Intermittent Communication

Mostafa Khoshnevisan, *Student Member, IEEE,* and J. Nicholas Laneman, *Senior Member, IEEE*

Abstract

We formulate a model for intermittent communications that can capture bursty transmissions or a sporadically available channel, where in either case the receiver does not know a priori when the transmissions will occur. Focusing on the point-to-point case, we develop two decoding structures and their achievable rates for such communication scenarios. One structure determines the transmitted codeword, and another scheme first detects the locations of codeword symbols and then uses them to decode. We introduce the concept of partial divergence and study some of its properties in order to obtain stronger achievability results. As the system becomes more intermittent, the achievable rates decrease due to the additional uncertainty about the positions of the codeword symbols at the decoder. Additionally, we provide upper bounds on the capacity of binary noiseless intermittent communication with the help of a genie-aided encoder and decoder. The upper bounds imply a tradeoff between the capacity and the intermittency rate of the communication system. Finally, we obtain lower and upper bounds on the capacity per unit cost of intermittent communication.

Index Terms

I. Introduction

Communication systems are traditionally analyzed assuming continuous transmission of encoded symbols through the channel. However, in many practical applications such an assumption is not appropriate, and transmitting a codeword can be intermittent due to lack of synchronization, shortage of transmission energy, or burstiness of the system. The challenge is that the receiver

This work was supported in part by NSF grants CCF05-46618, CCF11-17365, and CPS12-39222.

Mostafa Khoshnevisan and J. Nicholas Laneman are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, 46556 USA, (e-mails: {mkhoshne, jnl}@nd.edu).

does not explicitly know whether a given output symbol of the channel is the result of sending a symbol of the codeword or is simply a noise symbol containing no information about the message. Intermittent communication introduced in this paper provides one model for non-contiguous transmission of codewords in such settings.

If the intermittent process is considered to be part of the channel behavior, then intermittent communication models a sporadically available channel in which at some time slots a symbol from the codeword is sent, and at others the receiver observes only noise. The model can be interpreted as an insertion channel in which some number of noise symbols are inserted between the codeword symbols. As another application, if the intermittent process is considered to be part of the transmitter, then we say that the transmitter is intermittent. Practical examples include energy harvesting systems in which the transmitter harvests energy usually from a natural source and uses it for transmission. Assuming that there is a special input that can be transmitted with zero energy, the transmitter sends the symbols of the codeword if there is enough energy for transmission, and sends the zero-cost symbol otherwise.

A. Related Work

Conceptually, two sets of related work are asynchronous communication and insertion / deletion channels. The former corresponds to contiguous transmission of codeword symbols in which the receiver observes noise before and after transmission, and the latter corresponds to a channel model in which some number of symbols are inserted at the output of the channel, or some of the input symbols of the channel are deleted. A key assumption of these communication models is that the receiver does not know a priori the positions of codeword / noise / deleted symbols, capturing certain kinds of asynchronism. Generally, this asynchronism makes the task of the decoder more difficult since, in addition to the uncertainty about the message of interest, there is also uncertainty about the positions of the symbols.

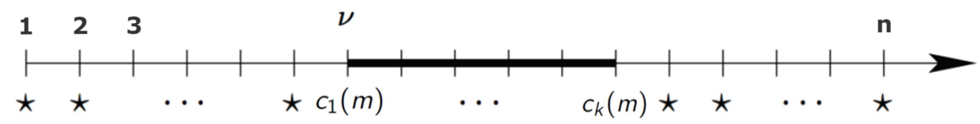
An information theoretic model for asynchronous communication is developed in [1]–[4] with a single block transmission of length k that starts at a random time v unknown to the receiver, within an exponentially large window $n = e^{\alpha k}$ known to the receiver, where α is called the asynchronism exponent. In this model, the transmission is contiguous; once it begins, the whole codeword is transmitted, and the receiver observes noise only both before and after transmission. Originally, the model appears in [5] with the goal being to locate a sync pattern at the output of a memoryless channel using a sequential detector without communicating a message to the receiver. The same author extend the framework for modeling asynchronous communication in [1], [2], where communication rate is defined with respect to the decoding delay, and give bounds on the capacity of the system.

Capacity per unit cost for asynchronous communication is studied in [3]. The capacity of asynchronous communication with a simpler notion of communication rate, i.e., with respect to the codeword length, is a special case of the result in [3]. In [4], it is shown that if the decoder is required to both decode the message and locate the codeword exactly, the capacity remains unchanged, and can be universally achieved. Also, it is shown that the channel dispersion is not affected because of asynchronism. In a recent work [6], the authors study the capacity per unit cost for asynchronous communication if the receiver is constrained to sample only a fraction of the channel outputs, and they show that the capacity remains unchanged under this constraint.

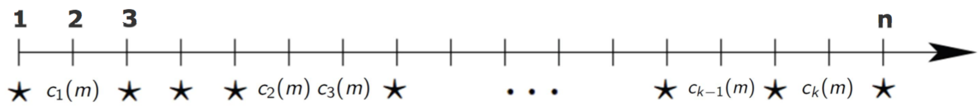
In [7], a slotted asynchronous channel model is investigated and fundamental limits of asynchronous communication in terms of miss detection and false alarm error exponents are studied. A more general result in the context of single-message unequal error protection (UEP) is given in [8].

Although we formally introduce the system model of intermittent communication in Section II, we would like to briefly compare asynchronous communication with intermittent communication. First, as contrasted in Figure 1, the transmission of codeword symbols is contiguous in asynchronous communication, whereas it is bursty in intermittent communication. Second, the receive window n is exponential with respect to the codeword length k in asynchronous communication, but as we will see it is linear in intermittent communication. Finally, in both models, we assume that the receiver does not know a priori the positions of codeword symbols.

Non-contiguous transmission of codeword symbols can be described by the following insertion channel: after the i^{th} symbol of the codeword, N_i fixed noise symbols $\star$ are inserted, where N_i, $i = 1, 2, ..., k$ are random variables, possibly independent and identically distributed (iid). The resulting sequence passes through a discrete memoryless channel, and the receiver should decode the message based on the output of the channel without knowing the positions of the codeword symbols. As will see later, if $N \geq k$ is the random variable denoting the total number of received symbols, the intermittency rate $N/k \xrightarrow{p} \alpha$ as $k \to \infty$ will be an important parameter of the system. To the best of our knowledge, this insertion channel model has not been studied before.



(a) Asynchronous communication: $n = e^{\alpha k}$.



(b) Intermittent communication: $n = \alpha k$.

Fig. 1: Comparison of the channel input sequence of asynchronous communication with intermittent communication.

A more general class of channels with synchronization errors is studied in [9] in which every transmitted symbol is independently replaced with a random number of the same symbol, possibly empty string to model a deletion event. Dobrushin [9] proved the following characterization of the capacity for such iid synchronization error channels.

Theorem 1. *([9]): For iid synchronization error channels, let $X^k := (X_1, X_2, ..., X_k)$ denote the channel input sequence of length k, and $Y^N := (Y_1, Y_2, ..., Y_N)$ denote the corresponding output sequence at the decoder, where length N is a random variable determined by the channel realization. The channel capacity is*

$$C = \lim_{k \to \infty} \max_{P_{X^k}} \frac{1}{k} \mathbb{I}(X^k; Y^N), \tag{1}$$

where $\mathbb{I}(X;Y)$ denotes the average mutual information [27]. Theorem 1 demonstrates that iid synchronization error channels are information stable. However, there are two difficulties related to computing the capacity through this characterization. First, it is challenging to compute the mutual information because of the memory inherent in the joint distribution of the input and output sequences. Second, the optimization over all the input distributions is computationally involved. A single-letter characterization for the capacity of the general class of synchronization error channels is still an open problem, even though there are many papers deriving bounds on the capacity of the insertion / deletion channels [10]–[19].

Focusing on the lower bounds, Gallager [10] considers a channel model with substitution and insertion / deletion errors and derives a lower bound on the channel capacity. In [11], codebooks from first-order Markov chains are used to improve the achievability results for deletion channels. The intuition is that it is helpful to put some memory in the codewords if the channel has some inherent memory. Improved achievability results are obtained in [12], [13], with the latter directly lower bounds the information capacity given by (1) for channels with iid deletions and duplications with an input distribution following a symmetric, first-order Markov chain.

Upper bounds on the capacity of iid insertion / deletion channels are obtained in [17]–[19], in which the capacity of an axillary channel obtained by a genie-aided decoder (and encoder) with access to side-information about the insertion / deletion process is numerically evaluated. Although our insertion channel model is different, we can apply some of these ideas and techniques to upper bound the capacity of intermittent communication.

Finally, bounds on the capacity per unit cost for a noisy channel with synchronization errors are given in a recent work [20], in which an encoding / decoding scheme with only loose transmitter-receiver synchronization is proposed. We use the same approach for obtaining a lower bound on the capacity per unit cost of intermittent communication.

B. Summary of Contributions

After introducing a model for intermittent communication in Section II, we develop two coding theorems for achievable rates to lower bound the capacity in Section III. Toward this end, we introduce the notion of partial divergence and its properties. We show that, as long as the ratio of the receive window to the codeword length is finite and the capacity of the DMC is not zero, rate $R = 0$ is achievable for intermittent communication; i.e., if there are only two messages, then the probability of decoding error vanishes as the codeword length becomes sufficiently large. By using decoding from exhaustive search and decoding from pattern detection, we obtain two achievable rates that are also valid for arbitrary intermittent processes.

Focusing on the binary-input binary-output noiseless case, we obtain upper bounds on the capacity of intermittent communication in Section IV by providing the encoder and the decoder with various amounts of side-information, and calculating or upper bounding the capacity of this genie-aided system. Although the gap between the achievable rates and upper bounds is fairly large, especially for large values of intermittency rate, the results suggest that linear scaling

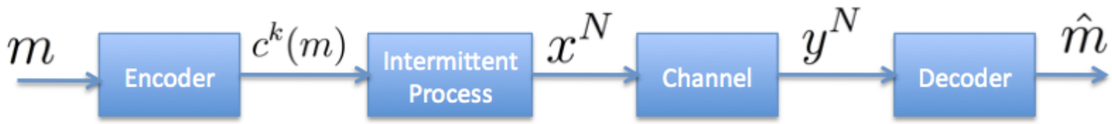


Fig. 2: System model for intermittent communication.

of the receive window with respect to the codeword length considered in the system model is relevant since the upper bounds imply a tradeoff between the capacity and the intermittency rate.

Finally, in Section V we obtain lower and upper bounds on the capacity per unit cost of intermittent communication. To obtain the lower bound, we use a similar approach to the one in [20], in which pulse-position modulation codewords are used at the encoder, and the decoder searches for the position of the pulse in the output sequence. The upper bound is the capacity per unit cost of the DMC.

II. System Model and Foundations

We consider a communication scenario in which a transmitter communicates a single message $m \in \{1, 2, ..., e^{kR} = M\}$ to a receiver over a discrete memoryless channel (DMC) with probability transition matrix W and input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let C_W denote the capacity of the DMC. Also, let $\star \in \mathcal{X}$ denote the "noise" symbol, which corresponds to the input of the channel when the transmitter is "silent". The transmitter encodes the message as a codeword $c^k(m)$ of length k, which is the input sequence of intermittent process as shown in Figure 2.

The intermittent process captures the burstiness of the channel or the transmitter and can be described as follows: After the i^{th} symbol from the codeword, N_i noise symbols $\star$ are inserted, where N_i's are iid geometric random variables with mean $\alpha - 1$, where $\alpha \geq 1$ is the intermittency rate. As will see later, if $N \geq k$ is the random variable denoting the total number of received symbols, the intermittency rate $N/k \xrightarrow{p} \alpha$ as $k \to \infty$ will be an important parameter of the system. In fact, the larger the value of α, the larger the receive window, and therefore, the more intermittent the system becomes with more uncertainty about the positions of the codeword symbols; if $\alpha = 1$, the system is not intermittent and corresponds to contiguous communication. We call this scenario *intermittent communication* and denote it by the tuple $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$.

This model corresponds to an iid insertion channel model in which at each time slot a codeword symbol is sent with probability $p_t := 1/\alpha$ and the noise symbol $\star$ is sent with probability $1-p_t$ until the whole codeword is transmitted.

The output of intermittent process then goes through a DMC. At the decoder, there are N symbols, where N is a random variable having a negative binomial distribution with parameters k and p_t:

$$P(N=n) = \binom{n-1}{k-1} p_t^k (1-p_t)^{n-k}, n \geq k, \tag{2}$$

with $\mathbb{E}[N] = \alpha k$, and we have

$$\frac{N}{k} = \frac{k + N_0 + N_1 + N_2 + ... + N_k}{k} \xrightarrow{p} 1 + \mathbb{E}(N_0) = \alpha, \text{ as } k \to \infty, \tag{3}$$

Therefore, the receive window N scales linearly with the codeword length k, as opposed to the exponential scaling in asynchronous communication. Intermittent communication model represents bursty communication in which either the transmitter or the channel is bursty. In a bursty communication scenario, the receiver usually does not know the realization of the bursts. Therefore, we assume that the receiver does not know the positions of the codeword symbols, making the decoder's task more involved.

