Limits of Reliable Communication with Low Probability of Detection on AWGN Channels

Boulat A. Bash, Dennis Goeckel, Don Towsley

UMass Technical Report UM-CS-2012-003

Abstract

We present a square root limit on the amount of information transmitted reliably and with low probability of detection (LPD) over additive white Gaussian noise (AWGN) channels. Specifically, if the transmitter has AWGN channels to an intended receiver and a warden, both with non-zero noise power, we prove that $o(\sqrt{n})$ bits can be sent from the transmitter to the receiver in n channel uses while lower-bounding $\alpha+\beta\geq 1-\epsilon$ for any $\epsilon>0$, where α and β respectively denote the warden's probabilities of a false alarm when the sender is not transmitting and a missed detection when the sender is transmitting. Moreover, in most practical scenarios, a lower bound on the noise power on the channel between the transmitter and the warden is known and $\mathcal{O}(\sqrt{n})$ bits can be sent in n LPD channel uses. 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 $n\to\infty$.

B. A. Bash and D. Towsley are with the School of Computer Science, University of Massachusetts, Amherst, Massachusetts.

D. Goeckel is with the Electrical and Computer Engineering Department, University of Massachusetts, Amherst, Massachusetts.

This research was sponsored by the National Science Foundation under grants CNS-0905349 and CNS-1018464, and by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

I. INTRODUCTION

Securing information transmitted over wireless links is of paramount concern for consumer, industrial, and military applications. Typically data transmitted in wireless networks is secured from interception by an eavesdropper using various encryption and key exchange protocols. However, there are many real-life scenarios where standard cryptographic security is not sufficient. Encrypted data arouses suspicion, and even the most theoretically robust encryption can often be defeated by a determined adversary using non-computational methods such as side-channel analysis. Such scenarios require *low probability of detection* (LPD) communication which prevents the detection of transmissions in the first place.

While practical LPD communications has been studied by the spread-spectrum community [1, Pt. 5, Ch. 1], the information-theoretic limits have not been explored. We thus develop fundamental bounds on LPD communication over wireless channels subject to additive white Gaussian noise (AWGN). In our scenario, Alice communicates with Bob over an AWGN channel, while passive eavesdropper Warden Willie attempts to detect her transmission. The channel between Alice and Willie is also AWGN and Willie is passive in that he does not actively jam Alice's channel. Alice transmits low-power signals to Bob that Willie attempts to classify 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. 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 across the LPD channel to Bob in n uses of the channel.

Our problem is related to imperfect steganography, which considers hiding information by altering the properties of fixed-size, finite-alphabet covertext objects (such as images or software binary code) while tolerating some fixed probability of detection of hidden information by the warden. The square root law of steganography in the passive warden environment states that

$\mathcal{O}(\sqrt{n})$ symbols in covertext of size n may safely be modified to hide an $\mathcal{O}(\sqrt{n}\log n)$-bit steganographic message [2, Ch. 13], where the $\log n$ factor stems directly from the fact that transmission to Bob is noiseless [2, Ch. 8]. In our scenario, Alice uses the noise on her channel to Willie instead of the statistical properties of the covertext to hide information. However, having to code against the noise on her channel to Bob allows only $\mathcal{O}(\sqrt{n})$ bits to be sent in n uses of the LPD channel.[1] The mathematics of statistical hypothesis testing yields a square root law in both problems, but as answers to different questions due to the fundamental differences in the communication channels. This relationship is discussed further at the end of Section III.

We state our main result that limits the amount of information that can be transmitted on the LPD channel between Alice and Bob using asymptotic notation [3, Ch. 3.1] where $f(n) = \mathcal{O}(g(n))$ denotes an asymptotically tight 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 $n \geq n_0$), and $f(n) = \omega(g(n))$ 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 $n \geq n_0$):

Theorem (Square root law)**.** *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. Denote by α the probability that Willie raises a false alarm when Alice is not transmitting, and by β the probability that Willie does not detect a transmission by Alice. 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 arbitrary low probability of decoding error) transmit $o(\sqrt{n})$ information bits to Bob in n channel uses while lower-bounding Willie's sum*

[1] The amount of information that could be transmitted by Alice to Bob using a *noiseless* LPD channel would be infinite due to it being continuously-valued, and a noiseless channel between Alice and Willie would preclude the existence of an LPD channel between Alice and Bob.

of the probabilities of detection errors $\alpha+\beta \geq 1-\epsilon$. 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 $\alpha+\beta \geq 1-\epsilon$. 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.

To enable LPD communication, Alice and Bob possess a common secret randomness resource. While in the information-theoretic analysis of encrypted communication such a resource is a one-time pad [4], in the construction of our proofs it is a secret codebook that is shared between Alice and Bob prior to communication and which is the only component of their system that is unknown to Willie. This follows "best practices" in security system design as the security of the LPD communication system depends only on the shared secret [5].

We also note that, since LPD 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 LPD 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 bits per channel use.

After introducing our channel model and hypothesis testing background in Section II, we prove the achievability of the square root law in Section III. We then prove the converse in Section IV. We discuss the relationship to previous work, the impact of Willie's prior knowledge of Alice's transmission state, and the mapping to the continuous-time channel in Section V, and conclude in Section VI.

II. PREREQUISITES

A. Channel Model

We use the discrete-time AWGN channel model with real-valued symbols (and defer discussion of the mapping to a continuous-time channel to Section V-C). Our formal system framework is depicted in Figure 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.) $z_i^{(b)} \sim \mathcal{N}(0, \sigma_b^2)$. Willie observes vector $\mathbf{y}_w = \{y_i^{(w)}\}_{i=1}^n$ where $y_i^{(w)} = f_i + z_i^{(w)}$, with i.i.d. $z_i^{(w)} \sim \mathcal{N}(0, \sigma_w^2)$. Willie uses statistical hypothesis tests on $\mathbf{y}_w$ to determine whether Alice is communicating, which we discuss next.

