Asymptotics of Input-Constrained Erasure Channel Capacity [*]

Yonglong Li
The University of Hong Kong
email: yonglong@hku.hk

Guangyue Han
The University of Hong Kong
email: ghan@hku.hk

May 3, 2016

Abstract

In this paper, we examine an input-constrained erasure channel and we characterize the asymptotics of its capacity when the erasure rate is low. More specifically, for a general memoryless erasure channel with its input supported on an irreducible finite-type constraint, we derive partial asymptotics of its capacity, using some series expansion type formulas of its mutual information rate; and for a binary erasure channel with its first-order Markovian input supported on the $(1, \infty)$-RLL constraint, based on the concavity of its mutual information rate with respect to some parameterization of the input, we numerically evaluate its first-order Markov capacity and further derive its full asymptotics. The asymptotics result obtained in this paper, when compared with the recently derived feedback capacity for a binary erasure channel with the same input constraint, enable us to draw the conclusion that feedback may increase the capacity of an input-constrained channel, even if the channel is memoryless.

Index Terms: erasure channel, input constraint, capacity, feedback.

1 Introduction

The primary concern of this paper is erasure channel, which is a common digital communication channel model that plays a fundamental role in coding and information theory. Throughout the paper, we assume that time is discrete and indexed by the integers. At time n, the erasure channel of interest can be described by the following equation:

$$Y_n = X_n E_n, \tag{1}$$

where the channel input $\{X_n\}$, supported on an irreducible finite-type constraint $\mathcal{S}$, is a stationary process taking values from the input alphabet $\mathcal{X} = \{1, 2, \cdots, K\}$, and the erasure process $\{E_n\}$, independent of $\{X_n\}$, is a binary stationary and ergodic process with *erasure*

[*] A preliminary version of this work has been presented in IEEE ISIT 2014.

rate $\varepsilon \triangleq P(E_1 = 0)$, and $\{Y_n\}$ is the channel output process over the output alphabet $\mathcal{Y} = \{0, 1, \cdots, K\}$. The word "erasure" as in the name of our channel naturally arises if a "0" is interpreted as an erasure at the receiving end of the channel; so, at time n, the channel output Y_n is nothing but the channel input X_n if $E_n = 1$, but an erasure if $E_n = 0$.

Let $\mathcal{X}^*$ denote the set of all the finite length words over $\mathcal{X}$. Let $\mathcal{F}$ be a finite subset of $\mathcal{X}^*$, and let $\mathcal{S}$ be the *finite-type constraint* with respect to $\mathcal{F}$, which is a subset of $\mathcal{X}^*$ consisting of all the finite length words over $\mathcal{X}$, each of which does not contain any element in $\mathcal{F}$ as a contiguous subsequence (or, roughly, elements in $\mathcal{F}$ are "forbidden" in $\mathcal{S}$). The most well known example is the (d, k)-run-length-limited (RLL) constraint over the alphabet $\{1, 2\}$, which forbids any sequence with fewer than d or more than k consecutive 1's in between two successive 2's; in particular, a prominent example is the $(1, \infty)$-RLL constraint, a widely used constraint in magnetic recording and data storage; see [30, 31]. For the (d, k)-RLL constraint with $k < \infty$, a forbidden set $\mathcal{F}$ is

$$\mathcal{F} = \{2\underbrace{1 \cdots 1}_{l}2 : 0 \leq l < d\} \cup \{\underbrace{0 \cdots 0}_{k+1}\}.$$

When $k = \infty$, one can choose $\mathcal{F}$ to be

$$\mathcal{F} = \{2\underbrace{1 \cdots 1}_{l}2 : 0 \leq l < d\};$$

in particular when $d = 1, k = \infty$, $\mathcal{F}$ can be chosen to be $\{22\}$. The *length* of $\mathcal{F}$ is defined to be that of the longest words in $\mathcal{F}$. Generally speaking, there may be many such $\mathcal{F}$'s with different lengths that give rise to the same constraint $\mathcal{S}$; the length of the shortest such $\mathcal{F}$'s minus 1 gives the *topological order* of $\mathcal{S}$. For example, the topological order of the $(1, \infty)$-RLL constraint, whose shortest $\mathcal{F}$ proves to be $\{22\}$, is 1. A finite-type constraint $\mathcal{S}$ is said to be *irreducible* if for any $u, v \in \mathcal{S}$, there is a $w \in \mathcal{S}$ such that $uwv \in \mathcal{S}$.

As mentioned before, the input process X of our channel (1) is assumed to be *supported* on an irreducible finite-type constraint $\mathcal{S}$, namely, $\mathcal{A}(X) \subseteq \mathcal{S}$, where

$$\mathcal{A}(X) \triangleq \{x_i^j \in \mathcal{X}^* : p_X(x_i^j) > 0\}.$$

The capacity of the channel (1), denoted by $C(\mathcal{S}, \varepsilon)$, can be computed by

$$C(\mathcal{S}, \varepsilon) = \sup_{\mathcal{A}(X) \subseteq \mathcal{S}} I(X; Y),$$

where the supremum is taken over all stationary processes X supported on $\mathcal{S}$. Here, we note that input-constraints [50] are widely used in various real-life applications such as magnetic and optical recording [31] and communications over band-limited channels with inter-symbol interference [12]. Particularly, we will pay special attention in this paper to a binary erasure channel with erasure rate ε (BEC(ε)) with the input supported on the $(1, \infty)$-RLL constraint, denoted by $\mathcal{S}_0$ throughout the paper.

When there is no constraint imposed on the input process X, that is, $\mathcal{S} = \mathcal{X}^*$, it is well known that $C(\mathcal{S}, \varepsilon) = (1 - \varepsilon) \log K$; see Theorem 5.1. When $\varepsilon = 0$, that is, when the channel is perfect with no erasures, $C(\mathcal{S}, \varepsilon)$ proves to be the *noiseless capacity* of the constraint $\mathcal{S}$, which can be achieved by a unique m-th order Markov chain $\hat{X}$ with $\mathcal{A}(\hat{X}) = \mathcal{S}$ [33].

On the other hand, other than these two above-mentioned "degenerated" cases, "explicit" analytic formulas of capacity for "non-degenerated" cases have remained evasive, and the problem of analytically characterizing the noisy constrained capacity is widely believed to be intractable.

The problem of numerically computing the capacity $C(\mathcal{S}, \varepsilon)$ seems to be as challenging: the computation of the capacity of a general channel with memory or input constraints is notoriously difficult and has been open for decades; and the fact that our erasure channel is only a special class of such ones does not appear to make the problem easier. Here, we note that for a discrete memoryless channel, Shannon gave a closed-form formula of the capacity in his celebrated paper [40], and Blahut [5] and Arimoto [1], independently proposed an algorithm which can efficiently compute the capacity and the capacity-achieving distribution simultaneously. However, unlike the discrete memoryless channels, the capacity of a channel with memory or input constraints in general admits no single-letter characterization and very little is known about the efficient computation of the channel capacity. To date, most known results in this regard have been in the forms of numerically computed bounds: for instance, numerically computed lower bounds by Arnold and Loeliger [2], A. Kavcic [25], Pfister, Soriaga and Siegel [35], Vontobel and Arnold [47].

One of the most effective strategies to compute the capacity of channels with memory or input constraints is the so-called *Markov approximation* scheme. The idea is that instead of maximizing the mutual information rate over all stationary processes, one can maximize the mutual information rate over all m-th order Markov processes to obtain the m-th order Markov capacity. Under suitable assumptions (see, e.g., [6]), when m tends to infinity, the corresponding sequence of Markov capacities will tend to the channel capacity. For our erasure channel, the *m-th order Markov capacity* is defined as

$$C^{(m)}(\mathcal{S}, \varepsilon) = \sup I(X; Y),$$

where the supremum is taken over all m-th order Markov chains supported on $\mathcal{S}$.

The main contributions of this work are the characterization of the asymptotics of the above-mentioned input-constrained erasure channel capacity. Of great relevance to this work are results by Han and Marcus [19], Jacquet and Szpankowski [24], which have characterized asymptotics of the capacity of the a binary symmetric channel with crossover probability ε (BSC(ε)) with the input supported on the $(1, \infty)$-RLL constraint. The approach in the above-mentioned work is to obtain the asymptotics of the mutual information rate first, and then apply some bounding argument to obtain that of the capacity. The approach in this work roughly follows the same strategy, however, as elaborated below, our approach differs from theirs to a great extent in terms of technical implementations.

Throughout the paper, we use the logarithm with base e in the proofs and we use the logarithms with base 2 in the numerical computations of the channel capacity. Below is a brief account of our results and methodology employed in this work.

The starting point of our approach is Lemma 2.1 in Section 2, a key lemma that expresses the conditional entropy $H(Y_0|Y_{-n}^{-1})$ in a form that is particularly effective for analyzing asymptotics of $C(\mathcal{S}, \varepsilon)$ when ε is close to 0. As elaborated in Theorem 2.2, Lemma 2.1 naturally gives a lower and upper bound on $C(\mathcal{S}, \varepsilon)$, where the lower bound gives a counterpart result of Wolf's conjecture for a BEC(ε). Moreover, when applied to the case when X is a Markov chain, Lemma 2.1 yields some explicit series expansion type formulas in Theorem 2.4 and Corollary 2.5, which aptly pave the way for characterizing the asymptotics of the input-constrained erasure channel capacity. Here we remark that the method in [19, 24] have been further developed for more general families of memory channels in [20, 21] via examining the contractiveness of an associated random dynamical system [4]. However, the methodology to derive asymptotics of the mutual information rate in this work capitalizes on certain characteristics that are in a sense unique to erasure channels.

In Section 3, we consider a memoryless erasure channel with the input supported on an irreducible finite-type constraint, and in Theorem 3.1, we derive partial asymptotics of its capacity $C(\mathcal{S}, \varepsilon)$ in the vicinity of $\varepsilon = 0$ where $C(\mathcal{S}, \varepsilon)$ is written as the sum of a constant term, a linear term in ε and an $O(\varepsilon^2)$-term. The lower bound part in the proof of this theorem follows from an easy application of Theorem 2.4, and the upper bound part hings on an adapted argument in [19].

In Section 4, we consider a BEC(ε) with the input being a first-order Markov process supported on the $(1, \infty)$-RLL constraint $\mathcal{S}_0$. Within this special setup, we show in Theorem 4.1 that the $I(X; Y)$ is strictly concave with respect to some parameterization of X. And in Section 4.2, we numerically evaluate $C^{(1)}(\mathcal{S}_0, \varepsilon)$ and the corresponding capacity-achieving distribution using the randomized algorithm proposed in [16] which proves to be convergent given the concavity of $I(X; Y)$. Moreover, the concavity of $I(X; Y)$ guarantees the uniqueness of the capacity achieving distribution, based on which we derive full asymptotics of the above input-constrained BEC(ε) around $\varepsilon = 0$ in Theorem 4.2, where $C^{(1)}(\mathcal{S}, \varepsilon)$ is expressed as an infinite sum of all $O(\varepsilon^k)$-terms.

In Section 5, we turn to the scenarios when there might be feedback in our erasure channel. We first prove in Theorem 5.1 that when there is no input constraint, the feedback does not increase the capacity of the erasure channel even with the presence of the channel memory. When the input constraint is not trivial, however, we show in Theorem 5.3 that feedback does increase the capacity using the example of a BEC(ε) with the $(1, \infty)$-RLL input constraint, and so feedback may increase the capacity of input-constrained erasure channels even if there is no channel memory. The results obtained in this section suggest the intricacy of the interplay between feedback, memory and input constraints.

2 A Key Lemma and Its Applications

In this section, we focus on the mutual information of the erasure channel (1) introduced in Section 1. The starting point of our approach is the following key lemma, which is particularly effective for analysis of input-constrained erasure channels.

Lemma 2.1. *For any $n \geq 1$, we have*

$$H(Y_0|Y_{-n}^{-1}) = H(E_0|E_{-n}^{-1}) + \sum_{D \subseteq [-n,-1]} H(X_0|X_D) P(E_0 = 1, E_D = 1, E_{D^c} = 0), \qquad (2)$$

where $[-n, -1] \triangleq \{-n, \cdots, -1\}$.

