The Feedback Capacity of Noisy Output Is the STate (NOST) Channels

Eli Shemuel, *Student Member, IEEE*, Oron Sabag, *Member, IEEE*, and Haim H. Permuter, *Senior Member, IEEE*

***Abstract*— We consider finite-state channels (FSCs) where the channel state is stochastically dependent on the previous channel output. We refer to these as Noisy Output is the STate (NOST) channels. We derive the feedback capacity of NOST channels in two scenarios: with and without causal state information (CSI) available at the encoder. If CSI is unavailable, the feedback capacity is $C_{\text{FB}} = \max_{P(x|y')} I(X;Y|Y')$, while if it is available at the encoder, the feedback capacity is $C_{\text{FB-CSI}} = \max_{P(u|y'),x(u,s')} I(U;Y|Y')$, where U is an auxiliary RV with finite cardinality. In both formulas, the output process is a Markov process with stationary distribution. The derived formulas generalize special known instances from the literature, such as where the state is i.i.d. and where it is a deterministic function of the output. C_{FB} and $C_{\text{FB-CSI}}$ are also shown to be computable via convex optimization problem formulations. Finally, we present an example of an interesting NOST channel for which CSI available at the encoder does not increase the feedback capacity.**

***Index Terms*— Channel capacity, channels with memory, convex optimization, feedback capacity, finite state channels.**

I. Introduction

THE popular model of finite-state channels (FSCs) [1]–[6] has been motivated by channels or systems with memory, common in wireless communication [7]–[13], molecular communication [14], [15] and magnetic recordings [16]. The memory of a channel or a system is encapsulated in a finite set of states in the FSC model. Although feedback does not increase the capacity of memoryless channels [17], [18], it can generally increase the capacity of channels with memory. Nonetheless, in the general case, both the capacity and the feedback capacity of FSCs are characterized by multi-letter expressions that are non-computable, and they still have no simple closed-form formulas. That is, a general capacity formula for channels with memory is given as the limit of the n-fold mutual information sequence [2], [19]–[21], whereas the feedback capacity is commonly expressed by the limit of the n-fold directed information [22]–[29].

The explicit capacity of channels with memory is known only in a few instances of channels where feedback does not increase the capacity, such as the POST(α) channel [30] and channels with certain symmetric properties [31]–[34]. Furthermore, a single-letter expression was derived in [35] for the capacity of FSCs with channel state information known at the receiver and delayed feedback in the absence of inter-symbol interference (ISI), i.e., the channel state is input-independent. There are some additional special cases of FSCs where the feedback capacity is known explicitly. One method to compute an explicit feedback capacity expression is by formulating it as a dynamic programming (DP) optimization problem, as was first introduced in Tatikonda's thesis [36] and then in [25], [27], [37]–[40]. This is beneficial in estimating the feedback capacity using efficient algorithms such as the value iteration algorithm [41], which, in turn, can help in generating a conjecture for the exact solution of the corresponding Bellman equation [42]. Thus, for a family of FSCs with ISI called unifilar FSCs, in which the new channel state is a deterministic, time-invariant function of the current state, the input and the output, the feedback capacity can be computed via DP, as was formulated in [27], and closed-form expressions or exact values for the feedback capacity of particular unifilar FSCs were derived in [27], [43]–[49]. For a sub-family of unifilar FSCs where the channel state is a deterministic function of the channel output, a single-letter feedback capacity expression was derived in [39]. Another method to compute explicit feedback capacity expressions is the Q-graph method that was introduced and utilized in [47], [49]–[51] for unifilar FSCs.

Motivated by the absence of a single-letter, computable feedback capacity formula for FSCs with ISI where the channel state evolves stochastically, in this paper, we investigate FSCs where the state is stochastically dependent on the output. This is a generalization of the unifilar FSCs studied in [39]. We refer this generalization as *Noisy Output is the STate (NOST) channels*. We study two settings of NOST channels subject to the availability of causal state information (CSI) at the encoder, as illustrated in Fig. 1. The

Manuscript received 13 July 2021; revised 20 January 2022; accepted 7 March 2022. Date of publication 7 April 2022; date of current version 13 July 2022. This work was supported in part by the German Research Foundation (DFG) via the German–Israeli Project Cooperation [DIP], in part by the Israel Science Foundation (ISF) Research Grant 899/21, and in part by the Israeli Innovation Authority as part of the Wireless Intelligent Networks (WIN) Consortium. The work of Eli Shemuel was supported by the Ministry of Science, Technology and Space of Israel. The work of Oron Sabag was supported in part by the Israel Scholarship Education Foundation (ISEF) Postdoctoral Fellowship. *(Corresponding author: Eli Shemuel.)*

Eli Shemuel and Haim H. Permuter are with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er Sheva 84105, Israel (e-mail: els@post.bgu.ac.il; haimp@post.bgu.ac.il).

Oron Sabag is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: oron@caltech.edu).

Communicated by L. Wang, Associate Editor for Shannon Theory and Information Measures.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3165538.

Digital Object Identifier 10.1109/TIT.2022.3165538

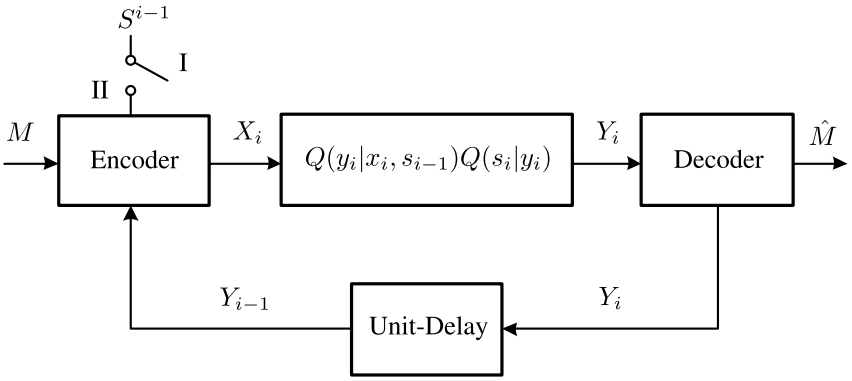

Fig. 1. NOST channels in the presence of feedback. Setting I (open switch) – no CSI is available. Setting II (closed switch) – CSI available at the encoder.

first setting is where CSI is unavailable, while the second is where it is available at the encoder. Our main contribution in this paper is single-letter, computable feedback capacity formulas. We show the computability of the feedback capacity formulas by formulating them as convex optimization.

The achievability of the feedback capacity of the first setting is based on rate-splitting and random coding, and a similar proof was also given in [35]. A posterior matching scheme [52], a principle that was also used in [46], can also be used for the achievability from the work of [53]. The converse of the feedback capacity is based on a recently developed technique to derive upper bounds with stationary distributions [54]. In fact, the first setting can be shown to be equivalent to the setting mentioned in [39], and our formula is identical to theirs; however, our result generalizes their capacity result in two ways. Firstly, the feedback capacity formula in [39] is subject to a restricting irreducibility assumption that we are able to relax due to our converse technique. This allows us to determine the feedback capacity of additional channels such as the POST(α) channel [30]. Secondly, as aforementioned, our convex optimization formulation of the formula enables us to compute the feedback capacity.

The second setting in this paper, where CSI is available at the encoder, is somewhat more challenging, and it reveals an interesting interface between the utilization of CSI and the memory of the channel. This side information may generally be beneficial for increasing the feedback capacity of channels with memory. The capacity problem of discrete memoryless channels (DMCs) with states known at the encoder dates back to Shannon's early work [55], followed by works of Kusnetsov and Tsybakov [56], Gel'fand and Pinsker [57], and Heegard and El Gamal [58], which paved the way to various recent works such as [40], [59]–[67]. Since Shannon showed in [17] that feedback does not increase the capacity of a DMC, his setting in [55], where the state process is i.i.d. and known causally at the encoder, is covered by our second setting by assuming, in particular, that the state is independent of the output. Furthermore, we show that the capacity expression of this Shannon's setting is covered by ours.

The remainder of the paper is organized as follows. Section II defines the notation and the settings. Section III presents the main results regarding the single-letter feedback capacity expressions and their convex optimization formulations. Section IV shows how the capacity expression of each setting covers the capacity characterization of several special cases from the literature, and provides an interesting example of a NOST channel for which CSI available at the encoder does not increase its feedback capacity; this example is referred to as *the noisy-POST*(α, η) *channel*, and is a generalization of the POST(α) channel [30]. Section V provides proofs and derivations of the main results. Finally, Section VI concludes this work.

II. Problem Definition

In this section, we introduce the notation and the communication setup.

A. Notation

Lowercase letters denote sample values (e.g. x, y), and uppercase letters denote discrete random variables (RVs) (e.g. X, Y). Subscripts and superscripts denote vectors in the following way: $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$ and $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$ for $1 \leq i \leq j$. x^n and X^n are shorthand for x_1^n and X_1^n, respectively. We use calligraphic letters (e.g. $\mathcal{X}, \mathcal{Y}$) to denote alphabets, and $|\cdot|$ (e.g. $|\mathcal{X}|$) to denote the cardinality of an alphabet. For two RVs X, Y the probability mass function (PMF) of X is denoted by $P_X(x)$, the conditional PMF of $X = x$ given $Y = y$ is denoted by $P_{X|Y}(x|y)$, and the joint PMF is denoted by $P_{X,Y}(x, y)$; the shorthand $P(x), P(x|y), P(x, y)$ are used for the above, respectively, when the RVs are clear from the context. The indicator function is denoted by $\mathbb{1}\{\cdot\}$. We define $\bar{a} \triangleq 1 - a$ for some $a \in [0, 1]$. For a pair of integers $n \leq m$, we define the discrete interval $[n : m] \triangleq \{n, n + 1, \ldots, m\}$.

B. The Communication Setup

We consider FSCs as shown in Fig. 1. An FSC [2] consists of finite input, output and channel state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, respectively. It is defined by the model ($\mathcal{X} \times \mathcal{S}$, $Q(y, s|x, s')$, $\mathcal{Y}$) where s', s are the channel state at the beginning and at the end of a transmission, respectively. The channel is stationary in the sense that when it is used n times with message M and inputs X^n, at time $i \in [1 : n]$ given the past, it has the Markov property

$$Q(y_i, s_i|x^i, s_0^{i-1}, y^{i-1}, m)$$

$$= Q_{Y,S|X,S'}(y_i, s_i|x_i, s_{i-1}) \quad (1)$$

$$= Q_{Y|X,S'}(y_i|x_i, s_{i-1})Q_{S|Y}(s_i|y_i), \quad (2)$$

where (1) holds for any general FSC and (2) is particularized for NOST channels. We also use the averaged channel defined by

$$Q_{Y|X,Y'}(y|x, y') = \sum_{s' \in \mathcal{S}} Q_{S|Y}(s'|y')Q_{Y|X,S'}(y|x, s'), \quad (3)$$

where $y', y \in \mathcal{Y}$ are interpreted as the channel outputs before and after a transmission, respectively. The initial channel state is assumed to be distributed according to $Q(s_0)$ and will be shown to have no effect on the feedback capacity solution subject to a connectivity assumption. We consider two communication scenarios of NOST channels: without and with CSI available at the encoder.

1) Setting I—No CSI: At time i, the encoder has access to the message $m \in \mathcal{M}$ and the outputs' feedback, where the message set is $\mathcal{M} = [1 : \lceil 2^{nR} \rceil]$ and M is assumed to be uniformly distributed over $\mathcal{M}$. The encoder's mapping at time i is denoted by

$$x_i : \mathcal{M} \times \mathcal{Y}^{i-1} \to \mathcal{X}, \quad (4)$$

and the decoder's mapping is

$$\hat{m} : \mathcal{Y}^n \to \mathcal{M}. \quad (5)$$

A $(2^{nR}, n)$ code is a pair of encoding and decoding mappings (4)-(5). A rate R is *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that the *average probability of error*, $P_e^{(n)} = \Pr(\hat{M} \neq M)$, tends to zero as $n \to \infty$. The *feedback capacity* is the supremum over all achievable rates, and is denoted by C_{FB}.

2) Setting II—CSI Available at the Encoder: Setting II is defined similarly to Setting I, except that at time i the encoder has also causal access to the channel states, that is,

$$x_i : \mathcal{M} \times \mathcal{S}_0^{i-1} \times \mathcal{Y}^{i-1} \to \mathcal{X}. \quad (6)$$

A $(2^{nR}, n)$ code is a pair of encoding and decoding mappings given by (6) and (5), respectively. The feedback capacity of Setting II is denoted by $C_{\text{FB-CSI}}$.

