_b = \{\mathbf{y}_b(t)\}_{t=1}^{T(n)}$ where $\mathbf{y}_b(t) = [y_{(t-1)n+1}^{(b)}, \ldots, y_{tn}^{(b)}]$ is a vector of observations of slot t, $y_{(t_A-1)n+i}^{(b)} =$

$f_i + z^{(b)}_{(t_A-1)n+i}$ and $y^{(b)}_{(t-1)n+i} = z^{(b)}_{(t-1)n+i}$ for all $t \neq t_A$, with an independent and identically distributed (i.i.d.) sequence $\{z^{(b)}_i\}_{i=1}^{nT(n)}$ of zero-mean Gaussian random variables with variance σ_b^2 (i.e., $z^{(b)}_i \sim \mathcal{N}(0, \sigma_b^2)$). Similarly, Willie observes vector $\mathbf{y}_w = \{\mathbf{y}_w(t)\}_{t=1}^{T(n)}$ where $\mathbf{y}_w(t) = [y^{(w)}_{(t-1)n+1}, \ldots, y^{(w)}_{tn}]$ is a vector of observations of slot t, $y^{(w)}_{(t_A-1)n+i} = f_i + z^{(w)}_{(t_A-1)n+i}$ and $y^{(w)}_{(t-1)n+i} = z^{(w)}_{(t-1)n+i}$ for all $t \neq t_A$, with an i.i.d. sequence $\{z^{(w)}_i\}_{i=1}^{nT(n)}$ of zero-mean Gaussian random variables with variance σ_w^2 (i.e., $z^{(w)}_i \sim \mathcal{N}(0, \sigma_w^2)$). Willie does not know t_A and has to examine the entire $\mathbf{y}_w$ to determine whether Alice is communicating.

5.2 Achievability

We first state the achievability theorem, then discuss the proof idea before proceeding with the proof.

Theorem 5.2.1 (Achievability)**.** *Suppose Alice has a slotted AWGN channel to Bob with $T(n) = \omega(1)$ slots, each containing n symbol periods. Then, provided that Alice and Bob have a secret of sufficient length, if Alice chooses to, she can transmit $\mathcal{O}\left(\min\{\sqrt{n \log T(n)}, n\}\right)$ bits in a single slot while $\lim_{n\to\infty} \mathbb{P}_e^{(w)} > \frac{1}{2} - \epsilon$ and $\lim_{n\to\infty} \mathbb{P}_e^{(b)} \leq \delta$ for arbitrary $\epsilon > 0$ and $\delta > 0$.*

Proof idea. The techniques used to bound the performance of Willie's optimal detector in the proofs of Theorems 4.2.1 and 4.2.2 are ineffective here, as the resulting bounds lack the necessary tightness. Therefore, we take a different approach by explicitly deriving the test statistic for Willie's optimal detector assuming that Alice's channel input distribution, as well as the distribution of the AWGN on the channel between Alice and Willie, are known to Willie. Furthermore, it is assumed that Alice confines her transmission (if she transmits) to one of the slots depicted in Figure 5.2 with the slot boundaries known to Willie. Since Willie has to discriminate between two simple hypotheses on Alice's transmission state, the optimal detector, given by the Neyman–Pearson lemma, employs the likelihood ratio test (LRT) [59, Ch. 3.2].

Willie's LRT statistic is a sum of $T(n)$ independent random variables. Regardless of Alice's transmission state, $T(n)-1$ terms of this sum are identically distributed since they correspond to the slots that Alice does not select for potential transmission (i.e., slots that are not t_A). We show that these $T(n)-1$ terms are i.i.d. zero-mean unit-variance random variables, each weighted by $\frac{1}{\sqrt{T(n)-1}}$. Thus, by the central limit theorem, asymptotically, their sum is a standard Gaussian random variable $Z \sim \mathcal{N}(0,1)$. The distribution of the term that corresponds to slot t_A depends on Alice's transmission state. This term effectively offsets Z's mean away from zero, however, we show that it converges to zero in probability under either hypothesis as long as Alice uses per-symbol power $P_f = \mathcal{O}\left(\frac{\sqrt{\log T(n)}}{\sqrt{n}}\right)$. This allows us to lower-bound $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ for a sufficiently large n and prove the covertness of Alice's transmission. We conclude the proof by extending the analysis of Bob's probability of error from the proof of Theorem 4.2.1 to $T(n)$ slots via a standard union bound.

Proof. (Theorem 5.2.1) **Construction:** Alice secretly selects slot t_A uniformly at random out of the $T(n)$ slots in which to communicate. Alice's channel encoder takes as input blocks of length M bits and encodes them into codewords of length n symbols. We employ random coding arguments and independently generate 2^M codewords $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^M\}$ from $\mathbb{R}^n$ for messages $\{W_k\}_{k=1}^{2^M}$, each according to $p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i)$, where $X \sim \mathcal{N}(0, P_f)$ and symbol power $P_f < \frac{\sigma_w^2}{2}$ is defined later. The codebook is used only to send a single message and, along with t_A, is the secret not revealed to Willie, though he knows how it is constructed, including the value P_f.

Another way of viewing the construction is as a choice of one of $T(n)$ codebooks, where the i^{th} codebook has a block of non-zero symbols in the i^{th} slot. Agreement on the timing is equivalent to selection of the t_A-th codebook and the message is encoded by choosing a codeword from the selected codebook.

Analysis (Willie): Willie is interested in performing the following hypothesis test on his vector of observations $\mathbf{y}_w$:

$$H_0 : \text{Alice does not transmit}$$
$$H_1 : \exists \text{ a slot } t_A \in \{1, \ldots, T(n)\} \text{ in which Alice transmits}$$

Let $Y_t = \sum_{y_i \in \mathbf{y}_w(t)} y_i^2$ be the power in slot t. Since Willie's channel from Alice is corrupted by AWGN with power σ_w^2, the likelihood function of the observations $\mathbf{y}_w$ under H_0 is:

$$f_0(\mathbf{y}_w) = \left(\frac{1}{2\pi\sigma_w^2}\right)^{\frac{nT(n)}{2}} \exp\left[-\frac{1}{2\sigma_w^2}\sum_{t=1}^{T(n)} Y_t\right]. \tag{5.1}$$

Since Willie does not know which of the $T(n)$ slots Alice and Bob randomly select for communication, nor the codebook they use, but knows that Alice's signal is Gaussian, the likelihood function of the observations $\mathbf{y}_w$ under H_1 is:

$$f_1(\mathbf{y}_w) = \frac{1}{A(n)(2\pi)^{\frac{nT(n)}{2}}T(n)} \sum_{t=1}^{T(n)} e^{-\frac{Y_t}{2(\sigma_w^2+P_f)} - \frac{B(t)}{2\sigma_w^2}}, \tag{5.2}$$

where $A(n) = \sigma_w^{(T(n)-1)n}(\sigma_w^2 + P_f)^{\frac{n}{2}}$ and $B(t) = \sum_{\substack{r=1 \\ r\neq t}}^{T(n)} Y_r$.

Since the test is between two simple hypotheses on Alice's transmission state, the likelihood ratio test (LRT) is optimal under the Neyman–Pearson criterion [59, Ch. 3.2]. Taking the ratio between (5.1) and (5.2), and re-arranging terms, we obtain:

$$\Lambda(\mathbf{y}_w) = \frac{f_1(\mathbf{y}_w)}{f_0(\mathbf{y}_w)} = \left(\frac{\sigma_w^2}{\sigma_w^2 + P_f}\right)^{\frac{n}{2}} \frac{1}{T(n)} \sum_{t=1}^{T(n)} e^{\frac{P_f Y_t}{2\sigma_w^2(\sigma_w^2+P_f)}}. \tag{5.3}$$

The likelihood ratio $\Lambda(\mathbf{y}_w)$ is compared to a threshold $\tau(n)$, which is a function of the information known to Willie, and H_0 or H_1 is chosen based on whether $\Lambda(\mathbf{y}_w)$ is

smaller or larger than $\tau(n)$ (if it is equal, a random decision is made):

$$\Lambda(\mathbf{y}_w) \underset{H_1}{\overset{H_0}{\lessgtr}} \tau(n) \tag{5.4}$$

When Alice does not transmit in the i^{th} symbol period, $y_i \sim \mathcal{N}(0, \sigma_w^2)$ since Willie observes AWGN; when Alice transmits, $y_i \sim \mathcal{N}(0, \sigma_w^2 + P_f)$ by construction. Let $\{X_t\}$, $X_t \sim \chi_n^2$, $t = 1, \ldots, T(n)$ be a sequence of i.i.d. chi-squared random variables with n degrees of freedom. Then $Y_t = \sigma_w^2 X_t$ for all $t \in \{1, \ldots, T(n)\}$ under H_0 and $t \in \{1, \ldots, T(n)\} \setminus \{t_A\}$ under H_1. However, under H_1, $Y_{t_A} = (\sigma_w^2 + P_f) X_{t_A}$.

Consider a random variable $L^{(n)}$ defined as follows:

$$L^{(n)} = \frac{M(n)T(n)\Lambda(\mathbf{y}_w) - (T(n)-1)M(n)}{\sqrt{V(n)}} \tag{5.5}$$

where $M(n) = \left(\frac{\sigma_w^2 + P_f}{\sigma_w^2}\right)^{\frac{n}{2}}$ and

$$V(n) = (T(n)-1)\left[\left(\frac{\sigma_w^2 + P_f}{\sigma_w^2 - P_f}\right)^{\frac{n}{2}} - \left(\frac{\sigma_w^2 + P_f}{\sigma_w^2}\right)^{n}\right]. \tag{5.6}$$

This is just a deterministically re-normalized LRT statistic. Since n, $T(n)$, σ_w^2, and P_f are known to Willie, and $M(n)$ and $V(n)$ are deterministic functions, the hypothesis test:

$$L^{(n)} = \frac{\sum_{t=1}^{T(n)} e^{\frac{P_f Y_t}{2\sigma_w^2(\sigma_w^2 + P_f)}} - (T(n)-1)M(n)}{\sqrt{V(n)}} \underset{H_1}{\overset{H_0}{\lessgtr}} S(n) \tag{5.7}$$

is equivalent to that in (5.4), with the threshold

$$S(n) = \frac{M(n)T(n)\tau(n) - (T(n)-1)M(n)}{\sqrt{V(n)}}. \tag{5.8}$$

The performance of both tests is equal. The probability of error is thus

$$\mathbb{P}_e^{(w)} = \frac{\mathbb{P}(L^{(n)} > S(n)|H_0 \text{ true}) + \mathbb{P}(L^{(n)} \leq S(n)|H_1 \text{ true})}{2} \tag{5.9}$$

When H_0 is true, we can write (5.7) as the normalized sum of $T(n)-1$ i.i.d. random variables $\{U_t\}_{t=1}^{T(n)-1}$ and an independent random variable $\frac{U_{T(n)}}{\sqrt{V(n)}}$ as follows:

$$L^{(n)} = \frac{1}{\sqrt{V(n)}} \sum_{t=1}^{T(n)-1} (U_t - M(n)) + \frac{U_{T(n)}}{\sqrt{V(n)}}, \tag{5.10}$$

where $U_{T(n)}$ is identical to U_t that is defined as

$$U_t = \exp\left[\frac{P_f X_t}{2(\sigma_w^2 + P_f)}\right]. \tag{5.11}$$

When H_1 is true, we can write (5.7) as the normalized sum of $T(n)-1$ i.i.d. random variables $\{U_t\}_{t=1,t\neq t_A}^{T(n)}$ and an independent random variable $\frac{U_{t_A}}{\sqrt{V(n)}}$ as follows:

$$L^{(n)} = \frac{1}{\sqrt{V(n)}} \sum_{t=1,t\neq t_A}^{T(n)} (U_t - M(n)) + \frac{U_{t_A}}{\sqrt{V(n)}}, \tag{5.12}$$

where U_t in the sum is defined as in (5.11), and

$$U_{t_A} = \exp\left[\frac{P_f X_{t_A}}{2\sigma_w^2}\right]. \tag{5.13}$$

We first show that the normalized sums in (5.10) and (5.12) contain i.i.d. zero-mean unit-variance random variables, thus both converging in distribution to the standard Gaussian distribution $\mathcal{N}(0,1)$ by the central limit theorem (CLT). We then show that, outside the sums, $\frac{U_{T(n)}}{\sqrt{V(n)}} \xrightarrow{\mathcal{P}} 0$ and $\frac{U_{t_A}}{\sqrt{V(n)}} \xrightarrow{\mathcal{P}} 0$, where $K_n \xrightarrow{\mathcal{P}} Q$ denotes

convergence of random variable K_n to random variable Q in probability. This allows us to lower bound Willie's probability of error for all values of threshold $S(n)$.

First let's calculate the moments of U_t defined in (5.11). The expectation of U_t is the moment generating function (MGF) $\mathcal{M}_{\chi_n^2}(x) = (1-2x)^{-n/2}$ of a chi-squared random variable evaluated at $x = \frac{P_f}{2(\sigma_w^2+P_f)}$:

$$\mathbb{E}[U_t] = \mathbb{E}\left[\exp\left(\frac{P_f X_t}{2(\sigma_w^2+P_f)}\right)\right] = \left(\frac{\sigma_w^2+P_f}{\sigma_w^2}\right)^{\frac{n}{2}} \tag{5.14}$$

Thus, $M(n) = \mathbb{E}[U_t]$, and the terms inside the sum in (5.10) and (5.12) have zero mean. The second moment of U_t is:

$$\mathbb{E}[U_t^2] = \mathbb{E}\left[\exp\left(\frac{P_f X_t}{\sigma_w^2+P_f}\right)\right] = \left(\frac{\sigma_w^2+P_f}{\sigma_w^2-P_f}\right)^{\frac{n}{2}} \tag{5.15}$$

Thus $V(n) = (T(n)-1)\operatorname{Var}[U_t]$, and, by the Lindenberg CLT for a triangular array [9, Th. 27.2], the normalized sums in both (5.10) and (5.12) converge in distribution to $\mathcal{N}(0,1)$.

The probability that the magnitude of $\frac{U_{T(n)}}{\sqrt{V(n)}}$ in (5.10) exceeds $\delta > 0$ is upper-bounded using the Chebyshev's inequality:

$$\mathbb{P}\left(\left|\frac{U_{T(n)}}{\sqrt{V(n)}}\right| > \delta\right) \leq \left(\delta\sqrt{T(n)-1} - R(n)\right)^{-2} \tag{5.16}$$

where $R(n) = \frac{\mathbb{E}[U_t]}{\sqrt{\operatorname{Var}[U_t]}} = \left[\left(\frac{\sigma_w^4}{\sigma_w^4-P_f^2}\right)^{\frac{n}{2}} - 1\right]^{-\frac{1}{2}}$. Since $T(n)$ is increasing and $P_f < \frac{\sigma_w^2}{2}$, $\frac{U_{T(n)}}{\sqrt{T(n)}} \xrightarrow{\mathcal{P}} 0$ as $n \to \infty$.

To show that $\frac{U_{t_A}}{\sqrt{V(n)}}$ in (5.12) also converges in probability to zero, we need the first two moments of U_{t_A} defined in (5.13). We use the MGF $\mathcal{M}_{\chi_n^2}(x) = (1-2x)^{-n/2}$ evaluated at $x = \frac{P_f}{2\sigma_w^2}$ to compute the expectation:

$$\mathbb{E}[U_{t_A}] = \mathbb{E}\left[\exp\left(\frac{P_f X_{t_A}}{2\sigma_w^2}\right)\right] = \left(\frac{\sigma_w^2}{\sigma_w^2-P_f}\right)^{\frac{n}{2}} \tag{5.17}$$

The second moment of U_{t_A} is:

$$\mathbb{E}[U_{t_A}^2] = \mathbb{E}\left[\exp\left(\frac{P_f X_{t_A}}{\sigma_w^2}\right)\right] = \left(\frac{\sigma_w^2}{\sigma_w^2 - 2P_f}\right)^{\frac{n}{2}} \tag{5.18}$$

The probability that the magnitude of the term $\frac{U_{t_A}}{\sqrt{V(n)}}$ in (5.12) exceeds $\delta > 0$ is upper-bounded using Chebyshev's inequality:

$$\mathbb{P}\left(\left|\frac{U_{t_A}}{\sqrt{V(n)}}\right| > \delta\right) \leq \frac{\mathrm{Var}[U_{t_A}]}{\left(\delta\sqrt{V(n)} - \mathbb{E}[U_{t_A}]\right)^2} \tag{5.19}$$

Dividing the numerator and denominator in the RHS of (5.19) by $\mathrm{Var}[U_{t_A}]$, we note that $\frac{\mathbb{E}[U_{t_A}]}{\sqrt{\mathrm{Var}[U_{t_A}]}} = \left(\left(1 + \frac{P_f^2}{\sigma_w^4 - 2P_f\sigma_w^2}\right)^{\frac{n}{2}} - 1\right)^{-\frac{1}{2}} < C$, with C a constant for $P_f < \frac{\sigma_w^2}{2}$. Also, $\frac{V(n)}{\mathrm{Var}[U_{t_A}]} \geq \frac{V(n)}{\mathbb{E}[U_{t_A}^2]}$ and

$$\frac{V(n)}{\mathbb{E}[U_{t_A}^2]} \geq (T(n) - 1)\left[\left(1 - \frac{2P_f^2}{\sigma_w^4}\right)^{\frac{n}{2}} - \left(1 - \frac{3P_f^2}{\sigma_w^4}\right)^{\frac{n}{2}}\right]. \tag{5.20}$$

The dominant term inside the square brackets in (5.20) is $\left(1 - \frac{2P_f^2}{\sigma_w^4}\right)^{\frac{n}{2}} = e^{\frac{n}{2}\log\left(1 - \frac{2P_f^2}{\sigma_w^4}\right)}$. When $T(n) = o(e^n)$, we demonstrate that $\frac{U_{t_A}}{\sqrt{V(n)}} \xrightarrow{\mathcal{P}} 0$ by setting the symbol power to $P_f = \frac{c_P\sigma_w^2\sqrt{\log T(n)}}{\sqrt{n}}$ for a constant $c_P \in (0, 1)$ and using the Taylor series expansion of $\log(1 - x)$ at $x = 0$. When $T(n) = \Omega(e^n)$, convergence is obtained with $P_f = \frac{c_P\sigma_w^2}{2}$. Thus, effectively, the symbol power that guarantees convergence is $P_f = c_P\sigma_w^2 \min\left\{\frac{\sqrt{\log T(n)}}{\sqrt{n}}, \frac{1}{2}\right\}$.

When $\left|\frac{U_{T(n)}}{\sqrt{V(n)}}\right| < \delta$, the false alarm probability is lower-bounded as follows:

$$\mathbb{P}(L^{(n)} > S(n)|H_0 \text{ is true}) \geq \mathbb{P}\left(E_g(S(n), \delta)\right), \tag{5.21}$$

where $E_g(S(n),\delta)$ denotes the event that $\frac{1}{\sqrt{V(n)}}\sum_{t=1}^{T(n)-1}(U_t - M(n)) \geq S(n)+\delta$. Similarly, when $\left|\frac{U_{t_A}}{\sqrt{V(n)}}\right| < \delta$, the probability of missed detection is lower-bounded as follows:

$$\mathbb{P}(L^{(n)} \leq S(n) | H_1 \text{ is true}) \geq \mathbb{P}\left(E_l(S(n),\delta)\right), \tag{5.22}$$

where $E_l(S(n),\delta)$ denotes the event that $\frac{1}{\sqrt{V(n)}}\sum_{t=1,t\neq t_A}^{T(n)}(U_t - M(n)) \leq S(n)-\delta$. Denote by $E_C(S(n),\delta)$ the event when either event $E_g(S(n),\delta)$ occurs when Alice is quiet or event $E_l(S(n),\delta)$ occurs when Alice transmits. Since we assume equiprobable priors,

$$\mathbb{P}(E_C(S(n),\delta)) = \frac{\mathbb{P}\left(E_g(S(n),\delta)\right) + \mathbb{P}\left(E_l(S(n),\delta)\right)}{2}. \tag{5.23}$$

By the CLT for triangular arrays in [9, Th. 27.2], the normalized sums in the events $E_g(S(n),\delta)$ and $E_l(S(n),\delta)$ converge in distribution to standard Gaussian random variables. This result only provides pointwise convergence in the argument of the distribution function, but $S(n)$ is the n^{th} value in an arbitrary sequence. Instead, in Appendix A.5, we exploit the uniform convergence on any finite number of points and the monotonicity of the distribution function to show that, for each normalized sum, setting $\delta = \epsilon\sqrt{2\pi}/9$ yields n_0 such that for all $n \geq n_0$ and any $S(n)$,

$$\mathbb{P}\left(E_C\left(S(n), \frac{\epsilon\sqrt{2\pi}}{9}\right)\right) \geq \frac{1}{2} - \frac{\epsilon}{3}. \tag{5.24}$$

By (5.16) and (5.19), there exists n_1 such that for all $n \geq n_1$, $\mathbb{P}\left(\left|\frac{U_{T(n)}}{\sqrt{V(n)}}\right| > \frac{\epsilon\sqrt{2\pi}}{9}\right) < \frac{\epsilon}{3}$ and $\mathbb{P}\left(\left|\frac{U_{t_A}}{\sqrt{V(n)}}\right| > \frac{\epsilon\sqrt{2\pi}}{9}\right) < \frac{\epsilon}{3}$. The intersection of these events and the event $E_C(S(n),\epsilon\sqrt{2\pi}/9)$ yields an error event. By combining their probabilities using DeMorgan's Law and the union bound, we can lower-bound $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ for all $n \geq \max\{n_0, n_1\}$, concluding the analysis of Willie's detector.

Analysis (Bob): Let Bob employ the maximum likelihood (ML) decoder (i.e., minimum distance decoder). If Bob knows the value of t_A, his probability of decoding error is given directly by (4.9). Since $P_f = c_P \sigma_w^2 \min\{\frac{\sqrt{\log T(n)}}{\sqrt{n}}, \frac{1}{2}\}$, Alice can covertly transmit $M = \frac{n\gamma}{2} \log_2 \left(1 + \frac{c_P \sigma_w^2}{2\sigma_b^2} \min\left\{\frac{\sqrt{\log T(n)}}{\sqrt{n}}, \frac{1}{2}\right\}\right)$ bits, where $\gamma \in (0,1)$ is a constant, with Bob's probability of decoding error (averaged over all the codebooks) decaying to zero as $n \to \infty$. Therefore, $\mathcal{O}(\min\{\sqrt{n \log T(n)}, n\})$ covert bits can be transmitted reliably using slot t_A.

However, knowledge of t_A is unnecessary for Bob if $T(n) < 2^{c_T n}$, where c_T is a constant. Let's augment Alice and Bob's Gaussian codebook with the origin $\mathbf{c}(0) = \{0, \ldots, 0\}$ (indicating "no transmission") and have Bob attempt to decode each of the $T(n)$ slots. The squared distance between a codeword $\mathbf{c}(W_k)$ and $\mathbf{c}(0)$ is $P_f X$ while the squared distance between any pair of codewords $\{\mathbf{c}(W_k), \mathbf{c}(W_i)\}$ is $2P_f X$, where $X \sim \chi_n^2$. Repeating the analysis that leads to (4.9) using the distance between $\mathbf{c}(W_k)$ and $\mathbf{c}(0)$ instead of $\mathbf{c}(W_i)$ yields a slightly looser upper bound on the probability of the decoding error in each slot. By the union bound over all $T(n)$ slots, the overall probability of error is:

$$\mathbb{P}_e^{(b)} \leq T(n) 2^{M - \frac{n}{2} \log_2 \left(1 + \frac{P_f}{4\sigma_b^2}\right)} \tag{5.25}$$

If $T(n) = o(e^n)$, then clearly Bob's decoding error probability decays to zero if Alice attempts to transmit $M = \frac{n\gamma}{2} \log_2 \left(1 + \frac{c_P \sigma_w^2 \sqrt{\log T(n)}}{4\sigma_b^2 \sqrt{n}}\right)$ bits in a randomly selected n-symbol slot t_A. If $T(n) = \Omega(e^n)$, then, $P_f = \frac{\sigma_w^2}{2}$, and $T(n) < 2^{c_T n}$ where $c_T = \frac{1-\gamma}{2} \log_2 \left(1 + \frac{\sigma_w^2}{8\sigma_b^2}\right)$ ensures that Bob's decoding error probability decays to zero if Alice attempts to transmit $M = \frac{n\gamma}{2} \log_2 \left(1 + \frac{c_P \sigma_w^2}{8\sigma_b^2}\right)$ bits in a randomly selected n-symbol slot t_A. □

5.3 Converse

Suppose Alice attempts to transmit one of 2^M (equally likely) M-bit messages reliably to Bob using a sequence of n consecutive symbol periods inside a sequence of $nT(n)$ symbol periods, where $M = \omega(\sqrt{n \log T(n)})$, while limiting Willie's ability to detect her transmission. She thus encodes each message arbitrarily into n symbols.

If Alice transmits, Willie's $nT(n)$ observations of his channel from Alice contain Alice's sequence of n consecutive channel uses, however, Willie is oblivious to the location of the start of Alice's transmission as well as other properties of her signal. Nevertheless, we prove that, by dividing his sequence of $nT(n)$ observations into a set of $T(n)$ non-overlapping subsequences, and employing a simple threshold detector on the maximum subsequence power, Willie can detect Alice if she attempts to transmit $\omega(\sqrt{n \log T(n)})$ bits reliably.

Theorem 5.3.1. *If Alice attempts to transmit $\omega(\sqrt{n \log T(n)})$ bits using a sequence of n consecutive symbol periods that are arbitrarily located inside a sequence of $nT(n)$ symbol periods, then, as $n \to \infty$, either Willie can detect her with high probability, or Bob cannot decode with arbitrarily low probability of error.*

Proof. Let Willie divide the sequence $\mathbf{y}_w$ of $nT(n)$ observations of his channel from Alice into a set of $T(n)$ non-overlapping subsequences $\{\mathbf{y}_w(t)\}_{t=1}^{T(n)}$, with each $\mathbf{y}_w(t)$ containing n consecutive observations. Denote by $Y_t = \sum_{y_i \in \mathbf{y}_w(t)} y_i^2$ the observed power in each subsequence and $Y_{\max} = \max_{t \in \{1,\ldots,T(n)\}} Y_t$. For a threshold S, Willie accuses Alice of transmitting if $Y_{\max} > S$.

Suppose Alice does not transmit. Willie's probability of false alarm is $\mathbb{P}(Y_{\max} > S)$. Let $S = \sigma_w^2(n + \sqrt{n}\delta)$. To find δ so that Willie's detector has an arbitrary probability of false alarm $\mathbb{P}_{FA}^*$ as $n \to \infty$, note that each $Y_t = \sigma_w^2 X_t$ where $\{X_t\}$, $X_t \sim \chi_n^2$, $t = 1, \ldots, T(n)$ is a sequence of i.i.d. chi-squared random variables each with n degrees of freedom. We have

$$\mathbb{P}(Y_{\max} > S) = 1 - \mathbb{P}\left(X_{\max} \leq S/\sigma_w^2\right) \tag{5.26}$$

$$= 1 - \left(1 - \mathbb{P}\left(X_1 > n + \sqrt{n}\delta\right)\right)^{T(n)} \tag{5.27}$$

where $X_{\max} = \max_{t\in\{1,\ldots,T\}} X_t$. For the desired $\mathbb{P}^*_{FA}$,

$$1 - (1 - \mathbb{P}^*_{FA})^{1/T(n)} = \mathbb{P}\left(X_1 > n + \sqrt{n}\delta\right). \tag{5.28}$$

Using a Chernoff bound for the tail of a chi-squared distribution [24, Lemma 2.2], we obtain:

$$\mathbb{P}\left(X_1 > n + \sqrt{n}\delta\right) \leq \left(1 + \delta/\sqrt{n}\right)^{n/2} e^{-\frac{\sqrt{n}\delta}{2}} \tag{5.29}$$

$$= e^{\frac{n}{2}\log\left(1+\frac{\delta}{\sqrt{n}}\right) - \frac{\sqrt{n}\delta}{2}} \tag{5.30}$$

$$= e^{-\delta^2/4 + \mathcal{O}(1/\sqrt{n})} \tag{5.31}$$

with (5.31) is from the Taylor series expansion of $\log(1+x)$ at $x = 0$. Discarding low order terms and solving (5.31) for δ yields $\delta = 2\sqrt{-\log\left(1 - (1 - \mathbb{P}^*_{FA})^{1/T(n)}\right)}$. Taylor series expansion of $1 - c^x$ at $x = 0$ yields $1 - (1 - \mathbb{P}^*_{FA})^{1/T(n)} = \frac{1}{T(n)} \log\left(\frac{1}{1-\mathbb{P}^*_{FA}}\right) + \mathcal{O}(\frac{1}{T^2(n)})$. Thus, setting $\delta = c\sqrt{\log T(n)}$ with some constant $c > 0$ yields the desired probability of false alarm $\mathbb{P}^*_{FA}$.