Proof. Note that

$$\begin{aligned} H(Y_0|Y_{-n}^{-1}) &= -\sum_{y_{-n}^0} p(y_{-n}^0) \log p(y_0|y_{-n}^{-1}) \\ &= T_1(n) + T_2(n), \end{aligned}$$

where

$$T_1(n) = -\sum_{y_{-n}^{-1}, y_0=0} p(y_{-n}^0) \log p(y_0|y_{-n}^{-1}) \qquad \text{and} \qquad T_2(n) = -\sum_{y_{-n}^{-1}, y_0\neq 0} p(y_{-n}^0) \log p(y_0|y_{-n}^{-1}).$$

From the independence of $\{X_n\}$ and $\{E_n\}$, it follows that

$$\begin{aligned} p(y_i^j) &= \sum_{x_i^j:\ x_k=y_k \text{ for } k\in\mathcal{I}(y_i^j)} p_X(x_i^j) P(E_{\mathcal{I}(y_i^j)} = 1, E_{\bar{\mathcal{I}}(y_i^j)} = 0) \\ &= p_X\left(y_{\mathcal{I}(y_i^j)}\right) P(E_{\mathcal{I}(y_i^j)} = 1, E_{\bar{\mathcal{I}}(y_i^j)} = 0). \end{aligned} \tag{3}$$

Here and throughout the paper, let $\mathcal{Y}^*$ be the set of all finite length words over $\mathcal{Y}$ and we define, for any $y_i^j \in \mathcal{Y}^*$,

$$\mathcal{I}(y_i^j) = \{k : i \leq k \leq j, y_k \neq 0\}, \quad \bar{\mathcal{I}}(y_i^j) = \{k : i \leq k \leq j, y_k = 0\}$$

and

$$y_{\mathcal{I}(y_i^j)} = \{y_k : k \in \mathcal{I}(y_i^j)\}.$$

For $y_0 \neq 0$,

$$\begin{aligned} p(y_0|y_{-n}^{-1}) &= \frac{p(y_{-n}^0)}{p(y_{-n}^{-1})} \\ &= \frac{p_X\left(y_{\mathcal{I}(y_{-n}^0)}\right) P(E_{\mathcal{I}(y_{-n}^0)} = 1, E_{\bar{\mathcal{I}}(y_{-n}^0)} = 0)}{p_X\left(y_{\mathcal{I}(y_{-n}^{-1})}\right) P(E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0)} \\ &\overset{(a)}{=} p_X\left(y_0|y_{\mathcal{I}(y_{-n}^{-1})}\right) P(E_0 = 1|E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0), \end{aligned}$$

where (a) follows from the fact that $\bar{\mathcal{I}}(y_{-n}^0) = \bar{\mathcal{I}}(y_{-n}^{-1})$. Similarly, for $y_0 = 0$,

$$\begin{aligned} p(y_0|y_{-n}^{-1}) &= \frac{p_X\left(y_{\mathcal{I}(y_{-n}^0)}\right) P(E_{\mathcal{I}(y_{-n}^0)} = 1, E_{\bar{\mathcal{I}}(y_{-n}^0)} = 0)}{p_X\left(y_{\mathcal{I}(y_{-n}^{-1})}\right) P(E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0)} \\ &\overset{(a)}{=} P(E_0 = 1|E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0), \end{aligned}$$

where (a) follows from the fact that $\mathcal{I}(y_{-n}^0) = \mathcal{I}(y_{-n}^{-1})$. Therefore,

$$\begin{aligned} T_1(n) &= -\sum_{y_{-n}^{-1}, y_0=0} p(y_{-n}^0) \log p(y_0|y_{-n}^{-1}) \\ &= -\sum_{y_{-n}^{-1}, y_0=0} p(y_{-n}^0) \log P(E_0 = 0|E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0) \\ &= -\sum_{D\subseteq[-n,-1]} \sum_{y_{-n}^0 : \mathcal{I}(y_{-n}^{-1})=D, y_0=0} p(y_{-n}^0) \log P(E_0 = 0|E_D) = 1, E_{D^c} = 0) \\ &\overset{(a)}{=} -\sum_{D\subseteq[-n,-1]} P(E_0 = 0, E_D = 1, E_{D^c} = 0) \log P(E_0 = 0|E_D = 1, E_{D^c} = 0), \end{aligned} \tag{4}$$

where (a) follows from the fact that for any given $D \subseteq [-n, -1]$,

$$\sum_{y_{-n}^{0}: \mathcal{I}(y_{-n}^{-1}) = D, y_0 = 0} p(y_{-n}^{0}) = P(E_0 = 0, E_D = 1, E_{D^c} = 0).$$

Also, we have

$$\begin{aligned}
T_2(n) &= - \sum_{y_{-n}^{-1}, y_0 \neq 0} p(y_{-n}^{0}) \log p(y_0 | y_{-n}^{-1}) \\
&= - \sum_{y_{-n}^{-1}, y_0 \neq 0} p(y_{-n}^{0}) \log p_X \left(y_0 | y_{\mathcal{I}(y_{-n}^{-1})} \right) P(E_0 = 1 | E_{\mathcal{I}(y_{-l}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-l}^{-1})} = 0) \\
&= T_3(n) - \sum_{y_{-n}^{-1}, y_0 \neq 0} p(y_{-n}^{0}) \log P(E_0 = 1 | E_{\mathcal{I}(y_{-n}^{-1})} = 1, E_{\bar{\mathcal{I}}(y_{-n}^{-1})} = 0) \\
&\overset{(a)}{=} T_3(n) - \sum_{D \subseteq [-n, -1]} P(E_0 = 1, E_D = 1, E_{D^c} = 0) \log P(E_0 = 1 | E_D = 1, E_{D^c} = 0), \quad (5)
\end{aligned}$$

where (a) follows from a similar argument as in the proof of (4) and

$$T_3(n) = - \sum_{y_{-n}^{-1}, y_0 \neq 0} p(y_{-n}^{0}) \log p_X(y_0 | y_{\mathcal{I}(y_{-n}^{-1})}).$$

From (3), it then follows that

$$\begin{aligned}
T_3(n) &= - \sum_{y_{-n}^{-1}, y_0 \neq 0} p(y_{-n}^{0}) \log p_X(y_0 | y_{\mathcal{I}(y_{-n}^{-1})}) \\
&= - \sum_{D \subseteq [-n, -1]} \sum_{y_{-n}^{0}: \mathcal{I}(y_{-n}^{0}) = D \cup \{0\}} p_X(y_D, y_0) P(E_0 = 1, E_D = 1, E_{D^c} = 0) \log p_X(y_0 | y_D) \\
&= \sum_{D \subseteq [-n, -1]} H(X_0 | X_D) P(E_0 = 1, E_D = 1, E_{D^c} = 0). \quad (6)
\end{aligned}$$

The desired formula for $H(Y_0 | Y_{-n}^{-1})$ then follows from (4), (5) and (6). □

One of the immediate applications of Lemma 2.1 is the following lower and upper bounds on $C(\mathcal{S}, \varepsilon)$.

Theorem 2.2.

$$(1 - \varepsilon) C(\mathcal{S}, 0) \leq C(\mathcal{S}, \varepsilon) \leq (1 - \varepsilon) \log K.$$

Proof. For the upper bound, it follows from Lemma 2.1 that

$$\begin{aligned}
I(X;Y) &= \lim_{n\to\infty} (H(Y_0|Y_{-n}^{-1}) - H(Y_0|Y_{-n}^{-1}, X_{-n}^{0})) \\
&= \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} H(X_0|X_D)P(E_0 = 1, E_D = 1, E_{D^c} = 0) \\
&\overset{(a)}{\leq} \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} H(X_0)P(E_0 = 1, E_D = 1, E_{D^c} = 0) \\
&\leq \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} P(E_0 = 1, E_D = 1, E_{D^c} = 0) \log K \\
&= P(E_0 = 1) \log K \\
&= (1-\varepsilon) \log K,
\end{aligned}$$

where we have used the fact that conditioning reduces entropy for (a).

Assume $\mathcal{S}$ is of topological order m, and let $\hat{X}$ be the m-order Markov chain that achieves the noiseless capacity $C(\mathcal{S}, 0)$ of the constraint $\mathcal{S}$. Again, it follows from Lemma 2.1 that

$$\begin{aligned}
I(\hat{X};Y) &= \lim_{n\to\infty} (H(Y_0|Y_{-n}^{-1}) - H(Y_0|Y_{-n}^{-1}, \hat{X}_{-n}^{0})) \\
&= \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} H(\hat{X}_0|\hat{X}_D)P(E_0 = 1, E_D = 1, E_{D^c} = 0) \\
&\geq \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} H(\hat{X}_0|\hat{X}_{-m}^{-1}, \hat{X}_D)P(E_0 = 1, E_D = 1, E_{D^c} = 0) \\
&\overset{(a)}{=} \lim_{n\to\infty} \sum_{D\subseteq[-n,-1]} H(\hat{X}_0|\hat{X}_{-m}^{-1})P(E_0 = 1, E_D = 1, E_{D^c} = 0) \\
&= P(E_0 = 1)H(\hat{X}_0|\hat{X}_{-m}^{-1}) \\
&= (1-\varepsilon)C(\mathcal{S}, 0),
\end{aligned}$$

where we have used the fact that $\{\hat{X}_n\}$ is an m-th order Markov chain for (a). □

Remark 2.3. The upper bound part of Theorem 2.2 also follows from the well-known fact that (see Theorem 5.1)

$$C(\mathcal{X}^*, \varepsilon) = (1-\varepsilon) \log K$$

and for any $\mathcal{S}$,

$$C(\mathcal{S}, \varepsilon) \leq C(\mathcal{X}^*, \varepsilon),$$

which is obviously true.

Let $C'(\mathcal{S}, \varepsilon)$ denote the capacity of a BSC(ε) with the (d, k)-RLL constraint. In [49] Wolf posed the following conjecture on $C'(\mathcal{S}, \varepsilon)$:

$$C'(\mathcal{S}, \varepsilon) \geq C'(\mathcal{S}, 0)(1 - H(\varepsilon)),$$

where $H(\varepsilon) \triangleq -\varepsilon \log \varepsilon - (1-\varepsilon) \log(1-\varepsilon)$. A weaker form of this bound has been established in [36] by counting the possible subcodes satisfying the (d, k)-RLL constraint in some linear coding scheme, but the conjecture for the general case still remains open.

It is well known that $1 - H(\varepsilon)$ is the capacity of a BSC(ε) without any input constraint, and $1 - \varepsilon$ is the capacity of a BEC(ε) without any input constraint. So, for an input-constrained BEC(ε), the lower bound part of Theorem 2.2 gives a counterpart result of Wolf's conjecture.

When applied to the channel with a Markovian input, Lemma 2.1 gives a relatively explicit series expansion type formula for the mutual information rate of (1).

Theorem 2.4. *Assume $\{X_n\}$ is an m-th order input Markov chain. Then,*

$$I(X;Y) \;=\; \sum_{k=0}^{\infty} \sum_{t=0}^{b(k-1,m)} \sum_{\{i_1^t\} \in B_2(k-1,t)} H(X_0 | X_{i_1^t}, X_{-k-m}^{-k-1}) P(E_{A(k,i_1^t)} = 1, E_{\bar{A}(k,i_1^t)} = 0), \quad (7)$$

where $A(k, i_1^t) = \{-k-m, \cdots, -k-1, i_1^t, 0\}$ and $\bar{A}(k, i_1^t) = \{-k-m, \cdots, 0\} - A(k, i_1^t)$ and $B_2(n, u) = \{\{i_1, \cdots, i_u\} \subseteq [-n, -1] :$ for all $j = 1, \cdots, u, \{i_j, i_j + 1, \cdots, i_j + m\} \nsubseteq \{i_1, \cdots, i_u\}\}$ and $b(k-1, m) = (m-1)\left\lfloor \frac{k-1}{m} \right\rfloor + R(k-1)$, here $R(k-1)$ denotes the remainder of $k-1$ divided by m.

