Learning Adversary's Actions for Secret Communication

Mehrdad Tahmasbi and Matthieu R. Bloch
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: mtahmasbi3@gatech.edu, matthieu.bloch@ece.gatech.edu

Aylin Yener
Electrical Engineering Department
The Pennsylvania State University
University Park, PA 16802
Email: yener@engr.psu.edu

***Abstract*—We analyze the problem of secure communication over a wiretap channel with an active adversary, in which the legitimate transmitter has the opportunity to sense and learn the adversary's actions. Specifically, the adversary has the ability to switch between two channels and to observe the corresponding output at every channel use; the encoder, however, has *causal* access to observations impacted by adversary's actions. We develop a joint learning/transmission scheme in which the legitimate users learn and adapt to the adversary's actions. For some channel models, we show that the achievable rates, which we define precisely, are arbitrarily close to those obtained *with hindsight*, had the transmitter known the actions ahead of time. This suggests that there is much to exploit and gain in physical-layer security by monitoring the environment.**

I. Introduction

While the pioneering work of Wyner [1] on the wiretap channel has provided the foundation for advances in information-theoretic security, the limits of the original model are now well recognized. In particular, several approaches have been developed to incorporate *active* attacks into the model and better capture some of the practical threats faced, e.g., in wireless systems. For instance, the wiretap channel Type II and its extensions [2]–[4] allow one to model the ability of an adversary to actively choose the best subset of observations. Another common approach is an adversarial model in which the adversary may not only observe the communication but also actively modify or jam the transmitted signals [5], [6]. The combination of eavesdropping and jamming attacks is also captured in arbitrarily varying wiretap channel models, in which both main and eavesdropper's channels depend on states under complete control of the adversary [7]–[10].

Despite notable success, the theoretical frameworks developed so far have a somewhat limited scope of application. This happens in large part because the models include assumptions that are either extremely optimistic with regards to what can be known about the adversary or overly pessimistic by endowing over-powerful abilities to the attacker. The rationale behind the present paper is that there might be a middle ground to develop adversarial yet realistic models. More precisely, we suggest that, although an adversary can potentially control channels, its actions are likely to come at a cost, i.e., the modification may induce some physical effects in the environment that can be detected by other parties; hence, legitimate parties may have the ability to *learn* the adversary's actions and accordingly adapt their coding scheme.

This work is supported in parts by NSF grants CCF1320298, CCF1527074, and CCF1319338.

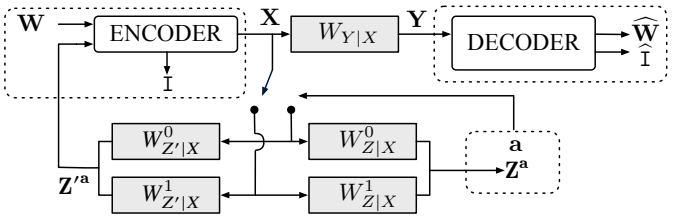

Fig. 1. Problem setup

As a first step in this direction, we study here a wiretap channel model in which an active adversary is able to improve its detection by selecting one of two channels at every channel use; however, the transmitter monitors the environment and *causally* receives a signal correlated to the adversary's observations. This consequently allows the legitimate parties to simultaneously "explore" the adversary's behavior and "exploit" it for secrecy. Our main result is that, for some channels, the legitimate parties achieve the same secrecy rates that they would have achieved *with hindsight*, had they known the attacker's actions non-causally. This result is close in spirit to similar ones in the context of multi-arm bandit problems [11]: without knowing the adversary's actions a priori, one can simultaneously exploit and explore to develop an asymptotically optimal strategy.

The remainder of the paper is organized as follows. In Section II, we introduce the channel model under investigation. In Section III, we develop the achievability proof; we omit the proof of the converse because of space constraints.