Assuming that the decoded message is denoted by $\hat{m}$, which is a function of the random sequence Y^N, and defining a code as in [27], we say that rate R is achievable if there exists a sequence of length k codes of size e^{kR} with $\frac{1}{M}\sum_{m=1}^{M}\mathbb{P}(\hat{m} \neq m) \to 0$ as $k \to \infty$. Note that the communication rate is defined as $\log M/k$. The capacity is the supremum of all the achievable rates. Rate region (R, α) is said to be achievable if the rate R is achievable for the corresponding scenario with the intermittency rate α.

It can be seen that the result of Theorem 1 for iid synchronization error channels applies to intermittent communication model, and therefore, the capacity equals $\lim_{k\to\infty} \max_{P_{C^k}} \frac{1}{k}\mathbb{I}(C^k; Y^N)$. Denting the binary entropy function by $h(p) := -p\log p - (1-p)\log(1-p)$, we have the following theorem.

Theorem 2. *For intermittent communication* $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$*, rates not exceeding* $R_1 := (C_W - \alpha h(1/\alpha))^+$ *are achievable.*

Proof: We show that R_1 is a lower bound for the capacity of intermittent communication by lower bounding the mutual information. Let vector $T^{k+1} := (N_0, N_1, ..., N_k)$ denote the number of noise insertions in between the codeword symbols, where the N_i's are iid geometric random variables with mean $\alpha - 1$. We have

$$\begin{aligned}
\mathbb{I}(C^k; Y^N) &= \mathbb{I}(C^k; Y^N, T^{k+1}) - \mathbb{I}(C^k; T^{k+1}|Y^N) \\
&= \mathbb{I}(X^k; Y^k) - \mathbb{I}(C^k; T^{k+1}|Y^N) \qquad (4) \\
&\geq k\mathbb{I}(X; Y) - H(T^{k+1}) \qquad (5) \\
&= k\mathbb{I}(X; Y) - (k+1)H(N_0) \qquad (6) \\
&= k\mathbb{I}(X; Y) - (k+1)\alpha h(\frac{1}{\alpha}), \qquad (7)
\end{aligned}$$

where (4) follows from the fact that C^k is independent of the insertion process T^{k+1}, and conditioned on the positions of noise symbols T^{k+1}, the mutual information between C^k and Y^N equals the mutual information between input and output sequences of the DMC without considering the noise insertions; where (5) follows by considering iid codewords and by the fact that conditioning cannot increase the entropy; and (6) and (7) follow from the fact that N_i's are iid geometric random variables. Finally, the result follows after dividing both sides by k and considering the capacity achieving input distribution of the DMC. ■

Although the lower bound on the capacity of intermittent communication in Theorem 2 is valid for the specific intermittent process described above, our achievability results in Section III apply to an arbitrary insertion process as long as $N/k \xrightarrow{p} \alpha$ as $k \to \infty$.

At this point, we summarize several notations that are used throughout the sequel. We use $o(\cdot)$ and $\text{poly}(\cdot)$ to denote quantities that grow strictly slower than their arguments and are polynomial in their arguments, respectively. By $X \sim P(x)$, we mean X is distributed according to P. For convenience, we define $W_\star(\cdot) := W(\cdot|X = \star)$, and more generally, $W_x(\cdot) := W(\cdot|X = x)$. In this paper, we use the convention that $\binom{n}{k} = 0$ if $k < 0$ or $n < k$, and the entropy $H(P) = -\infty$ if P is not a probability mass function, i.e., one of its elements is negative or the sum of its elements is larger than one. We also use the conventional definitions $x^+ := \max\{x, 0\}$, and if $0 \leq \rho \leq 1$, then $\bar{\rho} := 1 - \rho$. Additionally, $\doteq$ denotes an equality in exponential sense as $k \to \infty$, i.e., $\lim_{k\to\infty} \frac{1}{k} \log(\cdot)$ of both sides are equal.

We will make frequent use of the notations and results from the method of types, a powerful technique in large deviation theory developed by Csiszár and Körner [21]. Let $\mathcal{P}^{\mathcal{X}}$ denote the set of distributions over the finite alphabet $\mathcal{X}$. The empirical distribution (or type) of a sequence

$x^n \in \mathcal{X}^n$ is denoted by $\hat{P}_{x^n}$. A sequence x^n is said to have a type P if $\hat{P}_{x^n} = P$. The set of all sequences that have type P is denoted T_P^n, or more simply T_P. Joint empirical distributions are denoted similarly. The set of sequences y^n that have a conditional type W given x^n is denoted $T_W(x^n)$. A sequence $x^n \in \mathcal{X}^n$ is called P-typical with constant μ, denoted $x^n \in T_{[P]_\mu}$, if

$$|\hat{P}_{x^n}(x) - P(x)| \leq \mu \text{ for every } x \in \mathcal{X}.$$

Similarly, a sequence $y^n \in \mathcal{Y}^n$ is called W-typical conditioned on $x^n \in \mathcal{X}^n$ with constant μ, denoted $y^n \in T_{[W]_\mu}$, if

$$|\hat{P}_{x^n,y^n}(x,y) - \hat{P}_{x^n}(x)W(y|x)| \leq \mu \text{ for every } (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

For $P, P' \in \mathcal{P}^{\mathcal{X}}$ and $W, W' \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, the Kullback-Leibler divergence between P and P' is defined as

$$D(P\|P') := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{P'(x)},$$

and the conditional information divergence between W and W' conditioned on P is defined as

$$D(W\|W'|P) := \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{W'(y|x)}.$$

The average mutual information between $X \sim P$ and $Y \sim PW$ and coupled via $P_{Y|X} = W$ is denoted by $\mathbb{I}(P, W)$. With these definitions, we now state the following lemmas, which are used throughout the paper.

Lemma 1. *([21, Lemma 1.2.6]): If X^n is an iid sequence according to P', then the probability that it has a type P is bounded by*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} e^{-nD(P\|P')} \leq \mathbb{P}(X^n \in T_P) \leq e^{-nD(P\|P')}.$$

Also, if the input $x^n \in \mathcal{X}^n$ to a memoryless channel $W' \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$ has type P, then the probability that the observed channel output sequence Y^n has a conditional type W given x^n is bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} e^{-nD(W\|W'|P)} \leq \mathbb{P}(Y^n \in T_W(x^n)) \leq e^{-nD(W\|W'|P)}.$$

Lemma 2. *([21, Lemma 1.2.12]): If X^n is an iid sequence according to P, then*

$$\mathbb{P}(X^n \in T_{[P]_\mu}) \geq 1 - \frac{|\mathcal{X}|}{4n\mu^2}.$$

Also, if the input $x^n \in \mathcal{X}^n$ to a memoryless channel $W \in \mathcal{P}^{\mathcal{Y}|\mathcal{X}}$, and Y^n is the output, then

$$\mathbb{P}(Y^n \in T_{[W]_\mu}(x^n)) \geq 1 - \frac{|\mathcal{X}||\mathcal{Y}|}{4n\mu^2},$$

and here the terms subtracted from 1 *could be replaced by exponentially small terms* $2|\mathcal{X}|e^{-2n\mu^2}$ *and* $2|\mathcal{X}||\mathcal{Y}|e^{-2n\mu^2}$*, respectively.*

Finally, we state a stronger version of the packing lemma [22, Lemma 3.1] that will be useful in typicality decoding, and is proved in [2, Equations (24) and (25)] based on the method of types.

Lemma 3. *Assume that $C^n(m)$ and $\tilde{Y}^n$ are independent, $C^n(m)$ is generated iid according to P, then*

$$\mathbb{P}(\tilde{Y}^n \in T_{[W]_\mu}(C^n(m))) \leq \textit{poly}(n) e^{-n(\mathbb{I}(P,W)-\epsilon)}$$

for all n sufficiently large, where $\epsilon > 0$ can be made arbitrarily small by choosing a small enough typicality parameter μ.

III. Achievability

In this section, we obtain achievability results for intermittent communication based upon two decoding structures: *decoding from exhaustive search*, which attempts to decode the transmitted codeword from a selected set of output symbols without any attempt to first locate or detect the codeword symbols; and *decoding from pattern detection*, which attempts to decode the transmitted codeword only if the selected outputs appear to be a pattern of codeword symbols. In order to analyze the probability of error for the second decoding structure, which gives a larger achievable rate, we use a generalization of the method of types that leads to the notion of *partial divergence* and its properties described in Section III-A. We also show that rate $R = 0$ is achievable for intermittent communication with finite intermittency rate, using the properties of partial divergence. Although the system model in Section II assumes iid geometric insertions, the results of this section apply to a general intermittent process, as we point out in Remark 1.

A. Partial Divergence

We will see in Section III-C that a relevant function is $d_\rho(P\|Q)$, which we call partial divergence and view as a generalization of the Kullback-Leibler divergence. In this section, we examine some of the interesting properties of partial divergence that provide insights about the achievable rates in Section III-C. Loosely speaking, partial divergence is the normalized exponent of the probability that a sequence with independent elements generated partially according to one distribution and partially according to another distribution has a specific type. This exponent is useful in characterizing a decoder's ability to distinguish a sequence obtained partially from the codewords and partially from the noise from a codeword sequence or a noise sequence. The following lemma is a generalization of the method of types [21]. A different characterization and proof of Lemma 4 can be found in [23].

Lemma 4. *Consider the distributions $P, Q, Q' \in \mathcal{P}^{\mathcal{X}}$ on a finite alphabet $\mathcal{X}$. A random sequence X^k is generated as follows: ρk symbols iid according to Q and $\bar{\rho} k$ symbols iid according to Q', where $0 \leq \rho \leq 1$. The normalized exponent of the probability that X^k has type P is*

$$\begin{aligned} d(P, Q, Q', \rho) &:= \lim_{k \to \infty} -\frac{1}{k} \log \mathbb{P}(X^k \in T_P) \\ &= \min_{\substack{P_1, P_2 \in \mathcal{P}^{\mathcal{X}}: \\ \rho P_1 + \bar{\rho} P_2 = P}} \rho D(P_1 \| Q) + \bar{\rho} D(P_2 \| Q'). \end{aligned} \tag{8}$$

Proof: With a little abuse of notation, let $X_1^{\rho k}$ and $X_2^{\bar{\rho} k}$ be the sequence of symbols in X^k that are iid according to Q, Q', respectively. If these sequences have types P_1 and P_2, respectively, then the whole sequence X^k has type $\rho P_1 + \bar{\rho} P_2$. Therefore, we have

$$\begin{aligned} \mathbb{P}(X^k \in T_P) &= \mathbb{P}\left(\cup_{\substack{P_1, P_2 \in \mathcal{P}^{\mathcal{X}}: \\ \rho P_1 + \bar{\rho} P_2 = P}} \{X_1^{\rho k} \in T_{P_1}, X_2^{\bar{\rho} k} \in T_{P_2}\} \right) & (9) \\ &= \sum_{\substack{P_1, P_2 \in \mathcal{P}^{\mathcal{X}}: \\ \rho P_1 + \bar{\rho} P_2 = P}} \mathbb{P}(X_1^{\rho k} \in T_{P_1}, X_2^{\bar{\rho} k} \in T_{P_2}) & (10) \\ &\doteq \sum_{\substack{P_1, P_2 \in \mathcal{P}^{\mathcal{X}}: \\ \rho P_1 + \bar{\rho} P_2 = P}} \exp\{-k(\rho D(P_1 \| Q) + \bar{\rho} D(P_2 \| Q'))\}, & (11) \\ &\doteq \exp\{-k \min_{\substack{P_1, P_2 \in \mathcal{P}^{\mathcal{X}}: \\ \rho P_1 + \bar{\rho} P_2 = P}} \rho D(P_1 \| Q) + \bar{\rho} D(P_2 \| Q')\} & (12) \end{aligned}$$

where: (10) follows from the disjointness of the events in (9) since a sequence has a unique type; (11) follows from the independence of the events in (10) and obtaining the probability of each of them according to Lemma 1 to first order in the exponent; and (12) follows from the fact that the number of different types is polynomial in the length of the sequence [21], which makes the total number of terms in the summation (11) polynomial in k, and therefore, the exponent equals the largest exponent of the terms in the summation (11). ■

Specializing Lemma 4 for $Q' = Q$ results in Lemma 1, and we have $d(P, Q, Q, \rho) = D(P\|Q)$. However, we will be interested in the special case of Lemma 4 for which $Q' = P$. In other words, we need to upper bound the probability that a sequence has a type P if its elements are generated independently with a fraction ρ according to Q and the remaining fraction $\bar{\rho}$ according to P. For this case, we call $d_\rho(P\|Q) := d(P, Q, P, \rho)$ the partial divergence between P and Q with mismatch ratio $0 \leq \rho \leq 1$. Proposition 1 gives an explicit expression for the partial divergence by solving the optimization problem in (8) for the special case of $Q' = P$.

Proposition 1. *If $\mathcal{X} = \{1, 2, ..., |\mathcal{X}|\}$ and $P, Q \in \mathcal{P}^{\mathcal{X}}$, where $P := (p_1, p_2, ..., p_{|\mathcal{X}|})$ and $Q := (q_1, q_2, ..., q_{|\mathcal{X}|})$ and we assume that all values of the PMF Q are nonzero, then partial divergence can be written as*

$$d_\rho(P\|Q) = D(P\|Q) - \sum_{j=1}^{|\mathcal{X}|} p_j \log(c^* + \frac{p_j}{q_j}) + \rho \log c^* + h(\rho), \tag{13}$$

where c^* is a function of ρ, P, and Q, and can be uniquely determined from

$$c^* \sum_{j=1}^{|\mathcal{X}|} \frac{p_j q_j}{c^* q_j + p_j} = \rho. \tag{14}$$

A proof of the proposition uses Karush-Kuhn-Tucker (KKT) conditions since the optimization problem in (8) is convex, but is omitted due to space considerations. The next proposition states some of the properties of the partial divergence, which will be applied in the sequel.