Proof. Note that

$$\begin{aligned} H(Y_0 | X_{-n}^0, Y_{-n}^{-1}) &= H(X_0 E_0 | X_{-n}^0, E_{-n}^{-1}, Y_{-n}^{-1}) \\ &\overset{(a)}{=} H(E_0 | E_{-n}^{-1}), \end{aligned}$$

where (a) follows from the independence of $\{X_n\}$ and $\{E_n\}$. From Lemma 2.1, it then follows that

$$\begin{aligned} I(X;Y) &= \lim_{n\to\infty} (H(Y_0|Y_{-n}^{-1}) - H(Y_0|X_{-n}^0, Y_{-n}^{-1})) \\ &= \lim_{n\to\infty} \sum_{D \subseteq [-n,-1]} H(X_0|X_D) P(E_0 = 1, E_D = 1, E_{D^c} = 0). \end{aligned} \quad (8)$$

Now, letting

$$B(n, u) = \{D \subseteq [-n, -1] : |D| = u\} \text{ and } B_1(n, u) = B(n, u) - B_2(n, u),$$

we deduce that, for $\{i_1^t\} \in B_2(k-1, t)$

$$\sum_{D \subseteq [-n,-k-m-1]} P(E_{A(k,i_1^t)} = 1, E_D = 1, E_{\bar{A}(k,i_1^t)} = 0, E_{[-n,-k-m-1]-D} = 0) = P(E_{A(k,i_1^t)} = 1, E_{\bar{A}(k,i_1^t)} = 0).$$

and

$$\begin{aligned}
&\sum_{k=m}^{n} \sum_{\{i_1,\ldots,i_k\}\in B_1(n,k)} H(X_0|X_{i_1^k})P(E_0=1,E_{i_1^k}=1,E_{\bar{i}_1^k}=0) \\
&\quad= \sum_{k=0}^{n-m+1} \sum_{t=0}^{b(k-1,m)} \sum_{\{i_1^t\}\in B(k-1,t)} \sum_{D\subseteq[-n,-k-m-1]} \big\{ H(X_0|X_{A(k,i_1^t)},X_D) \\
&\qquad \times P(E_{A(k,i_1^t)}=1,E_D=1,E_{\bar{A}(k,i_1^t)}=0,E_{[-n,-k-m-1]-D}=0)\big\} \\
&\quad\overset{(a)}{=} \sum_{k=0}^{n-m+1} \sum_{t=0}^{b(k-1,m)} \sum_{\{i_1^t\}\in B(k-1,t)} \sum_{D\subseteq[-n,-k-m-1]} \big\{ H(X_0|X_{A(k,i_1^t)}) \\
&\qquad \times P(E_{A(k,i_1^t)}=1,E_D=1,E_{\bar{A}(k,i_1^t)}=0,E_{[-n,-k-m-1]-D}=0)\big\} \\
&\quad= \sum_{k=0}^{n-m+1} \sum_{t=0}^{b(k-1,m)} \sum_{\{i_1^t\}\in B(k-1,t)} H(X_0|X_{A(k,i_1^t)})P(E_{A(k,i_1^t)}=1,E_{\bar{A}(k,i_1^t)}=0),
\end{aligned}$$

where (a) follows from the fact that $\{X_n\}$ is an m-th order Markov chain. Then it follows that

$$\begin{aligned}
&\sum_{D\subseteq[-n,-1]} H(X_0|X_D)P(E_0=1,E_D=1,E_{D^c}=0) \\
&\quad= \sum_{k=0}^{n} \sum_{\{i_1,\ldots,i_k\}\in B(n,k)} H(X_0|X_{i_1^k})P(E_0=1,E_{i_1^k}=1,E_{\bar{i}_1^k}=0) \\
&\quad= \left(\sum_{k=m}^{n} \sum_{\{i_1,\ldots,i_k\}\in B_1(n,k)} + \sum_{k=0}^{b(n,m)} \sum_{\{i_1,\ldots,i_k\}\in B_2(n,k)} \right) H(X_0|X_{i_1^k})P(E_0=1,E_{i_1^k}=1,E_{\bar{i}_1^k}=0) \\
&\quad= \sum_{k=0}^{n-m+1} \sum_{t=0}^{b(k-1,m)} \sum_{\{i_1^t\}\in B(k-1,t)} H(X_0|X_{A(k,i_1^t)})P(E_{A(k,i_1^t)}=1,E_{\bar{A}(k,i_1^t)}=0)+T(n), \qquad (9)
\end{aligned}$$

where

$$T(n) \;=\; \sum_{k=0}^{b(n,m)} \sum_{\{i_1^k\}\in B_2(n,k)} H(X_0|X_{i_1^k})P(E_0=1,E_{i_1^k}=1,E_{\bar{i}_1^k}=0).$$

It follows from $H(X_0|X_{i_1^k}) \leq \log K$ that

$$T(n) \leq \sum_{k=0}^{b(n,m)} \sum_{\{i_1^k\}\in B_2(n,k)} P(E_0=1,E_{i_1^k}=1,E_{\bar{i}_1^k}=0)\log K \leq P(F_n)\log K,$$

where F_n is the event that "there is no m consecutive 1's in E_{-n}^{-1}". Now, let $W_i = (E_i,\cdots,E_{i-m+1})$ for $i \leq -1$. Then it follows from the assumption that W_i is also a stationary and ergodic process with $P(W_i=(1,\cdots,1))>0$. Using Poincare's recurrence theorem [10], we have that $P(W_i = (1, \cdots, 1)\ i.o.) = 1$, which implies that $P(F) = 0$, where F denotes the event that "there is no m consecutive 1's in $E_{-\infty}^{-1}$". This, together with the fact that $\lim_{n\to\infty} P(F_n) = P(F)$, implies that $\lim_{n\to\infty} T(n) = 0$, and therefore the proof of the theorem is complete. □

The following corollary can be readily deduced from Theorem 2.4.

Corollary 2.5. *Assume that $\{E_n\}$ is i.i.d. and $\{X_n\}$ is an m-th order Markov chain. Then*

$$I(X;Y) \;\;=\;\; (1-\varepsilon)^{m+1} \sum_{k=0}^{\infty} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^t \varepsilon^{k-t}, \tag{10}$$

where

$$a(k,t) = \sum_{\{i_1 \dots i_t\} \in B_2(k-1,t)} H(X_0 | X_{i_1^t}, X_{-k-m}^{-k-1}).$$

In particular, if $\{X_n\}$ is a first-order Markov chain,

$$I(X;Y) = (1-\varepsilon)^2 \sum_{k=0}^{\infty} H(X_0|X_{-k-1})\varepsilon^k. \tag{11}$$

Remark 2.6. A series expansion type formula for $H(X|Y)$ different from (11) is given in Theorem 12 of [46] for a discrete memoryless erasure channel with a first-order input Markov chain. It can be verified that these two formulas are "equivalent" in the sense that either one can be deduced from the other one via simple derivations. The form that our formula takes however makes it particularly effective for the capacity analysis of an input-constrained erasure channel.

3 Input-Constrained Memoryless Erasure Channel

In this section, we will focus on the case when $\{E_n\}$ is i.i.d. and $\mathcal{S}$ is an irreducible finite-type constraint of topological order m. With Lemma 2.1 and Corollary 2.5 established, we are ready to characterize the asymptotics of the capacity of this type of erasure channels.

As mentioned in Section 1, when $\varepsilon = 0$, it is well known [33] that there exists an mth-order Markov chain $\hat{X}$ with $\mathcal{A}(\hat{X}) = \mathcal{S}$ such that

$$H(\hat{X}) = H(\hat{X}_0|\hat{X}_{-m}^{-1}) = \max_{\mathcal{A}(X)\subseteq\mathcal{S}} H(X) = C(\mathcal{S}, 0), \tag{12}$$

where the maximization is over all stationary processes supported on $\mathcal{S}$. The following theorem characterizes the asymptotics of $C(\mathcal{S}, \varepsilon)$ near $\varepsilon = 0$.

Theorem 3.1. *Assume that $\{E_n\}$ is i.i.d. Then,*

$$C(\mathcal{S},\varepsilon) = C(\mathcal{S},0) - \left\{ (m+1)H(\hat{X}_0|\hat{X}_{-m}^{-1}) - \sum_{i=1}^{m} H(\hat{X}_0|\hat{X}_{-i+1}^{-1}, \hat{X}_{-i-m}^{-i-1}) \right\} \varepsilon + O(\varepsilon^2). \tag{13}$$

Moreover, for any $n \geq m$, $C^{(n)}(\mathcal{S}, \varepsilon)$ is of the same asymptotic form as in (3.1), namely,

$$C^{(n)}(\mathcal{S}, \varepsilon) = C(\mathcal{S}, 0) - \left\{ (m+1) H(\hat{X}_0 | \hat{X}_{-m}^{-1}) - \sum_{i=1}^{m} H(\hat{X}_0 | \hat{X}_{-i+1}^{-1}, \hat{X}_{-i-m}^{-i-1}) \right\} \varepsilon + O(\varepsilon^2). \quad (14)$$

Proof. To establish (13), we prove that $C(\mathcal{S}, \varepsilon)$ is lower and upper bounded by the same asymptotic form as in (13).

For the lower bound part, we consider the channel (1) with $\hat{X}$ as its input. Note that

$$P(\hat{F}_0) = (1-\varepsilon)^m \text{ and } P(\hat{F}_k) = \varepsilon(1-\varepsilon)^m \quad \text{for } 1 \leq k \leq m,$$

and furthermore, for $k \geq m+1$

$$P(\hat{F}_k) = \sum_{t=0}^{b(k-1,m)} |B_2(k-1, t)| (1-\varepsilon)^{t+m} \varepsilon^{k-t},$$

where we have defined

$$\hat{F}_k = \{E_{-k-1}^{-k-m} = 1, E_{-k} = 0, E_{-k+1}^{-1} \text{ contains no } m \text{ consecutive 1's}\}.$$

It then follows that

$$\begin{aligned}
&(1-\varepsilon)^{m+1} \sum_{k=m+1}^{\infty} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^t \varepsilon^{k-t} \\
&\quad \leq \quad (1-\varepsilon) \sum_{k=m+1}^{\infty} P(\hat{F}_k) \log K \\
&\quad \overset{(a)}{=} \quad (1-\varepsilon)(1 - \sum_{k=0}^{m} P(\hat{F}_k) \log K \\
&\quad = \quad (m\varepsilon^2 (1-\varepsilon)^m + \sum_{u=2}^{m} \binom{m}{u} (1-\varepsilon)^{m-u} \varepsilon^u) \log K \\
&\quad = \quad O(\varepsilon^2), \qquad (15)
\end{aligned}$$

where (a) follows from $P(\cup_{k \geq 0} \hat{F}_k) = 1$ and the constant in $O(\varepsilon^2)$ depends only on m and K. Then, from Corollary 2.5, it follows that

$$\begin{aligned}
C(\mathcal{S}, \varepsilon) \geq I(\hat{X}; Y) &= (1-\varepsilon)^{m+1} \sum_{k=0}^{\infty} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^t \varepsilon^{k-t} \\
&= (1-\varepsilon)^{m+1} \sum_{k=0}^{m} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^t \varepsilon^{k-t} + (1-\varepsilon)^{m+1} \sum_{k=m+1}^{\infty} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^t \varepsilon^{k-t} \\
&\overset{(b)}{=} H(\hat{X}_0 | \hat{X}_{-m}^{-1}) + \left\{ (m+1) H(\hat{X}_0 | \hat{X}_{-m}^{-1}) - \sum_{i=1}^{m} H(\hat{X}_0 | \hat{X}_{-i+1}^{-1}, \hat{X}_{-i-m}^{-i-1}) \right\} \varepsilon + O(\varepsilon^2),
\end{aligned}$$

where (b) follows from (15) and

$$\sum_{k=0}^{m} \sum_{t=0}^{b(k-1,m)} a(k,t)(1-\varepsilon)^{t+m+1}\varepsilon^{k-t} = H(\hat{X}_0|\hat{X}_{-m}^{-1}) + \left\{ (m+1)H(\hat{X}_0|\hat{X}_{-m}^{-1}) - \sum_{i=1}^{m} H(\hat{X}_0|\hat{X}_{-i+1}^{-1}, \hat{X}_{-i-m}^{-i-1}) \right\} \varepsilon + O(\varepsilon^2).$$

This, together with (12), establishes that $C(\mathcal{S}, \varepsilon)$ is lower bounded by the asymptotic form in (13).

