The Rate-and-State Capacity with Feedback

Shraga I. Bross, *Senior Member, IEEE*, and Amos Lapidoth, *Fellow, IEEE*

***Abstract*— The rate-and-state capacity of a state-dependent channel with a state-cognizant encoder is the highest possible rate of communication over the channel when the decoder—in addition to reliably decoding the data—must also reconstruct the state sequence with some required fidelity. Feedback from the channel output to the encoder is shown to increase this capacity even for channels that are memoryless with memoryless states. This capacity is calculated here for such channels with feedback when the state reconstruction fidelity is measured using a single-letter distortion function and the state sequence is revealed to the encoder in one of two different ways: strictly–causally or causally. For the noncausal case, we provide bounds on the capacity and identify a condition under which the bounds coincide. Feedback does not increase the rate-and-state capacity when the decoder must reconstruct the state sequence perfectly or, in some settings, when the channel is Gaussian and fidelity is measured in terms of mean squared-error.**

***Index Terms*— capacity, causal, distortion, feedback, fidelity, Gelfand-Pinsker, noncausal, rate, side-information, state.**

I. Introduction

THE Rate-and-State (RnS) capacity of a state-dependent discrete memoryless channel (SD-DMC) with a state-cognizant encoder is the highest rate at which data can be transmitted over the channel when the decoder—in addition to reliably decoding the data—must also reconstruct the state sequence with some required fidelity. As we shall see, unlike the Shannon capacity, it typically increases when a feedback link is introduced from the channel's output to the encoder. Noteworthy exceptions are when the state sequence is to be reconstructed losslessly or, in some settings, when the channel is Gaussian and fidelity is measured in terms of mean squared-error.

Here we compute the RnS capacity in the presence of feedback in two settings: when the state-information (SI) is revealed to the encoder *strictly-causally* and when it is revealed *causally*. We shall see that in both cases the RnS capacity can be achieved using a Block-Markov coding scheme with backward decoding, where in Block-b the encoder uses a blockcode to send fresh information and also a (lossy) description of the state sequence pertaining to Block-$(b-1)$. When forming this description, the describer is cognizant of the Block-$(b-1)$ channel outputs (via the feedback link) as well as of the Block-$(b-1)$ codeword. The description is required to be sufficiently detailed so as to allow a reconstructor that is also cognizant of these outputs and codeword to estimate the Block-$(b-1)$ state sequence with the desired fidelity. In decoding the Block-$(b-1)$ codeword, the receiver ignores the description that was sent in Block-b: it only uses the Block-$(b-1)$ output sequence. Once it has decoded the Block-$(b-1)$ codeword, it is in possession not only of the Block-$(b-1)$ channel outputs but also of the Block-$(b-1)$ codeword. It then uses these and the state description that was sent in Block-b to estimate the Block-$(b-1)$ state sequence.

We thus send two data streams in Block-b: fresh information and a description of the Block-$(b-1)$ state sequence. These are decoded based on the Block-b output sequence only. The sum of their rates is thus upper-bounded by the achievable rate on the channel (with the given input distribution and in the absence of state-reconstruction constraints). This achievable rate is known for all the models we consider: in the strictly-causal case it corresponds to $I(X;Y)$; in the causal case to $I(T;Y)$; and in the noncausal case to $I(T;Y)-I(T;S)$. To achieve the distortion D, the rate of the data stream describing the Block-$(b-1)$ state sequence can be (slightly more than) the conditional rate-distortion function, i.e., $R_{S|XY}(D)$ in the strictly-causal case; $R_{S|TY}(D)$ in the causal case; and $R_{S|TY}(D)$ in the noncausal case. Subtracting this rate from the total rate leaves us with an achievable data rate of $I(X;Y)-R_{S|XY}(D)$ in the strictly-causal case (*c.f.* (16)); $I(T;Y)-R_{S|TY}(D)$ in the causal case (*c.f.* (30)); and $I(T;Y)-I(T;S)-R_{S|TY}(D)$ in the noncausal case (*c.f.* (35)).

For the noncausal case, however, we do not have a matching upper bound and hence no proof of optimality. The upper bound that we do present is not always tight. Moreover, the noncausal case requires some extra care, because the classical coding scheme for this setting involves subcodes, with only the subcode corresponding to the transmitted message—as opposed to the transmitted codeword itself—typically decoded. In the Appendix we address this issue and show that the rate $I(T;Y)-I(T;S)$ can also be achieved when we insist on reliably decoding the codeword.[1]

There is an alternative Block-Markov scheme that we do not pursue here, because it is more complicated and yet leads to the same achievable rates. In Block-b of that scheme, in addition to fresh data, the encoder also transmits

Manuscript received March 28, 2017; revised August 15, 2017; accepted October 28, 2017. Date of publication November 23, 2017; date of current version February 15, 2018. S. Bross was supported by the Israel Science Foundation under Grant 455/14.

S. Bross is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: brosss@biu.ac.il).

A. Lapidoth is with ETH Zurich, 8092 Zurich, Switzerland (e-mail: lapidoth@isi.ee.ethz.ch).

This paper was presented in part at the 2016 IEEE International Symposium on Information Theory and the 2016 IEEE Convention of Electrical and Electronics Engineers, Israel.

Communicated by A. Khisti Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2017.2777389

[1] An analogous result for the "dirty-paper" setting was presented in [2].

a lossy description of the states *and codeword* pertaining to Block-$(b-1)$. For the purpose of this description, the channel outputs pertaining to Block-$(b-1)$ serve as side-information that is available (before Block-b commences) to both describer (via the feedback link) and reconstructor. Once the transmission in Block-b has been decoded, the receiver recovers the fresh information that was transmitted in that block as well as the said description pertaining to Block-$(b-1)$. Using the latter in combination with the Block-$(b-1)$ channel outputs, it then proceeds to decode the Block-$(b-1)$ codeword. (The description must be fine enough to allow this.) Using the description, the decoded Block-$(b-1)$ codeword, and the Block-$(b-1)$ channel outputs, the receiver then estimates the Block-$(b-1)$ state sequence to within the required fidelity.

The literature on the SD-DMC is extensive [9]. Noteworthy is [10], which considers RnS transmission without feedback when the reconstruction fidelity is replaced by a list size: in addition to decoding the data reliably, the decoder must form a list that with high probability contains the state sequence. The problem addressed in [10] is thus more of a "guessing" nature than an "estimation" nature. For this problem [10] characterizes the tension between the data rate and the exponential growth of the list-size in the blocklength. The converse in [10] is based on the extension of Fano's inequality to lists and is hence inapplicable to our setting. Particularly relevant to our setting are [19] and [3], which consider data transmission and state estimation without feedback: the first deals with the Gaussian channel with noncausal SI and mean squared-error state-reconstruction fidelity, and the second with a general SD-DMC with strictly-causal or causal SI and general single-letter state-reconstruction fidelity.

Although without feedback, [3] presents techniques that are very relevant to our setting, particularly to the strictly-causal case. In fact, for this case the with-feedback converse can be derived using the no-feedback converse of [3] in combination with the Functional Representation Lemma [20], [12]. But this does not apply to the causal case.

Related to the noncausal version of our RnS problem is the source-coding problem with a "vending machine" [14]. Indeed—in the special case where R is zero, i.e., when we only wish to transmit state information—our problem of finding the least achievable state-estimation distortion is nearly identical to the special case of the problem addressed in [14, Fig. 2, Sec. III] when we substitute zero for the description rate. In the terminology of the present paper, the zero-description-rate case in [14, Fig. 2, Sec. III] corresponds to $R = 0$ and no feedback. In the terminology of [14] our zero-rate problem corresponds to the case where the description-rate is zero; the encoder observes the side information strictly causally; and there are no cost constraints. Issues related to coordination over state-dependent channels are addressed in [11].