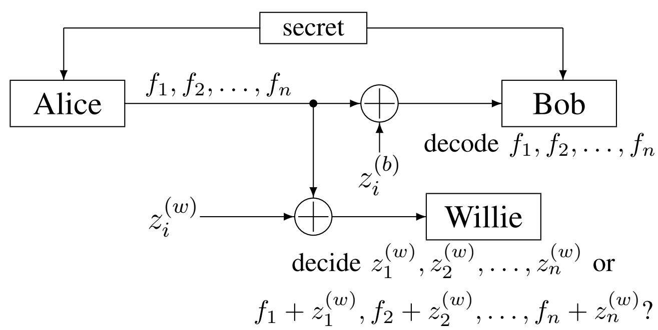

Fig. 1. System framework: Alice and Bob share a secret before the transmission. Alice encodes information into a vector of real symbols $\mathbf{f} = \{f_i\}_{i=1}^n$ and transmits it on an AWGN channel to Bob, while Willie attempts to classify his vector of observations of the channel from Alice $\mathbf{y}_w$ 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.

B. Hypothesis Testing

Willie expects vector $\mathbf{y}_w$ of n channel readings to be consistent with his channel noise model. He performs a statistical hypothesis test on this vector, with the null hypothesis H_0 being that Alice is not communicating. In this case each sample is i.i.d. $y_i^{(w)} \sim \mathcal{N}(0, \sigma_w^2)$. The alternate hypothesis H_1 is that Alice is transmitting, which corresponds to samples $y_i^{(w)}$ coming from a different distribution. Willie can tolerate some false positives, or cases when his statistical test incorrectly accuses Alice. This rejection of H_0 when it is true is known as the type I error (or false alarm), and, following the standard nomenclature, we denote its probability by α [6]. Willie's test may also miss Alice's transmissions. Acceptance of H_0 when it is false is known as the type II error (or missed detection), and we denote its probability by β. We assume that Willie uses classical hypothesis testing with equal prior probabilities of each hypothesis being true (and discuss the generalization to unequal prior probabilities in Section V-B). Then, the lower bound on the sum $\alpha+\beta$ characterizes the necessary trade-off between the false alarms and the missed detections in the design of a hypothesis test.

III. Achievability of Square Root Law

Willie's objective is to determine whether Alice transmits given the vector of observations $\mathbf{y}_w$ of his channel from Alice. 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 result, we assume 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. 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 sum of error probabilities $\alpha+\beta$ [6, Ch. 13]. The following holds for such a test:

Fact 1 (Theorem 13.1.1 in [6])**.** *For the optimal test,*

$$\alpha+\beta=1-\mathcal{V}_T(\mathbb{P}_0,\mathbb{P}_1)$$

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

Definition 1 (Total variation distance [6])**.** *The* total variation distance *between two continuous* *probability measures* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *is*

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

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

Implicit in the above is that the *a priori* probabilities of H_0 and H_1 are unknown to Willie. We discuss the inclusion of knowledge of prior probabilities in Section V-B.

Since total variation 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 total variation metric is unwieldy for products of probability measures, which are used in the analysis of the vectors of observations. We thus use Pinsker's inequality:

Fact 2 (Pinsker's inequality (Lemma 11.6.1 in [7]))**.**

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

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

Definition 2. *The* relative entropy *(also known as* Kullback–Leibler divergence*) between two probability measures* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *is:*

$$\mathcal{D}(\mathbb{P}_0\|\mathbb{P}_1) = \int_{\mathcal{X}} p_0(x) \ln \frac{p_0(x)}{p_1(x)} dx \tag{2}$$

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:

Fact 3 (Relative entropy product)**.** *From the chain rule for relative entropy [7, Eq. (2.67)]:*

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

Relative entropy is directly related to Neyman–Pearson hypothesis testing via the Chernoff–Stein Lemma [7, Ch. 11.8]: for a given $\alpha < \nu$ with $0 < \nu < \frac{1}{2}$, $\lim_{\nu \to 0} \lim_{n \to \infty} \frac{1}{n} \ln \beta^* =$

$-\mathcal{D}(\mathbb{P}_0\|\mathbb{P}_1)$ where $\beta^* = \min \beta$. 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 [2], [8]. However, we take the extra step of lower-bounding $\alpha+\beta$ since it has a natural signal processing interpretation via the receiver operating characteristic (ROC) curve [9, Ch. 2.2.2], which plots probability of detection $1-\beta$ versus α. Since $1-\beta \geq \alpha$ and $\alpha+\beta \geq 1-\epsilon$, small ϵ implies that the ROC curve lies very close to the line of no-discrimination (the diagonal line where $1-\beta=\alpha$) over the entire domain of α because $\alpha+\epsilon \geq 1-\beta \geq \alpha$.

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

$$\begin{aligned} f(v) =& f(u) + f'(u)(v-u) + \ldots + \frac{f^{(n)}(u)}{n!}(v-u)^n \\ &+ \frac{f^{(n+1)}(\xi)}{(n+1)!}(v-u)^{n+1} \end{aligned} \tag{3}$$

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

The proof can be found in, e.g. [10, 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 1.1 (Achievability)**.** *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 sum of the probabilities of detection errors $\alpha+\beta \geq 1-\epsilon$ 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 input blocks of length M bits and encodes them into codewords of length n at the rate of $R = M/n$ bits/symbol. We employ random coding arguments and independently generate 2^{nR} codewords $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^{nR}\}$ from $\mathbb{R}^n$ for messages $\{W_k\}_{k=1}^{2^{nR}}$, 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 in the remark following the proof of Theorem 1.2.

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 1, the sum of the probabilities of Willie's detector's errors is expressed by $\alpha + \beta = 1 - \mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1)$, where the total variation 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 the sum of the error probabilities by upper-bounding the total variation distance: $\mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1) \leq \epsilon$.