Proposition 2. *The partial divergence $d_\rho(P\|Q), 0 \leq \rho \leq 1$, where all of the elements of the PMF Q are nonzero, has the following properties:*

(a) $d_0(P\|Q) = 0$.

(b) $d_1(P\|Q) = D(P\|Q)$.

(c) Partial divergence is zero if $P = Q$, i.e., $d_\rho(P\|P) = 0$.

(d) Let $d'_\rho(P\|Q) := \frac{\partial d_\rho(P\|Q)}{\partial \rho}$ denote the derivative of the partial divergence with respect to ρ, then $d'_0(P\|Q) = 0$.

(e) If $P \neq Q$, then $d'_\rho(P\|Q) > 0$, for all $0 < \rho \leq 1$, i.e., partial divergence is increasing in ρ.

(f) If $P \neq Q$, then $d''_\rho(P\|Q) > 0$, for all $0 \leq \rho \leq 1$, i.e., partial divergence is convex in ρ.

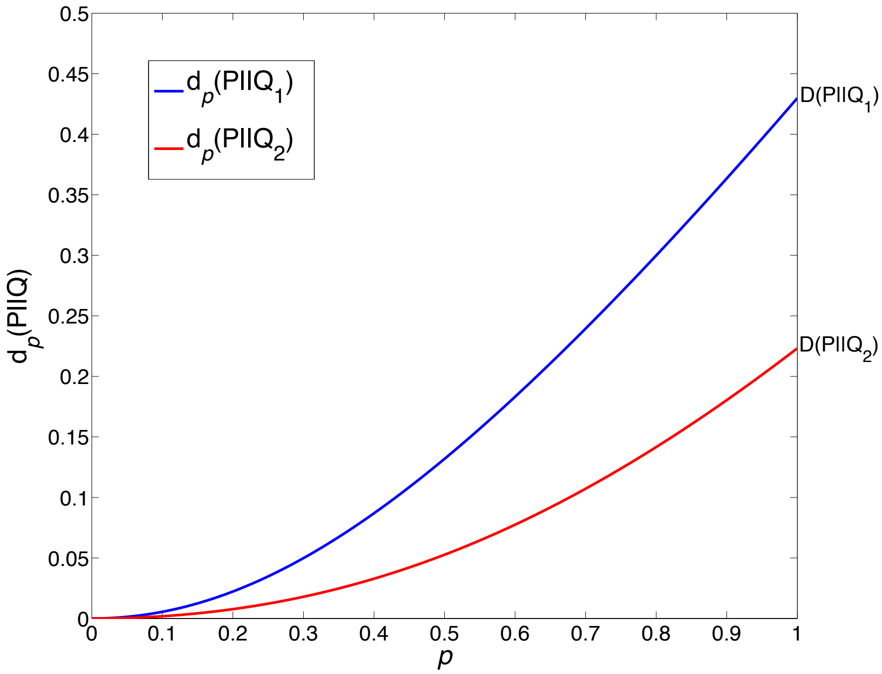


Fig. 3: Partial divergence $d_\rho(P\|Q)$ versus ρ for $P = (0.25, 0.25, 0.25, 0.25)$, $Q_1 = (0.1, 0.1, 0.1, 0.7)$, and $Q_2 = (0.1, 0.4, 0.1, 0.4)$.

(g) $0 \leq d_\rho(P\|Q) \leq \rho D(P\|Q)$.

Proof: See Appendix A. ■

Figure 3 shows examples of the partial divergence for PMF's with alphabets of size 4. Specifically, $d_\rho(P\|Q)$ versus ρ is sketched for $P = (0.25, 0.25, 0.25, 0.25)$, and two different Q's, $Q_1 = (0.1, 0.1, 0.1, 0.7)$ and $Q_2 = (0.1, 0.4, 0.1, 0.4)$. The properties of Proposition 2 are apparent in the figure for these examples.

Proposition 3. *The partial divergence* $d_\rho(P\|Q), 0 \leq \rho \leq 1$, *satisfies*

$$d_\rho(P\|Q) \geq D(P\|\rho Q + \bar{\rho} P).$$

Proof: From the definition of the partial divergence and (8), we have

$$d_\rho(P\|Q) = \min_{\substack{P_1,P_2\in\mathcal{P}^{\mathcal{X}}:\\ \rho P_1+\bar{\rho}P_2=P}} \rho D(P_1\|Q) + \bar{\rho} D(P_2\|P)$$

$$\geq \min_{\substack{P_1,P_2\in\mathcal{P}^{\mathcal{X}}:\\ \rho P_1+\bar{\rho}P_2=P}} D(\rho P_1 + \bar{\rho}P_2\|\rho Q + \bar{\rho}P) \quad (15)$$

$$= D(P\|\rho Q + \bar{\rho}P), \quad (16)$$

where (15) follows from the convexity of the Kullback-Leibler divergence; and (16) follows from the constraint $\rho P_1 + \bar{\rho}P_2 = P$ in the minimization. ■

The interpretation of Proposition 3 is that if all the elements of a sequence are generated independently according to a mixture probability $\rho Q + \bar{\rho}P$, then it is more probable that this sequence has type P than in the case that a fraction ρ of its elements are generated independently according to Q and the remaining fraction $\bar{\rho}$ are generated independently according to P. Since the partial divergence $d_\rho(P\|Q)$ is used to obtain achievability results in Theorem 5, it can be substituted with the less complicated function $D(P\|\rho Q + \bar{\rho}P)$ in (20) with the expense of loosening the bound according to Proposition 3.

Using the results on partial divergence, we now state a result about the achievability at rate $R = 0$ in the following theorem. The idea is that no matter how large the intermittency threshold becomes, as long as it is finite, the receiver can distinguish between two messages with vanishing probability of error.

Theorem 3. *If the intermittency rate is finite and the capacity of the DMC, C_W, is non-zero, then rate $R = 0$ is achievable.*

Proof: We need to show that the transmission of a message $m \in \{1, 2\}$ is reliable for intermittent communication $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$ as $k \to \infty$.

Encoding: If $m = 1$, then transmit symbol $\star$ at all the times, i.e., $c^k(1) = \star^k$. If $m = 2$, then transmit symbol $x \neq \star$ at all the times, i.e., $c^k(2) = x^k$, where we consider the symbol $x^* = \text{argmax}_x D(W_\star\|W_x)$. If the capacity of the DMC is nonzero, then $W_\star \neq W_{x^*}$.

Decoding: For arbitrarily small ϵ, if $|N/k - \alpha| > \epsilon$, then the decoder declares an error; otherwise, if the sequence y^N has type $W_\star$ with a fixed typicality parameter $\mu > 0$, i.e., $y^N \in T_{[W_\star]_\mu}$, then $\hat{m} = 1$; otherwise $\hat{m} = 2$.

Analysis of the probability of error: The average probability of error can be bounded as

$$p_e \leq \mathbb{P}(|N/k-\alpha|>\epsilon)$$
$$+\sum_{n=k(\alpha-\epsilon)-1}^{k(\alpha+\epsilon)+1} \mathbb{P}(N=n)\left(\mathbb{P}(Y^N \notin T_{[W_\star]_\mu}|m=1,N=n)+\mathbb{P}(Y^N \in T_{[W_\star]_\mu}|m=2,N=n)\right) \quad (17)$$
$$\leq o(1)+\max_{n:|n/k-\alpha|<\epsilon}(o(1)+e^{-nd_{k/n}(W_\star\|W_{x^*})}) \quad (18)$$
$$\leq o(1)+e^{-k(\alpha-\epsilon)d_{1/(\alpha+\epsilon)}(W_\star\|W_{x^*})} \to 0 \text{ as } k\to\infty, \quad (19)$$

where: (17) follows from the union bound; (18) follows from the fact that $\mathbb{P}(|N/k-\alpha|>\epsilon)\to 0$ as $k\to\infty$, Lemma 2 and Lemma 4; and (19) follows from the fact that $d_{1/(\alpha+\epsilon)}(W_\star\|W_{x^*})>0$ according to Proposition 2 since $W_\star \neq W_{x^*}$ and $1/(\alpha+\epsilon)>0$. ■

In order to prove achievability results for the case of an exponential number of messages, i.e., $R>0$, we introduce two decoding algorithms in the following section.

B. Decoding Algorithms

In this section, two decoding structures are introduced. The encoding structure is identical for both: Given an input distribution P, the codebook is randomly and independently generated, i.e., all $C_i(m), i\in\{1,2,...,k\}, m\in\{1,2,...,M\}$ are iid according to P. Although we focus on typicality for detection and decoding for ease of analyzing the probability of error, other algorithms such as maximum likelihood decoding could in principle be used in the context of these decoding structures. However, detailed specification and analysis of such structures and algorithms are beyond the scope of this paper. Note that the number of received symbols at the decoder, N, is a random variable. However, using the same procedure as in the proof of Theorem 3, we can focus on the case that $|N/k-\alpha|<\epsilon$, and essentially assume that the receive window is of length $n=\alpha k$, which makes the analysis of the probability of error for the decoding algorithms more concise.

Decoding from exhaustive search: In this structure, the decoder observes the n symbols of the output sequence y^n, chooses k of them, resulting in a subsequence denoted by $\tilde{y}^k$, and performs joint typicality decoding with a fixed typicality parameter $\mu>0$, i.e., checks if $\tilde{y}^k \in T_{[W]_\mu}(c^k(m))$ for a unique index m. In words, this condition corresponds to the joint type for codeword $c^k(m)$ and selected outputs $\tilde{y}^k$ being close to the joint distribution induced by $c^k(m)$ and the channel

$W(y|x)$. If the decoder finds a unique m satisfying this condition, it declares m as the transmitted message. Otherwise, it makes another choice for the k symbols from symbols of the sequence y^n and again attempts typicality decoding. If at the end of all $\binom{n}{k}$ choices the typicality decoding procedure does not declare any message as being transmitted, then the decoder declares an error.

Decoding from pattern detection: This structure involves two stages for each choice of the output symbols. As in decoding from exhaustive search, the decoder chooses k of the n symbols from the output sequence y^n. Let $\tilde{y}^k$ denote the subsequence of the chosen symbols, and $\hat{y}^{n-k}$ denote the subsequence of the other symbols. For each choice, the first stage checks if this choice of the output symbols is a good one, which consists of checking if $\tilde{y}^k$ is induced by a codeword, i.e., if $\tilde{y}^k \in T_{PW}$, and if $\hat{y}^{n-k}$ is induced by noise, i.e., if $\hat{y}^{n-k} \in T_{W_\star}$. If both of these conditions are satisfied, then we perform typicality decoding with $\tilde{y}^k$ over the codebook as in the decoding from exhaustive search, which is called the second stage here. Otherwise, we make another choice for the k symbols and repeat the two-stage decoding procedure. At any step that we run the second stage, if the typicality decoding declares a message as being sent, then decoding ends. If the decoder does not declare any message as being sent by the end of all $\binom{n}{k}$ choices, then the decoder declares an error. In this structure, we constrain the search domain for the typicality decoding (the second stage) only to typical patterns by checking that our choice of codeword symbols satisfies the conditions in the first stage.

In decoding from pattern detection, the first stage essentially distinguishes a sequence obtained partially from the codewords and partially from the noise from a codeword sequence or a noise sequence. As a result, in the analysis of the probability of error, partial divergence and its properties described in Section III-A play a role. This structure always outperforms the decoding from exhaustive search structure, and their difference in performance indicates how much the results on the partial divergence improve the achievable rates.

C. Achievable Rates

In this section, we obtain two achievable rates for intermittent communication using the decoding algorithms introduced in Section III-B.

Theorem 4. *Using decoding from exhaustive search for intermittent communication with* $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$, *rates not exceeding* $R_1 = (C_W - \alpha h(1/\alpha))^+$ *are achievable.*

The proof of the theorem is removed since the analysis of the probability of error is similar to that of Theorem 5, except that instead of breaking down the first term in (46) as in (49), we use the union bound over all the $\binom{n}{k}$ choices without trying to distinguish the output symbols based on their empirical distributions.

Note that R_1 is the same as the lower bound in Theorem 2, but here we introduced an explicit decoding structure which is also valid for a general intermittent process. The form of the achievable rate is reminiscent of communications overhead as the cost of constraints [25], where the constraint is the system's burstiness or intermittency, and the overhead cost is $\alpha h(1/\alpha)$. Note that the overhead cost is increasing in the intermittency rate α, is equal to zero at $\alpha = 1$, and approaches infinity as $\alpha \to \infty$. These observations suggest that increasing the receive window makes the decoder's task more difficult.

Theorem 5. *Using decoding from pattern detection for intermittent communication with* $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$, *rates not exceeding* $\max_P\{(\mathbb{I}(P, W) - f(P, W, \alpha))^+\}$ *are achievable, where*

$$f(P, W, \alpha) := \max_{0\leq\beta\leq1}\{(\alpha-1)h(\beta)+h((\alpha-1)\beta)-d_{(\alpha-1)\beta}(PW\|W_\star)-(\alpha-1)d_\beta(W_\star\|PW)\}. \quad (20)$$

Proof: See Appendix B. ∎

Remark 1. *The results of Theorems 3, 4, and 5 are valid for an arbitrary intermittent process in Figure 2 as long as* $N/k \xrightarrow{p} \alpha$ *as* $k \to \infty$.

