The Feedback Capacity of the Binary Erasure Channel With a No-Consecutive-Ones Input Constraint

Oron Sabag, *Student Member, IEEE*, Haim H. Permuter, *Senior Member, IEEE*, and Navin Kashyap, *Senior Member, IEEE*

***Abstract*— The input-constrained erasure channel with feedback is considered, where the binary input sequence contains no consecutive ones, i.e., it satisfies the $(1,\infty)$-RLL constraint. We derive the capacity for this setting, which can be expressed as $C_\epsilon = \max_{0\le p\le 0.5} \frac{(1-\epsilon)H_b(p)}{1+(1-\epsilon)p}$, where ϵ is the erasure probability and $H_b(\cdot)$ is the binary entropy function. Moreover, we prove that *a priori* knowledge of the erasure at the encoder does not increase the feedback capacity. The feedback capacity was calculated using an equivalent dynamic programming (DP) formulation with an optimal average-reward that is equal to the capacity. Furthermore, we obtained an optimal encoding procedure from the solution of the DP, leading to a capacity-achieving, zero-error coding scheme for our setting. DP is, thus, shown to be a tool not only for solving optimization problems, such as capacity calculation, but also for constructing optimal coding schemes. The derived capacity expression also serves as the only non-trivial upper bound known on the capacity of the input-constrained erasure channel without feedback, a problem that is still open.**

***Index Terms*— Feedback capacity, constrained coding, dynamic programming, binary erasure channel, runlength-limited (RLL) constraints.**

I. Introduction

MEMORYLESS channels have been the focus of research activity in information theory since they were introduced in 1948 by Shannon [2]. The capacity of a memoryless channel has an elegant, single-letter expression, $C = \sup_{p(x)} I(X;Y)$, and this can be calculated for a broad range of channels [3], [4]. When considering a memoryless channel with input that is constrained, the capacity is given by the maximum mutual information rate between the input and output sequences. The capacity calculation of such channels involves a calculation of the entropy rate of a Hidden Markov Model (HMM), since the transmission of a constrained sequence through a memoryless channel results in an output sequence that is described by an HMM. This makes the capacity of input-constrained memoryless channels difficult to compute [5]–[8].

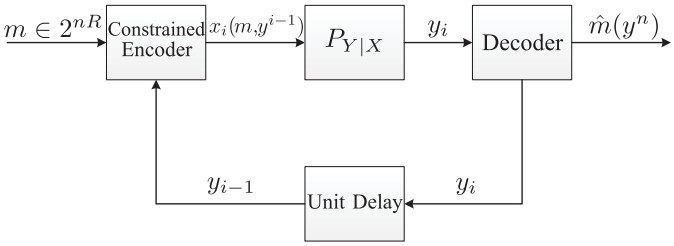

Fig. 1. System model for an input-constrained memoryless channel with perfect feedback.

Constrained coding arises naturally in many communication and recording systems [9], [10]; a common constraint that is useful in magnetic and optical recording is the (d,k)-runlength limited (RLL) constraint. A binary sequence satisfies this constraint if the number of zeros between any pair of successive ones is at least d and at most k. This constraint has also recently appeared in code designs for energy harvesting systems, where communication is used not only for information transfer but also for charging the receiver's battery [11]. In this paper, we focus on the special case of the $(1,\infty)$-RLL constraint, in which no consecutive ones are allowed.

It is well known that feedback does not increase the capacity of a memoryless channel, as shown by Shannon [12]. However, Shannon's argument does not apply to memoryless channels with constrained inputs, and special tools are required to determine the capacity of such channels with or without feedback.

We consider an $(1,\infty)$-RLL input-constrained binary erasure channel (BEC) with feedback, represented pictorially in Fig. 1, with the channel depicted in Fig. 2. Based on the message M and the previous channel outputs, y^{i-1}, the encoder chooses the input X_i, such that the input constraint is satisfied. The mechanism of the BEC is simple: each transmitted bit is transformed into an erasure symbol with probability ϵ or received successfully with its complementary probability. The decoder estimates the message $\hat{M}$ with low probability of error as a function of the output sequence Y^n. In this paper, we derive the explicit expression for the feedback capacity of the $(1,\infty)$-RLL input-constrained BEC.

The feedback capacity that is derived here also serves as an upper bound on the capacity of the $(1,\infty)$-RLL input-constrained BEC without feedback,

Manuscript received March 15, 2015; revised October 13, 2015; accepted October 13, 2015. Date of publication October 27, 2015; date of current version December 18, 2015. This work was supported by the Joint UGC–Israel Science Foundation Research Grant. O. Sabag and H. H. Permuter was supported by the European Research Council Starting Grant. This paper was presented at the 2015 Information Theory Workshop [1].

O. Sabag and H. H. Permuter are with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beersheba 8410501, Israel (e-mail: oronsa@post.bgu.ac.il; haimp@bgu.ac.il).

N. Kashyap is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: nkashyap@ece.iisc.ernet.in).

Communicated by J. Chen, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2015.2495239

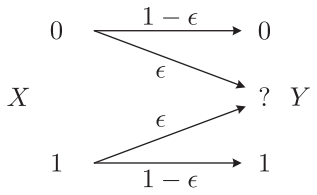

Fig. 2. Erasure channel with erasure probability ϵ.

a problem that is still open. A lower bound on the capacity of the non-feedback setting was derived in [13] by considering an input that is restricted to first-order Markov process (first-order capacity). The lower bound in [13] and our feedback capacity are presented in Fig. 3, and it can be seen that maximal gap is attained at $\epsilon = 0.71$, where the first-order capacity is approximately 0.2354 while the feedback capacity is approximately 0.2547. Based on the plots in Fig. 3, it is tempting to conjecture that feedback does not increase capacity in this input-constrained setting, however, this is not true at least for certain values of ϵ, as will be discussed in Section VIII.

The relation between feedback-capacity calculation and dynamic programming (DP) first appeared in Tatikonda's thesis [14]. Chen and Berger were the first to present in [15] a DP formulation of finite state channel with feedback that is computable. They established a recursive formula for the capacity of finite state channels with feedback where the state is a function of the output, and showed that the solution of the recursion is stationary, under mild conditions of the channels. The DP formulation of feedback capacity was also extended to several other channel models, such as channels with state that is determined by the channel input [16], Markov channels [17] and power-constrained Gaussian noise channels with memory [18]. To apply algorithms from DP, such as value and policy iteration, quantization is required, and therefore, only lower bounds were derived in the above papers.

In [19] and [20], the feedback capacities of the trapdoor and Ising channels, respectively, were found by solving their corresponding Bellman equations. The idea is that the feedback capacity is equal to the optimal reward of the DP, and therefore, it suffices to find a solution which satisfies the Bellman equation [21]. Besides reward optimality verification, the Bellman equation also establishes a mechanism for optimal policy verification, which is a significant additional benefit.

The novelty in our work is the derivation of the optimal input distribution from the Bellman equation solution. The optimal solution of the DP is then utilized to understand how the dynamic program evolves under an optimal policy. We show that converting the DP solution into channel coding terms results in a straightforward interpretation of optimal encoding procedure. This encoding procedure led us to an innovative and zero-error coding scheme for our input-constrained setting. This establishes that DP as a tool that is good not only for solving optimization problems, but also for deriving optimal coding schemes.

We also consider an input-constrained BEC where the encoder knows ahead of time if there is an erasure in the channel. Clearly, this non-causal setting is superior in terms of capacity compared to the feedback setting. We have managed to show that the capacity of this setting coincides with our feedback capacity expression, and therefore, a priori knowledge of the erasure in the channel does not increase the feedback capacity. Although this finding and the coding scheme for the feedback setting are sufficient for the feedback-capacity derivation, we argue that the capacity-achieving coding scheme is hard to construct without the DP solution.

The remainder of the paper is organized as follows. Section II includes notation and description of the problem. Section III states the main results of this paper. In Section IV, we provide a brief review of infinite-horizon DP and present the DP formulation of the feedback capacity. In Section V, the DP for the erasure channel is calculated, evaluated numerically and, finally, we prove that the Bellman equation is satisfied. In Section VI, we present the derivation of the optimal scheme from the solution of the DP. In Section VII, we derive the capacity of non-causal input-constrained BEC. Finally, the paper is concluded in Section VIII.

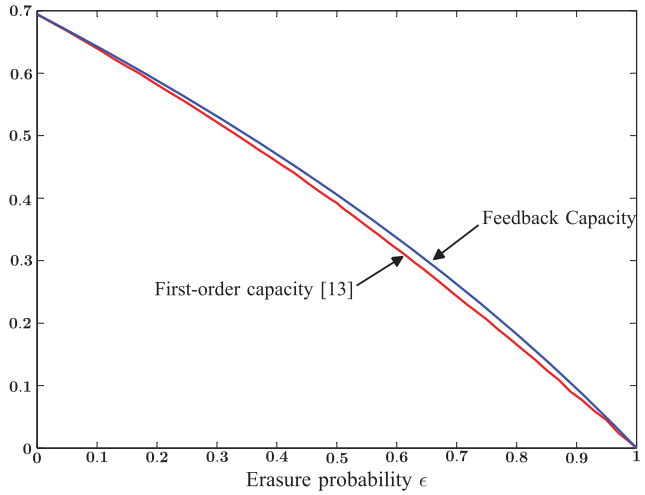

Fig. 3. Lower and upper bounds on the capacity of the input-constrained BEC without feedback.

II. Notation and Problem Definition

Throughout this paper, random variables will be denoted by upper-case letters, such as X, while realizations or specific values will be denoted by lower-case letters, e.g., x. Calligraphic letters will denote the alphabets of the random variables, e.g., $\mathcal{X}$. Let X^n denote the n-tuple $(X_1, \ldots, X_n)$. Let x^n denote vectors of n elements, i.e. $x^n = (x_1, x_2, \ldots, x_n)$, and x_i^j denote the $(j-i+1)$-tuple $(x_i, x_{i+1}, \ldots, x_j)$ when $j \geq i$, and an empty set otherwise. For any scalar $\alpha \in [0,1]$, $\bar{\alpha}$ stands for $\bar{\alpha} = 1 - \alpha$. Let $H_b(\alpha)$ denote the binary entropy for scalar $\alpha \in [0,1]$, i.e., $H_b(\alpha) = -\alpha \log_2 \alpha - \bar{\alpha} \log_2 \bar{\alpha}$. Let $H_{ter}(\alpha_1, \alpha_2, \alpha_3)$ denote the ternary entropy for scalars $\alpha_1, \alpha_2, \alpha_3 \in [0,1]$ such that $\sum_i \alpha_i = 1$, i.e., $H_{ter}(\alpha_1, \alpha_2, \alpha_3) = \sum_i -\alpha_i \log_2 \alpha_i$.

The communication setting of a memoryless channel with feedback is described in Fig. 1. A message M is drawn uniformly from the set $\{1, \ldots, 2^{nR}\}$ and made available to the encoder. The encoder at time i knows the message m and

the feedback samples y^{i-1}, and produces a binary output, $x_i \in \{0,1\}$, as a function of m and y^{i-1}. The sequence of encoder outputs, $x_1x_2x_3\ldots$, must satisfy the $(1,\infty)$-RLL input-constraint of the channel, namely, no two consecutive ones are allowed. The channel is memoryless in the sense that the output at time i, given the existing information in the system, depends only on the current input, i.e.,

$$p(y_i|x^i, y^{i-1}) = p(y_i|x_i), \quad \forall i. \tag{1}$$

We focus on the erasure channel, shown in Fig. 2. The input alphabet is $\mathcal{X} = \{0,1\}$, while the output can take values in $\mathcal{Y} = \{0,1,?\}$. The probability for erasure in the channel is ϵ and can take any value in $[0,1]$.