II. Problem Formulation and Main Results

We consider the channel model illustrated in Figure 1, in which two legitimate parties attempt to communicate over n channel uses of a Discrete Memoryless Channel (DMC) $(\mathcal{X}, W_{Y|X}, \mathcal{Y})$ in the presence of an adversary that observes the output of either one of DMCs $(\mathcal{X}, W^0_{Z|X}, \mathcal{Z})$ and $(\mathcal{X}, W^1_{Z|X}, \mathcal{Z})$. The adversary is allowed to actively and arbitrarily choose its channel at every channel use, which

we model with a sequence of actions $\mathbf{a} = (a_1, \cdots, a_n) \in \{0,1\}^n$; at every channel use $i \in [\![1, n]\!]$, the adversary's channel is therefore $W^{a_i}_{Z|X}$ and its corresponding output is denoted $Z_i^{a_i}$. We assume that the transmitter monitors the effect of the adversary's actions, which we model as *strictly causal* observations at the output of either one of two DMCs $(\mathcal{X}, W^0_{Z'|X}, \mathcal{Z}')$ and $(\mathcal{X}, W^1_{Z'|X}, \mathcal{Z}')$ chosen according to a_i for the channel use i. We assume that all channel outputs are conditionally independent given the input and that channel statistics are known to all parties. The transmitted sequence of input symbols is denoted $\mathbf{X} \triangleq (X_1, \cdots, X_n)$, while the corresponding observations of the receiver and adversary are $\mathbf{Y} \triangleq (Y_1, \cdots, Y_n)$ and $\mathbf{Z^a} \triangleq (Z_1^{a_1}, \cdots, Z_n^{a_n})$, respectively, and the monitored sequence is $\mathbf{Z'^a} \triangleq (Z_1'^{a_1}, \cdots, Z_n'^{a_n})$.

Formally, a code for this channel model operates as follows. Unlike traditional wiretap channel models, the number of message bits is not known at the beginning of transmission and potentially depends on the adversary's actions. Therefore, it is convenient to assume that the transmitter has access to K uniformly distributed bits $\mathbf{W} \triangleq (W_1, \cdots, W_K)$, with $K \leqslant n \max_{p_X} \mathbb{I}(X;Y)$ to be precisely defined later, and that only the first $\mathtt{I}$ bits will be transmitted.[1] The encoder consists of n possibly stochastic functions $\mathbf{f} = (f_1, \cdots, f_n)$, where $f_i : \mathcal{Z}'^{i-1} \times \{0,1\}^K \to \mathcal{X}$ outputs a symbol for the transmission over the channel. The total number of transmitted bits $\mathtt{I} : \mathcal{Z}'^n \to [\![0, K]\!]$ is a function of the transmitter's observations and can be determined after the n^{th} transmission. The decoder is a function $\phi : \mathcal{Y}^n \to \{0,1\}^K$, which allows the receiver to form an estimate $(\widehat{W_1}, \cdots, \widehat{W_K}) = \phi(\mathbf{Y})$ of the transmitted bits. The receiver is *not* required to reliably decode all bits, and therefore, we assume that there exists a function $\widehat{\mathtt{I}} : \mathcal{Y}^n \to [\![0, K]\!]$ that estimates the number of bits actually transmitted, so that the reliability is measured with a probability of error defined as

$$P_e(\mathbf{a}) \triangleq \mathbb{P}\Big(\widehat{\mathtt{I}} \neq \mathtt{I} \text{ or } \exists k \in [\![1, \widehat{\mathtt{I}}]\!] : \ \widehat{W_k} \neq W_k | \mathbf{a}\Big). \tag{1}$$

The functions $\mathbf{f}$, ϕ, $\mathtt{I}$, and $\widehat{\mathtt{I}}$ are assumed to be publicly known and secrecy is measured in terms of the average mutual information between the message bits $\mathbf{W}$ and the observations $\mathbf{Z^a}$ as

$$\mathtt{S}(\mathbf{a}) \triangleq \mathbb{I}(\mathbf{W}; \mathbf{Z^a} | \mathbf{a}). \tag{2}$$

Throughout the paper, we measure the information in bits and $\log(x)$ should be understood to be base 2; we use $\ln(x)$ for the logarithm base e. The rate of the code is a function of the adversary's actions and is a random variable defined as $\mathtt{R}(\mathbf{a}) \triangleq \frac{\widehat{\mathtt{I}}}{n}$. The quadruple $(\mathbf{f}, \phi, \mathtt{I}, \widehat{\mathtt{I}})$ defines a code $\mathcal{C}$.