To appreciate the benefits of feedback, it is instructive to consider a special kind of SD-DMC. Let us denote a generic SD-DMC by $\big(P_S, P_{\mathrm{c}}(y|x,s)\big)$, where P_S is the probability mass function (PMF) of the state, and where the transition law $P_{\mathrm{c}}(y|x,s)$ is the PMF induced on the output alphabet $\mathcal{Y}$ when the input to the channel is $x \in \mathcal{X}$ and the state of the channel is $s \in \mathcal{S}$. The special case to consider is when the output Y corresponds to a pair $(\tilde{Y}, \tilde{S})$, the state is S of PMF P_S, and the transition law factorizes as

$$P_{\mathrm{c}}(\tilde{y}, \tilde{s}|x, s) = \tilde{P}_{\mathrm{c}}(\tilde{y}|x)\, P_{\tilde{S}|S}(\tilde{s}|s). \tag{1}$$

In this case the state and input do not interact, and it is intuitively clear that this channel's RnS capacity is the difference between the Shannon capacity of the channel $\tilde{P}_{\mathrm{c}}\big(\tilde{y}|x\big)$ and the rate that is needed to describe the state to a reconstructor observing $\tilde{S}$. While the former is unaffected by feedback, the latter is: In the absence of feedback the $\tilde{S}$-sequence is only observed by the decoder, and the encoder is thus faced with a Wyner-Ziv problem [24] of describing S to a reconstructor that observes $\tilde{S}$. But in the presence of feedback the $\tilde{S}$-sequence—being part of the channel output $(\tilde{Y}, \tilde{S})$—is revealed also to the encoder, and the encoder is thus faced with a classical rate-distortion problem with side information $\tilde{S}$ *that is available to both describer and reconstructor*. Since this rate-distortion function is typically lower than the Wyner-Ziv rate [24, Sec. II], we conclude that—irrespective of whether the state is revealed to the encoder strictly-causally, causally, or noncausally—feedback can increase the RnS capacity.

II. The Set-Up

We are given a SD-DMC $\big(P_S, P_{\mathrm{c}}(y|x,s)\big)$ and a nonnegative distortion function $d\colon \mathcal{S} \times \hat{\mathcal{S}} \to \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the nonnegative reals; $\hat{\mathcal{S}}$ is the reconstruction alphabet; and the alphabets $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, $\hat{\mathcal{S}}$ are all finite. The maximum of d is finite and is denoted $d_{\max}$:

$$d_{\max} = \max_{(s,\hat{s}) \in \mathcal{S} \times \hat{\mathcal{S}}} d(s, \hat{s}). \tag{2}$$

The distortion $d(\mathbf{s}, \hat{\mathbf{s}})$ between an n-length source sequence $\mathbf{s} = (s_1, \ldots, s_n) \in \mathcal{S}^n$ and an n-length reconstruction sequence $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_n) \in \hat{\mathcal{S}}^n$ is defined as the average of the component-wise distortions

$$d(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{n} \sum_{k=1}^{n} d(s_k, \hat{s}_k). \tag{3}$$

Here and throughout we denote n-length sequences with bold letters, e.g., $\mathbf{Y}$ for $Y_1, \ldots, Y_n$. We use Y_i^j for $Y_i, \ldots, Y_j$, and we suppress i when it is 1.

Let n denote the blocklength, R the data rate, and $\mathcal{W} = \{1, \ldots, e^{nR}\}$ the set of messages.[2] In all our settings the decoder consists of two mappings. The first,

$$\phi_W\colon \mathcal{Y}^n \to \big\{1, \ldots, e^{nR}\big\}, \tag{4}$$

is used to decode the message, and we denote by $\hat{W}$ the result of applying it to the received sequence $\mathbf{Y}$, so $\hat{W} = \phi_W(\mathbf{Y})$. The second,

$$\phi_S\colon \mathcal{Y}^n \to \hat{\mathcal{S}}^n, \tag{5}$$

is used to reconstruct the state sequence, and we denote by $\hat{\mathbf{S}}$ the result of applying it to $\mathbf{Y}$, so $\hat{\mathbf{S}} = \phi_S(\mathbf{Y})$.

[2]Throughout this paper e^{nR} stands for $\lfloor e^{nR} \rfloor$, i.e., for the largest integer that does not exceed it.

The form of the encoder depends on the setting. In the strictly-causal setting with feedback the encoder comprises n mappings

$$f_k\colon \mathcal{W}\times\mathcal{S}^{k-1}\times\mathcal{Y}^{k-1}\to\mathcal{X},\quad k=1,\ldots,n \tag{6}$$

with the understanding that the time-k symbol X_k that the encoder produces in order to convey Message W after having observed the states S^{k-1} and the outputs Y^{k-1} is

$$X_k = f_k\big(W, S^{k-1}, Y^{k-1}\big),\quad k=1,\ldots,n. \tag{7}$$

In the causal case the domain in (6) is replaced by $\mathcal{W}\times\mathcal{S}^k\times\mathcal{Y}^{k-1}$ and the RHS of (7) is replaced by $f_k(W, S^k, Y^{k-1})$. And in the noncausal case the domain is $\mathcal{W}\times\mathcal{S}^n\times\mathcal{Y}^{k-1}$ and X_k is $f_k(W, S^n, Y^{k-1})$. Each of these cases also has a no-feedback counterpart, where $\mathcal{Y}^{k-1}$ is removed from the domain, and Y^{k-1} is removed from the definition of X_k. The arithmetic average of the probabilities of error associated with the different messages is denoted $P_{\mathrm{e}}^{(n)}$.

A pair (R, D) is *achievable* if for every $\varepsilon>0$ we can find some positive integer $n_0(\varepsilon)$ such that, for every blocklength n exceeding $n_0(\varepsilon)$, there exist an encoder whose rate exceeds $R-\varepsilon$ and decoding mappings ϕ_W and ϕ_S such that

$$\mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big] \le D+\varepsilon \tag{8}$$

and

$$\lim_{n\to\infty} P_{\mathrm{e}}^{(n)} = 0. \tag{9}$$

Here the allowed encoding functions are determined by the setting under consideration. We denote by $\mathscr{R}$ the set of achievable (R, D) pairs, and we note that it is a compact subset of $\mathbb{R}_+\times\mathbb{R}_+$. For every given maximal-allowed distortion D, we define the RnS capacity as the maximum over all rates R for which (R, D) is achievable, where throughout the paper we adopt the convention that the maximum over an empty set is $-\infty$. The maximum exists because $\mathscr{R}$ is compact. The different settings have RnS capacities

$$C^{\text{s-c}}(D),\ C^{\text{c}}(D),\ C^{\text{nc}}(D),\ C_{\text{FB}}^{\text{s-c}}(D),\ C_{\text{FB}}^{\text{c}}(D),\ C_{\text{FB}}^{\text{nc}}(D)$$

all of which are denoted C, with the subscript "FB" indicating feedback, and the superscript indicating how the state information is revealed to the encoder.

By the *lossless case* we refer to the case where the maximal-allowed distortion D is zero, and the distortion function is the Hamming distortion function $(s,\hat{s})\mapsto\mathbb{1}\{\hat{s}\neq s\}$. Here and throughout $\mathbb{1}\{\text{statement}\}$ is one or zero depending on whether or not the statement holds.

Remark 1: The lossless case is reminiscent of the case where Δ in [10] is $H(S)$. It is not identical because the latter case corresponds to a subexponential list, and our case corresponds to an arbitrarily small distortion in the sense of (8) (with D replaced by zero).

Finally, although not of finite alphabet, the Gaussian channel is also of interest to us. This is a memoryless channel, where

$$Y = x+S+Z, \tag{10a}$$

and where—irrespective of the (real) value of x—the random variables S and Z are independent centered Gaussians of respective variances σ_s^2 and N. The input is constrained to satisfy

$$\sum_{k=1}^{n}\mathbb{E}\big[X_k^2\big] \le nP \tag{10b}$$

for some given maximal-allowed average power P. Here the expectation is over the messages (under a uniform prior), the state sequence, and—in the case of feedback—over the noise sequence.

III. Main Results

Except when we discuss the Gaussian channel, we assume throughout a SD-DMC $\big(P_S, P_{\mathrm{c}}(y|x,s)\big)$ with finite alphabets and a (finite) nonnegative distortion function $d\colon\mathcal{S}\times\hat{\mathcal{S}}\to\mathbb{R}_+$. We begin with results on the case where the state information is revealed to the encoder strictly-causally.