For the upper bound part, we will adapt the argument in [19]. Let

$$S_n = \left\{ \mathbf{p}_n = (p(\hat{x}_{-n}^0) : \hat{x}_{-n}^0 \in \mathcal{A}(\hat{X}_{-n}^0)) : p(\hat{x}_{-n}^0) > 0, \sum_{\hat{x}_0^n \in \mathcal{A}(\hat{X}_{-n}^0)} p(\hat{x}_{-n}^0) = 1 \right\}$$

and

$$S_{n,\delta} = \{\mathbf{p}_n \in S_n : p(\hat{x}_{-n}^0) > \delta \text{ for any } \hat{x}_{-n}^0 \in \mathcal{A}(\hat{X}_{-n}^0)\},$$

where $\mathcal{A}(\hat{X}_{-n}^0) = \{\hat{x}_{-n}^0 : p(\hat{x}_n^0) > 0\}$. In this proof, we define

$$C_n(\varepsilon, \mathcal{S}) = \sup_{\mathbf{p}_n \in S_n} H(Y_0|Y_{-n}^{-1}) - H(\varepsilon).$$

It then follows from Lemma 2.1 that

$$C_n(\mathcal{S}, \varepsilon) = \sup_{\mathbf{p}_n \in S_n} f(\mathbf{p}_n, \varepsilon),$$

where

$$f(\mathbf{p}_n, \varepsilon) \triangleq \sum_{k=1}^{n} \sum_{D \subseteq [-n,-1], |D|=k} H(X_0|X_D)(1-\varepsilon)^{k+1}\varepsilon^{n-k}.$$

Let $\overline{\mathbf{p}}_n(\varepsilon)$ maximize $f(\mathbf{p}_n, \varepsilon)$. As $f(\mathbf{p}_n, \varepsilon)$ is continuous in $(\mathbf{p}_n, \varepsilon)$ and is maximized at $\hat{\mathbf{p}}_n$ when $\varepsilon = 0$, there exists some $\varepsilon_0 > 0$ (depends on n) and $\delta > 0$ such that for all $\varepsilon < \varepsilon_0$, $\overline{\mathbf{p}}_n(\varepsilon) \in S_{n,\delta}$. Then for $\varepsilon \leq \varepsilon_0$, there exists some constant M (depends on n) such that

$$C_n(\mathcal{S}, \varepsilon) \leq \sup_{\mathbf{p}_n \in S_{n,\delta}} \left\{ H(X_0|X_{-n}^{-1}) + \left((n+1)H(X_0|X_{-n}^{-1}) - \sum_{k=1}^{n} H(X_0|X_{-k+1}^{-1}, X_{-n}^{-k-1}) \right) \varepsilon \right\} + M\varepsilon^2.$$

From now on, we write

$$g_1(\mathbf{p}_n) = H(X_0|X_{-n}^{-1}), \quad g_2(\mathbf{p}_n) = \left((n+1)H(X_0|X_{-n}^{-1}) - \sum_{k=1}^{n} H(X_0|X_{-k+1}^{-1}, X_{-n}^{-k-1}) \right) \varepsilon.$$

Letting $\mathbf{H} = \mathbf{H}(\mathbf{p}_n)$ be the Hessian of $g_1(\mathbf{p}_n)$, we now expand $g_1(\mathbf{p}_n)$ and $g_2(\mathbf{p}_n)$ around $\mathbf{p}_n = \hat{\mathbf{p}}_n$ to obtain

$$g_1(\mathbf{p}_n) = g_1(\hat{\mathbf{p}}_n) + \frac{1}{2}\mathbf{q}_n^T \mathbf{H} \mathbf{q}_n + O(|\mathbf{q}_n|^2)$$

and

$$g_2(\mathbf{p}_n) = g_2(\hat{\mathbf{p}}_n) + \sum_{\hat{x}_{-n}^0 \in \mathcal{A}(\hat{X}_{-n}^0)} \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} q(\hat{x}_{-n}^0) + O(|\mathbf{q}_n|).$$

where $\mathbf{q}_n \triangleq \mathbf{p}_n - \hat{\mathbf{p}}_n$ contains all $q(\hat{x}_{-n}^0)$ as its coordinates. Since $\mathbf{H}$ is negative definite (see Lemma 3.1 [19]), we deduce that, for $|\mathbf{q}_n|$ sufficiently small,

$$g_1(\mathbf{p}_n) + g_2(\mathbf{p}_n)\varepsilon \leq g_1(\hat{\mathbf{p}}_n) + g_2(\hat{\mathbf{p}}_n)\varepsilon + \frac{1}{4}\mathbf{q}_n^T \mathbf{H} \mathbf{q}_n + 2 \sum_{\hat{x}_{-n}^0 \in \mathcal{A}(\hat{X}_{-n}^0)} \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} q(\hat{x}_{-n}^0) \right| \varepsilon.$$

Without loss of generality, we henceforth assume $\mathbf{H}$ is a diagonal matrix with all diagonal entries denoted $k(\hat{x}_{-n}^0) < 0$ (since otherwise we can diagonalize $\mathbf{H}$). Now, let

$$\mathcal{A}_1(\hat{X}_{-n}^0) = \left\{ \hat{x}_{-n}^0 : \frac{1}{4} k(\hat{x}_{-n}^0) \left| q(\hat{x}_{-n}^0) \right|^2 + 2 \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} \right| \left| q(\hat{x}_{-n}^0) \right| \varepsilon < 0 \right\}$$

and let $\mathcal{A}_2(\hat{X}_{-n}^0)$ denote the complement of $\mathcal{A}_1(\hat{X}_{-n}^0)$ in $\mathcal{A}(\hat{X}_{-n}^0)$:

$$\mathcal{A}_2(\hat{X}_{-n}^0) = \mathcal{A}(\hat{X}_{-n}^0) - \mathcal{A}_1(\hat{X}_{-n}^0).$$

Then, we have

$$\begin{aligned} f(\mathbf{p}_n, \varepsilon) &\leq g_1(\hat{\mathbf{p}}_n) + g_2(\hat{\mathbf{p}}_n)\varepsilon + \frac{1}{4}\mathbf{q}_n^T \mathbf{H} \mathbf{q}_n + 2 \sum_{\hat{x}_{-n}^0 \in \mathcal{A}(\hat{X}_{-n}^0)} \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0))} \right| \left| q(\hat{x}_{-n}^0) \right| \varepsilon + M\varepsilon^2 \\ &\leq g_1(\hat{\mathbf{p}}_n) + g_2(\hat{\mathbf{p}}_n)\varepsilon + \sum_{\hat{x}_{-n}^0 \in \mathcal{A}_2(\hat{X}_{-n}^0)} \left(\frac{1}{4} q(\hat{x}_{-n}^0)^2 k(\hat{x}_{-n}^0) + 2 \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} \right| \left| q(\hat{x}_{-n}^0) \right| \varepsilon \right) + M\varepsilon^2 \\ &\overset{(a)}{\leq} g_1(\hat{\mathbf{p}}_n) + g_2(\hat{\mathbf{p}}_n)\varepsilon + \sum_{\hat{x}_{-n}^0 \in \mathcal{A}_2(\hat{X}_{-n}^0)} \frac{4 \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} \right|^2 \varepsilon^2}{-k(\hat{x}_{-n}^0)} + M\varepsilon^2, \end{aligned}$$

where (a) follows from the easily verifiable fact that for any $\hat{x}_{-n}^0 \in \mathcal{A}_2(\hat{X}_{-n}^0)$,

$$\frac{1}{4} q(\hat{x}_{-n}^0)^2 k(\hat{x}_{-n}^0) + 2 \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} \right| \left| q(\hat{x}_{-n}^0) \right| \varepsilon \leq \frac{4 \left| \frac{\partial g_2(\hat{\mathbf{p}}_n)}{\partial p(\hat{x}_{-n}^0)} \right|^2 \varepsilon^2}{-k(\hat{x}_{-n}^0)}.$$

Since $\hat{X}_{-n}^0$ is an m-th order Markov chain, we deduce that

$$\begin{aligned} g_1(\hat{\mathbf{p}}_n) + g_2(\hat{\mathbf{p}}_n)\varepsilon &= H(\hat{X}_0 | \hat{X}_{-n}^{-1}) + \left((n+1) H(\hat{X}_0 | \hat{X}_{-n}^{-1}) - \sum_{k=1}^{n} H(\hat{X}_0 | \hat{X}_{-k+1}^{-1}, \hat{X}_{-n}^{-k-1}) \right) \varepsilon \\ &= H(\hat{X}_0 | \hat{X}_{-m}^{-1}) + \left((m+1) H(\hat{X}_0 | \hat{X}_{-m}^{-1}) - \sum_{k=1}^{m} H(\hat{X}_0 | \hat{X}_{-k+1}^{-1}, \hat{X}_{-m}^{-k-1}) \right) \varepsilon, \end{aligned}$$

We now ready to deduce that, for some positive constant M_1,

$$f(\mathbf{p}_n, \varepsilon) \leq H(\hat{X}_0 | \hat{X}_{-m}^{-1}) + \left((m+1) H(\hat{X}_0 | \hat{X}_{-m}^{-1}) - \sum_{k=1}^{m} H(\hat{X}_0 | \hat{X}_{-k+1}^{-1}, \hat{X}_{-m}^{-k-1}) \right) \varepsilon + M_1 \varepsilon^2,$$

which further implies that

$$C(\varepsilon, \mathcal{S}) \le C_n(\varepsilon, \mathcal{S}) \le H(\hat{X}_0|\hat{X}_{-m}^{-1})+\left((m+1)H(\hat{X}_0|\hat{X}_{-m}^{-1}) - \sum_{k=1}^{m} H(\hat{X}_0|\hat{X}_{-k+1}^{-1}, \hat{X}_{-m}^{-k-1})\right)\varepsilon+M_1\varepsilon^2.$$

The proof of (13) is then complete.

With $C(\mathcal{S}, \varepsilon)$ replaced with $C^{(m)}(\mathcal{S}, \varepsilon)$, the proof of (13) also establishes (14). □

Remark 3.2. In a fairly general setting (where input constraints are not considered), a similar asymptotic formula with a constant term, a term linear in ε and a residual $o(\varepsilon)$-term has been derived in Theorem 23 of [46].

As an immediate corollary of Theorem 3.1, the following result gives asymptotics of the capacity of a BEC(ε) with the input supported on the $(1, \infty)$-RLL constraint $\mathcal{S}_0$.