Definition 1: A $(n, 2^{nR}, (1,\infty))$ *code* for a constrained-input channel with feedback is defined by a set of encoding functions:

$$f_i : \{1,\ldots,2^{nR}\} \times \mathcal{Y}^{i-1} \to \mathcal{X}, \quad i = 1,\ldots,n,$$

satisfying $f_i(m, y^{i-1}) = 0$ if $f_{i-1}(m, y^{i-2}) = 1$ (the mapping $f_1(\cdot)$ is not constrained) for all (m, y^{i-1}) , and a decoding function:

$$\Psi : \mathcal{Y}^n \to \{1,\ldots,2^{nR}\}.$$

In addition, we define the non-causal $(1,\infty)$-RLL BEC. For this setting, all definitions remain the same as in the previous setting, but the encoder knows ahead of time whether there is an erasure in the channel. Formally, define θ_i as the indicator that corresponds to erasure in the channel at time i, namely, $\theta_i = 0$ if $x_i = y_i$ and $\theta_i = 1$ otherwise. The set of encoding functions for this setup is then defined as:

$$f_i : \{1,\ldots,2^{nR}\} \times \mathcal{Y}^{i-1} \times \{0,1\} \to \mathcal{X}_i, \quad i = 1,\ldots,n,$$

satisfying $f_i(m, y^{i-1}, \theta_i) = 0$ if $f_{i-1}(m, y^{i-2}, \theta_{i-1}) = 1$ for all $(m, y^{i-1}, \theta_{i-1}, \theta_i)$.

The *average probability of error* for a code is defined as $P_e^{(n)} = \Pr(M \neq \Psi(Y^n))$. A rate R is said to be $(1,\infty)$-*achievable* if there exists a sequence of $(n, 2^{nR}, (1,\infty))$ codes, such that $\lim_{n\to\infty} P_e^{(n)} = 0$. The *capacity*, C_ϵ^{fb}, defined to be the supremum over all $(1,\infty)$-achievable rates, is a function of the erasure probability ϵ. Let C_ϵ^{nc} denote the capacity for the non-causal $(1,\infty)$-RLL BEC. From operational considerations of the encoding functions for both settings, it is clear that $C_\epsilon^{\text{nc}} \geq C_\epsilon^{\text{fb}}$.

III. Main Results

The following is our main result concerning the capacity of the $(1,\infty)$-RLL constrained BEC with feedback.

Theorem 1: The capacity of the $(1,\infty)$-RLL input-constrained erasure channel with feedback is

$$C_\epsilon^{fb} = \max_{0\leq p\leq \frac{1}{2}} \frac{H_b(p)}{p + \frac{1}{1-\epsilon}}. \tag{2}$$

Furthermore, the capacity is achieved by an explicit zero-error coding scheme that is presented in Section VI-B, in Algorithm 1 and Algorithm 2.

In Fig. 4, the feedback capacity is evaluated for different values of erasure probability ϵ. As can be seen, the capacity is a decreasing function for an increasing value of ϵ. For $\epsilon = 0$, the capacity is $C_0^{\text{fb}} \approx 0.6942$, which can be represented as $\log_2\phi$, where ϕ is the golden ratio and is known as the entropy rate of a binary source with no consecutive ones. For $\epsilon = 1$, the capacity value is $C_1^{\text{fb}} = 0$, as expected.

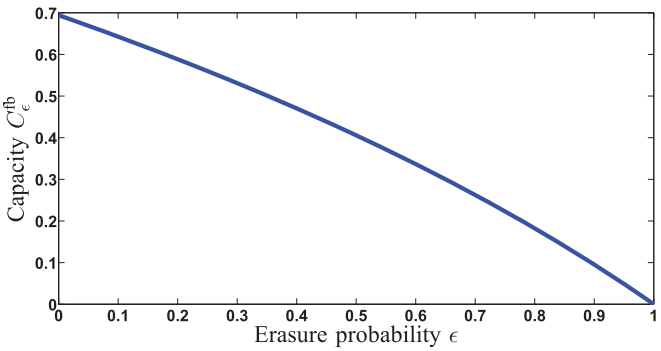

Fig. 4. The capacity C_ϵ^{fb}, as a function of ϵ, of the $(1,\infty)$-RLL input-constrained BEC with feedback.

The capacity of the non-constrained BEC can be expressed as $\max_{0\leq p\leq\frac{1}{2}} \frac{H_b(p)}{\frac{1}{1-\epsilon}} = 1-\epsilon$. Note that the only difference between this term and our capacity expression in (2) is the denominator. This fact hints that the capacity expressions of other input constraints may share a common structure.

The next theorem states that the non-causal $(1,\infty)$-RLL input-constrained BEC has the same capacity as the feedback setting.

Theorem 2: Non-causal knowledge of erasures does not increase the feedback capacity, i.e.,

$$C_\epsilon^{\text{nc}} = C_\epsilon^{\text{fb}}.$$

Next, we show the properties of the capacity expression (2).

Lemma 1: Define the function $f_\epsilon(p) = \frac{H_b(p)}{p+\frac{1}{1-\epsilon}}$, where $p \in [0,1]$. The following properties hold for $f_\epsilon(p)$:

- *The function $f_\epsilon(p)$ is concave on $[0,1]$, for any $\epsilon \geq 0$.*
- *The function $f_\epsilon(p)$ has only one maximum in $[0,1]$, which is the only real solution of the equation $p^{\frac{1}{1-\epsilon}} = (1-p)^{1+\frac{1}{1-\epsilon}}$. This maximum lies in $[0,\frac{1}{2}]$.*
- *Denote by p_ϵ the argument that achieves the maximum of $f_\epsilon(p)$. The capacity can also be expressed by*

$$C_\epsilon^{fb} = \frac{-\log_2(p_\epsilon)}{1+\frac{1}{1-\epsilon}}.$$

The proof of Lemma 1 is presented in Appendix B.

IV. Feedback Capacity and Dynamic Programming

The directed information was introduced by Massey in [22] as $I(X^n \to Y^n) = \sum_{i=1}^n I(X^i; Y_i|Y^{i-1})$. Massey showed that the maximum normalized directed information upper bounds the capacity of channels with feedback, and subsequently, it was proved that this expression indeed characterizes the feedback capacity for a broad class of channels [17], [19], [23]–[26]. The next theorem provides a multi-letter expression using the directed information for the case of memoryless channels with an $(1,\infty)$-RLL input constraint.

Theorem 3: The capacity of an $(1,\infty)$-RLL input-constrained memoryless channel with feedback can be written as:

$$C^{fb} = \sup \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} I(X_t; Y_t|Y^{t-1}), \tag{3}$$

where the supremum is taken with respect to $\{p(x_t|x_{t-1}, y^{t-1}) : p(x_t = 1|x_{t-1} = 1, y^{t-1}) = 0\}_{t\geq 1}$.

The proof of Theorem 3 appears in Appendix A. Having written the capacity of the input constrained channel with feedback as (3), we proceed to show that calculating the capacity can be formulated as an average-reward DP.

A. Average-Reward Dynamic Programs

Each DP is defined by the tuple $(\mathcal{Z}, \mathcal{U}, \mathcal{W}, F, P_Z, P_w, g)$. We consider a discrete-time dynamic system evolving according to:

$$z_t = F(z_{t-1}, u_t, w_t), \quad t = 1, 2, \dots$$

Each state, z_t, takes values in a Borel space $\mathcal{Z}$, each action, u_t, takes values in a compact subset $\mathcal{U}$ of a Borel space, and each disturbance, w_t, takes values in a measurable space $\mathcal{W}$. The initial state, z_0, is drawn from the distribution P_Z, and the disturbance, w_t, is drawn from $P_w(\cdot|z_{t-1}, u_t)$. The history, $h_t = (z_0, w_1^{t-1})$, summarizes all the information available to the controller at time t, prior to the selection of the tth action. The controller at time t chooses the action, u_t, by a function μ_t that maps histories to actions, i.e., $u_t = \mu_t(h_t)$. The collection of these functions is called a policy and is denoted as $\pi = \{\mu_1, \mu_2, \dots\}$. Note that given a policy, π, and the history, h_t, one can compute the actions vector, u^t, and the past states of the system, $z_1, z_2, \dots, z_{t-1}$.[1]

Our objective is to maximize the average reward given a bounded reward function $g : \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$. The average reward for a given policy π is given by:

$$\rho_\pi = \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_\pi \left[\sum_{t=1}^{N} g(Z_{t-1}, \mu_t(h_t)) \right],$$

where the subscript π indicates that actions u_t are generated by the policy π. The *optimal average reward* is defined as

$$\rho^* = \sup_\pi \rho_\pi .$$

B. Formulation of the Feedback Capacity as DP

The state of the dynamic programming, z_{t-1}, is defined as the conditional probability vector whose components are the elements $\beta_{t-1}(x_{t-1}) = p(x_{t-1}|y^{t-1})$, for $x_{t-1} \in \mathcal{X}$. The action space, $\mathcal{U}$, is the set of stochastic matrices, $p(x_t|x_{t-1})$, satisfying the input constraint. For a given policy and an initial state, the encoder at time t (prior to the selection of the channel input) can calculate the state, z_{t-1}, since the tuple y^{t-1} is available from the feedback. The disturbance is taken to be the channel output, $w_t = y_t$, and the reward gained at time t is chosen as $I(Y_t; X_t|y^{t-1})$. The formulation is summarized in Table I.

[1]Further details on the DP setting can be found at [27, Sec. 2.1], as our DP formulation follows their definitions.

TABLE I
FORMULATION OF CAPACITY AS DP

Input-constrained capacity	Dynamic programming
$p(x_{t-1}\|y^{t-1})$	z_{t-1} - state at time t
Constrained $p(x_t\|x_{t-1})$	u_t - action at time t
4 y_t	w_t - disturbance at time t
Equation (4)	$z_t = F(z_{t-1}, u_t, w_t)$ - system
$I(Y_t; X_t\|y^{t-1})$	$g(z_{t-1}, u_t)$ - reward at time t

1) Existence of System: We need to show that for a given policy, $\pi = \{\mu_1, \mu_2, \dots\}$, the state z_t can be calculated from the tuple (z_{t-1}, u_t, y_t). Consider the next chain of equalities for some $x_t \in \mathcal{X}$,

$$\begin{aligned}
&\beta_t(x_t) \\
&= p(x_t|y^t) \\
&= \sum_{x_{t-1}} p(x_t, x_{t-1}|y^t) \\
&= \frac{\sum_{x_{t-1}} p(x_t, x_{t-1}, y_t|y^{t-1})}{p(y_t|y^{t-1})} \\
&= \frac{\sum_{x_{t-1}} p(x_{t-1}|y^{t-1}) p(x_t|x_{t-1}, y^{t-1}) p(y_t|y^{t-1}, x_t, x_{t-1})}{\sum_{x'_t, x_{t-1}} p(y_t, x'_t, x_{t-1}|y^{t-1})} \\
&\overset{(a)}{=} \frac{\sum_{x_{t-1}} p(x_{t-1}|y^{t-1}) p(x_t|x_{t-1}, y^{t-1}) p(y_t|x_t)}{\sum_{x'_t, x_{t-1}} p(x_{t-1}|y^{t-1}) p(x'_t|x_{t-1}, y^{t-1}) p(y_t|x'_t)} \\
&= \frac{\sum_{x_{t-1}} \beta_{t-1}(x_{t-1}) u_t(x_t, x_{t-1}) p(y_t|x_t)}{\sum_{x'_t, x_{t-1}} \beta_{t-1}(x_{t-1}) u_t(x'_t, x_{t-1}) p(y_t|x'_t)},
\end{aligned} \tag{4}$$

where (a) follows from the memoryless property (1). Therefore, there exists a function F such that $\beta_t(x'_t) = F(z_{t-1}, u_t(x'_t, x_{t-1}), w_t)$, for all $x'_t \in \mathcal{X}$.