A. Strictly-Causal State Information

Theorem 1 (Strictly-Causal SI and Feedback):

$$C_{\text{FB}}^{\text{s-c}}(D) = \max_{P_X,\,P_{U|XSY}}\Big\{I(X;Y)-I(S;U|XY)\Big\}, \tag{11}$$

where the maximum is over all joint PMFs of the form

$$\begin{aligned}P_{SXYU}(s,x,y,u) = {}& P_S(s)\,P_X(x)\,P_{\mathrm{c}}(y|x,s)\\ &\cdot P_{U|XSY}(u|x,s,y)\end{aligned} \tag{12}$$

for which there exists a mappings

$$g\colon\mathcal{U}\times\mathcal{X}\times\mathcal{Y}\to\hat{\mathcal{S}} \tag{13}$$

satisfying

$$\mathbb{E}\big[d\big(S, g(U,X,Y)\big)\big] \le D, \tag{14}$$

where the expectation and the mutual informations are computed with respect to the above P_{SXYU}. The cardinality of the set $\mathcal{U}$ in which the auxiliary chance variable U takes values can be bounded by that of $\hat{\mathcal{S}}$

$$|\mathcal{U}| \le |\hat{\mathcal{S}}|. \tag{15}$$

Alternatively, $C_{\text{FB}}^{\text{s-c}}(D)$ can be expressed as

$$C_{\text{FB}}^{\text{s-c}}(D) = \max_{P_X}\big\{I(X;Y)-R_{S|XY}(D)\big\}, \tag{16}$$

where $R_{S|XY}(\cdot)$ is the rate-distortion function of the source S with respect to the distortion measure $d(s,\hat{s})$ when both encoder and reconstructor are cognizant of (X,Y), and $(X,S,Y)\sim P_X(x)\,P_S(s)\,P_{\mathrm{c}}(y|x,s)$ [1, eq. (6.1.21)], [6, eq. (11.2)].

Proof: The proof of (11) is in Section IV-A. To see that (11) is equivalent to (16), recall that $R_{S|XY}(\cdot)$ is the rate-distortion function of the source S when both encoder and reconstructor are cognizant of (X,Y). This situation can also be formulated as an instance of the Wyner-Ziv problem [6, Th. 11.3] when the side information (X,Y) is part of the source but need not be reconstructed. It is thus the Wyner-Ziv problem for the source $\big(S,(X,Y)\big)$ with the distortion $d(s,\hat{s})$ that depends on the source only via its first

component. Applying the Wyner-Ziv expression [6, Th. 11.3] yields (11). ∎

Discussion: With strictly-causal SI and feedback (or without it), the achievable rate corresponding to the input distribution P_X equals the input-output mutual information $I(X;Y)$. If, in addition to sending the message, the encoder must also provide a reconstructor that is cognizant of X and Y with a description of the state to within a desired distortion D, then the achievable rate is penalized by the conditional rate-distortion function $R_{S|XY}(D)$ for describing S when both the encoder and the decoder have access to (X, Y).

In the absence of feedback, the RnS capacity with strictly-causal SI was computed in [3, Th. 2]:

Theorem 2 (Strictly-Causal SI and No Feedback [3, Th. 2]):

$$C^{\text{s-c}}(D) = \max_{P_X,\, P_{U|XS}} \big\{ I(UX;Y) - I(S;U|X) \big\}, \tag{17}$$

where the maximum is over all joint PMFs of the form

$$\begin{aligned} P_{SXYU}(s,x,y,u) \;=\; & P_S(s)\, P_X(x)\, P_{\text{c}}(y|x,s) \\ & \cdot\, P_{U|XS}(u|x,s) \end{aligned} \tag{18}$$

for which there exists a mapping

$$g : \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \to \hat{\mathcal{S}} \tag{19}$$

satisfying

$$\mathbb{E}\big[d\big(S, g(U,X,Y)\big)\big] \le D, \tag{20}$$

where the expectation and the mutual informations are computed with respect to the above P_{SXYU}.

Discussion: This theorem—though not dealing with feedback—can be used to prove the converse part of Theorem 1 as follows. By the Functional Representation Lemma [20, eq. (44)], [12] we can represent the output Y_k as a deterministic function of (X_k, S_k, Θ_k), where $\{\Theta_k\}$ are IID, and Θ_k is independent of (X^{k-1}, S^n). The encoder, to whom the past inputs X^{k-1} are known, can thus compute the past outputs Y^{k-1} from (S^{k-1}, Θ^{k-1}), and, consequently, revealing the past states and outputs (S^{k-1}, Y^{k-1}) to the encoder is at most as informative as revealing (S^{k-1}, Θ^{k-1}) without the past outputs Y^{k-1}, i.e., without feedback. The latter scenario can be viewed as a no-feedback scenario with state $\tilde{S}_k = (S_k, \Theta_k)$; with only $\tilde{S}^{k-1}$ being fed to the encoder; and with the distortion measure

$$\tilde{d}\big((s,\theta),\hat{s}\big) = d(s,\hat{s}).$$

It thus falls under the setting addressed in [3, Th. 2], and we can apply [3, Th. 2], and specifically (17), to obtain the desired upper bound, because

$$\begin{aligned} I(UX;Y) - I(\tilde{S};U|X) \;&=\; I(UX;Y) - I(S\Theta;U|X) \\ &\overset{(a)}{=}\; I(UX;Y) - I(S\Theta Y;U|X) \\ &\overset{(b)}{=}\; I(X;Y) - I(S\Theta;U|XY) \\ &\le\; I(X;Y) - I(S;U|XY), \end{aligned} \tag{21}$$

where (a) holds because Y is a deterministic function of (X, S, Θ), and (b) follows from the chain rule.

This approach does not extend to the causal case: we can still argue that revealing $\tilde{S}^k$ is at least as informative as revealing (Y^{k-1}, S^k), but the application of [3, Th. 3] to the former setting would lead to a loose bound. Fortunately, the converse for the strictly-causal case that we present in Section IV-A does extend to the causal case.

In general, $C_{\text{FB}}^{\text{s-c}}(D)$ can exceed $C^{\text{s-c}}(D)$, but in the lossless case they are equal. This is perhaps not surprising because the Slepian-Wolf theorem on lossless source coding demonstrates that the side information is not always needed at the encoder.

Proposition 3 (Lossless Reconstruction: Strictly-Causal SI): In the lossless case with the state being revealed to the encoder strictly causally,

$$C^{\text{s-c}}(0) = C_{\text{FB}}^{\text{s-c}}(0) = \max_{P_X} \big\{ I(X;Y) - H(S|XY) \big\},$$

where the mutual information and conditional entropy are computed under the law $P_X(x)\, P_S(s)\, P_{\text{c}}(y|x,s)$.

Proof: See Section IV-B. ∎

Feedback also does not increase the RnS capacity on the Gaussian channel. This is perhaps not surprising because, for Gaussian sources with mean squared-error distortion, the Wyner-Ziv rate-distortion (corresponding to the case where the side-information is available to the reconstructor only) is equal to conditional rate-distortion (where the side-information is available to both encoder and reconstructor).

Proposition 4 (Gaussian Channel: Strictly-Causal SI): Consider the state-dependent Gaussian channel of noise-variance N, state-variance σ_s^2, and maximal-allowed average power P. Let the state-reconstruction distortion function be $(s,\hat{s}) \mapsto (s-\hat{s})^2$. If the state is revealed to the encoder strictly-causally, then $C^{\text{s-c}}(D) = C_{\text{FB}}^{\text{s-c}}(D)$ for every $D > 0$. Moreover, if we define for every $0 \le \gamma \le 1$ the quantities

$$R_\gamma \;=\; \frac{1-\gamma}{2} \log\left(\frac{P + \sigma_s^2 + N}{\sigma_s^2 + N} \right) \tag{22a}$$

$$D_\gamma \;=\; \sigma_s^2 \frac{N}{\sigma_s^2 + N} \left(\frac{\sigma_s^2 + N}{P + \sigma_s^2 + N} \right)^{\gamma}, \tag{22b}$$

then D_γ evaluates at $\gamma = 1$ to the least achievable distortion; R_γ evaluates at $\gamma = 0$ to the supremum of achievable rates; and $C_{\text{FB}}^{\text{s-c}}$ (and hence also $C^{\text{s-c}}$) is given parametrically by

$$C_{\text{FB}}^{\text{s-c}}(D_\gamma) = R_\gamma, \quad 0 \le \gamma \le 1. \tag{23}$$

A nonnegative tuple (R, D) is thus achievable if, and only if,

$$R + \frac{1}{2} \log^+\left(\frac{\sigma_s^2 N/(\sigma_s^2 + N)}{D} \right) \le \frac{1}{2} \log\left(1 + \frac{P}{\sigma_s^2 + N} \right), \tag{24}$$

where $\log^+(\xi) \triangleq \max\{0, \log \xi\}$.