Definition 1. *For a fixed sequence $\{\mathbf{a}_n \in \{0,1\}^n\}_{n \geqslant 1}$, we say that a sequence of codes $\{\mathcal{C}_n\}_{n \geqslant 1}$ achieves a rate R, if and only if , we have*

$$\lim_{n \to \infty} \mathtt{S}(\mathbf{a}_n) = 0, \tag{3}$$

[1]Despite conceptual similarities with layered secrecy coding [12], our problem formulation is different for technical reasons.

$$\lim_{n \to \infty} P_e(\mathbf{a}_n) = 0, \tag{4}$$

and for every $\epsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}(\mathtt{R}(\mathbf{a}_n) \leqslant R - \epsilon) = 0. \tag{5}$$

Remark 1. *The technical assumptions behind our model have concrete operational significance. Since* $\mathtt{I}$ *is publicly known, no secrecy is conveyed through the* number *of secret bits. Since only the transmitter monitors the environment, the receiver does not benefit from another channel observation that could potentially increase its reliability. Finally, since channel outputs are conditionally independent given the input, the transmitter only obtains information about the adversary's actions and not about the receiver or adversary's observations. These assumptions are not crucial in our achievability proof, but they are needed in the converse.*

It will be useful to define the following quantities.

Definition 2. *For $\alpha \in [0,1]$, define*

$$C_l(\alpha) \triangleq \sup I(U;Y) - \alpha I(U;Z^1) - (1-\alpha) I(U;Z^0) \tag{6}$$

where the supremum is taken over all distributions with $U - X - YZ^0Z^1Z'^0Z'^1$, $P_{Z'^1} \neq P_{Z'^0}$, and $I(U;Y) > \max[I(U,Z^0), I(U;Z^1), I(U;Z'^0), I(U;Z'^1)]$, and

$$C_u(\alpha) \triangleq \max_{U_0 U_1 - X - Y Z^0 Z^1} \alpha(I(U_1;Y) - I(U_1;Z^1)) - (1-\alpha)(I(U_0;Y) - I(U_0;Z^0)). \tag{7}$$

$C_l(\alpha)$ is the capacity of the wiretap channel analyzed in [3], in which the attacker is known to use channel $(\mathcal{X}, W^1_{Z|X}, \mathcal{Z})$ a fraction α of the time. On the other hand, $C_u(\alpha)$ represents the capacity of the wiretap channel in which the attacker uses the channel $(\mathcal{X}, W^1_{Z|X}, \mathcal{Z})$ a fraction α of the time and the uses are known non-causally. Our main results are then the following.

Theorem 1 (Converse)**.** *For a fixed sequence $\{\mathbf{a}_n \in \{0,1\}^n\}_{n \geqslant 1}$, if a sequence of codes $\{\mathcal{C}_n\}_{n \geqslant 1}$ achieves a rate R, then*

$$R \leqslant C_u\left(\lim_{n \to \infty} \frac{wt(\mathbf{a}_n)}{n}\right), \tag{8}$$

provided that $\lim_{n\to\infty} \frac{wt(\mathbf{a}_n)}{n}$ exists.

Proof. Omitted due to space restrictions. □

Theorem 2 (Achievability)**.** *For any $\epsilon > 0$, there exists a sequence of codes $\{\mathcal{C}_n\}_{n \geqslant 1}$ such that for every sequence $\{\mathbf{a}_n \in \{0,1\}^n\}_{n \geqslant 1}$, the rate*

$$C_l\left(\lim_{n \to \infty} \frac{wt(\mathbf{a}_n)}{n}\right) - \epsilon, \tag{9}$$

is achieved provided that $\lim_{n\to\infty} \frac{wt(\mathbf{a}_n)}{n}$ exists.

We prove Theorem 2 in Section III. Note that the main contribution of the theorem is in guaranteeing the existence of codes having good performance for *all* choices of $\mathbf{a}$. If

the maximum weight of the actions sequence $|\{i : a_i = 1\}|$ were known, the achievability would follow from [3]. Most importantly, we have the following corollary.

Corollary 1. *If the adversary's channels are degraded with respect to (w.r.t.) the main channel, if the capacity of all channels are obtained for the same capacity achieving input distribution, and if for the distribution that maximizes* (7) *there is no neighborhood in which $P_{Z'^0} = P_{Z'^1}$ or $I(U;Y) \leqslant \max[I(U,Z^0), I(U;Z^1), I(U;Z'^0), I(U;Z'^1)]$, then $C_l(\alpha) = C_u(\alpha)$ for all values of α.*

The condition $P_{Z'^0} \neq P_{Z'^1}$ expresses the idea that the distribution of the signal observed by the transmitter must be different to allow him to learn the actions. The condition $I(U;Y) > \max[I(U,Z^0), I(U;Z^1), I(U;Z'^0), I(U;Z'^1)]$ ensures that, irrespective of the channel used by the adversary, the secrecy capacity remains positive, and we can use channel resolvability results for the observations of the transmitter; this restriction is purely technical and can be removed, but it simplifies the proof and allows us to focus on the main conceptual aspects. Under these conditions, our results state that the achievable secrecy rates are arbitrarily close to the secrecy capacity *with hindsight*. We emphasize that there is little conceptual change when considering more than two adversarial channels, and we have mainly restricted the model for clarity.