Now suppose Alice uses an arbitrary codebook $\{\mathbf{c}(W_k), k = 1, \ldots, 2^{nR}\}$ and transmits codeword $\mathbf{c}(W_k)$ using n consecutive symbol periods. Denote average symbol power of $\mathbf{c}(W_k)$ by $P_f = \frac{\|\mathbf{c}(W_k)\|^2}{n}$. Since Alice uses n consecutive symbols, her transmission overlaps at most two of Willie's subsequences, which we denote t_A and t_B. Denote by P_A and P_B the power from Alice's transmission in subsequences t_A and t_B, respectively, with $P_A + P_B = nP_f$. Willie's probability of missing Alice's transmission is

$$\mathbb{P}_{MD}^{(k)} = \mathbb{P}(Y_{\max} \leq S) = \mathbb{P}(Y_{t_A} \leq S)\mathbb{P}(Y_{t_B} \leq S) \prod_{\substack{t=1 \\ t \notin \{t_A, t_B\}}}^{T(n)} \mathbb{P}(Y_t \leq S) \tag{5.32}$$

where the factorization in (5.32) is because Alice's codeword and the noise in other subsequences are independent. $\prod_{t=1, t\notin\{t_A,t_B\}}^{T(n)} \mathbb{P}(Y_t \leq S) \leq 1$ does not depend on Alice's codeword. However, since the codeword is an unknown deterministic signal that is added to AWGN on Willie's channel to Alice, $\frac{Y_{t_A}}{\sigma_w^2} \sim \chi_n^2(P_A)$ and $\frac{Y_{t_B}}{\sigma_w^2} \sim \chi_n^2(P_B)$ are non-central chi-squared random variables with n degrees of freedom and respective non-centrality parameters $\frac{P_A}{\sigma_w^2}$ and $\frac{P_B}{\sigma_w^2}$. Without loss of generality, assume that $P_A \geq P_B$. Thus, P_A satisfies $\frac{nP_f}{2} \leq P_A \leq nP_f$ and the expected value and variance of Y_{t_A} are bounded as follows [82, App. D.1]:

$$\mathbb{E}\left[Y_{t_A}\right] \geq \sigma_w^2 n + \frac{nP_f}{2} \tag{5.33}$$

$$\mathrm{Var}\left[Y_{t_A}\right] \leq 2n\sigma_w^4 + 4n\sigma_w^2 P_f \tag{5.34}$$

Since $\mathbb{P}(Y_{t_B} \leq S) \leq 1$, Chebyshev's inequality with (5.33) and (5.34) yields

$$\begin{aligned} \mathbb{P}_{MD}^{(k)} &\leq \mathbb{P}\left(|Y_{t_A} - \mathbb{E}[Y_{t_A}]| > \mathbb{E}[Y_{t_A}] - \sigma_w^2(n - c\sqrt{n\log T(n)})\right) \\ &\leq \frac{2\sigma_w^4 + 4\sigma_w^2 P_f}{\left(\frac{\sqrt{n}P_f}{2} - c\sigma_w^2\sqrt{\log T(n)}\right)^2}. \end{aligned} \tag{5.35}$$

If $P_f = \omega\left(\sqrt{\frac{\log T(n)}{n}}\right)$, as $n \to \infty$, Willie's average probability of error can be made arbitrarily low.

The proof of the non-zero lower bound on Bob's probability of decoding error if Alice tries to transmit $\omega(\sqrt{n\log T(n)})$ bits using average symbol power $P_f = \mathcal{O}\left(\sqrt{\frac{\log T(n)}{n}}\right)$ follows from a similar proof in Chapter 4.3. □

5.4 Relationship with Steganography

Steganographic systems discussed in Section 2.1 hide information by altering the properties of fixed-size, finite-alphabet covertext objects (e.g. images), and are subject to a similar square root law as covert communication: $\mathcal{O}(\sqrt{n})$ symbols in covertext of size n may safely be modified to hide an $\mathcal{O}(\sqrt{n}\log n)$-bit message [51]. The similarity between the square root laws in these disciplines is from the mathematics of statistical hypothesis testing, as discussed in Section 4.2.3.2. However, in steganography, the transmission to Bob is noiseless, which allows the extra $\log n$ factor.

Batch steganography uses multiple covertext objects to hide a message and is subject to the steganographic square root law described above [48, 47]. The batch steganography interpretation of covert communication using timing as described in this work is equivalent to using only one of $T(n)$ covertext objects of size n to embed a message. Willie, who knows that one covertext object is used but not which one, has to examine all of them. We are not aware of any work on this particular problem, but it is likely that our result extends to it.

CHAPTER 6

ANALYSIS OF COVERT OPTICAL COMMUNICATION

Optical signaling [11, 74] is particularly attractive for covert communication because of its narrow diffraction-limited beam spread in free space [30, 35] and the ease of detecting fiber taps using time-domain reflectometry [2]. Our information-theoretic analysis of covert communication on the AWGN channel in Chapter 4 also applies to a lossy optical channel with additive Gaussian noise when Alice uses a laser-light transmitter and both Bob and Willie use coherent-detection receivers. However, modern high-sensitivity optical communication components are primarily limited by noise of quantum-mechanical origin. Thus, recent studies on the performance of physical optical communication have focused on this quantum-limited regime [33, 89, 87]. We provide the background material on quantum information theory and quantum optics in Appendix B.

In this chapter we establish the quantum limits of covert communication. We begin by introducing our optical channel model in Section 6.1. In Section 6.2 we demonstrate that covert communication is impossible over a pure-loss channel. However, in Section 6.3 we show that, when the channel has any excess noise (e.g., the unavoidable thermal noise from the blackbody radiation at the operating temperature), Alice can reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits to Bob using n optical modes[1]

[1]This chapter and Chapter 7 address optical communication where we treat one spatio-temporal-polarization mode of the optical-frequency electromagnetic field as the fundamental transmission unit over the channel, which can be likened to "channel use" of the previous chapters. A mode is a spatio-temporal electromagnetic field pattern of a given polarization, which can act as a unit of communication over an optical channel. A mode can be excited in a *coherent state*—the quantum description of ideal laser light—of a given amplitude and phase, as is done in standard classical

even if Willie intercepts all the photons not reaching Bob and employs arbitrary quantum memory and measurements. This is achievable using standard laser-light modulation and homodyne detection (thus the Alice-Bob channel is still an AWGN channel). Thus, noise enables stealth. Indeed, we demonstrate in Section 6.4 that if Willie's detector contributes excess noise (e.g., dark counts in photon-counting detectors), Alice can covertly communicate to Bob, even when the channel itself is pure-loss. We conclude this chapter by showing that the square-root limit cannot be circumvented in Section 6.5.

6.1 Channel model

Consider a single-mode quasi-monochromatic lossy optical channel $\mathcal{E}_{\eta_b}^{\bar{n}_T}$ of transmissivity $\eta_b \in (0, 1]$ and thermal noise mean photon number per mode $\bar{n}_T \geq 0$, as depicted in Figure 6.1. Willie collects the entire $\eta_w = 1 - \eta_b$ fraction of Alice's photons that do not reach Bob but otherwise remains passive, not injecting any light into the channel. Later we argue that being active does not help Willie to detect Alice's transmissions. For a pure loss channel ($\bar{n}_T = 0$), the environment input is in the *vacuum* state $\hat{\rho}_0^E = |0\rangle\langle 0|^E$, corresponding to the minimum noise the channel must inject to preserve the Heisenberg inequality of quantum mechanics.

6.2 Pure loss insufficient for covert communication

Regardless of Alice's strategy, reliable and covert communication over a pure-loss channel to Bob is impossible, as Willie can effectively use an ideal *single photon detector* (SPD) on each mode to discriminate between an n-mode vacuum state and any

---

optical communication. On other hand, quantum optics allows for a mode to be excited in other, non-classical states of light such as a *squeezed state* or a *Fock state*. For example, each temporal mode of a single (spatial) mode optical fiber can carry one of the two coherent-state pulses of the binary phase shift keying (BPSK) modulation format. We provide a more formal description of optical modes in Appendix B.4.

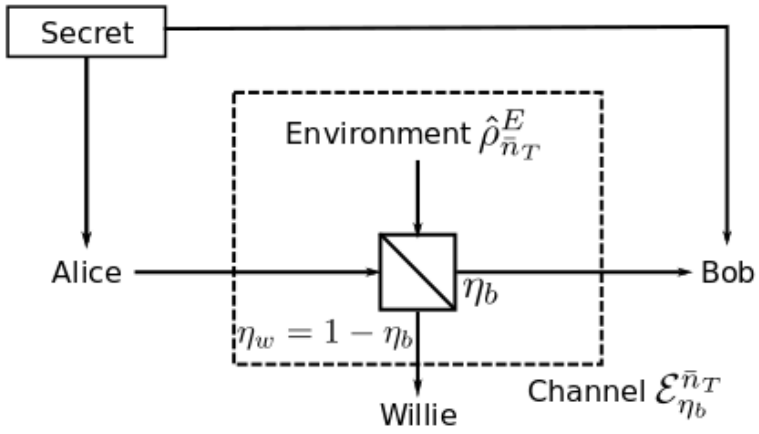


Figure 6.1: Optical channel model. The input-output relationship is captured by a beamsplitter of transmissivity η_b, with the transmitter Alice at one of the input ports and the intended receiver Bob at one of the output ports, and η_b being the fraction of Alice's signaling photons that reach Bob. The other input and output ports of the beamsplitter correspond to the environment and the adversary Willie. Willie collects the entire $\eta_w = 1 - \eta_b$ fraction of Alice's photons that do not reach Bob. This models single-spatial-mode free-space and single-mode fiber optical channels. Alice and Bob share a secret before the transmission.

non-vacuum state in Alice's codebook. Willie avoids false alarms since no photons impinge on his SPD when Alice is silent. However, a single *click*—detection of one or more photons—gives away Alice's transmission attempt regardless of the actual quantum state of Alice's signaling photons. Alice is thus constrained to codewords that are nearly indistinguishable from vacuum, rendering unreliable any communication attempt that is designed to be covert. Furthermore, any communication attempt that is designed to be reliable cannot remain covert, as Willie detects it with high probability for large n. This is true even when Alice and Bob have access to an infinitely-large pre-shared secret. The following theorem formally states this result.

Theorem 6.2.1. (Insufficiency of pure-loss for covert communication) *Suppose Willie has a pure-loss channel from Alice and is limited only by the laws of physics in his receiver measurement choice. Then Alice cannot communicate to Bob reliably and*

covertly even if Alice and Bob have access to a pre-shared secret of unbounded size, an unattenuated observation of the transmission, and a quantum-optimal receiver.

In the proof of this theorem we denote a tensor product of n Fock (or photon number) states by $|\mathbf{u}\rangle \equiv |u_1\rangle \otimes |u_2\rangle \otimes \cdots \otimes |u_n\rangle$, where vector $\mathbf{u} \in \mathbb{N}_0^n$ with $\mathbb{N}_0$ being the set of non-negative integers. Specifically, $|\mathbf{0}\rangle \equiv |0\rangle^{\otimes n}$. Before proceeding with the proof, we state the following lemma.

Lemma 6.1. *Given the input of n-mode vacuum state $|\mathbf{0}\rangle^{E^n}$ on the "environment" port and an n-mode entangled state $|\psi\rangle^{A^n} = \sum_{\mathbf{k}} a_{\mathbf{k}} |\mathbf{k}\rangle^{A^n}$ on the "Alice" port of a beamsplitter with transmissivity $\eta_b = 1 - \eta_w$, the diagonal elements of the output state ρ^{W^n} on the "Willie" port can be expressed in the n-fold Fock state basis as follows:*

$$ {}^{W^n}\langle \mathbf{s}| \hat{\rho}^{W^n} |\mathbf{s}\rangle^{W^n} = \sum_{\mathbf{k} \in \mathbb{N}_0^n} |a_{\mathbf{k}}|^2 \prod_{i=1}^{n} \binom{k_i}{s_i} (1-\eta_w)^{k_i - s_i} \eta_w^{s_i}. \tag{6.1} $$

Proof. See Appendix C.1 □

Proof. (Theorem 6.2.1) Alice sends one of 2^M (equally likely) M-bit messages by choosing an element from an arbitrary codebook $\{\hat{\rho}_x^{A^n}, x = 1, \ldots, 2^M\}$, where a state $\hat{\rho}_x^{A^n} = |\psi_x\rangle^{A^n A^n}\langle\psi_x|$ encodes an M-bit message W_x. $|\psi_x\rangle^{A^n} = \sum_{\mathbf{k} \in \mathbb{N}_0^n} a_{\mathbf{k}}(x) |\mathbf{k}\rangle^{A^n}$ is a general n-mode pure state, where $|\mathbf{k}\rangle \equiv |k_1\rangle \otimes |k_2\rangle \otimes \cdots \otimes |k_n\rangle$ is a tensor product of n Fock states. We limit our analysis to pure input states since, by convexity, using mixed states as inputs can only degrade the performance (since that is equivalent to transmitting a randomly chosen pure state from an ensemble and discarding the knowledge of that choice).

Let Willie use an ideal SPD on all n modes, given by positive operator-valued measure (POVM) $\left\{ |0\rangle\langle 0|, \sum_{j=1}^{\infty} |j\rangle\langle j| \right\}^{\otimes n}$. When W_u is transmitted, Willie's hypothesis test reduces to discriminating between the states

$$\hat{\rho}_0^{W^n} = |\mathbf{0}\rangle^{W^n W^n}\langle\mathbf{0}| \text{ and} \tag{6.2}$$

$$\hat{\rho}_1^{W^n} = \hat{\rho}_u^{W^n}, \tag{6.3}$$

where $\hat{\rho}_u^{W^n}$ is the output state of a pure-loss channel with transmissivity η_w corresponding to an input state $\hat{\rho}_u^{A^n}$. Thus, Willie's average error probability is:

$$\mathbb{P}_e^{(w)} = \frac{1}{2^{M+1}} \sum_{u=1}^{2^M} {}^{W^n}\langle\mathbf{0}|\, \hat{\rho}_u^{W^n}\, |\mathbf{0}\rangle^{W^n}, \tag{6.4}$$

since messages are sent equiprobably. Note that the error is entirely because of missed codeword detections, as Willie's receiver never raises a false alarm. By Lemma 6.1,

$$\begin{aligned} {}^{W^n}\langle\mathbf{0}|\, \hat{\rho}_u^{W}\, |\mathbf{0}\rangle^{W^n} &= \sum_{\mathbf{k}\in\mathbb{N}_0^n} |a_{\mathbf{k}}(u)|^2 \,(1-\eta_w)^{\sum_{i=1}^n k_i} \\ &\leq |a_{\mathbf{0}}(u)|^2 + (1-|a_{\mathbf{0}}(u)|^2)(1-\eta_w) \\ &= 1 - \eta_w \left(1 - |a_{\mathbf{0}}(u)|^2\right). \end{aligned} \tag{6.5}$$

Substituting equation (6.5) into equation (6.4) yields

$$\mathbb{P}_e^{(w)} \leq \frac{1}{2} - \frac{\eta_w}{2}\left(1 - \frac{1}{2^M}\sum_{u=1}^{2^M} |a_{\mathbf{0}}(u)|^2\right).$$

Thus, to ensure $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$, Alice must use a codebook with the probability of transmitting zero photons:

$$\frac{1}{2^M}\sum_{u=1}^{2^M} |a_{\mathbf{0}}(u)|^2 \geq 1 - \frac{2\epsilon}{\eta_w}. \tag{6.6}$$

Equation (6.6) can be restated as an upper bound on the probability of transmitting one or more photons:

$$\frac{1}{2^M}\sum_{u=1}^{2^M}\left(1-|a_{\mathbf{0}}(u)|^2\right)\leq\frac{2\epsilon}{\eta_w}. \tag{6.7}$$

Now we show that there exists an interval $(0,\epsilon_0]$, $\epsilon_0>0$ such that if $\epsilon\in(0,\epsilon_0]$, Bob's average decoding error probability $\mathbb{P}_e^{(b)}\geq\delta_0$ where $\delta_0>0$, thus making covert communication over a pure-loss channel unreliable.

Denote by $E_{u\to v}$ the event that the transmitted message W_u is decoded by Bob as $W_v\neq W_u$. Given that W_u is transmitted, the decoding error probability is the probability of the union of events $\cup_{v=0,v\neq u}^{2^M}E_{u\to v}$. Let Bob choose a POVM $\{\hat{\Lambda}_j^*\}$ that minimizes the average probability of error over n optical channel modes:

$$\mathbb{P}_e^{(b)}=\inf_{\{\hat{\Lambda}_j\}}\frac{1}{2^M}\sum_{u=1}^{2^M}\mathbb{P}\left(\cup_{v=0,v\neq u}^{2^M}E_{u\to v}\right). \tag{6.8}$$

Now consider a codebook that meets the necessary condition for covert communication given in equation (6.7). Define the subset of this codebook $\left\{\hat{\rho}_u^{A^n},u\in\mathcal{A}\right\}$ where $\mathcal{A}=\left\{u:1-|a_{\mathbf{0}}(u)|^2\leq\frac{4\epsilon}{\eta_w}\right\}$. We lower-bound (6.8) as follows:

$$\mathbb{P}_e^{(b)}=\frac{1}{2^M}\sum_{u\in\bar{\mathcal{A}}}\mathbb{P}\left(\cup_{v=0,v\neq u}^{2^M}E_{u\to v}\right)+\frac{1}{2^M}\sum_{u\in\mathcal{A}}\mathbb{P}\left(\cup_{v=0,v\neq u}^{2^M}E_{u\to v}\right) \tag{6.9}$$

$$\geq\frac{1}{2^M}\sum_{u\in\mathcal{A}}\mathbb{P}\left(\cup_{v=0,v\neq u}^{2^M}E_{u\to v}\right), \tag{6.10}$$

where the probabilities in equation (6.9) are with respect to the POVM $\{\hat{\Lambda}_j^*\}$ that minimizes equation (6.8) over the entire codebook. Without loss of generality, let's assume that $|\mathcal{A}|$ is even, and split $\mathcal{A}$ into two equal-sized non-overlapping subsets $\mathcal{A}^{(\text{left})}$ and $\mathcal{A}^{(\text{right})}$ (formally, $\mathcal{A}^{(\text{left})}\cup\mathcal{A}^{(\text{right})}=\mathcal{A}$, $\mathcal{A}^{(\text{left})}\cap\mathcal{A}^{(\text{right})}=\emptyset$, and $|\mathcal{A}^{(\text{left})}|=|\mathcal{A}^{(\text{right})}|$). Let $g:\mathcal{A}^{(\text{left})}\to\mathcal{A}^{(\text{right})}$ be a bijection. We can thus re-write (6.10):

$$\mathbb{P}_e^{(b)} \geq \frac{1}{2^M} \sum_{u \in \mathcal{A}^{(\text{left})}} 2 \left(\frac{\mathbb{P}\left(\cup_{v=0, v \neq u}^{2^M} E_{u \to v} \right)}{2} + \frac{\mathbb{P}\left(\cup_{v=0, v \neq g(u)}^{2^M} E_{g(u) \to v} \right)}{2} \right)$$
$$\geq \frac{1}{2^M} \sum_{u \in \mathcal{A}^{(\text{left})}} 2 \left(\frac{\mathbb{P}\left(E_{u \to g(u)} \right)}{2} + \frac{\mathbb{P}\left(E_{g(u) \to u} \right)}{2} \right), \tag{6.11}$$

where the second lower bound is because the events $E_{u \to g(u)}$ and $E_{g(u) \to u}$ are contained in the unions $\cup_{v=0, v \neq u}^{2^M} E_{u \to v}$ and $\cup_{v=0, v \neq g(u)}^{2^M} E_{g(u) \to v}$, respectively. The summation term in equation (6.11),

$$\mathbb{P}_e(u) \equiv \frac{\mathbb{P}\left(E_{u \to g(u)} \right)}{2} + \frac{\mathbb{P}\left(E_{g(u) \to u} \right)}{2}, \tag{6.12}$$

is Bob's average probability of error when Alice only sends messages W_u and $W_{g(u)}$ equiprobably. We thus reduce the analytically intractable problem of discriminating between many states in equation (6.8) to a quantum binary hypothesis test.

The lower bound on the probability of error in discriminating two received codewords is obtained by lower-bounding the probability of error in discriminating two codewords *before* they are sent (this is equivalent to Bob having an unattenuated unity-transmissivity channel from Alice). Recalling that $\hat{\rho}_u^{A^n} = |\psi_u\rangle^{A^n A^n}\langle\psi_u|$ and $\hat{\rho}_{g(u)}^{A^n} = \left|\psi_{g(u)}\right\rangle^{A^n A^n}\left\langle\psi_{g(u)}\right|$ are pure states, the lower bound on the probability of error in discriminating between $\left|\psi_u^{A^n}\right\rangle$ and $\left|\psi_{g(u)}^{A^n}\right\rangle$ is [41, Chapter IV.2 (c), Equation (2.34)]:

$$\mathbb{P}_e(u) \geq \left[1 - \sqrt{1 - F\left(|\psi_u\rangle^{A^n}, \left|\psi_{g(u)}\right\rangle^{A^n} \right)} \right] \Big/ 2\,, \tag{6.13}$$

where $F(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$ is the fidelity between the pure states $|\psi\rangle$ and $|\phi\rangle$. Lower-bounding $F\left(|\psi_u\rangle^{A^n}, \left|\psi_{g(u)}\right\rangle^{A^n} \right)$ lower-bounds the RHS of equation (6.13). For

pure states $|\psi\rangle$ and $|\phi\rangle$, $F(|\psi\rangle, |\phi\rangle) = 1 - \left(\frac{1}{2}\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \|_1\right)^2$, where $\|\rho - \sigma\|_1$ is the trace distance [88, Equation (9.134)]. Thus,

$$\begin{aligned} F\left(|\psi_u\rangle^{A^n}, \left|\psi_{g(u)}\right\rangle^{A^n}\right) &= 1 - \left(\frac{1}{2}\|\hat{\rho}_u^{A^n} - \hat{\rho}_{g(u)}^{A^n}\|_1\right)^2 \\ &\geq 1 - \left(\frac{\|\hat{\rho}_u^{A^n} - |\mathbf{0}\rangle^{A^n A^n}\langle\mathbf{0}|\,\|_1}{2} + \frac{\|\hat{\rho}_{g(u)}^{A^n} - |\mathbf{0}\rangle^{A^n A^n}\langle\mathbf{0}|\,\|_1}{2}\right)^2 \\ &= 1 - \left(\sqrt{1 - \left|{}^{A^n}\langle\mathbf{0}|\psi_u\rangle^{A^n}\right|^2} + \sqrt{1 - \left|{}^{A^n}\left\langle\mathbf{0}|\psi_{g(u)}\right\rangle^{A^n}\right|^2}\right)^2, \end{aligned} \tag{6.14}$$

where the inequality is from the triangle inequality for trace distance. Substituting (6.14) into (6.13) yields

$$\mathbb{P}_e(u) \geq \left[1 - \sqrt{1 - \left|{}^{A^n}\langle\mathbf{0}|\psi_u\rangle^{A^n}\right|^2} - \sqrt{1 - \left|{}^{A^n}\left\langle\mathbf{0}|\psi_{g(u)}\right\rangle^{A^n}\right|^2}\right] \Big/ 2\,. \tag{6.15}$$

Since $\left|{}^{A^n}\langle\mathbf{0}|\psi_u\rangle^{A^n}\right|^2 = |a_{\mathbf{0}}(u)|^2$ and, by the construction of $\mathcal{A}$, $1 - |a_{\mathbf{0}}(u)|^2 \leq \frac{4\epsilon}{\eta_w}$ and $1 - |a_{\mathbf{0}}(g(u))|^2 \leq \frac{4\epsilon}{\eta_w}$, we have

$$\mathbb{P}_e(u) \geq \frac{1}{2} - 2\sqrt{\frac{\epsilon}{\eta_w}}. \tag{6.16}$$

Recalling the definition of $\mathbb{P}_e(u)$ in equation (6.12), we substitute (6.16) into (6.11) to obtain

$$\mathbb{P}_e^{(b)} \geq \frac{|\mathcal{A}|}{2^M}\left(\frac{1}{2} - 2\sqrt{\frac{\epsilon}{\eta_w}}\right), \tag{6.17}$$

Now, re-stating the condition for covert communication (6.7) yields

$$\begin{aligned} \frac{2\epsilon}{\eta_w} &\geq \frac{1}{2^M}\sum_{u\in\overline{\mathcal{A}}}\left(1 - |a_{\mathbf{0}}(u)|^2\right) \\ &\geq \frac{\left(2^M - |\mathcal{A}|\right)}{2^M}\frac{4\epsilon}{\eta_w} \end{aligned} \tag{6.18}$$

with equality (6.18) because $1-|a_{\mathbf{0}}(u)|^2 > \frac{4\epsilon}{\eta_w}$ for all codewords in $\overline{\mathcal{A}}$ by the construction of $\mathcal{A}$. Solving inequality in (6.18) for $\frac{|\mathcal{A}|}{2^M}$ yields the lower bound on the fraction of the codewords in $\mathcal{A}$,

$$\frac{|\mathcal{A}|}{2^M} \geq \frac{1}{2}. \tag{6.19}$$

Combining equations (6.17) and (6.19) results in a positive lower bound on Bob's probability of decoding error $\mathbb{P}_e^{(b)} \geq \frac{1}{4} - \sqrt{\frac{\epsilon}{\eta_w}}$ for $\epsilon \in \left(0, \frac{\eta_w}{16}\right]$ and any n, and demonstrates that reliable covert communication over a pure-loss channel is impossible. □

Thus, if Willie controlled the environment (as assumed in QKD proofs), by setting it to vacuum, he could deny covert communication between Alice and Bob. However, a positive amount of non-adversarial excess noise—whether from the thermal background or the detector itself—is unavoidable, which enables covert communication.

6.3 Channel noise yields the square root law

Now consider a lossy bosonic channel $\mathcal{E}_{\eta_b}^{\bar{n}_T}$, where the environment mode is in a thermal state with mean photon number $\bar{n}_T > 0$. A thermal state $\hat{\rho}_{\bar{n}_T}^E$ is represented by a mixture of coherent states $|\alpha\rangle$—quantum descriptors of ideal laser-light—weighted by a Gaussian distribution over the field amplitude $\alpha \in \mathbb{C}$:

$$\hat{\rho}_{\bar{n}_T}^E = \frac{1}{\pi \bar{n}_T} \int_{\mathbb{C}} e^{-|\alpha|^2/\bar{n}_T} |\alpha\rangle\langle\alpha|^E \mathrm{d}^2\alpha.$$

This thermal noise masks Alice's transmission attempt, enabling covert communication even when Willie has arbitrary resources, such as any quantum-limited measurement on the *isometric extension* of the Alice-to-Bob quantum channel (i.e., access to all signaling photons not captured by Bob). The following theorem demonstrates that in this scenario Alice can use mean photon number per mode $\bar{n} = \mathcal{O}(1/\sqrt{n})$ to

reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using n optical modes to Bob, who needs only a conventional homodyne-detection receiver:

Theorem 6.3.1. (Square root law for the thermal noise channel) *Suppose Willie has access to an arbitrarily complex receiver measurement as permitted by the laws of quantum physics and can capture all the photons transmitted by Alice that do not reach Bob. Let Willie's channel from Alice be subject to noise from a thermal environment that injects $\bar{n}_T > 0$ photons per optical mode on average, and let Alice and Bob share a secret of sufficient length before communicating. Then Alice can lower-bound Willie's detection error probability $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ for any $\epsilon > 0$ while reliably transmitting $\mathcal{O}(\sqrt{n})$ bits to Bob in n optical modes even if Bob only has access to a (sub-optimal) coherent detection receiver, such as an optical homodyne detector.*

First, we define quantum relative entropy.

Definition 6.1. Quantum relative entropy *between states $\hat{\rho}_0$ and $\hat{\rho}_1$ is* $D(\hat{\rho}_0 \| \hat{\rho}_1) \equiv \mathrm{Tr}\{\hat{\rho}_0(\ln \hat{\rho}_0 - \ln \hat{\rho}_1)\}$.

The following lemma provides the expression for the quantum relative entropy between two thermal states.

Lemma 6.2. *If* $\hat{\rho}_0 = \sum_{n=0}^{\infty} \frac{\bar{n}_0^n}{(1+\bar{n}_0)^{1+n}} |n\rangle\langle n|$ *and* $\hat{\rho}_1 = \sum_{n=0}^{\infty} \frac{\bar{n}_1^n}{(1+\bar{n}_1)^{1+n}} |n\rangle\langle n|$, *then* $D(\hat{\rho}_0 \| \hat{\rho}_1) = \bar{n}_0 \ln \frac{\bar{n}_0(1+\bar{n}_1)}{\bar{n}_1(1+\bar{n}_0)} + \ln \frac{1+\bar{n}_1}{1+\bar{n}_0}$

Proof. See Appendix C.2. □

Proof. (Theorem 6.3.1) Construction: Let Alice use a zero-mean isotropic Gaussian-distributed coherent state input $\{p(\alpha), |\alpha\rangle\}$, where $\alpha \in \mathbb{C}$, $p(\alpha) = e^{-|\alpha|^2/\bar{n}}/\pi\bar{n}$ with mean photon number per symbol $\bar{n} = \int_{\mathbb{C}} |\alpha|^2 p(\alpha) \mathrm{d}^2\alpha$. Alice encodes M-bit blocks of input into codewords of length n symbols by generating 2^M codewords $\{\bigotimes_{i=1}^n |\alpha_i\rangle_k\}_{k=1}^{2^M}$, each according to $p(\bigotimes_{i=1}^n |\alpha_i\rangle) = \prod_{i=1}^n p(\alpha_i)$, where $\bigotimes_{i=1}^n |\alpha_i\rangle = |\alpha_1 \ldots \alpha_n\rangle$ is an n-mode tensor-product coherent state. The codebook is used only once to send a single message and is kept secret from Willie, though he knows how it is constructed.