Corollary 3.3. *Assume $K = 2$ and $\{E_n\}$ is i.i.d. Then, we have*

$$C(\mathcal{S}_0, \varepsilon) = \log \lambda - \frac{2 \log 2}{1+\lambda^2}\varepsilon + O(\varepsilon^2),$$

and for any $n \ge 1$, $C^{(n)}(\mathcal{S}_0, \varepsilon)$ is of the same asymptotic form, namely,

$$C^{(n)}(\mathcal{S}_0, \varepsilon) = \log \lambda - \frac{2 \log 2}{1+\lambda^2}\varepsilon + O(\varepsilon^2). \tag{16}$$

Proof. Let $\lambda = (1+\sqrt{5})/2$. It is well known [30] that the noiseless capacity

$$C(\mathcal{S}_0, 0) = \log \lambda$$

and the first-order Markov chain $\{\hat{X}_n\}$ with the following transition probability matrix

$$\Pi = \begin{bmatrix} 1-1/\lambda^2 & 1/\lambda^2 \\ 1 & 0 \end{bmatrix}, \tag{17}$$

achieves the noiseless capacity, that is, $H(\hat{X}) = C(\mathcal{S}_0, 0) = \log \lambda$. Furthermore, via straightforward computations, we deduce that

$$H(\hat{X}_0|\hat{X}_{-2}) = \frac{\lambda^2}{1+\lambda^2}H(2/\lambda^2) + \frac{1}{1+\lambda^2}H(1/\lambda^2) = 2\log\lambda - \frac{2\log 2}{1+\lambda^2},$$

which, together with the fact that

$$H(\hat{X}) = H(\hat{X}_0|\hat{X}_{-1}) = \log \lambda,$$

implies

$$C(\mathcal{S}_0, \varepsilon) = \log \lambda - \frac{2 \log 2}{1+\lambda^2}\varepsilon + O(\varepsilon^2),$$

as desired.

And the asymptotic form of $C^{(n)}(\mathcal{S}_0, \varepsilon)$ follows from similar computations. □

Remark 3.4. The asymptotic form in (16) only gives partial asymptotics of $C^{(n)}(\mathcal{S}_0, \varepsilon)$ for any $n \ge 1$. For the case $n = 1$, we will derive later on the full asymptotics of $C^{(1)}(\mathcal{S}_0, \varepsilon)$; see (24) in Section 4.

4 Input-Constrained Binary Erasure Channel

In this section, we will focus on a BEC(ε) with the input being a first-order Markov process supported on the $(1,\infty)$-RLL constraint $\mathcal{S}_0$. To be more precise, we assume that $K=2$, $\{E_n\}$ is i.i.d. and $\{X_n\}$ is a first-order Markov chain, taking values in $\{1,2\}$ and having the following transition probability matrix:

$$\Pi = \begin{bmatrix} 1-\theta & \theta \\ 1 & 0 \end{bmatrix}.$$

In Section 4.1, we will show that $I(X;Y)$ is concave with respect to θ, and in Section 4.2, we apply the algorithm in [16] to numerically evaluate $C^{(1)}(\mathcal{S}_0,\varepsilon)$, whose convergence is guaranteed by the above-mentioned concavity result. Finally, in Section 4.3, we characterize the full asymptotics of $C^{(1)}(\mathcal{S}_0,\varepsilon)$ around $\varepsilon=0$.

4.1 Concavity

The concavity of the mutual information rate of special families of finite-state machine channels (FSMCs) has been considered in [21] and [29]. The results therein actually imply that the concavity of $I(X;Y)$ with respect to some parameterization of the Markov chain X when ε is small enough. In this section, however, we will show that $I(X;Y)$ is concave with respect to θ, irrespective of the values of ε. Below is the main theorem of this section.

Theorem 4.1. *For all $\varepsilon\in[0,1)$, $I(X;Y)$ is strictly concave with respect to θ, $0\le\theta\le 1$.*

Proof. From Corollary 2.5, it follows that to prove the theorem, it suffices to show that for any $n\ge 1$, $H(X_0|X_{-n})$ is strictly concave with respect to θ, $0\le\theta\le 1$. To prove this, we will deal with the following several cases:

Case 1: $n=1$. Straightforward computations give

$$H(X_0|X_{-1}) = \frac{-\theta\log\theta-(1-\theta)\log(1-\theta)}{1+\theta}$$

and

$$H''(X_0|X_{-1}) = \frac{1}{(1+\theta)^3}\left\{2\log\theta-4\log(1-\theta)-\frac{1}{\theta}-\frac{4}{1-\theta}+1\right\}.$$

One checks that the function within the brace is negative and it takes the maximum at $\theta=1/2$. Therefore $H(X_0|X_{-1})$ is strictly concave in θ.

Case 2: $n\ge 2$. By definition,

$$H(X_0|X_{-n}) \;=\; P(X_{-n}=1)H(X_0|X_{-n}=1)+P(X_{-n}=2)H(X_0|X_{-n}=2).$$

The following facts can be verified easily:

(i) $f(\theta)\triangleq P(X_{-n}=1)=1-P(X_{-n}=2)=\frac{1}{1+\theta}$;

(ii) the n-step transition probability matrix of the Markov chain $\{X_n\}$ is

$$\Pi^n = \begin{bmatrix} g_{n+1}(\theta) & 1-g_{n+1}(\theta) \\ g_n(\theta) & 1-g_n(\theta) \end{bmatrix},$$

where

$$g_n(\theta) \triangleq \frac{1-(-\theta)^n}{1+\theta};$$

(iii) $H(y) = -y\log y - (1-y)\log(1-y)$ is strictly concave with respect to y for $y \in (0,1)$.

With the above notation, we have

$$H(X_0|X_{-n}) = f(\theta)H(g_{n+1}(\theta)) + (1-f(\theta))H(g_n(\theta))$$

and

$$\begin{aligned} H''(X_0|X_{-n}) = {} & fH''(g_{n+1})(g'_{n+1})^2 + (1-f)H''(g_n)(g'_n)^2 & (18)\\ & + f''(H(g_{n+1}) - H(g_n)) & (19)\\ & - 2f'H'(g_n)g'_n + (1-f)H'(g_n)g''_n & (20)\\ & + 2f'H'(g_{n+1})g'_{n+1} + fH'(g_{n+1})g''_{n+1}, & (21) \end{aligned}$$

where $f = f(\theta)$ and $g_n = g_n(\theta)$. It follows from (i) and (iii) that the term (18) is strictly negative. So, to prove the theorem, it suffices to show that $T \triangleq (19) + (20) + (21) \leq 0$.

By the mean value theorem,

$$\begin{aligned} (19) &= \frac{2}{(1+\theta)^3}(g_{n+1} - g_n)\log\frac{1-z_1}{z_1} \\ &= \frac{2(-\theta)^n(1+\theta)}{(1+\theta)^4}\log\frac{1-z_1}{z_1}, \end{aligned}$$

where z_1 lies between g_n and g_{n+1}. As a function of z_1, $\frac{2(-\theta)^n(1+\theta)}{(1+\theta)^4}\log\frac{1-z_1}{z_1}$ takes the maximum at g_n. It then follows that

$$\begin{aligned} T &\leq \frac{c_n\{2\theta - 2 + (-\theta)^{n-1}[(n^2-3n)\theta^2 + 2(n^2-n-2) + n^2+n]\}}{(1+\theta)^4} \\ &\quad + \frac{c_{n+1}\{4 - (-\theta)^{n-1}[(n^2-3n)\theta^2 + 2(n^2-n-2) + n^2+n]\}}{(1+\theta)^4} \\ &= \frac{c_n[2\theta - 2 + Q(n,\theta)(-\theta)^{n-1}]}{(1+\theta)^4} + \frac{c_{n+1}[4 - Q(n,\theta)(-\theta)^{n-1}]}{(1+\theta)^4}, \end{aligned}$$

where

$$Q(n,\theta) = (n^2-3n)\theta^2 + 2(n^2-n-2)\theta + n^2 + n$$

and

$$c_n = \log\frac{1-g_n}{g_n}.$$

We then consider the following several cases:

Case 2.1: n is a positive even number. We first consider the case that $g_n \leq \frac{1}{2}$. For this case, obviously we have $c_n \geq 0$, $g_{n+1} > \frac{1}{2}$ and $Q(n,\theta) > 0$, which further implies that

$$T \leq \frac{c_n[2\theta - 2 - Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} + \frac{c_{n+1}[4 + Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} < 0.$$

Now, for the case that $g_n > \frac{1}{2}$, again obviously we have $c_n < 0$ and furthermore,

$$c_{n+1} - c_n = \log \frac{1 - g_{n+1}}{g_{n+1}} - \log \frac{1 - g_n}{g_n} \leq 0,$$

where we have used the fact that $g_n \leq g_{n+1}$ for a positive even number n. Now, we are ready to deduce that

$$\begin{aligned} T &\leq \frac{c_n[2\theta - 2 - Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} + \frac{c_{n+1}[4 + Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} \\ &= \frac{(c_{n+1} - c_n)(4 + Q(n,\theta)\theta^{n-1})}{(1+\theta)^4} + \frac{c_n(2\theta+2)}{(1+\theta)^4} \\ &< 0. \end{aligned}$$

<u>**Case 2.2: n is a positive odd integer and $n \geq 3$.**</u> In this case, we have

$$\begin{aligned} T &\leq \frac{c_n[2\theta - 2 + Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} + \frac{c_{n+1}[4 - Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} \\ &= \frac{(c_{n+1} - c_n)[4 - Q(n,\theta)\theta^{n-1}]}{(1+\theta)^4} + \frac{c_n(2\theta+2)}{(1+\theta)^4} \\ &\leq \frac{1}{(1+\theta)^4}\left\{\frac{[4 - Q(n,\theta)\theta^{n-1}]\theta^n}{(1-y_2)y_2} + (2\theta+2)(1/g_n - 2)\right\}, \end{aligned}$$

where the last inequality follows from the mean value theorem, the inequality $\log z \leq z - 1$ for $z > 0$ and the fact that z_2 lies between g_n and g_{n+1}.

Let $y_2 = B_n/(1+\theta)$ and $C_n = 1 + \theta - B_n$, where $B_n \in [1 - \theta^{n+1}, 1 + \theta^n]$. Then

$$\begin{aligned} T &\leq \frac{1}{(1+\theta)^4}\left\{\frac{[4 - Q(n,\theta)\theta^{n-1}]\theta^n}{(1-g_2)g_2} + (2\theta+2)(1/g_n - 2)\right\} \\ &\leq \frac{(2\theta - 2 - 4\theta^n)B_nC_n + (1+\theta)(1+\theta^n)\theta^n(4 - Q(n,\theta)\theta^{n-1})}{B_nC_n(1+\theta)^3(1+\theta^n)}. \end{aligned}$$

Note that the above numerator, as a function of B_n, takes the maximum at $B_n = 1 + \theta^n$. Denote this maximum by $-2\theta(1+\theta^n)h_1(n,\theta)$, where

$$h_1(n,\theta) = 1 - \theta - 3\theta^{n-1} + \theta^n + \frac{Q(n,\theta)}{2}\theta^{2n-2} + \frac{Q(n,\theta) - 4}{2}\theta^{2n-1}.$$

To complete the proof, it suffices to prove $h_1(n,\theta) \geq 0$. Substituting $Q(n,\theta)$ into $h_1(n,\theta)$, we have

$$\begin{aligned} h_1(n,\theta) &= 1 - \theta - 3\theta^{n-1} + \theta^n + \frac{Q(n,\theta)}{2}\theta^{2n-2} + \frac{Q(n,\theta) - 4}{2}\theta^{2n-1} \\ &= 1 - \theta - 3\theta^{n-1} + \theta^n + \frac{\theta^{2n-2}}{2}[(n^2 - 3n)\theta^3 \\ &\quad +(3n^2 - 5n - 4)\theta^2 + (3n^2 - n - 8)\theta + n^2 + n] \\ &\geq h(n,\theta), \end{aligned}$$

where

$$h(n,\theta) = 1 - \theta - 3\theta^{n-1} + \theta^n + (4n^2 - 4n - 6)\theta^{2n+1}.$$

The following facts can be easily verified:

(a) $h(n, \theta)$ takes the minimum at some θ_0, where θ_0 satisfies the following equation

$$h'(n, \theta) = 0; \tag{22}$$

(b) $h'(n, \theta) < 0$ for $\theta \in [0, 1/2]$ and $n \geq 11$.