For the sake of simplicity when defining a property called *connectivity* as follows and writing proofs throughout the paper, without loss of generality, we can assume the existence of an initial output, Y_0. That is, except for the case where the state is the output, Y_0 is fictitious with some arbitrary distribution $Q(y_0)$ and is independent of S_0.

Now, for both Setting I and Setting II, we assume that the NOST channels are connected.

Definition 1 (Connectivity): A NOST channel (2) is *connected* if for any pair of outputs, $y', y \in \mathcal{Y}$, there exist T (shorthand for $T(y', y)$) and a sequence of channel inputs x^T (shorthand for $x^T(y', y)$) such that $Q_{Y_T|X^T,Y_0}(y|x^T, y') > 0$.

By Definition 1, we want to avoid scenarios where the initial state determines the set of accessible outputs for the entire transmission procedure. Alternatively, we may formulate Definition 1 as follows. Assume, without loss of generality, that $\mathcal{Y} = [1 : |\mathcal{Y}|]$, and denote by Q the $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix $[Q_{ij}], i, j \in \mathcal{Y}$, where

$$Q_{ij} \triangleq \max_{x \in \mathcal{X}} Q_{Y_k|X_k,Y_{k-1}}(j|x, i). \quad (7)$$

A NOST channel (2) is *connected* if for all $i, j \in \mathcal{Y}$ there exists an integer $T(i, j)$ such that Q satisfies

$$(\overbrace{Q \cdots Q}^{T(i,j) \text{ times}})_{ij} > 0. \quad (8)$$

This definition is equivalent to Definition 1 since both of them imply that for any initial output y' (or i) and any desired output y (or j), there exists a sequence of channel inputs such that there is a positive probability of reaching y from y'.

III. Main Results

In this section we present our main results pertaining to Setting I and Setting II.

A. Setting I—No CSI

The following theorem characterizes the capacity of Setting I, C_{FB}, as a single-letter expression.

Theorem 1 (Feedback Capacity of Setting I): The feedback capacity of a connected NOST channel without CSI is given by

$$C_{\text{FB}} = \max_{P(x|y')} I(X; Y|Y'), \quad (9)$$

where the joint distribution is $P(y', x, y) = \pi(y')P(x|y')Q(y|x, y')$, $Q(y|x, y')$ is defined in (3), and $\pi(y')$ is a stationary distribution induced by the Markov kernel $P(y|y') = \sum_x P(x|y')Q(y|x, y')$.

We note that the averaged channel $Q(y|x, y')$ (3), given as a part of the joint distribution, implies that $S' - Y' - X$ forms a Markov chain. By definition, $\pi(y')$ is a stationary distribution if it is a solution of $\pi P = \pi$, where π is a probability vector on $\mathcal{Y}$, and P denotes the probability transition matrix $P(y|y')$, whose rows and columns represent the previous and next outputs $y', y \in \mathcal{Y}$, respectively; thus $P_{Y'}(y') = \pi(y') = P_Y(y'), \forall y' \in \mathcal{Y}$. From the assumption that $\mathcal{Y}$ is a finite set, there is always at least one stationary distribution (see, e.g., [68, Chapter 5.5]) for any input distribution. If the stationary distribution is not unique, there are infinitely many stationary output distributions.[1] However, the following lemma states that the maximum in (9) can always be attained by an input distribution that induces a unique stationary output distribution.

Lemma 1: For a connected NOST channel without CSI,

$$\max_{P(x|y')} I(X; Y|Y') = \max_{P(x|y') \in \mathcal{P}_\pi} I(X; Y|Y'), \quad (10)$$

where $\mathcal{P}_\pi$ is defined as the set of input distributions that induce a unique stationary output distribution $\pi(y')$, and this set is non-empty.

[1]For instance, if $|\mathcal{Y}| = 2$ and $P(x|y')$ induces a transition matrix $P(y|y')$ given by the identity matrix of size 2, I_2, all output distributions are stationary, i.e., any $\pi \triangleq [p\ \bar{p}], p \in [0, 1]$ is a distribution solving $\pi I_2 = \pi$.

Given an optimal distribution $P(x|y') \in \mathcal{P}_\pi$, the stationary output distribution $\pi(y')$ is well-defined regardless of $Q(s_0)$, and therefore C_{FB} is independent of $Q(s_0)$.

A special case of Setting I is where the channel state is the output, i.e., $s_i = y_i$. Conversely, by Theorem 1, it can be seen that C_{FB} depends on the averaged channel $Q(y|x, y')$ (3); hence, Setting I and this special case are operationally equivalent. In other words, we may define a fictitious state y' and obtain a channel which is operationally equivalent to the special case $s_i = y_i$. We note for this special case, the upper bound on the feedback capacity in [47] coincide with C_{FB} (9). Furthermore, the special case $s_i = y_i$ has been studied in [39], and the feedback capacity was derived under some assumptions on the channel; nevertheless, our result generalizes upon their result in two ways. First, we relax the assumptions in [39] to a connectivity condition (Definition 1), which allows us to determine the feedback capacity of a wide family of channels, e.g., the POST(α) channel, whose feedback capacity was derived in [30]. We discuss the relaxation issue and demonstrate the connectivity condition in Section IV. Second, another contribution of our work regarding Setting I is a novel convex optimization formulation of C_{FB}, given in the following theorem.

From (9) it is not clear if $I(X;Y|Y')$ is a concave function of $P(x|y')$; however, Theorem 2 clarifies that it is a concave function of the joint distribution, $P(y', x)$.

Theorem 2 (Convex Optimization for C_{FB}): The feedback capacity of a connected NOST channel without CSI, C_{FB}, can be formulated as the following convex optimization problem:

$$\underset{P(y',x)\in\mathcal{P}(\mathcal{Y}\times\mathcal{X})}{\text{maximize}} \quad I(X;Y|Y') \tag{11a}$$

$$\text{subject to} \quad \sum_{x,y} P_{Y',X}(\tilde{y},x)Q_{Y|X,Y'}(y|x,\tilde{y}) - \sum_{y',x} P_{Y',X}(y',x)Q_{Y|X,Y'}(\tilde{y}|x,y') = 0, \quad \forall \tilde{y} \in \mathcal{Y}. \tag{11b}$$

The benefit in formulating the feedback capacity as a convex optimization problem is the ability thus afforded to compute it via implementing known convex optimization algorithms. The main idea in this formulation is to view the stationary distribution as the set of linear constraints on the joint distribution, as presented in Constraints (11b), which are equivalent to $P_{Y'}(\tilde{y}) = P_Y(\tilde{y})$. A similar approach for stationary constraints also appears in [50].

B. Setting II—CSI Available at the Encoder

The following theorem characterizes the capacity of Setting II, $C_{\text{FB-CSI}}$.

Theorem 3 (Feedback Capacity of Setting II): The feedback capacity of a connected NOST channel with CSI available at the encoder is

$$C_{\text{FB-CSI}} = \max_{P(u|y'),x=f(u,s')} I(U;Y|Y'), \tag{12}$$

where the joint distribution is $P(y', u, y) = \pi(y')P(u|y')P_f(y|u, y')$, in which

$$P_f(y|u,y') = \sum_{s',x} Q(s'|y')\mathbb{1}\{x = f(u,s')\}Q(y|x,s'), \tag{13}$$

$\pi(y')$ is a stationary distribution induced by the Markov kernel $P(y|y') = \sum_u P(u|y')P_f(y|u, y')$, and U is an auxiliary RV with $|\mathcal{U}| \le L \triangleq \min\{(|\mathcal{X}|^{|\mathcal{S}|}, (|\mathcal{X}| - 1)|\mathcal{S}||\mathcal{Y}| + 1, (|\mathcal{Y}| - 1)|\mathcal{Y}| + 1\}$.

We note that (13) implies that $S' - Y' - U$ forms a Markov chain. The feedback capacity expression in (12) is interesting, as it combines the idea of an auxiliary RV and stationary distributions. This is the first appearance in the literature of such an expression that results in a single-letter capacity expression. Any realization of the auxiliary RV $u \in \mathcal{U}$ represents a deterministic mapping from $\mathcal{S}$ to $\mathcal{X}$. Such mappings are called *strategies*, and were introduced in Shannon's work [55]. Furthermore, although there is generally a total of $|\mathcal{X}|^{|\mathcal{S}|}$ strategies, $C_{\text{FB-CSI}}$ can be achieved with at most L of them. Hence, the maximization on $x = f(u, s')$ in (12) is to choose a subset of L maximizing strategies from the set of all strategies (that is, $\binom{|\mathcal{X}|^{|\mathcal{S}|}}{L}$ ways to choose in total). We note that in the case where the state sequence is i.i.d., $C_{\text{FB-CSI}}$ recovers the capacity derived by Shannon [55] (feedback cannot increase the capacity of DMCs [17]) with the cardinality bound $|\mathcal{U}| \le \min\{(|\mathcal{X}| - 1)|\mathcal{S}| + 1, |\mathcal{Y}|\}$ (see, e.g., [69]). In comparison, our general cardinality bound, L, has a multiplication by $|\mathcal{Y}|$ because of the memory preserved by the previous output, but both cardinality bounds coincide in the case of i.i.d. states.

Analogically to Lemma 1 for the case without CSI, the following Lemma declares that the maximum in (12) can particularly be attained by an input distribution and a function $f : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$ that induce a uniqueness of the stationary output distribution.

Lemma 2: For a connected NOST channel with CSI available at the encoder,

$$\max_{P(u|y'),x=f(u,s')} I(U;Y|Y') = \max_{P(u|y'),x=f(u,s')\in\mathcal{P}_\pi} I(U;Y|Y'), \tag{14}$$

where $\mathcal{P}_\pi$ is defined as the set of $(P(u|y'), f)$ pairs that induce a unique stationary output distribution $\pi(y')$, and this set is non-empty.

The following theorem enables us to compute $C_{\text{FB-CSI}}$, since for any choice of $f(\cdot)$ the feedback capacity expression, $\max_{P(u|y')} I(U;Y|Y')$, can be formulated as a convex optimization problem similar to that of Theorem 2, yet with input U instead of X.

Theorem 4 (Convex Optimization for $C_{FB\text{-}CSI}$): For any $f : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$ with $|\mathcal{U}| = L$, the expression for the feedback capacity of a connected NOST channel with CSI available at the encoder, $C_{\text{FB-CSI}}$, given in (12), can be formulated as the following convex optimization problem:

$$\underset{P(y',u)\in\mathcal{P}(\mathcal{Y}\times\mathcal{U})}{\text{maximize}} \quad I(U;Y|Y') \tag{15a}$$

$$\text{subject to} \quad \sum_{u,y} P_{Y',U}(\tilde{y},u)P_f(y|u,\tilde{y}) - \sum_{y',u} P_{Y',U}(y',u)P_f(\tilde{y}|u,y') = 0, \quad \forall \tilde{y} \in \mathcal{Y}. \tag{15b}$$

where $P_f(y|u,y')$, given in (13), is determined by f and the NOST channel model.

As a consequence of Theorem 4, the feedback capacity $C_{\text{FB-CSI}}$ given in (12) can be readily computed, because the maximum over functions f that map $x = f(u,s')$ is equivalent to taking the maximum of the solutions of all $\binom{|\mathcal{X}|^{|\mathcal{S}|}}{L}$ convex optimization problems with $|\mathcal{U}| = L$.

In Section IV, we provide an example of a connected NOST channel whose C_{FB} and $C_{\text{FB-CSI}}$ given in the previous theorems are equal, and derive these capacity expressions there explicitly as detailed in Theorem 5.

IV. Examples

This section covers special cases of Setting I and Setting II, and shows how the feedback capacity expressions of each setting, C_{FB} and $C_{\text{FB-CSI}}$, subsume the corresponding capacity characterizations from the literature. Furthermore, we demonstrate the connectivity property (Definition 1) on the POST(α) channel [30]. Finally, we generalize this channel to one having a state that is stochastically dependent on the output, a noisy version we thus call "the noisy-POST(α,η) channel", and show that CSI available at the encoder does not increase its feedback capacity.