Proof: See Section IV-C. ∎

B. Causal State Information

Theorem 5 (Causal SI and Feedback):

$$C_{\text{FB}}^{\text{c}}(D) = \max_{P_T,\, P_{U|TSY},\, f} \big\{ I(T;Y) - I(S;U|TY) \big\}, \tag{25}$$

where U and T are auxiliary chance variables taking values in $\mathcal{U}$ and $\mathcal{T}$ respectively; the mapping f is from $\mathcal{T} \times \mathcal{S}$ to $\mathcal{X}$;

the mutual informations are computed with respect to the joint PMF

$$P_{STXYU}(s,t,x,y,u) = P_S(s)\,P_T(t)\,\mathbb{1}\{x = f(t,s)\} \cdot P_{\mathrm{c}}(y|x,s)\,P_{U|TSY}(u|t,s,y); \quad (26)$$

and it is required that there exist a mapping

$$g: \mathcal{U} \times \mathcal{T} \times \mathcal{Y} \to \hat{\mathcal{S}} \quad (27)$$

satisfying

$$\mathbb{E}\big[d\big(S, g(U,T,Y)\big)\big] \le D. \quad (28)$$

Moreover, in the above maximization we may restrict the cardinalities to

$$|\mathcal{U}| \le |\hat{\mathcal{S}}| \quad (29a)$$

and

$$|\mathcal{T}| \le \min\big\{|\mathcal{X}| \cdot |\mathcal{S}|, |\mathcal{Y}|\big\} + 1. \quad (29b)$$

Alternatively, $C_{\mathrm{FB}}^{\mathrm{c}}(D)$ can be expressed as

$$C_{\mathrm{FB}}^{\mathrm{c}}(D) = \max_{P_T,\, f} \big\{I(T;Y) - R_{S|TY}(D)\big\}, \quad (30)$$

where $I(T;Y)$ and $R_{S|TY}(D)$ are computed under the PMF

$$P_{STY}(s,t,y) = P_S(s)\,P_T(t)\,P_{\mathrm{c}}\big(y\big|f(t,s),s\big). \quad (31)$$

Proof: The cardinality bound (29a) can be imposed because, like in (15), the auxiliary chance variable U can be chosen to take values in $\hat{\mathcal{S}}$. As to T, we note that for a fixed mapping $f: \mathcal{T} \times \mathcal{S} \to \mathcal{X}$ and for a fixed conditional PMF $P_{U|STY}$, the conditional PMF

$$P_{USY|T=t}(u,s,y) = P_S(s)\,P_{\mathrm{c}}\big(y\big|f(s,t),s\big)\,P_{U|STY}(u|s,t,y) \quad (32)$$

can be viewed as a joint PMF on (U,S,Y) that is parameterized by t, and that the collection of all such PMFs is a connected compact subset of all the PMFs on (U,S,Y). Imposing $|\mathcal{Y}| - 1$ linear constraints as in [6, Appendix C] fixes the Y-marginal and hence $H(Y)$. Alternatively, we can fix the Y-marginal by fixing the (X,S)-marginal by imposing $|\mathcal{X}||\mathcal{S}| - 1$ linear constraints. Additionally imposing an expectation constraint on

$$H\left(\sum_{u,s} P_{USY|T=t}\right)$$

fixes $H(Y|T)$, with the result that $I(T;Y)$ is fixed. Lastly, an additional constraint on the expectation of

$$\mathbb{E}\big[d\big(S, g(U,T,Y)\big)\big|T = t\big]$$

fixes the estimation distortion. The number of constraints we imposed is $|\mathcal{Y}|+1$ or $|\mathcal{X}||\mathcal{S}|+1$, and it thus follows from the Support Lemma [6, Appendix C] that—irrespective of f, g, and $P_{U|STY}$—we can restrict the cardinality of $\mathcal{T}$ as in (29b).

The rest of the proof is in Section V-A. ∎

Discussion: With causal SI and feedback, the achievable rate equals the mutual information $I(T;Y)$ between the input "Shannon strategy" and the output. If, in addition to sending the message, the encoder must also provide a reconstructor that is cognizant of T and Y with a description of the state to within a desired distortion D, then the achievable rate is penalized by the conditional rate-distortion function $R_{S|TY}(D)$ for describing S when both the encoder and the decoder have access to (T,Y).

As with strictly-causal SI, in the lossless case with causal SI, feedback does not increase the RnS capacity:

Proposition 6 (Lossless Reconstruction: Causal SI): If the state is revealed to the encoder causally and we require lossless reconstruction, then

$$C^{\mathrm{c}}(0) = C_{\mathrm{FB}}^{\mathrm{c}}(0) = \max_{P_T,\, f} \big\{I(T;Y) - H(S|TY)\big\},$$

where I and H are computed with respect to the joint PMF

$$P_S(s)\,P_T(t)\,\mathbb{1}\big\{x = f(t,s)\big\}\,P_{\mathrm{c}}(y|x,s), \quad (33)$$

and the mapping f is as in Theorem 5.

Proof: See Section V-B. ∎

C. Noncausal State Information

For the noncausal case we only provide bounds on the feedback RnS capacity. For the purpose of stating the lower bound, define

$$R^{(\mathrm{l})} = \max_{P_{TX|S},\, P_{U|STY},\, g} \big\{I(T;Y) - I(T;S) - I(S;U|TY)\big\}, \quad (34)$$

$$= \max_{P_{TX|S},\, P_{U|STY},\, g} \big\{I(T;Y) - I(T;S) - R_{S|TY}(D)\big\}, \quad (35)$$

where U and T are auxiliary chance variables taking values in $\mathcal{U}$ and $\mathcal{T}$ respectively; the mutual informations are computed with respect to the joint PMF

$$P_{STXYU}(s,t,x,y,u) = P_S(s)\,P_{TX|S}(t,x|s)\,P_{\mathrm{c}}(y|x,s) \cdot P_{U|STY}(u|s,t,y); \quad (36)$$

and it is required that there be a mapping

$$g: \mathcal{U} \times \mathcal{T} \times \mathcal{Y} \to \hat{\mathcal{S}} \quad (37)$$

satisfying

$$\mathbb{E}\big[d\big(S, g(U,T,Y)\big)\big] \le D. \quad (38)$$

Note that in the above maximization we may restrict the cardinalities of $\mathcal{U}$ and $\mathcal{T}$ to satisfy

$$|\mathcal{U}| \le |\hat{\mathcal{S}}| \quad (39a)$$

and

$$|\mathcal{T}| \le |\mathcal{X}| \cdot |\mathcal{S}| + 1. \quad (39b)$$

Also note that we can also express $R^{(\mathrm{l})}$ as

$$R^{(\mathrm{l})} = \max_{P_{TX|S},\, P_{U|STXY},\, g} \big\{I(T;Y) - I(T;S) - I(SX;U|TY)\big\}, \quad (40a)$$

where the mutual informations are computed with respect to the joint PMF

$$P_{STXYU}(s,t,x,y,u) = P_S(s)\, P_{TX|S}(t,x|s)\, P_{\mathrm{c}}(y|x,s) \cdot P_{U|STXY}(u|s,t,x,y), \quad (40b)$$

and $g(\cdot)$ is as in (37) and satisfies (38). Indeed, the maximum in (40a) is achieved when U and X are conditionally independent given (S,T,Y), in which case (40a) reduces to (34). This form makes it easier to compare the lower bound with the upper bound that we present next.

To this end, define

$$R^{(\mathrm{u})} = \max_{P_{TX|S},\, P_{U|STXY},\, g} \min\big\{I(T;Y) - I(T;S),\; I(XS;Y) - I(S;UTY)\big\}, \quad (41)$$

where the mutual informations are computed w.r.t. (40b) under the constraint (38).

Theorem 7 (Noncausal SI and Feedback):

$$R^{(\mathrm{l})} \le C_{\mathrm{FB}}^{\mathrm{nc}}(D) \le R^{(\mathrm{u})}. \quad (42)$$

Moreover, if $R^{(\mathrm{u})}$ is attained by a law under which X and (U,Y) are conditionally independent given (S,T)—e.g., when X is a deterministic function of (S,T)—then the above inequalities both hold with equality.

Proof: See Section VI-A. ∎

Remark 2: The upper bound in (42) need not be tight.