The coding scheme behind the result of Theorem 2 is developed in Section III. At a high level, the idea is to perform a layered transmission of secret key bits. By transmitting layered key bits instead of message bits, the transmitter is able to *defer* its decision regarding the secrecy of the transmitted bits. Provided a correct decision is made, the extracted key can be used as a one-time-pad for the protection of subsequent messages, and the process continues iteratively from one block to the next.

III. Achievability Proof

A. One-shot results for regular wiretap channel

In this section, we develop basic tools for the one-shot analysis of a regular wiretap channel $(\mathcal{X}, W_{Y|X}, W_{Z|X}, \mathcal{Y}, \mathcal{Z})$, in which there exists only one eavesdropper's channel. We derive a super-exponential bound for the probability that a randomly chosen code is not secure for a wiretap channel; similar bounds have already been used to prove the achievability for the wiretap channel in [4].

Definition 3. *Consider a DMC $(\mathcal{X}, W_{Y|X}, \mathcal{Y})$ and a distribution P_X on $\mathcal{X}$. We define*

$$P_Y(y) \triangleq \sum_x P_X(x) W_{Y|X}(y|x), \tag{10}$$

$$I(P_X, W_{Y|X}) \triangleq I(X;Y), \tag{11}$$

$$F_{XY}(\gamma) \triangleq \mathbb{P}_{W_{Y|X}P_X}\left(\log \frac{W_{Y|X}(Y|X)}{P_Y(Y)} \leqslant \gamma\right), \tag{12}$$

$$\overline{F}_{XY}(\gamma) = 1 - F_{XY}(\gamma). \tag{13}$$

Furthermore, if A is any random variable, we define $\mu_A \triangleq \min_{a:\mathbb{P}(A=a)>0} \mathbb{P}(A=a)$.

Definition 4 (One-shot Wiretap Code)**.** *Given a wiretap channel $(\mathcal{X}, W_{Y|X}, W_{Z|X}, \mathcal{Y}, \mathcal{Z})$ and K bits $(W_1, \cdots, W_K)$, a code $\mathcal{C} = (f, \phi)$ consists of an encoder $f : \{0,1\}^K \to \mathcal{X}$ and a decoder $\phi : \mathcal{Y} \to \{0,1\}^K$. Upon receiving symbol y, the receiver estimates the bits as $(\widehat{W}_1, \cdots, \widehat{W}_K) \triangleq \phi(y)$. If $(W_1, \cdots, W_K)$ has uniform distribution on $\{0,1\}^K$, we define the probability of error as $P_e \triangleq \mathbb{P}\Big((\widehat{W}_1, \cdots, \widehat{W}_K) \neq (W_1, \cdots, W_K)\Big)$ and the secrecy for the first p bits as $I(W_1, \cdots W_p; Z)$.*

Note that the encoder defined in Definition 4 is deterministic; indeed, the randomness in bits $W_{p+1}, \cdots, W_K$ is used to secure the first p bits $W_1, \cdots, W_p$. For simplicity, we define $x_{w_1,\cdots,w_K} \triangleq f(w_1, \cdots, w_K)$. The next technical lemma characterizes the performance of randomly generated codes.

Lemma 1. *Let $(\mathcal{X}, W_{Y|X}, W_{Z|X}, \mathcal{Y}, \mathcal{Z})$ be a wiretap channel and P_X be a distribution on $\mathcal{X}$. Assume that $\{X_{\boldsymbol{w}} : \boldsymbol{w} \in \{0,1\}^K\}$ are independent and identically distributed (i.i.d.) with distribution P_X. Then, for every γ_1 and γ_2, we have*

$$\mathbb{E}(P_e) \leqslant F_{XY}(\gamma_1) + 2^{K-\gamma_1}, \tag{14}$$

and

$$\mathbb{E}(I(W_1, \cdots W_p; Z)) \leqslant 2^{\gamma_2 - K + p + 1} + \overline{F}_{XZ}(\gamma_2) \log\left(\frac{1}{\mu_Z} + 1\right). \tag{15}$$