A. Special Cases of Setting I—No CSI

1) The State Is a Deterministic Function of the Output: The case where $s_i = y_i$ is trivially a special case of Setting I. As explained after Theorem 1, Setting I can also be formulated with the channel state y' and therefore, operationally, both settings are equivalent. In [39], the special case $s_i = y_i$ was studied, but Theorem 1 generalizes their result by relaxing the assumption in [39]. In particular, [39] shows that the capacity expression is the one in Theorem 1 but subject to *strong irreducibility* and *strong aperiodicity* [39, Defs. 2,4].[2] Our derivations do not require any aperiodicity assumption, and the strong irreducibility is particularly relaxed: recall that by (7), our connectivity condition holds if and only if

$$\forall i,j \in \mathcal{Y},\ \exists T(i,j):\ (\overbrace{Q\cdots Q}^{T(i,j)\text{times}})_{ij} > 0. \tag{16}$$

The strong irreducibility in [39, Def. 2] can be written similarly by changing the maximum in (7) to a minimum. In words, strong irreducibility requires irreducibility (in the usual sense) of the output Markov process $\{Y_i | i = 0,1,\dots\}$ with respect to *all* input distributions, while Definition 1 only requires the *existence* of an input distribution that induces a path (a positive probability) between any two channel outputs. We proceed to show the significance of this relaxation via the following Example a), then we provide Example b) of a periodic, connected NOST channel.

[2]More accurately, [39] also assumed an additional, unnecessary condition ([39, Def. 6]) just for simplifying the proof, as it was remarked there that it was not crucial for the feedback capacity theorem.

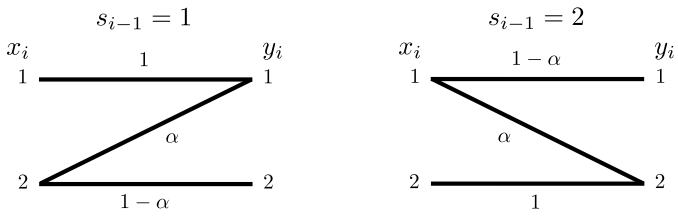

Fig. 2. The ZS-channel model characterizing the probability of $Q(y_i|x_i,s_{i-1})$, where $\alpha \in [0,1]$. For $s_{i-1} = 1$ we have the Z topology, and for $s_{i-1} = 2$ we have the S topology.

a) The POST(α) channel: The POST(α) channel studied in [30] is a simple, yet representative, example of an FSC. The alphabets $\mathcal{X},\mathcal{Y},\mathcal{S}$ are all binary, and the channel output depends on the input and the channel state as shown on Fig. 2. Specifically, if the input and the channel state are equal, the channel output is equal to them, while otherwise it is a random instance due to parameter $\alpha \in [0,1]$. The state evolution of the POST(α) channel is implied by its name, "Previous Output is the STate (POST)" [30], i.e., $s_i = y_i$. The POST(α) channel is not strongly irreducible under the definition of [39, Def. 2], but is a connected NOST channel under Definition 1, demonstrated as follows. For the POST(α) channel, Matrix Q defined in (7), and Matrix $\tilde{Q}$ defined by replacing the maximum in (7) with a minimum are, respectively,

$$Q = \begin{bmatrix} 1 & 1-\alpha \\ 1-\alpha & 1 \end{bmatrix}, \quad \tilde{Q} = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}.$$

On the one hand, for any power of $n = 1,2,\dots$, entries $(\tilde{Q}^n)_{12} = (\tilde{Q}^n)_{21} = 0$, i.e., there is no path from output $y' = 1$ to output $y = 2$ (and from $y' = 2$ to $y = 1$); and therefore, the POST(α) is not strongly irreducible. On the other hand, $Q_{ij} > 0$ for all $i,j \in \mathcal{Y}$ and $\alpha \in [0,1)$, and thus the POST(α) channel is connected (except for $\alpha = 1$, in which case the feedback capacity is trivially 0). Consequently, Theorem 1 recovers its known feedback capacity, which is the closed-form capacity expression of a simple Z channel, as was derived in [30].

It is compelling that many channel instances like the trapdoor [27], Ising [45] and POST(α) share the same channel characterization $Q(y|x,s')$. However, their feedback capacity is fundamentally different due to the channel state evolution. In Section IV-C, we generalize the POST(α) channel to have a stochastic state evolution, and study its feedback capacity with and without CSI available at the encoder. We now proceed to Example b) of a periodic connected NOST channel that does not satisfy strong aperiodicity [39, Def. 4].

b) A periodic NOST channel: Let $\mathcal{X} = \mathcal{S} = \{0,1\}$, $\mathcal{Y} = [0:3]$, where for both states a general binary-input binary-output channel (BIBO) is obtained, yet with different outputs, as given on the LHS of Fig. 3. The state s_i is a deterministic function of the output y_i, as given on the RHS of Fig. 3, and it induces a periodic Markov output process with period 2. Although the output Markov chain is periodic, Theorem 1 can determine the feedback capacity of

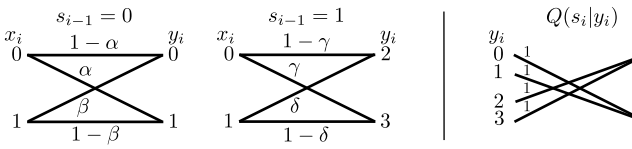

Fig. 3. An example of a periodic connected NOST channel. On the LHS: the conditional probabilities $Q(y_i|x_i, s_{i-1})$, i.e., for both $s_{i-1} = 0, 1$, general BIBO channels are obtained with some parameters $0 \leq \alpha, \beta, \gamma, \delta \leq 1$. On the RHS: the state evolution $Q(s_i|y_i)$, i.e., s_i is a deterministic function of y_i.

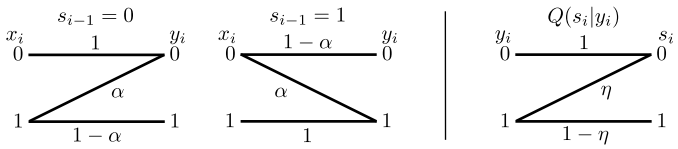

Fig. 4. The noisy-POST(α, η) channel. On the LHS: the ZS-channel model characterizing the probability of $Q(y_i|x_i, s_{i-1})$, where $\alpha \in [0, 1]$. On the RHS: the state evolution $Q(s_i|y_i)$ as the Z topology, where $\eta \in [0, 1]$.

this channel, because it is clearly a connected NOST channel. Denote the DMC capacities of the BIBOs in $s_{i-1} = 0, 1$ by C_1, C_2, respectively. Applying Theorem 1 gives that C_{FB} of this periodic channel example is the average of C_1 and C_2, because any stationary output distribution $\pi(y'), y' \in \mathcal{Y}$ satisfies $\pi_{Y'}(0) + \pi_{Y'}(1) = \pi_{Y'}(2) + \pi_{Y'}(3) = 0.5$; thus the feedback capacity is

$$\begin{aligned} C_{\text{FB}}^{\text{Per.}} &= \sum_{y' \in \mathcal{Y}} \pi(y') I(X; Y|Y' = y') \\ &= (\pi_{Y'}(0) + \pi_{Y'}(1))\, C_2 + (\pi_{Y'}(2) + \pi_{Y'}(3))\, C_1 \\ &= \frac{C_1 + C_2}{2}. \end{aligned} \tag{17}$$

2) The State Is Independent of the Output: In this special case, the channel state evolution satisfies $Q(s_i|y_i) = Q(s_i)$. Consequently, for this case, the averaged DMC $Q(y|x, y')$ in (3) does not depend on the previous channel input y', and can be written as $Q(y|x) \triangleq \sum_{s'} Q(s')Q(y|x, s')$ (which implies that X is independent of S'). The term for $Q(y|x)$ averages the DMCs $Q(y|x, s')$ over the state; the capacity is $C = \max_{P(x)} I(X; Y)$, and is not increased by feedback [17]. We show that C_{FB} is equal to C as follows. On the one hand, $Q(y|x, y') = Q(y|x)$ implies that $I(X; Y|Y') \leq I(X; Y)$, and on the other hand we have $I(X; Y|Y') \geq I(X; Y)$, which follows by considering $P(x|y') = P(x)$ as this implies that Y and Y' are independent due to $P(y|y') = \sum_x P(x|y')Q(y|y', x) = \sum_x P(x)Q(y|x) = P(y)$.

B. Special Cases of Setting II—CSI Available at the Encoder

We previously showed that Setting I (without CSI) was operationally equivalent to the setting where $s_i = y_i$ with feedback, by arguing that each one of them can be considered as a special case of the other. However, Setting II (CSI available at the encoder) cannot be considered a special case of the setting $s_i = y_i$ with feedback and CSI available at the encoder due to the following explanation. There is already a real state with a physical meaning, s', known at the encoder, and we cannot introduce a new fictitious state. When CSI is not available, it follows from Theorem 1 that the probability of Y_i given (X_i, Y_{i-1}) is determined by $Q(y|x, y')$ (3), i.e., it is fixed by the NOST channel model because of the Markov chain $S' - Y' - X$, which follows since the encoder does not have access to the states. However, this Markov chain does not necessarily hold when CSI is available at the encoder, and, therefore, the probability of Y_i given (X_i, Y_{i-1}) is given by $P(y|x, y')$ which is not fixed only by the NOST channel model, but also by the choice of an auxiliary RV U that maps the real state S' to a channel input X by some function $f : \mathcal{U} \times \mathcal{S} \to \mathcal{X}$, as shown in Theorem 3.

Another special case of Setting II is where the state is independent of the output, i.e., $Q(s_i|y_i) = Q(s_i)$. In this case, we obtain a new DMC, $P_f(y|u, y') = \sum_{s',x} Q(s')\mathbb{1}\{x = f(u, s')\}Q(y|x, s') = P_f(y|u)$, with input u and output y, as can be seen from (13). This implies that U and S' become independent. The capacity in this case was derived by Shannon [55] as $\max_{P(u), x(u,s')} I(U; Y)$, where U is, indeed, an auxiliary RV independent of S'. Feedback does not increase the capacity of DMCs, and it can be shown that $C_{\text{FB-CSI}}$ recovers Shannon's capacity expression by using the fact that $P_f(y|u, y') = P_f(y|u)$ and repeating the same arguments presented in Section IV-A.2 with U instead of X.

C. The Noisy-POST(α, η) Channel—Special Example for Which C_{FB} and $C_{FB\text{-}CSI}$ Are Equal

In this section, we introduce an interesting example of a NOST channel for which C_{FB} and $C_{\text{FB-CSI}}$ are equal, i.e., CSI available at the encoder does not increase its feedback capacity. This example is a generalization of the POST(α) channel, i.e., the channel output depends on the input and the channel state identically to the POST(α) channel, while the state evolution is generalized, as illustrated in Fig. 4. We emphasize that in all previous channel instances studied in the literature, such as the trapdoor, Ising and POST(α), the state evolves according to a deterministic rule and, thus, can be determined at the encoder, while here we focus on a noisy, new version of the POST(α) channel, in which the state evolves *stochastically* according to parameter $\eta \in [0, 1]$. In particular, the channel state depends on the output via a Z-channel, i.e., if the output is zero, the next state equals the channel output, and otherwise, the next state equals the output with probability $1 - \eta$. We call this generalized channel "the noisy-POST(α, η)"; note that when $\eta = 0$ we obtain the original "Previous Output is the STate (POST)" [30] channel. Similarly to the demonstration of the connectivity on the POST(α) in Section IV-A, it can be verified that the noisy-POST(α, η) channel is also connected under Definition 1.

In the remainder of this section, we study the feedback capacity of this noisy-POST(α, η) channel with or without CSI available at the encoder, denoted by $C_{\text{FB-CSI}}^{\text{N-POST}}(\alpha, \eta)$ and $C_{\text{FB}}^{\text{N-POST}}(\alpha, \eta)$, respectively. For simplicity, we arbitrarily focus on the case of $\alpha = 0.5, \eta \in [0, 1]$ as summarized in the following theorem, and analyze it.

Theorem 5: For the noisy-POST(α, η) channel with any $\alpha, \eta \in [0, 1]$, CSI available at the encoder does not increase

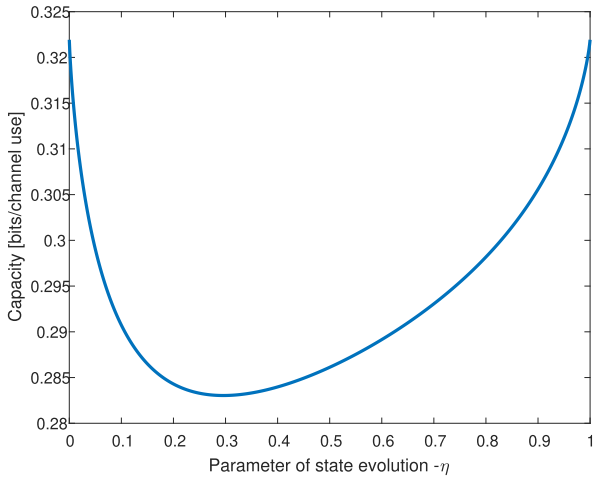

Fig. 5. The feedback capacity of the noisy-POST$(0.5, \eta)$ channel with or without CSI available at the encoder.