Proof: See Section VI-B. ∎

In the lossless case with noncausal SI the bounds in Theorem 7 coincide:

Proposition 8 (Lossless Reconstruction: Noncausal SI): In the lossless case with noncausal SI

$$C^{\mathrm{nc}}(0) = C_{\mathrm{FB}}^{\mathrm{nc}}(0) = \max_{P_{TX|S}} \big\{I(T;Y) - I(T;S) - H(S|TY)\big\},$$

where I and H are computed with respect to the joint PMF

$$P_{STXY}(s,t,x,y) = P_S(s)\, P_{TX|S}(t|s)\, P_{\mathrm{c}}(y|x,s), \quad (43)$$

and the maximization over $P_{TX|S}$ can be restricted to conditional law $P_{TX|S}$ of the form $P_{T|S}\, P_{X|TS}$ with $P_{X|TS}$ being 0-1 valued, i.e., with X being a deterministic function of (S,T).

Proof: See Section VI-C. ∎

Remark 3 (Gaussian Channel: Noncausal SI): On the Gaussian channel with mean squared-error reconstruction distortion and noncausal SI, feedback does not increase the RnS capacity. The characterization of the achievable (R,D) pairs in [19, Th. 2] thus also holds in the presence of feedback.

Proof: This can be shown in either of the following ways.

- Modify the converse proof in [19, Sec. III.B] to account for feedback.
- Study—as we do in Section VI-D—the upper bound $R^{(\mathrm{u})}$ and show that it coincides with the (R,D) trade-off characterization of [19, Th. 2].
- Show that $R^{(\mathrm{u})}$ can be achieved with X being a deterministic function of (S,T), and then invoke Theorem 7 to deduce its tightness.[3] ∎

IV. Proofs—Strictly-Causal State Information

This section provides the proofs for the results related to strictly-causal state information. We begin with Theorem 1.

A. Proof of Theorem 1

Before proving Theorem 1, we denote the RHS of (11) by $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D)$ and study some of its properties. Recall that the maximum there is unaltered if we restrict, as we shall, the cardinality of $\mathcal{U}$ to that of $\hat{\mathcal{S}}$, which is finite. Let $C_{X\to Y}$ denote the channel's Shannon capacity when the state is revealed to neither encoder nor decoder, so

$$C_{X\to Y} \triangleq \max_{P_X} I(X;Y). \quad (44)$$

Proposition 9: The function $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\colon \mathbb{R}_+ \to \mathbb{R}_+$ is monotonically nondecreasing and upper bounded by the channel's Shannon capacity,

$$\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D) \le C_{X\to Y}, \quad D \in \mathbb{R}_+, \quad (45)$$

with equality whenever $D \ge d_{\max}$. Moreover, it is concave and continuous.

Proof: Monotonicity holds because the feasible set in the maximization defining $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D)$ is enlarged (or is unchanged) when the maximal-allowed distortion D is increased. When D is greater-equal $d_{\max}$ the constraint (14) is inactive, and the maximization in (11) is thus unconstrained. The maximum is then achieved by choosing U deterministic (so that the nonnegative term $I(S;U|XY)$ be zero) and by choosing P_X to maximize $I(X;Y)$. With this choice $I(X;Y) - I(S;U|XY)$ is equal to $C_{X\to Y}$, thus demonstrating that $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D)$ is equal to $C_{X\to Y}$ whenever $D \ge d_{\max}$. This and the monotonicity establishes (45) and the sufficient condition for equality.

To establish concavity, let $D^{(1)}, D^{(2)} \in \mathbb{R}_+$ and $0 < \lambda < 1$ be given. For each $\nu \in \{1,2\}$ let $P_X^{(\nu)}$, $P_{U^{(\nu)}|XSY}$, $g^{(\nu)}$ achieve $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D^{(\nu)})$, where $U^{(\nu)}$ takes values in the finite set $\mathcal{U}^{(\nu)}$. Let Q be a time-sharing random variable that is independent of S and that takes on the values 1 and 2 with probabilities λ and $1-\lambda$. Define

$$P_{X|Q}(x|q) = P_X^{(q)}(x) \quad (46a)$$

$$\tilde{\mathcal{U}} = \big\{(u^{(q)}, q)\colon q \in \{1,2\}, u^{(q)} \in \mathcal{U}^{(q)}\big\} \quad (46b)$$

$$\tilde{U} = (U^{(Q)}, Q) \quad (46c)$$

$$P_{\tilde{U}|XSY}(u^{(q)}, q|x,s,y) = Q(q) P_{U^{(q)}|XSY}(u^{(q)}|x,s,y), \quad (46d)$$

$$g(u^{(q)}, q, x, y) = g^{(q)}(u^{(q)}, x, y). \quad (46e)$$

[3] To show that $R^{(\mathrm{u})}$ can be achieved by such an X, one can proceed as follows: First choose $\rho \in (0,1)$ so that the RHS of (170a) will equal the RHS of (170b). Next note that (170b) holds with equality when (S,T,X,U) are jointly Gaussian, with U being equal to $\hat{S}$. When this is the case, we can express X as $\alpha S + \beta T + W'$, where W' is Gaussian and independent of (S,T). Conclude the argument by noting that choosing W' to be zero renders (163) tight.

Note that under the law $P_S\, P_X\, P_{\mathrm{c}}(y|x,s)\, P_{\tilde{U}|XSY}$

$$Q \multimap (X,Y) \multimap S \tag{47}$$

(i.e., forms a Markov chain) because integrating $U^{(Q)}$ out shows that $(Q,S,X,Y) \sim P_Q\, P_S\, P_{X|Q}\, P_{\mathrm{c}}(y|x,s)$, which factorizes as $P_{XQ}(x,q)\cdot\big(P_S(s)P_{\mathrm{c}}(y|x,s)\big)$, with $P_{XQ}(x,q)$ being a function of q and (x,y), and with $P_S(s)P_{\mathrm{c}}(y|x,s)$ being a function of s and (x,y).

Since $P_X^{(\nu)}$, $P_{U^{(\nu)}|XSY}$, $g^{(\nu)}$ give rise to a distortion that does not exceed $D^{(\nu)}$, the choice P_X, $P_{\tilde{U}|XSY}$, g gives rise to a distortion that does not exceed $\lambda D^{(1)}+(1-\lambda)D^{(2)}$ and is thus feasible for the maximization problem defining $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(\lambda D^{(1)}+(1-\lambda)D^{(2)}\big)$. By the concavity of mutual information in the input law, the mutual information corresponding to P_X is at least the λ-weighted average of the mutual informations corresponding to $P_X^{(1)}$ and $P_X^{(2)}$. And as to $I(S;\tilde{U}|XY)$,

$$\begin{aligned} I\big(S;Q,U^{(Q)}\big|XY\big) &= I(S;Q|XY)+I\big(S;U^{(Q)}\big|XYQ\big)\\ &= I\big(S;U^{(Q)}\big|XYQ\big)\\ &= \lambda I\big(S;U^{(1)}\big|XY,Q=1\big)\\ &\quad +(1-\lambda)I\big(S;U^{(2)}\big|XY,Q=2\big), \end{aligned} \tag{48}$$

where the second equality follows from (47).

The above choice of $P_X, \tilde{U}, P_{\tilde{U}|XSY}, g$ thus satisfies the $\lambda D^{(1)}+(1-\lambda)D^{(2)}$ distortion constraint and satisfies

$$I(X;Y)-I(S;\tilde{U}|XY)\geq \lambda\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(D^{(1)}\big)+(1-\lambda)\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(D^{(2)}\big).$$

Since this choice need not be optimal, the LHS of the above is only a lower bound on $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(\lambda D^{(1)}+(1-\lambda)D^{(2)}\big)$, and

$$\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(\lambda D^{(1)}+(1-\lambda)D^{(2)}\big)\geq \lambda\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(D^{(1)}\big)+(1-\lambda)\tilde{C}_{\mathrm{FB}}^{\text{s-c}}\big(D^{(2)}\big),$$

which establishes concavity.

Continuity on $(0,\infty)$ is a consequence of the concavity, so it remains to prove continuity (from above) at $D=0$. By the definition in [16, Sec. 10, p. 84], $\mathbb{R}_+$ is locally simplicial, hence the concavity of $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}$ implies its lower-semicontinuity relative to $\mathbb{R}_+$. It thus remains to prove upper-semicontinuity relative to $\mathbb{R}_+$, i.e., that if

$$D^{(\kappa)}\downarrow 0 \text{ as } \kappa\to\infty$$

then there exists a subsequence $\{\kappa_\nu\}$ such that

$$\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(0)\geq \lim_{\nu\to\infty}\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D^{(\kappa_\nu)}).$$

Let $P_X^{(\kappa)}$, $P_{U|XSY}^{(\kappa)}$, and $g^{(\kappa)}$, achieve $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D^{(\kappa)})$ with $\mathcal{U}=\{1,\ldots,|\hat{\mathcal{S}}|\}$. Since the number of functions from $\mathcal{Y}\times\mathcal{U}\times\mathcal{X}$ to $\hat{\mathcal{S}}$ is finite, we can choose a subsequence $\{\kappa_\nu\}$ such that the mappings $g^{(\kappa_\nu)}$ do not depend on ν and can therefore be denoted g; the input distributions $P_X^{(\kappa_\nu)}$ converge to some $P_X^{(0)}$; and the conditional laws $P_{U|XSY}^{(\kappa_\nu)}$ to some $P_{U|XSY}^{(0)}$.