We also have

$$\mathbb{P}\big(P_e \geqslant \lambda \left(F_{XY}(\gamma_1) + 2^{K-\gamma_1}\right)\big) \leqslant \frac{1}{\lambda}, \tag{16}$$

and

$$\mathbb{P}(I(W_1, \cdots W_p; Z) \geqslant \mathbb{E}(I(W_1, \cdots W_p; Z)) + \lambda_2) \leqslant 2^{-\frac{2^{p+1}\lambda_2^2}{\log^2 \frac{1}{\mu_Z}}}. \tag{17}$$

Proof. Omitted due to space restrictions. □

Lemma 1 allows us to analyze the secrecy when the transmission happens over a class of compound wiretap channels indexed by a set $\mathcal{I}$ and with the same input and output alphabet. Here, the goal is to find a code such that, if the channel corresponding to the index $i \in \mathcal{I}$ is used, the first $p_i \in [\![0, K]\!]$ bits are securely transmitted.

Lemma 2. *Consider a family of wiretap channels $\{(\mathcal{X}, W_{Y|X}, W^i_{Z|X}, \mathcal{Y}, \mathcal{Z})\}_{i\in\mathcal{I}}$ with the same main channel and different adversarial channels. Let P_X be a distribution on $\mathcal{X}$. If $(W_1, \cdots, W_K)$ are K bits and $\{X_{\boldsymbol{w}} : \boldsymbol{w} \in \{0,1\}^K\}$ are i.i.d. with distribution P_X, then we have*

$$\mathbb{P}\big(P_e \geqslant \lambda_1 \left(F_{XY}(\gamma_1) + 2^{K-\gamma_1}\right)\big) \leqslant \frac{1}{\lambda_1}, \tag{18}$$

and

$$\mathbb{P}(\exists i \in \mathcal{I} :\ I_i \geqslant \mathbb{E}(I_i) + \lambda_2) \leqslant |\mathcal{I}| 2^{-\frac{2^{p_{\min}+1}\lambda_2^2}{\log^2 \frac{1}{\mu_Z}}}, \tag{19}$$

where $I_i \triangleq I(W_1, \cdots, W_{p_i}; Z)$ and $p_{\min} \triangleq \min_{i\in\mathcal{I}} p_i$.

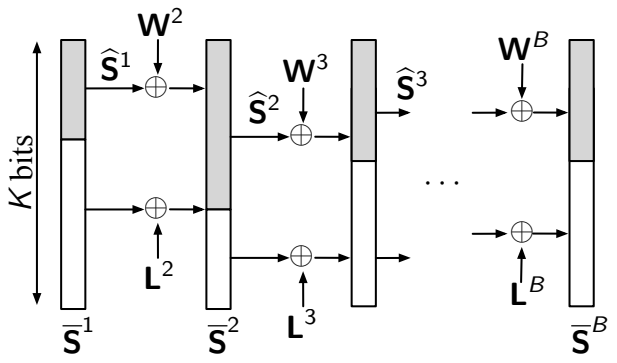

Fig. 2. Principle of the coding scheme. Gray bits represent the bits that are know to be secret at the end of the block transmission and are used to one-time pad the next message.

Proof. Since the probability of error is the same for all channels, (18) follows from (14). Furthermore, (17) and the union bound imply (19). □

B. Multiple uses of two wiretap channels

We now revert back to the setup introduced in Section II. We also fix a distribution P_X on $\mathcal{X}$ and without loss of generality assume that $I(P_X, W^1_{Z|X}) \geqslant I(P_X, W^0_{Z|X})$. For simplicity, we define for any $\alpha \in [0,1]$

$$\mathbb{I}_Y \triangleq I(P_X, W_{Y|X}), \quad \mathbb{I}^0_Z \triangleq I(P_X, W^0_{Z|X}), \tag{20}$$

$$\mathbb{I}^1_Z \triangleq I(P_X, W^1_{Z|X}), \quad \mathbb{I}^\alpha_Z \triangleq \alpha \mathbb{I}^1_Z + (1-\alpha)\mathbb{I}^0_Z. \tag{21}$$