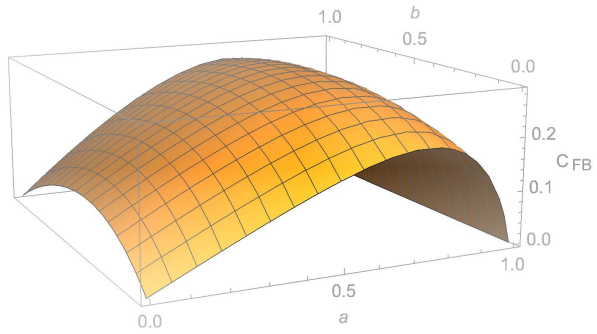

Fig. 6. The objective function of (18) evaluated for all values of $a, b \in [0, 1]$ and arbitrary $\eta = 0.5$.

the feedback capacity, and

$$C_{\text{FB}}^{\text{N-POST}}(0.5, \eta) = \max_{a,b\in[0,1]} \frac{b+\eta}{a+b+\eta}\left(H\left(\frac{a}{2}\right) - a\right) + \frac{a}{a+b+\eta}\left(H\left(\frac{b+\eta}{2}\right) - bH\left(\frac{\overline{\eta}}{2}\right) - \overline{b}H\left(\frac{\eta}{2}\right)\right). \quad (18)$$

The first result of Theorem 5, i.e., $C_{\text{FB-CSI}}^{\text{N-POST}}(\alpha, \eta) = C_{\text{FB}}^{\text{N-POST}}(\alpha, \eta)$, is proved at the end of this section. The second result of Theorem 5, i.e., Eq. (18), follows straightforwardly from applying Theorem 1 on the noisy-POST$(0.5, \eta)$ channel where $a \triangleq P_{X|Y'}(1|0)$ and $b \triangleq P_{X|Y'}(0|1)$ are the optimization variables; its derivation is tedious and thus is omitted.

In Fig. 5, $C_{\text{FB}}^{\text{N-POST}}(\alpha, \eta)$ is evaluated for $\eta \in [0, 1]$ using the convex optimization problem in Theorem 2. It can be seen that it is a convex function of η. In particular, for $\eta = 0$, the POST(0.5) channel is obtained, and for $\eta = 1$ the Z-channel with parameter 0.5 is obtained; in both cases, the feedback capacity is $-\log_2(0.8) \approx 0.3219$. In the case where $\eta \in (0, 1)$, it can be seen that the feedback capacity is less than the capacity of the Z-channel. This reflects the rate-loss due to the fact that state is known at the encoder, but not at the decoder.

For the special case $\eta = 0$, the feedback capacity is achieved with $a = b = 0.4$, and for $\eta = 1$ it is achieves with $a = 0.4, b = 0.6$. For general $\eta \neq 0, 1$, deriving a simpler capacity expression than (18) is challenging. In Fig. 6, we evaluate the objective function of (18) as a function of the optimization variables a and b, and $\eta = 0.5$. It is interesting to note that although we prove the concavity of the feedback capacity in $P(y', x)$ in Theorem 2, Fig. 6 suggests that the feedback capacity of the noisy-POST$(0.5, 0.5)$ is also a concave function of $P(x|y')$. A similar phenomenon is observed for other values of $\eta \in (0, 1)$ as well.

TABLE I

ALL THE STRATEGIES OF BINARY INPUT AND BINARY STATE ALPHABETS, $\mathcal{S} = \mathcal{X} = \{0, 1\}$

$x(u, s')$	$s' = 0$	$s' = 1$
u_0	0	0
u_1	0	1
u_2	1	0
u_3	1	1

Next, we prove Theorem 5.

Proof of Theorem 5: We prove here that CSI at the encoder does not increase the feedback capacity of the noisy-POST(α, η) channel, i.e., $C_{\text{FB}}^{\text{N-POST}}(\alpha, \eta) = C_{\text{FB-CSI}}^{\text{N-POST}}(\alpha, \eta)$. Consider $|\mathcal{U}| = |\mathcal{X}|^{|\mathcal{S}|} = 4$ with all possible strategies as detailed in Table I. By Theorem 3, assume that $I_{P_1}(U; Y|Y')$ is the feedback capacity of the noisy-POST(α, η) channel with CSI available at the encoder, induced by some input distribution $P_1(u|y')$ with the corresponding joint distribution

$$P_1(y', u, x, y) = \pi_1(y') \sum_{s'} Q(s'|y') P_1(u|y') \mathbb{1}\{x = f(u, s')\} Q(y|x, s'),$$

where $\pi_1(y')$ is a stationary output distribution induced from the conditional output distribution $P_1(y|y')$. We construct an input distribution $P_2(x|y')$ with the corresponding conditional mutual information satisfying $I_{P_2}(X; Y|Y') = I_{P_1}(U; Y|Y')$ induced by the joint distribution

$$P_2(y', x, y) = \pi_2(y') P(x|y') Q(y|x, y'),$$

where $\pi_2(y')$ is a stationary output distribution induced from the conditional output distribution $P_2(y|y')$ (see Theorem 1). Clearly, $I_{P_1}(U; Y|Y') \geq I_{P_2}(X; Y|Y')$; thus, our goal is to show that $I_{P_1}(U; Y|Y') \leq I_{P_2}(X; Y|Y')$. In the construction of $P_2(x|y')$, we only demand that it satisfies

$$P_2(x|y') = P_1(x|y') \quad \forall x \in \mathcal{X}, y' \in \mathcal{Y}, \quad (19)$$

where $P_1(x|y')$ is the input distribution induced by $P_1(u|y')$, and given by

$$P_1(x|y') = \sum_{u,s'} P_1(u, s', x|y') = \sum_{u,s'} P_1(u|y') Q(s'|y') \mathbb{1}\{x = f(u, s')\}.$$

Hence, for the noisy-POST(α, η) we obtain

$$P_2(X = 1|Y' = 0) \triangleq P_1(u_2|Y' = 0) + P_1(u_3|Y' = 0), \quad (20)$$

$$P_2(X = 0|Y' = 1) \triangleq P_1(u_0|Y' = 1) + \eta P_1(u_1|Y' = 1) + (1 - \eta) P_1(u_2|Y' = 1). \quad (21)$$

From the construction in (19), it follows that the conditional output distributions are also equal, i.e.,

$$\begin{aligned} P_2(y|y') &= \sum_x P_2(x|y')Q(y|x,y') \\ &= \sum_x P_1(x|y')Q(y|x,y') \\ &= P_1(y|y'), \end{aligned}$$

for all $y', y \in \mathcal{Y}$. Consequently, $\pi_2(y') = \pi_1(y'),\ \forall y' \in \mathcal{Y}$ and $H_{P_2}(Y|Y') = H_{P_1}(Y|Y')$ hold; thus,

$$\begin{aligned} I_{P_1}(U;Y|Y') &= H_{P_2}(Y|Y') - H_{P_1}(Y|Y',U) \\ &\overset{(a)}{\leq} H_{P_2}(Y|Y') - H_{P_2}(Y|Y',X) \\ &= I_{P_2}(X;Y|Y'), \end{aligned}$$

where (a) follows from defining $q \triangleq H_{P_1}(Y|Y',U) - H_{P_2}(Y|Y',X) \geq 0$. We show that $q \geq 0$ by applying the noisy-POST(α,η) channel model on

$$\begin{aligned} H_{P_1}(Y|Y',U) &= \sum_{y'} \pi_1(y') H_{P_1}(Y|Y'=y',U), \\ H_{P_2}(Y|Y',X) &= \sum_{y'} \pi_2(y') H_{P_2}(Y|Y'=y',X) \\ &= \sum_{y'} \pi_1(y') H_{P_2}(Y|Y'=y',X), \\ P_f(y|u,y') &= \sum_{s',x} Q(s'|y') \mathbb{1}\{x = f(u,s')\} Q(y|x,s'), \end{aligned}$$

giving the following identities:

$$\begin{aligned} &H_{P_1}(Y|y'=0,U) \\ &= P_1(X=0|Y'=0) \\ &\overset{(a)}{=} P_2(X=0|Y'=0) \\ &= H_{P_2}(Y|y'=0,X), \\ &H_{P_1}(Y|y'=1,U) \\ &= P_1(u_0|y'=1)H(\tfrac{1-\eta}{2}) + P_1(u_1|y'=1)H(\eta) \\ &\quad + P_1(u_2|y'=1) + P_1(u_3|y'=1)H(\tfrac{\eta}{2}) \\ &\overset{(b)}{=} P_1(u_0|y'=1)\left(H(\tfrac{1-\eta}{2}) - H(\tfrac{\eta}{2})\right) \\ &\quad + P_1(u_1|y'=1)\left(H(\eta) - H(\tfrac{\eta}{2})\right) \\ &\quad + P_1(u_2|y'=1)\left(1 - H(\tfrac{\eta}{2})\right) + H(\tfrac{\eta}{2}), \\ &H_{P_2}(Y|y'=1,U) \\ &= P_2(X=0|Y'=1)H(\tfrac{1-\eta}{2}) \\ &\quad + P_2(X=1|Y'=1)H(\tfrac{\eta}{2}) \\ &\overset{(c)}{=} P_2(X=0|Y'=1)H(\tfrac{1-\eta}{2}) \\ &\quad + (1 - P_2(X=0|Y'=1))\,H(\tfrac{\eta}{2}) \\ &\overset{(d)}{=} P_1(u_0|y'=1)\left(H(\tfrac{1-\eta}{2}) - H(\tfrac{\eta}{2})\right) \\ &\quad + P_1(u_1|y'=1)\eta\left(H(\tfrac{1-\eta}{2}) - H(\tfrac{\eta}{2})\right) \\ &\quad + P_1(u_2|y'=1)(1-\eta)\left(H(\tfrac{1-\eta}{2}) - H(\tfrac{\eta}{2})\right) + H(\tfrac{\eta}{2}), \end{aligned}$$

where (a) and (d) follow from the construction of $P_2(x|y')$ in (20)-(21); and (b) and (c) follow from substituting $P_1(u_3|y'=1)$ and $P_2(X=1|Y'=1)$, respectively, with their complementary distribution to 1. Hence, we deduce that

$$\begin{aligned} q &= P_1(Y'=1) \\ &\quad \times \left(H_{P_1}(Y|y'=1,U) - H_{P_2}(Y|y'=1,X)\right) \geq 0, \end{aligned}$$

because

$$\begin{aligned} &H_{P_1}(Y|y'=1,U) - H_{P_2}(Y|y'=1,X) \\ &= P_1(u_1|y'=1)\left(H(\eta) - (1-\eta)H(\tfrac{\eta}{2}) - \eta H(\tfrac{1-\eta}{2})\right) \\ &\quad + P_1(u_2|y'=1)(1 - \eta H(\tfrac{\eta}{2}) - (1-\eta)H(\tfrac{1-\eta}{2})) \\ &\geq P_1(u_1|y'=1)\left(H(\eta) - (1-\eta)H(\tfrac{\eta}{2}) - \eta H(\tfrac{1-\eta}{2})\right) \\ &\overset{(a)}{\geq} P_1(u_1|y'=1)\left(H(\eta) - H(\eta(1-\eta))\right) \\ &\overset{(b)}{\geq} 0, \end{aligned}$$

where (a) follows from the concavity of the binary entropy, and (b) is due to $H(\eta) \geq H(\eta(1-\eta))$, which is trivial for $\eta \in [0, 0.5]$, and for $\eta \in [0.5, 1]$ it is also trivial after using $H(\eta) = H(1-\eta)$.

To conclude, $I_{P_2}(X;Y|Y') = I_{P_1}(U;Y|Y')$, which implies that CSI available at the encoder does not increase the feedback capacity of the noisy-POST(α,η) channel. □

V. Proofs

In this section, we prove our main results given in Section III. In particular, the proofs of the feedback capacity expression, i.e., Theorems 1 and 3, are given in Sections V-A and V-B, respectively. We note that Lemmas 1 and 2 are used to establish the achievability proofs of the mentioned Theorems 1 and 3, respectively. As Lemma 2 generalizes Lemma 1, we only prove the former in Section V-C. The proof of the cardinality bound of Theorem 3 is provided in Section V-D. Finally, Section V-E proves the convex optimization formulations of the feedback capacity expressions, i.e., Theorems 2 and 4. Before all these proofs are given, we introduce the following useful lemma, whose proof is given in Appendix A.