The expectation on the LHS of (14)—when evaluated under $P_S\, P_X^{(\kappa_\nu)}\, P_{\mathrm{c}}\, P_{U|XSY}^{(\kappa_\nu)}$ and g—is sandwiched from below by zero and from above by $D^{(\kappa_\nu)}$ and hence converges to zero. Consequently, zero must also be its evaluation with respect to $P_S\, P_X^{(0)}\, P_{\mathrm{c}}\, P_{U|XSY}^{(0)}$ and g because the expectation is continuous with respect to P_X and $P_{U|XSY}$, and $P_X^{(\kappa_\nu)}$ converge to $P_X^{(0)}$ and $P_{U|XSY}^{(\kappa_\nu)}$ to $P_{U|XSY}^{(0)}$. Consequently, the triple $\big(P_X^{(0)}, P_{U|XSY}^{(0)}, g\big)$ is in the feasible set defining $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(0)$. The continuity of the mutual information implies that the limit $\lim_{\nu\to\infty}\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D^{(\kappa_\nu)})$ is equal to $I(X;Y)-I(S;U|XY)$ evaluated with respect to $P_S\, P_X^{(0)}\, P_{\mathrm{c}}(y|x,s)\, P_{U|XSY}^{(0)}$. And $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}(0)$ cannot be smaller than this limit because $\big(P_X^{(0)}, P_{U|XSY}^{(0)}, g\big)$ is in the feasible set defining it. ∎

We are now ready to prove the converse part of Theorem 1.

Proof (Proof of the converse part of Theorem 1): Consider any achievable pair (R,D), and let $\varepsilon>0$ be arbitrarily small but for now fixed. We will show that the achievability of (R,D) implies that

$$R-\varepsilon\leq \tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D+\varepsilon). \tag{49}$$

Since $\tilde{C}_{\mathrm{FB}}^{\text{s-c}}$ is continuous (Proposition 9), this implies (upon letting ε tend to zero from above) that

$$R\leq \tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D) \tag{50}$$

and thus establishes the converse.

To establish (49), let $n_0=n_0(\varepsilon)$ be sufficiently large so that for all $n\geq n_0$ there exists a blocklength-n code $\big(\{f_k\}_{k=1}^n,\phi_W,\phi_S\big)$ of rate

$$\frac{1}{n}\log|\mathcal{W}|\geq R-\varepsilon; \tag{51a}$$

fidelity

$$\mathbb{E}\big[d\big(S^n,\hat{S}^n\big)\big]\leq D+\varepsilon; \tag{51b}$$

and average probability of error $P_{\mathrm{e}}^{(n)}$ satisfying

$$\lim_{n\to\infty}P_{\mathrm{e}}^{(n)}=0. \tag{51c}$$

We will show that

$$\limsup_{n\to\infty}\frac{1}{n}\log|\mathcal{W}|\leq \tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D+\varepsilon), \tag{52}$$

from which (49) will follow using (51a).

It thus remains to establish (52). To simplify the typography, we denote the rate of the code $\big(\{f_k\}_{k=1}^n,\phi_W,\phi_S\big)$ by R', so

$$R'\triangleq\frac{1}{n}\log|\mathcal{W}|. \tag{53}$$

Draw W uniformly over $\mathcal{W}$, and let the random n-tuples S^n, X^n, Y^n, and $\hat{S}^n$ be the result of transmitting W over the channel using the encoder $X_k=f_k(W,S^{k-1},Y^{k-1})$ and of estimating the state sequence using ϕ_S. We first address the reliable transmission of the data by invoking Fano's inequality and (51c) to obtain

$$n(R'-\eta_n)\leq I(W;Y^n), \tag{54}$$

where

$$\lim_{n\to\infty}\eta_n=0. \tag{55}$$

We next address the state estimation. It is tempting to account for the computability of $\hat{S}^n$ from Y^n using the data processing inequality

$$I(\hat{S}^n;S^n)\leq I(Y^n;S^n) \tag{56}$$

and to then relate the estimation fidelity to $I(\hat{S}^n; S^n)$. But this will not do. To see why, consider the example of (1), when the state $S = (S^{(a)}, S^{(b)})$ has two independent components and likewise $\tilde{S} = (\tilde{S}^{(a)}, \tilde{S}^{(b)})$. Assume the factorization

$$P_{\tilde{S}|S} = P_{\tilde{S}^{(a)}|S^{(a)}} \, P_{\tilde{S}^{(b)}|S^{(b)}}$$

and that $\hat{S} = (\hat{S}^{(a)}, \hat{S}^{(b)})$ with $d(s, \hat{s})$ depending only on $(s^{(a)}, \hat{s}^{(a)})$. In this case it is intuitively clear that $\tilde{S}^{(b)}$ can be ignored by the encoder and that its joint law with $S^{(b)}$ is immaterial. But this joint law does influence $I(Y^n; S^n)$. Indeed, the latter is maximized when $\tilde{S}^{(b)}$ equals $S^{(b)}$, and it is minimized when the two are independent. The data processing inequality is in the former case thus too loose. To overcome this problem, we shall replace it with an identity and have to deal with the correction term. Additionally (for other reasons), we shall have to consider a conditional version thereof.

To derive the required identity, we begin by using the chain rule and the fact that $\hat{S}_k$ is a function of Y^n to obtain

$$\begin{aligned} &I(Y^n \hat{S}_k; S_k | W S^{k-1}) \\ &\quad = I(Y^n; S_k | W S^{k-1}) + I(\hat{S}_k; S_k | W Y^n S^{k-1}) \\ &\quad = I(Y^n; S_k | W S^{k-1}). \end{aligned} \tag{57}$$

Next, expanding the same term in the other order we obtain

$$\begin{aligned} &I(Y^n \hat{S}_k; S_k | W S^{k-1}) \\ &\quad = I(\hat{S}_k; S_k | W S^{k-1}) + I(Y^n; S_k | W \hat{S}_k S^{k-1}). \end{aligned} \tag{58}$$

From the two equally-valid expansions (57) and (58) we obtain the desired identity

$$\begin{aligned} &I(\hat{S}_k; S_k | W S^{k-1}) \\ &\quad = I(Y^n; S_k | W S^{k-1}) - I(Y^n; S_k | \hat{S}_k W S^{k-1}). \end{aligned} \tag{59}$$

To continue with the converse, define the auxiliary random variables

$$V_k \triangleq (W, S^{k-1}), \tag{60a}$$

$$U_k \triangleq Y^{n\setminus k} = (Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n), \tag{60b}$$

and note that for every $k \in [1 : n]$ the time-k state S_k is independent of V_k, and that $\hat{S}_k$ is a deterministic function (which depends on k) of (U_k, Y_k).