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. Therefore, Willie's probability distribution of the transmitted symbols is of zero-mean i.i.d. Gaussian random variables with variance P_f. Since noise is independent of the transmitted symbols, 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 2 and 3:

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

In our case the relative entropy is:

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

Since the first three derivatives of $\mathcal{D}(\mathbb{P}_w\|\mathbb{P}_s)$ with respect to P_f are continuous, we can apply Lemma 1. 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:

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

That relative entropy is locally quadratic is well-known [11, Ch. 2.6]; in fact $\left.\frac{\partial^2\mathcal{D}(\mathbb{P}_w\|\mathbb{P}_s)}{\partial P_f^2}\right|_{P_f=0}=\frac{1}{2\sigma_w^4}$ is the Fisher information that an observation of noise carries about its power. Now, the remainder term is:

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

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 $c=2\epsilon\sqrt{2}$ 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:

$$\mathcal{V}_T(\mathbb{P}_w^n,\mathbb{P}_s^n)\leq\frac{P_f}{2\sigma_w^2}\sqrt{\frac{n}{2}}\leq\frac{\epsilon f(n)}{\sigma_w^2}\tag{4}$$

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 $\mathcal{V}_T(\mathbb{P}_w^n,\mathbb{P}_s^n)\leq\epsilon$, limiting the performance of Willie's detector.

Next we examine the probability $\mathbb{P}_e$ 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:

$$\begin{aligned}
\mathbb{P}_e &= \mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(\cup_{i=0,i\neq k}^{2^{nR}} E_i(\mathbf{c}(W_k))\right)\right] \\
&\leq \mathbb{E}_{\mathbf{c}(W_k)}\left[\sum_{i=0,i\neq k}^{2^{nR}} \mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right] \qquad (5)\\
&= \sum_{i=0,i\neq k}^{2^{nR}} \mathbb{E}_{\mathbf{c}(W_k)}\left[\mathbb{P}\left(E_i(\mathbf{c}(W_k))\right)\right] \qquad (6)
\end{aligned}$$

where $\mathbb{E}_X[\cdot]$ denotes the expectation operator over random variable X and (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 [12, 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 $Q(x) \leq \frac{1}{2}e^{-x^2/2}$ [13, Eq. (5)] and $P_f = \frac{cf(n)}{\sqrt{n}}$:

$$\begin{aligned}
\mathbb{E}_U\left[Q\left(\sqrt{\frac{P_f U}{2\sigma_b^2}}\right)\right] &\leq \mathbb{E}_U\left[\exp\left(-\frac{cf(n)U}{4\sqrt{n}\sigma_b^2}\right)\right] \\
&= \int_0^\infty \frac{e^{-\frac{cf(n)u}{4\sqrt{n}\sigma_b^2}-\frac{u}{2}} 2^{-\frac{n}{2}} u^{\frac{n}{2}-1}}{\Gamma(n/2)} du \qquad (7)\\
&= 2^{-n/2}\left(\frac{1}{2} + \frac{cf(n)}{4\sqrt{n}\sigma_b^2}\right)^{-n/2} \qquad (8)
\end{aligned}$$

where (8) is from the substitution $v = u\left(\frac{1}{2} + \frac{cf(n)}{4\sqrt{n}\sigma_b^2}\right)$ in (7) and the definition of the Gamma function $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$. Since $\frac{1}{2} + \frac{cf(n)}{4\sqrt{n}\sigma_b^2} = 2^{\log_2\left(\frac{1}{2}+\frac{cf(n)}{4\sqrt{n}\sigma_b^2}\right)}$:

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

for all i, and (6) becomes:

$$\mathbb{P}_e \leq 2^{nR-\frac{n}{2}\log_2\left(1+\frac{cf(n)}{2\sqrt{n}\sigma_b^2}\right)} \tag{9}$$

Since $f(n) = \omega(1/\sqrt{n})$, if rate $R = \frac{\rho}{2}\log_2\left(1 + \frac{cf(n)}{2\sqrt{n}\sigma_b^2}\right)$ for a constant $\rho < 1$, as n increases, the probability of Bob's decoding error averaged over all codebooks decays exponentially to zero and Bob obtains $nR = n\frac{\rho}{2}\log_2\left(1 + \frac{cf(n)}{2\sqrt{n}\sigma_b^2}\right)$ LPD bits in n channel uses. Since $\ln(1+x) \leq x$ with equality when $x = 0$, $nR \leq \frac{\sqrt{n}\rho cf(n)}{4\sigma_b^2\ln 2}$, approaching equality as n gets large. 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 [7, 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 are using is effective. In such a scenario, the construction of Theorem 1.2 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 1.2 (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 sum of the probabilities of detection errors* $\alpha + \beta \geq 1 - \epsilon$ *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 1.2, we introduce a variant of the Leibniz integral rule as a lemma:

Lemma 2 (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* $\int_{x_0}^{\infty} g(x)dx < \infty$*;*
4) *There is a function* $h(x)$ *such that* $|\frac{\partial f(x,a)}{\partial a}| \leq h(x)$ *and* $\int_{x_0}^{\infty} h(x)dx < \infty$.

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

The proof of Lemma 2 is available in [10, Ch. XIII.3]. We now prove Theorem 1.2.

Proof (Theorem 1.2): **Construction:** Alice encodes the input in blocks of length M bits into codewords of length n at the rate $R = M/n$ bits/symbol 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^{nR} codewords $\{\mathbf{c}(W_k), k = 1, 2, \ldots, 2^{nR}\}$ 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 1.1, 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:

$$\mathcal{D}(\mathbb{P}_w \| \mathbb{P}_s) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\frac{1}{2}\left(e^{-\frac{(x+a)^2}{2\sigma_w^2}} + e^{-\frac{(x-a)^2}{2\sigma_w^2}}\right)} dx \tag{10}$$

Since (10) is an even function, we assume $a \geq 0$.