Lemma 3. *For any $\zeta \in]0, \frac{\mathbb{I}_Y - \mathbb{I}^1_Z}{\mathbb{I}_Y + \mathbb{I}^1_Z}[$ there exists a sequence of codes $\{\mathcal{C}_n\}_{n\geqslant 1}$ and two constants $\zeta_1, \zeta_2 > 0$ such that for large enough n, we have*

$$P_e \leqslant 2^{-\zeta_1 n} \tag{22}$$

$$\forall \mathbf{a} \in \{0,1\}^n : \quad I(W_1, \cdots, W_{p_\mathbf{a}}; \mathbf{Z}^\mathbf{a}) \leqslant 2^{-\zeta_2 n}, \tag{23}$$

with

$$p_\mathbf{a} \triangleq K_n - \lceil (1+\zeta)(\mathbb{I}^1_Z wt(\mathbf{a}) + \mathbb{I}^0_Z(n - wt(\mathbf{a}))) \rceil \tag{24}$$

with $K_n \triangleq \lfloor (1-\zeta)\mathbb{I}_Y n \rfloor$ the number of bits for $\mathcal{C}_n$.

Lemma 3 proves a result similar to the layered coding for wiretap channels [12], but note that we adopt a different approach and establish strong secrecy. The lemma does not make any statement about rate, and only characterizes the *number* of secrecy bits with hindsight after revealing the action sequence $\mathbf{a}$. This is used to construct a coding scheme in Subsection III-C.

Proof. Omitted due to space restrictions. □

C. A coding scheme for an active eavesdropper

Our coding scheme, which is illustrated in Figure 2, consists of a transmission over B blocks of length n. If $\mathbf{a}^b$ denotes the adversary's actions in block b, we assume for now that the transmitter will obtain the exact value of $\mathrm{wt}(\mathbf{a}^b)$ at the *end* of the bth block. We will remove this assumption in Subsection III-D. Intuitively, the scheme operates as follows. In every block b, we transmit K random bits $S^b_1, \cdots, S^b_K$ that do not convey information by their own; however, in the next block $b+1$, the number of secure bits is figured out and those bits are used as a key to one-time-pad information bits. The use of a layered key transmission scheme is crucial to enable the extraction of secrecy *with hindsight*. A formal description of the scheme is as follows.

Message Set: For every $b \in [\![1, B]\!]$ and $k \in [\![1, K]\!]$, assume that W^b_k and L^b_k are two random variables uniformly distributed on $\{0,1\}$ and all these random variables are independent of each other. Furthermore, for all $(k,b) \in [\![1,K]\!] \times [\![1,B]\!]$, we assume that W^b_k contains useful information but L^b_k is just an auxiliary bit.

Encoding: We use the sequence of codes $\{\mathcal{C}_n\}_{n\geqslant 1}$ introduced in Lemma 3. For the sake of simplicity, let us define $S^0_1 = S^0_2 = \cdots = S^0_K = 0$. Then, for the block $b \in [\![1,B]\!]$, code $\mathcal{C}_n$ is used for $\overline{\mathbf{S}}^b = (S^b_1, S^b_2, \cdots, S^b_K)$ where we have

$$S^b_k = \begin{cases} S^{b-1}_k \oplus W^b_k & k \in [\![1, \ell_b]\!] \\ S^{b-1}_k \oplus L^b_k & k \in [\![\ell_b + 1, K]\!] \end{cases}, \tag{25}$$

and ℓ_b is the number of secure bits in the block $b-1$, i.e.,

$$\ell_b \triangleq \begin{cases} 0 & \text{if } b = 1 \\ p_{\mathbf{a}^{b-1}} & \text{if } b \in [\![2, B]\!] \end{cases}. \tag{26}$$

Decoding: First, let $\widehat{S}^0_k = S^0_k = 0$ for all $k \in [\![1,K]\!]$. After receiving block $b \in [\![1,B]\!]$, the receiver can form the estimate $(\widehat{S}^b_1, \cdots, \widehat{S}^b_K)$ using the decoder corresponding to $\mathcal{C}_n$. Then, it can estimate W^k_b as

$$\widehat{W}^b_k = \widehat{S}^b_k \oplus \widehat{S}^{b-1}_k, \quad \forall k \in [\![1, \ell_b]\!]. \tag{27}$$

Reliability Analysis: If the receiver decodes all blocks successfully, it can decode all bits correctly. Thus, the probability of error is less than the probability that the decoder fails in at least one block. By the union bound and Lemma 3, this probability is at most $B2^{-\zeta_1 n}$.