2) Disturbance: Let us show that the disturbance distribution depends on the current state and action only, with no dependence on past information, i.e., $p(w_t|w^{t-1}, z^{t-1}, u^t) = p(w_t|z_{t-1}, u_t)$. Consider,

$$\begin{aligned}
&p(w_t|w^{t-1}, z^{t-1}, u^t) \\
&\quad = p(y_t|y^{t-1}, \beta^{t-1}, u^t) \\
&\quad = \sum_{x_t, x_{t-1}} p(y_t, x_t, x_{t-1}|y^{t-1}, \beta^{t-1}, u^t) \\
&\quad \overset{(a)}{=} \sum_{x_t, x_{t-1}} p(x_{t-1}|y^{t-1}, \beta^{t-1}, u^t)) \\
&\qquad \times p(x_t|x_{t-1}, y^{t-1}, \beta^{t-1}, u^t) p(y_t|x_t) \\
&\quad \overset{(b)}{=} \sum_{x_t, x_{t-1}} p(x_{t-1}|\beta_{t-1}, u_t\, p(x_t|x_{t-1}, \beta_{t-1}, u_t) p(y_t|x_t) \\
&\quad = \sum_{x_t, x_{t-1}} p(y_t, x_t, x_{t-1}|\beta_{t-1}, u_t) \\
&\quad = p(y_t|\beta_{t-1}, u_t) \\
&\quad = p(w_t|z_{t-1}, u_t),
\end{aligned}$$

where (a) follows from the fact that the channel is memoryless, and (b) follows from the fact that the value of $p(x_{t-1}|y^{t-1}, \beta^{t-1}, u^t)$ is determined by $\beta_{t-1}(x_{t-1})$, and the fact that x_t depends only on the triplet $(x_{t-1}, \beta_{t-1}, u_t)$.

3) Reward: We need to show that the reward, $I(Y_t; X_t|y^{t-1})$, that is achieved at time t is a function of the current state, z_{t-1}, and of the chosen action u_t. Note that the term of the reward depends on the conditional distribution $p(y_t, x_t|y^{t-1})$ only.

Let us show that the reward achieved at time t depends on the current state, action and the channel characterization,

$$\begin{aligned} p(y_t, x_t|y^{t-1}) &= \sum_{x_{t-1}} p(y_t, x_t, x_{t-1}|y^{t-1}) \\ &\overset{(a)}{=} \sum_{x_{t-1}} p(x_{t-1}|y^{t-1}) p(x_t|x_{t-1}, y^{t-1}) p(y_t|x_t) \\ &= \sum_{x_{t-1}} \beta_{t-1}(x_{t-1}) u_t(x_t, x_{t-1}) p(y_t|x_t), \end{aligned}$$

where (a) follows from the chain rule and the memoryless property (1). Recall that the term $p(y_t|x_t)$ is given by the channel characterization, and thus, the reward depends on the DP state, z_{t-1} (which contains all elements of the vector β_{t-1}), and the chosen action, u_t. Therefore, the reward at time t can be written as:

$$g(z_{t-1}, u_t) = I(Y_t; X_t|z_{t-1}, u_t).$$

It then follows that the optimal average reward of the defined DP is:

$$\rho^* = \sup_{\pi} \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} I_\pi(Y_t; X_t|Y^{t-1}),$$

where the subscript π indicates that the mutual information is calculated with respect to the policy π. As ρ^* is equal to the feedback capacity of the input-constrained memoryless channel in Theorem 3, one can conclude that the optimal average reward for the above DP formulation is equal to the capacity.

V. Solution for the Erasure Channel

This section is organized as follows: Section V-A formulates feedback capacity of the BEC as DP using the notation from Section IV-B. In Section V-B, we evaluate a numerical solution using the value iteration algorithm, and finally, in Section V-C, we present the Bellman equation and its solution for the BEC. The solution of the Bellman equation concludes the derivation of the feedback capacity expression in Theorem. 1.

TABLE II

The Conditional Distribution $p(x_t, x_{t-1}, y_t|z_{t-1}, u_t)$

x_t	x_{t-1}	$y_t = 0$	$y_t = ?$	$y_t = 1$
0	0	$z_{t-1}u_t(1,1)\bar{\epsilon}$	$z_{t-1}u_t(1,1)\epsilon$	0
1	0	0	$z_{t-1}u_t(1,2)\epsilon$	$z_{t-1}u_t(1,2)\bar{\epsilon}$
0	1	$(1-z_{t-1})\bar{\epsilon}$	$(1-z_{t-1})\epsilon$	0

A. Formulation of the Erasure Channel as DP

The state of the DP at time $t-1$, z_{t-1}, is the probability vector $[p(x_{t-1} = 0|y^{t-1}), p(x_{t-1} = 1|y^{t-1})]$. With some abuse of notation, we refer from now on to $z_{t-1} \triangleq p(x_{t-1} = 0|y^{t-1})$ as the first component of the vector, which also determines the second component, since they sum to 1. Each action, u_t, is a constrained 2×2 stochastic matrix, $p(x_t|x_{t-1})$, of the form:

$$u_t = \begin{bmatrix} p(x_t = 0|x_{t-1} = 0) & p(x_t = 1|x_{t-1} = 0) \\ 1 & 0 \end{bmatrix}.$$

The disturbance w_t is the channel output, y_t, and can take values in $\{0, 1, ?\}$. With the above definitions, and the system equation that is given in (4), the state of the DP is calculated at the bottom of this page in (5); Substituting specific values of y_t into (5) gives the explicit system equation:

$$z_t = \begin{cases} 1 & \text{if } w_t = 0, \\ 1 - z_{t-1} + z_{t-1}u_t(1,1) & \text{if } w_t = ?, \\ 0 & \text{if } w_t = 1. \end{cases} \tag{6}$$

At this point, to simplify notations we note that $1 - z_{t-1} + z_{t-1}u_t(1,1)$ can be written as $1 - z_{t-1}u_t(1,2)$. We denote $\delta_t \triangleq z_{t-1}u_t(1,2)$, and this implies the constraint $0 \le \delta_t \le z_{t-1}$, since u_t, by definition, must be a stochastic matrix. Furthermore, when investigating the relation of DP and encoding procedures, u_t has to be recovered from δ_t, given z_{t-1}. This calculation is trivial for $z_{t-1} \neq 0$, while for $z_{t-1} = 0$, we note that $u_t(1,2)$ has no effect on the DP, and therefore, $u_t(1,2)$ can be fixed to zero.

To calculate the reward, the conditional distribution $p(x_t, x_{t-1}, y_t|z_{t-1}, u_t)$ is described in Table II, and it follows that the reward is:

$$\begin{aligned} g(z_{t-1}, u_t) &= I(Y_t; X_t|z_{t-1}, u_t) \\ &= H(Y_t|z_{t-1}, u_t) - H(Y_t|X_t, z_{t-1}, u_t) \\ &\overset{(a)}{=} H_{ter}((1-\delta_t)\bar{\epsilon}, \epsilon, \delta_t\bar{\epsilon}) - H_b(\epsilon) \\ &\overset{(b)}{=} H_b(\epsilon) + \bar{\epsilon}H_b(\delta_t) - H_b(\epsilon) \\ &= \bar{\epsilon}H_b(\delta_t), \end{aligned} \tag{7}$$

where (a) follows from the marginal distribution $p(y_t|z_{t-1}, u_t)$ in Table II and the definition of δ_t,

$$\begin{aligned} z_t &= p(x_t = 0|y^t) \\ &= \frac{\sum_{x_{t-1}} \beta_{t-1}(x_{t-1}) u_t(x_t = 0, x_{t-1}) p(y_t|x_t = 0)}{\sum_{x'_t, x_{t-1}} \beta_{t-1}(x_{t-1}) u_t(x'_t, x_{t-1}) p(y_t|x'_t)} \\ &= \frac{z_{t-1}u_t(1,1)p(y_t|x_t = 0) + (1 - z_{t-1})u_t(2,1)p(y_t|x_t = 0)}{z_{t-1}[u_t(1,1)p(y_t|x_t = 0) + u_t(1,2)p(y_t|x_t = 1)] + (1 - z_{t-1})u_t(2,1)p(y_t|x_t = 0)} \end{aligned} \tag{5}$$

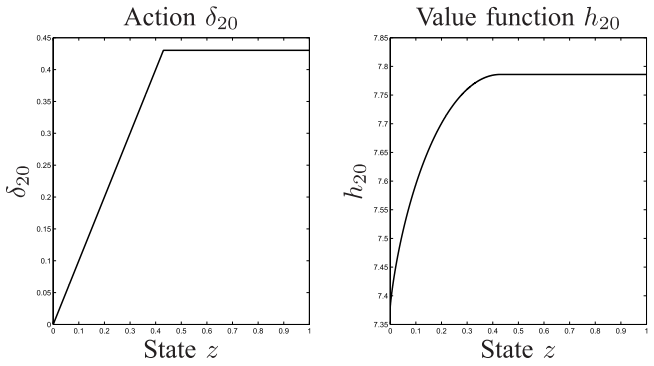

Fig. 5. Value iteration evaluation for the erasure channel with $\epsilon = 0.5$. The algorithm was implemented with 20 iterations and quantization of 5000 points for both action and state.

while (b) follows from an easily verifiable identity: $H_{ter}(a\bar{b}, \bar{a}\bar{b}, b) = H_b(b) + \bar{b}H_b(a)$, for all $a, b \in [0, 1]$.

To apply the value iteration in the next subsection, it is convenient to define the operator of the DP:

$$(Th)(z) = \sup_{u \in \mathcal{U}} g(z, u) + \int P_W(dw|z, u)h(F(z, u, w)), \tag{8}$$

for all functions $h : \mathcal{Z} \to \mathbb{R}$. Note that the disturbance W takes values in the finite set $\mathcal{Y}$, therefore, one can replace the Lebesgue integration with a summation over the conditioned disturbance distribution.

By substituting the reward from (7), the marginal conditioned distribution from Table II, and the system function from (6), we can calculate (8) as follows:

$$\begin{aligned}&(Th_\epsilon)(z)\\ &= \sup_{u \in \mathcal{U}} g(z, u) + \int P_W(dw|z, u)h_\epsilon(F(z, u, w))\\ &= \sup_{0 \le \delta \le z} \bar{\epsilon}H_b(\delta) + (1-\delta)\bar{\epsilon}h_\epsilon(1) + \epsilon h_\epsilon(1-\delta) + \delta\bar{\epsilon}h_\epsilon(0),\end{aligned} \tag{9}$$

for all $h_\epsilon : [0, 1] \to \mathbb{R}$, parameterized by ϵ.

B. Numerical Evaluation

Now that we have the DP formulation for our problem, we can apply the value iteration algorithm to estimate the optimal average reward. The value iteration algorithm is simply applying the DP operator from (9) successively, and it has the form $h_k(z) = (Th_{k-1})(z)$ with $h_0(z) = 0$. The state of the DP and the values in the action matrices are continuous, which cannot be implemented by a finite-precision computer. To this end, a quantization of 5000 points in the unit interval for both z_t and δ_t was performed, and the results after 20 iterations are presented in Fig. 5 for erasure probability $\epsilon = 0.5$.

We also simulated the system with the estimated optimal action δ_{20}. The initial state, z_0, was chosen to be zero and the action was taken according to δ_{20} which led to a gained reward. The disturbance was generated randomly according to the induced distribution from Table II. Having in hand the current state, action and disturbance, the new state was calculated and the process was repeated 10^6 times. This simulation led to an approximate average reward of 0.4056 and the histogram of the states is shown in Fig. 6. The significant importance of a discrete histogram will be discussed in Section VI, where it is explained how the DP simulation leads us to derive an optimal coding scheme for our channel setting.

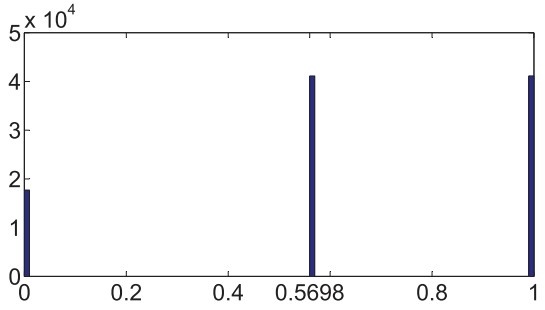

Fig. 6. Histogram of system states after 10^6 runs.

C. The Bellman Equation

In dynamic programming, the Bellman equation suggests a sufficient condition for average reward optimality. This equation establishes a mechanism for verifying that a given average reward is optimal. The next result encapsulates the Bellman equation and can be found in [27].