Remark 2. *The achievable rate in Theorem 5 can be expressed as follows: Rate* R *is achievable if for any mismatch* $0 \leq \beta \leq 1$, *we have*

$$R + (\alpha - 1)h(\beta) + h((\alpha - 1)\beta) < \mathbb{I}(P, W) + d_{(\alpha-1)\beta}(PW\|W_\star) + (\alpha - 1)d_\beta(W_\star\|PW).$$

The interpretation is that the total amount of uncertainty should be smaller than the total amount of information. Specifically, R *and* $(\alpha - 1)h(\beta) + h((\alpha - 1)\beta)$ *are the amount of uncertainty in codewords and patterns, respectively, and* $\mathbb{I}(P, W)$ *and* $d_{(\alpha-1)\beta}(PW\|W_\star) + (\alpha-1)d_\beta(W_\star\|PW)$ *are the amount of information about the codewords and patterns, respectively.*

The achievable rate in Theorem 5 is larger than the one in Theorem 4, because decoding from pattern detection utilizes the fact that the choice of the codeword symbols at the receiver might not be a good one, and therefore, restricts the typicality decoding only to the typical patterns and decreases the search domain. In Theorem 5, the overhead cost for a fixed input distribution is $f(P, W, \alpha)$. Using the properties of partial divergence, we state some properties of this overhead cost in the next proposition.

Proposition 4. *The overhead cost* $f(P, W, \alpha)$ *in* (20) *has the following properties:*

(a) The maximum of the term in (20) *occurs in the interval* $[0, 1/\alpha]$*, i.e., instead of the maximization over* $0 \leq \beta \leq 1$*,* $f(P, W, \alpha)$ *can be found by the same maximization problem over* $0 \leq \beta \leq 1/\alpha$*.*

(b) $f(P, W, \alpha)$ *is increasing in* α*.*

(c) $f(P, W, 1) = 0$*.*

(d) If $D(PW \| W_\star)$ *is finite, then* $f(P, W, \alpha) \to \infty$ *as* $\alpha \to \infty$*.*

Proof: See Appendix C. ■

Note that part (b) in Proposition 4 indicates that increasing the intermittency rate or the receive window increases the overhead cost, resulting in a smaller achievable rate. Parts (c) and (d) show that the achievable rate is equal to the capacity of the channel for $\alpha = 1$ and approaches zero as $\alpha \to \infty$.

Remark 3. *The results in this section can be extended to packet-level intermittent communication in which the intermittency is modeled at the packet level, and regimes that lie between the non-contiguous transmission of codeword symbols in intermittent communication, and the contiguous transmission of codeword symbols in asynchronous communication are explored. Specifically, the system model in this paper can be generalized to small packet, medium packet, and large packet intermittent communication. See, for example, [24].*

Now consider a binary symmetric channel (BSC) for the DMC in Figure 2 with the crossover probability $0 \leq p \leq 0.5$, and the noise symbol $\star = 0$. Figure 4 shows the value of the achievable rates for different p, versus α. $R_{\text{Insertion}}$ denotes the achievable rate obtained from Theorem 5 if the channel is noiseless ($p = 0$), and can be proven to be equal to $\max_{0 \leq p_0 \leq 1}\{2h(p_0) - \max_{0 \leq \beta \leq 1}\{(\alpha-1)h(\beta) + h((\alpha-1)\beta) + (1-(\alpha-1)\beta)h(\frac{p_0-(\alpha-1)\beta}{1-(\alpha-1)\beta})\}\}$.

As we can see from the plot, the achievable rate in Theorem 5 (indicated by "R_2") is always larger than the one in Theorem 4 (indicated by "R_1") since decoding from pattern detection takes advantage of the fact that the choice of the k output symbols might not be a good one.

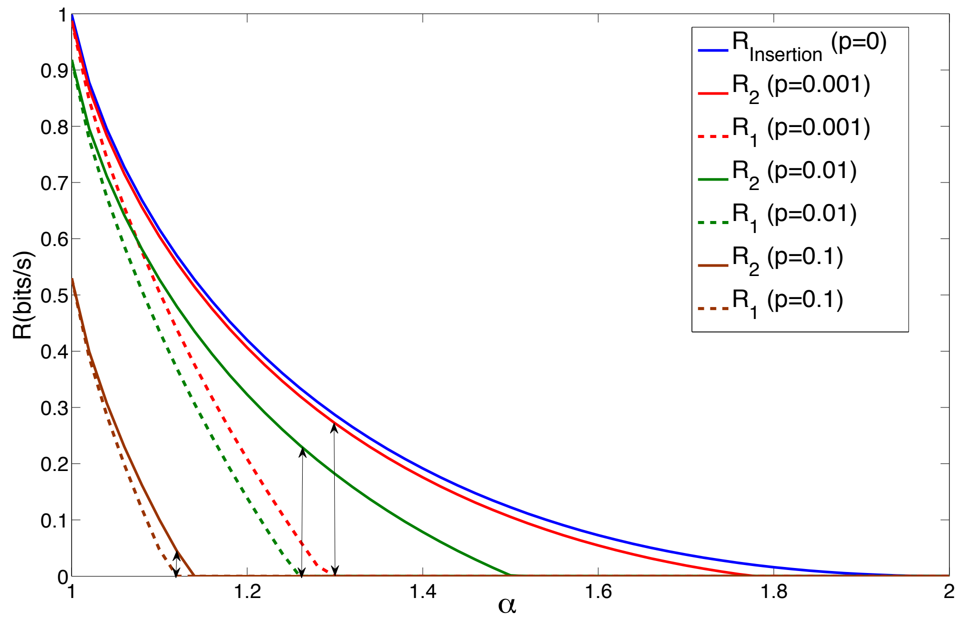


Fig. 4: Achievable rate region (R, α) for the BSC for different cross over probability p's.

Specifically, the exponent obtained in Lemma 4 in terms of the partial divergence helps the decoder detect the right symbols, and therefore, achieve a larger rate. The arrows in Figure 4 show this difference and suggest that the benefit of using decoding from pattern detection is larger for increasing α. Note that the larger α is, the smaller the achievable rate would be for a fixed p. Not surprisingly, as $\alpha \to 1$, the capacity of the BSC is approached for both of the achievable rates. In this example, we cannot achieve a positive rate if $\alpha \geq 2$, even for the case of a noiseless channel ($p = 0$). However, this is not true in general, because even the first achievable rate can be positive for a large α, if the capacity of the channel C_W is sufficiently large. The results suggest that, as communication becomes more intermittent and α becomes larger, the achievable rate is decreased due to the additional uncertainty about the positions of the codeword symbols at the decoder.

IV. Upper Bounds

In this section, we focus on obtaining upper bounds on the capacity of a special case of intermittent communication in which the DMC in Figure 2 is binary-input binary-output noiseless

A \ B	000	001	010	100	011	101	110	111
00	1	0	0	0	0	0	0	0
01	0	2/3	1/3	0	0	0	0	0
10	0	0	1/3	2/3	0	0	0	0
11	0	0	0	0	1/3	1/3	1/3	0

TABLE I: Transition probabilities $P(\mathbf{B}|\mathbf{A})$ for the auxiliary channel with $a = 2$ and $b = 3$.

with the noise symbol $\star = 0$. The achievable rate for this case is denoted by $R_{\text{Insertion}}$ in Section III-C, and is shown by the blue curve in Figure 4. Similar to [17], upper bounds are obtained by providing the encoder and the decoder with various amounts of side-information, and calculating or upper bounding the capacity of this genie-aided system. After introducing a useful function $g(a, b)$ in Section IV-A, we obtain upper bounds in Section IV-B. The techniques of in this section can in principle be applied to non-binary and noisy channels as well; however, the computational complexity for numerical evaluation of the genie-aided system grows very rapidly in the size of the alphabets.

A. Auxiliary Channel: Uniform Insertion

Let a and b be two integer numbers such that $0 \leq a \leq b$, and consider a discrete memoryless channel for which at each channel use the input consists of a sequence of a bits and the output consists of a sequence of b bits, i.e., the input and output of this channel are $\mathbf{A} \in \{0, 1\}^a$ and $\mathbf{B} \in \{0, 1\}^b$, respectively. At each channel use, $b - a$ zeroes are inserted randomly and uniformly among the input symbols. The positions at which the insertions occur randomly takes on each of the possible $\binom{b}{b-a} = \binom{b}{a}$ realizations with equal probability, and is unknown to the transmitter and the receiver. As an example, the transition probability matrix of this channel for the case of $a = 2$ and $b = 3$ is reported in Table I.

The capacity of the auxiliary channel is defined as

$$g(a, b) := \max_{P(\mathbf{A})} \mathbb{I}(\mathbf{A}; \mathbf{B}), \quad 0 \leq a \leq b, \tag{21}$$

where $P(\mathbf{A})$ is the input distribution. The exact value of the function $g(a, b)$ for finite a and b can be numerically computed by evaluating the transition probabilities $P(\mathbf{B}|\mathbf{A})$ and using the Blahut algorithm [26] to maximize the mutual information. The computational complexity of the Blahut algorithm increases exponentially for large values of a and b since the transition probability matrix is of size $2^a \times 2^b$. In order to partially overcome this issue, we recall the following lemma.

Lemma 5. *([27, Problem 7.28]): Consider a channel that is the union of i memoryless channels $(\mathcal{X}_1, P_1(y_1|x_1), \mathcal{Y}_1), ..., (\mathcal{X}_i, P_i(y_i|x_i), \mathcal{Y}_i)$ with capacities $C_1, ..., C_i$, where at each time one can send a symbol over exactly one of the channels. If the output alphabets are distinct and do not intersect, then the capacity C of this channel can be characterized in terms of $C_1, ..., C_i$ in bits per channel use as*

$$2^C = 2^{C_1} + ... + 2^{C_i}.$$

Now, notice that the function $g(a, b)$ can be evaluated by considering the union of $a + 1$ memoryless channels with distinct input and output alphabets, where the input of the i^{th} channel is binary sequences with length a and weight $i - 1$, $i = 1, ..., a + 1$, and the output is binary sequences with length b obtained from the input sequence by inserting $b-a$ zeroes uniformly. The weight of the sequences remains fixed after zero insertions, and therefore, the output alphabets are also distinct and do not intersect. Assuming that the capacity of the i^{th} channel is $g_i(a, b)$ and applying Lemma 5, we have

$$2^{g(a,b)} = 2^{g_1(a,b)} + ... + 2^{g_{a+1}(a,b)}. \tag{22}$$

It is easy to see that $g_1(a, b) = 0$ and $g_{a+1}(a, b) = 0$. For other values of i, the capacity $g_i(a, b), i = 2, ..., a$ can be evaluated numerically using the Blahut algorithm, where input and output alphabets have sizes $\binom{a}{i-1}$ and $\binom{b}{i-1}$, respectively, which are considerably less than those of the original alphabet sizes. This reduction allows us to obtain the function $g(a, b)$ for larger values of a and b, which will be useful in Section IV-B. The largest value of b for which we are able to evaluate the function $g(a, b)$ for all values of $a \leq b$ is $b = 17$. Although we cannot obtain a closed-form expression for the function $g(a, b)$, we can find a closed-form upper bound by expanding the mutual information in (21) and bounding some of its terms. As a result, we can find upper bounds on the function $g(a, b)$ for larger values of a and b. First, we introduce some notation.

For a binary sequence $x^a \in \{0, 1\}^a$, let $w(x^a)$ denote the weight of the sequence, i.e., number of 1's. Also, let the vector $r^0(x^a) := (r_1^0, ..., r_{l_0}^0)$ of length l_0 denote the length of consecutive

0's in the sequence x^a such that $r_1^0 = 0$ if $x_1 = 1$, i.e., the binary sequence starts with 1, and $r_{l_0}^0 = 0$ if $x_k = 1$, i.e., the binary sequence ends with 1, and all the other elements of the vector $r^0(x^a)$ are positive integers. In addition, let the vector $r^1(x^a) := (r_1^1, ..., r_{l_1}^1)$ of length l_1 denote the length of consecutive 1's in the sequence x^a with length larger than one, i.e., runs of 1's with length one not counted. Finally, let $l(x^a) := l_0 + l_1$. If it is clear from the context, we drop the argument x^a from these functions. For example, if $x^a = 0010111000011$, then $w = 6$, $r^0 = (2, 1, 4, 0)$, $r^1 = (3, 2)$, and $l = 4 + 2 = 6$. As another example, if $x^a = 10011101$, then $w = 5$, $r^0 = (0, 2, 1, 0)$, $r^1 = (3)$, and $l = 4 + 1 = 5$. Now, we define the following function, which will be used for expressing the upper bound,

$$F(x^a) := \sum_{i_1,...,i_l \in \mathbb{N} \cup \{0\} : \sum_{j=1}^{l} i_j = b-a} p_{i_1,...,i_l}(r^0, r^1) h_{i_1,...,i_l}(r^0), \tag{23}$$

where l, r^0, and r^1 are a function of x^a as defined before, and we have

$$p_{i_1,...,i_l}(r^0, r^1) := \frac{1}{\binom{b}{a}} \prod_{j=1}^{l_0} \binom{r_j^0 + i_j}{i_j} \prod_{j=1}^{l_1} \binom{r_j^1 - 2 + i_{j+l_0}}{i_{j+l_0}}, \tag{24}$$

$$h_{i_1,...,i_l}(r^0) := \sum_{j=1}^{l_0} \log \binom{r_j^0 + i_j}{i_j}. \tag{25}$$

Proposition 5. *The function $g(a, b)$ in (21) satisfies*

$$g(a, b) \leq \log \sum_{j=0}^{a} \binom{b}{j} 2^{\max_{x^a : w(x^a) = j} F(x^a)} - \log \binom{b}{a} \tag{26}$$

where $F(\cdot)$ is defined in (23).