It then follows from (a) and (b) that $\theta_0 \geq 1/2$ for $n > 11$. Solving (22) in θ^{n-2}, we have

$$\theta_0^{n-2} = \frac{(3 - \theta_0)n - 3 + \sqrt{((3 - \theta_0)n - 3)^2 + 4(2n + 1)(4n^2 - 4n - 6)\theta_0^5}}{2(2n + 1)(4n^2 - 4n - 6)\theta_0^5}.$$

For $n \geq 11$, substituting θ_0^{n-2} into $h(n, \theta)$, we have

$$\begin{aligned} h_1(n, \theta) &\geq h(n, \theta) \\ &\geq h(n, \theta_0) \\ &= 1 - \theta_0 - 3\theta_0^{n-1} + \theta_0^n + (4n^2 - 4n - 6)\theta_0^{2n+1} \\ &\geq \theta_0^{n-1}[-3 + \theta_0 + (4n^2 - 4n - 6)\theta_0^4 \cdot \theta_0^{n-2}] \\ &\geq \theta_0^{n-1} v(n), \end{aligned}$$

where

$$v(n) = -\frac{5}{2} + \frac{2n - 3 + \sqrt{(2n - 3)^2 + (2n + 1)(4n^2 - 4n - 6)2^{-3}}}{2(2n + 1)}.$$

One checks that $v(n) > 0$ for $n \geq 65$. Now, with the fact that $h_1(n, \theta) > 0$ for $3 \leq n \leq 65$ (this can be verified via tedious yet straightforward computations since $h_1(n, \theta)$ is an elementary function), we conclude that for $h_1(n, \theta) \geq 0$ for all $n \geq 3$ and $\theta \in [0, 1]$. □

4.2 Numerical evaluation of $C^{(1)}(\mathcal{S}_0, \varepsilon)$

When $\varepsilon = 0$, that is, when the channel is "perfect" with no erasures, both $C(\mathcal{S}, \varepsilon)$ and $C^{(m)}(\mathcal{S}, \varepsilon)$ boil down to the noiseless capacity of the constraint $\mathcal{S}$, which can be explicitly computed [33]; however, little progress has been made for the case when $\varepsilon > 0$ due to the lack of simple and explicit characterization for $C(\mathcal{S}, \varepsilon)$ and $C^{(m)}(\mathcal{S}, \varepsilon)$. In terms of numerically computing $C(\mathcal{S}, \varepsilon)$ and $C^{(m)}(\mathcal{S}, \varepsilon)$, relevant work can be found in the subject of FSMCs, as input-constrained memoryless erasure channels can be regarded as special cases of FSMCs. Unfortunately, the capacity of an FSMC is still largely unknown and the fact that our channel is only a special FSMC does not seem to make the problem easier.

Recently, Vontobel *et al.* [48] proposed a generalized Blahut-Arimoto algoritm (GBAA) to compute the capacity of an FSMC; and in [16], Han also proposed a randomized algorithm to compute the capacity of an FSMC. For both algorithms, the concavity of the mutual information rate is a desired property for the convergence (the convergence of the GBAA requires, in addition, the concavity of certain conditional entropy rate). On the other hand, as elaborated in [29], such a desired property, albeit established for a few special cases [21, 29], is not true in general.

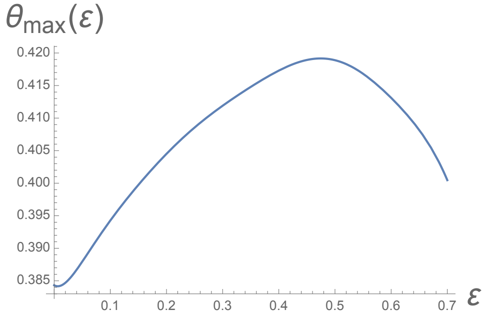

Figure 1: Capacity-achieving Distribution

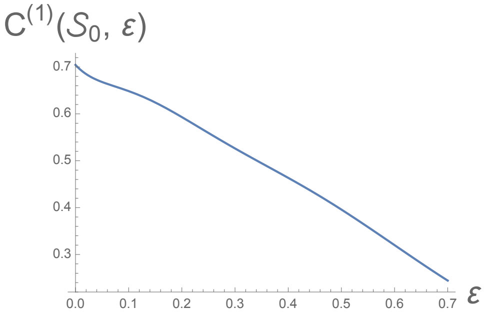

Figure 2: Capacity

The concavity established in the previous section allows us to numerically compute $C^{(1)}(\mathcal{S}_0, \varepsilon)$ using the algorithm in [16]. The randomized algorithm proposed in [16] iteratively compute $\{\theta_n\}$ in the following way:

$$\theta_{n+1} = \begin{cases} \theta_n, & \text{if } \theta_n + a_n g_{n^b}(\theta_n) \in [0,1], \\ \theta_n + a_n g_{n^b}(\theta_n), & \text{otherwise,} \end{cases}$$

where $g_{n^b}(\theta_n)$ is a simulator for $I'(X;Y)$ (for details, see [16]). The author shows that $\{\theta_n\}$ converges to the first-order capacity-achieving distribution if $I(X;Y)$ is concave with respect to θ, which has been proven in Theorem 4.1. Therefore, with proven convergence, this algorithm can be used to compute the first-order capacity-achieving distribution $\theta(\varepsilon)$ and the first-order capacity $C^{(1)}(\mathcal{S}_0, \varepsilon)$ (in bits), which are shown in Fig. 1 and Fig. 2, respectively.

4.3 Full Asymptotics

As in Section 4.1, the noiseless capacity of $(1,\infty)$-RLL constraint $\mathcal{S}_0$ is achieved by the first-order Markov chain with the transition probability matrix (17). So, we have

$$\theta_{\max}(\varepsilon) = \operatorname*{argmax}_{\theta} I(X;Y) = 1/\lambda^2,$$

where $\lambda = (1+\sqrt{5})/2$. In this section, we give a full asymptotic formula for $\theta_{max}(\varepsilon)$ around $\varepsilon = 0$, which further leads to a full asymptotic formula for $C^{(1)}(\mathcal{S}_0, \varepsilon)$ around $\varepsilon = 0$.

The following theorem gives the Taylor series of $\theta_{\max}$ in ε around $\varepsilon = 0$, which leads to an explicit formula for the n-th derivative of $C^{(1)}(\mathcal{S}_0, \varepsilon)$ at $\varepsilon = 0$, whose coefficients can be explicitly computed.

Theorem 4.2. *a) $\theta_{max}(\varepsilon)$ is analytic in ε for $\varepsilon \in [0, 1)$ and*

$$\begin{aligned}\theta^{(n)}_{max}(0) = &- \left(\frac{\mathrm{d}^2 H(X_0|X_{-1})}{\mathrm{d}\theta^2} \left(\frac{1}{\lambda^2}\right) \right)^{-1} \\ &\times \left\{ \sum_{k=1}^{n} \binom{n}{k} k! \sum_{m_1, m_2, \cdots, m_{n-k}} a(m_1, \cdots, m_{n-k}) \quad \frac{\mathrm{d}^{m_1+\cdots+m_{n-k}+1} H(X_0|X_{-k-1})}{\mathrm{d}\theta^{m_1+\cdots+m_{n-k}+1}} \left(\frac{1}{\lambda^2}\right) \prod_{j=1}^{n-k} (\theta^{(j)}_{max}(0))^{m_j} \right. \\ &\left. + \sum_{m_1, m_2, \cdots, m_n = 0} a(m_1, \cdots, m_n) \frac{\mathrm{d}^{m_1+\cdots+m_n+1} H(X_0|X_{-1})}{\mathrm{d}\theta^{m_1+\cdots+m_n+1}} \left(\frac{1}{\lambda^2}\right) \prod_{j=1}^{n} (\theta^{(j)}_{max}(0))^{m_j} \right\},\end{aligned} \tag{23}$$

where $\sum_{m_1, m_2, \cdots, m_{n-k}}$ is taken over all nonnegative intergers $m_1, \cdots, m_{n-k}$ satisfying the constraint

$$m_1 + 2m_2 + \cdots + (n-k)m_{n-k} = n - k$$

and

$$a(m_1, \cdots, m_{n-k}) = \frac{(n-k)!}{m_1! 1^{m_1} \cdots m_{n-k}! (n-k)^{m_{n-k}}}.$$

b) $C^{(1)}(\mathcal{S}_0, \varepsilon)$ is analytic in ε for $\varepsilon \in [0, 1)$ with the following Taylor series expansion around $\varepsilon = 0$:

$$C^{(1)}(\mathcal{S}_0, \varepsilon) = \sum_{n=0}^{\infty} \left(\left.\frac{\mathrm{d}^n G_0(\varepsilon)}{\mathrm{d}\varepsilon^n}\right|_{\varepsilon=0} + \left.\frac{d^{n-1}(G_1(\varepsilon) - 2G_0(\varepsilon))}{\mathrm{d}\varepsilon^{n-1}}\right|_{\varepsilon=0} + \sum_{k=2}^{n} \binom{n}{k} \left.\frac{\mathrm{d}^{n-k}}{\mathrm{d}\varepsilon^{n-k}} \left\{ (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon)) \right\}\right|_{\varepsilon=0} \right) \varepsilon^n, \tag{24}$$

where $G_k(\varepsilon) = H(X_0|X_{-k-1})(\theta_{max}(\varepsilon))$.

Proof. a) For $\varepsilon > 0$,

$$I(X;Y) = \begin{cases} 0 & \theta = 0 \text{ or } 1, \\ > 0 & \theta \in (0, 1). \end{cases}$$

With Theorem 4.1 establishing the concavity of $I(X;Y)$, $\theta_{\max}$ should be the unique zero point of the derivative of the mutual information rate. So, $\theta_{\max}(\varepsilon) \in (0, 1)$ and satisfies

$$0 = \frac{\mathrm{d}I(X;Y)}{\mathrm{d}\theta} = (1-\varepsilon)^2 \sum_{k=0}^{\infty} \frac{\mathrm{d}H(X_0|X_{-k-1})}{\mathrm{d}\theta} \varepsilon^k. \tag{25}$$

According to the analytic implicit function theorem [27], $\theta_{\max}(\varepsilon)$ is analytic in ε for $\varepsilon \in [0, 1)$. In the following the s-th order derivative of $\theta_{\max}(\varepsilon)$ at $\varepsilon = 0$ is computed. It follows from the Leibniz formula and the Faa di Bruno formula [7] that

$$
\begin{aligned}
0 &= \sum_{k=0}^{\infty} \frac{\mathrm{d}^n}{\mathrm{d}\varepsilon^n} \left\{ \frac{\mathrm{d}H(X_0|X_{-k-1})}{\mathrm{d}\theta} \varepsilon^k \right\} \Bigg|_{\varepsilon=0} \\
&= \sum_{k=0}^{n} k! \binom{n}{k} \frac{\mathrm{d}^{(n-k)}}{\mathrm{d}\varepsilon^{n-k}} \left\{ \frac{\mathrm{d}H(X_0|X_{-k-1})}{\mathrm{d}\theta} \right\} \Bigg|_{\varepsilon=0} \\
&= \sum_{k=0}^{n} k! \binom{n}{k} \sum_{m_1, m_2, \cdots, m_{n-k}} a(m_1, \cdots, m_{n-k}) \frac{\mathrm{d}^{m_1+\cdots+m_{n-k}+1} H(X_0|X_{-k-1})}{\mathrm{d}\theta^{m_1+\cdots+m_{n-k}+1}} \left(\frac{1}{\lambda^2} \right) \prod_{j=1}^{n-k} (\theta_{\max}^{(j)}(0))^{m_j}
\end{aligned}
\tag{26}
$$

which immediately implies a).

b) Note that

$$
\begin{aligned}
C^{(1)}(\mathcal{S}_0, \varepsilon) &= (1-\varepsilon)^2 \sum_{k=0}^{\infty} G_k(\varepsilon) \varepsilon^k \\
&= G_0(\varepsilon) + (G_1(\varepsilon) - 2G_0(\varepsilon))\varepsilon + \sum_{k=2}^{\infty} (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon))\varepsilon^k.
\end{aligned}
$$