Theorem 4 [27, Th. 6.2]: If $\rho \in \mathbb{R}$ and a bounded function $h : \mathcal{Z} \to \mathbb{R}$ satisfies for all $z \in \mathcal{Z}$:

$$\rho + h(z) = \sup_{u \in \mathcal{U}} g(z, u) + \int P_W(dw|z, u)h(F(z, u, w)), \tag{10}$$

then $\rho^* = \rho$. Furthermore, if there is a function $\mu : \mathcal{Z} \to \mathcal{U}$ such that $\mu(z)$ attains the supremum for each z, then $\rho^* = \rho_\pi$ for $\pi = \{\mu_0, \mu_1, \dots\}$ with $\mu_t(h_t) = \mu(z_{t-1})$ for each t.

This result is a direct consequence of [27, Th. 6.2]; specifically, the triplet $(\rho, h(\cdot), \mu(\cdot))$ is a canonical triplet by Theorem 6.2 since it satisfies (10). Now, as a canonical triplet defines the N-stage optimal reward and policy under terminal cost $h(\cdot)$, for all N, it can be concluded that a canonical triplet also defines the optimal reward and policy in the infinite horizon regime, since in this case the bounded terminal cost has a negligible affect.

As our DP formulation is parameterized with the parameter ϵ, it is clear from the context that one should solve the Bellman equation for all values of ϵ. Moreover, note that the right hand side of (10) coincides with the DP operator definition in (8).

Let us denote two constants $\tilde{\rho}_\epsilon$ and p_ϵ,

$$\begin{aligned}\tilde{\rho}_\epsilon &= \max_{0 \le p \le \frac{1}{2}} \frac{H_b(p)}{p + \frac{1}{1-\epsilon}},\\ p_\epsilon &= \arg\max_{0 \le p \le \frac{1}{2}} \frac{H_b(p)}{p + \frac{1}{1-\epsilon}},\end{aligned} \tag{11}$$

and a bounded function,

$$\tilde{h}_\epsilon(z) = \begin{cases} \bar{\epsilon}H_b(z) - z\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} & \text{if } 0 \le z < p_\epsilon \\ \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} & \text{if } p_\epsilon \le z \le 1. \end{cases} \tag{12}$$

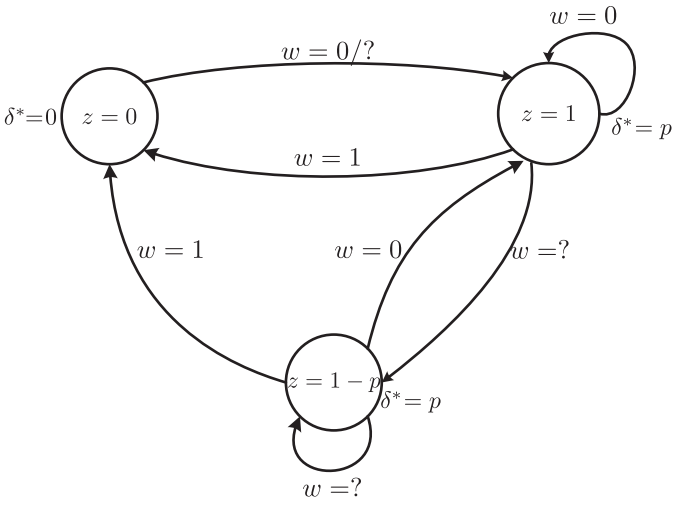

Fig. 7. State diagram of the DP for the input-constrained BEC under an optimal policy.

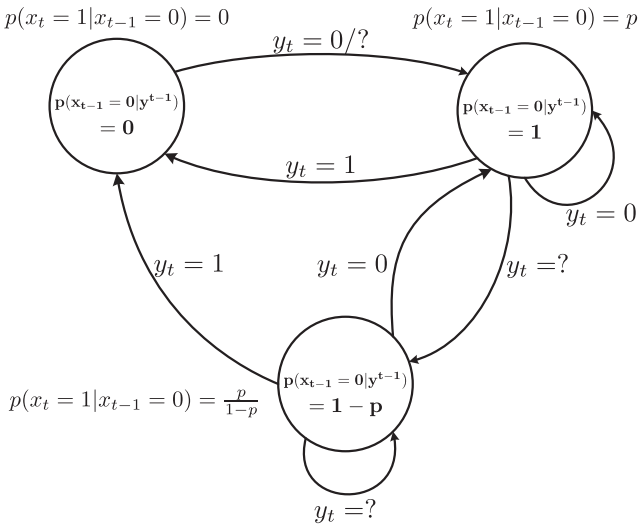

Fig. 8. Optimal encoding procedure for the input-constrained BEC. This encoding procedure was achieved from Fig. 7 by converting states, actions and disturbances into their corresponding channel coding terms.

We proceed to show the DP solution by explicitly solving (10) for our problem.

Theorem 5: The constant $\tilde{\rho}_\epsilon$ and the function $\tilde{h}_\epsilon(z)$ given in (11) *and* (12)*, respectively, satisfy the Bellman equation* (10) *for each ϵ. Therefore, $\tilde{\rho}_\epsilon$ is the optimal average reward.*

As the optimal average reward is equal to the capacity expression (2), Theorem 5 concludes the proof for the first part of Theorem 1. The proof of Theorem 5 is presented in Appendix C.

VI. Derivation of the Capacity-Achieving Coding Scheme From the DP Solution

In this section, we derive the optimal coding scheme using the DP solution and finally show that this leads to a capacity-achieving coding scheme. The method comprises recovering the optimal constrained input distributions $\{p(x_t|x_{t-1}, y^{t-1})\}_{t\geq 1}$ from the solution of the DP.

A. Relation of the Coding Scheme to DP Results

The histogram for $\epsilon = 0.5$, in Fig. 6, shows that under an optimal policy, δ^*, the system evolves between three steady states. Moreover, the solution of the Bellman equation indicates that there exists an optimal stationary policy, and therefore, we look at the stationary phase of the DP. The states, z, take values in the finite set $\{0, 1-p, 1\}$, with $p \triangleq p_\epsilon$ (Eq. (11)); the subscript ϵ is omitted for convenience, but all details are discussed for a fixed $\epsilon \in [0, 1]$ and its corresponding p_ϵ. For each state, the optimal policy, δ^*, is known from the Bellman equation and arrows can be drawn between the states as a function of the disturbance. The state diagram for our DP is presented in Fig. 7.

Converting the state diagram in Fig. 7 into channel coding terms, using the formulation described in Table I, results in an encoding procedure as described in Fig. 8. Specifically, the states, $p(x_{t-1} = 0|y^{t-1})$, take values from $\{0, 1-p, 1\}$. Each state has its corresponding action, $p(x_t = 1|x_{t-1} = 0)$, and the encoding procedure evolves as a function of the output y_t. Recall that $p(x_t = 0|x_{t-1} = 1) = 1$, and therefore, the action $p(x_t = 1|x_{t-1} = 0)$ is sufficient to determine the transfer matrix from X_{t-1} and X_t.

Let us explain how the encoding procedure evolves. We refer to the state $p(x_{t-1} = 0|y^{t-1}) = 1$ as the *ground state*, since this indicates that '0' was received at the decoder and, therefore, the encoder is allowed to transmit any input to the channel. For the ground state, the next transmitted bit is distributed according to Ber(p) and it is shown to be the optimal action.

Upon receiving $y_t = 0$ at the decoder, the system remains at the ground state and the encoding procedure starts over again. When the output is $y_t = 1$, the system moves to the state $p(x_{t-1} = 0|y^{t-1}) = 0$. At this state, since the last input was necessarily '1', the encoder is forced to transmit '0'. Therefore, the decoder knows that '0' is the only legitimate input, and the system returns to the ground state regardless of whether the input was erased or not.

The remaining scenario to examine begins at the ground state and is followed by $y_t =?$. The optimal action at the lower state, $p(x_{t-1} = 0|y^{t-1}) = 1-p$, suggests that if '0' is erased, the new transmitted bit should be distributed according to Ber$\left(\frac{p}{1-p}\right)$. The term $\frac{p}{1-p}$ is in the unit interval, since $p \leq \frac{1}{2}$. Additionally, the input constraint implies that if '1' was erased then '0' should be transmitted. Upon consecutive erasures, the encoder continues to transmit bits according to this policy. When an output is not an erasure, the system returns to the ground state, and this might take one or two time instances, depending on whether the (unerased) output bit is '0' or '1'.

The main challenge is to understand how this encoding procedure can be interpreted as transmitting a message by the encoder. Let the messages be points in the unit interval, i.e., messages take values in the set $\mathcal{M} \triangleq \{\frac{k}{2^{nR}}\}_{k=0}^{2^{nR}-1}$. At each time instance, the unit interval contains sub-intervals with labels that can be '0' or '1', and the input to the channel is simply

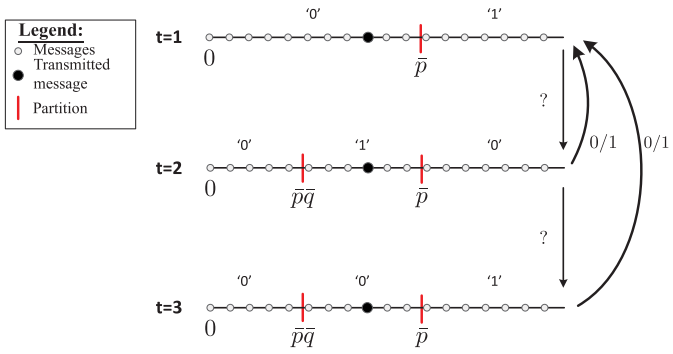

Fig. 9. Example for transmitting the black-dot message using the encoding procedure in Fig. 8 for 3 time instances. The initial partition at the ground state is according to p, and the encoder transmits '0' since the black-dot message falls within $[0, \bar{p})$. Upon a successful transmission, the encoder moves back to the ground state and a new procedure begins. In case of erasure, we move to $t = 2$, and the interval that was labelled '0' is partitioned according to $q = \frac{p}{1-p}$. The input constraint is preserved since the interval $[\bar{p}, 1)$, that was labelled '1', is now flipped to '0'. The encoder transmits '1' since the message falls within $[\bar{p}\bar{q}, \bar{p})$. In case of another erasure, a partition of q should be performed for the intervals that are labelled '0'. These intervals are $[0, \bar{p}\bar{q})$ and $[\bar{p}, 1)$, which are sum up to $1 - p$. Since $q = \frac{p}{1-p}$, we simply change the label of $[\bar{p}, 1)$ (which has length of p) to '1', and the label of $[0, \bar{p}\bar{q})$ remains '0'. The input-constraint is preserved since $[\bar{p}\bar{q}, \bar{p})$ is re-labelled as '0'. Upon another erasures, the labelling will be exchanged between the ones presented in $t = 2$ and $t = 3$ until a successful transmission. Note that the labelling at $t = 1$ and $t = 3$ are essentially the same.

the label of the sub-interval containing the message. Such an association of messages into a specified interval has been done before in [28]–[31].

The partition into sub-intervals will be according to parameters p and $q \triangleq \frac{p}{1-p}$, as described in Fig. 8. When performing a partition at the ground state, the lower interval is labelled '0' while the upper interval is labelled '1'. Before providing the precise encoding algorithm, it will be convenient to understand the labelling process in the example described in Fig. 9.

As can be seen in Fig. 9, all the proposed partitions in Fig. 8 can be encapsulated into two possible labellings. We denote the labelling at $t = 1$ as L_1, and the labelling at $t = 2$ as L_2. The initial labelling at the ground state is chosen as L_1, and upon erasure, the current labelling will be replaced with the other labelling. Note that changing the labelling L_i with L_j for $i \neq j$ preserves the input constraint and can be done simply by exchanging the labels of $[\bar{p}\bar{q}, \bar{p})$ and $[\bar{p}, 1)$, while the label of $[0, \bar{p}\bar{q})$ remains '0'.