Analysis (Willie): Since Willie does not have access to Alice's codebook, Willie has to discriminate between the following n-copy quantum states:

$$\hat{\rho}_0^{\otimes n} = \left(\sum_{i=0}^{\infty} \frac{(\eta_b \bar{n}_T)^i}{(1+\eta_b \bar{n}_T)^{1+i}} |i\rangle \langle i| \right)^{\otimes n},$$

and

$$\hat{\rho}_1^{\otimes n} = \left(\sum_{i=0}^{\infty} \frac{(\eta_w \bar{n} + \eta_b \bar{n}_T)^i}{(1+\eta_w \bar{n} + \eta^{(n)} \bar{n}_T)^{1+i}} |i\rangle \langle i| \right)^{\otimes n}.$$

Assuming equal prior probabilities, by Lemma B.2, Willie's average probability of error in discriminating between $\hat{\rho}_0^{\otimes n}$ and $\hat{\rho}_1^{\otimes n}$ is:

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} \left[1 - \frac{1}{2} \| \hat{\rho}_1^{\otimes n} - \hat{\rho}_0^{\otimes n} \|_1 \right],$$

where the minimum in this case is attained by a PNR detection. The trace distance $\|\hat{\rho}_0 - \hat{\rho}_1\|_1$ between states $\hat{\rho}_1$ and $\hat{\rho}_1$ (see Appendix B.2) is upper-bounded the quantum relative entropy (QRE) using quantum Pinsker's Inequality [88, Theorem 11.9.2] as follows:

$$\|\hat{\rho}_0 - \hat{\rho}_1\|_1 \leq \sqrt{2 D(\hat{\rho}_0 \| \hat{\rho}_1)},$$

Thus,

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \sqrt{\frac{1}{8} D(\hat{\rho}_0^{\otimes n} \| \hat{\rho}_1^{\otimes n})}. \tag{6.20}$$

QRE is additive for tensor product states:

$$D(\hat{\rho}_0^{\otimes n} \| \hat{\rho}_1^{\otimes n}) = n D(\hat{\rho}_0 \| \hat{\rho}_1). \tag{6.21}$$

By Lemma 6.2,

$$D(\hat{\rho}_0\|\hat{\rho}_1) = \eta_b\bar{n}_T \ln\frac{(1+\eta_w\bar{n}+\eta_b\bar{n}_T)\eta_b\bar{n}_T}{(\eta_w\bar{n}+\eta_b\bar{n}_T)(1+\eta_b\bar{n}_T)} + \ln\frac{1+\eta_w\bar{n}+\eta_b\bar{n}_T}{1+\eta_b\bar{n}_T}. \tag{6.22}$$

The first two terms of the Taylor series expansion of the RHS of (6.22) with respect to $\bar{n}$ at $\bar{n}=0$ are zero and the fourth term is negative. Thus, using Taylor's Theorem with the remainder, we can upper-bound equation (6.22) by the third term as follows:

$$D(\hat{\rho}_0\|\hat{\rho}_1) \leq \frac{\eta_w^2\bar{n}^2}{2\eta_b\bar{n}_T(1+\eta_b\bar{n}_T)}. \tag{6.23}$$

Combining equations (6.20), (6.21), and (6.23) yields

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \frac{\eta_w\bar{n}\sqrt{n}}{4\sqrt{\eta_b\bar{n}_T(1+\eta_b\bar{n}_T)}} \tag{6.24}$$

Therefore, setting

$$\bar{n} = \frac{4\epsilon\sqrt{\eta_b\bar{n}_T(1+\eta_b\bar{n}_T)}}{\sqrt{n}\eta_w} \tag{6.25}$$

ensures that Willie's error probability is lower-bounded by $\mathbb{P}_e^{(w)} \geq \frac{1}{2}-\epsilon$ over n optical modes.

Analysis (Bob): Suppose Bob uses a coherent detection receiver. A homodyne receiver, which is more efficient than a heterodyne receiver in the low photon number regime [33], induces an AWGN channel with noise power $\sigma_b^2 = \frac{2(1-\eta_b)\bar{n}_T+1}{4\eta_b}$. Since Alice uses Gaussian modulation with symbol power $\bar{n}$ defined in equation (6.25), we can upper-bound $\mathbb{P}_e^{(b)}$ by equation (4.9):

$$\mathbb{P}_e^{(b)} \leq 2^{M-\frac{n}{2}\log_2\left(1+\bar{n}/2\sigma_b^2\right)}. \tag{6.26}$$

Substitution of $\bar{n}$ from (6.25) into (6.26) shows that $\mathcal{O}(\sqrt{n})$ bits can be covertly transmitted from Alice to Bob with $\mathbb{P}_e^{(b)} < \delta$ for arbitrary $\delta > 0$ given large enough n. □

6.4 Detector noise also enables covert communication

While any $\bar{n}_T > 0$ enables covert communication, the number of covertly-transmitted bits decreases with $\bar{n}_T$. Blackbody radiation is negligible at optical frequencies (e.g., a typical daytime value of $\bar{n}_T \approx 10^{-6}$ photons per mode at the optical telecom wavelength of 1.55μm [53]). However, other sources of excess noise can also hide the transmissions (e.g. detector dark counts and Johnson noise).

To illustrate the information-hiding capabilities of these noise sources, we consider the (hypothetical) pure-loss channel. Willie's task reduces to that of discriminating between the states corresponding to n-mode vacuum state and the output state $\hat{\rho}_u^{W^n}$ of a pure-loss channel with transmissivity η_w corresponding to an input state $\hat{\rho}_u^{A^n}$, i.e., the states given by equations (6.2) and (6.3) in the proof of Theorem 6.2.1. The minimum probability of discrimination error satisfies [70, Section III]:

$$\frac{1-\sqrt{1-{}^{W^n}\langle \mathbf{0}|\,\hat{\rho}_u^{W^n}\,|\mathbf{0}\rangle^{W^n}}}{2} \leq \min \mathbb{P}_e^{(w)} \leq \frac{1}{2}\,{}^{W^n}\langle \mathbf{0}|\,\hat{\rho}_u^{W^n}\,|\mathbf{0}\rangle^{W^n}\,.$$

Since $\frac{{}^{W^n}\langle \mathbf{0}|\hat{\rho}_u^{W^n}|\mathbf{0}\rangle^{W^n}}{4} \leq \frac{1-\sqrt{1-{}^{W^n}\langle \mathbf{0}|\hat{\rho}_u^{W^n}|\mathbf{0}\rangle^{W^n}}}{2}$, the error probability for the SPD is at most twice that of an optimal detector. Thus, the SPD is an asymptotically optimal detector when the channel from Alice is pure-loss. Since the photon number resolving (PNR) receiver, given by the POVM elements $\{|0\rangle\langle 0|, |1\rangle\langle 1|, |2\rangle\langle 2|, \ldots\}^{\otimes n}$, could be used to mimic the SPD with the detection event threshold set at one photon, the PNR receiver is also asymptotically optimal in this scenario.

Thus, we consider a pure-loss channel where Willie is equipped with a PNR detector. However, any practical implementation of a PNR receiver has a non-zero dark

current, with the noise from the resulting dark counts enabling covert communication even over a pure-loss channel. We model the dark counts per mode in Willie's PNR detector as a Poisson process with average number of dark counts per mode λ_w.

We note that, since his receiver is fixed, lower-bounding Willie's probability of detection error is a classical problem, and we, therefore, can apply Lemmas 4.1 and 4.2. The following theorem demonstrates that, using an on-off keying (OOK) coherent state modulation where Alice transmits the *on* symbol $|\alpha\rangle$ with probability $q = \mathcal{O}(1/\sqrt{n})$ and the *off* symbol $|0\rangle$ with probability $1 - q$, Alice can reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using n OOK symbols:

Theorem 6.4.1. (Dark counts yield square root law) *Suppose that Willie has a pure-loss channel from Alice, captures all photons transmitted by Alice that do not reach Bob, but is limited to a receiver with a non-zero dark current. Let Alice and Bob share a secret of sufficient length before communicating. Then Alice can lower-bound Willie's detection error probability* $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ *for any* $\epsilon > 0$ *while reliably transmitting* $\mathcal{O}(\sqrt{n})$ *bits to Bob in* n *optical modes.*

Proof. Construction: Let Alice use a coherent state on-off keying (OOK) modulation $\{\pi_i, |\psi_i\rangle\langle\psi_i|\}$, $i = 1, 2$, where $\pi_1 = 1 - q$, $\pi_2 = q$, $|\psi_1\rangle = |0\rangle$, $|\psi_2\rangle = |\alpha\rangle$. Alice and Bob generate a random codebook with each codeword symbol chosen i.i.d. from the above binary OOK constellation.

Analysis (Willie): Willie records vector $\mathbf{y}_w = [y_1, \ldots, y_n]$, where y_i is the number of photons observed in the i^{th} mode. Denote by $\mathbb{P}_0$ the distribution of $\mathbf{y}_w$ when Alice does not transmit and by $\mathbb{P}_1$ the distribution when she transmits. When Alice does not transmit, Willie's receiver observes a Poisson dark count process with rate λ_w photons per mode. Thus, $\{y_i\}$ is independent and identically distributed (i.i.d.) sequence of Poisson random variables with rate λ_w, and $\mathbb{P}_0 = \mathbb{P}_w^n$ where $\mathbb{P}_w = \text{Poisson}(\lambda_w)$. When Alice transmits, although Willie captures all of her transmitted energy that does not reach Bob, he does not have access to Alice's and Bob's codebook. Since the

dark counts are independent of the transmitted pulses, each observation is a mixture of two independent Poisson random variables. Thus, each $y_i \sim \mathbb{P}_s$ is i.i.d., with $\mathbb{P}_s = (1-q)\text{Poisson}(\lambda_w) + q\text{Poisson}(\lambda_w + \eta_w|\alpha|^2)$ and $\mathbb{P}_1 = \mathbb{P}_s^n$. By Lemmas 4.1 and 4.2, $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \sqrt{\frac{1}{8}D(\mathbb{P}_0\|\mathbb{P}_1)}$. By Lemma 4.3, $D(\mathbb{P}_0\|\mathbb{P}_1) = nD(\mathbb{P}_w\|\mathbb{P}_s)$. Now,

$$\begin{aligned} D(\mathbb{P}_w\|\mathbb{P}_s) &= -\sum_{y=0}^{\infty} \frac{\lambda_w^y e^{-\lambda_w}}{y!} \log\left[1 - q + q\left(1 + \frac{\eta_w|\alpha|^2}{\lambda_w}\right)e^{-\eta_w|\alpha|^2}\right] \qquad (6.27) \\ &\leq \frac{q^2\left(e^{(\eta_w|\alpha|^2)^2/\lambda_w} - 1\right)}{2} \end{aligned}$$

where the inequality is from the application of Lemma 4.4 to the Taylor series expansion of equation (6.27) with respect to q at $q = 0$. Thus,

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \frac{q}{4}\sqrt{n\left(e^{(\eta_w|\alpha|^2)^2/\lambda_w} - 1\right)}. \qquad (6.28)$$

Therefore, to ensure that $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$, Alice sets

$$q = \frac{4\epsilon}{\sqrt{n\left(e^{(\eta_w|\alpha|^2)^2/\lambda_w} - 1\right)}}. \qquad (6.29)$$

Analysis (Bob): Suppose Bob uses a practical single photon detector (SPD) receiver with probability of a dark click per mode $p_D^{(b)}$. This induces a binary asymmetric channel between Alice and Bob, where the click probabilities, conditioned on the input, are $\mathbb{P}(\text{click} \mid \text{input } |0\rangle) = p_D^{(b)}$ and $\mathbb{P}(\text{click} \mid \text{input } |\alpha\rangle) = 1 - e^{-\eta_b|\alpha|^2}(1 - p_D^{(b)}) \equiv p_C^{(b)}$, with corresponding no-click probabilities $\mathbb{P}(\text{no-click} \mid \text{input } |0\rangle) = 1 - p_D^{(b)}$ and $\mathbb{P}(\text{no-click} \mid \text{input } |\alpha\rangle) = e^{-\eta_b|\alpha|^2}(1 - p_D^{(b)}) \equiv 1 - p_C^{(b)}$. At each mode, a click corresponds to "1" and a no-click to "0". Let Bob use a maximum likelihood decoder on this sequence. Then the standard upper bound on Bob's average decoding error probability is[2] $\mathbb{P}_e^{(b)} \leq e^{M - nE_0}$, where

[2]We use [32, Theorem 5.6.2], setting parameter $\rho = 1$.

$$E_0 = -\ln\left[\left(q\sqrt{p_C^{(b)}} + (1-q)\sqrt{p_D^{(b)}}\right)^2 + \left(q\sqrt{1-p_C^{(b)}} + (1-q)\sqrt{1-p_D^{(b)}}\right)^2\right].$$

The Taylor series expansion of E_0 with respect to q at $q = 0$ yields $E_0 = qC + \mathcal{O}(q^2)$, where

$$C = 2e^{-\eta_b|\alpha|^2/2}\left(e^{\eta_b|\alpha|^2/2} - 1 + p_D^{(b)} - \sqrt{p_D^{(b)}\left(e^{\eta_b|\alpha|^2/2} - 1 + p_D^{(b)}\right)}\right)$$

is a positive constant. Since $q = \mathcal{O}(1/\sqrt{n})$, this demonstrates that $\mathcal{O}(\sqrt{n})$ bits can be covertly transmitted from Alice to Bob with $\mathbb{P}_e^{(b)} < \delta$ for arbitrary $\delta > 0$ given large enough n. □

6.5 Quantum-strong converse of the square root law

Finally, we claim the ultimate unsurmountability of the square root law. We assume non-zero thermal noise ($\bar{n}_T > 0$) in the channel and non-zero dark count rate ($\lambda_w > 0$) in Willie's detector. We restrict the photon number variance of Alice's input states to $\sigma_x^2 = \mathcal{O}(n)$. However, this restriction is not onerous since it subsumes all well-known quantum states of a bosonic mode. However, proving this theorem for input states with unbounded photon number variance per mode remains an open problem. Setting $\lambda_w = 0$ yields the converse for Theorem 6.3.1, and setting $\bar{n}_T = 0$ yields the converse for Theorems 6.4.1 and 7.1.1. Setting $\lambda_w = 0$ and $\bar{n}_T = 0$ yields the conditions for Theorem 6.2.1.

Theorem 6.5.1. (Converse of the square root law) *Suppose Alice only uses n-mode codewords with total photon number variance $\sigma_x^2 = \mathcal{O}(n)$. Then, if she attempts to transmit $\omega(\sqrt{n})$ bits in n modes, as $n \to \infty$, she is either detected by Willie with arbitrarily low detection error probability, or Bob cannot decode with arbitrarily low decoding error probability.*

Proof. As in the proof of Theorem 6.2.1, Alice sends one of 2^M (equally likely) M-bit messages by choosing an element from an arbitrary codebook $\{\hat{\rho}_x^{A^n}\}_{x=1}^{2^M}$, where a state $\hat{\rho}_x^{A^n} = |\psi_x\rangle^{A^n A^n}\langle\psi_x|$ encodes an M-bit message W_x. $|\psi_x\rangle^{A^n} = \sum_{\mathbf{k}\in\mathbb{N}_0^n} a_{\mathbf{k}}(x)\,|\mathbf{k}\rangle$ is a general n-mode pure state, where $|\mathbf{k}\rangle \equiv |k_1\rangle\otimes|k_2\rangle\otimes\cdots\otimes|k_n\rangle$ is a tensor product of n Fock states. The mean photon number of a codeword $\hat{\rho}_x^{A^n}$ is $\bar{n}_x = \sum_{\mathbf{k}\in\mathbb{N}_0^n}(\sum_{i=1}^n k_i)|a_{\mathbf{k}}(x)|^2$, and the photon number variance is $\sigma_x^2 = \sum_{\mathbf{k}\in\mathbb{N}_0^n}(\sum_{i=1}^n k_i)^2|a_{\mathbf{k}}(x)|^2 - \bar{n}_x^2 = \mathcal{O}(n)$. We limit our analysis to pure input states since, by convexity, using mixed states as inputs can only deteriorate the performance (since that is equivalent to transmitting a randomly chosen pure state from an ensemble and discarding the knowledge of that choice).

Willie uses a noisy PNR receiver to observe his channel from Alice, and records the total photon count X_{tot} over n modes. For some threshold S that we discuss later, Willie declares that Alice transmitted when $X_{tot} \geq S$, and did not transmit when $X_{tot} < S$. When Alice does not transmit, Willie observes noise: $X_{tot}^{(0)} = X_D + X_T$, where X_D is the number of dark counts from the spontaneous emission process at the detector, and X_T is the number of photons observed from the thermal background. Since the dark counts are modeled by a Poisson process with rate λ_w photons per mode, both the mean and variance of the observed dark counts per mode is λ_w. The mean of the number of photons observed per mode from the thermal background with mean photon number per mode $\bar{n}_T$ is $(1-\eta_w)\bar{n}_T$ and the variance is $(1-\eta_w)^2(\bar{n}_T+\bar{n}_T^2)$. Thus, the mean of the total number of noise photons observed per mode is $\mu_N = \lambda_w + (1-\eta_w)\bar{n}_T$, and, because of the statistical independence of the noise processes, the variance is $\sigma_N^2 = \lambda_w + (1-\eta_w)^2(\bar{n}_T + \bar{n}_T^2)$. We upper-bound the false alarm probability using Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}_{\text{FA}} &= \mathbb{P}(X_{tot}^{(0)} \geq S) \\ &\leq \frac{n\sigma_N^2}{(S-n\mu_N)^2},\end{aligned} \tag{6.30}$$

where equation (6.30) is because of the memorylessness of the noise processes. Thus, to obtain the desired $\mathbb{P}^*_{\text{FA}}$, Willie sets threshold $S = n\mu_N + \sqrt{n\sigma_N^2/\mathbb{P}^*_{\text{FA}}}$.

When Alice transmits codeword $\hat{\rho}_u^{A^n}$ corresponding to message W_u, Willie observes $X_{tot}^{(1)} = X_u + X_D + X_T$, where X_u is the count from Alice's transmission. We upper-bound the missed detection probability using Chebyshev's inequality:

$$\begin{aligned}
\mathbb{P}_{\text{MD}} &= \mathbb{P}(X_{tot}^{(1)} < S) \\
&\leq \mathbb{P}\left(|X_{tot}^{(1)} - \eta_w \bar{n}_u - n\mu_N| \geq \eta_w \bar{n}_u - \sqrt{\frac{n\sigma_N^2}{\mathbb{P}^*_{\text{FA}}}}\right) \\
&\leq \frac{n\sigma_N^2 + \eta_w^2 \sigma_u^2}{(\eta_w \bar{n}_u - \sqrt{n\sigma_N^2/\mathbb{P}^*_{\text{FA}}})^2},
\end{aligned} \tag{6.31}$$

where equation (6.31) is because the noise and Alice's codeword are independent. Since $\sigma_u^2 = \mathcal{O}(n)$, if $\bar{n}_u = \omega(\sqrt{n})$, then $\lim_{n\to\infty} \mathbb{P}_{\text{MD}} = 0$. Thus, given large enough n, Willie can detect Alice's codewords that have mean photon number $\bar{n}_u = \omega(\sqrt{n})$ with probability of error $\mathbb{P}_e^{(w)} \leq \epsilon$ for any $\epsilon > 0$.

If Alice wants to lower-bound $\mathbb{P}_e^{(w)}$, her codebook must contain a positive fraction of codewords with mean photon number upper-bounded by $\bar{n}_{\mathcal{U}} = \mathcal{O}(\sqrt{n})$. Formally, there must exist a subset of the codebook $\left\{\hat{\rho}_u^{A^n}, u \in \mathcal{U}\right\}$, where $\mathcal{U} = \{u : \bar{n}_u \leq \bar{n}_{\mathcal{U}}\}$, with $\frac{|\mathcal{U}|}{2^M} \geq \kappa$ and $\kappa > 0$. Suppose Bob has an unattenuated pure-loss channel from Alice ($\eta_b = 0$ and $\bar{n}_T = 0$) and access to any receiver allowed by quantum mechanics. The decoding error probability $\mathbb{P}_e^{(b)}$ in such scenario clearly lower-bounds the decoding error probability in a practical scenario where the channel from Alice is lossy and either the channel or the receiver are noisy. Denote by $E_{a\to k}$ the event that a transmitted message W_a is decoded as $W_k \neq W_a$. Since the messages are equiprobable, the average probability of error for the codebook containing only the codewords in $\mathcal{U}$ is

$$\mathbb{P}_e^{(b)}(\mathcal{U}) = \frac{1}{|\mathcal{U}|} \sum_{a \in \mathcal{U}} \mathbb{P}\left(\cup_{k \in \mathcal{U} \setminus \{a\}} E_{a \to k}\right). \tag{6.32}$$

Since the probability that a message is sent from $\mathcal{U}$ is κ,

$$\mathbb{P}_e^{(b)} \geq \kappa \mathbb{P}_e^{(b)}(\mathcal{U}). \tag{6.33}$$

Equality holds only when Bob receives messages that are not in $\mathcal{U}$ error-free and knows when the messages from $\mathcal{U}$ are sent (in other words, equality holds when the set of messages on which decoder can err is reduced to $\mathcal{U}$). Denote by W_a, $a \in \mathcal{U}$, the message transmitted by Alice, and by $\hat{W}_a$ Bob's decoding of W_a. Then, since each message is equiprobable and $|\mathcal{U}| = \kappa 2^M$,

$$\begin{aligned}
\log_2 \kappa + M &= H(W_a) && (6.34)\\
&= I(W_a; \hat{W}_a) + H(W_a|\hat{W}_a) && (6.35)\\
&\leq I(W_a; \hat{W}_a) + 1 + (\log_2 \kappa + M)\mathbb{P}_e^{(b)}(\mathcal{U}) && (6.36)\\
&\leq \chi\left(\left\{\frac{1}{|\mathcal{U}|}, \hat{\rho}_u^{A^n}\right\}\right) + 1 + (\log_2 \kappa + M)\mathbb{P}_e^{(b)}(\mathcal{U}) && (6.37)
\end{aligned}$$

where (6.35) is from the definition of mutual information, (6.36) is because of classical Fano's inequality [19, Equation (9.37)], and (6.37) is the Holevo bound $I(X;Y) \leq \chi(\{p_X(x), \hat{\rho}_x\})$ (see Appendix B.3.6). The mutual information $I(X;Y)$ is between a classical input X and a classical output Y, which is a function of the prior probability distribution $p_X(x)$, and the conditional probability distribution $p_{Y|X}(y|x)$, with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The classical input x maps to a quantum state $\hat{\rho}_x$. A *specific choice* of a quantum measurement, described by POVM elements $\{\hat{\Pi}_y, y \in \mathcal{Y}\}$, induces the conditional probability distribution $p_{Y|X}(y|x) = \mathrm{Tr}\left[\hat{\Pi}_y \hat{\rho}_x\right]$. The Holevo information, $\chi(\{p_x, \hat{\rho}_x\}) = S\left(\sum_{x \in \mathcal{X}} p_x \hat{\rho}_x\right) - \sum_{x \in \mathcal{X}} p_x S(\hat{\rho}_x)$, where $S(\hat{\rho}) \equiv -\mathrm{Tr}[\hat{\rho} \ln \hat{\rho}]$ is the von Neumann entropy of the state $\hat{\rho}$, is not a function of the quantum measurement.

Since $\hat{\rho}_u^{A^n} = |\psi_u\rangle^{A^n A^n}\langle\psi_u|$ is a pure state, $\chi\left(\{\frac{1}{|\mathcal{U}|}, \hat{\rho}_u^{A^n}\}\right) = S\left(\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} |\psi_u\rangle^{A^n A^n}\langle\psi_u|\right)$. Denote the "average codeword" in $\mathcal{U}$ by $\bar{\rho}^{A^n} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} |\psi_u\rangle^{A^n A^n}\langle\psi_u|$, and the state of the j^{th} mode of $\bar{\rho}^{A^n}$ by $\bar{\rho}_j^{A^n}$. We obtain $\bar{\rho}_j^{A^n}$ by taking the partial trace over all the other modes in $\bar{\rho}^{A^n}$ and denote its mean photon number by $\bar{n}_j$ (i.e. $\bar{n}_j$ is the mean photon number of the j^{th} mode of $\bar{\rho}^{A^n}$). Finally, denote a coherent state ensemble with a zero-mean circularly-symmetric Gaussian distribution by $\hat{\rho}_{\bar{n}}^T = \frac{1}{\pi\bar{n}}\int e^{-|\alpha|^2/\bar{n}}|\alpha\rangle\langle\alpha|\mathrm{d}^2\alpha$. The von Neumann entropy of $\hat{\rho}_{\bar{n}}^T$, $S\left(\hat{\rho}_{\bar{n}}^T\right) = \log_2(1+\bar{n}) + \bar{n}\log_2\left(1+\frac{1}{\bar{n}}\right)$. Now,

$$S\left(\bar{\rho}^{A^n}\right) \leq \sum_{j=1}^{n} S\left(\bar{\rho}_j^{A^n}\right) \tag{6.38}$$

$$\leq \sum_{i=1}^{n} \log_2(1+\bar{n}_j) + \bar{n}_j \log_2\left(1+\frac{1}{\bar{n}_j}\right) \tag{6.39}$$

$$\leq n\left(\log_2\left(1+\frac{\bar{n}_{\mathcal{U}}}{n}\right) + \frac{\bar{n}_{\mathcal{U}}}{n}\log_2\left(1+\frac{n}{\bar{n}_{\mathcal{U}}}\right)\right), \tag{6.40}$$

where (6.38) follows from the subadditivity of the von Neumann entropy (see Appendix B.3.2) and (6.39) is because $\hat{\rho}_{\bar{n}}^T$ maximizes the von Neumann entropy of a single-mode state with mean photon number constraint $\bar{n}$ [33]. Now, $S\left(\hat{\rho}_{\bar{n}}^T\right)$ is concave and increasing for $\bar{n} > 0$, and, since $\sum_{j=1}^{n}\bar{n}_j \leq \bar{n}_{\mathcal{U}}$ by construction of $\mathcal{U}$, the application of Jensen's inequality yields (6.40). Combining (6.37) and (6.40) and solving for $\mathbb{P}_e^{(b)}(\mathcal{U})$ yields:

$$\mathbb{P}_e^{(b)}(\mathcal{U}) \geq 1 - \frac{\log_2\left(1+\frac{\bar{n}_{\mathcal{U}}}{n}\right) + \frac{\bar{n}_{\mathcal{U}}}{n}\log_2\left(1+\frac{n}{\bar{n}_{\mathcal{U}}}\right) + \frac{1}{n}}{\frac{\log_2\kappa}{n} + \frac{M}{n}}. \tag{6.41}$$

Substituting (6.41) into (6.33) yields the following lower bound on Bob's decoding error probability:

$$\mathbb{P}_e^{(b)} \geq \kappa\left[1 - \frac{\log_2\left(1+\frac{\bar{n}_{\mathcal{U}}}{n}\right) + \frac{\bar{n}_{\mathcal{U}}}{n}\log_2\left(1+\frac{n}{\bar{n}_{\mathcal{U}}}\right) + \frac{1}{n}}{\frac{\log_2\kappa}{n} + \frac{M}{n}}\right]. \tag{6.42}$$

Since Alice transmits $\omega(\sqrt{n})$ bits in n modes, $M/n = \omega(1/\sqrt{n})$ bits/symbol. However, since $\bar{n}_{\mathcal{U}} = \mathcal{O}(\sqrt{n})$, as $n \to \infty$, $\mathbb{P}_e^{(b)}$ is bounded away from zero for any $\kappa > 0$. Thus, Alice cannot transmit $\omega(\sqrt{n})$ bits in n optical modes both covertly and reliably. $\square$

CHAPTER 7

EXPERIMENTAL EVALUATION OF COVERT OPTICAL COMMUNICATION

In this chapter we corroborate the theoretical results from the previous chapter with a proof-of-concept experiment, where the excess noise in Willie's detection is emulated by dark counts of his single photon detector. This is the first known implementation of a truly quantum-information-theoretically secure covert communication system that allows communication when all transmissions are prohibited.

We use pulse position modulation (PPM) in our experiments, as it allows us to use a practical error correction code (ECC). In Section 7.1 we prove that it can be used for achieving covert communication, and in Section 7.2 we describe our testbed and report the results of our experiments demonstrating the feasibility of quantum-information-theoretically secure covert communication.

7.1 A structured strategy for covert communication

The assumptions of Section 6.4 (hypothetical pure-loss channel, with detector afflicted by the dark counts) describe many optical communication scenarios. While Theorem 6.4.1 states that such settings allow Alice to covertly communicate with Bob, however the skewed on-off duty cycle of OOK modulation used in the proof makes construction of an efficient ECC challenging. Constraining OOK signaling to Q-ary pulse position modulation (PPM) addresses this issue by sacrificing a constant fraction of throughput. Each PPM symbol uses a PPM *frame* to transmit a sequence of Q coherent state pulses, $|0\rangle \dots |\alpha\rangle \dots |0\rangle$, encoding message $i \in \{1, 2, \dots, Q\}$ by

transmitting $|\alpha\rangle$ in the i^{th} mode of the PPM frame. Thus, instead of $\mathcal{O}(n)$ bits allowed by OOK, PPM lets $\mathcal{O}\left(\frac{n}{Q}\log Q\right)$ bits be transmitted in n optical modes. However, PPM performs well in the low photon number regime [86] and the symmetry of its symbols enables us to use any one of efficient ECCs.

To communicate covertly, Alice and Bob use a fraction $\zeta = \mathcal{O}\left(\sqrt{Q/n}\right)$ of n/Q available PPM frames on average, effectively using $\bar{n} = \mathcal{O}(1/\sqrt{n})$ photons per mode. By keeping secret which frames they use, Alice and Bob force Willie to examine all of them, increasing the likelihood of dark counts. An ECC ensures the reliability of the communication between Alice and Bob (if it occurs). While the ECC is revealed to Willie, the transmitted pulse positions are scrambled within the corresponding PPM frames via an operation resembling one-time pad encryption [76], preventing Willie's exploitation of the ECC's structure for detection (rather than protecting the message content). The following theorem demonstrates that, using this scheme, Alice reliably transmits $\mathcal{O}\left(\sqrt{\frac{n}{Q}}\log Q\right)$ covert bits at the cost of pre-sharing $\mathcal{O}\left(\sqrt{\frac{n}{Q}}\log n\right)$ secret bits:

Theorem 7.1.1. (Dark counts yield square root law under structured modulation) *Suppose that Willie has a pure-loss channel from Alice, can capture all photons transmitted by Alice that do not reach Bob, but is limited to a PNR receiver with a non-zero dark current. Let Alice and Bob share a secret of sufficient length before communicating. Then Alice can lower-bound Willie's detection error probability* $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$ *for any* $\epsilon > 0$ *while reliably transmitting* $\mathcal{O}(\sqrt{\frac{n}{Q}}\log Q)$ *bits to Bob using* n *optical modes and a* Q*-ary PPM constellation.*

Proof. Construction: Prior to communication, Alice and Bob secretly choose a random subset $\mathcal{S}$ of PPM frames to use for transmission by selecting each of n/Q available PPM frames independently with probability ζ. Alice and Bob then secretly generate a vector $\mathbf{k}$ containing $|\mathcal{S}|$ integers selected independently uniformly at random from $\{0, 1, \ldots, Q-1\}$, where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. Alice encodes a message

into a codeword of size $|\mathcal{S}|$ using an ECC that may be known to Willie. She adds $\mathbf{k}$ modulo Q to this message and transmits it on the PPM frames in $\mathcal{S}$.