By (54) and (59),

$$\begin{aligned} &n(R' - \eta_n) + \sum_{k=1}^{n} I(\hat{S}_k; S_k | W S^{k-1}) \\ &\le I(W; Y^n) \\ &\quad + \sum_{k=1}^{n} \Big[I(Y^n; S_k | W S^{k-1}) - I(Y^n; S_k | W \hat{S}_k S^{k-1}) \Big] \\ &= I(W; Y^n) + I(Y^n; S^n | W) - \sum_{k=1}^{n} I(Y^n; S_k | W \hat{S}_k S^{k-1}) \\ &= I(W S^n; Y^n) - \sum_{k=1}^{n} I(Y^n; S_k | W \hat{S}_k S^{k-1}) \\ &= \sum_{k=1}^{n} \Big[I(Y_k; W S^n | Y^{k-1}) - I(S_k; Y^n | W \hat{S}_k S^{k-1}) \Big] \\ &\overset{(a)}{=} \sum_{k=1}^{n} \Big[I(Y_k; W X_k S^n | Y^{k-1}) - I(S_k; Y^n | W \hat{S}_k S^{k-1}) \Big] \\ &= \sum_{k=1}^{n} \Big[H(Y_k | Y^{k-1}) - H(Y_k | X_k W S^n Y^{k-1}) \\ &\qquad - I(S_k; Y^n | W \hat{S}_k S^{k-1}) \Big] \\ &\overset{(b)}{\le} \sum_{k=1}^{n} \Big[H(Y_k) - H(Y_k | X_k S_k) - I(S_k; Y^n | W \hat{S}_k S^{k-1}) \Big] \\ &= \sum_{k=1}^{n} \Big[I(Y_k; X_k S_k) - I(S_k; Y_k U_k | V_k \hat{S}_k) \Big] \\ &= \sum_{k=1}^{n} \Big[I(X_k; Y_k) + I(S_k; Y_k | X_k) - I(S_k; Y_k U_k | V_k \hat{S}_k) \Big] \\ &\overset{(c)}{=} \sum_{k=1}^{n} \Big[I(X_k; Y_k) + H(S_k) - H(S_k | Y_k X_k) - H(S_k | V_k \hat{S}_k) \\ &\qquad + H(S_k | X_k Y_k U_k V_k \hat{S}_k) \Big] \\ &\le \sum_{k=1}^{n} \Big[I(X_k; Y_k) + H(S_k) - H(S_k | Y_k X_k) - H(S_k | V_k \hat{S}_k) \\ &\qquad + H(S_k | X_k Y_k U_k) \Big] \\ &= \sum_{k=1}^{n} \Big[I(X_k; Y_k) - I(S_k; U_k | Y_k X_k) + I(S_k; V_k \hat{S}_k) \Big] \\ &= \sum_{k=1}^{n} \Big[I(X_k; Y_k) - I(S_k; U_k | Y_k X_k) + I(S_k; W S^{k-1} \hat{S}_k) \Big] \\ &\overset{(d)}{=} \sum_{k=1}^{n} \Big[I(X_k; Y_k) - I(S_k; U_k | Y_k X_k) + I(S_k; \hat{S}_k | W S^{k-1}) \Big]. \end{aligned} \tag{61}$$

Here

- (a) follows since X_k is a function of (W, S^{k-1}, Y^{k-1});
- (b) follows since $(W S^{n\setminus k} Y^{k-1})$ —○— $(X_k S_k)$ —○— Y_k forms a Markov chain, and conditioning cannot increase entropy;
- (c) follows since S_k and X_k are independent, and X_k is a function of (W, S^{k-1}, Y^{k-1}); and
- (d) follows because S_k is independent of (W, S^{k-1}).

Subtracting the sum that appears on both sides of (61), we obtain

$$n(R' - \eta_n) \le \sum_{k=1}^{n} \Big[I(X_k; Y_k) - I(S_k; U_k | Y_k X_k) \Big]. \tag{62}$$

Draw J uniformly from $\{1, \ldots, n\}$ independently of $\{(X_k, Y_k, S_k, U_k, \hat{S}_k),\ k = 1, \ldots, n\}$, and define the chance variables $U \triangleq (U_J, J)$, $S \triangleq S_J$, $Y \triangleq Y_J$, $X \triangleq X_J$, and $\hat{S} \triangleq \hat{S}_J$. Further define the function

$$g\big((u_j, j), x, y\big) = \phi_S^{(j)}\big(y^n\big)$$

so

$$\hat{S} = g(U, X, Y), \tag{63}$$

where $\phi_S^{(j)}\big(y^n\big)$—being the j-th component of the result of applying ϕ_S to y^n—is computable from ϕ_S and the tuple

$\big((u_j, j), x, y_j\big)$, because the tuple fully specifies both j and y^n. The value of $g\big((u_j, j), x, y\big)$ does not depend on x, but we have added x as an argument so that the mapping have the form (13) that appears in the direct part.

Using J we may express (62) as

$$\begin{aligned} R' - \eta_n &\le \frac{1}{n}\sum_{k=1}^{n}\Big[I(X_k; Y_k) - I(S_k; U_k|Y_k X_k)\Big] \\ &= I(X_J; Y_J|J) - I(S_J; U_J|Y_J, X_J, J) \\ &= I(X_J; Y_J|J) - I(S_J; U_J, J|Y_J, X_J) \\ &\quad + I(S_J; J|Y_J, X_J) \\ &\overset{(e)}{\le} I(X; Y) - I(S; U|XY). \end{aligned} \tag{64}$$

Here (e) follows because $I(X; Y) = I(P_X; W_{Y|X})$ is concave in P_X, and because the factorization

$$\begin{aligned} P_{J S_J X_J Y_J}(j, s, x, y) &= P_J(j)\, P_{S_J X_J Y_J|J}(s, x, y|j) \\ &= P_J(j)\, P_{X_j}(x)\, P_S(s)\, P_{\mathrm{c}}(y|x, s) \end{aligned} \tag{65}$$

shows that $J \multimap (X_J, Y_J) \multimap S_J$ and hence that $I(S_J; J|Y_J, X_J) = 0$.

Notice that the factorization (65) also implies (upon substituting $1/n$ for $P_J(j)$ and summing over j) that

$$P_{SXY}(s, x, y) = P_S(s)\left(\frac{1}{n}\sum_{j=1}^{n} P_{X_j}(x)\right) P_{\mathrm{c}}(y|x, s) \tag{66}$$

as in (12).

As to the expected distortion, starting from (51b),

$$\begin{aligned} D + \varepsilon &\ge \mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big] \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d\big(S_k, \hat{S}_k\big)\big] \\ &= \mathbb{E}\big[d\big(S_J, \hat{S}_J\big)\big] \\ &= \mathbb{E}\big[d\big(S, \hat{S}\big)\big] \\ &= \mathbb{E}\big[d\big(S, g(U, X, Y)\big)\big], \end{aligned} \tag{67}$$

where the second line follows from (3); the third by conditioning on J and then averaging over it; the fourth from the definition of the chance variables S and $\hat{S}$; and the last by (63).

It now follows from (64), (67), and the fact that by (66) the joint law of S, X, Y, U factorizes as in (12) that

$$R' - \eta_n \le \tilde{C}_{\mathrm{FB}}^{\text{s-c}}(D + \varepsilon), \tag{68}$$

which, in view of (55), establishes (52) and hence concludes the proof of (50). ■

Having established the converse part of Theorem 1, we now prove its direct part.

Proof of the direct part of Theorem 1: A sketch of the proof of the direct part was presented in the introduction. Here we provide some of the missing technical details.

Our coding scheme comprises B blocks, each of n channel uses. No attempt is made to estimate the state sequence pertaining to the last block. This may contribute up to $d_{\max}/B$ to our overall average reconstruction distortion. We choose B sufficiently large so that this penalty be negligible.

In Block-b we generate a blocklength-n, rate-$(R + R_s)$ random codebook whose codewords are chosen IID, with the components of each codeword being drawn IID according to the PMF P_X. This codebook, in combination with joint-typicality decoding, is used to send rate-R fresh information and a rate-R_s description of the Block-$(b-1)$ state sequence. The description is designed so that, based on this description, a reconstructor cognizant of the Block-$(b-1)$ codeword and output sequence will be able to reconstruct the Block-$(b-1)$ state sequence with average distortion that does not exceed D. When forming this description as Block-b is about to begin, the describer is cognizant of the Block-$(b-1)$ channel outputs (via the feedback link) and of the Block-$(b-1)$ codeword it sent in Block-$(b-1)$.

After observing the Block-b output sequence, the decoder can recover the fresh information and the state-description information as in the classical (stateless) channel coding theorem provided that $R + R_s$ is smaller than $I(X; Y)$ (and the blocklength is large enough).

The decoder then proceeds to guess the codeword that was sent in Block-$(b-1)$. Having done so, it assumes that its guess is correct and uses this guess, the Block-$(b-1)$ output sequence, and the state-description information that it decoded in Block-b to estimate the Block-$(b-1)$ state sequence.