To summarize at this point, at each time instant, we have two possible labellings (which depend on the value of ϵ) of the unit interval which define uniquely the mapping from messages to the channel input. The current labelling is determined only by the output tuple, y^{t-1}, and therefore, the decoder and encoder both agree on the latter.

B. Capacity-Achieving Coding Scheme

At time instance $t - 1$, the *set of possible messages* is defined as $\mathcal{M}_{t-1} = \{m \in \mathcal{M} : p(m|y^{t-1}) > 0\}$, with $\mathcal{M}_0 = \mathcal{M}$. The conditional distribution $p(m|y^{t-1})$ is calculated using Bayes' rule, using the fact that the encoding procedure and both labellings are revealed to all parties before transmission begins. Note that the set of possible messages can also be calculated at the encoder, since the output tuple, y^{t-1}, is available from the feedback.

Algorithm 1 Encoding Procedure

while Set of possible messages contains more than one message **do**
  Label the unit interval according to L_1.
  Transmit the label of the sub-interval containing the message.
  while Received symbol is an erasure **do**
    Exchange the labels of $[\bar{p}\bar{q}, \bar{p})$ and $[\bar{p}, 1)$.
    Transmit the label of the sub-interval containing the message.
  end while
  if Received symbol is '0' **then**
    Denote the messages within sub-intervals which are labelled '0' as the set of possible messages.
  else
    Denote the messages within sub-intervals which are labelled '1' as the set of possible messages
    Transmit '0'.
  end if
  Expand the set of possible messages to the unit interval.
end while

Algorithm 2 Decoding Procedure

while Set of possible messages contains more than one message **do**
  Label the unit interval according to L_1.
  while Received symbol is an erasure **do**
    Exchange the labels of $[\bar{p}\bar{q}, \bar{p})$ and $[\bar{p}, 1)$.
  end while
  if Received symbol is '0' **then**
    Denote the messages within sub-intervals which are labelled '0' as the set of possible messages.
  else
    Denote the messages within sub-intervals which are labelled '1' as the set of possible messages.
    Ignore the next received symbol.
  end if
  Expand the set of possible messages to the unit interval.
end while

Any received symbol at the decoder might reduce the set of potential messages, and a *successful transmission* is defined as a transmission where the size of the set of possible messages is changed, namely, $|\mathcal{M}_t| < |\mathcal{M}_{t-1}|$. Specifically, a successful transmission can occur in one of two scenarios; the first is $y_t = 1$, and the second is where $y_t = 0$ and $y_{t-1} \neq 1$. Upon a successful transmission, the set of possible messages is calculated and expanded uniformly to the unit interval. To be precise, the messages in the set $\mathcal{M}_t$ take values in $\{\frac{k}{|\mathcal{M}_t|}\}_{k=0}^{|\mathcal{M}_t|-1}$. This transmission procedure continues repeatedly until the set of possible messages contains one message. The detailed encoding and decoding procedures are described in Algorithms 1 and 2.

Rate Analysis: The main feature of this coding scheme is that the length of the sub-interval that is labelled by '1' is p. This property is recorded as Lemma 2.

Lemma 2: At any step of the message transmission process, the lengths of the sub-intervals that are labelled by '1' sum up to p.

Proof: Throughout transmission, there are two possible labellings; for L_1, the interval $[\bar{p}, 1)$ that is labelled '1' has length of p, while for L_2, the interval $[\bar{p}\bar{q}, \bar{p})$ has length of $\bar{p}q = p$. ∎

From Lemma 2, we note that the encoder transmits '1' if message falls within sub-interval that has length of p. However, the messages are discrete points and a partition might fall between two messages. This implies that the transmitted bit is distributed as $\text{Ber}(p + e_i)$, where e_i is a correction factor. In Appendix D, it is shown that the correction factor has a negligible effect on the rate of the coding scheme. To simplify the derivations here, with some loss of accuracy, we say that each transmitted bit is distributed according to $\text{Ber}(p)$.

In the next lemma, we show that each successful transmission reduces the expected number of bits that is required to describe the set of possible messages by $H_b(p)$.

Lemma 3: With each successful transmission, the expected number of bits that describe the set of possible messages is reduced by $H_b(p)$.

Proof: Assume that the set of possible messages is of size k; upon a successful transmission, if '0' is received then the new set of possible messages has size $\bar{p}k$, and if '1' is received then its new size is pk. The expected number of bits that is required to describe the new set of possible messages is $\bar{p}\log_2(\bar{p}k) + p\log_2(pk) = \log_2 k - H_b(p)$. ∎

The next step is to calculate the expected number of channel uses for a *complete procedure*. We define a complete procedure to consist of all transmissions by the encoder starting at some time t at which it is in the ground state, and ending at the first time $t' > t$ at which it returns to the ground state. In other words, a procedure is completed when a '0' or '1' is received at the decoder, including one extra channel use in the case when a '1' has been received and has to be followed by '0'.

Let N be a random variable corresponding to the number of channel uses within a complete procedure. The expected value of N will be calculated by the law of total expectation. Define an indicator function

$$\theta = \begin{cases} 0 & \text{if the received bit is '0'} \\ 1 & \text{if the received bit is '1',} \end{cases}$$

and consider,

$$\begin{aligned} \mathbb{E}[N] &\overset{(a)}{=} \mathbb{E}[\mathbb{E}[N|\theta]] \\ &\overset{(b)}{=} \mathbb{E}\Big[\frac{1}{1-\epsilon} + \theta\Big] \\ &\overset{(c)}{=} \frac{1}{1-\epsilon} + p, \end{aligned}$$

where (a) follows from the law of total expectation, (b) follows from the fact that channel is memoryless and, therefore, $\frac{1}{1-\epsilon}$ is the expected value of time to receive a symbol which is not an erasure, and (c) follows from $\mathbb{E}[\theta] = \Pr(\theta = 1)$.

Finally, we prove the second part of Theorem 1, specifically, the rate of this coding scheme can be arbitrary close to the capacity expression, C_ϵ^{fb}.

Proof: It follows from the law of large numbers that the rate of our coding scheme can be arbitrarily close to the expected number of received bits within a complete procedure divided by the expected number of channel uses within a complete procedure. In Lemma 3, we showed that within a successful transmission, the expected number of received bits is $H_b(p)$. Moreover, the expected number of channel uses within a complete procedure is $\mathbb{E}[N] = \frac{1}{1-\epsilon} + p$. Therefore, the rate of the code can be arbitrarily close to $R = \frac{H_b(p)}{p + \frac{1}{1-\epsilon}}$. ∎

The above proof and Theorem 5 conclude the proof of our main result Theorem 1.

VII. Non-Causal Capacity

In this section, we prove Theorem 2 by showing that $C_\epsilon^{\text{nc}} = \max_{0 \le p \le \frac{1}{2}} \frac{H_b(p)}{p + \frac{1}{1-\epsilon}}$. Operational considerations of non-causal and feedback capacities reveal the trivial inequality $C_\epsilon^{\text{nc}} \ge C_\epsilon^{\text{fb}}$. Furthermore, we derive in this section an upper-bound on C_ϵ^{nc}, which is equal to C_ϵ^{fb}, and this concludes the proof of Theorem 2 with $C_\epsilon^{\text{nc}} = C_\epsilon^{\text{fb}}$.

The next lemma shows that it is sufficient to consider encoders which transmit '0' if erasure occurs, i.e., $x_i = 0$ if $\theta_i = 1$. The intuition behind this lemma is that replacing erased ones with zeros does not affect the output sequence, while the input-constraint is not violated.

Lemma 4: For any $(1, 2^{nR}, (1, \infty))$ code $\mathcal{C}$ with probability of error $P_e^{(n)}$, there exists a $(n, 2^{nR}, (1, \infty))$ code $\mathcal{C}'$ with probability of error $P_e^{(n)}$, satisfying

$$f_i(m, y^{i-1}, \theta_i = 1) = 0, \quad i = 1, \dots, n \quad \forall (m, y^{i-1}).$$

Proof: For any $(1, 2^{nR}, (1, \infty))$ code $\mathcal{C}$ consisting of encoding functions, $\{f_i(\cdot)\}_{i=1}^n$, and a decoding function $\Psi(\cdot)$ with probability of error $P_e^{(n)}$, define a new sequence of encoding functions as follows:

$$\tilde{f}_i(m, y^{i-1}, \theta_i) = \begin{cases} f_i(m, y^{i-1}, \theta_i) & \text{if } \theta_i = 0, \\ 0 & \text{if } \theta_i = 1, \end{cases}$$

for all (m, y^{i-1}) and $i = 1, \dots, n$. We argue that $\{\tilde{f}_i(\cdot)\}_{i=1}^n$ and the original decoding function $\Psi(\cdot)$ determine a new code with the same probability of error $P_e^{(n)}$. First, the set of encoding functions, $\{\tilde{f}_i(\cdot)\}_{i=1}^n$, satisfies the input constraint, since we replaced ones with zeros. Further, the output sequence is not affected by our modification, since we replaced only bits that are erased, and therefore, our new code also has probability of error $P_e^{(n)}$. ∎

We introduce $(1, \infty, \text{Ber}(\epsilon))$-RLL *encoder*, which outputs sequences X^n that satisfies two constraints:

1) The $(1, \infty)$-RLL constraint.
2) $X_i = 0$ if $\theta_i = 1$ (the constraint induced by Lemma 4).

The second constraint can be viewed as a "random constraint" since $\theta_i \sim \text{Ber}(\epsilon)$, while the first constraint is a deterministic constraint. Thus, the $(1, \infty, \text{Ber}(\epsilon))$-RLL encoder combines both deterministic and random constraints.

The entropy rate of $(1,\infty,\mathrm{Ber}(\epsilon))$-RLL encoder is measured by $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}H(X_i|X^{i-1},\theta^i)$ since this is the available information at the encoder. The next lemma provides an upper bound on the entropy rate of sequences that can be generated by a $(1,\infty,\mathrm{Ber}(\epsilon))$-RLL encoder.

Lemma 5: The entropy rate of sequences that are generated by a $(1,\infty,Ber(\epsilon))$-RLL encoder is upper bounded by $\max_{0\le p\le\frac{1}{2}}\frac{H_b(p)}{p+\frac{1}{1-\epsilon}}$.

Proof: Recall that the encoder can choose its output bit, x_i, only if $x_{i-1}=\theta_i=0$; we parameterize this by $p(x_i=1|x_{i-1}=0,\theta_i=0)=p$, where $p\in[0,1]$. Now, consider the transition probability matrix of the chain X^n,

$$\mathbf{Q}=\begin{bmatrix}\epsilon+\bar{\epsilon}\,\bar{p} & \bar{\epsilon}\,p\\ 1 & 0\end{bmatrix},$$

where the transition probability $\epsilon+\bar{\epsilon}\,\bar{p}$ was calculated by

$$p(x_i=0|x_{i-1}=0)=\sum_{\theta_i}p(x_i=0,\theta_i|x_{i-1}=0).$$

The stationary distribution of this chain is $[x^*(0)\ x^*(1)]=[\frac{1}{1+\bar{\epsilon}p}\ \ \frac{\bar{\epsilon}p}{1+\bar{\epsilon}p}]$.

Consider the next upper bound for some i,

$$\begin{aligned}&H(X_i|X^{i-1},\theta^i)\\ &\overset{(a)}{\le} H(X_i|X_{i-1},\theta_i)\\ &\overset{(b)}{=} \bar{\epsilon}H(X_i|X_{i-1},\theta_i=0)\\ &\overset{(c)}{=} \bar{\epsilon}H(X_i|x_{i-1}=0,\theta_i=0)p(x_{i-1}=0|\theta_i=0)\\ &\overset{(d)}{=} \bar{\epsilon}H_b(p)p(x_{i-1}=0)\end{aligned}\tag{13}$$

where (a) follows conditioning reduces entropy, (b) follows from $H(X_i|X_{i-1},\theta_i=1)=0$, (c) follows from $H(X_i|x_{i-1}=1,\theta_i=0)=0$, and (d) follows from the fact that X_{i-1} is independent of θ_i and substituting the parameter p.