Analysis (Willie): Willie detects each PPM frame received from Alice, recording the photon counts in $\mathbf{y}_w = [\mathbf{y}_1^{(w)}, \ldots, \mathbf{y}_n^{(w)}]$ where $\mathbf{y}_i^{(w)} = [y_{i,1}^{(w)}, \ldots, y_{i,Q}^{(w)}]$ and $y_{i,j}^{(w)}$ it the number of photons observed in the j^{th} mode of the i^{th} PPM frame. Denote by $\mathbb{P}_0$ the distribution of $\mathbf{y}_w$ when Alice does not transmit and by $\mathbb{P}_1$ the distribution when she transmits. When Alice does not transmit, Willie's receiver observes a Poisson dark count process with rate λ_w photons per mode, implying that $\mathbf{y}_w$ is a vector of nQ i.i.d. Poisson(λ_w) random variables. Therefore, $\{\mathbf{y}_i^{(w)}\}$ is i.i.d. with $\mathbf{y}_i^{(w)} \sim \mathbb{P}_w$ and $\mathbb{P}_0 = \mathbb{P}_w^n$, where $\mathbb{P}_w$ is the distribution of Q i.i.d. Poisson(λ_w) random variables with p.m.f.:

$$p_0(\mathbf{y}_i^{(w)}) = \prod_{j=1}^{Q} \frac{\lambda_w^{y_{i,j}^{(w)}} e^{-\lambda_w}}{y_{i,j}^{(w)}!}. \tag{7.1}$$

When Alice transmits, by construction, each PPM frame is randomly selected for transmission with probability ζ. In each selected PPM frame, a pulse is transmitted using one of Q modes chosen equiprobably. Therefore, in this case $\{\mathbf{y}_i^{(w)}\}$ is also i.i.d. with $\mathbf{y}_i^{(w)} \sim \mathbb{P}_s$ and $\mathbb{P}_1 = \mathbb{P}_s^n$, where the p.m.f. of $\mathbb{P}_s$ is:

$$p_1(\mathbf{y}_i^{(w)}) = (1-\zeta) \prod_{j=1}^{Q} \frac{\lambda_w^{y_{i,j}^{(w)}} e^{-\lambda_w}}{y_{i,j}^{(w)}!} + \frac{\zeta}{Q} \sum_{m=1}^{Q} \frac{(\eta_w |\alpha|^2 + \lambda_w)^{y_{i,m}^{(w)}} e^{-\eta_w |\alpha|^2 - \lambda_w}}{y_{i,m}^{(w)}!} \prod_{\substack{j=1 \\ j \neq m}}^{Q} \frac{\lambda_w^{y_{i,j}^{(w)}} e^{-\lambda_w}}{y_{i,j}^{(w)}!}. \tag{7.2}$$

By Lemma 4.3, $D(\mathbb{P}_0 \| \mathbb{P}_1) = \frac{n}{Q} D(\mathbb{P}_w \| \mathbb{P}_s)$. Now, denoting by $\mathbf{x} = [x_1, \cdots, x_Q]$ where $x_j \in \mathbb{N}_0$, we have

$$D(\mathbb{P}_w \| \mathbb{P}_s) = -\sum_{\mathbf{x}\in\mathbb{N}_0^Q} \prod_{j=1}^{Q} \frac{\lambda_w^{x_j} e^{-\lambda_w}}{x_j!} \log\left[1-\zeta+\frac{\zeta}{Q}\sum_{m=1}^{Q}\left(1+\frac{\eta_w|\alpha|^2}{\lambda_w}\right)^{x_m} e^{-\eta_w|\alpha|^2}\right] \quad (7.3)$$
$$\leq \frac{\zeta^2\left(e^{\frac{(\eta_w|\alpha|^2)^2}{\lambda_w}}-1\right)}{2Q}$$

where the inequality is from Lemma 4.4 applied to the Taylor series expansion of equation (7.3) with respect to ζ at $\zeta = 0$. By Lemmas 4.1 and 4.2, $\zeta = \frac{4\epsilon Q}{\sqrt{n\left(e^{\frac{(\eta_w|\alpha|^2)^2}{\lambda_w}}-1\right)}}$ ensures that Willie's error probability is lower-bounded by $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \epsilon$.

Analysis (Bob): As in the proof of Theorem 6.4.1, Bob uses a practical SPD receiver with probability of a dark click $p_D^{(b)}$. Bob examines only the PPM frames in $\mathcal{S}$. If two or more clicks are detected in a PPM frame, a PPM symbol is assigned by selecting one of the clicks uniformly at random. If no clicks are detected, the PPM frame is labeled as an *erasure*. After subtracting $\mathbf{k}$ modulo Q from this vector of PPM symbols (subtraction is not performed on erasures), the resultant vector is passed to the decoder. A random coding argument [32, Theorem 5.6.2] yields reliable transmission of $\mathcal{O}\left(\sqrt{\frac{n}{Q}}\log Q\right)$ covert bits. □

7.2 Implementation of experimental covert optical communication system

7.2.1 System design and implementation

To demonstrate the square-root law of covert optical communication we realized a proof-of-concept test-bed implementation. Here we describe its design and implementation.

7.2.1.1 Alice's encoder

Alice and Bob engage in an n-mode communication session consisting of n/Q Q-ary PPM frames with $Q = 32$. As in the construction of the coding scheme in

the proof of Theorem 7.1.1, prior to communication, Alice and Bob secretly select a random subset $\mathcal{S}$ of PPM frames to use for transmission: each of the n/Q available PPM frames is selected independently with probability ζ. Alice and Bob then secretly generate a vector $\mathbf{k}$ containing $|\mathcal{S}|$ integers selected independently uniformly at random from $\{0, 1, \ldots, Q-1\}$, where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. However, instead of using a random codebook as in the proof of Theorem 7.1.1, Alice encodes a message into a codeword of size $|\mathcal{S}|$ using a $(31, 15)$ Reed-Solomon (RS) error correction code. She adds $\mathbf{k}$ modulo Q to this message and transmits it on the PPM frames in $\mathcal{S}$. RS codes perform well on channels dominated by *erasures*, which occur in low receive-power scenarios, e.g., covert and deep space communication [68].

7.2.1.2 Generation of transmitted symbols

Alice generates the binary sequence of length n describing the transmitted signal, with a "1" at a given location indicating a pulse in that mode, and a "0" indicating the absence of a pulse. First, Alice encodes random data, organized into Q-ary symbols, with an RS code and modulo-Q addition of $\mathbf{k}$ to produce a coded sequence of Q-ary symbols. The value of the i^{th} symbol in this sequence indicates which mode in the i^{th} PPM symbol in the set $\mathcal{S}$ contains a pulse, whereas all modes of the PPM frames not in $\mathcal{S}$ remain empty. Mapping occupied modes to "1" and unoccupied modes to "0" results in the desired length-n binary sequence.

To accurately estimate Willie's detection error probability in the face of optical power fluctuations, the above binary sequence is alternated with a sequence of n "0"s, to produce a final length-$2n$ sequence that is passed to the experimental setup. Willie gets a "clean" look at the channel when Alice is silent using these interleaved "0"s, thus allowing the estimation of both the false alarm and the missed detection probabilities under the same conditions. Bob simply discards the interleaved "0"s.

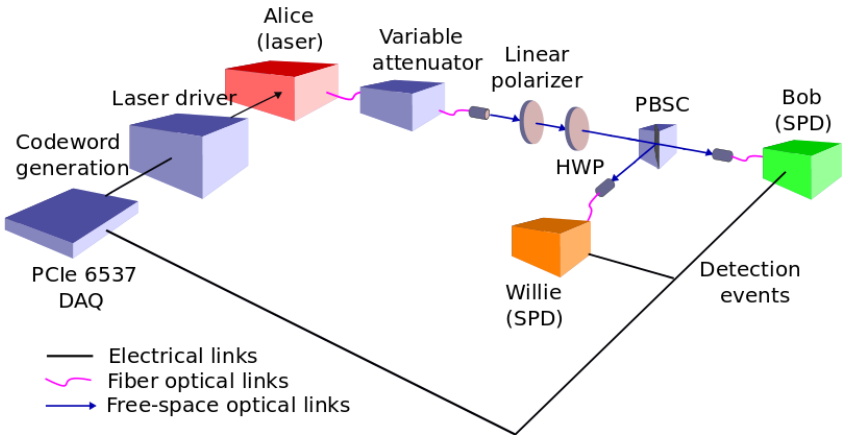


Figure 7.1: Experimental setup. A National Instruments PCIe-6537 data acquisition card (DAQ), driven by a 1 MHz clock, controlled the experiment, generating transmissions and reading detection events. Alice generated 1 ns optical pulses using a temperature-stabilized laser diode with center wavelength 1550.2 nm. The pulses were sent into a free-space optical channel, where a half-wave plate (HWP) and polarizing beamsplitter cube (PBSC) sent a fraction η_b of light to Bob, and the remaining light to Willie. Bob and Willie's receivers operated InGaAs Geiger-mode avalanche photodiode SPDs that were gated with 1 ns reverse bias triggered to match the arrival of Alice's pulses.

We varied n from 3.2×10^6 to 3.2×10^7 in several communication regimes: "careful Alice" ($\zeta = 0.25\sqrt{Q/n}$), "careless Alice" ($\zeta = 0.03\sqrt[4]{Q/n}$), and "dangerously careless Alice" ($\zeta = 0.003$ and $\zeta = 0.008$). For each (n, ζ) pair we conducted 100 experiments and 10^5 Monte-Carlo simulations, measuring Bob's total number of bits received and Willie's detection error probability.

7.2.1.3 Implementation

The experiment was conducted using a mixture of fiber-based and free-space optical elements implementing channels from Alice to both Bob and Willie. As depicted by the schematic in Figure 7.1, we used a National Instruments PCIe-6537 data acquisition card, driven by a 1 MHz clock, to control the experiment. Alice generated 1 ns optical pulses using a temperature-stabilized laser diode with center wavelength 1550.2 nm, the standard optical telecom wavelength. The pulses were sent into a

Table 7.1: Optical channel characteristics

	Willie		Bob	
Experimental estimates	$p_D^{(w)}$	$\bar{n}_{det}^{(w)}$	$p_D^{(b)}$	$\bar{n}_{det}^{(b)}$
$\zeta = 0.25\sqrt{Q/n}$	9.15×10^{-5}	0.036	2.96×10^{-6}	1.52
$\zeta = 0.03\sqrt[4]{Q/n}$	9.12×10^{-5}	0.031	2.54×10^{-6}	1.14
$\zeta = 0.003$	9.29×10^{-5}	0.033	2.62×10^{-6}	1.19
$\zeta = 0.008$	9.28×10^{-5}	0.028	2.63×10^{-6}	1.05
Target:	9×10^{-5}	0.03	3×10^{-6}	1.4

free-space optical channel, where a half-wave plate and polarizing beamsplitter cube were employed to send a fraction η_b of light to Bob, and the remaining light to Willie. Because of the low intensity of Alice's pulses, direct detection using single photon detectors (SPDs), rather than PNR receivers, was sufficient. Bob and Willie's receivers operated InGaAs Geiger-mode avalanche photodiode SPDs that were gated with 1 ns reverse bias triggered to match the arrival of Alice's pulses. Geiger-mode photodiodes have to reset after each detection event, resulting in a deterministic number of no-clicks always following a click [45]. This is known as the *dead time* t_d of a detector, and, in our experiment, $t_d = 16$ observation periods.[1,2]

While some thermal noise is unavoidable, in order to control the experimental environment, several configurations were considered for implementing the background

[1]We note that t_d is adjustable. The detector is forced to reset in order to suppress the *afterpulses*: a sequence of erroneous clicks that immediately follow detection events in Geiger-mode avalanche photodiodes because of the imperfections in their circuitry. To verify our choice of $t_d = 16$ we performed Pearson chi-squared tests of independence [36] between observations x_i and x_{i+k} on each of the four click records corresponding to different communication regimes (i.e., values of ζ) in our experiments. None of the tests rejected the null hypothesis (independence between x_i and x_{i+k}) for $k = 17$ at 5% significance level, while rejecting it for smaller k, thus providing evidence for the correctness of $t_d = 16$ used in our study.

[2]While we account for the dead time in our evaluation of the experiments and the Monte-Carlo simulation, in Appendix D.2 we argue that its impact on Willie's detector's performance is insignificant. We thus do not account for the detector dead time in the proof of Theorem 7.1.1: the construction of Alice and Bob's signaling scheme ensures covert communication since the positive dead time only hurts Willie's detector performance.

noise at the receivers. We provided noise only during the gating period of the detectors since continuous wave light irradiating Geiger-mode avalanche photodiodes (APDs) suppresses detection efficiency [64]. Instead of providing extraneous optical pulses during the gating window of the APD, we emulated optical noise at the detectors by increasing the detector gate voltage, thus increasing the detector's dark click probability. While the APD dark counts are Poisson-distributed with mean rate $\bar{n}_N$ photons per mode, when $\bar{n}_N \ll 1$, the dark click probability $1-e^{-\bar{n}_N}$ is close to $\frac{\bar{n}_N}{1+\bar{n}_N}$, the probability that an incoherent thermal background with mean photon number per mode $\bar{n}_N$ produces a click. In Table 7.1 we report the experimentally-observed estimates and targeted values of dark click probabilities $p_D^{(b)}$ and $p_D^{(w)}$ of Bob's and Willie's detectors, as well as the mean number of photons detected by Bob $\bar{n}_{det}^{(b)} = \eta_b \eta_{\mathrm{QE}}^{(b)} \bar{n}$ and Willie $\bar{n}_{det}^{(w)} = (1-\eta_b)\eta_{\mathrm{QE}}^{(w)} \bar{n}$, where $\bar{n} = 5$ is the mean photon number of Alice's pulses, $\eta_b = 0.97$ is the fraction of light sent to Bob, and $\eta_{\mathrm{QE}}^{(b)}$ and $\eta_{\mathrm{QE}}^{(w)}$ are the quantum efficiencies of Bob's and Willie's detectors, which can be approximated using the estimates in Table 7.1 (we note that quantum efficiency is strongly correlated with the detector's dark click probability [71]). We provide the details of estimating the dark click probability in Appendix D.4. While we observed slight temporal variations in dark counts during our experiments, in Appendix D.4 we argue that their effect on our analysis is minimal.

The amount of transmitted information, with other parameters fixed, is proportional to $\bar{n}_{det}^{(b)}/\bar{n}_{det}^{(w)}$. Our choice of $\bar{n}_{det}^{(b)} \gg \bar{n}_{det}^{(w)}$ allowed the experiment to gather a statistically meaningful data sample in a reasonable duration. In an operational free-space laser communication system, a directional transmitter will likely yield just such an asymmetry in coupling between Bob and Willie; however, we note that the only fundamental requirement for implementing information-theoretically secure covert communication is $p_D^{(w)} > 0$, or $\bar{n}_T > 0$.

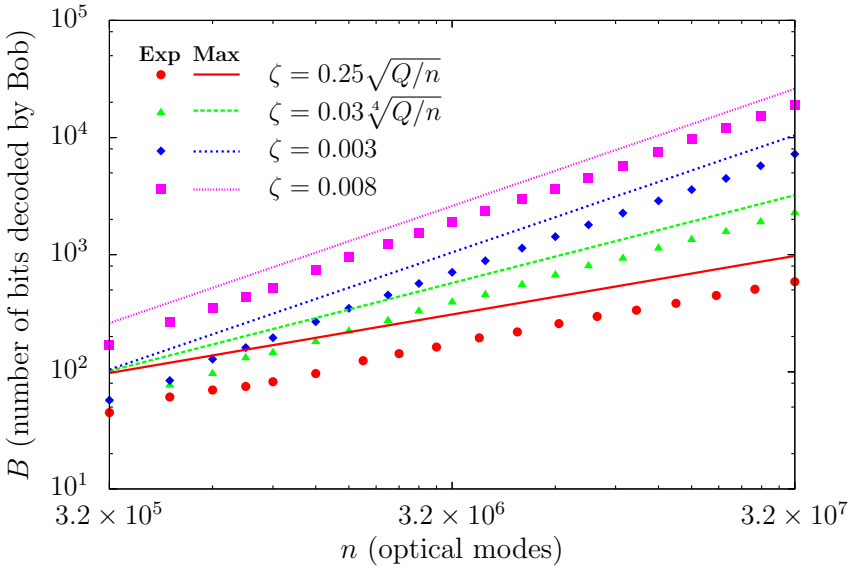


Figure 7.2: Number of bits decoded by Bob. Each data point is an average from 100 experiments, with negligibly small 95% confidence intervals. The symbol error rates are: 1.1×10^{-4} for $\zeta = 0.25\sqrt{Q/n}$, 8.3×10^{-3} for $\zeta = 0.03/\sqrt[4]{Q/n}$, 4.5×10^{-3} for $\zeta = 0.003$, and 1.8×10^{-2} for $\zeta = 0.008$. We also report the maximum throughput $\frac{C_s \zeta n}{Q}$ computed using the experimentally-observed values from Table 7.1, where C_s is the per-symbol Shannon capacity [75]. Given the low observed symbol error rate for $\zeta = 0.25\sqrt{Q/n}$, we note that a square root scaling is achievable even using a relatively short RS code; Figure 7.3 demonstrates that this is achieved covertly.

7.2.2 Analysis

7.2.2.1 Bob's decoder

Bob examines only the PPM frames in $\mathcal{S}$. If two or more pulses are detected in a PPM frame, one of them is selected uniformly at random. If no pulses are detected, it is labeled as an *erasure*. After subtracting $\mathbf{k}$ modulo Q from this vector of PPM symbols (subtraction is not performed on erasures), the resultant vector is passed to the RS decoder.

For each experiment we record the total number of bits in the successfully-decoded codewords; the undecoded codewords are discarded. For each pair of parameters (ζ, n) we report the mean of the total number of bits decoded by Bob over 100 experiments in Figure 7.2. For each communication regime we report the symbol error rate in the caption of Figure 7.2. The symbol error rate is the total number of lost data symbols

during all the experiments at the specified communication regime divided by the total number of data symbols transmitted during these experiments. We also report Bob's maximum throughput from Alice in Figure 7.2, which is the per-symbol Shannon capacity C_s multiplied by the expected number of transmitted PPM symbols $\frac{\zeta n}{Q}$. C_s is calculated for each regime using the experimentally-observed channel characteristics in Table 7.1), with the details of the calculation deferred to Appendix D.1.

Our relatively short $(31, 15)$ RS code achieves between 45% and 60% of the maximum throughput in the "careful Alice" regime and between 55% and 75% of the maximum in other regimes at reasonable error rates, showing that even a basic code demonstrates our theoretical scaling.

7.2.2.2 Willie's detector

Estimation of $\mathbb{P}_e^{(w)}$—Willie's detection problem can be reduced to a test between two simple hypotheses where the log-likelihood ratio test minimizes $\mathbb{P}_e^{(w)}$ [59, Theorem 13.1.1]. The test statistic for the log-likelihood ratio test is derived in Appendix D.2 and is simply the total number of clicks Y observed by Willie. Willie compares Y to a threshold S, accusing Alice if $Y \geq S$. Willie chooses the value of S that minimizes Willie's detection error probability $\mathbb{P}_e^{(w)}$.

For each pair of parameters (n, ζ) as well as Alice's transmission state, we perform m experiments, recording the observed number of clicks Y. We denote by $\left\{Y_i^{(0)}\right\}_{i=1}^m$ and $\left\{Y_i^{(1)}\right\}_{i=1}^m$ the sequences of experimentally observed click counts when Alice does not transmit and transmits, respectively. To estimate Willie's detection error probability $\mathbb{P}_e^{(w)}$, we construct empirical distribution functions $\hat{F}_m^{(0)}(x) = \frac{1}{n}\sum_{i=1}^m \mathbf{1}_{Y_i^{(0)} \leq x}(x)$ and $\hat{F}_m^{(1)}(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{Y_i^{(1)} \leq x}(x)$, where $\mathbf{1}_{\mathcal{A}}(x) = \{1 \text{ if } x \in \mathcal{A};\ 0 \text{ if } x \notin \mathcal{A}\}$ denotes the indicator function. The estimated detection error probability is then

$$\hat{\mathbb{P}}_e^{(w)} = \frac{1}{2}\min_S (1 - \hat{F}_m^{(0)}(S) + \hat{F}_m^{(1)}(S)). \tag{7.4}$$

Monte-Carlo simulation and Gaussian approximation—We perform a Monte-Carlo study using 10^5 simulations per (n, ζ) pair. We generate, encode, and detect the messages as in the physical experiment, and use equation (7.4) to estimate Willie's probability of error, but simulate the optical channel induced by our choice of a laser-light transmitter and an SPD using its estimated characteristics reported in Table 7.1. Similarly, we use the values in Table 7.1 for our analytical Gaussian approximation of $\mathbb{P}_e^{(w)}$ described in Appendix D.3.

Confidence intervals—We compute the confidence intervals for the estimate in equation (7.4) using Dvoretzky-Keifer-Wolfowitz inequality [25, 66], which relates the distribution function $F_X(x)$ of random variable X to the empirical distribution function $\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{X_i \leq x}(x)$ associated with a sequence $\{X_i\}_{i=1}^m$ of m i.i.d. draws of the random variable X as follows:

$$\mathbb{P}(\sup_x |\hat{F}_m(x) - F_X(x)| > \xi) \leq 2e^{-2m\xi^2}, \tag{7.5}$$

where $\xi > 0$. For x_0, the $(1-\alpha)$ confidence interval for the empirical estimate of $F(x_0)$ is given by $[\max\{\hat{F}_m(x_0) - \xi, 0\}, \min\{\hat{F}_m(x_0) + \xi, 1\}]$ where $\xi = \sqrt{\frac{\log(2/\alpha)}{2m}}$. Thus, $\pm\xi$ is used for reporting the confidence intervals in Figure 7.3.

Results—Figure 7.3 reports Willie's probability of error estimated from the experiments and the Monte-Carlo study, as well as its analytical Gaussian approximation. Monte-Carlo simulations show that the Gaussian approximation is accurate. More importantly, Figure 7.3 highlights Alice's safety when she obeys the square root law and her peril when she does not. When $\zeta = \mathcal{O}(1/\sqrt{n})$, $\mathbb{P}_e^{(w)}$ remains constant as n increases. However, for asymptotically larger ζ, $\mathbb{P}_e^{(w)}$ drops at a rate that depends on Alice's carelessness. The drop at $\zeta = 0.008$ vividly demonstrates our converse.

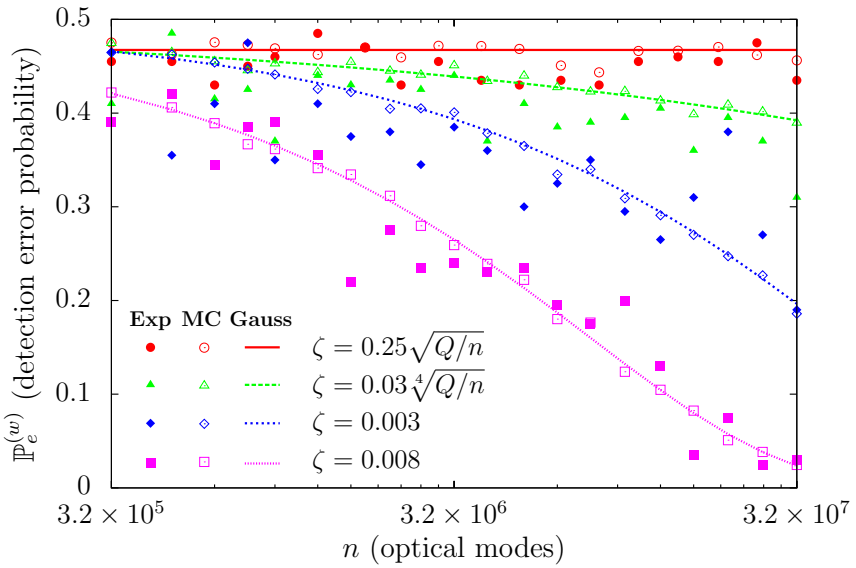


Figure 7.3: Willie's error probability. Estimates from 100 experiments have solid fill; estimates from 10^5 Monte-Carlo simulations have clear fill; and Gaussian approximations are lines. The 95% confidence intervals for the experimental estimates are ± 0.136; for the Monte-Carlo simulations they are ± 0.014. Alice transmits $\zeta n/Q$ PPM symbols on average and Willie's error probability remains constant when Alice obeys the square root law and uses $\zeta = \mathcal{O}(\sqrt{Q/n})$; it drops as n increases if Alice breaks the square root law by using an asymptotically larger ζ.

CHAPTER 8

CONCLUSION AND FURTHER WORK

8.1 Summary

In this thesis we developed the information-theoretic foundation for covert communication. We established the square root limit on the amount of information that can be transmitted both over the AWGN and noisy optical channels. Specifically, we determined that $\mathcal{O}(\sqrt{n})$ covert bits can be sent reliably in n uses of either an AWGN or a single-mode noisy optical channel. Conversely, attempting to transmit more than that either results in detection by the warden with probability one, or a non-zero probability of decoding error as $n \to \infty$. We have also shown that the warden's ignorance of the transmission time provides additional throughput gain, and demonstrated that additive noise is critical for establishing covert communication, whether the source of this noise is the channel or the warden's detection equipment. We corroborated our theory in a proof-of-concept experiment on an optical testbed, which, to our knowledge, is the first known implementation of information-theoretically secure covert communication system.

8.2 Further Work

The field of covert communication is rich with research opportunities. As discussed in the introduction, there are ongoing projects to reduce the size of the secret shared between the parties prior to the communication attempt, as well as to eliminate it completely in some scenarios [16, 54]. The converse of the square root law for covert

optical communication can also be improved by lifting the restriction on the photon number variance in the signaling states.

The investigation of the impact of Alice's message being short relative to the amount of time that she has to send it combined with warden's ignorance of her choice of the transmission time should also be continued. The results in Chapter 5 are for AWGN channels, and the throughput improvement is shown only using the Gaussian random coding with average power constraint. The proof of Theorem 5.2.1 needs to extended to peak power constrained signaling, as well as other channel models (such as DMCs). This would also allow significant reduction in the size of the pre-shared secret, as the exchange of full Gaussian codebooks would no longer be needed. Furthermore, the constraint on the length of Alice's message could be beneficial to Willie in that the time when Alice is not transmitting could be used to calibrate Willie's detector (even when the time of the actual transmission is unknown). For example, this could allow him to accurately estimate his receiver's noise power, and thus nullify the positive covert communication rate when his knowledge of his receiver's noise power is incomplete as in [57, 58]

Our ultimate objective is to enable a "shadow network", illustrated in Figure 1.1, comprised of transmitters, receivers, and friendly jammers that generate artificial noise, impairing wardens' ability to detect transmissions. Relays in covert networks are valuable and require protection while the jammers are cheap, numerous, and disposable (i.e., a warden can silence a particular jammer easily, but, because of their great numbers, silencing enough of them to produce a significant impact is infeasible). Jammer activities are independent from the relay transmission states: that is, wardens cannot detect transmissions by listening to the jammers. Thus, jammers have a parasitic effect on the wardens' SNRs and are a nuisance.

It is important to characterize the scaling behavior of such a network, akin to how [14, 34, 84] extend the results of [28, 39] to the secure (but not covert) multi-

path unicast communication in a large wireless network. The first step towards this goal is extending the covert communication scenario of this thesis to point-to-point jammer-assisted covert communication in the presence of multiple wardens. Preliminary results [79] assume that jammers operate at a constant power, and the signal propagation model accounts only for path loss and AWGN. However, as [57, 58] demonstrate, uncertainty in noise experienced by the warden is beneficial to Alice. Thus, variable jamming power and multipath fading should be incorporated into the jammer-assisted covert communication model, as it may enable covert communication at a positive rate. Completing the characterization of the point-to-point covert link in a multi-warden multi-jammer environment is an important step towards understanding the behavior of "shadow networks", and their eventual implementation.

APPENDIX A

CLASSICAL COVERT COMMUNICATION MISCELLANEA

A.1 Impact of Warden's *a priori* Knowledge of the Transmission State

Our proofs assume that Willie has no prior knowledge on whether Alice transmits or not. Here we argue that the assumption of a non-trivial prior distribution on Alice's transmission state does not impact our asymptotic results. Suppose that Willie knows that Alice will not transmit (i.e. H_0 is true) with probability π_0 and that she will transmit (i.e. H_0 is true) with probability $\pi_1 = 1 - \pi_0$. Denote the probability distribution of Willie's channel observations conditioned on Alice not transmitting (i.e. on H_0 being true) as $\mathbb{P}_0$, and the probability distribution of the observations conditioned on Alice transmitting (i.e. on H_1 being true) as $\mathbb{P}_1$. The following generalized version of Lemma 4.1 then holds:

Lemma A.1 (Generalized Lemma 4.1)**.** $\mathbb{P}_e^{(w)} \geq \min(\pi_0, \pi_1) - \max(\pi_0, \pi_1) V(\mathbb{P}_0, \mathbb{P}_1)$

Proof. Upon observing x, Willie's hypothesis test selects either the null hypothesis H_0 or the alternate hypothesis H_1. Denote by $p_0(x) = p(x|H_0)$ and $p_1(x) = p(x|H_1)$ the probability density functions of x conditioned on either hypothesis H_0 or H_1 being true; $p_0(x)$ and $p_1(x)$ are therefore the probability density functions of $\mathbb{P}_0$ and $\mathbb{P}_1$. Denote by $p(H_0|x)$ and $p(H_1|x)$ the probabilities of hypotheses H_0 and H_0 being true conditioned on the observation x. Since the optimal hypothesis test uses the

maximum *a posteriori* probability rule, the probability $\mathbb{P}_c^{(b)}$ of Willie's optimal test being correct, averaged over all observations, is as follows:

$$\mathbb{P}_c^{(w)} = \int_{\mathcal{X}} \max(p(H_0|x), p(H_1|x))p(x)\mathrm{d}x \tag{A.1}$$

$$= \int_{\mathcal{X}} \max(\pi_0 p_0(x), \pi_1 p_1(x))\mathrm{d}x \tag{A.2}$$

where $\mathcal{X}$ is the support of $p_0(x)$ and $p_1(x)$, and (A.2) follows from Bayes' theorem. Denote the error probability of Willie's optimal test by $\mathbb{P}_e^{(w)} = 1 - \mathbb{P}_c = 1 - \int_{\mathcal{X}} \max(\pi_0 p_0(x), \pi_1 p_1(x))\mathrm{d}x$. Now, since $\max(a, b) = \frac{a+b+|a-b|}{2}$, $\mathbb{P}_e^{(w)}$ can be expressed as follows:

$$\mathbb{P}_e^{(w)} = 1 - \frac{1}{2}\left(\pi_0 \int_{\mathcal{X}} p_0(x)\mathrm{d}x + \pi_1 \int_{\mathcal{X}} p_1(x)\mathrm{d}x\right) - \frac{1}{2}\int_{\mathcal{X}} |\pi_0 p_0(x) - \pi_1 p_1(x)|\mathrm{d}x \tag{A.3}$$

$$= \frac{1}{2} - \frac{1}{2}\|\pi_0 p_0(x) - \pi_1 p_1(x)\|_1 \tag{A.4}$$

where (A.4) is because of the probability densities integrating to one over their supports in the first two integrals of (A.3), $\pi_0 + \pi_1 = 1$, and the last integral in (A.3) being the $\mathcal{L}_1$ norm.