Lemma 3: For any NOST channel (2) in Setting I (without CSI),

$$\begin{aligned} Q(y_i|x^i, y^{i-1}, m) &= \sum_{s_{i-1} \in \mathcal{S}} Q(s_{i-1}|y_{i-1}) Q(y_i|x_i, s_{i-1}) \\ &= Q(y_i|x_i, y_{i-1}). \end{aligned} \tag{22}$$

A. Proof of Theorem 1

1) Proof of Converse: Throughout the proof, the initial output, y_0, is assumed to be available at both the encoder and the decoder.

For a fixed sequence of $(2^{nR}, n)$ codes, where R is an achievable rate, we bound R as

$$\begin{aligned} R - \epsilon_n &\overset{(a)}{\leq} \frac{1}{n} I(M;Y^n) \\ &\overset{(b)}{=} \frac{1}{n} \sum_{i=1}^{n} I(M, X_i; Y_i|Y^{i-1}) \end{aligned}$$

$$\overset{(c)}{\leq} \frac{1}{n}\sum_{i=1}^{n} H(Y_i|Y_{i-1}) - H(Y_i|Y_{i-1}, X_i)$$
$$\overset{(d)}{\leq} \max_{\{P(x_i|y_{i-1})\}_{i=1}^{n}} \frac{1}{n}\sum_{i=1}^{n} I(X_i; Y_i|Y_{i-1}), \tag{23}$$

where ϵ_n tends to zero as $n \to \infty$, and

(a) follows from Fano's inequality;
(b) follows from the fact that x_i is a deterministic function of (m, y^{i-1}); and
(c) follows from the fact that conditioning reduces entropy and from Lemma 3;
(d) follows from the fact that for any k, the joint distribution $P(y_{k-1}, x_k, y_k) = P(y_{k-1})P(x_k|y_{k-1})Q(y_k|x_k, y_{k-1})$ is determined by $\{P(x_i|y_{i-1})\}_{i=1}^{k}$, which can be shown by induction.

In Section V-B, we show a fundamental result on the optimality of time-invariant input distributions in Lemma 4. To avoid repetition, we refer the reader to follow the proof of Lemma 4 with x instead of u, the joint distribution $P(y', x, y) = P(y', x)Q(y|x, y')$ and the modified set

$$\mathcal{D}_\epsilon \triangleq \{P(y', x) \in \mathcal{P}_{\mathcal{Y}\times\mathcal{X}} : |P_{Y'}(y) - \sum_{y', x} P_{Y', X}(y', x)Q(y|x, y')| \leq \epsilon, \forall y\}, \tag{24}$$

in order to deduce that any achievable rate R must satisfy $R \leq \max_{P(x|y')} I(X; Y|Y')$. □

2) Proof of Achievability: We need to prove that rates satisfying $R < \max_{P(x|y')} I(X; Y|Y')$ are achievable. Nevertheless, by recalling Lemma 1, which states that it is sufficient to maximize over input distributions that induce a unique stationary output distribution, we prove, for simplicity, that rates satisfying $R < \max_{P(x|y')\in\mathcal{P}_\pi} I(X; Y|Y')$ are achievable. The proof uses rate-splitting where Y^n is treated as a time-sharing sequence, and the previous channel output, $Y_{i-1} = y' \in \mathcal{Y}$, determines one of $|\mathcal{Y}|$ DMCs that are multiplexed at the encoder and demultiplexed at the decoder.

Proof: At time $i = 1$, the encoder transmits an arbitrary input symbol X_1, and afterwards Y_1 is known both at the decoder and at the encoder by the feedback. More generally, from time $i = 2$ on, the previous channel output Y_{i-1} is known at both parties before each transmission. According to the known $Y_{i-1} = y'$, a DMC characterized by $Q_{Y|X} = Q_{Y|X,Y'=y'}$ with message $M_{y'}$ is treated in the current channel use.

a) Rate-splitting and code construction: Fix an input distribution $P(x|y') \in \mathcal{P}_\pi$ that achieves C_{FB} in (9), i.e., a collection of conditional PMFs $P(x|y')$ on $\mathcal{X}$ for every $y' \in \mathcal{Y}$ is to be determined such that a unique stationary distribution on the outputs, $\pi(y')$, is induced. By Lemma 1, such $P(x|y')$ always exists on account of the connectivity assumption (Definition 1) and the assumption that $|\mathcal{Y}|$ is finite. Each message M consists of $|\mathcal{Y}|$ independent sub-messages $M_{y'} \in [1 : 2^{nR_{y'}}]$, $y' \in \mathcal{Y}$. This implies that $R = \sum_{y'} R_{y'}$. From the achievability of the channel coding theorem for DMCs, in each DMC $Q_{Y|X,Y'=y'}, y' \in \mathcal{Y}$, every rate $R_{y'} < I(X; Y|Y' = y')$ is achievable, where the joint distribution is determined by the fixed conditional input distribution given y', i.e., $P(x, y|y') = P(x|y')Q(y|x, y')$. That is, there exists a sequence of $(2^{nR_{y'}}, n)$ codes with an average probability of error $P(\hat{M}_{y'} \neq M_{y'})$ that tends to zero as $n \to \infty$. For a block length n and $y' \in \mathcal{Y}$, denote the codebook of the nth code of such a sequence by $\mathcal{C}_{n,y'}$. Each $\mathcal{C}_{n,y'}$ consists of $2^{nR_{y'}}$ codewords $x^n(m_{y'})$.

Returning to our connected NOST channel: to send message $m = \{m_{y'}|y' \in \mathcal{Y}\}$, at time $i \in [2, n+1]$, with known previous output $Y_{i-1} = y'$, the encoder transmits the next unsent symbol of codeword $x^n(m_{y'}) \in \mathcal{C}_{n,y'}$. Upon receiving the entire output sequence y^{n+1}, the decoder demultiplexes it into $|\mathcal{Y}|$ sub-sequences of outputs $y^{n_{y'}}(y')$ whose previous output is $y' \in \mathcal{Y}$, where $n_{y'}$ is the number of times that output y' was "visited" during times $i \in [1 : n]$, i.e.,

$$n_{y'} = \sum_{i=2}^{n+1} \mathbb{1}\{Y_{i-1} = y'\}, \tag{25}$$

thus $\sum_{y'} n_{y'} = n$. For each sub-sequence $y^{n_{y'}}(y'), y' \in \mathcal{Y}$ of DMC $Q_{Y|X,Y=y'}$, the receiver decodes $m_{y'}$ as in the aforementioned direct coding theorem for DMCs (joint typicality decoding).

b) Analysis of the probability of error: Using this theorem, it follows that the probability of error in decoding each $m_{y'}$ tends to zero as $n \to \infty$ if

$$R_{y'} \leq \lim_{n\to\infty} \frac{n_{y'}}{n} I(X; Y|Y' = y') - \delta(\epsilon),$$
$$\overset{(a)}{=} \pi(y')I(X; Y|Y' = y') - \delta(\epsilon), \tag{26}$$

where $\delta(\epsilon)$ tends to zero as $\epsilon \to 0$. Step (a) follows from Birkhoff's ergodic theorem on Markov chains with a unique stationary distribution (see, e.g., [68]), since the fixed $P(x|y')$ induces a homogeneous Markov chain $\{Y_i|i = 0, 1, \dots\}$ with a unique stationary output distribution, $\pi(y')$. Now, the total probability of error in decoding the message $m = \{m_{y'}|y' \in \mathcal{Y}\}$ tends to zero as $n \to \infty$ if

$$R = \sum_{y'} R_{y'}$$
$$\leq \sum_{y'} \pi(y')I(X; Y|Y' = y') - \tilde{\delta}(\epsilon)$$
$$= I(X; Y|Y') - \tilde{\delta}(\epsilon), \tag{27}$$

where $\tilde{\delta}(\epsilon)$ tends to zero as $\epsilon \to 0$. This completes the proof of achievability. □

We note that Theorem 1 can be derived by another approach based on the directed information, which generally characterizes the capacity of channels with feedback, as given in Appendix B.

B. Proof of Theorem 3

1) Proof of Converse: Here, we prove that an achievable rate R must satisfy $R \leq \max_{P(u|y')} I(U; Y|Y')$, where, without loss of generality, $\mathcal{U}$ is the set of all strategies. In this proof, the initial output, y_0, is assumed to be available at both the encoder and the decoder, and the proof consists of two

parts. In the first part, we show that for achievable rates

$$R \leq \max_{\{P(u_i|y_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) + \epsilon_n, \tag{28}$$

where $U_i \in \mathcal{U}$ enumerates all possible strategies and maps S_{i-1} to X_i, and $\epsilon_n \to 0$ as $n \to \infty$. The second part of the proof, stated by the following Lemma 4, is to show that it is sufficient to maximize over time-invariant conditional distributions, $P(u|y')$.

Lemma 4: For a connected NOST channel with CSI available at the encoder,

$$\lim_{n\to\infty} \max_{\{P(u_i|y_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) \leq \max_{P(u|y')} I(U;Y|Y'), \tag{29}$$

where $U_i \in \mathcal{U}$ enumerates all possible mappings from $\mathcal{S}$ to $\mathcal{X}$, the joint distribution on the RHS is $P(y',u,y) = \pi(y')P(u|y')P_f(y|u,y')$, and $P_f(y|u,y')$ is given in (13).

The proof of Lemma 4 is based on a method developed in [54], and it is given next in the second part of Section V-B. The first part of the converse, i.e., Inequality (28), is now shown.

For a fixed sequence of $(2^{nR},n)$ codes such that the probability of error $P_e^{(n)} \to 0$ as $n \to \infty$, we bound

$$\begin{aligned}
&R - \epsilon_n \\
&\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^n H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},M) \\
&\leq \frac{1}{n}\sum_{i=1}^n H(Y_i|Y_{i-1}) - H(Y_i|Y^{i-1},M) \\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) \\
&\overset{(c)}{\leq} \max_{\{P(u_i|y_{i-1}),P(x_i|u_i,s_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) \\
&\overset{(d)}{=} \max_{\{P(u_i|y_{i-1}),P(v_i),x_i=f_i(u_i,v_i,s_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) \\
&\overset{(e)}{\leq} \max_{\{P(\tilde{u}_i|y_{i-1}),x_i=f_i(\tilde{u}_i,s_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(\tilde{U}_i;Y_i|Y_{i-1}) \\
&\overset{(f)}{=} \max_{\{P(\tilde{u}_i|y_{i-1}),x_i=f(s_{i-1},\tilde{u}_i,i)\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(\tilde{U}_i;Y_i|Y_{i-1}) \\
&\overset{(g)}{\leq} \max_{\{P(\tilde{\tilde{u}}_i|y_{i-1}),x_i=f(\tilde{\tilde{u}}_i,s_{i-1})\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n I(\tilde{\tilde{U}}_i;Y_i|Y_{i-1})
\end{aligned} \tag{30}$$

where $\epsilon_n \to 0$ as $n \to \infty$, and

(a) follows from Fano's inequality;

(b) follows from defining $U_i \triangleq (M, Y^{i-1})$ for every $i \in [1:n]$; this definition also satisfies the Markov chain $(Y_i,S_i) - (X_i,S_{i-1}) - U_i$ due to the assumption that the channel is an FSC;

(c) follows from the following lemma, whose proof is given in the next part of Section V-B:

Lemma 5: For any k, the joint distribution $P(y_{k-1},u_k,y_k)$ is determined by

$$\{P(u_i|y_{i-1})P(x_i|u_i,s_{i-1})\}_{i=1}^k;$$

(d) follows from the Functional Representation Lemma [69], i.e., for every $i \in [1:n]$ there exists a RV V_i, such that X_i can be represented as a function of (U_i,S_{i-1},V_i), where V_i is independent of (U_i,S_{i-1}), and the Markov chain $(Y_i,S_i) - (U_i,S_{i-1},X_i) - V_i$ holds (hence $(Y_i,S_i) - (X_i,S_{i-1}) - (V_i,U_i)$ holds as well), and from the following lemma, whose proof uses the aforementioned properties of V_i and is similar to that of Lemma 5, and therefore it is omitted:

Lemma 6: For any k, the joint distribution $P(y_{k-1},u_k,y_k)$ is determined by

$$\{P(u_i|y_{i-1})P(v_i)x_i(v_i,u_i,s_{i-1})\}_{i=1}^k;$$

(e) follows from defining $\tilde{U}_i \triangleq (U_i,V_i)$; hence,

$$P(\tilde{u}_i|y_{i-1}) = P(v_i|y_{i-1})P(u_i|v_i,y_{i-1}),$$

as $P(u_i|y_{i-1})$ and $P(v_i)$ are sub-domains of $P(u_i|v_i,y_{i-1})$ and $P(v_i|y_{i-1})$, respectively;

(f) follows since there exists a time-invariant function f such that $f(\tilde{u},s,i) = f_i(\tilde{u}_i,s_{i-1})$; and

(g) follows from defining $\tilde{\tilde{U}} = (\tilde{U}_i, T=i)$, where T represents the time index.