Secrecy Analysis: The main ideas of the proof is as follows. We define

$$\mathbf{W}^b \triangleq \{W^b_k : k \in [\![1, \ell_b]\!]\}, \tag{28}$$

$$\mathbf{L}^b \triangleq \{L^b_k : k \in [\![\ell_b + 1, K]\!]\}, \tag{29}$$

$$\mathbf{S}^b \triangleq \{S^b_k : k \in [\![1, \ell_{b+1}]\!]\}, \tag{30}$$

$$\widetilde{\mathbf{S}}^b \triangleq \{S^b_k : k \in [\![\ell_{b+1} + 1, K]\!]\}. \tag{31}$$

If $\mathbf{Z}^b$ denotes the sequence received by the adversary in block b, our goal is to upper bound

$$I(\mathbf{W}^1, \cdots, \mathbf{W}^B; \mathbf{Z}^1, \cdots, \mathbf{Z}^B | \mathbf{a}^1, \cdots, \mathbf{a}^B). \tag{32}$$

Using (23) and induction one can check that we have $I(\mathbf{S}^b; \mathbf{Z}^1, \cdots, \mathbf{Z}^b) \leqslant b2^{-\zeta_2 n}$. Using the above inequality and standard chaining arguments, we obtain

$$I(\mathbf{W}^b; \mathbf{Z}^1, \cdots, \mathbf{Z}^B | \mathbf{W}^{b+1}, \cdots, \mathbf{W}^B \mathbf{a}^1, \cdots, \mathbf{a}^B) \leqslant b2^{-\zeta_2 n}. \tag{33}$$

Finally, we use the chain rule to complete the proof.

Rate Analysis: In this case, the rate is not random, since we know that the bits $\{W_k^b : b \in [\![1, B]\!], k \in [\![1, \ell_b]\!]\}$ are supposed to be transmitted. Hence, if we define $\alpha \triangleq \frac{\sum_{b=1}^{B} \mathrm{wt}(\mathbf{a}^b)}{nB}$, then one can check the rate, $\mathtt{R}(\mathbf{a})$, is equal to

$$\sum_{b=1}^{B} \frac{\ell_b}{nB} \geqslant \mathbb{I}_Y - \mathbb{I}_Z^{\alpha} - O\left(\zeta + \frac{1}{n} + \frac{1}{B}\right). \tag{34}$$

D. Estimation of adversary's actions

Notice that encoding and decoding only require knowledge of ℓ_b. Therefore, in this subsection, we give an estimate of ℓ_b denoted by $\widehat{\ell}_b$ based on the transmitter's observations $\mathbf{Z}'^{b-1}$ in block $b-1$. Before discussing $\widehat{\ell}_b$, we prove the existence of codes with an additional property:

Lemma 4. *Assume that the conditions in Lemma 3 hold. Then, there exists a sequence of codes $\{\mathcal{C}_n\}_{n\geqslant 1}$ satisfying* (22) *and* (23) *in addition to*

$$\forall \mathbf{a} \in \{0,1\}^n \quad \mathbb{D}\left(\widehat{P}^{\mathbf{a}}_{\mathbf{Z}'} \| \prod_{i=1}^{n} P^{a_i}_{Z'}\right) \leqslant e^{-\zeta_2' n}, \tag{35}$$

for sufficiently large n where $\widehat{P}^{\mathbf{a}}_{\mathbf{Z}'}$ is the induced distribution on $\mathbf{Z}'$ by the code given that adversary's actions are $\mathbf{a}$.

Proof. Omitted due to space restrictions. □

Lemma 5. *Suppose $P_{Z'^0}$ and $P_{Z'^1}$ are the output distributions of the channels $W^0_{Z'|X}$ and $W^1_{Z'|X}$ with the input distribution P_X. Furthermore, assume that there exists one symbol $\tilde{z} \in \mathcal{Z}'$ such that $P_{Z'^1}(\tilde{z}) \neq P_{Z'^0}(\tilde{z})$ and without loss of generality, it is assumed that $P_{Z'^1}(\tilde{z}) > P_{Z'^0}(\tilde{z})$. For fixed adversary's actions $\mathbf{a}^1, \cdots, \mathbf{a}^B$ and $\delta > 0$, if we estimate ℓ_b for $b \in [\![2, B]\!]$ as*