When the prior probabilities of the hypotheses are equal: $\pi_0 = \pi_1 = \frac{1}{2}$, (A.4) yields the proof of Lemma A.1. When $\pi_0 \neq \pi_1$, we can lower-bound (A.4) using the triangle inequality for the $\mathcal{L}_1$ norm:

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \frac{1}{2}\left(\|\pi_0 p_0(x) - \pi_0 p_1(x)\|_1 + \|\pi_0 p_1(x) - \pi_1 p_1(x)\|_1\right) \tag{A.5}$$

$$= \frac{1}{2} - \frac{|\pi_0 - \pi_1|}{2} - \frac{\pi_0}{2}\|p_0(x) - p_1(x)\|_1 \tag{A.6}$$

where (A.6) follows from the $\mathcal{L}_1$ norm of a probability density function evaluating to one and $\pi_0 > 0$. If $\pi_1 > \pi_0$, the following application of the triangle inequality yields a tighter bound:

$$\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \frac{1}{2}\left(\|\pi_1 p_1(x) - \pi_1 p_0(x)\|_1 + \|\pi_1 p_0(x) - \pi_0 p_0(x)\|_1\right) \tag{A.7}$$

$$= \frac{1}{2} - \frac{|\pi_0 - \pi_1|}{2} - \frac{\pi_1}{2}\|p_0(x) - p_1(x)\|_1 \tag{A.8}$$

By Definition 4.1, $\frac{1}{2}\|p_0(x) - p_1(x)\|_1 = V(\mathbb{P}_0, \mathbb{P}_1)$. Since $\min(a, b) = \frac{a+b-|a-b|}{2}$, we can combine (A.6) and (A.8) to yield

$$\mathbb{P}_e^{(w)} \geq \min(\pi_0, \pi_1) - \max(\pi_0, \pi_1) V(\mathbb{P}_0, \mathbb{P}_1) \tag{A.9}$$

which completes the proof. □

Thus, while Lemma A.1 demonstrates that additional information about the likelihood of Alice transmitting (in the form of unequal prior probabilities $\pi_0 \neq \pi_1$) helps Willie, the square root law still holds via the bounds on the variational distance $V(\mathbb{P}_0, \mathbb{P}_1)$.

A.2 Mapping to a Continuous-time Channel

We employ a discrete-time model in Chapters 4 and 5. However, while this is commonly assumed without loss of generality in standard communication theory, it is important to consider whether we have missed some aspect of the covert communication problem by focusing on discrete time.

Consider the standard communication system model, where Alice's (baseband) continuous-time waveform is given in terms of her discrete time transmitted sequence by:

$$x(t) = \sum_{i=1}^{n} f_i \; p(t - iT_s)$$

where T_s is the symbol period and $p(\cdot)$ is the pulse shaping waveform. Consider a (baseband) system bandwidth constraint of W Hz. Now, if Alice chooses $p(\cdot)$ ideally

as $\mathrm{sinc}(2Wt)$, where $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$, then the natural choice of $T_s = 1/2W$ results in no intersymbol interference (ISI). From the Nyquist sampling criterion, both Willie (and Bob) can extract all of the information from the signaling band by sampling at a rate of $2W$ samples/second, which then leads directly to the discrete-time model of Sections 4.1 and 5.1, and suits our demonstration of the fundamental limits to Alice's covert communication capabilities over AWGN channels. However, when $p(\cdot)$ is chosen in a more practical fashion, for example, as a raised cosine pulse with some excess bandwidth, then sampling at a rate higher than $2W$ has utility for signal detection even if the Nyquist ISI criterion is satisfied. In particular, techniques involving cyclostationary detection are now applicable, and we consider such a scenario a promising area for future work.

A.3 $D(\mathbb{P}_w \| \mathbb{P}_s)$ in the Proof of Theorem 4.2.2 Meets the Conditions of Lemmas 4.4 and 4.5

Here we show that the expression (4.10) for the relative entropy $D(\mathbb{P}_w \| \mathbb{P}_s)$ between the distributions $\mathbb{P}_w$ and $\mathbb{P}_s$ of an observation on Willie's channel from Alice corresponding to Alice's transmission state meets the regularity conditions of Taylor's theorem (Lemma 4.4) and Leibniz integral rule (Lemma 4.5). Specifically, we need to show that $D(\mathbb{P}_w \| \mathbb{P}_s)$ and its first six derivatives are continuous on $[0, \sqrt{b}]$ and integrable.

Re-arranging the terms of (4.10) results in the following expression:

$$D(\mathbb{P}_w \| \mathbb{P}_s) = \frac{a^2}{2\sigma_w^2} - \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln \cosh\left(\frac{ax}{\sigma_w^2}\right) \mathrm{d}x \tag{A.10}$$

where $\cosh(x) = \frac{e^x + e^{-x}}{2}$ is the hyperbolic cosine function. Since $\frac{a^2}{2\sigma_w^2}$ is clearly continuous and differentiable with respect to a, we focus on the integral in (A.10), specifically on its integrand:

$$K(x,a) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln\cosh\left(\frac{ax}{\sigma_w^2}\right) \tag{A.11}$$

Because of the peak power constraint, $0 \leq a \leq \sqrt{b}$. Also, $\ln\cosh(x) \leq |x|$ since $\ln\left(\frac{e^x+e^{-x}}{2}\right) - |x| = \ln\left(\frac{1+e^{-2|x|}}{2}\right) \leq 0$. Therefore, $g(x) = \frac{\sqrt{b}|x|e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w^3} \geq |K(x,a)|$, in other words, $g(x)$ dominates $K(x,a)$. $g(x)$ is integrable since $\int_{-\infty}^{\infty} g(x)\mathrm{d}x = \sqrt{\frac{2b}{\pi\sigma_w^2}} < \infty$.

The derivatives of $K(x,a)$ with respect to a can be written in the following form:

$$\frac{\partial^i K(x,a)}{\partial a^i} = \begin{cases} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{x^i}{\sigma_w^{2i}}\tanh\left(\frac{ax}{\sigma_w^2}\right)\sum_{k=1}^{(i-1)/2} c_{i,k}\operatorname{sech}^{2k}\left(\frac{ax}{\sigma_w^2}\right) & , \ i \text{ odd} \\ \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{x^i}{\sigma_w^{2i}}\sum_{k=1}^{i/2} c_{i,k}\operatorname{sech}^{2k}\left(\frac{ax}{\sigma_w^2}\right) & , \ i \text{ even} \end{cases} \tag{A.12}$$

where $\operatorname{sech}(x) = \frac{2}{e^x+e^{-x}}$ and $\tanh(x) = \frac{e^x-e^{-x}}{e^x+e^{-x}}$ are the hyperbolic secant and tangent functions, respectively, $c_{i,k}$ are constants, and $\sum_{k=1}^{0} c_{i,k} = 1$. The first six derivatives of $K(x,a)$ with respect to a are as follows:

$$\frac{\partial K(x,a)}{\partial a} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{x}{\sigma_w^2}\tanh\left(\frac{ax}{\sigma_w^2}\right) \tag{A.13}$$

$$\frac{\partial^2 K(x,a)}{\partial a^2} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{x^2}{\sigma_w^4}\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right) \tag{A.14}$$

$$\frac{\partial^3 K(x,a)}{\partial a^3} = -\frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{2x^3}{\sigma_w^6}\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right)\tanh\left(\frac{ax}{\sigma_w^2}\right) \tag{A.15}$$

$$\frac{\partial^4 K(x,a)}{\partial a^4} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{2x^4}{\sigma_w^8}\left(2\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right) - 3\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{A.16}$$

$$\frac{\partial^5 K(x,a)}{\partial a^5} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{8x^5\tanh\left(\frac{ax}{\sigma_w^2}\right)}{\sigma_w^{10}}\left(3\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right) - \operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{A.17}$$

$$\frac{\partial^6 K(x,a)}{\partial a^6} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\frac{8x^6}{\sigma_w^{12}}\left(15\operatorname{sech}^6\left(\frac{ax}{\sigma_w^2}\right) - 15\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right) + 2\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{A.18}$$

Clearly, $K(x,a)$ and its derivatives are continuous, satisfying conditions 1 and 2 of Lemma 4.5. Since $-1 \leq \tanh(x) \leq 1$ and $0 \leq \operatorname{sech}(x) \leq 1$ for all real x, we can use the triangle inequality to show that $\left|\frac{\partial^i K(x,a)}{\partial a^i}\right| \leq h_i(x)$ where

$$h_i(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|^i}{\sigma_w^{2i}} \sum_{k=1}^{\lfloor i/2 \rfloor} |c_{i,k}| \tag{A.19}$$

with $\lfloor x \rfloor$ denoting the largest integer $y \leq x$. Therefore, the following relations show dominating functions of the corresponding derivatives of $K(x,a)$:

$$\left|\frac{\partial K(x,a)}{\partial a}\right| \leq h_1(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|}{\sigma_w^2} \tag{A.20}$$

$$\left|\frac{\partial^2 K(x,a)}{\partial a^2}\right| \leq h_2(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|^2}{\sigma_w^4} \tag{A.21}$$

$$\left|\frac{\partial^3 K(x,a)}{\partial a^3}\right| \leq h_3(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{2|x|^3}{\sigma_w^6} \tag{A.22}$$

$$\left|\frac{\partial^4 K(x,a)}{\partial a^4}\right| \leq h_4(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{10|x|^4}{\sigma_w^8} \tag{A.23}$$

$$\left|\frac{\partial^5 K(x,a)}{\partial a^5}\right| \leq h_5(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{32|x|^5}{\sigma_w^{10}} \tag{A.24}$$

$$\left|\frac{\partial^6 K(x,a)}{\partial a^6}\right| \leq h_6(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{256|x|^6}{\sigma_w^{12}} \tag{A.25}$$

Clearly, the above functions are integrable since they are found in the integrands of the central absolute moments of the Gaussian distribution. Therefore, conditions 3 and 4 of Lemma 4.5 are met by the integrand of (4.10) and the integrand's derivatives.

The use of Lemma 4.4 is conditional on the integrals over x of $K(x,a)$ and its derivatives in (A.12) being continuous on $a \in [0, \sqrt{b}]$. To prove the continuity of a function $f(x)$ on the interval $[u,v]$, it is sufficient to show that $\lim_{x \to x_0} f(x) = f(x_0)$ for all $x_0 \in [u,v]$. We prove that $\int_{-\infty}^{\infty} K(x,a)\mathrm{d}x$ is continuous as follows:

$$\lim_{a\to a_0}\int_{-\infty}^{\infty} K(x,a)\mathrm{d}x = \int_{-\infty}^{\infty}\lim_{a\to a_0} K(x,a)\mathrm{d}x = \int_{-\infty}^{\infty} K(x,a_0)\mathrm{d}x \tag{A.26}$$

where the first equality is because of the application of the dominated convergence theorem, which is valid since we provide the function $g(x)$ above that dominates $K(x,a)$ and is integrable, and the second equality is because of the continuity of $K(x,a)$. Similar steps can be used to prove the continuity of the integrals of the derivatives of $K(x,a)$, with the ultimate result being the satisfaction of the continuity condition of Lemma 4.4.

A.4 Size of Secret for Covert Communication over AWGN Channels is $\mathcal{O}(\sqrt{n}\log n)$ Bits

Here we demonstrate how Alice and Bob can construct a binary coding scheme for covert communication over the AWGN channel described in Chapter 4 that, on average, requires an $\mathcal{O}(\sqrt{n}\log n)$-bit secret. Figure A.1 depicts the construction and operation of this scheme.

The scheme is constructed in two stages. First, Alice and Bob randomly select the symbol periods that they will use for their transmission by flipping a biased coin n times, with probability of heads τ to be assigned later. The i^{th} symbol period is selected if the i^{th} flip is heads. Denote the number of selected symbol periods by n_s and note that $\mathbb{E}[n_s] = \tau n$. Alice and Bob then use the best public binary codebook with codewords of length n_s on these selected n_s symbol periods. They also generate and share a random binary vector $\mathbf{k}$ where $p_{\mathbf{K}}(\mathbf{k}) = \prod_{i=1}^{n_s} p_K(k_i)$ with $p_K(0) = p_K(1) = \frac{1}{2}$. Alice XORs $\mathbf{k}$ and the binary representation of the codeword $\mathbf{c}(W_k)$. The symbol location selection is independent of both the symbol and the channel noise. When Alice is transmitting a codeword, the distribution of each of Willie's observations is $\mathbb{P}_s = (1-\tau)\mathcal{N}(0,\sigma_w^2) + \frac{\tau}{2}\left(\mathcal{N}(-a,\sigma_w^2) + \mathcal{N}(a,\sigma_w^2)\right)$ and, thus,

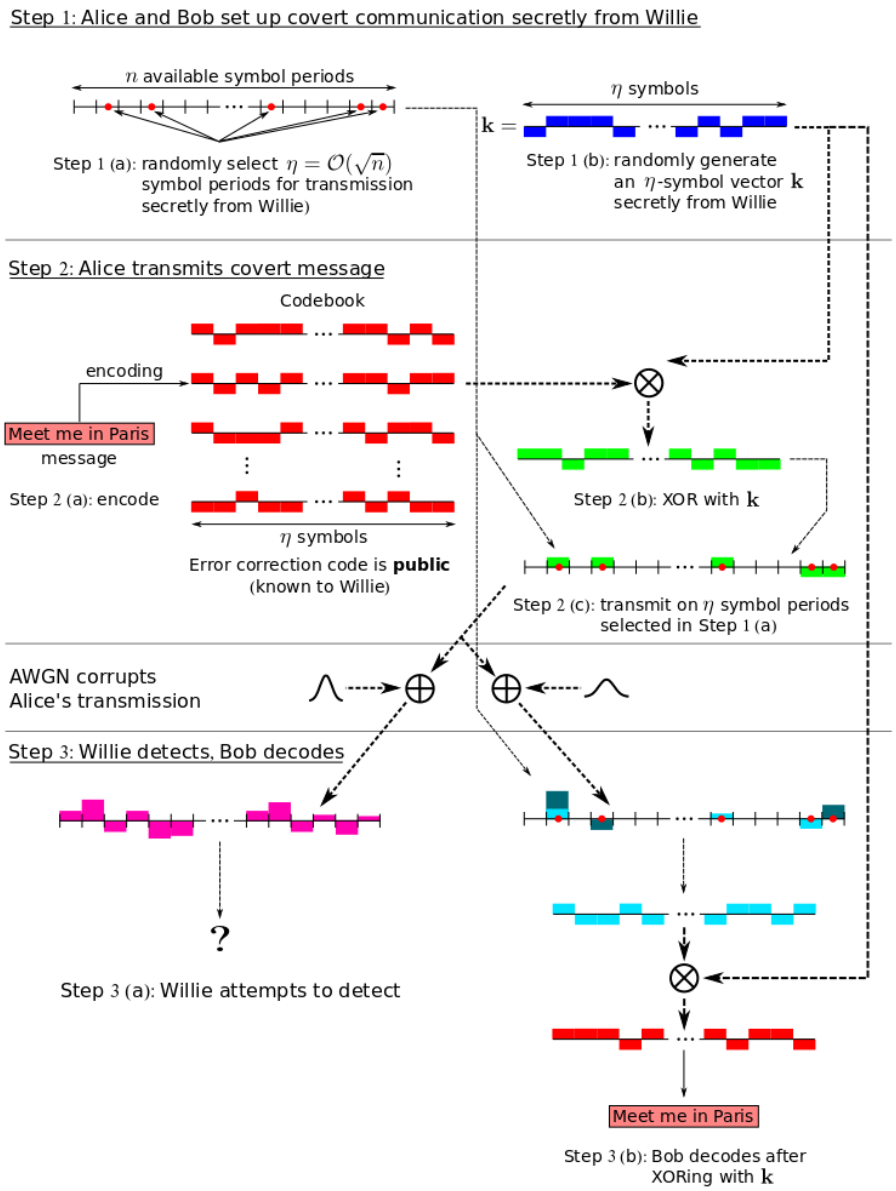


Figure A.1: Design of a covert communication system that allows Alice and Bob to use any error-correction codes (including those known to Willie) to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits using $\mathcal{O}(\sqrt{n}\log n)$ pre-shared secret bits. Step 1 (a) effectively constructs a frequency/time-hopping pattern, as m symbol periods are selected to be used in the transmission by flipping biased random coin n times, with probability of heads $\mathcal{O}(\sqrt{n})$: the i^{th} symbol period is chosen if the i^{th} flip is heads. On average, $\mathcal{O}(\sqrt{n})$ symbol periods is selected . Bob simply ignores the discarded symbol periods, however, Willie cannot do so and thus observes mostly noise. Furthermore, XORing by vector $\mathbf{k}$ prevents Willie's exploitation of the error correction code's structure to detect Alice (rather than protects the message content). Note that in Chapter 7 the extension of this scheme to Q-ary pulse-position modulation is implemented on an optical testbed.

$$D(\mathbb{P}_w \| \mathbb{P}_s) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln \frac{e^{-\frac{x^2}{2\sigma_w^2}}/\sqrt{2\pi}\sigma_w}{\frac{(1-\tau)e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} + \frac{\tau}{2}\left(\frac{e^{-\frac{(x+a)^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} + \frac{e^{-\frac{(x-a)^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\right)} \mathrm{d}x \qquad \text{(A.27)}$$

There is no closed-form expression for (A.27), but we can upper-bound it using Lemma 4.4. The Taylor series expansion with respect to a around $a = 0$ can be done using Lemma 4.5, with conditions for Lemmas 4.4 and 4.5 proven similarly as in Theorem 4.2.2. This yields the following bound:

$$V(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \frac{\tau a^2}{2\sigma_w^2}\sqrt{\frac{n}{2}} \qquad \text{(A.28)}$$

The only difference in (A.28) from (4.12) is τ in the numerator. Thus, if Alice sets the product $\tau a^2 \leq \frac{cf(n)}{\sqrt{n}}$, with c and $f(n)$ as previously defined, she limits the performance of Willie's detector. This product is the average symbol power used by Alice. Now fix a and set $\tau = \mathcal{O}(1/\sqrt{n})$. Since, on average, τn symbol periods are selected, it takes (again, on average) $\mathcal{O}(\sqrt{n})$ positive integers to enumerate the selected symbols. There are n total symbols, and, thus, it takes at most $\log(n)$ bits to represent each selected symbol location and $\mathcal{O}(\sqrt{n}\log n)$ bits to represent all the locations of selected symbols. Also, the average length of the secret binary vector $\mathbf{k}$ is $\mathcal{O}(\sqrt{n})$ bits. Thus, on average, Alice and Bob need to share $\mathcal{O}(\sqrt{n}\log n)$ secret bits for Alice to reliably transmit $\mathcal{O}(\sqrt{n})$ covert bits in n channel uses employing this coding scheme.

A.5 Derivation of (5.24)

The normalized sums $\frac{1}{\sqrt{V(n)}}\sum_{t=1,t\neq t_A}^{T(n)}(U_t - M(n))$ and $\frac{1}{\sqrt{V(n)}}\sum_{t=1}^{T(n)-1}(U_t - M(n))$ in the events $E_g(S(n),\delta)$ and $E_l(S(n),\delta)$ are identically distributed. Thus, we denote both of them by $Z(n)$, and the distribution function of $Z(n)$ by $F_{Z(n)}(z)$. Then (5.23) can be re-written as follows:

$$\mathbb{P}(E_C(S(n),\delta)) = \frac{1 - F_{Z(n)}(S(n)+\delta) + F_{Z(n)}(S(n)-\delta)}{2}. \tag{A.29}$$

Denote the standard Gaussian distribution function by $\Phi(z) = \int_{-\infty}^{z} \phi(t)dt$ where $\phi(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}}$ is the standard Gaussian density function. The convergence of $F_{Z(n)}(z)$ to $\Phi(z)$ as provided by the CLT for the triangular arrays in [9, Th. 27.2] is pointwise in the argument z, and, since $S(n)$ is the n^{th} value in an arbitrary sequence, we cannot use this result directly.

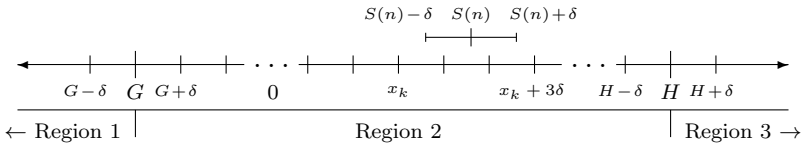


Figure A.2: The real number line partitioned into three regions for the analysis of $\mathbb{P}(E_C(S(n),\delta))$. G, H and δ are the constants that we select. $S(n)$ satisfying $G \le S(n) \le H$ is illustrated.

However, let's choose finite constants $G < 0$ and $H > 0$, and partition the real number line into three regions as shown in Figure A.2. Clearly, for any n, $S(n)$ is in one of these regions. Next we demonstrate that (5.24) holds for an arbitrary $S(n)$ by appropriately selecting G, H, and δ.

Consider $S(n) < G$, or region 1 in Figure A.2:

$$\mathbb{P}(E_C(S(n),\delta)) \ge \frac{1}{2}\left(1 - F_{Z(n)}(S(n)+\delta)\right) \tag{A.30}$$

$$\ge \frac{1}{2}\left(1 - F_{Z(n)}(G+\delta)\right). \tag{A.31}$$

Because the convergence of $F_{Z(n)}(z)$ to $\Phi(z)$ is pointwise, given δ, ϵ, and $G = \Phi^{-1}(\epsilon/3) - \delta$, there exists n_2 such that, for all $n \ge n_2$,

$$\mathbb{P}(E_C(S(n),\delta)) \ge \frac{1}{2}\left(1 - \Phi(G+\delta) - \frac{\epsilon}{3}\right) = \frac{1}{2} - \frac{\epsilon}{3} \tag{A.32}$$

when $S(n) < G$. Similarly for $S(n) > H$, or region 3 in Figure A.2:

$$\mathbb{P}(E_C(S(n), \delta)) \geq \frac{1}{2} F_{Z(n)}(S(n) - \delta) \tag{A.33}$$

$$\geq \frac{1}{2} F_{Z(n)}(H + \delta). \tag{A.34}$$

Again, because the convergence of $F_{Z(n)}(z)$ to $\Phi(z)$ is pointwise, given δ, ϵ, and $H = \Phi^{-1}(1 - \epsilon/3) + \delta$, there exists n_3 such that, for all $n \geq n_3$,

$$\mathbb{P}(E_C(S(n), \delta)) \geq \frac{1}{2} \left(\Phi(H + \delta) - \frac{\epsilon}{3} \right) = \frac{1}{2} - \frac{\epsilon}{3} \tag{A.35}$$

when $S(n) > H$.

Finally, consider $S(n)$ satisfying $G \leq S(n) \leq H$, or region 2 in Figure A.2). Let's assume that H and G are selected so that $H - G$ is an integer multiple of δ (e.g., using larger H than necessary, which results in the RHS of (A.35) being smaller). Consider a sequence $(x_k)_{k=0}^{(H-G)/\delta+2}$ where $x_0 = G - \delta$, $x_1 = G$, $x_2 = G + \delta$, $x_3 = G + 2\delta, \ldots, x_{(H-G)/\delta} = H - \delta$, $x_{(H-G)/\delta+1} = H$, $x_{(H-G)/\delta+2} = H + \delta$. Sequence $(x_k)_{k=0}^{(H-G)/\delta+2}$ partitions region 2 into $\frac{H-G}{\delta} + 2$ subregions, and, for any $S(n)$ satisfying $G \leq S(n) \leq H$, there exists $k \in \left\{0, \ldots, \frac{H-G}{\delta} + 2\right\}$ such that $x_k \leq S(n) - \delta < S(n) + \delta \leq x_k + 3\delta$, as illustrated in Figure A.2. Therefore, since $F_{Z(n)}(z)$ is monotonic,

$$\mathbb{P}(E_C(S(n), \delta)) \geq \frac{1 - F_{Z(n)}(x_k + 3\delta) + F_{Z(n)}(x_k)}{2}. \tag{A.36}$$

Since the convergence of $F_{Z(n)}(z)$ to $\Phi(z)$ is pointwise, for a given x_k, δ, and ϵ, there exists m_k such that for all $n \geq m_k$,

$$\mathbb{P}(E_C(S(n), \delta)) \geq \frac{1 - \left(\Phi(x_k + 3\delta) + \frac{\epsilon}{6}\right) + \left(\Phi(x_k) - \frac{\epsilon}{6}\right)}{2}$$

$$= \frac{1}{2} \left(1 - \int_{x_k}^{x_k + 3\delta} \phi(t) dt - \frac{\epsilon}{3} \right) \tag{A.37}$$

$$\geq \frac{1}{2} - \frac{3\delta}{2\sqrt{2\pi}} - \frac{\epsilon}{6} \tag{A.38}$$

where (A.38) follows from $\phi(t) \leq \frac{1}{\sqrt{2\pi}}$. Setting $\delta = \frac{\epsilon\sqrt{2\pi}}{9}$ and $n_4 = \max_{\left\{0,\ldots,\frac{H-G}{\delta}\right\}}(m_k)$ yields the desired lower bound for all $n \geq n_4$ when $S(n)$ satisfies $G \leq S(n) \leq H$.

Therefore, for an arbitrary $S(n)$ when $n \geq n_0$ where $n_0 = \max(n_2, n_3, n_4)$,

$$\mathbb{P}\left(E_C\left(S(n), \frac{\epsilon\sqrt{2\pi}}{9}\right)\right) \geq \frac{1}{2} - \frac{\epsilon}{3}. \tag{A.39}$$

APPENDIX B

QUANTUM COMMUNICATION AND INFORMATION THEORY PRELIMINARIES

This appendix provides a brief background on quantum mechanics, quantum optics, and quantum information theory that will be useful in reading this thesis.

B.1 Quantum Mechanics: States, Evolution, and Measurement[1]

While the foundations of quantum mechanics date to the early 1800s, the modern discipline began with Max Planck's work in the early 1900s. Max Planck discovered that the energy of electromagnetic waves must be described as consisting of small packets of energy or "quanta" in order to explain the spectrum of black-body radiation. He postulated that a radiating body consisted of an enormous number of elementary electronic oscillators, some vibrating at one frequency and some at another, with all frequencies from zero to infinity being represented. The energy E of any one oscillator was not permitted to take on any arbitrary value, but was proportional to an integer multiple of the frequency f of the oscillator, i.e., $E = hf$, where $h = 6.626 \times 10^{-34}$ Joule seconds is the Planck's constant. In 1905, Albert Einstein used Planck's constant to explain the photoelectric effect by postulating that the energy in a beam of light occurs in concentrations that he called "light quanta," which later became known as *photons*. This led to a theory that established a duality

[1]The content of this section was adapted from [38, Appendix A.1] with permission of the author.

between subatomic particles and electromagnetic waves in which particles and waves were neither one nor the other, but had certain properties of both.

The acceptance of quantum mechanics[2] by the general physics community stems from its accurate prediction of the physical behavior of systems, particularly of systems showing previously unexplained phenomena in which Newtonian mechanics fails, such as the black body radiation, photoelectric effect, and stable electron orbits. Most of classical physics is now recognized to be composed of special cases of quantum mechanics and/or relativity theory. Paul Dirac brought relativity theory to bear on quantum physics so that it could properly deal with events that occur at a substantial fraction of the speed of light. Classical physics, however, also deals with gravitational forces, and no one has yet been able to bring gravity into a unified theory with the relativized quantum theory.

Here we provide a very brief account of the mathematical formulation of quantum mechanics that serves as a background for the material in Chapter 6 of this thesis. The detailed study of quantum mechanics is available in many popular texts on the subject, such as [37] and [72].

B.1.1 Pure and Mixed States

A *pure state* in quantum mechanics is the entirety of information that can be known about a physical system. Mathematically, a pure state is a unit length vector $|\psi\rangle$ (known as a *ket* in Dirac notation) that lives in a complex Hilbert space $\mathcal{H}$ of possible states for that system. Expressed in terms of a set of complete basis vectors $\{|\phi_n\rangle\} \in \mathcal{H}$, $|\psi\rangle = \sum_n c_n|\phi_n\rangle$ becomes a column vector of (a possibly infinite) set of complex numbers c_n, where $\sum_n |c_n|^2 = 1$. With each pure state $|\psi\rangle$ we associate its Hermitian conjugate vector (known as a *bra*) $\langle\psi|$, which is a row vector when expressed in a basis of $\mathcal{H}$. The simplest example of a pure state is the state of a two-

[2]The term "quantum mechanics" was coined by Max Born in 1924.

level system known as a *qubit*, which is the fundamental unit of quantum information, in analogy with a *bit* of classical information. A qubit lives in the two-dimensional complex vector space $\mathbb{C}^2$ spanned by two orthonormal vectors $|0\rangle$ and $|1\rangle$, and can be expressed as $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where $\alpha, \beta \in \mathbb{C}$, and $|\alpha|^2 + |\beta|^2 = 1$.