For simplicity of appearance, we replace $\tilde{\tilde{U}}_i$ with U_i and obtain from (30) that any achievable rate must satisfy (28), where $\mathcal{U}$ is the aforementioned set of all strategies, i.e., $|\mathcal{U}| = |\mathcal{X}|^{|\mathcal{S}|}$ (increasing the cardinality of $\mathcal{U}$ beyond $|\mathcal{X}|^{|\mathcal{S}|}$ cannot increase the objective function further). Finally, the proof is completed by Lemma 4. □

2) Proofs of Technical Lemmas 4-5:

Proof of Lemma 4: The proof is divided into two parts. In the first part, $\frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1})$ is upper bounded for any n and joint distribution on (U^n,Y^n). Subsequently, in the second part of the proof we take the limit of this bound when n tends to infinity in order to obtain (29).

The first part of the proof is as follows. For any n and $\{P(y_{i-1},u_i)\}_{i=1}^n$,

$$\frac{1}{n}\sum_{i=1}^n I(U_i;Y_i|Y_{i-1}) \overset{(a)}{\leq} I(U;Y|Y') \overset{(b)}{\leq} \max_{P\in\mathcal{D}_{\frac{1}{n}}} I(U;Y|Y'), \tag{31}$$

where for Steps

(a) the joint distribution on the RHS is $\tilde{P}(y',u,y) = \tilde{P}(y',u)\sum_{s',x} Q(s'|y')\mathbb{1}\{x = f(u,s')\}Q(y|x,s')$, in which

$$\tilde{P}(y',u) \triangleq \frac{1}{n}\sum_{i=1}^n P_{Y_{i-1},U_i}(y',u), \tag{32}$$

$\mathcal{U}$ is defined identically to all $\mathcal{U}_i$, i.e., it is the set of all strategies mapping s' to x by the deterministic

function f, and this step follows from the fact that $I(U;Y|Y')$ is concave in the joint distribution $P(y',u)$ as is explained in particular in the proof of Theorem 4 given in Section V-E;

(b) the notation $\mathcal{D}_\epsilon$ denotes the set

$$\mathcal{D}_\epsilon \triangleq \{P(y',u) \in \mathcal{P}_{\mathcal{Y}\times\mathcal{U}} : |P_{Y'}(y) - \sum_{y',u} P_{Y',U}(y',u)Q(y|u,y')| \le \epsilon, \forall y\}, \quad (33)$$

where

$$Q(y|u,y') \triangleq \sum_{s',x} Q(s'|y')\mathbb{1}\{x = f(u,s')\}Q(y|x,s'),$$

and for any codebook of length n, its induced probability, $\tilde{P}(y',u)$, lies in $\mathcal{D}_{\frac{1}{n}}$, i.e., $|\tilde{P}_{Y'}(y) - \sum_{y',u}\tilde{P}_{Y',U}(y',u)Q(y|u,y')| \le \frac{1}{n}$ for all y, because by using the definition of $\tilde{P}(y',u)$ given in (32) we obtain

$$\begin{aligned}
&|\tilde{P}_{Y'}(y) - \sum_{y',u}\tilde{P}_{Y',U}(y',u)Q(y|u,y')| \\
=&\frac{1}{n}|\sum_{y',u}\sum_{i=1}^{n} P_{Y_{i-1}}(y) - P_{Y_{i-1},U_i}(y',u)Q(y|u,y')| \\
=&\frac{1}{n}|\sum_{i=1}^{n}\sum_{y',u} P_{Y_{i-1}}(y) - P_{Y_{i-1},U_i}(y',u)Q(y|u,y')| \\
=&\frac{1}{n}|\sum_{i=2}^{n}\sum_{y',u} P_{Y_{i-1}}(y) - P_{Y_{i-2},U_{i-1}}(y',u)Q(y|u,y') \\
&+ P_{Y_0}(y) - P_{Y_{n-1},U_n}(y',u)Q(y|u,y')| \\
=&\frac{1}{n}\,|P_{Y_0}(y) - P_{Y_n}(y)| \\
\le&\frac{1}{n}, \qquad (34)
\end{aligned}$$

which follows from $\sum_{y',u} P_{Y_{i-1},U_i}(y',u)Q(y|u,y') = P_{Y_i}(y)$ for any $i \in [1:n]$.

This completes the first part of the proof. In the second part of the proof, we obtain from (31)

$$\begin{aligned}
&\lim_{n\to\infty}\max_{\{P(u_i|y_{i-1})\}_{i=1}^n}\frac{1}{n}\sum_{i=1}^{n} I(U_i;Y_i|Y_{i-1}) \\
&\le \lim_{n\to\infty}\max_{P\in\mathcal{D}_{\frac{1}{n}}} I(U;Y|Y') \\
&\overset{(a)}{=} \max_{P\in\mathcal{D}_0} I(U;Y|Y') \\
&\overset{(b)}{=} \max_{P(u|y')} I(U;Y|Y'), \qquad (35)
\end{aligned}$$

where

(a) follows since any $P \in \mathcal{D}_0$ satisfies $P \in \cap_{n=1}^{\infty}\mathcal{D}_{\frac{1}{n}}$ due to the fact that $\frac{1}{n}$ is positive for all n, and vice versa, i.e., any $P \in \cap_{n=1}^{\infty}\mathcal{D}_{\frac{1}{n}}$ satisfies $P \in \mathcal{D}_0$ because $\frac{1}{n}$ monotonically decreases in n; hence, $\lim_{n\to\infty}\mathcal{D}_{\frac{1}{n}} = \mathcal{D}_0$;

(b) follows since $\mathcal{D}_0$ implies the set of all $P(y',u)$ that have a stationary output distribution, i.e., $P_{Y'}(y') = P_Y(y')$; recall that since the output set $\mathcal{Y}$ is assumed to be finite, there always exists a stationary output distribution (not necessarily unique) with regard to any $P(u|y')$, thus $\mathcal{D}_0$ is non-empty.

This concludes the proof. □

Proof of Lemma 5: We prove by induction that the joint distribution $P(y_{k-1},u_k,y_k)$ is determined by $\{P(u_i|y_{i-1})P(x_i|u_i,s_{i-1})\}_{i=1}^k$, where $U_i \triangleq (M,Y^{i-1},y_0)$ and y_0 is assumed to be known at both the encoder and the decoder. For $k=1$,

$$\begin{aligned}
&P(u_1,y_1|y_0) \\
&= \sum_{s_0,x_1} P(s_0,u_1,x_1,y_1|y_0) \\
&= \sum_{s_0,x_1} Q(s_0)P(u_1|y_0)P(x_1|u_1,s_0)Q(y_1|x_1,s_0),
\end{aligned}$$

which follows from the facts that: $U_1 = (M,y_0)$, where M and S_0 are independent and $Y_1 - (X_1,S_0) - M$ forms a Markov chain due to the FSC Markov property (1). Suppose that the lemma is true for $k-1$, i.e., $P(y_{k-2},u_{k-1},y_{k-1})$ is determined by $\{P(u_i|y_{i-1}),P(x_i|u_i,s_{i-1})\}_{i=1}^{k-1}$. Then, for k we have

$$\begin{aligned}
&P(y_{k-1},u_k,y_k) \\
&= \sum_{s_{k-1},x_k} P(y_{k-1},s_{k-1},u_k,x_k,y_k) \\
&= \sum_{s_{k-1},x_k} P(y_{k-1})Q(s_{k-1}|y_{k-1}) \\
&\times P(u_k|y_{k-1})P(x_k|u_k,s_{k-1})Q(y_k|x_k,s_{k-1}), \qquad (36)
\end{aligned}$$

which follows from the NOST channel Markov property (2) and the definition of U_k. From the induction hypothesis, $P(y_{k-1})$ is determined by $\{P(u_i|y_{i-1}),P(x_i|u_i,s_{i-1})\}_{i=1}^{k-1}$. Hence, from (36) it can be seen that $P(y_{k-1},u_k,y_k)$ is determined by $\{P(u_i|y_{i-1}),P(x_i|u_i,s_{i-1})\}_{i=1}^{k}$, which completes the proof. □

3) Proof of Achievability: We prove that every rate $R < \max_{P(u|y')} I(U;Y|Y')$, where $\mathcal{U}$ is the set of all strategies, is achievable. This is shown by converting Setting II into a setting of type I, i.e., where no CSI is available, as is shown in Fig. 7. In particular, at time $i \in [2:n]$ (the communication setting starts at time $i=2$ for the same reason given in the proof of achievability of Setting I), the channel input is U_i, which is a function of the message and feedback only, without the state, i.e., $U_i(M,Y^{i-1})$. Then, given the current state S_{i-1}, the strategy U_i maps S_{i-1} to input X_i, thus inducing a new NOST channel, $Q(y|u,s')Q(s|y)$, with input U_i (rather than X_i), in the presence of feedback. The new NOST channel is also connected, because $\mathcal{U}$ specifically includes all $|\mathcal{X}|$ strategies that map all states to an input $x \in \mathcal{X}$. This allows us to use the achievability of Theorem 1 (in which Lemma 2 should be used instead of Lemma 1) and deduce that rates that satisfy $R < \max_{P(u|y')} I(U;Y|Y')$ are achievable. □

C. Proofs of Lemma 1 and Lemma 2

Here, we prove Lemma 2 which also generalizes Lemma 1.

Proof of Lemma 2: Without loss of generality, we assume that $\mathcal{U}$ is the set of all strategies, thus all strategies are chosen

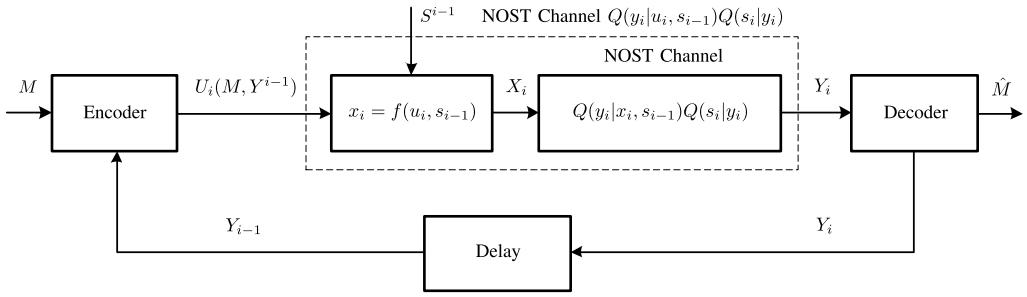

Fig. 7. An equivalent setting for the lower bound formulation with a new NOST channel $Q(y|u,s')Q(s|y)$ where CSI is unavailable.

and $|\mathcal{U}|=|\mathcal{X}|^{|\mathcal{S}|}$. We need to prove that

$$\max_{P(u|y')} I(U;Y|Y') = \max_{P(u|y')\in\mathcal{P}_\pi} I(U;Y|Y'), \qquad (37)$$

where $\mathcal{P}_\pi$ denotes the non-empty set of input distributions $P(u|y')$ that induce a unique stationary output distribution. We prove it by constructing some $P(u|y')\in\mathcal{P}_\pi$ that achieves the maximum on the LHS of (37).

Since $\mathcal{Y}$ is assumed to be a finite set, any input distribution induces at least one stationary output distribution (see, e.g., [70, p. 239]). Let $P^*(u|y')$ be an optimal input distribution that achieves the maximum on the LHS of (37), denoted by $I^*(U;Y|Y')$, and induces at least one stationary output distribution, i.e., $P^*_{Y'}(y')=P^*_Y(y'),\forall y'\in\mathcal{Y}$. If $P^*(u|y')$ induces a probability transition matrix $P^*(y|y')$ whose stationary output distribution is unique, the proof is concluded. Hence, we assume, otherwise, that $P^*(y|y')$ has infinitely many stationary output distributions. We show that there always exists $\tilde{\tilde{P}}(u|y')\in\mathcal{P}_\pi$, i.e., an input distribution that induces a unique stationary output distribution $\tilde{\tilde{\pi}}(y')=\tilde{\tilde{P}}_{Y'}(y')=\tilde{\tilde{P}}_Y(y'),\forall y'\in\mathcal{Y}$ with the corresponding conditional mutual information $\tilde{\tilde{I}}(U;Y|Y')$, such that $\tilde{\tilde{I}}(U;Y|Y')=I^*(U;Y|Y')$.