Proof: See Appendix D. ∎

Similarly, it is possible to obtain a lower bound on the function $g(a, b)$. The lower bound and a numerical comparison between the lower bound, the exact value, and the upper bound on the function $g(a, b)$ can be found in [28]. The following definition will be useful in expressing the upper bounds in Section IV-B.

$$\phi(a, b) := a - g(a, b), \tag{27}$$

Note that the function $\phi(a, b)$ quantifies the loss in capacity due to the uncertainty about the positions of the insertions, and cannot be negative. The following proposition characterizes some of the properties of the functions $g(a, b)$ and $\phi(a, b)$, which will be used later.

Proposition 6. *The functions $g(a,b)$ and $\phi(a,b)$ have the following properties:*

(a) $g(a,b) \leq a,\ \phi(a,b) \geq 0.$

(b) $g(a,a) = a,\ \phi(a,a) = 0.$

(c) $g(1,b) = 1,\ \phi(1,b) = 0.$

(d) $g(a,b+1) \leq g(a,b),\ \phi(a,b+1) \geq \phi(a,b).$

(e) $g(a+1,b+1) \leq 1 + g(a,b),\ \phi(a+1,b+1) \geq \phi(a,b).$

Proof: See Appendix E. ∎

B. Genie-Aided System and Numerical Upper Bounds

In this section, we focus on upper bounds on the capacity of binary-input binary-output noiseless intermittent communication. The procedure is similar to [17]. Specifically, we obtain upper bounds by giving some kind of side-information to the encoder and decoder, and calculating or upper bounding the capacity of this genie-aided channel.

Now we introduce one form of side-information. Assume that the position of the $[(s+1)i]^{th}$ codeword symbol in the output sequence is given to the encoder and decoder for all $i = 1, 2, ...$ and a fixed integer number $s \geq 1$. We assume that the codeword length is a multiple of $s+1$, so that $t = k/(s+1)$ is an integer, and is equal to the total number of positions that are provided as side-information. This assumption does not impact the asymptotic behavior of the channel as $k \to \infty$. We define the random sequence $\{Z_i\}_{i=1}^{t}$ as follows: Z_1 is equal to the position of the $[s+1]^{th}$ codeword symbol in the output sequence, and for $i \in \{2, 3, ..., t\}$, Z_i is equal to the difference between the positions of the $[(s+1)i]^{th}$ codeword symbol and $[(s+1)(i-1)]^{th}$ codeword symbol in the output sequence.

Since we assumed iid insertions, the random sequence $\{Z_i\}_{i=1}^{t}$ is iid as well with negative binomial distribution:

$$P(Z_i = b+1) = \binom{b}{s}(1-p_t)^{b-s} p_t^{s+1}, b \geq s, \tag{28}$$

with mean $\mathbb{E}[Z_i] = (s+1)/p_t$. Also, note that as $k \to \infty$, by the law of large numbers, we have

$$\frac{N}{t} \xrightarrow{p} \mathbb{E}[Z_i] = \frac{s+1}{p_t}. \tag{29}$$

Let C_1 denote the capacity of the channel if we provide the encoder and decoder with side-information on the random sequence $\{Z_i\}_{i=1}^{t}$, which is clearly an upper bound on the capacity of the original channel. With this side-information, we essentially partition the transmitted and received sequences into t contiguous blocks that are independent from each other. In the i^{th} block the place of the $[s+1]^{th}$ codeword symbol is given, which can convey one bit of information. Other than that, the i^{th} block has s input bits and $Z_i - 1$ output bits with uniform 0 insertions. Therefore, the information that can be conveyed through the i^{th} block equals $g(s, Z_i - 1) + 1$. Thus, we have

$$
\begin{aligned}
C_1 &= \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{t} g(s, Z_i - 1) + 1 \\
&= \lim_{k\to\infty} \frac{N}{k} \frac{t}{N} \frac{1}{t} \sum_{i=1}^{t} g(s, Z_i - 1) + 1 \\
&= \frac{1}{s+1} \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} g(s, Z_i - 1) + 1 && (30) \\
&= \frac{1}{s+1} \mathbb{E}\left[g(s, Z_i - 1) + 1\right] && (31) \\
&= \frac{1}{s+1} \left[1 + \sum_{b=s}^{\infty} \binom{b}{s} (1-p_t)^{b-s} p_t^{s+1} g(s,b)\right] && (32) \\
&= 1 - \frac{1}{s+1} \sum_{b=s}^{\infty} \binom{b}{s} (1-p_t)^{b-s} p_t^{s+1} \phi(s,b), && (33)
\end{aligned}
$$

where: (30) follows from (29); (31) follows from the law of large numbers; (32) follows from the distribution of Z_i's given in (28); and (33) follows from the definition (27). Note that the capacity C_1 cannot be larger than 1, since the coefficients $\phi(\cdot,\cdot)$ cannot be negative. The negative term in (33) can be interpreted as a lower bound on the communication overhead as the cost of intermittency in the context of [25].

The expression in (33) gives an upper bound on the capacity of the original channel with $p_t = 1/\alpha$. However, it is infeasible to numerically evaluate the coefficients $\phi(s,b)$ for large values of b. As we discussed before, the largest value of b for which we are able to evaluate the function $\phi(s,b)$ is $b_{max} = 17$. The following upper bound on C_1 results by truncating the summation in (33) and using part (d) of Proposition 6.

$$
C_1 \leq 1 - \frac{\phi(s, b_{max})}{s+1} + \frac{1}{s+1} \sum_{b=s}^{b_{max}} \binom{b}{s} p_t^{s+1} (1-p_t)^{b-s} \left(\phi(s, b_{max}) - \phi(s,b)\right), \tag{34}
$$

The expression (34), which we denote by C_1', gives a nontrivial and computable upper bound for each value of $s = 2, 3, ..., b_{max} - 1$ on C_1, and therefore, an upper bound on the capacity

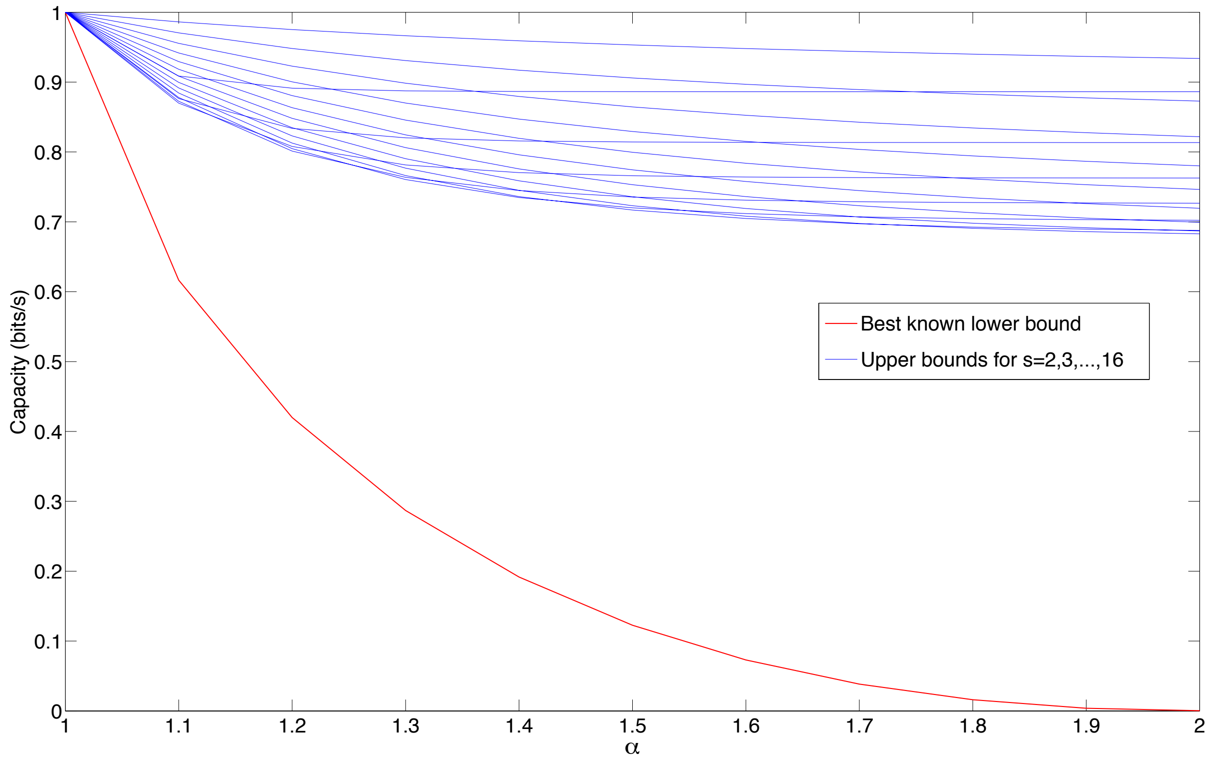


Fig. 5: Comparison between the best achievability result with different upper bounds obtained from (34) for $b_{max} = 17$ and $s = 2, 3, ..., 16$, versus the intermittency rate α.

of the original channel with $p_t = 1/\alpha$. Figure 5 shows the upper bounds for $b_{max} = 17$ and $s = 2, 3, ..., 16$ versus the intermittency rate α, along with the the achievability result.

Next, we introduce a second form of side-information. Assume that for consecutive blocks of length s of the output sequence, the number of codeword symbols within that block is given to the encoder and decoder as side-information, i.e., the number of codeword symbols in the sequence $(y_{(i-1)s+1}, y_{(i-1)s+2}, ..., y_{is})$, $i = 1, 2, ...$ for a fixed integer number $s \geq 2$. Let C_2 denote the capacity of the channel if we provide the encoder and decoder with this side-information. Using a similar procedure, we obtain

$$C_2 = 1 - \frac{1}{sp_t} \sum_{a=0}^{s} \binom{s}{a} p_t^a (1 - p_t)^{s-a} \phi(a, s). \tag{35}$$

Note that the summation in (35) is finite, and we do not need to upper bound C_2 as we did

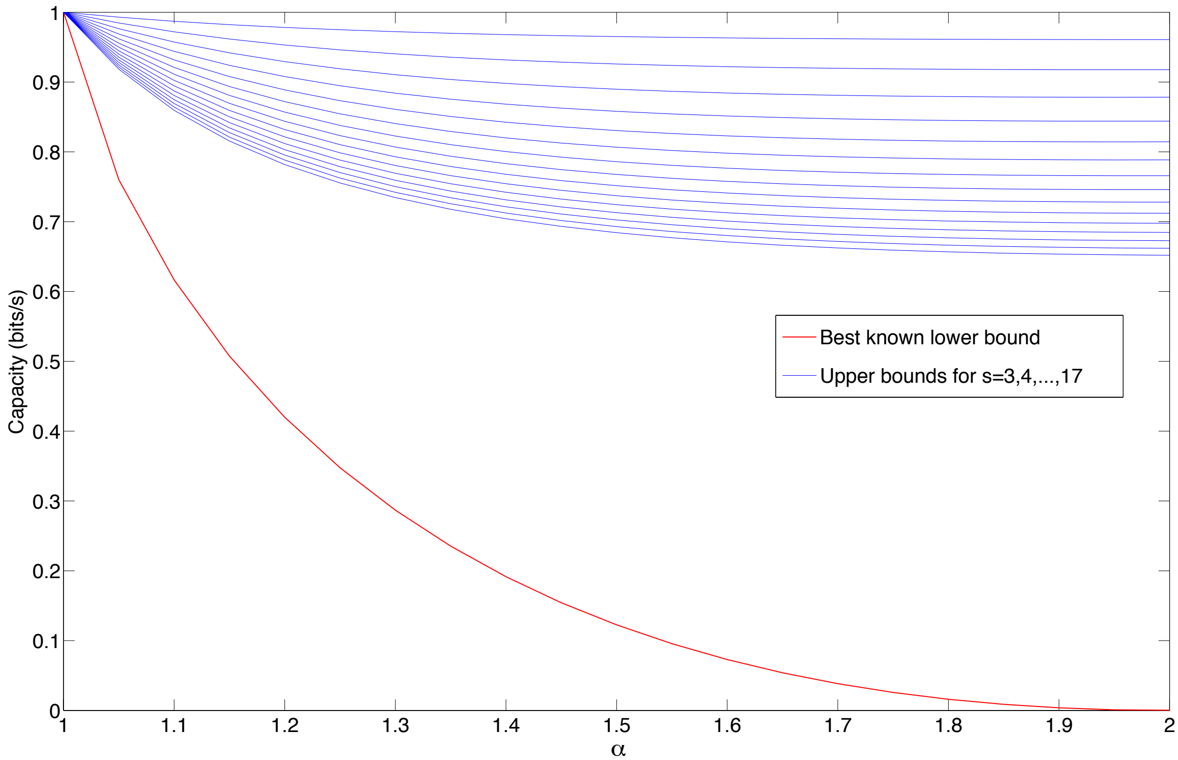


Fig. 6: Comparison between the best achievability result with different upper bounds obtained from (35) for $s = 3, 4, ..., 17$, versus the intermittency rate α.

for C_1. The value of C_2 gives nontrivial and computable upper bounds on the capacity of the original channel. Figure 6 shows the upper bounds for $s = 3, 4, ..., 17$ versus the intermittency rate α, along with the the achievability result. The upper bound corresponding to $s = 17$ is tighter than others for all ranges of α, i.e., (35) is decreasing in s. Intuitively, this is because by decreasing s, we provide the side-information more frequently, and therefore, the capacity of the resulting genie-aided system becomes larger.