While there is no closed-form expression for (10), its integrand is well-behaved, allowing the application of Lemma 1 to (10). The Taylor series expansion with respect to a around $a = 0$ can be performed using Lemma 2. We demonstrate that the conditions for Lemmas 1 and 2 hold in [14, Appendix B]. The zeroth through third order terms of the Taylor series expansion of (10) are zero, as is the fifth term. The fourth order term is:

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

Suppose Alice sets $a^2 \leq \frac{cf(n)}{\sqrt{n}}$, where c and $f(n)$ are defined as in Theorem 1.1. The sixth derivative of (10) with respect to a, is derived in [14, Appendix B], where we also show that it is continuous with respect to a and negative when evaluated at $a = 0$. Thus, there exists a neighborhood $[0, \mu]$ such that, for all $\xi \in [0, \mu]$, the remainder term $\frac{a^6}{6!} \times \left.\frac{\partial^6 \mathcal{D}(\mathbb{P}_w \| \mathbb{P}_s)}{\partial a^6}\right|_{a=\xi} \leq 0$. Then, for n large enough, we can apply Lemma 1 to upper-bound relative entropy with the fourth order term as follows:

$$\mathcal{V}_T(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \frac{a^2}{2\sigma_w^2}\sqrt{\frac{n}{2}} \leq \frac{\epsilon f(n)}{\sigma_w^2} \tag{11}$$

Since the power of Alice's symbol is $a^2 = P_f$, (11) is identical to (4) and Alice obtains the upper bound $\mathcal{V}_T(\mathbb{P}_w^n, \mathbb{P}_s^n) \leq \epsilon$, limiting the performance of Willie's detector.

Next let's examine the probability $\mathbb{P}_e$ of Bob's decoding error averaged over all possible codebooks. As in Theorem 1.1, 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 $\mathbb{P}_e \leq 2^{nR-n(1-\mathcal{H}(p_e))}$ [15], where $\mathcal{H}(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. We expand the analysis in [16, Section I.2.1] to characterize the rate R. We use Lemma 1 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 $p_e^{(UB)}$ is 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 $\mathcal{H}(p)$ is a monotonically increasing function on the interval $\left[0, \frac{1}{2}\right]$, $\mathcal{H}(p_e) \leq \mathcal{H}(p_e^{(UB)})$. The Taylor series expansion of $\mathcal{H}(p_e^{(UB)})$ with respect to a around $a = 0$ yields $\mathcal{H}(p_e^{(UB)}) = 1 - \frac{a^2}{\sigma_b^2\pi\ln 2} + \mathcal{O}(a^4)$. Substituting $a^2 = \frac{cf(n)}{\sqrt{n}}$, we obtain $\mathbb{P}_e \leq 2^{nR - \frac{\sqrt{n}cf(n)}{\sigma_b^2\pi\ln 2} + \mathcal{O}(1)}$. Since $f(n) = \omega(1/\sqrt{n})$, if rate $R = \frac{\rho c f(n)}{\sqrt{n}\sigma_b^2\pi\ln 2}$ bits/symbol for a constant $\rho < 1$, the probability of Bob's decoding error averaged over all codebooks decays exponentially to zero as n increases and Bob obtains $nR = o(\sqrt{n})$ bits in n channel uses, and $\mathcal{O}(\sqrt{n})$ bits in n channel uses if $f(n) = \hat{\sigma}_w^2$. ■

Remarks

Employing the best codebook: The proof of Theorem 1.2 guarantees Bob's decoding error performance averaged over all binary codebooks. Following the standard coding arguments [7, 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 1.2 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 [4], 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 1.2 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. The development of LPD communication with a shared secret either linear or sublinear in the message size is a subject of future research.

Relationship with Square Root Law in Steganography: The LPD communication problem is related to steganography. A comprehensive review of steganography is available in a book by Fridrich [2]. 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 [2, Ch. 13] [17]. This result was extended to Markov covertext [18] and was shown to either require a key linear in the size of the message [19] or encryption of the message prior to embedding [20].

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 [11, Ch. 2.6]: $\mathcal{D}(\mathbb{P}_0\|\mathbb{P}_1) = \frac{\delta^2}{2}\mathcal{J}(\theta)+\mathcal{O}(\delta^3)$, where $\mathcal{J}(\theta) = \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 f(x;\theta)dx$ 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 [21], [22].

In a typical steganography scenario with a passive warden, coding techniques similar to Hamming codes allow embedding of $\log(n)$ bits per changed symbol [2, Ch. 8], which make hiding $\mathcal{O}(\sqrt{n}\log n)$ bits in n symbols possible. However, due to the noise on the channel between Alice and Bob, and the resultant need for error correction, the LPD channel only allows $\mathcal{O}(\sqrt{n})$ bits to be transmitted in n channel uses, as we prove in the following section.

IV. Converse

Here, as in the proof of achievability, the channel between Alice and Bob is AWGN with power σ_b^2. Alice's objective is to transmit a message W_k that is $M = \omega(\sqrt{n})$ bits long to Bob in n channel uses with arbitrarily small probability of decoding error as n gets large, while limiting Willie's ability to detect her transmission. Alice encodes each message W_k arbitrarily into n symbols at the rate $R = M/n$ symbols/bit. For an upper bound on the reduction in entropy, the messages are chosen equiprobably.