It then follows from the Leibniz formula that

$$
\begin{aligned}
&\frac{\mathrm{d}^n}{\mathrm{d}\varepsilon^n} \left\{ \sum_{k=2}^{\infty} (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon))\varepsilon^k \right\} \\
&\qquad = \sum_{k=2}^{\infty} \sum_{t=0}^{n} \binom{n}{t} \frac{\mathrm{d}^{n-t}}{\mathrm{d}\varepsilon^{n-t}} \left\{ (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon)) \right\} \varepsilon^{k-t}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{\mathrm{d}^n C^{(1)}(\mathcal{S}_0, \varepsilon)}{\mathrm{d}\varepsilon^s} \Bigg|_{\varepsilon=0} &= \frac{\mathrm{d}^n G_0(\varepsilon)}{d\varepsilon^n} \Bigg|_{\varepsilon=0} + \frac{\mathrm{d}^{n-1}(G_1(\varepsilon) - 2G_0(\varepsilon))}{\mathrm{d}\varepsilon^{n-1}} \Bigg|_{\varepsilon=0} \\
&\quad + \sum_{k=2}^{\infty} \sum_{t=0}^{n} \binom{n}{t} \frac{\mathrm{d}^{n-t}}{\mathrm{d}\varepsilon^{n-t}} \left\{ (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon)) \right\} \varepsilon^{k-t} \Bigg|_{\varepsilon=0} \\
&= \frac{\mathrm{d}^n G_0(\varepsilon)}{\mathrm{d}\varepsilon^n} \Bigg|_{\varepsilon=0} + \frac{\mathrm{d}^{n-1}(G_1(\varepsilon) - 2G_0(\varepsilon))}{\mathrm{d}\varepsilon^{n-1}} \Bigg|_{\varepsilon=0} \\
&\quad + \sum_{k=2}^{n} \binom{n}{k} \frac{\mathrm{d}^{n-k}}{\mathrm{d}\varepsilon^{n-k}} \left\{ (G_k(\varepsilon) + G_{k-2}(\varepsilon) - 2G_{k-1}(\varepsilon)) \right\} \Bigg|_{\varepsilon=0},
\end{aligned}
$$

which immediately implies b). □

Despite their convoluted looks, (23) and (24) are explicit and computable. Below, we list the coefficients of $C^{(1)}(\mathcal{S}_0, \varepsilon)$ (in bits) and $\theta_{\max}(\varepsilon)$ up to the third order, which are numerically computed according to (23) and (24) and rounded off to the ten thousandths decimal digit:

Table 1:

	ε^0	ε^1	ε^2	ε^3
$\theta_{\max}(\varepsilon)$	0.3820	0.0462	0.1586	0.2455
$C_1(\varepsilon)$	0.6942	-0.6322	0.0159	-0.0625

5 Feedback with Input-Constraint

In this section, we consider the input-constrained erasure channel (1) as in Section 1 however with possible feedback, and we are interested in comparing its feedback capacity $C_{FB}(\mathcal{S}, \varepsilon)$ and its non-feedback capacity $C(\mathcal{S}, \varepsilon)$. The following theorem states that for the erasure channel without any input-constraint, feedback does not increase the capacity and both of them can be computed explicitly. This result is in fact implied by Theorem 12 in [46], where a random coding argument has been employed in the proof; we nonetheless give an alternative proof in Appendix A for completeness.

Theorem 5.1. *For the erasure channel (1) without any input constraints, feedback does not increase the capacity, and we have*

$$C_{FB}(\mathcal{X}^*, \varepsilon) = C(\mathcal{X}^*, \varepsilon) = (1-\varepsilon)\log K.$$

On the other hand, we will show in the following that feedback may increase the capacity when the input constraint in the erasure channel is non-trivial. As elaborated below, this is achieved by comparing the asymptotics of the feedback capacity and the non-feedback capacity for a special input-constrained erasure channel.

In [37], Sabag *et al.* computed an explicit formula of feedback capacity for BEC with $(1,\infty)$-RLL input constraint $\mathcal{S}_0$.

Theorem 5.2. *[37] The feedback capacity of the* $(1,\infty)$*-RLL input-constrained erasure channel is*

$$C_{FB}(\mathcal{S}_0, \varepsilon) = \max_{0\le p\le \frac{1}{2}} \frac{H(p)}{p + \frac{1}{1-\varepsilon}},$$

where the unique maximizer $p(\varepsilon)$ *satisfies* $p = (1-p)^{2-\varepsilon}$.

Clearly, the explicit formula in Theorem 5.2 readily gives the asymptotics of the feedback capacity.

To see this, note that $p(0) = 1/\lambda^2$ and $p(1) = 1/2$. Straightforward computations yield

$$\frac{\mathrm{d}\log p(\varepsilon)}{\mathrm{d}\varepsilon} = -\frac{(1-p(\varepsilon))\log p(\varepsilon)}{(1-p(\varepsilon)+p(\varepsilon)(2-\varepsilon))(2-\varepsilon)}.$$

Hence,

$$\begin{aligned}
C_{FB}(\mathcal{S}_0, \varepsilon) &= -\frac{H(p(\varepsilon))}{p(\varepsilon) + \frac{1}{1-\varepsilon}} \\
&= -\frac{(1-\varepsilon)\log p(\varepsilon)}{2-\varepsilon} \\
&= -\frac{1}{2}\log p(\varepsilon) + \frac{1}{2}\log p(\varepsilon)\sum_{k=1}^{\infty}\left(\frac{\varepsilon}{2}\right)^k \\
&= -\frac{1}{2}\log p(0) - \frac{1}{2}\left.\frac{\mathrm{d}\log p(\varepsilon)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0}\varepsilon + \frac{\varepsilon}{4}\log p(0) + O(\varepsilon^2) \\
&= \log\lambda - \frac{\lambda^2}{\lambda^2+1}\log\lambda\cdot\varepsilon + O(\varepsilon^2).
\end{aligned}$$

It then follows from straightforward computations that for the case when ε is close to 0, $C(\mathcal{S}_0, \varepsilon) < C_{FB}(\mathcal{S}_0, \varepsilon)$. So, we have proven the following theorem:

Theorem 5.3. *For a BEC(ε) with the $(1, \infty)$-RLL input constraint, feedback increases the channel capacity when ε is small enough.*

Remark 5.4. An independent work in [44] also found that feedback does increase the capacity of a BEC(ε) with the same input constraint $\mathcal{S}_0$, by comparing a tighter bound of non-feedback capacity $C(\mathcal{S}_0, \varepsilon)$, obtained via a dual capacity approach, with the feedback capacity $C_{FB}(\mathcal{S}_0, \varepsilon)$.

Remark 5.5. Recently, Sabag *et al.* [38] also computed an explicit asymptotic formula for the feedback capacity of a BSC(ε) with the input supported on the $(1, \infty)$-RLL constraint. By comparing the asymptotics of the feedback capacity with the that of non-feedback capacity [19], they showed that feedback does increase the channel capacity in the high SNR regime.

It is well known that for any memoryless channel without any input constraint, feedback does not increase the channel capacity. Theorem 5.1 states when there is no input constraint, the feedback does not increase the capacity of the erasure channel even with the presence of the channel memory. Theorem 5.3 says that feedback may increase the capacity of input-constrained erasure channels even if there is no channel memory. These two theorems, together with the results in [38, 44], suggest the intricacy of the interplay between feedback, memory and input constraints.

Appendices

A Proof of Theorem 5.1

We first prove that

$$C(\mathcal{X}^*, \varepsilon) = (1-\varepsilon)\log K. \tag{27}$$

A similar argument using the independence of $\{X_i\}$ and $\{E_i\}$ as in the proof of (3) yields that

$$p(y_1^n) = P(E_{\mathcal{I}(y_1^n)} = 1, E_{\bar{\mathcal{I}}(y_1^n)} = 0)P(X_{\mathcal{I}(y_1^n)} = y_{\mathcal{I}(y_1^n)}).$$

It then follows that

$$\begin{aligned}
H(Y_1^n) &= -\sum_{y_1^n} p(y_1^n) \log p(y_1^n) \\
&= -\sum_{y_1^n} p(y_1^n) \log P(E_{\mathcal{I}(y_1^n)} = 1, E_{\bar{\mathcal{I}}(y_1^n)} = 0) - \sum_{y_1^n} P(y_1^n) \log P(X_{\mathcal{I}(y_1^n)} = y_{\mathcal{I}(y_1^n)}) \\
&= -\sum_{D \subseteq [1,n]} \sum_{y_1^n : \mathcal{I}(y_1^n) = D} P(E_D = 1, E_{D^C} = 0) P(X_D = y_D) \log P(E_D = 1, E_{D^c} = 0) \\
&\quad - \sum_{D \subseteq [1,n]} \sum_{y_1^n : \mathcal{I}(y_1^n) = D} P(E_D = 1, E_{D^c} = 0) P(X_D = y_D) \log P(X_D = y_D) \\
&= -\sum_{D \subseteq [1,n]} P(E_D = 1, E_{D^c} = 0) \log P(E_D = 1, E_{D^c} = 0) + \sum_{D \subseteq [1,n]} P(E_D = 1, E_{D^c} = 0) H(X_D) \\
&= \sum_{D \subseteq [1,n]} P(E_D = 1, E_{D^c} = 0) H(X_D) + H(E_1^n) \\
&\leq H(E_1^n) + \sum_{D \subseteq [1,n]} P(E_D = 1, E_{D^c} = 0) |D| \log K \\
&= H(E_1^n) + \mathbf{E}[E_1 + \cdots + E_n] \log K, \qquad (28)
\end{aligned}$$

where the only inequality becomes equality if $\{X_n\}$ is i.i.d. with the uniform distribution. It then further follows that

$$\begin{aligned}
C(\mathcal{X}^*, \varepsilon) &= \lim_{n \to \infty} \frac{1}{n} \sup_{p(x_1^n)} I(X_1^n; Y_1^n) \\
&= \lim_{n \to \infty} \frac{1}{n} \sup_{p(x_1^n)} (H(Y_1^n) - H(Y_1^n | X_1^n)) \\
&= \lim_{n \to \infty} \frac{1}{n} \sup_{p(x_1^n)} (H(Y_1^n) - H(E_1^n)) \\
&\overset{(a)}{\leq} \lim_{n \to \infty} \frac{1}{n} \mathbf{E}[E_1 + \cdots + E_n] \log K \\
&\overset{(b)}{=} P(E_1 = 1) \log K \\
&= (1 - \varepsilon) \log K,
\end{aligned}$$

where (a) follows from (28) and (b) follows from the ergodicity of $\{E_n\}$. The desired (27) then follows from the fact that the only inequality (a) becomes equality if $\{X_n\}$ is i.i.d. with the uniform distribution.

We next prove that

$$C_{FB}(\mathcal{X}^*, \varepsilon) \leq (1 - \varepsilon) \log K,$$

which, together with (27), immediately implies the theorem.

Let W, independent of $\{E_i\}$, be the message to be sent and $X_i(W, Y_1^{i-1})$ denote the encoding function. As shown in [42],

$$C_{FB}(\mathcal{X}^*, \varepsilon) = \lim_{n \to \infty} \frac{1}{n} \sup_{\{p(X_i = \cdot | X_1^{i-1}, Y_1^{i-1}) : i = 1, \cdots, n\}} I(W; Y_1^n).$$

Using the chain rule for entropy, we have

$$\begin{aligned}
H(Y_1^n|W) &= \sum_{i=1}^n H(Y_i|W, Y_1^{i-1}) \\
&\overset{(a)}{=} \sum_{i=1}^n H(E_iX_i|W, Y_1^{i-1}, X_1^i, E_1^{i-1}) \\
&= \sum_{i=1}^n H(E_i|W, Y_1^{i-1}, X_1^i, E_1^{i-1}) \\
&\overset{(b)}{=} \sum_{i=1}^n H(E_i|E_1^{i-1}) \\
&= H(E_1^n), \qquad (29)
\end{aligned}$$

where (a) follows from the fact that X_i is a function of W and Y_1^{i-1} and $E_i = 0$ if and only if $Y_i = 0$, (b) follows from the independence of W and $\{E_i\}$.