It seems that (35) gives better upper bounds for the range of α shown in the figures ($1 < \alpha \leq 2$). However, the other upper bound C_1' can give better results for the limiting values of $\alpha \to \infty$

or $p_t \to 0$. We have

$$\lim_{\alpha\to\infty} C_1' = 1 - \frac{\phi(s, b_{max})}{s+1}, \tag{36}$$

$$\lim_{\alpha\to\infty} C_2 = 1.$$

This is because of the fact that by increasing α, and thus decreasing p_t, we have more zero insertions and the first kind of genie-aided system provides side-information less frequently leading to tighter upper bounds. The best upper bound for the limiting case of $\alpha \to \infty$ found by (36) is 0.6739 bits/s. In principle, we can use the upper bound on $g(a, b)$ in Proposition 5 to upper bound C_1 and C_2. By doing so, we can find the bounds for larger values of s and b_{max}, because we can calculate the upper bound (26) for larger arguments. It seems that this does not improve the upper bounds significantly for the range of α shown in the figures. However, by upper bounding (36) via (26), we can tighten the upper bound for the limiting case of $\alpha \to \infty$ to 0.6307 bits/s.

Although the gap between the achievable rates and upper bounds is not particularly tight, especially for large values of intermittency rate α, the upper bounds suggest that the linear scaling of the receive window with respect to the codeword length considered in the system model is natural since there is a tradeoff between the capacity of the channel and the intermittency rate. By contrast, in asynchronous communication [1], [2], where the transmission of the codeword is contiguous, only exponential scaling $n = e^{\alpha k}$ induces a tradeoff between capacity and asynchronism.

V. Bounds on Capacity Per Unit Cost

In this section, we obtain bounds on the capacity per unit cost of intermittent communication. Let $\gamma : \mathcal{X} \to [0, \infty]$ be a cost function that assigns a non-negative value to each channel input. We assume that the noise symbol has zero cost, i.e., $\gamma(\star) = 0$. The cost of a codeword is defined as

$$\Gamma(c^k(m)) = \sum_{i=1}^{k} \gamma(c_i(m)).$$

A (k, M, P, ε) code consists of M codewords of length k, $c^k(m), m \in \{1, 2, ..., M\}$, each having cost at most P with average probability of decoding error at most ε, where the intermittent process is the same as in Section II. Note that the cost of the input and output sequences of the intermittent process shown in Figure 2 is the same since the cost of the noise symbols is zero. We say rate $\hat{R}$ bits per unit cost is achievable if for every $\varepsilon > 0$ and large enough M there exists a (k, M, P, ε) code with $\log(M)/P \geq \hat{R}$. For intermittent communication $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$, the capacity per unit cost $\hat{C}_\alpha$ is the supremum of achievable rates per unit cost.

It is shown in [29] that the capacity per unit cost of a general DMC is

$$\max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(W_x \| W_\star)}{\gamma(x)},$$

where we assumed that $\gamma(\star) = 0$, and the optimization is over the input alphabet instead of over the set of all input distributions. The asynchronous capacity per unit cost for asynchronous communication with timing uncertainty per information bit β has been shown to be [3]

$$\frac{1}{1+\beta} \max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(W_x \| W_\star)}{\gamma(x)}.$$

Therefore, comparing to the capacity per unit cost of a DMC, the rate is penalized by a factor of $1/(1+\beta)$ due to asynchronism. For a channel with iid synchronization errors with average number of duplications equal to μ concatenated with a DMC, bounds on the capacity per unit cost, $\hat{C}_\mu$, have been obtained in [20] as

$$\frac{\mu}{2} \max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(W_x \| W_\star)}{\gamma(x)} \leq \hat{C}_\mu \leq \mu \max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(W_x \| W_\star)}{\gamma(x)},$$

where the lower bound is obtained by using a type of pulse position modulation at the encoder and searching for the position of the pulse at the decoder. Using similar encoding and decoding schemes, we obtain a lower bound for the capacity per unit cost of intermittent communication.

Theorem 6. *The capacity per unit cost $\hat{C}_\alpha$ for intermittent communication $(\mathcal{X}, \mathcal{Y}, W, \star, \alpha)$ satisfies*

$$\frac{\alpha}{2} \max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(\frac{1}{\alpha} W_x + (1 - \frac{1}{\alpha}) W_\star \| W_\star)}{\gamma(x)} \leq \hat{C}_\alpha \leq \max_{x \in \mathcal{X} \setminus \{\star\}} \frac{D(W_x \| W_\star)}{\gamma(x)} \tag{37}$$

Sketch of the Proof: The upper bound in (37) is the capacity per unit cost of the DMC W, and follows by providing the decoder with side-information about the positions of inserted noise symbols $\star$. The derivation of the lower bound is similar to the one in [20, Theorem 3]. Essentially, the encoder uses pulse position modulation, i.e., to transmit message m, it transmits a burst of symbols x of length

$$B := \frac{2 \log(M)}{\alpha D(\frac{1}{\alpha} W_x + (1 - \frac{1}{\alpha}) W_\star \| W_\star)},$$

at a position corresponding to this message and transmits the zero-cost noise symbol $\star$ at the other $k-B$ positions before and after this burst, so that each codeword has cost $P = B\gamma(x)$. In order to decode the message, we search for the location of the pulse using a sliding window with an appropriate length looking for a subsequence that has a type equal to $\frac{1}{\alpha}W_x + (1-\frac{1}{\alpha})W_\star$, because at the receiver, we expect to have approximately $B(\alpha-1)$ inserted noise symbols $\star$ in between the B burst symbols x. Similar to the analysis of [20, Theorem 3], it can be shown that the probability of decoding error vanishes as $M \to \infty$, and the rate per unit cost is

$$\hat{R} = \frac{\log(M)}{P} = \frac{\alpha}{2}\frac{D(\frac{1}{\alpha}W_x + (1-\frac{1}{\alpha})W_\star \| W_\star)}{\gamma(x)}.$$

Finally, by choosing the optimum input symbol x, rate per unit cost equal to the left-hand side of (37) can be achieved. □

From the convexity of the Kullback-Leibler divergence, it can be seen that the lower bound is always smaller than half of the upper bound. Consider the BSC example with crossover probability $p = 0.1$ and input costs $\gamma(\star = 0) = 0$ and $\gamma(1) = 1$. The upper bound in (37) equals 2.536 bits per unit cost, and the lower bound in (37) is plotted in Figure 7 versus the intermittency rate α. As we would expect, the lower bound decreases as the intermittency increases.

VI. Conclusion

We formulated a model for intermittent communications that can capture bursty transmissions or a sporadically available channel by inserting a random number of silent symbols between each codeword symbol so that the receiver does not know a priori when the transmissions will occur. First, we specified two decoding schemes in order to develop achievable rates. Interestingly, decoding from pattern detection, which achieves a larger rate, is based on a generalization of the method of types and properties of partial divergence. As the system becomes more intermittent, the achievable rates decrease due to the additional uncertainty about the positions of the codeword symbols at the decoder. We also showed that as long as the intermittency rate α is finite and the capacity of the DMC is not zero, rate $R = 0$ is achievable for intermittent communication. For the case of binary-input binary-output noiseless channel, we obtained upper bounds on the capacity of intermittent communication by providing the encoder and the decoder with various amounts of side-information, and calculating or upper bounding the capacity of this genie-aided system. The results suggest that the linear scaling of the receive window with respect to the codeword length

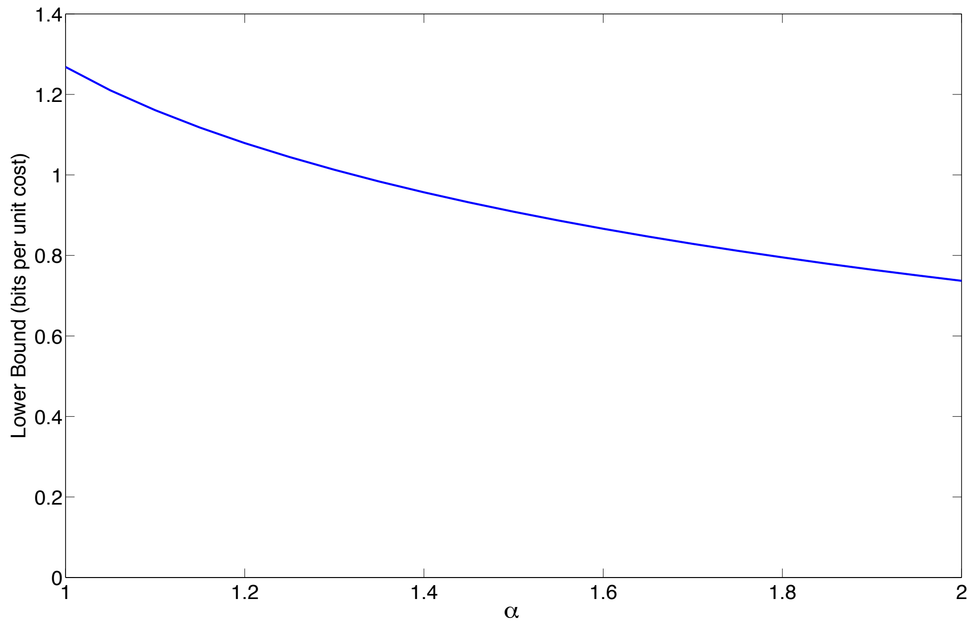


Fig. 7: The lower bound on the capacity per unit cost of intermittent communication versus the intermittency rate α.

considered in the system model is relevant since the upper bounds imply a tradeoff between the capacity and the intermittency rate. Finally, we derived bounds on the capacity per unit cost of intermittent communication. To obtain the lower bound, we used pulse-position modulation at the encoder, and searched for the position of the pulse at the decoder.

Appendix A
Proof of Proposition 2

(a) From (14), we observe that $c^* \to 0$ as $\rho \to 0$. Therefore, from (13), we have

$$d_0(P\|Q) = D(P\|Q) - \sum_{j=1}^{|\mathcal{X}|} p_j \log \frac{p_j}{q_j} + c^* \log c^* + h(0) = 0,$$

for $c^* \to 0$.

(b) From (14), we observe that $c^* \to \infty$ as $\rho \to 1$. Therefore, from (13), we have

$$d_1(P\|Q) = D(P\|Q) - \sum_{j=1}^{|\mathcal{X}|} p_j \log(1 + \frac{p_j}{c^* q_j}) + h(1) = D(P\|Q),$$

for $c^* \to \infty$.

(c) If $P = Q$, then (14) simplifies to $\frac{c^*}{c^*+1} = \rho$, and therefore $c^* = \rho/\bar{\rho}$. By substituting c^* into (13) and because $D(P\|P) = 0$, we obtain

$$d_\rho(P\|P) = -\log(\frac{\rho}{\bar{\rho}} + 1) + \rho \log \frac{\rho}{\bar{\rho}} + h(\rho) = 0.$$

(d) By taking the derivative of (13) with respect to ρ, we obtain

$$\begin{aligned} d'_\rho(P\|Q) &= -\frac{\partial c^*}{\partial \rho} \sum_{j=1}^{|\mathcal{X}|} \frac{p_j q_j}{c^* q_j + p_j} + \log c^* + \frac{\partial c^*}{\partial \rho} \frac{\rho}{c^*} + \log \frac{\bar{\rho}}{\rho} \\ &= \frac{\partial c^*}{\partial \rho} (\frac{\rho}{c^*} - \sum_{j=1}^{|\mathcal{X}|} \frac{p_j q_j}{c^* q_j + p_j}) + \log(c^* \frac{\bar{\rho}}{\rho}) \\ &= \log(c^* \frac{\bar{\rho}}{\rho}), \end{aligned} \tag{38}$$

where (38) is obtained by using (14). Therefore,

$$d'_0(P\|Q) = \lim_{\rho \to 0} d'_\rho(P\|Q) = \lim_{\rho \to 0} \log(c^* \frac{\bar{\rho}}{\rho}) = 0,$$

because from (14), we have $\lim_{\rho \to 0} \frac{\rho}{c^*} = \sum_{j=1}^{|\mathcal{X}|} \frac{p_j q_j}{p_j} = 1$.