A *mixed state* in quantum mechanics represents classical (statistical) uncertainty about a physical system. Mathematically, a mixed state is represented by a *density matrix* (or a density operator) $\hat{\rho}$, which is a positive definite, unit-trace operator in $\mathcal{H}$. The canonical form of a density matrix is

$$\hat{\rho} = \sum_k p_k |\psi_k\rangle\langle\psi_k| \tag{B.1}$$

for any collection of pure states $\{|\psi_k\rangle\}$, and $\sum_k p_k = 1$. The mixed state $\hat{\rho}$ can be thought of as a statistical mixture of pure states $|\psi_k\rangle$, where the projection $|\psi_k\rangle\langle\psi_k|$ is the density operator for the pure state $|\psi_k\rangle$, though it is worth pointing out that the decomposition of a mixed state $\hat{\rho}$ as a mixture of pure states (B.1) is by no means unique. A positive definite operator $\hat{\rho}$ has a spectral decomposition $\hat{\rho} = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ in terms of the eigenkets $|\lambda_i\rangle$, with the unit-trace condition on $\hat{\rho}$ requiring that the eigenvalues λ_i form a probability distribution.

B.1.2 Composite Quantum Systems

We shall henceforth use symbols such as A, B, C to refer to quantum systems, with $\mathcal{H}_A$ referring to the Hilbert space whose unit vectors are the pure states of the quantum system A. Given two systems A and B, the pure states of the composite system AB correspond to unit vectors in $\mathcal{H}_{AB} \equiv \mathcal{H}_A \otimes \mathcal{H}_B$. We use superscripts on pure state vectors and density matrices to identify the quantum system with which they are associated. For a multipartite density matrix $\hat{\rho}^{ABC}$, we use the notation $\hat{\rho}^{AB} = \mathrm{Tr}_C\{\hat{\rho}^{ABC}\} \equiv \sum_n {}^C\langle\phi_n|\hat{\rho}^{ABC}|\phi_n\rangle^C$ to denote the *partial trace* over one of the constituent quantum systems.

Let $\{|\phi_m\rangle^A\}$ and $\{|\phi_n\rangle^B\}$ represent sets of basis vectors for the state spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ of quantum systems A and B respectively. Pure states $|\psi\rangle^{AB}$ and mixed states $\hat{\rho}^{AB}$ of the composite system AB are defined similarly with an underlying set of basis vectors $|\phi_{mn}\rangle^{AB} \triangleq |\phi_m\rangle^A \otimes |\phi_n\rangle^B \in \mathcal{H}_{AB}$, viz.,

$$|\psi\rangle^{AB} = \sum_{mn} c_{mn}|\phi_{mn}\rangle^{AB}, \quad \text{with } \sum_{mn} |c_{mn}|^2 = 1, \quad \text{and} \tag{B.2}$$

$$\hat{\rho}^{AB} = \sum_{k} p_k |\psi_k\rangle^{AB\,AB}\langle\psi_k|, \quad \text{with } p_k \geq 0, \quad \sum_k p_k = 1, \tag{B.3}$$

for pure states $|\psi_k\rangle^{AB} \in \mathcal{H}_{AB}$.

A pure state $|\psi\rangle^{AB} \in \mathcal{H}_{AB}$ of a composite system AB can be classified into:

1. A product state: when $|\psi\rangle^{AB}$ can be decomposed into a tensor product of two pure states in A and B, i.e., $|\psi\rangle^{AB} = |\psi\rangle^A \otimes |\psi\rangle^B$.

2. An entangled state: when $|\psi\rangle^{AB}$ cannot be expressed as a tensor product of two pure states in A and B (for instance, the state $(|0\rangle|0\rangle + |1\rangle|1\rangle)/\sqrt{2}$ is a pure entangled state of a two-qubit system).[3]

A mixed state $\hat{\rho}^{AB} \in B(\mathcal{H}_{AB})$ of a composite system[4] AB can be classified into:

1. A product state: when $\hat{\rho}^{AB}$ can be decomposed into a tensor product of two states in A and B, i.e. $\hat{\rho}^{AB} = \hat{\rho}^A \otimes \hat{\rho}^B$, with at least one of $\hat{\rho}^A$ or $\hat{\rho}^B$ being a mixed state.

2. A classically-correlated state: when $\hat{\rho}^{AB}$ is not a product state, but nevertheless can be expressed as a statistical mixture of product pure states of the systems

[3]Entanglement is inherently a quantum-mechanical property of composite physical systems and is stronger than any probabilistic correlation between the constituent systems that classical physics permits. The individual states of the systems A and B, when their joint state $|\psi\rangle^{AB}$ is pure and entangled, are mixed states, which are obtained by taking a partial trace over the other system, i.e., $\hat{\rho}^A = \mathrm{Tr}_B\{\hat{\rho}^{AB}\} = \mathrm{Tr}_B\{|\psi\rangle^{AB\,AB}\langle\psi|\} \equiv \sum_n {}^B\langle\phi_n|\hat{\rho}^{AB}|\phi_n\rangle^B$, and vice versa.

[4]$B(\mathcal{H})$ is the set of all bounded operators in $\mathcal{H}$.

A and B, i.e. $\hat{\rho}^{AB} = \sum_k p_k (|\alpha_k\rangle^A \otimes |\beta_k\rangle^B)({}^A\langle\alpha_k| \otimes {}^B\langle\beta_k|)$ for any set of pure states $|\alpha_k\rangle \in \mathcal{H}_\mathcal{A}$ and $|\beta_k\rangle \in \mathcal{H}_B$, with $p_k \geq 0$ and $\sum_k p_k = 1$.

3. An entangled state: when $\hat{\rho}^{AB}$ is a mixed state of the composite system AB which is neither a product state nor a classically-correlated state, i.e., the joint state of the composite system has a correlation between the systems A and B, which is stronger than any (classical) probabilistic correlation. For instance, consider equal mixtures of the Bell states $|\alpha\rangle = (|0\rangle|0\rangle + |1\rangle|1\rangle)/\sqrt{2}$ and $|\beta\rangle = (|1\rangle|0\rangle + |0\rangle|1\rangle)/\sqrt{2}$. This is a mixed entangled state, $(|\alpha\rangle\langle\alpha| + |\beta\rangle\langle\beta|)/2$, of a two-qubit system.[5]

B.1.3 Evolution

The time evolution of a *closed system* is defined in terms of the unitary time-evolution operator $\hat{U}(t, t_0) = \exp(-i\hat{H}(t - t_0)/\hbar)$, where $\hat{H}$ is the time-independent Hamiltonian of the closed system. The evolution of the system when it is in a pure state $|\psi(t_0)\rangle$ at time t_0, and when it is in a mixed state $\hat{\rho}(t_0)$ at time t_0 are respectively given by:

$$|\psi(t)\rangle = \hat{U}(t, t_0)|\psi(t_0)\rangle, \quad \text{and} \tag{B.4}$$

$$\hat{\rho}(t) = \hat{U}(t, t_0)\hat{\rho}(t_0)\hat{U}^\dagger(t, t_0). \tag{B.5}$$

The time evolution of a general *open system*, i.e., a system that interacts with an environment, is not unitary. While the joint state of the system and the environment is a closed system and hence follows a unitary evolution,[6] the evolution of the state of

[5] We reiterate that if a mixed state $\hat{\rho}^{AB}$ is not decomposable into a tensor product of mixed states, i.e. $\hat{\rho}^{AB} \neq \hat{\rho}^A \otimes \hat{\rho}^B$, the joint state $\hat{\rho}^{AB}$ is NOT necessarily entangled, and it could just have classical correlations between the two constituent systems.

[6] We use a unitary transformation describing the beamsplitter model of the optical channel to prove Lemma 6.1 in Appendix C.1. However, in that particular case, the derivation of the expression

the system alone is non-unitary and is represented by a *trace-preserving, completely-positive* (TPCP) map. A TPCP map $\mathcal{E}$ takes density operator $\hat{\rho}_{\text{in}} \in B(\mathcal{H}_{\text{in}})$ to density operator $\hat{\rho}_{\text{out}} \in B(\mathcal{H}_{\text{out}})$, and satisfies the following properties:

1. $\mathcal{E}$ preserves the trace, i.e., $\text{Tr}\{\mathcal{E}(\hat{\rho})\} = 1$ for any $\hat{\rho}_{\text{in}} \in B(\mathcal{H}_{\text{in}})$.

2. $\mathcal{E}$ is a convex linear map on the set of density operators $\hat{\rho}_{\text{in}} \in B(\mathcal{H}_{\text{in}})$, i.e. $\mathcal{E}(\sum_k p_k \hat{\rho}_k) = \sum_k p_k \mathcal{E}(\hat{\rho}_k)$ for any probability distribution $\{p_k\}$.

3. $\mathcal{E}$ is a completely positive map. This means that $\mathcal{E}$ maps positive operators in $B(\mathcal{H}_{\text{in}})$ to positive operators on $B(\mathcal{H}_{\text{out}})$, and, for any reference system R and for any positive operator $\hat{\rho} \in B(\mathcal{H}_{\text{in}} \otimes R)$, we have that $(\mathcal{E} \otimes I_R)\hat{\rho} \geq 0$, where $\hat{I}_R$ is the identity operator on R.

It can be shown that any TPCP map can be expressed in an *operator sum representation* [69], $\mathcal{E}(\hat{\rho}) = \sum_k \hat{A}_k \hat{\rho} \hat{A}_k^\dagger$, where the Kraus operators A_k must satisfy $\sum_k \hat{A}_k^\dagger \hat{A}_k = \hat{I}$ in order to preserve the trace of $\mathcal{E}(\hat{\rho})$.

B.1.4 Observables and Measurement

In quantum mechanics, each dynamical observable (for instance position, momentum, energy, angular momentum, etc.) is represented by a Hermitian operator $\hat{M}$. Being a Hermitian operator, $\hat{M}$ must have a complete orthonormal set of eigenvectors $\{|\phi_m\rangle\}$ with associated real eigenvalues ϕ_m that satisfy $\hat{M}|\phi_m\rangle = \phi_m|\phi_m\rangle$. The outcome of a measurement of $\hat{M}$ on a quantum state $\hat{\rho}$ always leads to an eigenvalue ϕ_n with probability $p(n) = \langle\phi_n|\hat{\rho}|\phi_n\rangle$. Given that the measurement result obtained is ϕ_n, the post-measurement state of the system is the eigenstate $|\phi_n\rangle$ corresponding to the eigenvalue ϕ_n. This phenomenon is known as the "collapse" of the wave function. Thus, if the system is in an eigenstate of a measurement operator $\hat{M}$ to begin with,

---

for the output state of the beamsplitter is relatively simple since the environment port is in a vacuum state; the derivation is substantially more complicated for a non-vacuum environment.

the measurement result is known with certainty and the measurement of $\hat{M}$ does not alter the state of the system. The Hermitian operator $\hat{H}$ corresponding to measuring the total energy of a closed quantum system is known as the Hamiltonian for the system. The measurement of an observable as described above is also known as a *projective measurement*, as the measurement projects the state onto an eigenspace of the measurement operator.

In analogy to the evolution of an open system described above, a more general measurement on a system entails a projective measurement performed on the joint state of the system in question along with an auxiliary environment prepared in some initial state. This general measurement scheme can be described by a set of positive semi-definite operators $\left\{\hat{\Pi}_m\right\}$ that satisfy $\sum_m \hat{\Pi}_m = \hat{I}$. If a measurement is performed on a quantum state $\hat{\rho}$, the outcome of the measurement is n with probability $p(n) = \mathrm{Tr}\{\hat{\rho}\hat{\Pi}_n\}$. The above description of a quantum measurement is known as the *positive operator-valued measure* (POVM) formalism and the operators $\left\{\hat{\Pi}_m\right\}$ are known as POVM operators. The POVM operators by themselves do not determine a post-measurement state. POVM formalism is crucial to quantum hypothesis testing which is why we use it extensively in Chapter 6.

B.2 Trace Distance and Quantum Binary Hypothesis Testing

Willie has to perform a quantum binary hypothesis test to determine whether Alice is transmitting. Here we develop the *trace distance* lower bound on the probability of error in discriminating between two quantum states $\hat{\rho}_0$ and $\hat{\rho}_1$, which we use in the proof of Theorem 6.3.1. We prove the quantum analog of Lemma A.1, thus arguing that the assumption of a non-trivial prior distribution on Alice's transmission state does not impact our asymptotic results.

First, the trace distance between two quantum states is defined as follows:

Definition B.1 (Trace distance [88])**.** *The* trace distance *between two density operators $\hat{\sigma}$ and $\hat{\rho}$ is*

$$\|\hat{\sigma} - \hat{\rho}\|_1 = \mathrm{Tr}\{\sqrt{(\hat{\sigma} - \hat{\rho})^\dagger (\hat{\sigma} - \hat{\rho})}\}. \tag{B.6}$$

Trace distance relates to the probability of successful discrimination between two quantum states via the following lemma:

Lemma B.1. *One half of the trace distance $\frac{1}{2}\|\hat{\rho} - \hat{\sigma}\|_1$ between quantum states $\hat{\rho}$ and $\hat{\sigma}$ is equal to the largest probability difference that two states $\hat{\rho}$ and $\hat{\sigma}$ could give to the outcome of the same measurement given by the positive semi-definite operator $\hat{\Lambda}$ with eigenvalues upper-bounded by one:*

$$\frac{1}{2}\|\hat{\rho} - \hat{\sigma}\|_1 = \max_{0 \leq \hat{\Lambda} \leq \hat{I}} \mathrm{Tr}\left\{\Lambda\left(\hat{\rho} - \hat{\sigma}\right)\right\}. \tag{B.7}$$

Proof. See [88, Lemma 9.1.1]. □

Denote by $\hat{\rho}_0$ and $\hat{\rho}_1$ the respective quantum states that Willie observes on his channel from Alice when she does not transmit and transmits. Willie constructs a binary POVM $\{\hat{\Lambda}_0, \hat{\Lambda}_1\}$ to discriminate between these states. Suppose that Willie knows that Alice will not transmit (i.e., H_0 is true) with probability π_0 and that she will transmit (i.e., H_1 is true) with probability $\pi_1 = 1 - \pi_0$. Thus, π_0 and π_1 denote the prior probabilities of states $\hat{\rho}_0$ and $\hat{\rho}_1$, respectively, and the probability of error in discriminating between $\hat{\rho}_0$ and $\hat{\rho}_1$ is

$$\mathbb{P}_e^{(w)} = \pi_0 \,\mathrm{Tr}\{\hat{\Lambda}_1 \hat{\rho}_0\} + \pi_1 \,\mathrm{Tr}\{\hat{\Lambda}_0 \hat{\rho}_1\}. \tag{B.8}$$

The following lemma generalizes the result for $\pi_0 = \pi_1 = \frac{1}{2}$ given in [88, Section 9.1.4] to $\pi_0 \neq \pi_1$:

Lemma B.2 (Quantum version of Lemma A.1)**.**

$$\mathbb{P}_e^{(w)} \geq \min(\pi_0, \pi_1) - \frac{\max(\pi_0, \pi_1)}{2} \|\hat{\rho}_0 - \hat{\rho}_1\|_1. \tag{B.9}$$

Proof. First, suppose that $\pi_0 \leq \pi_1$. Since $\{\hat{\Lambda}_0, \hat{\Lambda}_1\}$ is a POVM, $\hat{\Lambda}_0 + \hat{\Lambda}_1 = \hat{I}$. Substituting $\hat{\Lambda}_1 = \hat{I} - \hat{\Lambda}_0$ in (B.8) and re-arranging the terms, we obtain

$$\mathbb{P}_e^{(w)} = \pi_0 \operatorname{Tr}\{\hat{\rho}_0\} - \pi_0 \operatorname{Tr}\{\hat{\Lambda}_0 \hat{\rho}_0\} + \pi_1 \operatorname{Tr}\{\hat{\Lambda}_0 \hat{\rho}_1\} \tag{B.10}$$

$$= \pi_0 - \operatorname{Tr}\{\hat{\Lambda}_0(\pi_0 \hat{\rho}_0 - \pi_1 \hat{\rho}_1)\}, \tag{B.11}$$

where (B.11) is because the eigenvalues of a density operator sum to one. When the prior probabilities of the hypotheses are equal: $\pi_0 = \pi_1 = \frac{1}{2}$, an application of Lemma B.1 yields the lower bound $\mathbb{P}_e^{(w)} \geq \frac{1}{2} - \frac{1}{4}\|\hat{\rho}_0 - \hat{\rho}_1\|_1$. Now,

$$\mathbb{P}_e^{(w)} = \pi_0 - \operatorname{Tr}\{\hat{\Lambda}_0(\pi_0 \hat{\rho}_0 - \pi_1 \hat{\rho}_1 + \pi_1 \hat{\rho}_0 - \pi_1 \hat{\rho}_0)\} \tag{B.12}$$

$$= \pi_0 - \pi_1 \operatorname{Tr}\{\hat{\Lambda}_0(\hat{\rho}_0 - \hat{\rho}_1)\} + (\pi_1 - \pi_0) \operatorname{Tr}\{\hat{\Lambda}_0 \hat{\rho}_0\} \tag{B.13}$$

$$\geq \pi_0 - \pi_1 \operatorname{Tr}\{\hat{\Lambda}_0(\hat{\rho}_0 - \hat{\rho}_1)\} \tag{B.14}$$

$$\geq \pi_0 - \frac{\pi_1}{2}\|\hat{\rho}_0 - \hat{\rho}_1\|_1, \tag{B.15}$$

where (B.14) follows since $\pi_0 \leq \pi_1$ and $\operatorname{Tr}\{\hat{\Lambda}_0 \hat{\rho}_0\} \geq 0$, and (B.15) follows by Lemma B.1.

When $\pi_1 \leq \pi_0$, the same steps are used with substitution of $\hat{\Lambda}_0 = \hat{I} - \hat{\Lambda}_1$ in (B.8) instead of $\hat{\Lambda}_1 = \hat{I} - \hat{\Lambda}_0$, and replacement of $\pi_1 \hat{\rho}_0 - \pi_1 \hat{\rho}_0$ with $\pi_0 \hat{\rho}_1 - \pi_0 \hat{\rho}_1$ inside the trace operator in (B.12). This yields

$$\mathbb{P}_e^{(w)} \geq \pi_1 - \frac{\pi_0}{2}\|\hat{\rho}_0 - \hat{\rho}_1\|_1, \tag{B.16}$$

and the lemma. □

Thus, while Lemma B.2 demonstrates that additional information about the likelihood of Alice transmitting (in the form of unequal prior probabilities $\pi_0 \neq \pi_1$) helps Willie, the square root law still holds via the bounds on the trace distance $\|\hat{\rho}_0 - \hat{\rho}_1\|_1$.

B.3 Quantum Entropy and Information Measures[7]

Amongst various measures of how *mixed* is a quantum state $\hat{\rho}$, the most relevant one information-theoretically is the *von Neumann entropy* $S(\hat{\rho})$, which is defined as

$$S(\hat{\rho}) = -\operatorname{Tr}\{\hat{\rho}\ln\hat{\rho}\} \tag{B.17}$$

$$= H(\{\lambda_n\}), \tag{B.18}$$

where $H(\{\lambda_n\}) \equiv -\sum_n \lambda_n \ln \lambda_n$ is the Shannon entropy of the eigenvalues λ_n of $\hat{\rho}$. Hence, it is obvious that the von Neumann entropy of a pure state is zero, i.e., $S(|\psi\rangle\langle\psi|) = 0$. Most of quantum information theory is built around the von Neumann entropy measure of a quantum state. We now review a few important properties of von Neumann entropy.

B.3.1 Data Compression

In analogy with the role that Shannon entropy plays in classical information theory, it can be shown that $S(\hat{\rho}^A)$ is the optimal compression rate on the quantum system A in the state $\hat{\rho}^A \in B(\mathcal{H}_A)$. In other words, for large n, the density matrix $\hat{\rho}^{A\otimes n}$ has nearly all of its support on a subspace of $\mathcal{H}_A^{\otimes n}$ (called the *typical subspace*) of dimension $2^{nS(\hat{\rho}^A)}$. We will henceforth use the notation $S(A)$ interchangeably with $S(\hat{\rho}^A)$ to mean von Neumann entropy of the system A (or the von Neumann entropy of the state $\hat{\rho}^A$). If A is a classical random variable, we use the function $H(A)$ to denote the Shannon entropy of A.

---

[7]The content of this section was adapted from [38, Appendix A.2] with permission of the author.

B.3.2 Subadditivity

The joint entropy $S(A, B)$ of a bipartite system AB is always upper bounded by the sum of the entropies of the individual systems A and B, i.e.,

$$S(A, B) \leq S(A) + S(B), \tag{B.19}$$

with equality when the joint state of AB is a product state, i.e., $\hat{\rho}^{AB} = \hat{\rho}^{A} \otimes \hat{\rho}^{B}$. We use the subadditivity of von Neumann entropy in the proof of Theorem 6.5.1.

B.3.3 Joint and Conditional Entropy

The entropy of a bipartite system AB in a joint state $\hat{\rho}^{AB}$ is defined as $S(A, B) = -\operatorname{Tr}\{\hat{\rho}^{AB} \ln \hat{\rho}^{AB}\}$. Even though there is no direct definition of quantum conditional entropy as in classical information theory, one may define a conditional entropy (in analogy to its classical counterpart) as $S(A|B) = S(A, B) - S(B)$. The quantum conditional entropy can be negative, contrary to its classical counterpart.[8] However, like its classical counterpart, conditioning can only reduce entropy, i.e., $S(A|B, C) \leq S(A|B)$. Discarding a quantum system can never increase quantum mutual information (see Section B.3.5, i.e., $I(A; B) \leq I(A; B, C)$.

B.3.4 Classical-quantum States

Here we define the notion of classical-quantum states and classical-quantum channels. We associate any classical set $\mathcal{X}$ with a Hilbert space $\mathcal{H}_X$ having orthonormal basis $\left\{|x\rangle^X\right\}_{x\in\mathcal{X}}$. Thus, for any classical random variable X which takes the values $x \in \mathcal{X}$ with probability $p(x)$, we can write a density matrix

---

[8] For the bipartite two-qubit Bell state $|\psi\rangle^{AB} = (|00\rangle + |11\rangle)/\sqrt{2}$, $S(A|B) = S(A, B) - S(B) = 0 - 1 = -1$. The joint state of the system AB is a pure state, hence $S(A, B) = 0$, whereas the state of system B, $\hat{\rho}^{B} = \operatorname{Tr}_A\{\hat{\rho}^{AB}\} = (|0\rangle\langle 0| + |1\rangle\langle 1|)/2$ is a mixed state with entropy $S(B) = 1$.

$$\hat{\rho}^X = \sum_x p(x)|x\rangle\langle x|^X \equiv \bigoplus_x p(x) \tag{B.20}$$

which is diagonal in the orthonormal basis $\left\{|x\rangle^X\right\}_{x\in\mathcal{X}}$. Similarly, an ensemble of quantum states $\left\{\hat{\rho}_x^B, p(x)\right\}$ can be associated with a block diagonal *classical-quantum (cq)* state for the system XB:

$$\hat{\rho}^{XB} = \sum_x p(x)|x\rangle\langle x|^X \otimes \hat{\rho}_x^B \equiv \bigoplus_x p(x)\hat{\rho}_x^B, \tag{B.21}$$

where X is a classical random variable and B is a quantum system with conditional density matrices $\hat{\rho}_x^B$. The conditional entropy $S(B|X)$ is then

$$S(B|X) = \sum_x p(x)S(\hat{\rho}_x^B). \tag{B.22}$$

B.3.5 Quantum Mutual Information

The quantum mutual information $I(A;B)$ of a bipartite system AB is defined analogously to Shannon mutual information as:

$$I(A;B) = S(A) + S(B) - S(A,B) \tag{B.23}$$
$$= S(A) - S(A|B) \tag{B.24}$$
$$= S(B) - S(B|A). \tag{B.25}$$

A bipartite product mixed state $\hat{\rho}^A \otimes \hat{\rho}^B$ has zero quantum mutual information. The quantum mutual information of a cq-state (B.21) is given by

$$I(X;B) = S(B) - S(B|X) \tag{B.26}$$
$$= S\left(\sum_x p(x)\hat{\rho}_x^B\right) - \sum_x p(x)S(\hat{\rho}_x^B) \tag{B.27}$$
$$\triangleq \chi\left(p(x), \hat{\rho}_x^B\right), \tag{B.28}$$

where $\chi\left(p(x), \hat{\rho}_x^B\right)$ is defined as the *Holevo information* of the ensemble of states $\left\{p(x), \hat{\rho}_x^B\right\}$.

B.3.6 The Holevo Bound

Suppose Alice chooses a classical message index $x \in \mathcal{X}$ with probability $p(x)$ and encodes x by preparing a quantum state $\hat{\rho}_x^A$. She sends her state to Bob through a channel $\mathcal{E}$, which then produces a state $\hat{\rho}_x^B = \mathcal{E}(\hat{\rho}_x^A)$ at Bob's end, conditioned on the classical index x. In order to obtain information about x, Bob measures his state $\hat{\rho}_x^B$ using a POVM $\left\{\hat{\Pi}_y\right\}$. The probability that the outcome of his POVM measurement is y given that Alice sent x is $p(y|x) = \mathrm{Tr}\{\hat{\rho}_x^B \hat{\Pi}_y\}$. Using X and Y to denote the random variables of which x and y are instances, we know from Shannon information theory that, when Bob uses the POVM $\left\{\hat{\Pi}_y\right\}$, the maximum rate at which Alice can transmit information to Bob using a suitable encoding and decoding scheme is given by the maximum of the mutual information $I(X;Y)$ over all input distributions $p(x)$. Holevo, Schumacher and Westmoreland showed [43, 40, 73] that for a given prior $p(x)$ and POVM $\left\{\hat{\Pi}_y\right\}$, the single-use Holevo information is an upper bound on Shannon mutual information,

$$I(X;Y) \leq \chi\left(p(x), \hat{\rho}_x^B\right), \tag{B.29}$$

which is known as the *Holevo bound*. Maximizing over $p(x)$ on both sides, one obtains

$$\max_{p(x)} I(X;Y) \leq \max_{p(x)} \chi\left(p(x), \mathcal{E}(\hat{\rho}_x^A)\right). \tag{B.30}$$

As the right-hand side does not depend on the choice of the POVM elements $\left\{\hat{\Pi}_y\right\}$, the inequality is preserved by a further maximization of the left hand side over the measurements,

$$\max_{p(x),\left\{\hat{\Pi}_y\right\}} I(X;Y) \leq \max_{p(x)} \chi\left(p(x), \mathcal{E}(\hat{\rho}_x^A)\right), \quad \text{or} \tag{B.31}$$

$$C_{1,1}(\mathcal{E}) \leq C_{1,\infty}(\mathcal{E}), \tag{B.32}$$

where $C_{1,1}(\mathcal{E})$ is the maximum value of the Shannon Information $I(X;Y)$ optimized over all possible symbol-by-symbol POVM measurements $\left\{\hat{\Pi}_y\right\}$. $C_{1,\infty}(\mathcal{E})$ on the other hand, is the maximum value of the Shannon Information $I(X;Y)$ optimized not only over all possible symbol-by-symbol POVM measurements, but also over arbitrary multiple-channel-use POVM measurements. As we see below, $C_{1,\infty}(\mathcal{E})$ is the capacity of the channel $\mathcal{E}$ for transmission of classical information if Alice is limited to single channel uses (i.e., symbols $\hat{\rho}_x^A$) and Bob may choose any joint measurement at the receiver.

B.3.7 Ultimate Classical Communication Capacity: the HSW Theorem

The classical capacity of a quantum channel is established by random coding arguments akin to those employed in classical information theory. A set of symbols $\{j\}$ is represented by a collection of input states $\{\hat{\rho}_j\}$ that are selected according to some prior distribution $\{p_j\}$. The output states $\{\hat{\rho}'_j\}$ are obtained by applying the channel's TPCP map $\mathcal{E}(\cdot)$ to these input symbols. According to the HSW Theorem, the capacity of this channel, in nats per use, is

$$C = \sup_n (C_{n,\infty}/n) = \sup_n \{\max_{\{p_j,\hat{\rho}_j\}} [\chi(p_j, \mathcal{E}^{\otimes n}(\hat{\rho}_j))/n]\}, \tag{B.33}$$

where $C_{n,\infty}$ is the capacity achieved when coding is performed over n-channel-use symbols and arbitrary joint-detection measurement is used at the receiver. The supremum over n is necessitated by the fact that channel capacity may be superadditive, viz., $C_{n,\infty} > nC_{1,\infty}$ is possible for quantum channels, whereas such is not the case for classical channels. The HSW Theorem tells us that Holevo information plays the role

for classical information transmission over a quantum channel that Shannon mutual information does for a classical channel.

Neither (B.28) nor (B.33) have any explicit dependence on the quantum measurement used at the receiver, thus, measurement optimization is implicit within the HSW Theorem. To obtain the same capacity C by maximizing Shannon mutual information we can introduce a positive-operator-valued measure (POVM) [69], representing the multi-symbol quantum measurement (a joint measurement over an entire codeword) performed at the receiver. For example, if single-use encoding is performed with priors $\{p_j\}$, the probability of receiving a particular m-symbol codeword, $\mathbf{k} \equiv (k_1, k_1, \ldots, k_m)$, given that $\mathbf{j} \equiv (j_1, j_2, \ldots, j_m)$ was sent is

$$\Pr(\,\mathbf{k}\,|\,\mathbf{j}\,) \equiv \mathrm{Tr}\left\{ \hat{\Pi}_{\mathbf{k}} \left[\bigotimes_{l=1}^{m} \mathcal{E}(\hat{\rho}_{j_l}) \right] \right\}, \tag{B.34}$$

where the POVM, $\{\hat{\Pi}_{\mathbf{k}}\}$, is a set of Hermitian operators on the Hilbert space of output states for m channel uses that resolve the identity (i.e., $\sum_{\mathbf{k}} \hat{\Pi}_{\mathbf{k}} = \hat{I}$). From $\{\, p_j, \Pr(\,\mathbf{k}\,\mid\,\mathbf{j}\,)\}$ we can then write down Shannon mutual information for single-use encoding and m-symbol codewords that must be maximized. Ultimately, by allowing for n-channel-use symbols and optimizing over the priors, the signal states, *and* the POVM, we would arrive at the capacity predicted by the HSW Theorem. Evidently, determining capacity is easier via the HSW Theorem than it is via Shannon mutual information, because one less optimization is required. However, finding a practical system that can approach capacity requires paying attention to the receiver measurement.