Note that for $y_i \neq 0$,

$$\begin{aligned}
p(y_i|w, y_1^{i-1}) &= P(X_i(w, y_1^{i-1}) = y_i, E_i = 1|w, y_1^{i-1}) \\
&= P(X_i(w, y_1^{i-1}) = y_i|w, y_1^{i-1})P(E_i = 1|X_i(w, y_1^{i-1}) = y_i, w, y_1^{i-1}) \\
&= P(X_i(w, y_1^{i-1}) = y_i|w, y_1^{i-1})P(E_i = 1|E_{\mathcal{I}(y_1^{i-1})} = 1, E_{\bar{\mathcal{I}}(y_1^{i-1})} = 0). \qquad (30)
\end{aligned}$$

And for $y_i = 0$,

$$\begin{aligned}
p(y_i|w, y_1^{i-1}) &= P(E_i = 0|E_{\mathcal{I}(y_1^{i-1})} = 1, E_{\bar{\mathcal{I}}(y_1^{i-1})} = 0, w, y_1^{i-1}) \\
&\overset{(a)}{=} P(E_i = 0|E_{\mathcal{I}(y_1^{i-1})} = 1, E_{\bar{\mathcal{I}}(y_1^{i-1})} = 0), \qquad (31)
\end{aligned}$$

where (a) follows from the independence of W and $\{E_i\}$. It then follows that

$$\begin{aligned}
p(y_1^n) &= \sum_w p(w)p(y_1^n|w) \\
&= \sum_w p(w) \prod_{i=1}^n p(y_i|w, y_1^{i-1}) \\
&\overset{(a)}{=} \sum_w p(w) \prod_{i\in\mathcal{I}(y_1^n)} p(y_i|w, y_1^{i-1}) \prod_{i\in\bar{\mathcal{I}}(y_1^n)} p(y_i|w, y_1^{i-1}) \\
&\overset{(b)}{=} \sum_w p(w) \prod_{i\in\mathcal{I}(y_1^n)} p(y_i|w, y_1^{i-1}) \prod_{i\in\bar{\mathcal{I}}(y_1^n)} P(E_i = 0|E_{\mathcal{I}(y_1^{i-1})} = 1, E_{\bar{\mathcal{I}}(y_1^{i-1})} = 0) \\
&= \left\{ \sum_w p(w) \prod_{i\in\mathcal{I}(y_1^n)} P(X_i(w, y_1^{i-1}) = y_i|w, y_1^{i-1}) \right\} P(E_{\mathcal{I}(y_1^n)} = 1, E_{\bar{\mathcal{I}}(y_1^n)} = 0),
\end{aligned}$$

where (a) follows from (30) and (b) follows from (31). Since for any $D \subseteq [1, n]$,

$$\sum_{y_1^n:\mathcal{I}(y_1^n)=D} p(y_1^n) = P(E_D = 1, E_{D^c} = 0),$$

which implies that

$$\left\{q(\tilde{y}_1^n)) \triangleq \sum_w p(w) \prod_{i \in \mathcal{I}(\tilde{y}_1^n)} P(X_i(w, \tilde{y}_1^{i-1}) = \tilde{y}_i | w, \tilde{y}_1^{i-1}) : \mathcal{I}(\tilde{y}_1^n) = \mathcal{I}(y_1^n)\right\}$$

is an $M^{|\mathcal{I}(y_1^n)|}$-dimensional probability mass function. Therefore, through a similar argument as before, we have

$$\begin{aligned}
H(Y_1^n) &= -\sum_{y_1^n} p(y_1^n) \log p(y_1^n) \\
&= -\sum_{y_1^n} p(y_1^n) \log P(E_{\mathcal{I}(y_1^n)} = 1, E_{\bar{\mathcal{I}}(y_1^n)} = 0) - \sum_{y_1^n} P(E_{\mathcal{I}(y_1^n)} = 1, E_{\bar{\mathcal{I}}(y_1^n)} = 0) q(y_1^n) \log q(y_1^n) \\
&= H(E_1^n) - \sum_{D \subseteq [1,n]} \sum_{y_1^n : \mathcal{I}(y_1^n) = D} P(E_D = 1, E_{D^c} = 0) q(y_1^n) \log q(y_1^n) \\
&\leq H(E_1^n) + \sum_{D \subseteq [1,n]} P(E_D = 1, E_{D^c} = 0) |D| \log K \\
&= H(E_1^n) + \mathbf{E}[E_1 + \cdots + E_n] \log K, \qquad (32)
\end{aligned}$$

where the inequality follows from the fact that $q(y_1^n)$ is an $M^{|D|}$-dimensional probability mass function.

Combining (29) and (32), we have

$$\begin{aligned}
C_{FB}(\mathcal{X}^*, \varepsilon) &= \lim_{n \to \infty} \frac{1}{n} \sup_{\{p(X_i = \cdot | X_1^{i-1}, y_1^{i-1}) : i = 1, \cdots, n\}} I(W; Y_1^n) \\
&\leq \lim_{n \to \infty} \frac{1}{n} \mathbf{E}[E_1 + \cdots + E_n] \log K \\
&= P(E_1 = 1) \log K \\
&= (1 - \varepsilon) \log K,
\end{aligned}$$

as desired.

Acknowledgement. We would like to thank Navin Kashyap, Haim Permuter, Oron Sabag and Wenyi Zhang for insightful discussions and suggestions and for pointing out relevant references that result in great improvements in many aspects of this work.

References

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[2] D. M. Arnold and H. A. Loeliger, "On the information rate of binary-input channels with memory," in *Proceedings of IEEE International Conference on Communications*, vol. 9, pp. 2692–2695, Jun. 2001.

[3] D. M. Arnold, H. A. Loeliger, P. O. Vontobel, A. Kavcic and W. Zeng, "Simulation-Based Computation of Information Rates for Channels With Memory," *IEEE. Trans. Inf. Theory*, vol.52, no.8, pp. 3498–3508, Aug. 2006.

[4] D. Blackwell, "The entropy of functions of finite-state markov chains," *Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, pp. 13–20, 1957.

[5] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Apr. 1972.

[6] J. Chen and P. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1295–1303, Mar. 2008.

[7] G. Constantine and T. Savits. "A Multivariate Faa Di Bruno Formula with Applications", *Trans. Amer. Math. Soc.*, vol. 348, no. 2, pp. 503–520, Feb. 1996.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[9] A. Dembo, "On Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1072–1076, Sep. 1989.

[10] R. Durrett, *Probability: Theory and Examples*, Cambridge University Press, 2010.

[11] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.

[12] G. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 363–378, Mar. 1972.

[13] R. Gallager, *Information theory and reliable communication.* New York: Wiley, 1968.

[14] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, Mar. 1996.

[15] R. M. Gray, *Entropy and Information Theory.* Springer US, 2011.

[16] G. Han, "A randomized algorithm for the capacity of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3651-3669, July 2015.

[17] G. Han and B. Marcus, "Analyticity of entropy rate of hidden Markov chains," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5251–5266, Dec. 2006.

[18] G. Han and B. Marcus, "Derivatives of entropy rate in special families of hidden Markov chains," *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2642–2652, Jul. 2007.

[19] G. Han and B. Marcus, "Asymptotics of input-constrained binary symmetric channel capacity," *Ann. Appl. Probab.*, vol. 19, no. 3, pp. 1063–1091, 2009.

[20] G. Han and B. Marcus, "Asymptotics of entropy rate in special families of hidden Markov chains," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1287–1295, Mar. 2010.

[21] G. Han and B. Marcus. "Concavity of the mutual information rate for input-restricted memoryless channels at high SNR," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1534–1548, Mar. 2012.

[22] W. Hirt and J. L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 34, no. 3, pp. 380–388, May 1988.

[23] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," *Theoret. Comput. Sci.*, vol. 395, no. 2-3, pp. 203–219, 2008.

[24] P. Jacquet and W. Szpankowski, "Noisy constrained capacity for BSC channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5412–5423, Nov. 2010.

[25] A. Kavcic, "On the capacity of Markov sources over noisy channels," in *Proceedings of 2001 IEEE Global Telecommunications Conference*, vol. 5, pp. 2997–3001, Nov. 2001.

[26] H. Kobayashi and D. T. Tang, "Application of partial-response channel coding to magnetic recording systems," *IBM Journal of Research and Development*, vol. 14, no. 4, pp. 368–375, 1970.

[27] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem: History, Theory, and Applications,* Springer New York, 2013.

[28] Y. Li and G. Han, "Input-constrained erasure channels: Mutual information and capacity," in *Proceedings of the IEEE International Symposium on Information Theory*, pp. 3072-3076, Jul. 2014.

[29] Y. Li and G. Han, "Concavity of mutual information rate of finite-state channels," in *Proceedings of IEEE International Symposium on Information Theory*, pp. 2114–2118, Jul. 2013.

[30] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, 1995.

[31] B. Marcus, R. Roth, and P. Siegel, "Constrained systems and coding for recording channels," *Handbook of coding theory*, Vol. I, II. Amsterdam: North-Holland, pp. 1635–1764, 1998.

[32] M. Mushkin and I. Bar-David, "Capacity and coding for the gilbert-elliott channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277 –1290, Jun. 1989.

[33] W. Parry, "Intrinsic markov chains," *Transactions of the American Mathematical Society*, vol. 112, no. 1, pp. 55–66, 1964.

[34] H. D. Pfister, "The capacity of finite-state channels in the high-noise regime," *Entropy of hidden Markov processes and connections to dynamical systems*, London Math. Soc. Lecture Note Series. Cambridge: Cambridge Univ. Press, 2011, vol. 385, pp. 179–222.

[35] H. Pfister, J. B. Soriaga, and P. Siegel, "On the achievable information rates of finite state ISI channels," in *Proceedings of IEEE Global Telecommunications Conference*, vol. 5, pp. 2992–2996, Nov. 2001.

[36] A. Patapoutian and P. Vijay Kumar, "The (d, k) Subcode of a Linear Block Code," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1375–1382, Jul. 1992.

[37] O. Sabag, H. Permuter and N. Kashyap, "The Feedback Capacity of the $(1, \infty)$-RLL Input-Constrained Erasure Channel," http://arxiv.org/abs/1503.03359.

[38] O. Sabag, H. Permuter and N. Kashyap, "The feedback capacity of the binary symmetric channel with a no-consecutive-ones input constraint," in 50th Ann. Allerton Conf. Urbana, IL, 2015, to appear.

[39] E. Seneta, *Non-negative matrices and Markov chains*, Springer Series in Statistics. New York: Springer, 1981.

[40] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech.J.*, vol. 27, pp. 379–423, 623–656, 1948.

[41] C. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[42] S. Tatikonda and S. Mitter, "The Capacity of Channels With Feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[43] J. Taylor, *Several Complex Variables With Connections to Algebraic Geometry and Lie Groups*, Graduate Studies in Mathematics. American Mathematical Society, 2002.

[44] A. Thangaraj Dual capacity upper bounds for noisy runlength constrained channels. Submitted to ITW 2016.

[45] Yang, Shaohua and Kavcic, A. and Tatikonda, S.,"Feedback Capacity of Stationary Sources over Gaussian Intersymbol Interference Channels," *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE*,vol., no., pp.1–6, Nov. 27 2006-Dec. 1 2006.

[46] S. Verdu and T. Weissman, "The Information Lost in Erasures," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5030–5058, Nov. 2008.

[47] P. O. Vontobel and D. M. Arnold, "An upper bound on the capacity of channels with memory and constraint input," in *Proceedings of IEEE Information Theory Workshop*, pp. 147–149, Sep.2001.

[48] P. O. Vontobel, A. Kavčić, D. M. Arnold, and H. A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1887–1918, May 2008.

[49] J. K. Wolf, Invited talk on the Magnetic recording channel presented at the Twenty Sixth Ann. Allerton Conf., Urbana, IL, Sep. 1988.

[50] E. Zehavi and J. Wolf, “On runlength codes,” *IEEE Trans. Inf. Theory*, vol. 34, no. 1, pp. 45–54, Jan. 1988.