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 2. *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 there exists a detector that Willie can use to detect her with arbitrarily low sum of error probabilities $\alpha + \beta$, or Bob cannot decode with arbitrarily low probability of error.*

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

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

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 α and β 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:

$$\mathbb{E}\left[S\right] = \sigma_w^2 \tag{12}$$
$$\mathrm{Var}\left[S\right] = \frac{2\sigma_w^4}{n} \tag{13}$$

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 $\mathbb{E}\left[y_i^2\right] = \sigma_w^2 + \left(f_i^{(k)}\right)^2$ and the variance is $\mathrm{Var}\left[y_i^2\right] = \mathbb{E}\left[y_i^4\right] - \left(\mathbb{E}\left[y_i^2\right]\right)^2 = 4\left(f_i^{(k)}\right)^2 \sigma_w^2 + 2\sigma_w^4$. 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:

$$\mathbb{E}\left[S\right] = \sigma_w^2 + P_k \tag{14}$$
$$\mathrm{Var}\left[S\right] = \frac{4P_k\sigma_w^2 + 2\sigma_w^4}{n} \tag{15}$$

The variance of Willie's test statistic (15) is computed by adding the variances conditioned on $\mathbf{c}(W_k)$ of the squared individual observations $\mathrm{Var}\left[y_i^2\right]$ (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 (12). 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 α^*, which is the probability that $S \geq \sigma_w^2 + t$ when H_0 is true. We bound it using (12) and (13) with Chebyshev's Inequality [7, Eq. (3.32)]:

$$\alpha = \mathbb{P}_0\left(S \geq \sigma_w^2 + t\right) \leq \mathbb{P}_0\left(|S - \sigma_w^2| \geq t\right) \leq \frac{2\sigma_w^4}{nt^2}$$

Thus, to obtain α^*, Willie sets $t = \frac{d}{\sqrt{n}}$, where $d = \frac{\sqrt{2}\sigma_w^2}{\sqrt{\alpha^*}}$ 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 $\beta^{(k)}$ is the probability that $S < \sigma_w^2 + t$, where $t = \frac{d}{\sqrt{n}}$. We bound $\beta^{(k)}$ using (14) and (15) with Chebyshev's Inequality:

$$\begin{aligned}\beta^{(k)} &= \mathbb{P}_1^{(k)}\left(S < \sigma_w^2 + t\right) \leq \mathbb{P}_1^{(k)}\left(\left|S - \sigma_w^2 - P_k\right| \geq P_k - t\right) \\ &\leq \frac{4P_k\sigma_w^2 + 2\sigma_w^4}{(\sqrt{n}P_k - d)^2} \qquad (16)\end{aligned}$$

If the average symbol power $P_k = \omega(1/\sqrt{n})$, $\lim_{n\to\infty} \beta^{(k)} = 0$. 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 (16) 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 the sum of the probabilities of error of Willie's statistical test by $\alpha + \beta \geq \zeta > 0$, 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 $\mathbb{P}_e$. The probability that a message from set $\mathcal{U}$ is sent is $\mathbb{P}(\mathcal{U}) = \gamma$, as all messages are equiprobable. We bound $\mathbb{P}_e = \mathbb{P}_e(\mathcal{U})\mathbb{P}(\mathcal{U}) + \mathbb{P}_e\left(\overline{\mathcal{U}}\right)\mathbb{P}\left(\overline{\mathcal{U}}\right) \geq \gamma\mathbb{P}_e(\mathcal{U})$, 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:

$$\mathbb{P}_e(\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{W\in\mathcal{U}} \mathbb{P}_e(\mathbf{c}(W)\text{ sent}) \tag{17}$$

where $\mathbb{P}_e(\mathbf{c}(W)\text{ sent})$ is the probability of error when codeword $\mathbf{c}(W)$ is transmitted, $|\cdot|$ denotes the set cardinality operator, and (17) holds because all messages are equiprobable.

When Bob uses the optimal decoder, $\mathbb{P}_e(\mathbf{c}(W)\text{ sent})$ 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 $W_j \neq W$:

$$\begin{aligned}\mathbb{P}_e(\mathbf{c}(W)\text{ sent}) &= \mathbb{P}\left(\cup_{j=1,W_j\neq W}^{2^{nR}} E_j\right)\\ &\geq \mathbb{P}\left(\cup_{W_j\in\mathcal{U}\setminus\{W\}} E_j\right) \triangleq \mathbb{P}_e^{(\mathcal{U})}\end{aligned} \tag{18}$$

Here the inequality in (18) is due to the observation that the sets in the second union are contained in the first. From the decoder perspective, this is due to the decrease in the decoding error probability if Bob knew that the message came from $\mathcal{U}$ (reducing the set of messages on which the decoder can err).

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 [7, 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 [7]. Rather, we explicitly use:

$$H(W|\hat{W}) \leq 1 + (\log_2|\mathcal{U}|)\mathbb{P}_e^{(\mathcal{U})} \tag{19}$$

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 $\gamma 2^{nR}$ and that, since each message is equiprobable, the entropy of a message W from $\mathcal{U}$ is $H(W) = \log_2|\mathcal{U}| =$

$\log_2 \gamma + nR$, we utilize (19) and carry out steps (9.38)–(9.53) in [7] to obtain:

$$\mathbb{P}_e^{(\mathcal{U})} \geq 1 - \frac{P_{\mathcal{U}}/2\sigma_b^2 + 1/n}{\frac{\log_2 \gamma}{n} + R} \tag{20}$$

Since Alice transmits $\omega(\sqrt{n})$ bits in n channel uses, her rate is $R = \omega(1/\sqrt{n})$ bits/symbol. However, $P_{\mathcal{U}} = O(1/\sqrt{n})$, and, as $n \to \infty$, $\mathbb{P}_e^{(\mathcal{U})}$ is bounded away from zero. Since $\gamma > 0$, $\mathbb{P}_e$ 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 (20), the probability of her message being successfully decoded by Bob is $\mathbb{P}_s^{(\mathcal{U})} = 1 - \mathbb{P}_e^{(\mathcal{U})} = \mathcal{O}\left(\frac{1}{\sqrt{n}R}\right)$ and the goodput is $G(n) = \gamma \mathbb{P}_s^{(\mathcal{U})} Rn = \mathcal{O}(\sqrt{n})$. 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.

V. Discussion

A. Relationship to Previous Work in Communications

The relationship of our work to steganography has already been discussed in the remark at the end of Section III. Here we relate our problem to other work in communication.