From Markov theory (see, e.g., [68, Th. 5.3.3 and Th. 5.5.12]), since $\mathcal{Y}$ is finite, if $P^*(y|y')$ induces only one irreducible subset of $\mathcal{Y}$, there is a unique stationary output distribution on $\mathcal{Y}$; this contradicts our assumption. Therefore, $P^*(y|y')$ decomposes $\mathcal{Y}$ into at least two disjoint irreducible closed sets. Assume that $P^*(u|y')$ induces exactly two irreducible closed subsets of $\mathcal{Y}$, denoted by $\mathcal{C}_i, i\in\{1,2\}$, with the corresponding probability transition matrices $P^*_{\mathcal{C}_i}(y|y'),\forall y',y\in\mathcal{C}_i$ which are derived from $P^*(y|y')$. Let $\pi^*_{\mathcal{C}_i}(y'),\forall y'\in\mathcal{C}_i$, where $\sum_{y'\in\mathcal{C}_i}\pi^*_{\mathcal{C}_i}(y')=1$, be the unique stationary distribution induced by $P^*_{\mathcal{C}_i}(y|y')$, and denote the corresponding maximal conditional mutual information of each $\mathcal{C}_i$ by $I^*_{\mathcal{C}_i}(U;Y|Y')\triangleq\sum_{y'\in C_i}\pi^*_{\mathcal{C}_i}(y')I^*(U;Y|Y'=y')$. It follows that $I^*(U;Y|Y')=\max_{i\in\{1,2\}} I^*_{\mathcal{C}_i}(U;Y|Y')$ since if, without loss of generality, $I^*_{\mathcal{C}_1}(U;Y|Y')\le I^*_{\mathcal{C}_2}(U;Y|Y')$, then

$$P^*_{Y'}(y')=\begin{cases}\pi^*_{\mathcal{C}_2}(y'), & y'\in\mathcal{C}_2\\ 0, & \text{otherwise}\end{cases} \qquad (38)$$

is a legitimate stationary distribution, i.e., it satisfies $P^*_{Y'}(y')=P^*_Y(y'),\forall y'\in\mathcal{Y}$. Furthermore, we can construct $\tilde{\tilde{P}}(u|y')\in\mathcal{P}_\pi$ such that $\tilde{\tilde{I}}(U;Y|Y')=I^*_{\mathcal{C}_2}(U;Y|Y')$ as follows. Construct $\tilde{\tilde{P}}(u|y')$ exactly as $P^*(u|y')$ for all $y'\in\mathcal{Y}$, but with $\tilde{\tilde{P}}(u|y'_1), y'_1\in\mathcal{C}_1$ that induces a positive probability to reach an arbitrary $y_2\in\mathcal{C}_2$ from an arbitrary initial $y'_1\in\mathcal{C}_1$ with some input sequence. This construction is legitimate because the NOST channel is assumed to be connected, and the $|\mathcal{X}|$ strategies that map all states to a specific input x are also a part of the set of all strategies. This construction of $\tilde{\tilde{P}}(u|y')$ renders all outputs in $\mathcal{C}_1$ transient and $\mathcal{C}_2$ a unique irreducible closed subset in $\mathcal{Y}$. That is, $\tilde{\tilde{P}}(u|y')$ induces a unique stationary output distribution on $\mathcal{Y}$ as given in (38), and $\tilde{\tilde{I}}(U;Y|Y'=y')=I^*_{\mathcal{C}_2}(U;Y|Y'=y')$ for any $y'\in\mathcal{C}_2$. Hence, $\tilde{\tilde{I}}(U;Y|Y')=I^*(U;Y|Y')$ as desired.

The construction can be extended in the case of multiple disjoint irreducible closed subsets of $\mathcal{Y}$ as there can be at most $|\mathcal{Y}|$ subsets, which is a finite number. Hence, it can be deduced that (37) holds.

Finally, using the cardinality bound of Theorem 3, whose proof (given next in this section) shows that $I^*(U;Y|Y')$ can be achieved with at most $L\le|\mathcal{X}|^{|\mathcal{S}|}$ strategies such that $P^*(y|y')$ is preserved, we conclude that $I^*(U;Y|Y')$ can always be achieved with some $P(u|y'),x(u,s')\in\mathcal{P}_\pi$ where $|\mathcal{U}|\le L$. This concludes the proof. □

D. Cardinality Bound

As the cardinality bound $|\mathcal{U}|\le|\mathcal{X}|^{|\mathcal{S}|}$ is trivial (there are $|\mathcal{X}|^{|\mathcal{S}|}$ strategies in total), here we prove the non-trivial cardinality bounds on $\mathcal{U}$ in Theorem 3, i.e., $|\mathcal{U}|\le(|\mathcal{X}|-1)|\mathcal{S}||\mathcal{Y}|+1$ and $|\mathcal{U}|\le(|\mathcal{Y}|-1)|\mathcal{Y}|+1$. If either of them is less than $|\mathcal{X}|^{|\mathcal{S}|}$, it implicitly means that not all strategies are required in order to achieve the feedback capacity, but only L of them at most.

Proof of Cardinality Bounds: We invoke the support lemma [69, p. 631], which is a consequence of the Fenchel-Eggle-ston-Caratheodory theorem [71], twice, for the auxiliary RV U. In each use, we show how the measures of the feedback capacity, i.e., the conditional entropies in $I(U;Y|Y')=H(Y|Y')-H(Y|Y',U)$, are preserved, thereby implying both non-trivial cardinality bounds. In other words, assuming U takes values in an arbitrary alphabet $\mathcal{U}$, we prove that given any (Y',S',U,X), there exists $(Y',S',\tilde{U},X)$ with $|\tilde{\mathcal{U}}|\le\min\{(|\mathcal{X}|-1)|\mathcal{S}||\mathcal{Y}|+1,(|\mathcal{Y}|-1)|\mathcal{Y}|+1\}$ such that $I(U;Y|Y')=I(\tilde{U};Y|Y')$.

We begin with proving $|\mathcal{U}|\le(|\mathcal{X}|-1)|\mathcal{S}||\mathcal{Y}|+1$. $\tilde{U}$ must have $(|\mathcal{X}|-1)|\mathcal{S}||\mathcal{Y}|$ letters to preserve $P(x|s',y')$ for all $s',y'\in\mathcal{S}\times\mathcal{Y}$. If $P(x|s',y')$ is preserved, $P(y',y)$ is preserved as well, because $P(y',y)=\pi(y')P(y|y')$, where $\pi(y')$ is a stationary distribution induced from the probability transition

matrix $P(y|y')$ that can also be expressed by

$$P(y|y') = \sum_{s',x} Q(s'|y')P(x|s',y')Q(y|x,s').$$

Additionally, if $P(y',y)$ is preserved, $H(Y|Y')$ is preserved, too. Finally, $\tilde{U}$ must have another letter to preserve $H(Y|Y',U)$. This concludes the the proof of the first non-trivial cardinality bound.

Following so, we prove $|\mathcal{U}| \leq (|\mathcal{Y}|-1)|\mathcal{Y}|+1$. However, this time, we aim to preserve $P(y|y')$ for all $y', y \in \mathcal{Y}$, directly. To address this, $\tilde{U}$ must have $(|\mathcal{Y}|-1)|\mathcal{Y}|$ letters, thereby preserving $H(Y|Y')$. Finally, $\tilde{U}$ must have another letter to preserve $H(Y|Y',U)$, which concludes the proof. □

E. Convex Optimization Formulations

Firstly, we prove Theorem 2, and subsequently Theorem 4 follows likewise.

Proof of Theorem 2: In this proof we show that (11a)-(11b) is a convex optimization problem, i.e., the maximization domain is convex, Constraints (11b) are linear and the objective (11a) is concave. The maximization is over the probability simplex $\mathcal{P}_{\mathcal{Y}\times\mathcal{X}}$, which is convex, and Constraints (11b) are linear in $\mathcal{P}_{\mathcal{Y}\times\mathcal{X}}$ because $Q(y|x,y')$ is the averaged channel in (3). Finally, we show that the objective function in (11a) is concave as follows. The conditional mutual information characterizing the feedback capacity can be written as $I(X;Y|Y') = H(Y|Y') - H(Y|Y',X)$. The first conditional entropy can be expressed as $H(Y|Y') = \log|\mathcal{Y}| - D(P(y',y)||P(y')U(y))$, where $U(y) = \frac{1}{|\mathcal{Y}|}$ is the uniform distribution over $\mathcal{Y}$, because

$$\begin{aligned} D(P(y',y)||P(y')U(y)) &= \sum_{y',y} P(y',y)\log\frac{P(y|y')}{U(y)} \\ &= \log|\mathcal{Y}| - H(Y|Y'). \end{aligned}$$

(This conditional entropy identity is an extension of the known entropy identity $D(P(y)||U(y)) = \log|\mathcal{Y}| - H(Y)$, see [72, Eq. (2.93)]). The relative entropy above is convex in the pair $(P(y',y), P(y')U(y))$, which are linear in $P(y',x)$ due to $P(y',y) = \sum_x P(y',x)Q(y|x,y')$ and $P(y') = \sum_x P(y',x)$, and thus $H(Y|Y')$ is concave in $P(y',x)$. On the other hand, the second conditional entropy is

$$H(Y|Y',X) = \sum_{y',x} P(y',x)H_Q(Y|x,y'), \tag{39}$$

where $H_Q(Y|x,y')$ is a constant determined by $Q(y|x,y')$; i.e., $H(Y|Y',X)$ is linear in $P(y',x)$. Thus, the difference between both conditional entropies is concave in $P(y',x)$, which completes the proof. □

The proof of Theorem 4 is similar to that of Theorem 2, but with replacing all occurrences of x with u and referring to $P_f(y|u,y')$ (13), which is determined by the NOST channel model for a fixed f, instead of referring to the averaged channel $Q(y|x,y')$.

VI. Conclusion and Further Work

This work is part of an ongoing progress on the feedback capacity of FSCs, which model channels with memory. It is challenging to derive a computable formula for the feedback capacity of FSCs with a stochastic channel state evolution and ISI. And yet, for the introduced family of FSCs called *NOST channels*, where the channel state is stochastically dependent on the channel output, the feedback capacities were derived as single-letter formulas under a connectivity condition in two scenarios: without and with CSI available at the encoder. These formulas were shown to be computable via convex optimization formulations. Furthermore, it was demonstrated via the noisy-POST(α,η) channel that CSI at the encoder may not always increase the feedback capacity.

We remark here on several interesting research directions that follow from the current work. It may be possible to extend our capacity results to a countable channel output alphabet $\mathcal{Y}$. This is not straightforward from the current derivation, as we explain. We previously mentioned, from Markov theory, that for a finite Markov chain there always exists at least one stationary distribution. However, for an infinite Markov chain, there may be no stationary distributions at all.[3] Another interesting research direction is to study NOST channels without feedback or in the regimes of noisy and delayed feedback links.

Appendix A
Proof of Lemma 3

$$\begin{aligned} &Q(y_i|x^i,y^{i-1},m) \\ &= \sum_{s_{i-1}\in S} Q(s_{i-1}|x^i,y^{i-1},m)Q(y_i|x^i,y^{i-1},s_{i-1},m) \\ &\overset{(a)}{=} \sum_{s_{i-1}\in S} Q(s_{i-1}|y^{i-1},m)Q(y_i|x^i,y^{i-1},s_{i-1},m) \\ &\overset{(b)}{=} \sum_{s_{i-1}\in S} Q(s_{i-1}|y_{i-1})Q(y_i|x_i,s_{i-1}) \\ &\overset{(c)}{=} Q(y_i|x_i,y_{i-1}), \end{aligned} \tag{40}$$

where

(a) follows from (4), i.e., for each time i, x_i is a function of (m,y^{i-1});
(b) follows from the NOST channel model (2);
(c) follows from the fact that (x^{i-1},y^{i-2},m) do not appear in the summation.

□

Appendix B
Proof of Theorem 1 Based on the Directed Information

Before presenting the proof, we recall the definitions of the *directed information* and the *causally conditional distribution*.

[3] A simple such known example is the symmetric random walk on the integers: $P_{Y|Y'}(i-1|i) = P_{Y|Y'}(i+1|i) = 0.5, \forall i \in \mathcal{Y} \triangleq \mathbb{Z}$, which has no solution π of $\pi P = \pi$ that is a legitimate distribution. We note that in this example, $\mathcal{Y}$ is irreducible. While for a finite Markov chain, irreducibility induces the existence of a unique stationary distribution, a fact we use throughout our derivations, this is not necessarily true for an infinite set. In fact, an irreducible Markov chain has a stationary distribution if and only if it is positive recurrent (see, e.g., [68, Th. 5.5.12]).