By substituting the stationary distribution $p(x_{i-1}=0)=x^*(0)$ into (13), we see that the entropy rate of the chain is upper bounded by $\frac{\bar{\epsilon}H_b(p)}{1+\bar{\epsilon}p}$, for some $p\in[0,1]$. This term can also be written as $\frac{H_b(p)}{\frac{1}{1-\epsilon}+p}$, and the parameter p need be maximized only on $[0,0.5]$ from Lemma 1. ∎

The rate of the message M is upper bounded by the entropy rate of sequences that can be generated by a $(1,\infty,\mathrm{Ber}(\epsilon))$-RLL encoder, and this concludes the proof of Theorem 2 with

$$\begin{aligned}C_\epsilon^{\mathrm{nc}} &\le \max_{0\le p\le\frac{1}{2}}\frac{H_b(p)}{p+\frac{1}{1-\epsilon}}\\ &= C_\epsilon^{\mathrm{fb}}.\end{aligned}$$

VIII. Conclusions

We considered the setup of an input-constrained erasure channel with feedback and found its capacity using equivalent DP. We then pursued the complementary derivation of a simple and error-free capacity-achieving coding scheme, which we found using the strong relation between optimal policies in DP and encoding procedures in channel coding. Moreover, we have shown that the capacity remains the same even if the erasure is known non-causally to the encoder.

Following the theorem that feedback does not increase the capacity of a memoryless channel [12], Shannon also argued that this theorem can be extended to channels with memory if the channel state can be computed at the encoder. Shannon, however, omitted the proof of his assertion, so that it remained a conjecture until now. Very recently, Y. Li used the expression for feedback capacity in our Theorem 1 along with tools from [32] to show that Shannon's conjecture is false [33]. To be precise, Li showed that close to $\epsilon=0$, the expression in Theorem 1 strictly exceeds the corresponding capacity without feedback. Additionally, in a parallel work [34], Shaviv et al. gave an example of a finite-state channel model formulated in an energy-harvesting setting, in which it was shown that the feedback capacity is greater than the feed-forward capacity.

As Shannon's conjecture does not hold for our input-constrained erasure channel, it could be interesting to derive the capacity of the input-constrained erasure channel with delayed feedback, namely, when the input to the channel at time i depends on the message and the tuple $Y^{i-\nu}$, where ν is the delay of the feedback. DP formulation for the delayed-feedback capacity is feasible and could yield tighter upper bounds on the capacity of the input-constrained erasure channel without feedback, a problem that is wide open.

Appendix A
Proof of Theorem 3

Recall the statement of Theorem 3:

Theorem 6: The feedback capacity of a $(1,\infty)$-RLL input-constrained memoryless channel can be expressed as:

$$C^{fb}=\sup\liminf_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}I(X_t;Y_t|Y^{t-1}),$$

where the supremum is taken with respect to $\{p(x_t|x_{t-1},y^{t-1}):p(x_t=1|x_{t-1}=1,y^{t-1})=0\}_{t\ge1}$.

This theorem essentially can be deduced from [19, Th. 1]; however, as it is not a special case of their work, we follow similar lines to those taken in their proof, and conclude the capacity expression for our case.

Throughout this section, we use the common notation of directed information, i.e. $I(X^N\to Y^N)=\sum_{t=1}^{N}I(X^t;Y_t|Y^{t-1})$. The notation $Q(x^N||y^{N-1})=\prod_{t=1}^{N}q(x_t|x^{t-1},y^{t-1})$ stands for the causal conditioning, but restricted to satisfy the input constraint. We also define $Q_0(x^N||y^{N-1})$ as $Q(x^N||y^{N-1})$ with the initial condition $q(x_1=0)=1$. In a similar manner, we define P^∞ as the infinite sequence of input distributions $\{p(x_t|x_{t-1},y^{t-1}):p(x_t=1|x_{t-1}=1,y^{t-1})=0\}_{t\ge1}$, and P_0^∞ is defined as P^∞ with $p(x_1=0)=1$.

We also use the notation:

$$\overline{C}_N\triangleq\max_{Q(x^N||y^{N-1})}\frac{1}{N}I(X^N\to Y^N),$$

$$\underline{C}_N\triangleq\max_{Q_0(x^N||y^{N-1})}\frac{1}{N}I(X^N\to Y^N),$$

where the only difference between the defined sequences is the maximization domain.

The following lemma establishes upper and lower bounds on the feedback capacity:

Lemma 6: The capacity can be bounded by:

$$\lim_{N\to\infty} \underline{C}_N \leq C^{fb} \leq \lim_{N\to\infty} \overline{C}_N,$$

and the limit of both sequences exist.

The proof of Lemma 6 appears in Subsection A of this appendix.

Proof of Theorem 3: Equations (14)-(17) below summarize the four main steps of our proof:

$$C^{\text{fb}} = \lim_{N\to\infty} \max_{Q_0(x^N||y^{N-1})} \frac{1}{N} I(X^N \to Y^N) \tag{14}$$

$$= \lim_{N\to\infty} \sup \frac{1}{N} \sum_{t=1}^{N} I(X_t; Y_t|Y^{t-1}) \tag{15}$$

$$= \sup_{P_0^\infty} \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} I(X_t; Y_t|Y^{t-1}) \tag{16}$$

$$= \sup_{P^\infty} \liminf_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} I(X_t; Y_t|Y^{t-1}), \tag{17}$$

where the supremum in (15) is taken over $\left\{p(x_i|x_{i-1}, y^{i-1}) : p(x_i = 1|x_{i-1} = 1, y^{i-1}) = 0, p(x_1 = 0) = 1\right\}_{i=1}^{N}$.

The first step above (Eq. (14)) will be proven by showing the equality:

$$\lim_{N\to\infty} \underline{C}_N = \lim_{N\to\infty} \overline{C}_N,$$

which implies that both bounds provided in Lemma 6 are tight, and are equal to the capacity. We then proceed to the next steps of the proof with $\lim_{N\to\infty} \underline{C}_N$ instead of $\lim_{N\to\infty} \overline{C}_N$, since the exchange of the limit and the supremum (Eq. (16)) will follow in a direct manner from [19, Lemma 4]. Finally, in equality (17), we show that restricting the input distributions to $p(x_1 = 0) = 1$ has no affect on the limit, and this concludes the derivation with the required maximization.

Proof of Equality (14): As mentioned, the difference between the lower and the upper bounds in Lemma 6 is the maximization domain. We show that the limits are equal by constructing for each $Q(x^N||y^{N-1})$, an input distribution from $Q_0(x^N||y^{N-1})$, such that the limit under both distributions is equal. For given $Q(x^N||y^{N-1})$, construct $Q_0(x^N||y^{N-1}) = \prod_{t=1}^{N} q_0(x_t|x^{t-1}, y^{t-1})$ as follows:

$$q_0(x_t|x^{t-1}, y^{t-1}) = \begin{cases} \mathbb{1}_{x_1=0} & \text{if } t = 1, \\ q(x_{t-1}|x^{t-2}, y^{t-2}) & \text{if } t = 2, \ldots, N. \end{cases} \tag{18}$$

We now show that the difference between the normalized directed informations that are induced by the defined distributions vanishes with N,

$$\begin{aligned} &\frac{1}{N} I_Q(X^N \to Y^N) - \frac{1}{N} I_{Q_0}(X^N \to Y^N) \\ &\quad \overset{(a)}{=} \frac{1}{N} I_Q(X^N \to Y^N) - \frac{1}{N} I_Q(X^{N-1} \to Y^{N-1}) \\ &\quad \overset{(b)}{=} \frac{1}{N} I_Q(X^N; Y_N|Y^{N-1}) \\ &\quad \leq \frac{\log|\mathcal{Y}|}{N}, \end{aligned}$$

where (a) is due to the structure of $Q_0(x^N||y^{N-1})$ in (18), and (b) follows by decomposing the directed information into a sum of mutual information instances.

Proof of Equality (15): Each instance in the directed information can be expressed as:

$$\begin{aligned} I(X^t; Y_t|Y^{t-1}) &= H(Y_t|Y^{t-1}) - H(Y_t|X^t, Y^{t-1}) \\ &\overset{(a)}{=} H(Y_t|Y^{t-1}) - H(Y_t|X_t, Y^{t-1}) \\ &= I(X_t; Y_t|Y^{t-1}), \end{aligned}$$

where (a) follows from the memoryless channel property (1).

Now, when the tth instance is determined by $p(x_t, y^t)$ only, it is sufficient to maximize over $\{p(x_i|x_{i-1}, y^{i-1}) : p(x_i = 1|x_{i-1} = 1, y^{i-1}) = 0, p(x_1 = 0) = 1\}_{i=1}^{N}$. The proof of this step is omitted here, and can be found in the justification of Equality 18 in [19, Th. 3].

Proof of Equality (16): The proof of this step follows directly from the proof of [19, Lemma 4]. More specifically, their proof requires two conditions that are satisfied here:

- Super-additive property of the sequence $\underline{C}_N$, which follows from the achievability part in the proof of Lemma 6 (Eq. (19)).
- The concatenation of two input distributions should yield a new input distribution with the same properties. This follows from the fact that our input distributions satisfy $x_1 = 0$ with probability 1; hence, any concatenation of distributions will satisfy the input constraint as well as $x_1 = 0$ with probability 1.

Proof of Equality (17): The last step follows from the same arguments used in equality (14). Specifically, we take the maximizer in Eq. (17) that is from the set $\{p(x_t|x_{t-1}, y^{t-1}) : p(x_t = 1|x_{t-1} = 1, y^{t-1}) = 0\}_{t\geq 1}$ and concatenate it with $p(x_1 = 0) = 1$ as its first instance. Clearly, this new distribution is from the set $\{p(x_t|x_{t-1}, y^{t-1}) : p(x_t = 1|x_{t-1} = 1, y^{t-1}) = 0, p(x_1 = 0) = 1\}_{t\geq 1}$ and has the same limiting expression. ■

A. Proof of Lemma 6

The converse (upper bound) is derived directly in our setting using Fano's inequality and the fact that a $(1, \infty)$-RLL constrained code must induce an input distribution from $Q(x^N||y^{N-1})$. For the second part of the achievability, we define an FSC without input constraint and apply the achievability from [25] with constrained input distributions, i.e. $Q_0(x^N||y^{N-1})$. This gives us a design of a code for the defined FSC, and then we argue that this code can be also implemented in our constrained setting and achieve the same probability of error.

1) Converse: For a $(N, 2^{NR}, (1,\infty))$ code with achievable rate R, consider the following chain of inequalities

$$\begin{aligned} NR &\overset{(a)}{\leq} I(X^N \to Y^N) + N\epsilon_N \\ &\overset{(b)}{\leq} \max_{Q(x^N||y^{N-1})} I(X^N \to Y^N) + N\epsilon_N, \end{aligned}$$

where (a) is a standard derivation by using Fano's inequality with $\epsilon_N \to 0$, (b) is due to the fact that any $(N, 2^{NR}, (1,\infty))$ code that satisfies the input constraint induces an input distribution that is restricted to $Q(x^N||y^{N-1})$. Let us show this property precisely by calculating the conditional probability $p(x_1 = 1|x_{i-1} = 1, y^{i-1})$ induced by any $(N, 2^{NR}, (1,\infty))$ code (and a random message M), for any output tuple y^{i-1}:

$$\begin{aligned} &p(x_i = 1|x_{i-1} = 1, y^{i-1}) \\ &= \sum_m p(m, x_i = 1|x_{i-1} = 1, y^{i-1}) \\ &= \sum_m p(m|x_{i-1} = 1, y^{i-1})\, p(x_i = 1|m, x_{i-1} = 1, y^{i-1}) \\ &\overset{(a)}{=} \sum_m p(m|x_{i-1} = 1, y^{i-1})\, \mathbb{1}_{\{f_i(m,y^{i-1})=1\}\cap\{f_{i-1}(m,y^{i-2})=1\}} \\ &\overset{(b)}{=} 0, \end{aligned}$$

where (a) follows from the notation of encoding functions, and (b) follows from the code constraints in Definition 1.