Spread Spectrum Communications: As wireless communication became prevalent, militaries sought methods to protect their signals from being detected by the enemy, leading to the development of spread spectrum communication. Spread spectrum communication provides an LPD capability as well as resistance to jamming by transmitting a signal that requires bandwidth W_M on a much wider bandwidth $W_s \gg W_M$, thereby reducing the power spectral density. Most spread spectrum results address the practical aspects of spread spectrum architectures and a comprehensive review [1] is available. We are not aware of any prior work studying the fundamental limits on the information that can be transmitted with low probability of detection using spread spectrum technology. However, we note that, while we present our result for narrowband channels, our analysis trivially translates to wideband channels as well: Alice can reliably transmit $\mathcal{O}(\sqrt{W_s n})$ LPD bits per n uses of a channel with bandwidth W_s. Thus, spread spectrum systems are also limited by the square root law.

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 [23]. Cheong and Hellman extended this result to Gaussian channels [24], and Csiszár and Körner generalized it to broadcast channels [25]. 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, recent succeeding work [26] claims that if the adversary and the legitimate receiver each has a *binary symmetric channel* (BSC) to the transmitter, with the adversary having a significantly noisier channel (i.e. a wire-tap BSC with positive secrecy rate), then the square-root law of LPD communication is achievable without the use of a secret codebook.

Anonymous communication: Our problem is related to that of anonymous communication [27], 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 [28]. 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.

Cognitive Radio: The LPD communication problem is also related to that of establishing a cognitive radio (CR) network [29]. An aspect of the CR problem is limiting the interference from the secondary users' radios to the primary users of the network. The LPD problem with a passive warden can be cast within this framework by having primary users only listen [30]. However, the properties of the secondary signal that allow smooth operation of the primary network are very different from those of an undetectable signal. While there is a lot of work on the former topic, we are not aware of work by the CR community on the latter issue.

B. *Impact of Adversary's* a priori *Knowledge of the Transmission State on Achievability*

The proofs of achievability (Theorems 1.1 and 1.2) in Section III 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 does not transmit (i.e. H_0 is true) with probability π_0 and that she transmits (i.e. H_0 is true) with probability $\pi_1 = 1-\pi_0$. Let $\mathbb{P}_e$ denote the probability that Willie's hypothesis test makes an error averaged over all observations. The following generalized version of Fact 1 then holds:

Fact 4 (Generalized Fact 1)**.** $\mathbb{P}_e \geq \min(\pi_0, \pi_1) - \max(\pi_0, \pi_1)\mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1)$

where, as in Section III, we 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 proof is in Appendix C. Thus, while Fact 4 demonstrates that additional information about the likelihood of Alice transmitting helps Willie, the square root law still holds via the bounds on the total variation distance $\mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1)$.

C. Mapping to a Continuous-time Channel

We employ a discrete-time model throughout the paper. However, while this is commonly assumed without loss of generality in standard communication theory, it is important to consider whether some aspect of the LPD problem has been missed 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 Section II and suits our demonstration of the fundamental limits to Alice's LPD channel capabilities. 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.

VI. Conclusion

Practitioners have always known that LPD communication requires the use of low power in order to blend in with the noise on the adversary's channel. However, the specific requirements for achieving LPD communication and resulting achievable performance have not been analyzed prior to this work. We quantified the conditions for existence and maintenance of an LPD channel by proving that the LPD communication is subject to a square root law in that the number of LPD bits that can be transmitted in n channel uses is $\mathcal{O}(\sqrt{n})$.

There are a number of avenues for future research. The key efficiency and, specifically, LPD communication with a secret linear in the message length is an open theoretical research problem. Practical network settings and the implications of the square root law on the LPD transmission of packets under additional constraints such as delay should be analyzed. The impact of dynamism in the network should also be examined, as well as more realistic scenarios that include channel artifacts such as fading and interference from other nodes. One may be able to improve LPD communication by employing nodes that perform friendly jamming. Eventually, we would like to answer this fundamental question: is it possible to establish and maintain a "shadow" wireless network in the presence of both active and passive wardens?

Appendix

A. Using an $\mathcal{O}(\sqrt{n}\log n)$-bit secret

Here we demonstrate how Alice and Bob can construct a binary coding scheme that, on average, requires an $\mathcal{O}(\sqrt{n}\log n)$-bit secret. This is done 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 η and note that $\mathbb{E}\left[\eta\right] = \tau n$. Alice and Bob then use the best public binary codebook with codewords of length η on these selected η symbol periods. They also generate and share a random binary vector $\mathbf{k}$ where $p_{\mathbf{K}}(\mathbf{k}) = \prod_{i=1}^{\eta} p_K(k_i)$ 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,

$$\mathcal{D}(\mathbb{P}_w\|\mathbb{P}_s) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln \frac{e^{-\frac{x^2}{2\sigma_w^2}}/\sqrt{2\pi}\sigma_w}{\frac{(1-\tau)e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} + \frac{\tau}{2}\left(\frac{e^{-\frac{(x+a)^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} + \frac{e^{-\frac{(x-a)^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w}\right)} dx \tag{21}$$

There is no closed-form expression for (21), but we can upper-bound it using Lemma 1. The Taylor series expansion with respect to a around $a=0$ can be done using Lemma 2, with conditions for Lemmas 1 and 2 proven similarly as in Theorem 1.2. This yields the following bound:

$$\mathcal{V}_T(\mathbb{P}_w^n, \mathbb{P}_s^n) \le \frac{\tau a^2}{2\sigma_w^2}\sqrt{\frac{n}{2}} \tag{22}$$

The only difference in (22) from (11) is τ in the numerator. Thus, if Alice sets the product $\tau a^2 \le \frac{cf(n)}{\sqrt{n}}$, 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 $\mathcal{O}(\sqrt{n})$ bits in n LPD channel uses employing this coding scheme.

B. $\mathcal{D}(\mathbb{P}_w\|\mathbb{P}_s)$ in the proof of Theorem 1.2 meets the conditions of Lemmas 1 and 2

Re-arranging the terms of (10) results in the following expression:

$$\mathcal{D}(\mathbb{P}_w\|\mathbb{P}_s) = \frac{a^2}{2\sigma_w^2} - \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln\cosh\left(\frac{ax}{\sigma_w^2}\right) dx \tag{23}$$

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 (23), specifically on its integrand:

$$K(x,a) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \ln\cosh\left(\frac{ax}{\sigma_w^2}\right) \tag{24}$$

Due to 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 $\int_{-\infty}^{\infty} g(x)dx = \sqrt{\frac{2b}{\pi\sigma_w^2}} < \infty$.