(e) According to (38), in order to prove $d'_\rho(P\|Q) > 0, 0 < \rho \leq 1$, it is enough to show that $\frac{\rho}{\bar{\rho}} < c^*, 0 < \rho \leq 1$:

$$\begin{aligned} 1 &= (\sum_{j=1}^{|\mathcal{X}|} p_j)^2 \\ &< \sum_{j=1}^{|\mathcal{X}|} (\sqrt{c^* q_j + p_j})^2 \cdot \sum_{j=1}^{|\mathcal{X}|} (\frac{p_j}{\sqrt{c^* q_j + p_j}})^2 \qquad (39) \\ &= (c^* + 1) \sum_{j=1}^{|\mathcal{X}|} \frac{p_j^2}{c^* q_j + p_j} \\ &= (c^* + 1)\bar{\rho}, \qquad (40) \end{aligned}$$

where (39) holds according to the Cauchy-Schwarz inequality, and (40) is true because

$$\bar{\rho} = 1 - \rho = \sum_{j=1}^{|\mathcal{X}|} p_j - \rho$$
$$= \sum_{j=1}^{|\mathcal{X}|} p_j - \sum_{j=1}^{|\mathcal{X}|} \frac{c^* p_j q_j}{c^* q_j + p_j} \tag{41}$$
$$= \sum_{j=1}^{|\mathcal{X}|} \frac{p_j^2}{c^* q_j + p_j},$$

where (41) comes from (14). Note that the Cauchy-Schwarz inequality in (39) cannot hold with equality for $0 < \rho$ (and therefore, for $0 < c^*$), because otherwise, $p_j = q_j, j = 1, 2, ..., |\mathcal{X}|$, and $P = Q$. From (40), $1 < (c^* + 1)\bar{\rho}$, which results in the desirable inequality $\frac{\rho}{\bar{\rho}} < c^*$.

(f) By taking the derivative of (38) with respect to ρ, it can easily be seen that

$$d''_\rho(P\|Q) = \frac{1}{c^*}\frac{\partial c^*}{\partial \rho} - \frac{1}{\rho\bar{\rho}}. \tag{42}$$

Also, by taking the derivative of (14) with respect to ρ and after some calculation, we have

$$\sum_{j=1}^{|\mathcal{X}|} \frac{(c^* q_j)^2 p_j}{(c^* q_j + p_j)^2} = \rho - \frac{c^*}{\frac{\partial c^*}{\partial \rho}}. \tag{43}$$

Therefore,

$$\rho^2 = \left(\sum_{j=1}^{|\mathcal{X}|} \frac{c^* p_j q_j}{c^* q_j + p_j} \right)^2$$
$$< \sum_{j=1}^{|\mathcal{X}|} \frac{(c^* q_j)^2 p_j}{(c^* q_j + p_j)^2} \cdot \sum_{j=1}^{|\mathcal{X}|} p_j \tag{44}$$
$$= \rho - \frac{c^*}{\frac{\partial c^*}{\partial \rho}}, \tag{45}$$

where (44) holds according to the Cauchy-Schwarz inequality, which cannot hold with equality since otherwise $P = Q$, and where (45) follows from (43). From (45), $\rho^2 < \rho - \frac{c^*}{\frac{\partial c^*}{\partial \rho}}$, which implies that $d''_\rho(P\|Q) > 0$ according to (42).

(g) From part (f), $d_\rho(P\|Q)$ is convex in ρ, and therefore, $\frac{d_\rho(P\|Q)}{\rho}$ is increasing in ρ. In addition, from part (d), $\lim_{\rho\to 0} \frac{d_\rho(P\|Q)}{\rho} = d'_0(P\|Q) = 0$, and from part (b), $\lim_{\rho\to 1} \frac{d_\rho(P\|Q)}{\rho} = D(P\|Q)$. Consequently, $0 \le d_\rho(P\|Q) \le \rho D(P\|Q)$.

APPENDIX B
PROOF OF THEOREM 5

Fix the input distribution P, and consider decoding from pattern detection described in Sections III-B. For any $\epsilon > 0$, we prove that if $R = \mathbb{I}(P, W) - f(P, W, \alpha) - 2\epsilon$, then the average probability of error vanishes as $k \to \infty$. We have

$$p_e^{avg} \leq \mathbb{P}(\hat{m} \in \{2, 3, ..., M\} | m = 1) + \mathbb{P}(\hat{m} = e | m = 1), \tag{46}$$

where (46) follows from the union bound in which the second term is the probability that the decoder declares an error (does not find any message) at the end of all $\binom{n}{k}$ choices, which implies that even if we pick the correct output symbols, the decoder either does not pass the first stage or does not declare $m = 1$ in the second stage. Therefore,

$$\mathbb{P}(\hat{m} = e | m = 1) \leq \mathbb{P}(Y^k \notin T_{[PW]_\mu}) + \mathbb{P}(Y_\star^{n-k} \notin T_{[W_\star]_\mu}) + \mathbb{P}(Y^k \notin T_{[W]_\mu}(C^k(1))) \tag{47}$$

$$\to 0, \text{ as } k \to \infty, \tag{48}$$

where Y^k is the output of the channel if the input is $C^k(1)$, and $Y_\star$ is the output of the channel if the input is the noise symbol, and where we use the union bound to establish (47). The limit (48) follows because all the three terms in (47) vanish as $k \to \infty$ according to Lemma 2.

The first term in (46) is more challenging. It is the probability that for at least one choice of the output symbols, the decoder passes the first stage and then the typicality decoder declares an incorrect message. We characterize the $\binom{n}{k}$ choices based on the number of incorrectly chosen output symbols, i.e., the number of symbols in $\tilde{y}^k$ that are in fact output symbols corresponding to a noise symbol, which is equal to the number of symbols in $\hat{y}^{n-k}$ that are in fact output symbols corresponding to a codeword symbol. For any $0 \leq k_1 \leq n - k$, there are $\binom{k}{k_1}\binom{n-k}{k_1}$ choices. [1] Using the union bound for all the choices and all the messages $\hat{m} \neq 1$, we have

$$\mathbb{P}(\hat{m} \in \{2, 3, ..., M\} | m = 1) \leq (e^{kR} - 1) \sum_{k_1=0}^{n-k} \binom{k}{k_1} \binom{n-k}{k_1} \mathbb{P}_{k_1}(\hat{m} = 2 | m = 1), \tag{49}$$

where the index k_1 in (49) denotes the condition that the number of wrongly chosen output symbols is equal to k_1. Note that message $\hat{m} = 2$ is declared at the decoder only if it passes the

[1] According to Vandermonde's identity, we have $\sum_{k_1=0}^{n-k} \binom{k}{k_1}\binom{n-k}{k_1} = \binom{n}{k}$.

first and the second stage. Therefore,

$$\begin{aligned}
&\mathbb{P}_{k_1}(\hat{m}=2|m=1)\\
&=\mathbb{P}_{k_1}\left(\{\tilde{Y}^k\in T_{[PW]_\mu}\}\cap\{\hat{Y}^{n-k}\in T_{[W_\star]_\mu}\}\cap\{\tilde{Y}^k\in T_{[W]_\mu}(C^k(2))\}|m=1\right)\\
&=\mathbb{P}_{k_1}(\tilde{Y}^k\in T_{[PW]_\mu})\cdot\mathbb{P}_{k_1}(\hat{Y}^{n-k}\in T_{[W_\star]_\mu})\cdot\mathbb{P}(\tilde{Y}^k\in T_{[W]_\mu}(C^k(2))|m{=}1,\tilde{Y}^k{\in}T_{[PW]_\mu},\hat{Y}^{n-k}{\in}T_{[W_\star]_\mu}) \quad (50)\\
&\le e^{o(k)}e^{-kd_{k_1/k}(PW\|W_\star)}e^{-(n-k)d_{k_1/(n-k)}(W_\star\|PW)}e^{-k(\mathbb{I}(P,W)-\epsilon)}, \quad (51)
\end{aligned}$$

where (50) follows from the independence of the events $\{\tilde{Y}^k\in T_{[PW]_\mu}\}$ and $\{\hat{Y}^{n-k}\in T_{[W_\star]_\mu}\}$ conditioned on k_1 wrongly chosen output symbols, and (51) follows from using Lemma 4 for the first two terms in (50) with mismatch ratios k_1/k and $k_1/(n-k)$, respectively, and using Lemma 3 for the last term in (50), because conditioned on message $m=1$ being sent, $C^k(2)$ and $\tilde{Y}^k$ are independent regardless of the other conditions in the last term. Substituting (51) into the summation in (49), we have

$$\begin{aligned}
&\mathbb{P}(\hat{m}\in\{2,3,...,M\}|m=1)\\
&\le e^{o(k)}(e^{kR}-1)e^{-k(\mathbb{I}(P,W)-\epsilon)}\sum_{k_1=0}^{n-k}\binom{k}{k_1}\binom{n-k}{k_1}e^{-kd_{k_1/k}(PW\|W_\star)-(n-k)d_{k_1/(n-k)}(W_\star\|PW)} \quad (52)\\
&\le e^{o(k)}e^{kR}e^{-k(\mathbb{I}(P,W)-\epsilon)}e^{kf(P,W,\alpha)} \quad (53)\\
&=e^{o(k)}e^{-k\epsilon}\quad\to 0 \text{ as } k\to\infty, \quad (54)
\end{aligned}$$

where (54) is obtained by substituting $R=\mathbb{I}(P,W)-f(P,W,\alpha)-2\epsilon$, and where (53) is obtained by finding the exponent of the sum in (52) as follows

$$\begin{aligned}
&\lim_{k\to\infty}\frac{1}{k}\log\sum_{k_1=0}^{n-k}\tbinom{k}{k_1}\tbinom{n-k}{k_1}e^{-kd_{k_1/k}(PW\|W_\star)-(n-k)d_{k_1/(n-k)}(W_\star\|PW)}\\
&=\lim_{k\to\infty}\frac{1}{k}\log\sum_{k_1=0}^{n-k}\exp\{kh(\tfrac{k_1}{k})+(n-k)h(\tfrac{k_1}{n-k})-kd_{\frac{k_1}{k}}(PW\|W_\star)-(n-k)d_{\frac{k_1}{n-k}}(W_\star\|PW)\} \quad (55)\\
&=\lim_{k\to\infty}\frac{1}{k}\max_{k_1=0,\dots,n-k}\{kh(\tfrac{k_1}{k})+(n-k)h(\tfrac{k_1}{n-k})-kd_{\frac{k_1}{k}}(PW\|W_\star)-(n-k)d_{\frac{k_1}{n-k}}(W_\star\|PW)\} \quad (56)\\
&\le\max_{0\le\beta\le1}\{(\alpha-1)h(\beta)+h((\alpha-1)\beta)-d_{(\alpha-1)\beta}(PW\|W_\star)-(\alpha-1)d_\beta(W_\star\|PW)\} \quad (57)\\
&=f(P,W,\alpha), \quad (58)
\end{aligned}$$

where (55) follows by using Stirling's approximation for the binomial terms; where (56) follows by noticing that the exponent of the summation is equal to the largest exponent of each term in the summation, since the number of terms is polynomial in k; where (57) is obtained by letting $\beta := k_1/(n-k)$ $(0 \leq \beta \leq 1)$ and substituting $n = \alpha k$; and where (58) follows from the definition (20).

Now, combining (46), (48), and (54), we have $p_e^{avg} \to 0$ as $k \to \infty$, which proves the Theorem.

APPENDIX C
PROOF OF PROPOSITION 4

(a) The term $(\alpha-1)h(\beta)+h((\alpha-1)\beta)$ in (20) is maximized at $\beta = 1/\alpha$, because it is concave in β and its derivative with respect to β is zero at $1/\alpha$. Thus, this term is decreasing in β in the interval $[1/\alpha, 1]$. Also, note that the partial divergence terms in (20) are increasing with respect to β according to Proposition 2 (e). Therefore, the term in the max operator in (20) is decreasing in β in the interval $[1/\alpha, 1]$, and the maximum occurs in the interval $[0, 1/\alpha]$.

(b) The term in the max operator in (20) is concave in β, because $h(\beta)$ is concave in β and $d_\beta(\cdot\|\cdot)$ is convex in β according to Proposition 2 (f). Therefore, the term is maximized at point β^* where the derivative with respect to β is equal to zero. Thus, we have

$$\log\frac{1-\beta^*}{\beta^*} + \log\frac{1-(\alpha-1)\beta^*}{(\alpha-1)\beta^*} - \log c_1\frac{1-(\alpha-1)\beta^*}{(\alpha-1)\beta^*} - \log c_2\frac{1-\beta^*}{\beta^*} = 0, \tag{59}$$

where (38) is used to derive (59), and where c_1 and c_2 are the corresponding c^*'s in (14) for the two partial divergence terms in (20). Taking derivative of (20) with respect to α assuming that the maximum occurs at β^*, we obtain

$$\begin{aligned}
&\frac{\partial f(P,W,\alpha)}{\partial\alpha}\\
&=\frac{\partial\beta^*}{\partial\alpha}(\alpha-1)(\cdot)+h(\beta^*)-d_{\beta^*}(W_\star\|PW)+\beta^*\Big(\log\frac{1-(\alpha-1)\beta^*}{(\alpha-1)\beta^*}-\log c_1\frac{1-(\alpha-1)\beta^*}{(\alpha-1)\beta^*}\Big)\\
&=h(\beta^*)-d_{\beta^*}(W_\star\|PW)+\beta^*\Big(\log c_2\frac{1-\beta^*}{\beta^*}-\log\frac{1-\beta^*}{\beta^*}\Big) &(60)\\
&=-\log(1-\beta^*)+\left(\frac{\partial d_{\beta^*}(W_\star\|PW)}{\partial\beta^*}-d_{\beta^*}(W_\star\|PW)\right) &(61)\\
&\geq 0, &(62)
\end{aligned}$$

where $(\cdot)$ in the first line is the left-side of (59), which is equal to zero, and where (60) follows from (59), and (61) follows from (38). Finally, (62) follows from the fact that

$-\log(1-\beta^*)$ is always positive for $0 \leq \beta^* \leq 1$ and $\partial d_{\beta^*}(W_\star \| PW)/\partial\beta^* - d_{\beta^*}(W_\star \| PW)$ is also always positive, because the partial divergence $d_{\beta^*}(W_\star \| PW)$ is convex in β^* according to Proposition 2 (f).