B.4 Basic Description of Optical Modes

An optical mode is an optical field function that couples an input at the transmitter with an output at the receiver. To define the number of optical modes n

available to Alice and Bob more formally, consider a free-space L meter line-of-sight optical channel with areas of transmitter and receiver apertures A_t and A_r, respectively. Let's assume quasi-monochromatic transmission at center-wavelength λ. The free-space Fresnel number product of this channel is $D_f = \frac{A_t A_r}{(\lambda L)^2}$. When $D_f \ll 1$, the channel is in the *far-field* regime, and only one spatial mode—the spatial optical field function at the transmitter aperture—can couple any appreciable fraction of power into the receiver aperture. In this case, the transmitter-to-receiver power transmissivity $\eta \approx D_f$. On the other hand, when $D_f \gg 1$, the channel is in the *near-field* regime, and there are approximately D_f mutually-orthogonal transmitter-receiver spatial modes, with each having near-unit receiver-to-transmitter power transmissivity. These spatial modes are analogous to parallel channels.

Suppose the transmitter employs M orthogonal spatial modes with transmissivities $\eta_1, \eta_2, \ldots, \eta_M$. Now consider a time-bandwidth product $K = WT$, where T is the length of the total transmission window (in seconds), and W is the total frequency bandwidth (in Hz and determined by bandwidths of the transmitter and the receiver). Thus, there are K mutually-orthogonal temporal modes that can be accommodated within that time-bandwidth product. A burst of communication that uses K temporal modes on each of M spatial modes transmits using $n = MK = MWT$ spatio-temporal modes. Furthermore, both orthogonal polarizations of light can be used to increase the total to $n = 2MWT$ spatio-temporal-polarization modes.

Typically, single-mode fiber is used in commercial applications. A single-mode L meter fiber link supports communication using a single spatial mode ($M = 1$) with power transmissivity $\eta = e^{-\alpha L}$, where α denotes the fiber's loss coefficient. In that case $n = 2WT$, assuming the fiber can transmit both polarizations.

APPENDIX C

QUANTUM COVERT COMMUNICATION MISCELLANEA

C.1 Proof of Lemma 6.1

A beamsplitter can be described as a unitary transformation U_{BS} from the two input modes (Alice's and the environment's ports) to the two output modes (Bob's and Willie's ports). Given a Fock state input $|t\rangle^A$ on Alice's port and vacuum input $|0\rangle^E$ on the environment's port, the output at Bob's and Willie's ports is described as follows [13, Section IV.D]:

$$U_{BS}\,|t\rangle^A\,|0\rangle^E = \sum_{m=0}^{t} \sqrt{\binom{t}{m}\eta_w^m(1-\eta_w)^{t-m}}\,|m\rangle^W\,|t-m\rangle^B\,.$$

Thus,

$$U_{BS}^{\otimes n}\,|\mathbf{t}\rangle^{A^n}\,|\mathbf{0}\rangle^{E^n} = \bigotimes_{i=1}^{n}\sum_{m_i=0}^{t_i}\sqrt{\binom{t_i}{m_i}\eta_w^{m_i}(1-\eta_w)^{t_i-m_i}}\,|m_i\rangle^{W_i}\,|t_i-m_i\rangle^{B_i}\,,$$

which implies

$$\begin{aligned} U_{BS}^{\otimes n}\,|\psi\rangle^{A^n}\,|\mathbf{0}\rangle^{E^n} &= \sum_{\mathbf{t}\in\mathbb{N}_0^n} a_{\mathbf{t}} \bigotimes_{i=1}^{n}\sum_{m_i=0}^{t_i}\sqrt{\binom{t_i}{m_i}\eta_w^{m_i}(1-\eta_w)^{t_i-m_i}}\,|m_i\rangle^{W_i}\,|t_i-m_i\rangle^{B_i} \\ &\equiv |\phi\rangle^{W^nB^n}\,. \end{aligned}$$

Now, the partial trace of the output state $\rho^{BW} = |\phi\rangle^{W^n B^n}$ over Bob's system reveals Willie's output state:

$$\begin{aligned}\rho^{W^n} &= \mathrm{Tr}_{B^n}\left[|\phi\rangle^{W^n B^n\ W^n B^n}\langle\phi|\right] \\ &= \sum_{\mathbf{x}\in\mathbb{N}_0^n} \left|{}^{B^n}\langle\mathbf{x}|\phi\rangle^{W^n B^n}\right|^2,\end{aligned}$$

with

$$\begin{aligned}{}^{B^n}\langle\mathbf{x}|\phi\rangle^{W^n B^n} &= \sum_{\mathbf{t}\in\mathbb{N}_0^n} a_{\mathbf{t}} \bigotimes_{i=1}^{n} \sum_{m_i=0}^{t_i} \sqrt{\binom{t_i}{m_i}\eta_w^{m_i}(1-\eta_w)^{t_i-m_i}}\,|m_i\rangle^{W_i\ B_i}\langle x_i|t_i-m_i\rangle^{B_i} \\ &= \sum_{\mathbf{t}\in\mathbb{N}_0^n} a_{\mathbf{t}} \bigotimes_{i=1}^{n} \sqrt{\binom{t_i}{x_i}\eta_w^{t_i-x_i}(1-\eta_w)^{x_i}}\,|t_i-x_i\rangle^{W_i}, \qquad \text{(C.1)}\end{aligned}$$

where equation (C.1) is because the Fock states are orthogonal. Thus,

$${}^{W^n}\langle\mathbf{s}|\,\hat{\rho}^{W^n}\,|\mathbf{s}\rangle^{W^n} = \sum_{\mathbf{x}\in\mathbb{N}_0^n} \left|{}^{W^n}\langle\mathbf{s}|\,{}^{B^n}\langle\mathbf{x}|\phi\rangle^{W^n B^n}\right|^2, \qquad \text{(C.2)}$$

where

$$\begin{aligned}{}^{W^n}\langle\mathbf{s}|\,{}^{B^n}\langle\mathbf{x}|\phi\rangle^{W^n B^n} &= \sum_{\mathbf{t}\in\mathbb{N}_0^n} a_{\mathbf{t}} \prod_{i=1}^{n} \sqrt{\binom{t_i}{x_i}\eta_w^{t_i-x_i}(1-\eta_w)^{x_i}}\,\delta_{s_i,t_i-x_i} \\ &= a_{\mathbf{x}+\mathbf{s}} \prod_{i=1}^{n} \sqrt{\binom{x_i+s_i}{x_i}\eta_w^{s_i}(1-\eta_w)^{x_i}}, \qquad \text{(C.3)}\end{aligned}$$

with $\delta_{a,b} = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{otherwise} \end{cases}$. Substituting $\mathbf{k} = \mathbf{x}+\mathbf{s}$ into equation (C.3) and substituting the right-hand side (RHS) of (C.3) into equation (C.2) yields

$$\begin{aligned}{}^{W^n}\langle\mathbf{s}|\,\hat{\rho}^{W^n}\,|\mathbf{s}\rangle^{W^n} &= \sum_{\mathbf{k}\in\mathbb{N}_0^n} \left| a_{\mathbf{k}} \prod_{i=1}^{n} \sqrt{\binom{k_i}{s_i}\eta_w^{s_i}(1-\eta_w)^{k_i-s_i}} \right|^2 \\ &= \sum_{\mathbf{k}\in\mathbb{N}_0^n} |a_{\mathbf{k}}|^2 \prod_{i=1}^{n} \binom{k_i}{s_i}\eta_w^{s_i}(1-\eta_w)^{k_i-s_i} \qquad \text{(C.4)}\end{aligned}$$

where equation (C.4) is because $\eta_w \in [0, 1)$. $\square$

C.2 Proof of Lemma 6.2

Express $D(\hat{\rho}_0 \| \hat{\rho}_1) = -\operatorname{Tr}\{\hat{\rho}_0 \ln \hat{\rho}_1\} - S(\hat{\rho}_0)$, where $S(\hat{\rho}_0) \equiv -\operatorname{Tr}[\hat{\rho}_0 \ln \hat{\rho}_0]$ is the von Neumann entropy of the state $\hat{\rho}_0$:

$$S(\hat{\rho}_0) = \ln(1 + \bar{n}_0) + \bar{n}_0 \ln\left(1 + \frac{1}{\bar{n}_0}\right). \tag{C.5}$$

Now,

$$\begin{aligned}
\operatorname{Tr}[\hat{\rho}_0 \ln \hat{\rho}_1] &= \operatorname{Tr}\left[\left(\sum_{n=0}^{\infty} \frac{\bar{n}_0^n}{(1+\bar{n}_0)^{1+n}} |n\rangle \langle n|\right)\left(\sum_{n=0}^{\infty} \ln \frac{\bar{n}_1^n}{(1+\bar{n}_1)^{1+n}} |n\rangle \langle n|\right)\right] \\
&= \sum_{n=0}^{\infty} \frac{\bar{n}_0^n}{(1+\bar{n}_0)^{1+n}} \ln \frac{\bar{n}_1^n}{(1+\bar{n}_1)^{1+n}} \\
&= \frac{1}{1+\bar{n}_0} \ln \frac{1}{1+\bar{n}_1} \sum_{n=0}^{\infty} \left(\frac{\bar{n}_0}{1+\bar{n}_0}\right)^n + \ln \frac{\bar{n}_1}{1+\bar{n}_1} \sum_{n=0}^{\infty} \frac{n}{1+\bar{n}_0} \cdot \left(\frac{\bar{n}_0}{1+\bar{n}_0}\right)^n \\
&= \ln \frac{1}{1+\bar{n}_1} + \bar{n}_0 \ln \frac{\bar{n}_1}{1+\bar{n}_1}, \qquad \text{(C.6)}
\end{aligned}$$

where (C.6) is because the geometric series

$$\sum_{n=0}^{\infty} \left(\frac{\bar{n}_0}{1+\bar{n}_0}\right)^n = \left(1 - \frac{\bar{n}_0}{1+\bar{n}_0}\right)^{-1}$$

and

$$\sum_{n=0}^{\infty} \frac{n}{1+\bar{n}_0} \left(\frac{\bar{n}_0}{1+\bar{n}_0}\right)^n = \bar{n}_0$$

is the expression for the mean of geometrically-distributed random variable $X \sim \text{Geom}\left(\frac{1}{1+\bar{n}_0}\right)$. Combining (C.5) and (C.6) yields the lemma. $\square$

APPENDIX D

EXPERIMENTAL MISCELLANEA

D.1 Calculation of Bob's Maximum Throughput

The Q-ary PPM signaling combined with Bob's device for assigning symbols to received PPM frames induces a discrete memoryless channel described by a conditional distribution $\mathbb{P}(Y|X)$, where $X \in \{1, \ldots, Q\}$ is Alice's input symbol and $Y \in \{1, \ldots, Q, \mathcal{E}\}$ is Bob's output symbol with $\mathcal{E}$ indicating an erasure. Since Bob observes Alice's pulse with probability $1 - e^{-\bar{n}_{det}^{(b)}}$, $\mathbb{P}(Y|X)$ is characterized as follows:

$$\begin{aligned}
\mathbb{P}(Y = x|X = x) &= \left(1 - e^{-\bar{n}_{det}^{(b)}}\right) \sum_{i=0}^{Q-1} \frac{1}{i+1} \left(p_D^{(b)}\right)^i \left(1 - p_D^{(b)}\right)^{Q-1-i} \\
&\quad + e^{-\bar{n}_{det}^{(b)}} \sum_{i=1}^{Q} \frac{1}{i} \left(p_D^{(b)}\right)^i \left(1 - p_D^{(b)}\right)^{Q-i} \\
\mathbb{P}(Y = \mathcal{E}|X = x) &= e^{-\bar{n}_{det}^{(b)}} \left(1 - p_D^{(b)}\right)^Q \\
\mathbb{P}(Y = y, y \notin \{x, \mathcal{E}\}|X = x) &= \frac{1 - \mathbb{P}(Y = x|x) - \mathbb{P}(Y = \mathcal{E}|x)}{Q - 1}
\end{aligned}$$

The symmetry of this channel allows straightforward computation of its Shannon capacity [75] $C_s = I(X;Y)$, where $\mathbb{P}(X = x) = \frac{1}{Q}$ for $x = 1, \ldots, Q$ and $I(X;Y)$ is the mutual information between X and Y. We use the estimates from Table 7.1 to compute C_s for each regime, and plot $\frac{C_s \zeta n}{Q}$ in Figure 7.2 since $\frac{\zeta n}{Q}$ is the expected number of PPM frames selected for transmission.

D.2 Mathematical Details of Willie's Hypothesis Test

Willie collects the *click record*: a binary sequence $\mathbf{x}_w = \left\{x_j^{(w)}\right\}_{j=1}^{n}$, $x_j^{(w)} \in \{0,1\}$ of his single photon detector's (SPD's) observations of the channel from Alice, where "0" and "1" indicate the absence and the presence of a click, respectively. The hypothesis test on Alice's transmission state is between two point hypotheses, as Alice is either not transmitting (H_0) or transmitting (H_1). Thus, the log-likelihood ratio test minimizes Willie's detection error probability $\mathbb{P}_e^{(w)}$ [59, Theorem 13.1.1]. Here we derive Willie's test statistic.

The log-likelihood ratio is $L = \ln \frac{f_1(\mathbf{x}_w)}{f_0(\mathbf{x}_w)}$, where $f_0(\mathbf{x}_w)$ and $f_1(\mathbf{x}_w)$ are the likelihood functions of the click record $\mathbf{x}_w$ corresponding respectively to Alice being quiet and transmitting. When Alice does not transmit, Willie only observes dark clicks. Suppose that the dead time of Willie's SPD $t_d = 0$. Then Willie's detector readings are independent, and $\mathbf{x}_w$ is a vector of i.i.d. Bernoulli $\left(p_D^{(w)}\right)$ random variables. The likelihood function of $\mathbf{x}_w$ under H_0 is then:

$$f_0(\mathbf{x}_w) = \left(p_D^{(w)}\right)^{\sum_{j=1}^{n} x_j^{(w)}} \left(1 - p_D^{(w)}\right)^{\sum_{j=1}^{n} 1 - x_j^{(w)}} \tag{D.1}$$

$$= \prod_{i=1}^{n/Q} \left(p_D^{(w)}\right)^{\sum_{j=1}^{Q} x_{l_{ij}}^{(w)}} \left(1 - p_D^{(w)}\right)^{\sum_{j=1}^{Q} 1 - x_{l_{ij}}^{(w)}}, \tag{D.2}$$

where $l_{ij} = (i-1)Q + j$ and in (D.2) we evaluate each PPM symbol separately, as that would be convenient later.

When the dead time of the detectors $t_d > 0$, as is the case for our SPDs, the observations are not completely independent. Since a click is always followed by t_d observations without any clicks, each occurrence of "1" in $\mathbf{x}_w$ is followed by t_d occurrences of "0" with probability one. Thus, the likelihood function given in (D.2) has to be adjusted as follows:

$$f_0(\mathbf{x}_w) = \prod_{i=1}^{n/Q} \left(p_D^{(w)}\right)^{\sum_{j=1}^{Q} x_{l_{ij}}^{(w)}} \left(1 - p_D^{(w)}\right)^{\sum_{j=1}^{Q} \overline{\overline{x}}_{l_{ij}}^{(w)}}, \tag{D.3}$$

where $\bar{\bar{x}}_{l_{ij}}^{(w)} = (1 - x_{l_{ij}}^{(w)})(1 - \sum_{k=1}^{t_d} x_{l_{ij}-k})$, and the sum $\sum_{k=1}^{t_d} x_{l_{ij}-k} \in \{0, 1\}$ for $i = 1, \ldots, n/Q$ and $j = 1, \ldots, Q$, since the clicks are at least t_d observations apart. We note that, adopting the convention $\sum_{i=1}^{0} f(i) = 0$, equations (D.2) and (D.3) are identical when $t_d = 0$.

Now consider the scenario when Alice transmits. Again, we first derive the likelihood function assuming the dead time $t_d = 0$, and then adjust for $t_d > 0$. The secret shared between Alice and Bob identifies the random subset $\mathcal{S}$ of the PPM frames used for transmission, and a random vector $\mathbf{k}$ which is modulo-added to the codeword. Modulo addition of $\mathbf{k}$ effectively selects a random pulse location within each PPM frame. Note that, while both the construction in the proof of Theorem 7.1.1 and Alice's encoder described in Section 7.2.1.1 generate $\mathcal{S}$ first and then $\mathbf{k}$, the order of these operations can be reversed: we can first fix a random location of a pulse in each of n/Q PPM frames, and then select a random subset of these frames. Consider Willie's observation of the i^{th} PPM frame, and assume that the m^{th} mode is used if the frame is selected for transmission. Denote the probability of Willie's detector observing Alice's pulse by $p_r^{(w)} = 1 - e^{-\bar{n}_{det}^{(w)}}$. By construction, frames are selected for transmission independent of each other with probability ζ. Willie's detector registers a click on the m^{th} mode of the i^{th} PPM frame when one of the following disjoint events occurs:

- The i^{th} PPM frame is selected and pulse is detected by Willie in the m^{th} mode of this frame (probability $\zeta p_r^{(w)}$);
- The i^{th} PPM frame is selected, but Willie, instead of detecting the pulse, records a dark click in the m^{th} mode of this frame (probability $\zeta \left(1 - p_r^{(w)}\right) p_D^{(w)}$); and,
- Even though the i^{th} PPM frame is not selected, Willie records a dark click in the m^{th} mode of this frame (probability $(1 - \zeta) p_D^{(w)}$).

The probability of the union of these events is

$$p_s^{(w)} = \zeta p_r^{(w)} \left(1 - p_D^{(w)}\right) + p_D^{(w)}. \tag{D.4}$$

Therefore, assuming detector dead time $t_d = 0$, Willie observes an independent Bernoulli $\left(p_s^{(w)}\right)$ random variable in the m^{th} mode of the i^{th} PPM frame. Since Alice only uses the m^{th} mode for transmission, in modes other than the m^{th}, Willie observes a set of $Q-1$ i.i.d. Bernoulli $\left(p_D^{(w)}\right)$ random variables corresponding to dark clicks, again, assuming $t_d = 0$. Thus, the likelihood function of $\mathbf{x}_w$ under H_1 when $t_d = 0$ is

$$f_1(\mathbf{x}_w) = \prod_{i=1}^{n/Q} \frac{1}{Q} \sum_{m=1}^{Q} \left(p_s^{(w)}\right)^{x_{l_{im}}^{(w)}} \left(1 - p_s^{(w)}\right)^{1 - x_{l_{im}}^{(w)}} \left(p_D^{(w)}\right)^{\sum_{\substack{j=1 \\ j \neq l}}^{Q} x_{l_{ij}}^{(w)}} \left(1 - p_D^{(w)}\right)^{\sum_{\substack{j=1 \\ j \neq l}}^{Q} 1 - x_{l_{ij}}^{(w)}},$$

where, as before, $l_{ij} = (i-1)Q + j$. Adjustment for $t_d > 0$ yields

$$f_1(\mathbf{x}_w) = \prod_{i=1}^{n/Q} \frac{1}{Q} \sum_{m=1}^{Q} \left(p_s^{(w)}\right)^{x_{l_{im}}^{(w)}} \left(1 - p_s^{(w)}\right)^{\overline{\overline{x}}_{l_{im}}^{(w)}} \left(p_D^{(w)}\right)^{\sum_{\substack{j=1 \\ j \neq l}}^{Q} x_{l_{ij}}^{(w)}} \left(1 - p_D^{(w)}\right)^{\sum_{\substack{j=1 \\ j \neq l}}^{Q} \overline{\overline{x}}_{l_{ij}}^{(w)}},$$

where, as before, $\overline{\overline{x}}_{l_{ij}}^{(w)} = (1 - x_{l_{ij}}^{(w)})(1 - \sum_{k=1}^{t_d} x_{l_{ij}-k})$.

The likelihood ratio is

$$\frac{f_1(\mathbf{x}_w)}{f_0(\mathbf{x}_w)} = \prod_{i=1}^{n/Q} \frac{1}{Q} \sum_{m=1}^{Q} \left(\frac{p_s^{(w)}}{p_D^{(w)}}\right)^{x_{l_{im}}^{(w)}} \left(\frac{1 - p_s^{(w)}}{1 - p_D^{(w)}}\right)^{(1 - x_{l_{im}}^{(w)})(1 - \sum_{k=1}^{t_d} x_{l_{im}-k})}. \tag{D.5}$$

Now, when $x_{l_{im}}^{(w)} = 1$, the corresponding summation term in (D.5) is $\frac{p_s^{(w)}}{p_D^{(w)}}$. When $x_{l_{im}}^{(w)} = 0$ and the no-click event is within the detector dead time (that is, $1 - \sum_{k=1}^{t_d} x_{l_{im}-k} = 0$), then the corresponding summation term is one; otherwise the corresponding summation term is $\frac{1 - p_s^{(w)}}{1 - p_D^{(w)}}$. Denote by $y_i^{(w)} = \sum_{m=1}^{Q} x_{l_{im}}^{(w)}$ the number of clicks observed in the i^{th} PPM frame, and by $\overline{\overline{y}}_i^{(w)}$ the number of no-click events in the i^{th} PPM frame

within the detector dead time (where the click that triggered the detector reset may not necessarily be in the same frame). Equation (D.5) can thus be simplified as follows:

$$\begin{aligned}\frac{f_1(\mathbf{x}_w)}{f_0(\mathbf{x}_w)} &= \prod_{i=1}^{n/Q} \frac{1}{Q}\left[\frac{p_s^{(w)} y_i^{(w)}}{p_D^{(w)}} + \overline{\overline{y}}_i^{(w)} + \frac{\left(1-p_s^{(w)}\right)\left(Q - y_i^{(w)} - \overline{\overline{y}}_i^{(w)}\right)}{1-p_D^{(w)}}\right] \\ &= \prod_{i=1}^{n/Q}\left[1 + \zeta p_r^{(w)}\left(\frac{y_i^{(w)}}{Q p_D^{(w)}} + \frac{\overline{\overline{y}}_i^{(w)}}{Q} - 1\right)\right]. \end{aligned} \tag{D.6}$$

where equation (D.6) is obtained by noting that $\frac{p_s^{(w)}}{p_D^{(w)}} = \frac{\zeta p_r^{(w)}(1-p_D^{(w)})}{p_D^{(w)}} + 1$ and $\frac{1-p_s^{(w)}}{1-p_D^{(w)}} = 1 - \zeta p_r^{(w)}$. Taking the logarithm of (D.6) yields the log-likelihood ratio

$$\ln \frac{f_1(\mathbf{x}_w)}{f_0(\mathbf{x}_w)} = \sum_{i=1}^{n/Q} \ln\left[1 + \zeta p_r^{(w)}\left(\frac{y_i^{(w)}}{Q p_D^{(w)}} + \frac{\overline{\overline{y}}_i^{(w)}}{Q} - 1\right)\right]. \tag{D.7}$$

For small ζ, the Taylor series expansion of the summand in (D.7) at $\zeta = 0$ yields

$$\ln\left[1 + \zeta p_r^{(w)}\left(\frac{y_i^{(w)}}{Q p_D^{(w)}} + \frac{\overline{\overline{y}}_i^{(w)}}{Q} - 1\right)\right] \approx \zeta p_r^{(w)}\left(\frac{y_i^{(w)}}{Q p_D^{(w)}} + \frac{\overline{\overline{y}}_i^{(w)}}{Q} - 1\right). \tag{D.8}$$

Since $\sum_{i=1}^{n/Q} \overline{\overline{y}}_i^{(w)} = Y t_d$, where $Y = \sum_{i=1}^{n/Q} y_i^{(w)}$ is the total click count, the log-likelihood ratio can be approximated as follows:

$$\ln \frac{f_1(\mathbf{x}_w)}{f_0(\mathbf{x}_w)} \approx \frac{\zeta p_r^{(w)}}{Q p_D^{(w)}}(Y(1 + p_D^{(w)} t_d) - n p_D^{(w)}). \tag{D.9}$$

Since the approximation in (D.9) is an invertible function of the total click count Y, we use it as the test statistic for Willie. We also note that, because $p_D^{(w)} t_d \ll 1$ in our experiments, the dead time causes only a minimal performance degradation for Willie's detector.[1]

[1]We also conjecture that the availability of a detector with shorter dead time to Bob would only increase the number of covert bits that Alice can reliably transmit by a multiplicative constant related to Willie's detector dead time and not affect the square root scaling law.

D.3 Gaussian Approximation of $\mathbb{P}_e^{(w)}$

At the end of the previous section we argued that the detector dead time does not substantially impact Willie's performance, as the dead time is short relative to the average time between clicks. Here we show how setting $t_d = 0$ yields a useful analytical approximation of $\mathbb{P}_e^{(w)}$.

First consider the case when Alice does not transmit. Since $t_d = 0$, the total click count is a binomial random variable $Y \sim \mathcal{B}\left(y; p_D^{(w)}, n\right)$ whose distribution, for large n, can be approximated using the central limit theorem by a Gaussian distribution $\Phi\left(y; \mu_0, \sigma_0^2\right)$ with $\mu_0 = np_D^{(w)}$ and $\sigma_0^2 = np_D^{(w)}\left(1 - p_D^{(w)}\right)$, where $\Phi\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{|t-\mu|^2}{2\sigma^2}} dt$ is the distribution function of a Gaussian random variable $\mathcal{N}(x; \mu, \sigma^2)$.

Now consider the case when Alice transmits. Since $\mathcal{S}$ and $\mathbf{k}$ are unknown to Willie, the total click count is the sum of two independent but not identical binomial random variables $Y = X + Z$, where $X \sim \mathcal{B}\left(x; p_D^{(w)}, n - \frac{n}{Q}\right)$ is the number of dark clicks in the $n - \frac{n}{Q}$ modes that Alice never uses in a PPM scheme and $Z \sim \mathcal{B}\left(z; p_s^{(w)}, \frac{n}{Q}\right)$ is the contribution from the $\frac{n}{Q}$ modes that Alice can use to transmit, with $p_s^{(w)}$ defined in (D.4). By the central limit theorem, for large n, the distribution of X can be approximated using a Gaussian distribution $\Phi(x; \mu_X, \sigma_X^2)$ where $\mu_X = \left(n - \frac{n}{Q}\right) p_D^{(w)}$ and $\sigma_X^2 = \left(n - \frac{n}{Q}\right) p_D^{(w)}\left(1 - p_D^{(w)}\right)$. Similarly, the distribution of Z can be approximated by a Gaussian distribution $\Phi\left(z; \mu_Z, \sigma_Z^2\right)$ where $\mu_Z = \frac{n}{Q}\left(\zeta p_r^{(w)} + \left(1 - \zeta p_r^{(w)}\right) p_D^{(w)}\right)$ and $\sigma_Z^2 = \frac{n}{Q}\left(\zeta p_r^{(w)} + \left(1 - \zeta p_r^{(w)}\right) p_D^{(w)}\right)\left(1 - \zeta p_r^{(w)}\right)\left(1 - p_D^{(w)}\right)$. Thus, the distribution of Y can be approximated by a Gaussian distribution $\Phi\left(y; \mu_1, \sigma_1^2\right)$ with $\mu_1 = \mu_X + \mu_Z$ and $\sigma_1^2 = \sigma_X^2 + \sigma_Z^2$ via the additivity property of independent Gaussian random variables. Willie's probability of error is thus approximated by:

$$\tilde{\mathbb{P}}_e^{(w)} = \frac{1}{2}\min_S (1 - \Phi(S; \mu_0, \sigma_0^2) + \Phi(S; \mu_1, \sigma_1^2)). \tag{D.10}$$

The value of the threshold S^* that minimizes the RHS of (D.10) satisfies $\frac{|S^*-\mu_0|^2}{\sigma_0^2} - \log(\sigma_1^2/\sigma_0^2) = \frac{|S^*-\mu_1|^2}{\sigma_1^2}$.