The *directed information* from X to Y conditioned on S, introduced by Massey [22] and employed with conditioning in [26], is defined as

$$I(X^n \to Y^n|S) \triangleq \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}, S). \tag{41}$$

The *causally conditional distribution*, introduced in [73], [74], is defined as

$$P(x^n||y^{n-1}) \triangleq \prod_{i=1}^{n} P(x_i|x^{i-1}, y^{i-1}). \tag{42}$$

The feedback capacity of any FSC was shown in [26] to be bounded by

$$\lim_{n\to\infty} \frac{1}{n} \max_{P(x^n||y^{n-1})} \min_{s_0} I(X^n \to Y^n|s_0) \leq C_{\text{FB}} \leq \lim_{n\to\infty} \frac{1}{n} \max_{P(x^n||y^{n-1})} \max_{s_0} I(X^n \to Y^n|s_0). \tag{43}$$

Alternative Proof of Theorem 1: For Setting I (connected NOST channels without CSI), the LHS of (43) is

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \max_{P(x^n||y^{n-1})} \min_{s_0} I(X^n \to Y^n|s_0) \\
&\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n} \max_{P(x^n||y^{n-1})} \min_{s_0} \left[\sum_{i=1}^{n} H(Y_i|Y^{i-1}, s_0) - H(Y_i|Y_{i-1}, X_i) \right] \\
&\overset{(b)}{=} \lim_{n\to\infty} \max_{\{P(x_i|y_{i-1})\}_{i=1}^{n}} \min_{s_0} \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y_i|Y_{i-1}) \\
&\overset{(c)}{=} \max_{P(x|y')} I(X; Y|Y'),
\end{aligned} \tag{44}$$

where,

(a) follows from Lemma 3;

(b) is explained by justifying the direct ($\geq$) and the converse ($\leq$): the direct follows from maximizing over $\{P(x_i|y_{i-1})\}_{i=1}^{n}$ for all i, which is a sub-domain of $P(x^n||y^{n-1})$, hence, $H(Y_i|Y^{i-1}, s_0) = H(Y_i|Y_{i-1})$ for $i > 1$ due to the Markov chain

$$\begin{aligned}
&P(y_i|y^{i-1}, s_0) \\
&= \sum_{s_{i-1}, x_i} Q(s_{i-1}|y_{i-1}) P(x_i|y_{i-1}) Q(y_i|x_i, s_{i-1}) \\
&= P(y_i|y_{i-1});
\end{aligned} \tag{45}$$

the converse follows from $H(Y_i|Y^{i-1}, s_0) \leq H(Y_i|Y_{i-1})$, and then identifying that for all i, the summand $I(X_i; Y_i|Y_{i-1})$ is induced by $P(y_{i-1}, x_i, y_i) = P(y_{i-1})P(x_i|y_{i-1})Q(y_i|x_i, y_{i-1})$; for any i, this joint distribution is determined by $\{P(x_j|y_{j-1})\}_{j=1}^{i}$, which can be shown by induction; and

(c) is also explained by justifying the direct and the converse as follows: the direct follows from maximizing over time-invariant input distributions that induce a unique stationary output distribution, then applying Lemma 1, which utilizes the connectivity assumption (Definition 1); the converse follows from maximizing over all time-invariant input distributions and applying Lemma 4 with x instead of u.

Note that up to Step (b), the expression depends on s_0 which affects the distribution on Y_1, which in turn affects the distribution on Y_2, and so on, while in Step (c) the expression is independent of s_0. This chain of equalities can be repeated also with $\max_{s_0}$ instead of $\min_{s_0}$, and hence we conclude that $\max_{P(x|y')} I(X; Y|Y')$ is the feedback capacity of a connected NOST channel, and it is not affected by $Q(s_0)$. □

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their valuable and constructive comments, which helped to improve this paper.

REFERENCES

[1] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 29, no. 4, pp. 1209–1220, Dec. 1958.

[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[3] 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.

[4] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 4, pp. 453–460, Jul. 1985.

[5] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state ISI channels," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2001, pp. 2992–2996.

[6] R. Gray, M. Dunham, and R. Gobbi, "Ergodicity of Markov channels," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 5, pp. 656–664, Sep. 1987.

[7] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels—A survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.

[8] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[9] W. Turin, *Performance Analysis of Digital Transmission Systems*. New York, NY, USA: Computer Science Press, Mar. 1990.

[10] M. Hassan, M. M. Krunz, and I. Matta, "Markov-based channel characterization for tractable performance analysis in wireless packet networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 3, pp. 821–831, May 2004.

[11] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[12] C. Pimentel, T. H. Falk, and L. Lisbôa, "Finite-state Markov modeling of correlated Rician-fading channels," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1491–1501, Sep. 2004.

[13] L. Zhong, F. Alajaji, and G. Takahara, "A model for correlated Rician fading channels based on a finite queue," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 79–89, Jan. 2008.

[14] L. Galluccio, A. Lombardo, G. Morabito, S. Palazzo, C. Panarello, and G. Schembra, "Capacity of a binary droplet-based microfluidic channel with memory and anticipation for flow-induced molecular communications," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 194–208, Jan. 2018.

[15] N. Farsad, H. B. Yilmaz, A. Eckford, C. B. Chae, and W. Guo, "A comprehensive survey of recent advancements in molecular communication," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1887–1919, 3rd Quart., 2016.

[16] K. A. S. Immink, P. H. Siegel, and J. K. Wolf, "Codes for digital recorders," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2260–2299, Oct. 1998.

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

[18] T. T. Kadota, M. Zakai, and J. Ziv, "Capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 372–378, Jul. 1971.

[19] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," (in Russian), *Uspekhi Matematicheskikh Nauk*, vol. 14, pp. 3–104, 1959.

[20] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Moscow, Russia: Izvestiya Rossiiskoi Akademii Nauk, 1960.

[21] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[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, "Directed information for channels with feedback," Ph.D. dissertation, ETH Zurich, Zürich, Switzerland, 1998.

[24] Y. H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 25, no. 4, pp. 1488–1499, Apr. 2008.

[25] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[26] 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.

[27] 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.

[28] B. Shrader and H. Permuter, "Feedback capacity of the compound channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3629–3644, Aug. 2009.

[29] R. Dabora and A. J. Goldsmith, "On the capacity of indecomposable finite-state channels with feedback," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 193–203, Jan. 2013.

[30] H. H. Permuter, H. Asnani, and T. Weissman, "Capacity of a post channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6041–6057, Oct. 2014.

[31] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 546–549, Mar. 1995.

[32] F. Alajaji and T. Fuja, "Effect of feedback on the capacity of discrete additive channels with memory," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 1994, p. 464.

[33] L. Song, F. Alajaji, and T. Linder, "Capacity of burst noise-erasure channels with and without feedback and input cost," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 276–291, Jan. 2019.

[34] N. Sen, F. Alajaji, and S. Yuksel, "Feedback capacity of a class of symmetric finite-state Markov channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4110–4122, Jul. 2011.

[35] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 761–771, Mar. 1999.

[36] S. C. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2000.

[37] 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.

[38] S. Yang, A. Kavčić, and S. 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.

[39] 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.

[40] E. Shemuel, O. Sabag, and H. Permuter, "Finite-state channel with feedback and causal state information available at the encoder," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 1081–1088.

[41] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vols. 1–2. Belmont, MA, USA: Athena Scientific, 2000.

[42] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.

[43] J. Wu and A. Anastasopoulos, "On the capacity of the general trapdoor channel with feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 2256–2260.

[44] O. Sabag, H. H. Permuter, and N. Kashyap, "The feedback capacity of the binary erasure channel with a no-consecutive-ones input constraint," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 8–22, Jan. 2016.

[45] 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.

[46] O. Sabag, H. H. Permuter, and N. Kashyap, "Feedback capacity and coding for the BIBO channel with a no-repeated-ones input constraint," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 4940–4961, Jul. 2018.

[47] O. Sabag, H. H. Permuter, and H. D. Pfister, "A single-letter upper bound on the feedback capacity of unifilar finite-state channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1392–1409, Mar. 2017.

[48] O. Peled, O. Sabag, and H. H. Permuter, "Feedback capacity and coding for the $(0, k)$-RLL input-constrained BEC," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4097–4114, Jul. 2019.

[49] Z. Aharoni, O. Sabag, and H. H. Permuter, "Computing the feedback capacity of finite state channels using reinforcement learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 837–841.

[50] O. Sabag, B. Huleihel, and H. H. Permuter, "Graph-based encoders and their performance for finite-state channels with feedback," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2106–2117, Apr. 2020.

[51] O. Sabag and H. H. Permuter, "An achievable rate region for the two-way channel with common output," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 527–531.

[52] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.

[53] J. H. Bae and A. Anastasopoulos, "A posterior matching scheme for finite-state channels with feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2010, pp. 2338–2342.

[54] O. Sabag, V. Kostina, and B. Hassibi, "On converse bounds with stationary distributions," submitted for publication.

[55] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, no. 4, pp. 289–293, Oct. 1958.

[56] A. V. Kuznetsov and B. S. Tsybakov, "Coding in a memory with defective cells," *Problemy Peredachi Informatsii*, vol. 10, no. 2, pp. 52–60, Apr./Jun. 1974.

[57] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Problems Control Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[58] C. Heegard and A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 731–739, Sep. 1983.

[59] T. Weissman, "Capacity of channels with action-dependent states," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5396–5411, Nov. 2010.

[60] C. Choudhuri, Y.-H. Kim, and U. Mitra, "Causal state communication," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3709–3719, Jun. 2013.

[61] Y.-K. Chia and A. El Gamal, "Wiretap channel with causal state information," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2838–2849, May 2012.

[62] A. Bracher and A. Lapidoth, "Feedback, cribbing, and causal state information on the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7627–7654, Dec. 2014.

[63] S. Kotagiri and J. N. Laneman, "Multiaccess channels with state known to one encoder: A case of degraded message sets," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2007, pp. 1566–1570.

[64] A. Zaidi, S. P. Kotagiri, J. N. Laneman, and L. Vandendorpe, "Multiaccess channels with state known to one encoder: Another case of degraded message sets," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2009, pp. 2376–2380.

[65] A. Zaidi, P. Piantanida, and S. S. Shitz, "Multiple access channel with states known noncausally at one encoder and only strictly causally at the other encoder," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2011, pp. 2801–2805.

[66] E. Shemuel, O. Sabag, and H. Permuter, "Feedback capacity of finite-state channels with causal state known at the encoder," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2120–2125.

[67] Z. Goldfeld, P. Cuff, and H. H. Permuter, "Wiretap channels with random states non-causally available at the encoder," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1497–1519, Mar. 2020.

[68] R. Durrett, *Probability: Theory and Examples*, 5th ed. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[69] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[70] R. B. Ash, *Basic Probability Theory*. Chelmsford, MA, USA: Courier Corporation, 2008.

[71] H. G. Eggleston, *Convexity*. Cambridge, U.K.: Cambridge Univ. Press, 1958.

[72] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[73] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[74] H. Permuter, T. Weissman, and A. Goldsmith, "Capacity of finite-state channels with time-invariant deterministic feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2006, pp. 64–68.

Eli Shemuel (Student Member, IEEE) received the B.Sc. *(cum laude)* and M.Sc. *(summa cum laude)* degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering, under the direct track for honor students. His research interests include information theory and machine learning. He was a recipient of the Ze'ev Jabotinsky Fellowship from the Ministry of Science, Technology and Space of Israel for outstanding Ph.D. students in the direct track.

Oron Sabag (Member, IEEE) received the B.Sc. *(cum laude)*, M.Sc. *(summa cum laude)*, and Ph.D. degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2013, 2016, and 2019, respectively. He is currently a Post-Doctoral Fellow with the Department of Electrical Engineering, California Institute of Technology (Caltech). His research interests include control theory, information theory, and reinforcement learning. He was a recipient of several awards, among them are the ISEF Post-Doctoral Fellowship, the Lachish Fellowship, the ISIT-2017 Best Student Paper Award, the SPCOM-2016 Best Student Paper Award, the Feder Family Award for Outstanding Research in Communications, and the Kaufman Award.

Haim H. Permuter (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees *(summa cum laude)* in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 1997 and 2003, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2008. From 1997 to 2004, he was an Officer with the Research and Development Unit, Israeli Defense Forces. Since 2009, he has been with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, where he is currently a Professor and the Luck-Hille Chair of electrical engineering. He also serves as the Head for the Communication Track in his department. He was a recipient of several awards, including the Fullbright Fellowship, the Stanford Graduate Fellowship (SGF), the Allon Fellowship, and the U.S.–Israel Binational Science Foundation Bergmann Memorial Award. He has served on the Editorial Board for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2013 to 2016.