The limit existence of $\overline{C}_N$ follows from the sub-additivity property of the sequence $N\overline{C}_N + \log|\mathcal{X}|$, i.e.

$$N\overline{C}_N + \log|\mathcal{X}| \leq l\overline{C}_l + \log|\mathcal{X}| + (N-l)\overline{C}_{N-l} + \log|\mathcal{X}|,$$

for all integers l, N such that $l \leq N$.

This property follows from the following chain of inequalities,

$$\begin{aligned} &N\overline{C}_N \\ &= \max_{Q(x^N||y^{N-1})} \sum_{i=1}^{N} I(X^i; Y_i|Y^{i-1}) \\ &\leq l\overline{C}_l + \max_{Q(x^N||y^{N-1})} \sum_{i=l+1}^{N} I(X^i; Y_i|Y^{i-1}) \\ &\leq l\overline{C}_l + \max_{Q(x^N||y^{N-1})} \sum_{i=l+1}^{N} I(X^i; Y_i|Y^{i-1}, X_l) + \log|\mathcal{X}| \\ &\overset{(a)}{\leq} l\overline{C}_l + \max_{Q(x^N||y^{N-1})} \sum_{i=l+1}^{N} I(X_{l+1}^i; Y_i|Y_{l+1}^{i-1}, X_l) + \log|\mathcal{X}| \\ &\leq l\overline{C}_l + \max_{x_l} \max_{Q(x_{l+1}^N||\{y_{l+1}^{N-1}, x_l\})} \sum_{i=l+1}^{N} I(X_{l+1}^i; Y_i|Y_{l+1}^{i-1}, x_l) \\ &\quad + \log|\mathcal{X}| \\ &\overset{(b)}{=} l\overline{C}_l + \max_{Q(x_{l+1}^N||\{y_{l+1}^{N-1}, x_l=0\})} \sum_{i=l+1}^{N} I(X_{l+1}^i; Y_i|Y_{l+1}^{i-1}, x_l = 0) \\ &\quad + \log|\mathcal{X}| \\ &= l\overline{C}_l + \max_{Q(x^{N-l}||y^{N-l-1})} \sum_{i=1}^{N-l} I(X^i; Y_i|Y^{i-1}) + \log|\mathcal{X}| \\ &= l\overline{C}_l + (N-l)\overline{C}_{N-l} + \log|\mathcal{X}|, \end{aligned}$$

where (a) follows from the fact that conditioning reduces entropy and the memoryless property of the channel, and (b) is due to the fact that $x_l = 1$ restricts the maximization domain, while $x_l = 0$ has no affect on the maximization domain. Thus, by Fekete's sub-additive lemma, the limit of the sequence $\overline{C}_N + \frac{\log|\mathcal{X}|}{N}$ exists, which proves the existence of the limit $\lim_{N\to\infty} \overline{C}_N$.

2) Achievability: In the achievability, we use a slightly modified version of an existing result on FSCs from [25, Sec. III]. Specifically, we consider the achievability in [25] for FSCs but with an input distribution that satisfies the input constraint, i.e., it has the form of $Q(x^N||y^{N-1})$.

However, throughout their proof, it is required that a concatenation of two constrained input distributions also yield a constrained input distribution, and this might fail if we use $Q(x^N||y^{N-1})$. To this end, we use $Q_0(x^N||y^{N-1})$ which is the constrained input distribution that begins with the symbol '0' with probability 1. This additional restriction ensures that a concatenation of any such distributions will result in a third input distribution that satisfies the input constraint and the initial condition $q_0(x_1 = 0) = 1$.

The FSC we consider has a single state, and the channel from X to Y is BEC; a direct consequence of applying the achievability in [25, Sec. III] is the following lower bound on the capacity of this FSC:

$$C^{\text{FSC}} \geq \lim_{N\to\infty} \max_{Q_0(x^N||y^{N-1})} \min_{s_0} \frac{1}{N} I(X^N \to Y^N|s_0), \quad (19)$$

where the existence of the limit above follows from the superadditivity property, which is shown in [25, eq. (36)]. Now, as there is only one state in the defined FSC, (19) can be re-written as:

$$\begin{aligned} C^{\text{FSC}} &\geq \lim_{N\to\infty} \max_{Q_0(x^N||y^{N-1})} \frac{1}{N} I(X^N \to Y^N) \\ &= \lim_{N\to\infty} \underline{C}_N. \end{aligned}$$

Note that each code designed for the FSC can be implemented in our original channel since the input distribution implies that the input sequence satisfies the input constraint. Moreover, the channel in the FSC has same characterization as in our original BEC, and therefore, the probability of error will remain the same if we implement the same code on our channel. ■

APPENDIX B
PROOF OF LEMMA 1

- A sufficient condition for the concavity of a function $f(p)$ is that the second derivative is negative for any value of p. We denote $k = \frac{1}{1-\epsilon}$ and find a condition on k such that the second derivative is negative. To simplify the derivations, we take $H_b(\cdot)$ to be the binary entropy with the natural logarithm base, since multiplication by a constant does not effect concavity. Calculation shows that

$$\frac{d^2}{dp^2}\left(\frac{H_b(p)}{p+k}\right) = \frac{\frac{(p+k)^2}{p(p-1)} - 2k\ln\left(\frac{1-p}{p}\right) - 2\ln(1-p)}{p^3}. \quad (20)$$

It suffices to examine the sign of the numerator, since $p^3 \geq 0$. Define $g(p) \triangleq \frac{(p+k)^2}{p(p-1)} - 2k \ln\left(\frac{1-p}{p}\right) - 2\ln(1-p)$. Derivation of the maximum for $g(p)$ shows that it has only one maximum, which is at $p = \frac{1}{2}$. Substituting $g(\frac{1}{2}) = -4(\frac{1}{2}+k)^2 + 2\ln 2$. It then follows that $g(p) \leq 0, \forall p \in [0,1]$ if and only if $k \geq \sqrt{\frac{1}{2}\ln 2} - \frac{1}{2} \sim 0.088$.

- Derivation of the first derivative of $f(p)$ shows that the derivative is equal to zero if and only if $p^{\frac{1}{1-\epsilon}} = (1-p)^{1+\frac{1}{1-\epsilon}}$ holds. The uniqueness of the maximum point follows from the fact that $p^{\frac{1}{1-\epsilon}}$ increases as p grows, while $(1-p)^{1+\frac{1}{1-\epsilon}}$ decreases with a growing p. Now, assume that the maximum is $p_m \in (\frac{1}{2}, 1]$. Symmetry of the binary entropy function implies $H_b(p_m) = H_b(\bar{p}_m)$, and therefore, it is sufficient to examine the denominator. Since both arguments $p_m, \bar{p}_m \in [0,1]$, it then follows that $f(p_m) < f(\bar{p}_m)$, which is a contradiction.
- This property follows from substituting the relation $p^{\frac{1}{1-\epsilon}} = (1-p)^{1+\frac{1}{1-\epsilon}}$ into the function $f(p)$.

Appendix C
Proof of Theorem 5

The next lemma is technical and will be useful in the proof of Theorem 5.

Lemma 7: The function $f_\epsilon(z) = \bar{\epsilon} H_b(z) - z\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}$ *is concave on* $[0,1]$ *and its maximum is at* $z = p_\epsilon$*, where* $p_\epsilon = \arg\max_{0 \leq p \leq \frac{1}{2}} \frac{H_b(p)}{p + \frac{1}{1-\epsilon}}$.

Proof of Lemma 7: The concavity of $f_\epsilon(z)$ on $z \in [0,1]$ follows from the concavity of the binary entropy function, and therefore, it suffices to show that the first derivative of $f_\epsilon(z)$ at p_ϵ is equal to zero. The definition of p_ϵ, (11), and Lemma 1 imply the relation, $\frac{d}{dz}\left[\frac{H_b(z)}{z+\frac{1}{1-\epsilon}}\right]\Big|_{z=p_\epsilon} = 0$, which is equivalent to

$$H_b'(p_\epsilon)\left(p_\epsilon + \frac{1}{1-\epsilon}\right) - H_b(p_\epsilon) = 0. \tag{21}$$

The first derivative of $f_\epsilon(z)$ at the point p_ϵ is:

$$\begin{aligned}
&\frac{d}{dz}\left[\bar{\epsilon} H_b(z) - z\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}\right]\Big|_{z=p_\epsilon} \\
&\quad = \left(\bar{\epsilon} H_b'(z) - \bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}\right)\Big|_{z=p_\epsilon} \\
&\quad = \frac{\bar{\epsilon} H_b'(z)(p_\epsilon + \frac{1}{1-\epsilon}) - \bar{\epsilon} H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad \overset{(a)}{=} 0,
\end{aligned}$$

where (a) follows from (21). ∎

We proceed to the proof of Theorem 5.

Proof of Theorem 5: Recall that the Bellman equation is satisfied with the pair $\tilde{\rho}_\epsilon$ and $\tilde{h}_\epsilon(z)$ if

$$\tilde{\rho}_\epsilon + \tilde{h}_\epsilon(z) = (T\tilde{h}_\epsilon)(z) \tag{22}$$

holds; while the left hand side of (22) is given explicitly in (11)-(12), one should calculate the expression $(T\tilde{h}_\epsilon)(z)$ which can be simplified as follows:

$$\begin{aligned}
&(T\tilde{h}_\epsilon)(z) \\
&\quad = \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) + \bar{\epsilon}(1-\delta)\tilde{h}_\epsilon(1) + \epsilon\tilde{h}_\epsilon(1-\delta) + \bar{\epsilon}\delta\tilde{h}_\epsilon(0) \\
&\quad \overset{(a)}{=} \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \epsilon\tilde{h}_\epsilon(1-\delta),
\end{aligned}$$

where (a) follows from the definition of $\tilde{h}_\epsilon(z)$ in (12), specifically, $\tilde{h}_\epsilon(0) = 0$ and $\tilde{h}_\epsilon(1) = \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}$.

The term $\tilde{h}_\epsilon(1-\delta)$ is now calculated for two cases:

$$\tilde{h}_\epsilon(1-\delta) = \begin{cases} \bar{\epsilon} H_b(\delta) - (1-\delta)\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} & \text{if } 1-\delta < p_\epsilon \\ \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} & \text{if } 1-\delta \geq p_\epsilon. \end{cases} \tag{23}$$

To complete the proof, we have three cases for calculating the operator $(T\tilde{h}_\epsilon)(z)$:

- For $0 \leq z < p_\epsilon$, the constraint $0 \leq \delta \leq z$ implies that $0 \leq \delta < p_\epsilon$, and from (23), we have $\tilde{h}_\epsilon(1-\delta) = \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}$. Let us show that (22) is satisfied:

$$\begin{aligned}
&(T\tilde{h}_\epsilon)(z) \\
&\quad = \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \epsilon\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad = \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) - \delta\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad \overset{(a)}{=} \bar{\epsilon} H_b(z) - z\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad = \tilde{h}_\epsilon(z) + \tilde{\rho}_\epsilon,
\end{aligned}$$

where (a) follows from Lemma 7.

- For $p_\epsilon \leq z < 1 - p_\epsilon$, the same calculation as for the previous interval shows that $\tilde{h}_\epsilon(1-\delta) = \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}}$ for all $\delta \in [0, 1-p_\epsilon]$. Let us show that (22) is satisfied:

$$\begin{aligned}
&(T\tilde{h}_\epsilon)(z) \\
&\quad = \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \epsilon\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad = \sup_{0 \leq \delta \leq z} \bar{\epsilon} H_b(\delta) - \delta\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad \overset{(a)}{=} \bar{\epsilon} H_b(p_\epsilon) - p_\epsilon\bar{\epsilon}\frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad = \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} + \frac{H_b(p_\epsilon)}{p_\epsilon + \frac{1}{1-\epsilon}} \\
&\quad = \tilde{h}_\epsilon(z) + \tilde{\rho}_\epsilon,
\end{aligned}$$

where (a) follows from Lemma 7.