(c) Substituting $\alpha = 1$ in (20), all the terms would be zero, because $h(0) = 0$ and $d_0(P \| Q) = 0$ according to Proposition 2 (a).

(d) Consider that the maximum in (20) occurs at β^*. According to part (a), $0 \leq \beta^* \leq 1/\alpha$, and therefore, $\beta^* \to 0$ as $\alpha \to \infty$. Using Proposition 2 (d) and (59), we have $\alpha\beta^* \to 1$ as $\alpha \to \infty$. Substitiuting $\alpha = 1/\beta^*$ in (20), we obtain

$$\begin{aligned}
&\lim_{\alpha\to\infty} f(P, W, \alpha) \\
&= \lim_{\beta^*\to 0} \frac{h(\beta^*)}{\beta^*} + h(1) - d_1(PW \| W_\star) - \frac{d_{\beta^*}(W_\star \| PW)}{\beta^*} \\
&= \lim_{\beta^*\to 0} \frac{h(\beta^*)}{\beta^*} - D(PW \| W_\star) - d_0'(P \| Q) && (63) \\
&= \lim_{\beta^*\to 0} \frac{h(\beta^*)}{\beta^*} - D(PW \| W_\star) && (64) \\
&\to \infty, && (65)
\end{aligned}$$

where (63) follows from Proposition 2 (b), and where (64) follows from Proposition 2 (d), and where (65) follows from the definition of the binary entropy function and the assumption that $D(PW \| W_\star)$ is finite.

(e) This part follows directly from the definition (20).

Appendix D
Proof of Proposition 5

Let P^b denote the random vector describing the positions of the insertions in the output sequence Y^b, such that $P_i = 1$ if and only if Y_i is one of the $b - a$ inserted 0's. We have

$$\begin{aligned}
\mathbb{I}(X^a; Y^b) &= H(X^a) - H(X^a | Y^b) \\
&= H(X^a) - H(P^b | Y^b) + H(P^b | X^a, Y^b), && (66)
\end{aligned}$$

where (66) follows by the general identity $H(X^a|Y^b) + H(P^b|X^a, Y^b) = H(P^b|Y^b) + H(X^a|P^b, Y^b)$ and noticing that for this choice of P^b, we have $H(X^a|P^b, Y^b) = 0$. For the term $H(P^b|X^a, Y^b)$

in (66), we have

$$H(P^b|X^a,Y^b) = \sum_{x^a} P(x^a) \sum_{y^b} P(y^b|x^a) H(P^b|X^a = x^a, Y^b = y^b) \tag{67}$$

$$= \sum_{x^a} P(x^a) F(x^a), \tag{68}$$

where $F(\cdot)$ is defined in (23); and where (68) is because instead of the summation over y^b, we can sum over the possible 0 insertions in between the runs of a fixed input sequence x^a such that there are total of $b-a$ insertions. If we denote the number of insertions in between the runs of zeros by $i_1, ..., i_{l_0}$, and the number of insertions in between the runs of ones by $i_{1+l_0}, ..., i_{l_1+l_0}$, then we have $i_1, ..., i_l \in \mathbb{Z}_{\geq 0} : \sum_{j=1}^{l} i_j = b-a$. Given these number of insertions, it is easy to see that $P(y^b|x^a)$ in (67) is equal to $p_{i_1,...,i_l}(r^0, r^1)$ in (24). Also, $H(P^b|X^a = x^a, Y^b = y^b)$ is equal to $h_{i_1,...,i_l}(r^0)$ in (25), because given the input and output sequences, the only uncertainty about the position sequence is where there is a run of zeros in the input sequence, i.e., for a run of ones, we know that all the zeros in between them are insertions. Also, the uncertainty is uniformly distributed over all the possible choices. Note that from (22), we have

$$g(a,b) = \log \sum_{j=0}^{a} 2^{\max_{P(x^a)} \mathbb{I}_j(X^a;Y^b)}, \tag{69}$$

where $\mathbb{I}_j(X^a;Y^b)$ denotes the mutual information if the input sequence, and therefore, the output sequence have weight j, and the maximization is over the distribution of all such input sequences. Using the chain rule, we have

$$H(P^b|Y^b) = H(Y^b|P^b) + H(P^b) - H(Y^b)$$

$$= H(X^a) + \log \binom{b}{a} - H(Y^b), \tag{70}$$

where (70) is because the entropy of the output sequence given the insertion positions equals the entropy of the input sequence, and because the entropy of the position sequence equals $\log \binom{b}{a}$ due to the uniform insertions. Combining (66), (68), and (70), we have

$$\mathbb{I}_j(X^a;Y^b) = H_j(Y^b) - \log \binom{b}{a} + \sum_{x^a : w(x^a) = j} P(x^a) F(x^a)$$

$$\leq \log \binom{b}{j} - \log \binom{b}{a} + \max_{x^a : w(x^a) = j} F(x^a), \tag{71}$$

where $H_j(Y^b)$ denotes the entropy of the output sequence if it has weight j; and where (71) follows from the fact that the uniform distribution maximizes the entropy and by maximizing $F(x^a)$ over all input sequences with weight j. Finally, by combining (69) and (71), we get the upper bound (26).

APPENDIX E
PROOF OF PROPOSITION 6

We prove the properties for the capacity function $g(a, b)$. The corresponding properties for the function $\phi(a, b)$ easily follows from (27).

(a) Since the cardinality of the input alphabet of this channel is 2^a, the capacity of this channel is at most a bits/s.

(b) There are no insertions. Therefore, it is a noiseless channel with input and output alphabets of sizes 2^a and capacity a bits/s.

(c) The input alphabet is $\{0, 1\}$, and the output consists of binary sequences with length b and weight 0 or 1, because only 0's can be inserted in the sequence. Considering all the output sequences with weight 1 as a super-symbol, the channel becomes binary noiseless with capacity 1 bits/s.

(d) The capacity $g(a, b+1)$ cannot decrease if, at each channel use, the decoder knows exactly one of the positions at which an insertion occurs, and the capacity of the channel with this genie-aided encoder and decoder becomes $g(a, b)$. Therefore, $g(a, b+1) \leq g(a, b)$.

(e) The capacity $g(a+1, b+1)$ cannot decrease if, at each channel use, the encoder and decoder know exactly one of the positions at which an input bit remains unchanged, so that it can be transmitted uncoded and the capacity of the channel with this genie-aided encoder and decoder becomes $1 + g(a, b)$. Therefore, $g(a+1, b+1) \leq 1 + g(a, b)$.

ACKNOWLEDGMENT

The authors wish to thank Dr. Aslan Tchamkerten and Shyam Kumar for the useful discussion, which leaded to a more compact expression and a shorter proof for the result in Lemma 4.

REFERENCES

[1] A. Tchamkerten, V. Chandar, and G. W. Wornell, "Communication Under Strong Asynchronism," *IEEE Trans. on Inf. Theory,* vol. 55, no. 10, pp. 4508-4528, Oct. 2009.

[2] ——, “Asynchronous Communication: Capacity Bounds and Suboptimality of Training,” *IEEE Trans. on Inf. Theory*, vol. 59, no. 3, pp. 1227-1255, Mar. 2013.

[3] V. Chandar, A. Tchamkerten, and D. N. C. Tse, “Asynchronous Capacity Per Unit Cost,” *IEEE Trans. on Inf. Theory*, vol. 59, no. 3, pp. 1213-1226, Mar. 2013.

[4] Y. Polyanskiy, “Asynchronous Communication: Exact Synchronization, Universality, and Dispersion,” *IEEE Trans. on Inf. Theory*, vol. 59, no. 3, pp. 1256-1270, Mar. 2013.

[5] V. Chandar, A. Tchamkerten, and G. Wornell, “Optimal Sequential Frame Synchronization,” *IEEE Trans. on Inf. Theory,* vol. 54, no. 8, pp. 3725-3728, Aug. 2008.

[6] A. Tchamkerten, V. Chandar, and G. Caire, “Energy and Sampling Constrained Asynchronous Communication,” *IEEE Trans. on Inf. Theory*, submitted 2013. [Online]. Available: http://arxiv.org/pdf/1302.6574v2.pdf

[7] D. Wang, V. Chandar, S.-Y. Chung, and G. Wornell, “Error Exponents in Asynchronous Communication,” in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, St. Petersburg, Russia, Aug. 2011.

[8] D. Wang, V. Chandar, S.-Y. Chung, and G. Wornell, “On Reliability Functions for Single-Message Unequal Error Protection,” in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Cambridge, MA, USA, July 2012.

[9] R. L. Dobrushin, “Shannon’s Theorems for Channels with Synchronization Errors,” *Problems of Information Transmission,* vol. 3, no. 4, pp. 11-26, 1967.

[10] R. Gallager, “Sequential Decoding for Binary Channels with Noise and Synchronization Errors,” *Tech. Rep., MIT Lincoln Lab. Group Report,* 1961.

[11] S. Diggavi and M. Grossglauser, “On Information Transmission over a Finite Buffer Channel,” *IEEE Trans. on Inf. Theory,* vol. 52, no. 3, pp. 1226-1237, Mar. 2006.

[12] E. Drinea and M. Mitzenmacher, “Improved Lower Bounds for the Capacity of iid Deletion and Duplication Channels,” *IEEE Trans. on Inf. Theory,* vol. 53, no. 8, pp. 2693-2714, Aug. 2007.

[13] A. Kirsch and E. Drinea, “Directly Lower Bounding the Information Capacity for Channels with iid Deletions and Duplications,” *IEEE Trans. on Inf. Theory,* vol. 56, no. 1, pp. 86-102, Jan. 2010.

[14] M. Mitzenmacher, “A Survey of Results for Deletion Channels and Related Synchronization Channels,” *Probability Surveys,* vol. 6, pp. 1-33, 2009.

[15] M. Mitzenmacher, “Capacity Bounds for Sticky Channels,” *IEEE Trans. on Inf. Theory*, vol. 54, pp. 72-77, Jan. 2008.

[16] R. Venkataramanan, S. Tatikonda, and K. Ramchandran, “Achievable Rates for Channels with Deletions and Insertions,” in *Proc. IEEE Int. Symp. Information Theory (ISIT),* St. Petersburg, Russia, Aug. 2011.

[17] D. Fertonani and T. M. Duman, “Novel Bounds on the Capacity of the Binary Deletion Channel,” *IEEE Trans. on Inf. Theory,* vol. 56, no. 6, pp. 2753-2765, 2010.

[18] D. Fertonani, T. M. Duman, and M. F. Erden, “Bounds on the Capacity of Channels with Insertions, Deletions and Substitutions,” *IEEE Transactions on Communications,* vol. 59, no. 1, pp. 2-6, Jan. 2011.

[19] S. Diggavi, M. Mitzenmacher, and H. D. Pfister, “Capacity Upper Bounds for the Deletion Channels,” in *Proc. IEEE Int. Symp. Information Theory (ISIT),* pp. 1716-1720, Nice, France, June 2007.

[20] Y.-C. Huang, U. Niesen, and P. Gupta, “Energy-Efficient Communication in the Presence of Synchronization Errors,” *IEEE Trans. on Inf. Theory*, submitted 2013. [Online]. Available: http://arxiv.org/pdf/1301.6589v1.pdf

[21] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* New York: Academic, second edition, 2011.

[22] A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press, UK, 2012.

[23] M. Khoshnevisan and J. N. Laneman, “Achievable Rates for Intermittent Communication,” in *Proc. IEEE Int. Symp. Information Theory (ISIT),* pp. 1346-1350, Cambridge, MA, USA, July 2012.

[24] M. Khoshnevisan and J. N. Laneman, “Intermittent Communication and Partial Divergence,” in *Proc. Allerton Conf. Communications, Control, and Computing,* Monticello, IL, Oct. 2012.

[25] J. N. Laneman and B. P. Dunn, “Communications Overhead as the Cost of Constraints,” in *Proc. IEEE Information Theory Workshop (ITW)*, Paraty, Brazil, Oct. 2011.

[26] R. E. Blahut, "Computation of Channel Capacity and Rate Distortion Functions," *IEEE Trans. on Inf. Theory,* vol. 18, no. 1, pp. 460-473, Jan. 1972.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, Second Edition, 2006.

[28] M. Khoshnevisan and J. N. Laneman, "Upper Bounds on the Capacity of Binary Intermittent Communication," in *Proc. Workshop on Information Theory and its Applications,* San Diego, CA, Feb. 2013.

[29] S. Verdu, "On Channel Capacity Per Unit Cost," *IEEE Trans. on Inf. Theory,* vol. 36, pp. 1019-1030, Sept. 1990.