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

$$\text{odd } i \ : \frac{\partial^i K(x,a)}{\partial a^i} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{x^i}{\sigma_w^{2i}} \tanh\left(\frac{ax}{\sigma_w^2}\right) \sum_{k=1}^{(i-1)/2} c_{i,k} \operatorname{sech}^{2k}\left(\frac{ax}{\sigma_w^2}\right) \tag{25}$$

$$\text{even } i : \ \frac{\partial^i K(x,a)}{\partial a^i} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{x^i}{\sigma_w^{2i}} \sum_{k=1}^{i/2} c_{i,k} \operatorname{sech}^{2k}\left(\frac{ax}{\sigma_w^2}\right) \tag{26}$$

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 the "empty" sum $\sum_{k=1}^{0} c_{i,k} = 1$. The first six derivatives of $K(x,a)$ with respect to a are as follows:

$$\frac{\partial K(x,a)}{\partial a} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{x}{\sigma_w^2} \tanh\left(\frac{ax}{\sigma_w^2}\right) \tag{27}$$

$$\frac{\partial^2 K(x,a)}{\partial a^2} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{x^2}{\sigma_w^4} \operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right) \tag{28}$$

$$\frac{\partial^3 K(x,a)}{\partial a^3} = -\frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{2x^3}{\sigma_w^6} \operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right) \tanh\left(\frac{ax}{\sigma_w^2}\right) \tag{29}$$

$$\frac{\partial^4 K(x,a)}{\partial a^4} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{2x^4}{\sigma_w^8} \left(2\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right) - 3\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{30}$$

$$\frac{\partial^5 K(x,a)}{\partial a^5} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{8x^5 \tanh\left(\frac{ax}{\sigma_w^2}\right)}{\sigma_w^{10}} \left(3\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right) - \operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{31}$$

$$\frac{\partial^6 K(x,a)}{\partial a^6} = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{8x^6}{\sigma_w^{12}} \left(15\operatorname{sech}^6\left(\frac{ax}{\sigma_w^2}\right) - 15\operatorname{sech}^4\left(\frac{ax}{\sigma_w^2}\right) + 2\operatorname{sech}^2\left(\frac{ax}{\sigma_w^2}\right)\right) \tag{32}$$

Clearly, $K(x,a)$ and its derivatives are continuous, satisfying conditions 1 and 2 of Lemma 2. 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

$$h_i(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|^i}{\sigma_w^{2i}} \sum_{k=1}^{\lfloor i/2 \rfloor} |c_{i,k}| \tag{33}$$

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

$$\left|\frac{\partial K(x,a)}{\partial a}\right| \leq h_1(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|}{\sigma_w^2} \tag{34}$$

$$\left|\frac{\partial^2 K(x,a)}{\partial a^2}\right| \leq h_2(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{|x|^2}{\sigma_w^4} \tag{35}$$

$$\left|\frac{\partial^3 K(x,a)}{\partial a^3}\right| \leq h_3(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{2|x|^3}{\sigma_w^6} \tag{36}$$

$$\left|\frac{\partial^4 K(x,a)}{\partial a^4}\right| \leq h_4(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{10|x|^4}{\sigma_w^8} \tag{37}$$

$$\left|\frac{\partial^5 K(x,a)}{\partial a^5}\right| \leq h_5(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{32|x|^5}{\sigma_w^{10}} \tag{38}$$

$$\left|\frac{\partial^6 K(x,a)}{\partial a^6}\right| \leq h_6(x) = \frac{e^{-\frac{x^2}{2\sigma_w^2}}}{\sqrt{2\pi}\sigma_w} \frac{256|x|^6}{\sigma_w^{12}} \tag{39}$$

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 2 are met by the integrand of (10) and the integrand's derivatives.

The use of Lemma 1 is conditional on the integrals over x of $K(x,a)$ and its derivatives in (25) and (26) 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 $\int_{-\infty}^{\infty} K(x,a)dx$ is continuous as follows:

$$\lim_{a\to a_0} \int_{-\infty}^{\infty} K(x,a)dx = \int_{-\infty}^{\infty} \lim_{a\to a_0} K(x,a)dx = \int_{-\infty}^{\infty} K(x,a_0)dx \tag{40}$$

where the first equality is due to 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 due to 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 1.

C. Proof of the generalized version of Fact 1

Proof (Fact 4): 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 $\mathbb{P}_c$ of Willie's optimal test being correct, averaged over all observations, is as follows:

$$\mathbb{P}_c = \int_{\mathcal{X}} \max(p(H_0|x), p(H_1|x))p(x)dx \tag{41}$$

$$= \int_{\mathcal{X}} \max(\pi_0 p_0(x), \pi_1 p_1(x))dx \tag{42}$$

where $\mathcal{X}$ is the support of $p_0(x)$ and $p_1(x)$, and (42) follows from Bayes' theorem. Let $\mathbb{P}_e = 1 - \mathbb{P}_c = 1 - \int_{\mathcal{X}} \max(\pi_0 p_0(x), \pi_1 p_1(x))dx$ denote the error probability of Willie's optimal test. Now, since $\max(a,b) = \frac{a+b+|a-b|}{2}$, $\mathbb{P}_e$ can be expressed as follows:

$$\mathbb{P}_e = 1 - \frac{1}{2}\left(\pi_0 \int_{\mathcal{X}} p_0(x)dx + \pi_1 \int_{\mathcal{X}} p_1(x)dx\right) - \frac{1}{2}\int_{\mathcal{X}} |\pi_0 p_0(x) - \pi_1 p_1(x)|dx \tag{43}$$