- The last calculation is for $1 - p_\epsilon \leq z \leq 1$; in this case, the expression $\tilde{h}_\epsilon(1-\delta)$ might be equal to different functions according to the value of δ. However, we show

the maximization on δ can be restricted to $[0, 1-p_\epsilon]$ which determine uniquely the function $\tilde{h}_\epsilon(1-\delta)$:

$$\begin{aligned}
&(T\tilde{h}_\epsilon)(z)\\
&= \sup_{0\le\delta\le z} \bar{\epsilon}H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}} + \epsilon\tilde{h}_\epsilon(1-\delta)\\
&\overset{(a)}{=} \sup_{0\le\delta\le 1-p_\epsilon} \bar{\epsilon}H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}} + \epsilon\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&\overset{(b)}{=} 2\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&= \tilde{h}_\epsilon(z) + \tilde{\rho}_\epsilon,
\end{aligned}$$

where (b) follows from Lemma 7, and (a) follows from the following upper bound:

$$\begin{aligned}
&\sup_{1-p_\epsilon<\delta\le z} \bar{\epsilon}H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}} + \epsilon\tilde{h}_\epsilon(1-\delta)\\
&= \sup_{1-p_\epsilon<\delta\le z} \bar{\epsilon}(1+\epsilon)H_b(\delta) + \bar{\epsilon}\bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&\le \sup_{1-p_\epsilon\le\delta\le z} \bar{\epsilon}(1+\epsilon)H_b(\delta) + \sup_{1-p_\epsilon\le\delta\le z} \bar{\epsilon}\bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&= \bar{\epsilon}(1+\epsilon)H_b(1-p_\epsilon) + \bar{\epsilon}\bar{\epsilon}(1-(1-p_\epsilon))\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&= \frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\left[2\bar{\epsilon}p_\epsilon + 1 + \epsilon\right]\\
&\le 2\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}\\
&= \sup_{0\le\delta\le 1-p_\epsilon} \bar{\epsilon}H_b(\delta) + \bar{\epsilon}(1-\delta)\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}} + \epsilon\frac{H_b(p_\epsilon)}{p_\epsilon+\frac{1}{1-\epsilon}}.
\end{aligned}$$

■

APPENDIX D
ACCURATE RATE ANALYSIS

The rate analysis in Section VI was simplified by assuming that each transmitted bit is Ber(p). Here, we show precisely that our coding scheme can be arbitrarily close to C_ϵ^{fb}. The idea is to separate the coding scheme into two parts using a parameter λ, which is a fixed constant. First, we use the coding scheme from Section VI-B to transmit a large number, $nR-\lambda$, of message bits, while a different coding scheme will be used to transmit the remaining λ bits. We show that the rate of the overall scheme is essentially determined by the rate of the first coding scheme. The next lemma will be used for the rate analysis of the first coding scheme,

Lemma 8: Each transmitted bit, X_i, can be chosen to be distributed as Ber$(p-e_i)$, where $0 \le e_i < \frac{1}{|\mathcal{M}_{i-1}|}$.

Proof: Assume that at time i, a procedure begins and its corresponding set of possible messages is $\mathcal{M}_{i-1}$. According to L_1, the number of messages that are labelled '1' is $\lfloor p|\mathcal{M}_{i-1}|\rfloor$, where $\lfloor\cdot\rfloor$ is the floor operator. The resulting input distribution is $X_i \sim \text{Ber}\left(\frac{\lfloor p|\mathcal{M}_{i-1}|\rfloor}{|\mathcal{M}_{i-1}|}\right)$, which can be written also as $X_i \sim \text{Ber}(p-e_i)$ since $p - \frac{1}{|\mathcal{M}_{i-1}|} < \frac{\lfloor p|\mathcal{M}_{i-1}|\rfloor}{|\mathcal{M}_{i-1}|} \le p$.

In case of erasure at time i, recall that the number of messages that were labelled '0' in L_1 is greater than the number of messages labelled '1', and thus, we are able to construct the labelling L_2 as follows; $\lfloor p|\mathcal{M}_{i-1}|\rfloor$ messages that were labelled '0' at the previous transmission are flipped to '1', and all the remaining messages are labelled '0'. It is clear that the input distribution is preserved in this case, and upon consecutive erasures, L_1 and L_2 are being exchanged and the input distribution is not changed. Note that the choices of labelling are made in advance and both encoder and decoder agree on current labelling. ■

The encoding procedure occurs repeatedly and is over when the set of possible messages is less or equal than 2^λ. Denote by $e_1, e_2, \ldots, e_k$ the correction factors for the k successful transmissions until the scheme is over. Following the same derivations in Section VI, it follows that the rate is $\tilde{R} = \frac{\sum_{i=1}^k H_b(p-e_i)}{k(\frac{1}{1-\epsilon}+p)-\sum_{i=1}^k e_i}$.

For the λ remaining bits, we perform a code where a bit of message is followed by zero and this pair is transmitted repeatedly until a successful transmission. Thus, to send the message bit '0', the pair '00' is repeated until '00' or '0?' are received, and to send the message bit '1', the bits $'10'$ are repeatedly transmitted until a '1' is received. The decoding for this scheme is straightforward, and calculation of the rate gives that $\bar{R} = \frac{1-\epsilon}{2}$.

To summarize, the average rate for the overall coding scheme is

$$R = \left(\frac{nR-\lambda}{nR}\right)\tilde{R} + \left(\frac{\lambda}{nR}\right)\bar{R}.$$

Consider the next lower bound on R,

$$\begin{aligned}
R &= \left(\frac{nR-\lambda}{nR}\right)\frac{\sum_{i=1}^k H_b(p-e_i)}{k(\frac{1}{1-\epsilon}+p)-\sum_{i=1}^k e_i} + \left(\frac{\lambda}{nR}\right)\frac{1-\epsilon}{2}\\
&\ge \left(\frac{nR-\lambda}{nR}\right)\frac{k\min_i H_b(p-e_i)}{k(\frac{1}{1-\epsilon}+p)-k\min_i e_i} + \left(\frac{\lambda}{nR}\right)\frac{1-\epsilon}{2}\\
&\overset{(a)}{\ge} \left(\frac{nR-\lambda}{nR}\right)\frac{H_b(p-2^{-\lambda})}{\frac{1}{1-\epsilon}+p} + \left(\frac{\lambda}{nR}\right)\frac{1-\epsilon}{2},
\end{aligned}$$

where (a) follows from Lemma 8, namely, $e_i \in [0, 2^{-\lambda})$ for $i = 1, \ldots, k$.

Letting $n \to \infty$, we see that $R^* = \frac{H_b(p-2^{-\lambda})}{\frac{1}{1-\epsilon}+p}$ is achievable. Thus, by choosing λ to be arbitrarily large (but still finite), we can make R^* arbitrarily close to the capacity C_ϵ^{fb}.

ACKNOWLEDGMENT

The authors would like to thank Yonglong Li and Prof. Guangyue Han for providing us the data for the lower bound plotted in Fig. 3. They would also like to thank Prof. Paul Cuff, for providing an insightful observation which led to a great simplification of the proposed coding scheme. Finally, they would like to thank the Associate Editor and anonymous reviewers, for their valuable and constructive comments, which helped to improve this paper.

REFERENCES

[1] O. Sabag, H. H. Permuter, and N. Kashyap, "Capacity of the (1, ∞)-RLL input-constrained erasure channel with feedback," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr./May 2015, pp. 1–5.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[3] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.
[4] 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.
[5] P. O. Vontobel, A. Kavcic, 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.
[6] G. Han and B. H. 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.
[7] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Inf. Theory*, vol. 34, no. 1, pp. 45–54, Jan. 1988.
[8] G. Han and B. H. 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.
[9] B. H. Marcus, R. M. Roth, and P. H. Siegel, "Constrained systems and coding for recording channels," in *Handbook of Coding Theory*, V. S. Pless and W. C. Huffman, Eds. Amsterdam, The Netherlands: Elsevier, 1998, pp. 1635–1764.
[10] K. Immink, *Codes for Mass Data Storage Systems*. Rotterdam, The Netherlands: Shannon Foundation, 2004.
[11] A. M. Fouladgar, O. Simeone, and E. Erkip, "Constrained codes for joint energy and information transfer," *IEEE Trans. Commun.*, vol. 62, no. 6, pp. 2121–2131, Jun. 2014.
[12] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.
[13] Y. Li and G. Han, "Input-constrained erasure channels: Mutual information and capacity," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 3072–3076.
[14] S. C. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2000.
[15] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.
[16] S. Yang, A. Kavčić, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.
[17] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
[18] S. Yang, A. Kavčić, and S. C. Tatikonda, "On the feedback capacity of power-constrained Gaussian noise channels with memory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 929–954, Mar. 2007.
[19] H. H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2008.
[20] O. Elishco and H. Permuter, "Capacity and coding for the Ising channel with feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5138–5149, Sep. 2014.
[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vols. 1–2, 3rd ed. Belmont, MA, USA: Athena Scientific, 2005.
[22] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Nov. 1990, pp. 303–305.
[23] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[24] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 54, no. 4, pp. 1488–1499, Apr. 2008.
[25] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.
[26] B. Shrader and H. Permuter, "Feedback capacity of the compound channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3629–3644, Aug. 2009.
[27] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.
[28] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 9, no. 3, pp. 136–143, Jul. 1963.
[29] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback—Part I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, Apr. 1966.
[30] Y.-H. Kim, "Feedback capacity of the first-order moving average Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3063–3079, Jul. 2006.
[31] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.
[32] G. Han and B. Marcus, "Asymptotics of input-constrained binary symmetric channel capacity," *Ann. Appl. Probab.*, vol. 19, no. 3, pp. 1063–1091, Jun. 2009.
[33] Y. Li and G. Han, private communication, Jun. 2015.
[34] D. Shaviv, A. Ozgur, and H. Permuter, "Can feedback increase the capacity of the energy harvesting channel?" in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr./May 2015, pp. 1–5. [Online]. Available: http://arxiv.org/abs/1506.02026

Oron Sabag (S'14) received his B.Sc. (cum laude) degree in Electrical and Computer Engineering from the Ben-Gurion University of the Negev, Israel, in 2013. He is currently pursuing his Ph.D in the direct track for honor students in Electrical and Computer Engineering at the same institution. Oron is a recipient of the Lachish Fellowship for honor students in the direct Ph.D. Program and excellence scholarship from the Electrical and Computer Engineering department, Ben-Gurion University of the Negev, Israel.

Haim H. Permuter (M'08–SM'13) received his B.Sc. (summa cum laude) and M.Sc. (summa cum laude) degrees in Electrical and Computer Engineering from the Ben-Gurion University, Israel, in 1997 and 2003, respectively, and the Ph.D. degree in Electrical Engineering from Stanford University, California in 2008. Between 1997 and 2004, he was an officer at a research and development unit of the Israeli Defense Forces. Since 2009 he is with the department of Electrical and Computer Engineering at Ben-Gurion University where he is currently an associate professor. Prof. Permuter is a recipient of several awards, among them the Fullbright Fellowship, the Stanford Graduate Fellowship (SGF), Allon Fellowship, and and the U.S.-Israel Binational Science Foundation Bergmann Memorial Award. Haim is currently serving on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY.

Navin Kashyap (S'97–M'02–SM'07) received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay, in 1995, the M.S. degree in Electrical Engineering from the University of Missouri-Rolla in 1997, and the M.S. degree in Mathematics and the Ph.D. degree in Electrical Engineering from the University of Michigan, Ann Arbor, in 2001. From November 2001 to November 2003, he was a postdoctoral research associate at the University of California, San Diego. From 2004 to 2010, he was at the Department of Mathematics and Statistics at Queen's University, Kingston, Ontario, first as an Assistant Professor, then as an Associate Professor. In January 2011, he joined the Department of Electrical Communication Engineering at the Indian Institute of Science as an Associate Professor. His research interests lie primarily in the application of combinatorial and probabilistic methods in information and coding theory. Prof. Kashyap served on the editorial board of the IEEE TRANSACTIONS ON INFORMATION THEORY during the period 2009-2014.