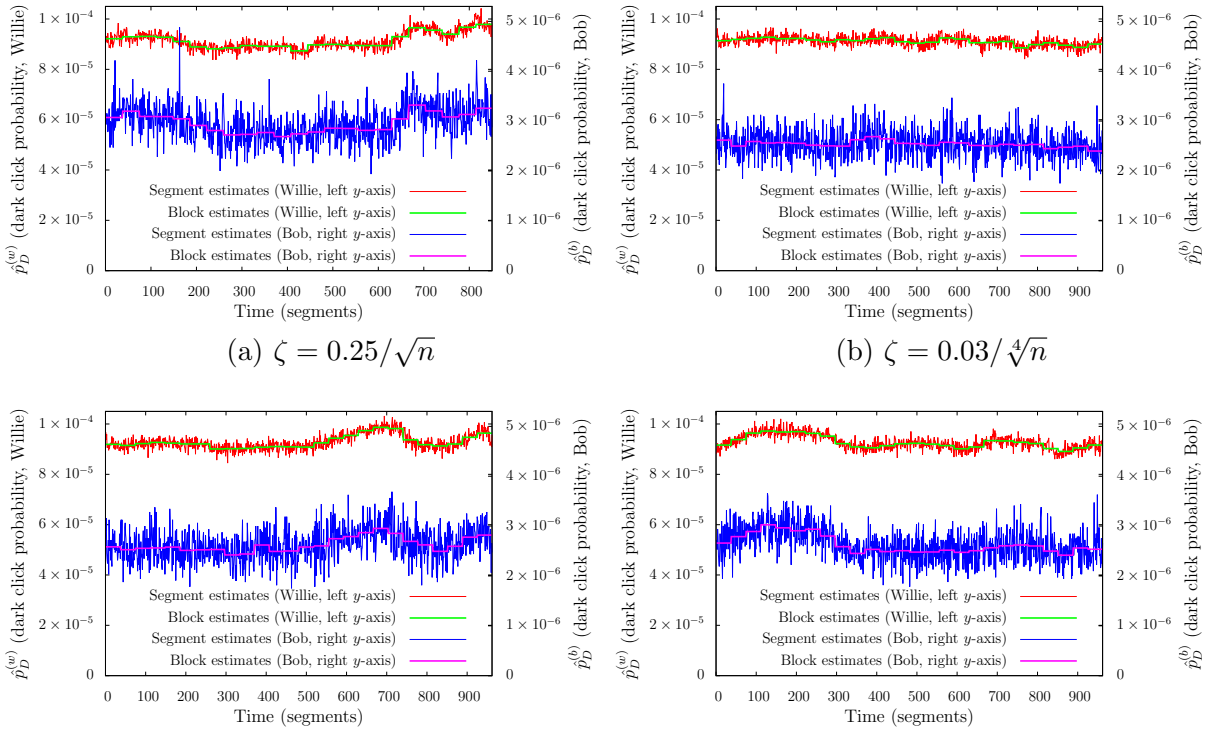
D.4 Analysis of the Detector Dark Clicks

Here we provide the detailed analysis of detector dark clicks, focusing on how their temporal variation affected our experiments. While we took great care in maintaining uniform conditions throughout our experiments, controlling every aspect of our environment was beyond our capabilities. However, we argue that the temporal variation in the dark click probability that we experienced had no significant impact on our results.

We maximize the logarithm of the likelihood function in equation (D.1) and obtain the following maximum likelihood estimator of the dark click probability:

$$\hat{p}_D = \frac{\sum_{i=1}^{n_D} x_i}{n_D - t_d \sum_{i=1}^{n_D} x_i}, \tag{D.11}$$

where $x_1, \ldots, x_{n_D}$ is the sequence of n_D observations where only the dark clicks can be observed, i.e, it is the experimental click record that excludes the observations of Alice's transmissions as well as the dead time following the detected transmissions. The entire click record contains 100 experiments at each value of n for both Alice using and not using the channel, totalling 2.72×10^{10} observations when $\zeta = 0.25/\sqrt{n}$, and 3.0784×10^{10} observations when $\zeta \in \{0.03/\sqrt[4]{n}, 0.003, 0.008\}$. For each of the four communication regimes, we divide the click record (sorted by time) into *segments* of $n_{D_s} = 3.2 \times 10^7$ consecutive observations. The estimates of Willie's and Bob's dark click probabilities for these segments are denoted by $\hat{p}_{D,s}^{(w)}(j)$ and $\hat{p}_{D,s}^{(b)}$, $j = 1, \ldots, n_{S,s}(\zeta)$, where $n_{S,s}(0.25/\sqrt{n}) = 850$ and $n_{S,s}(0.03/\sqrt[4]{n}) = n_{S,s}(0.003) = n_{S,s}(0.008) = 962$. The plots of $\hat{p}_{D,s}^{(w)}(j)$ and $\hat{p}_{D,s}^{(b)}(j)$ in Figure D.1 illustrate the temporal variations in dark click probability. However, they also show homogeneity over



(a) $\zeta = 0.25/\sqrt{n}$ (b) $\zeta = 0.03/\sqrt[4]{n}$

(c) $\zeta = 0.003$ (d) $\zeta = 0.008$

Figure D.1: Temporal variation in dark click probability. The estimates of Willie's dark click probability are plotted using the left y-axis; the estimates of Bob's dark click probability are plotted using the right y-axis. Dark click probability is estimated using equation (D.11) for consecutive segments, each containing $n_{D_s} = 3.2\times10^7$ consecutive observations, as well as for consecutive blocks, each containing $n_{D_b} = 1.184 \times 10^9$ consecutive observations (with the exception of the last block for $\zeta = 0.25/\sqrt{n}$ in panel (a) where $n_{D_b} = 1.152 \times 10^9$ observations).

relatively long periods of time. We thus estimate the dark click probability for *blocks* of 37 consecutive segments using $n_{D_b} = 1.184 \times 10^9$ observations (except for the last, 23$^{\text{rd}}$, block of the click record for $\zeta = 0.25/\sqrt{n}$ containing 36 ($n_{D_b} = 1.152 \times 10^9$ observations) segments instead of 37). The estimates of Willie's and Bob's dark click probabilities for these blocks are denoted by $\hat{p}_{D,b}^{(w)}(k)$ and $\hat{p}_{D,b}^{(b)}(k)$, $k = 1, \ldots, n_{S,b}(\zeta)$, where $n_{S,b}(0.25/\sqrt{n}) = 23$ and $n_{S,s}(0.03/\sqrt[4]{n}) = n_{S,s}(0.003) = n_{S,s}(0.008) = 26$. The block dark click probability estimate is close to the average of the estimates for its

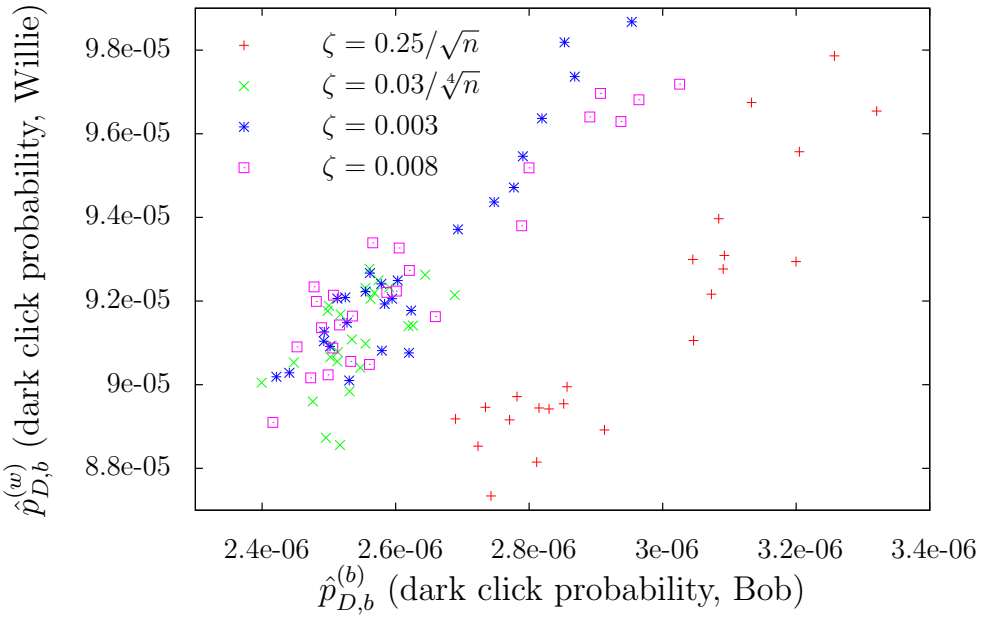


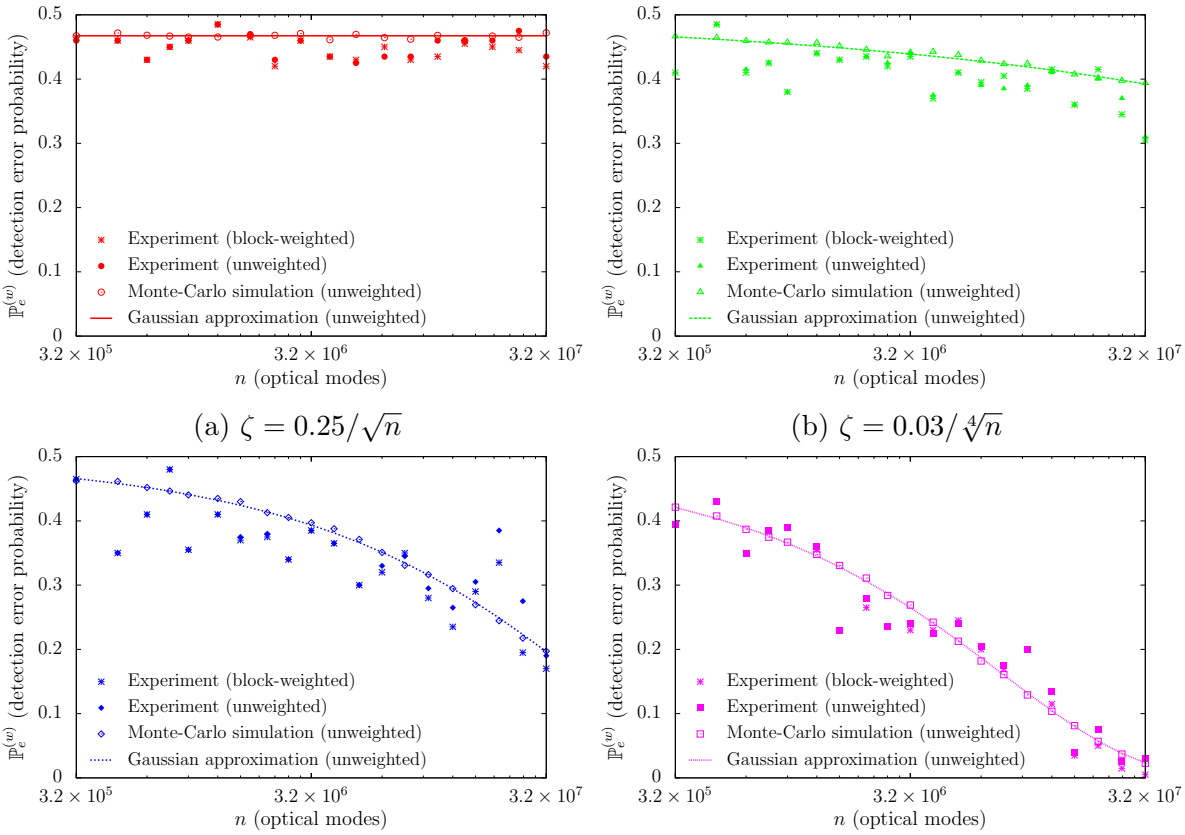
Figure D.2: Correlation in Willie's and Bob's dark click probabilities. A glitch in the PCIe-6537 data acquisition card resulted in a slight change in setup between the experiments corresponding to $\zeta = 0.25/\sqrt{n}$ and $\zeta \in \{0.03/\sqrt[4]{n}, 0.003, 0.008\}$.

component segments, and over most of the 101 blocks, the segment estimates of the dark click probability are homogeneous.[2]

We also plot $\hat{p}^{(w)}_{D,b}(j)$ versus $\hat{p}^{(b)}_{D,b}(j)$ in Figure D.2, revealing strong correlation between the dark click probabilities of Bob's and Willie's detectors. Thus the temporal variations in the dark click probabilities likely stem from the external environmental factors (such as laboratory temperature changes) rather than the detectors themselves.

Intuitively, clicks observed under less noisy channel conditions carry more evidence for the hypothesis that Alice is transmitting than clicks observed when the channel

[2]Assuming that the dark click probability stays constant over the period of time corresponding to a segment, the number of observed dark clicks would be binomially-distributed if our detectors had zero dead time. However, we argue in Appendix D.2 that the dead time has a minimal impact on our experiment, making the binomial distribution a good approximation for the distribution of the number of observed dark clicks. Thus, for each of blocks, we performed the Pearson chi-squared test for homogeneity [36] in the estimated dark click probability of its component segments. We found that the test rejects the null hypothesis (that the estimates are homogeneous) in only 29 out of 202 blocks, consistent with the visual inspection of Figure D.1.



(a) $\zeta = 0.25/\sqrt{n}$

(b) $\zeta = 0.03/\sqrt[4]{n}$

(c) $\zeta = 0.003$

(d) $\zeta = 0.008$

Figure D.3: Impact of variations in dark click probability on the estimates of Willie's detection error. The probability of detection error estimated from the block-weighted test statistic given in equation (D.12) is plotted using the asterisks alongside the plots from Figure 7.3 of the estimates using the (unweighted) total click count. Weighting does not substantially change the detection error probability estimates.

is noisier. Indeed, in the derivation of the total click count Y as Willie's test statistic in Section D.2, the contribution to Y from the i^{th} channel observation $x_i^{(w)} \in \{0, 1\}$ is effectively weighted by $1/p_D^{(w)}$ (we ignore the term corresponding to the detector dead time for simplicity of exposition and since it has no tangible impact on our experimental results). In the analysis of Section 7.2 we used the average dark click probability, however, if the dark click probability $p_D^{(w)}(i)$ is available for $i = 1, \ldots, n$, $Y_{\text{weighted}} = \sum_{i=1}^{n} \frac{x_i^{(w)}}{p_D^{(w)}(i)}$ is a better test statistic. Since the exact $p_D^{(w)}(i)$ is unavailable, and since the estimates are (mostly) homogeneous over the duration of the blocks,

we study the impact of temporal variation of dark clicks on Willie's detection error by weighting the observations by the block estimates $\hat{p}_{D,b}^{(w)}(j)$. Denoting the set of observations in the j^{th} block by $\mathcal{W}_j$, we block-weight Willie's test statistic as follows:

$$Y_{\text{block-weighted}} = \sum_{j=1}^{n_{S,l}(\zeta)} \frac{1}{\hat{p}_{D,b}^{(w)}(j)} \sum_{x_i^{(w)} \in \mathcal{W}_j} x_i^{(w)}. \tag{D.12}$$

We plot the estimates of detection error probability that are calculated using the block-weighted test statistic given by (D.12) in Figure D.3 alongside the estimates from Figure 7.3 that are calculated using the (unweighted) total click count. While the estimated probability of detection error decreases in some cases (and increases or remains the same in others), the overall impact is small. The square root scaling law is unaffected since Alice and Bob can design their covert communication using a lower bound on $p_D^{(w)}$ (e.g., the dark click probability for the best available photon detector operating in near-ideal conditions). However, since the random fluctuations in noise power have been shown to yield positive-rate covert communication in AWGN channel setting [57, 58], Alice and Bob could potentially exploit the random process governing $p_D^{(w)}$ to transmit covert information at a positive rate.

BIBLIOGRAPHY

[1] Abdo, Baleegh, Sliwa, Katrina, Shankar, S., Hatridge, Michael, Frunzio, Luigi, Schoelkopf, Robert, and Devoret, Michel. Josephson directional amplifier for quantum measurement of superconducting circuits. *Phys. Rev. Lett. 112* (Apr 2014), 167701.

[2] Anderson, Duwayne R., Johnson, Larry M., and Bell, Florian G. *Troubleshooting Optical Fiber Networks: Understanding and Using Optical Time-Domain Reflectometers*, 2nd ed. Elsevier Academic Press, 2004.

[3] Bash, Boulat A., Gheorghe, Andrei H., Patel, Monika, Habif, Jonathan L., Goeckel, Dennis, Towsley, Don, and Guha, Saikat. Covert optical communication. arXiv:1404:7347, University of Massachusetts Technical Report UM-CS-2014-004, 2014.

[4] Bash, Boulat A., Goeckel, Dennis, and Towsley, Don. Limits of reliable communication with low probability of detection on AWGN channels. *IEEE J. Selected Areas Commun. 31*, 9 (2013), 1921–1930. Originally presented at ISIT 2012, Cambridge MA.

[5] Bash, Boulat A., Goeckel, Dennis, and Towsley, Don. LPD Communication when the Warden Does Not Know When. In *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)* (Honolulu, HI, July 2014). arXiv:1403.1013.

[6] BBC. Edward Snowden: Leaks that exposed US spy programme. `http://www.bbc.com/news/world-us-canada-23123964`, Jan. 2014.

[7] Bennett, C. H., and Brassard, G. Quantum Cryptography: Public Key Distribution and Coin Tossing. In *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing* (New York, 1984), IEEE Press, pp. 175–179.

[8] Berrou, C., Glavieux, A., and Thitimajshima, P. Near shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Proceedings of IEEE International Conference on Communications* (Geneva, May 1993), vol. 2, pp. 1064–1070 vol.2.

[9] Billingsley, Patrick. *Probability and Measure*, 3rd ed. Wiley, New York, 1995.

[10] Bloch, M.R., and Laneman, J.N. Strong secrecy from channel resolvability. *IEEE Trans. Inf. Theory 59*, 12 (Dec 2013), 8077–8098.

[11] Bouchet, O., Sizun, H., Boisrobert, C., de Fornel, F., and Favennec, P.N. *Free-Space Optics: Propagation and Communication*. Wiley, 2010.

[12] Cachin, Christian. An information-theoretic model for steganography. *Information and Computation 192*, 1 (2004), 41–56.

[13] Campos, Richard A., Saleh, Bahaa E. A., and Teich, Malvin C. Quantum-mechanical lossless beam splitter: SU(2) symmetry and photon statistics. *Phys. Rev. A 40* (Aug 1989), 1371–1384.

[14] Capar, C., Goeckel, D., Liu, Benyuan, and Towsley, D. Secret communication in large wireless networks without eavesdropper location information. In *2012 Proceedings of IEEE INFOCOM* (March 2012), pp. 1152–1160.

[15] Che, Pak Hou, Bakshi, Mayank, and Jaggi, Sidharth. Reliable deniable communication: Hiding messages in noise. arXiv:1304.6693, 2013.

[16] Che, Pak Hou, Bakshi, Mayank, and Jaggi, Sidharth. Reliable deniable communication: Hiding messages in noise. In *Proc. of IEEE International Symposium on Information Theory (ISIT)* (Istanbul, Turkey, July 2013). arXiv:1304.6693.

[17] Chiani, Marco, Dardari, Davide, and Simon, Marvin K. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications 2*, 4 (July 2003), 840–845.

[18] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L., and Stein, Clifford. *Introduction to Algorithms*, 2nd ed. MIT Press, Cambridge, Massachusetts, 2001.

[19] Cover, Thomas M., and Thomas, Joy A. *Elements of Information Theory*, 2nd. ed. John Wiley & Sons, Hoboken, NJ, 2002.

[20] Craver, Scott, and Yu, Jun. Subset selection circumvents the square root law. vol. 7541, pp. 754103–1–754103–6.

[21] Csiszár, I., and Körner, J. Broadcast channels with confidential messages. *IEEE Transactions on Information Theory 24*, 3 (May 1978), 339–348.

[22] Csiszár, Imre, and Körner, János. Broadcast channels with confidential messages. *Information Theory, IEEE Transactions on 24*, 3 (May 1978), 339–348.

[23] Danezis, G., and Diaz, C. A survey of anonymous communication channels. Tech. Rep. TR-2008-35, Microsoft Research, 2008.

[24] Dasgupta, Sanjoy, and Gupta, Anupam. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms 22*, 1 (Jan. 2003), 60–65.

[25] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist. 27* (1956), 642–669.

[26] Filler, Tomáš, and Fridrich, Jessica. Fisher information determines capacity of ϵ-secure steganography. In *Information Hiding*, Stefan Katzenbeisser and Ahmad-Reza Sadeghi, Eds., vol. 5806 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 31–47.

[27] Filler, Tomáš, Ker, Andrew D., and Fridrich, Jessica. The square root law of steganographic capacity for markov covers. *Media Forensics and Security 7254*, 1 (2009).

[28] Franceschetti, Massimo, Dousse, Olivier, Tse, David N. C., and Thiran, Patrick. Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Transactions on Information Theory 53*, 3, 1009–1018.

[29] Fridrich, Jessica. *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed. Cambridge University Press, New York, 2009.

[30] Gagliardi, R.M., and Karp, S. *Optical Communications*, 2nd ed. Wiley, 1995.

[31] Gallager, Robert G. *Low-Density Parity-Check Codes*, vol. 21 of *Research monograph series*. M.I.T. Press, Cambridge, MA, 1963.

[32] Gallager, Robert G. *Information Theory and Reliable Communication*. John Wiley and Sons, Inc., New York, 1968.

[33] Giovannetti, V., Guha, S., Lloyd, S., Maccone, L., Shapiro, J. H., and Yuen, H. P. Classical capacity of the lossy bosonic channel: The exact solution. *Phys. Rev. Lett. 92* (Jan 2004), 027902.

[34] Goeckel, D., Vasudevan, S., Towsley, D., Adams, S., Ding, Z., and Leung, K. Artificial noise generation from cooperative relays for everlasting secrecy in two-hop wireless networks. *IEEE Journal on Selected Areas in Communications 29*, 10 (December 2011), 2067–2076.

[35] Goodman, J.W. *Introduction to Fourier Optics*, 3rd ed. Roberts & Company, 2005.

[36] Greenwood, P.E., and Nikulin, M.S. *A guide to chi-squared testing*. Wiley, New York, NY, 1996.

[37] Griffiths, D.J. *Introduction to Quantum Mechanics*, 2nd. ed. Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

[38] Guha, Saikat. *Multiple-User Quantum Information Theory for Optical Communication Channels*. PhD thesis, Massachusetts Institute of Technology, 2008.

[39] Gupta, P., and Kumar, P.R. The capacity of wireless networks. *IEEE Transactions on Information Theory 46*, 2 (Mar 2000), 388–404.

[40] Hausladen, Paul, Jozsa, Richard, Schumacher, Benjamin, Westmoreland, Michael, and Wootters, William. Classical information capacity of a quantum channel. *Phys. Rev. A 54* (Sep 1996), 1869–1876.

[41] Helstrom, Carl W. *Quantum Detection and Estimation Theory*. Academic Press, Inc., New York, NY, USA, 1976.

[42] Herodotus. c. 440 BCE. 5.35 and 7.239.

[43] Holevo, A. S. The capacity of the quantum channel with general signal states. *IEEE Trans. Inf. Theory 44* (1998), 269–273.

[44] Hou, Jie, and Kramer, Gerhard. Effective secrecy: Reliability, confusion and stealth. In *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)* (Honolulu, HI, July 2014). arXiv:1311.1411.

[45] Itzler, M.A., Entwistle, M., and Jiang, Xudong. High-rate photon counting with geiger-mode APDs. In *In Proc. IEEE Photonics Conference (PHO)* (Oct 2011), pp. 348–349.

[46] Kadhe, Swanand, Jaggi, Sidharth, Bakshi, Mayank, and Sprintson, Alex. Reliable, deniable, and hidable communication over multipath networks. In *Proc. of IEEE International Symposium on Information Theory (ISIT)* (Honolulu, HI, July 2014). arXiv:1401.4451.

[47] Ker, Andrew D. Batch steganography and pooled steganalysis. In *Information Hiding*, Jan L. Camenisch, Christian S. Collberg, Neil F. Johnson, and Phil Sallee, Eds., vol. 4437 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2007, pp. 265–281.

[48] Ker, Andrew D. A capacity result for batch steganography. *IEEE Signal Processing Letters 14*, 8 (Aug. 2007), 525–528.

[49] Ker, Andrew D. Estimating steganographic fisher information in real images. In *Information Hiding*, Stefan Katzenbeisser and Ahmad-Reza Sadeghi, Eds., vol. 5806 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 73–88.

[50] Ker, Andrew D. The square root law requires a linear key. In *Proceedings of the 11th ACM workshop on Multimedia and security* (Princeton, NJ, USA, 2009), MM&Sec '09, pp. 85–92.

[51] Ker, Andrew D. The square root law does not require a linear key. In *Proceedings of the 12th ACM workshop on Multimedia and Security* (Roma, Italy, 2010), MM&Sec '10, pp. 213–224.

[52] Kerckhoffs, Auguste. La cryptographie militaire. *Journal des sciences militaires IX* (English translation), 5–83, January 1883, pp. 161–191, February 1883.

[53] Kopeika, N.S., and Bordogna, J. Background noise in optical communication systems. *Proc. of the IEEE 58*, 10 (Oct. 1970), 1571–1577.

[54] Kramer, Gerhard. Secret-less covert communication on the discrete memoryless channels. Personal communication, 2014.

[55] Kullback, Solomon. *Information Theory and Statistics*. Wiley, New York, NY, 1959.

[56] Lang, Serge. *Undergraduate Analysis*, 2nd ed. Springer-Verlag, New York, NY, 1997.

[57] Lee, Seonwoo, and Baxley, R.J. Achieving positive rate with undetectable communication over AWGN and Rayleigh channels. In *Proc. of IEEE Int. Conf. Commun. (ICC)* (June 2014), pp. 780–785.

[58] Lee, Seonwoo, Baxley, R.J., McMahon, J.B., and Frazier, R.S. Achieving positive rate with undetectable communication over mimo rayleigh channels. In *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)* (June 2014), pp. 257–260.

[59] Lehmann, Erich, and Romano, Joseph. *Testing Statistical Hypotheses*, 3rd. ed. Springer, New York, 2005.

[60] Leung-Yan-Cheong, S., and Hellman, M. The gaussian wire-tap channel. *IEEE Transactions on Information Theory 24*, 4 (July 1978), 451–456.

[61] MacKay, D.J.C., and Neal, R.M. Near shannon limit performance of low density parity check codes. *Electronics Letters 33*, 6 (Mar 1997), 457–458.

[62] Madhow, Upamanyu. *Fundamentals of Digital Communication*. Cambridge University Press, Cambridge, UK, 2008.

[63] Majani, Eric E. *A model for the study of very noisy channels, and applications*. PhD thesis, California Institute of Technology, 1988.

[64] Makarov, Vadim. Controlling passively quenched single photon detectors by bright light. *New Journal of Physics 11*, 6 (2009), 065003.

[65] Marano, S., Matta, V., He, T., and Tong, L. The embedding capacity of information flows under renewal traffic. *IEEE Transactions on Information Theory 59*, 3 (2013), 1724–1739.

[66] Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability 18*, 3 (07 1990), 1269–1283.

[67] Menezes, Alfred J., Vanstone, Scott A., and Oorschot, Paul C. Van. *Handbook of Applied Cryptography*, 1st ed. CRC Press, Inc., Boca Raton, FL, USA, 1996.

[68] Moision, Bruce, Hamkins, Jon, and Cheng, Michael. Design of a coded modulation for deep space optical communications. In *Information Theory and its Applications (ITA) Workshop* (2006), University of California San Diego.

[69] Nielsen, Michael A., and Chuang, Isaac L. *Quantum Computation and Quantum Information*. Cambridge University Press, New York, NY, USA, 2000.

[70] Pirandola, Stefano, and Lloyd, Seth. Computable bounds for the discrimination of gaussian states. *Phys. Rev. A 78* (Jul 2008), 012331.

[71] Ribordy, Gregoire, Gautier, Jean-Daniel, Zbinden, Hugo, and Gisin, Nicolas. Performance of ingaas/inp avalanche photodiodes as gated-mode photon counters. *Applied Optics 37*, 12 (1998), 2272–2277.

[72] Sakurai, J.J. *Modern Quantum Mechanics*, 2nd. ed. Addison-Wesley, Reading, MA, 1993.

[73] Schumacher, Benjamin, and Westmoreland, Michael. Sending classical information via noisy quantum channels. *Phys. Rev. A 56* (Jul 1997), 131–138.

[74] Senior, John. *Optical Fiber Communications*, 3rd ed. Pearson Education, 2009.

[75] Shannon, Claude E. A mathematical theory of communication. *Bell System Technical Journal 27* (1948).

[76] Shannon, Claude E. Communication theory of security. *Bell System Technical Journal 28* (1949), 656–715.

[77] Shaw, Bilal A., and Brun, Todd A. Quantum steganography with noisy quantum channels. *Phys. Rev. A 83* (Feb 2011), 022310.

[78] Simon, Marvin K., Omura, Jim K., Scholtz, Robert A., and Levitt, Barry K. *Spread Spectrum Communications Handbook*. McGraw-Hill, 1994.

[79] Soltani, Ramin, Bash, Boulat A., Goeckel, Dennis, Guha, Saikat, and Towsley, Don. Covert single-hop communication in a wireless network with distributed artificial noise generation. In *Proc. of Conf. on Commun., Control, Comp. (Allerton)* (Monticello, IL, 2014).

[80] Talbot, J., and Welsh, D.J.A. *Complexity and Cryptography: An Introduction*. Cambridge University Press, 2006.

[81] Варакин, Л. Е. [Varakin, L. E.]. *Системы связи с шумоподобными сигналами [Sistemy svyazi s shumopodobnymi signalami] (Spread Spectrum Communication Systems)*. Радио и связь [Radio i Svyaz'], Moscow, USSR, 1985. (in Russian).

[82] Torrieri, Don. *Principles of Spread-spectrum Communication Systems*. Springer, Boston, MA, USA, 2005.

[83] Van Trees, Harry L. *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, Inc., New York, 2001.

[84] Vasudevan, Sudarshan, Goeckel, Dennis, and Towsley, Donald F. Security-capacity trade-off in large wireless networks using keyless secrecy. In *Proceedings of the Eleventh ACM International Symposium on Mobile Ad Hoc Networking and Computing* (Chicago, Illinois, USA, 2010), pp. 21–30.

[85] Viterbi, Andrew J. *CDMA: Principles of Spread Spectrum Communication*. Addison-Wesley, Reading, MA, 4 1995.

[86] Wang, Ligong, and Wornell, Gregory W. Refined analysis of the poisson channel in the high-photon-efficiency regime. In *Proc. of IEEE Information Theory Workshop (ITW)* (Lausanne, Switzerland, 2012), pp. 582–586.

[87] Wilde, Mark M., Hayden, Patrick, and Guha, Saikat. Information trade-offs for optical quantum communication. *Phys. Rev. Lett. 108* (Apr 2012), 140501.

[88] Wilde, M.M. *Quantum Information Theory*. Cambridge University Press, 2013.

[89] Wolf, Michael M., Pérez-García, David, and Giedke, Geza. Quantum capacities of bosonic channels. *Phys. Rev. Lett. 98* (Mar 2007), 130501.

[90] Wyner, Aaron D. The wire-tap channel. *Bell System Technical Journal 54* (1975), 1355–1387.

[91] Yucek, T., and Arslan, H. A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys & Tutorials 11*, 1 (First Qtr 2009), 116–130.