$$= \frac{1}{2} - \frac{1}{2}\|\pi_0 p_0(x) - \pi_1 p_1(x)\|_1 \tag{44}$$

where (44) is due to the probability densities integrating to one over their supports in the first two integrals of (43), $\pi_0 + \pi_1 = 1$, and the last integral in (43) being the $\mathcal{L}_1$ norm. We can lower-bound (44) using the triangle inequality for the $\mathcal{L}_1$ norm:

$$\mathbb{P}_e \geq \frac{1}{2} - \frac{1}{2}\left(\|\pi_0 p_0(x) - \pi_0 p_1(x)\|_1 + \|\pi_0 p_1(x) - \pi_1 p_1(x)\|_1\right) \tag{45}$$

$$= \frac{1}{2} - \frac{|\pi_0 - \pi_1|}{2} - \frac{\pi_0}{2}\|p_0(x) - p_1(x)\|_1 \tag{46}$$

where (46) 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:

$$\mathbb{P}_e \geq \frac{1}{2} - \frac{1}{2}\left(\|\pi_1 p_1(x) - \pi_1 p_0(x)\|_1 + \|\pi_1 p_0(x) - \pi_0 p_0(x)\|_1\right) \tag{47}$$

$$= \frac{1}{2} - \frac{|\pi_0 - \pi_1|}{2} - \frac{\pi_1}{2}\|p_0(x) - p_1(x)\|_1 \tag{48}$$

By Definition 1, $\frac{1}{2}\|p_0(x) - p_1(x)\|_1 = \mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1)$. Since $\min(a,b) = \frac{a+b-|a-b|}{2}$, we can combine (46) and (48) to yield

$$\mathbb{P}_e \geq \min(\pi_0, \pi_1) - \max(\pi_0, \pi_1)\mathcal{V}_T(\mathbb{P}_0, \mathbb{P}_1) \tag{49}$$

which completes the proof. ■

References

[1] M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*. McGraw-Hill, 1994.

[2] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed. New York, NY, USA: Cambridge University Press, 2009.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. Cambridge, Massachusetts: MIT Press, 2001.

[4] C. E. Shannon, “Communication theory of security,” *Bell System Technical Journal*, vol. 28, pp. 656–715, 1949.

[5] A. J. Menezes, S. A. Vanstone, and P. C. V. Oorschot, *Handbook of Applied Cryptography*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 1996.

[6] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York: Springer, 2005.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Hoboken, NJ, 2002.

[8] C. Cachin, “An information-theoretic model for steganography,” *Information and Computation*, vol. 192, no. 1, pp. 41–56, 2004.

[9] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. New York: John Wiley & Sons, Inc., 2001.

[10] S. Lang, *Undergraduate Analysis*, 2nd ed. New York, NY: Springer-Verlag, 1997.

[11] S. Kullback, *Information Theory and Statistics*. New York, NY: Wiley, 1959.

[12] U. Madhow, *Fundamentals of Digital Communication*. Cambridge, UK: Cambridge University Press, 2008.

[13] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 840–845, Jul. 2003.

[14] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on AWGN channels," University of Massachusetts, Tech. Rep. UM-CS-2012-003, 2012.

[15] A. Barg and G. D. Forney, Jr., "Random codes: minimum distances and error exponents," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2568–2573, Sep. 2002.

[16] E. E. Majani, "A model for the study of very noisy channels, and applications," Ph.D. dissertation, California Institute of Technology, 1988.

[17] A. D. Ker, "A capacity result for batch steganography," *IEEE Signal Processing Letters*, vol. 14, no. 8, pp. 525–528, Aug. 2007.

[18] T. Filler, A. D. Ker, and J. Fridrich, "The square root law of steganographic capacity for markov covers," *Media Forensics and Security*, vol. 7254, no. 1, 2009.

[19] A. D. Ker, "The square root law requires a linear key," in *Proceedings of the 11th ACM workshop on Multimedia and security*, ser. MM&Sec '09. New York, NY, USA: ACM, 2009, pp. 85–92.

[20] ——, "The square root law does not require a linear key," in *Proceedings of the 12th ACM workshop on Multimedia and security*, ser. MM&Sec '10. New York, NY, USA: ACM, 2010, pp. 213–224.

[21] T. Filler and J. Fridrich, "Fisher information determines capacity of ϵ-secure steganography," in *Information Hiding*, ser. Lecture Notes in Computer Science, S. Katzenbeisser and A.-R. Sadeghi, Eds. Springer Berlin / Heidelberg, 2009, vol. 5806, pp. 31–47.

[22] A. D. Ker, "Estimating steganographic fisher information in real images," in *Information Hiding*, ser. Lecture Notes in Computer Science, S. Katzenbeisser and A.-R. Sadeghi, Eds. Springer Berlin / Heidelberg, 2009, vol. 5806, pp. 73–88.

[23] A. D. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, pp. 1355–1387, 1975.

[24] S. Leung-Yan-Cheong and M. Hellman, "The gaussian wire-tap channel," *IEEE Transactions on Information Theory*, vol. 24, no. 4, pp. 451–456, Jul. 1978.

[25] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 339–348, May 1978.

[26] P. H. Che, M. Bakshi, and S. Jaggi, "Reliable deniable communication: Hiding messages in noise," arXiv:1304.6693, 2013.

[27] G. Danezis and C. Diaz, "A survey of anonymous communication channels," Microsoft Research, Tech. Rep. TR-2008-35, 2008.

[28] S. Marano, V. Matta, T. He, and L. Tong, “The embedding capacity of information flows under renewal traffic,” *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1724–1739, 2013.

[29] T. Yucek and H. Arslan, “A survey of spectrum sensing algorithms for cognitive radio applications,” *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 116–130, First Qtr 2009.

[30] W. Ren, A. Swami, and Q. Zhao, “Coexistence, connectivity and delay in heterogeneous networks,” in *Proceedings of the 27th Army Science Conference*, 2011.