$$\widehat{\ell}_b \triangleq \lfloor (1-\zeta)\mathbb{I}_Y n \rfloor - \left\lceil (1+\zeta)((\mathbb{I}_Z^1 - \mathbb{I}_Z^0) \sum_{i=1}^{n} \frac{\mathbb{1}\{Z_i'^{b-1} = \tilde{z}\} - P_{Z'^0}(\tilde{z})}{P_{Z'^1}(\tilde{z}) - P_{Z'^0}(\tilde{z})} + \mathbb{I}_Z^0 n \right\rceil - \lceil \delta n \rceil \tag{36}$$

then we have

$$\mathbb{P}\left(\forall b \in [\![2, B]\!] : \ell_b - 2\delta n \leqslant \widehat{\ell}_b \leqslant \ell_b\right) \geqslant 1 - 2^{-\zeta_3 n}, \tag{37}$$

for some $\zeta_3 > 0$ and large enough n.

Proof. By (35), if we assume that the input distribution is i.i.d., the probabilty of events are *almost* the same. Then, using Hoeffding's Inequality yields the result. □

Lemma 5 shows that with high probability, our estimation is accurate at least to the first order in n. This can be used to argue that the secrecy and reliability criteria still hold without losing rate.

E. Proof of Theorem 2

In this subsection, we briefly discuss how to combine the results proved in the previous subsections to establish Theorem 2. First, note that the auxiliary random variable U is merely the result of channel prefixing. Then, for any block-length $N \geqslant 1$, we choose $n = \lfloor N^{\frac{2}{3}} \rfloor$ and $B = \lfloor N^{\frac{1}{3}} \rfloor - 1$ and consider the coding scheme described in Subsection III-C with ℓ_b estimated as $\widehat{\ell}_b$ introduced in (36). After sending the B blocks, we send another block containing all $\widehat{\ell}_2, \cdots, \widehat{\ell}_B$. Note that there are $O(n) = O\left(N^{\frac{2}{3}}\right)$ available bits in the last block, which is more than enough to transmit the $O(B \log n) = O\left(N^{\frac{1}{3}} \log N\right)$ bits representing $\widehat{\ell}_2, \cdots, \widehat{\ell}_B$. Since the decoder obtains $\widehat{\ell}_2, \cdots, \widehat{\ell}_B$ with high probability, one can show that estimations errors have no impact on asympotic performance and that any rate less than $\mathbb{I}_Y - \mathbb{I}_Z^{\alpha}$ is achievable.

REFERENCES

[1] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. Journal*, vol. 54, pp. 1355–1387, 1975.

[2] L. H. Ozarow and A. D. Wyner, "Wire-tap channel II," *AT&T Bell Laboratories technical journal*, vol. 63, no. 10, pp. 2135–2157, 1984.

[3] M. Nafea and A. Yener, "A new wiretap channel model and its strong secrecy capacity," in *Proc. of IEEE International Symposium on Information Theory*, 2016, pp. 2804–2808.

[4] Z. Goldfeld, P. Cuff, and H. H. Permuter, "Semantic-security capacity for wiretap channels of type II," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3863–3879, Jul 2016.

[5] V. Aggarwal, L. Lai, R. Calderbank, and H. V. Poor, "Wiretap channel type II with an active eavesdropper," in *Proc. of IEEE International Symposium on Information Theory*, Seoul, Korea, July 2009, pp. 1944–1948.

[6] P. Wang and R. Safavi-Naini, "A model for adversarial wiretap channels," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 970–983, Feb. 2016.

[7] E. MolavianJazi, M. Bloch, and J. N. Laneman, "Arbitrary jamming can preclude secure communications," in *Proc. 47th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2009, pp. 1069–1075.

[8] X. He and A. Yener, "MIMO wiretap channels with unknown and varying eavesdropper channel states," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6844–6869, Nov. 2014.

[9] R. F. Schaefer, H. Boche, and H. V. Poor, "Secure communication under channel uncertainty and adversarial attacks," *Proceedings of the IEEE*, vol. 103, no. 10, pp. 1796–1813, Oct. 2015.

[10] Z. Goldfeld, P. Cuff, and H. H. Permuter, "Arbitrarily varying wiretap channels with type constrained states," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7216–7244, Dec. 2016.

[11] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[12] S. Zou, Y. Liang, L. Lai, and S. Shamai, "Layered decoding and secrecy over degraded broadcast channels," in *Proc. of IEEE 14th Workshop on Signal Processing Advances in Wireless Communications*, 2013, pp. 679–683.