Roughly speaking, the decoding and the state estimation steps should be successful if $R + R_s$ is smaller than $I(X; Y)$ and R_s exceeds $R_{S|XY}(D)$. There is, however, a delicate dependence issue that needs to be addressed: conditional on the Block-$(b-1)$ codeword being correctly decoded, the Block's state sequence is no longer IID P_S, so the application of the (conditional) rate-distortion theorem to this setting is tricky. This can be addressed by resorting to a "genie-aided decoder" [21, p. 419], [15]. The genie-aided decoder decodes the Block-$(b-1)$ codeword like we do and hence has the same probability of error in decoding the fresh information. However—unlike our decoder, which feeds this guess (and the Block-$(b-1)$ output sequence) to the estimator of the Block-$(b-1)$ state sequence—it feeds the *correct* Block-$(b-1)$ codeword to the state-estimation circuitry. The two decoders thus produce the same estimate of the state sequence whenever our decoder does not err. The difference in the distortions they incur is thus upper bounded by the product of the maximal distortion $d_{\max}$ and the maximal probability of decoding error. It thus tends to zero. ■

B. Proof of Proposition 3

The converse, which we prove with feedback, follows from Theorem 1: when the maximal-allowed Hamming distortion is zero, the distortion constraint (14) translates to $H(S|UXY)$ being zero and hence to $I(S; U|XY)$ being $H(S|XY)$, in which case (11) yields $R \le I(X; Y) - H(S|XY)$.

For the direct part we use the RnS capacity expression in Theorem 2, which does not utilize feedback, wherein the choice $U = S$ in (19) yields $D = 0$ and hence by (17) the rate $R = I(X; Y) - H(S|XY)$ is achievable.

C. Proof of Proposition 4

We prove the proposition by proving the converse with feedback and the achievability without it. For the converse we sketch two proofs. The first is based on (16) of Theorem 1[4]. The second is based on the data processing inequality $I(\hat{S}^n; S^n) \le I(Y^n; S^n)$, which here—unlike in the general case—is tight. Both allow for feedback, i.e., allow X_k to be of the form $X_k = X_k(W, S^{k-1}, Y^{k-1})$.

Proof 1 of the converse part of Proposition 4: To derive the converse from (16) we note that there is a one-to-one correspondence between the pairs (X, Y) and $(X, Y - X)$, so

$$\begin{aligned} R_{S|XY}(D) &= R_{S|X,Y-X}(D) \\ &= R_{S|S+Z}(D) \\ &= \frac{1}{2}\log^+\left(\frac{\sigma^2_{S|S+Z}}{D}\right) \\ &= \frac{1}{2}\log^+\left(\frac{\sigma_s^2\, N}{(\sigma_s^2+N)D}\right), \end{aligned} \tag{69}$$

where the second equality holds because X is independent of (S, Z); the third because S and Z are independent Gaussians, so the conditional distribution of S given $S + Z$ is Gaussian; and the fourth equality holds because the conditional variance is

$$\sigma^2_{S|S+Z} = \frac{\sigma_s^2\, N}{\sigma_s^2 + N}. \tag{70}$$

As to the term $I(X; Y)$ on the RHS of (16), we note that it is maximized by the Gaussian distribution, so

$$\left(\mathbb{E}\left[X^2\right] \le P\right) \implies \left(I(X;Y) \le \frac{1}{2}\log\left(1 + \frac{P}{\sigma_s^2 + N}\right)\right). \tag{71}$$

Combining (71), (69) and (16) yields the desired converse. ∎

Proof 2 of the converse part of Proposition 4: Our second proof is based on the data processing inequality. Let $R(\cdot)$ be the Rate-Distortion function for describing S_k in mean squared-error.

$$R(D) \triangleq \min_{P_{\hat{S}|S}:\, \mathbb{E}[d(S,\hat{S})] \le D} I(S; \hat{S}). \tag{72}$$

Since the state sequence is IID, it follows as in [4, eqs. (10.61)–(10.70)] that

$$\left(\mathbb{E}\left[d\left(S^n, \hat{S}^n\right)\right] \le D\right) \implies \left(nR(D) \le I(S^n; \hat{S}^n)\right). \tag{73}$$

[4]The use of (16) requires some justification, because Theorem 1 assumes finite alphabets and no cost constraints. Nevertheless, the converse does go through when we restrict D to be strictly positive (with continuity hence following from convexity).

By (54), (73) and the data processing inequality (56)

$$\begin{aligned} n(R' - \eta_n) + nR(D) &\le I(W; Y^n) + I(S^n; Y^n) \\ &\overset{(a)}{\le} I(W; Y^n|S^n) + I(Y^n; S^n) \\ &= I(WS^n; Y^n) \\ &= \sum_{k=1}^n I\left(Y_k; WS^n|Y^{k-1}\right) \\ &= \sum_{k=1}^n \Big[h(Y_k|Y^{k-1}) - h(Y_k|WS^nY^{k-1})\Big] \\ &\overset{(b)}{=} \sum_{k=1}^n \Big[h\left(Y_k|Y^{k-1}\right) - h\left(Y_k|X_kWS^nY^{k-1}\right)\Big] \\ &\overset{(c)}{=} \sum_{k=1}^n \Big[h\left(Y_k|Y^{k-1}\right) - h\left(Y_k|X_kS_k\right)\Big] \\ &\overset{(d)}{\le} \sum_{k=1}^n \Big[h(Y_k) - h(Y_k|X_kS_k)\Big] \\ &= \sum_{k=1}^n I(Y_k; X_kS_k) \\ &\overset{(e)}{\le} \frac{n}{2}\log\left(\frac{P + \sigma_s^2 + N}{N}\right). \end{aligned} \tag{74}$$

Here

- (a) holds because S^n is independent of W, and conditioning cannot increase entropy;
- (b) holds because $X_k = X_k(W, S^{k-1}, Y^{k-1})$ by (7);
- (c) holds because $(W, S^{n\setminus k}, Y^{k-1}) \multimap (X_k, S_k) \multimap Y_k$ is a Markov chain;
- (d) holds because conditioning cannot increase entropy; and
- (e) holds because X, S, and Z are independent, so the second moment of $X + S + Z$ cannot exceed $P + \sigma_s^2 + N$.

For the Gaussian memoryless source $\{S_k\}_{k=1}^\infty$ the rate-distortion function is

$$R(D) = \frac{1}{2}\log^+\left(\frac{\sigma_s^2}{D}\right), \tag{75}$$

and (74) therefore yields the desired converse

$$R_\gamma + \frac{1}{2}\log^+\left(\frac{\sigma_s^2}{D_\gamma}\right) \le \frac{1}{2}\log\left(\frac{P + \sigma_s^2 + N}{N}\right). \tag{76}$$

∎

Proof of the direct part of Proposition 4: We prove achievability using a scheme that does not utilize the feedback, so $X_k = X_k(W, S^{k-1})$. As in the direct part of Theorem 1, we use a Block-Markov scheme of B blocks. In Block b the transmitter uses a Gaussian codebook to transmit fresh information $W^{(b)}$ at rate R as well as a description $W_s^{(b)}$ at rate R_s of the Block-$(b-1)$ state sequence $\mathbf{S}^{(b-1)}$. The description is *à la* Wyner-Ziv: the transmitter describes $\mathbf{S}^{(b-1)}$ to a reconstructor that is cognizant of the difference between the Block-$(b-1)$ outputs $\mathbf{Y}^{(b-1)}$ and the Block-$(b-1)$ inputs $\mathbf{X}^{(b-1)}$. (Since we do not utilize feedback, the transmitter is incognizant of these outputs, and though it is cognizant of the inputs, it ignores this knowledge.) The difference between

the Block-$(b-1)$ outputs $\mathbf{Y}^{(b-1)}$ and the Block-$(b-1)$ inputs $\mathbf{X}^{(b-1)}$ thus serves as side information that is available to the reconstructor (once the transmitted symbols have been decoded) but not to the describer. And since the codebooks are Gaussian, this side-information is jointly Gaussian with the state $\mathbf{S}^{(b-1)}$. The Wyner-Ziv setting at hand is thus the Gaussian setting where performance is as good as if the side-information were also available to the describer.

In decoding the Block-$(b-1)$ transmission, the Wyner-Ziv description of $\mathbf{S}^{(b-1)}$ that was sent in Block-b is ignored. Reliable decoding can thus be achieved whenever

$$R + R_s < I(X;Y) = \frac{1}{2}\log\left(\frac{P+\sigma_s^2+N}{\sigma_s^2+N}\right). \tag{77}$$

Once the receiver decodes $\mathbf{X}^{(b-1)}$, it can subtract it from $\mathbf{Y}^{(b-1)}$ to obtain the side information. It then uses this side information and the Wyner-Ziv description $W_s^{(b)}$ of $\mathbf{S}^{(b-1)}$ that was sent in Block-b to estimate $\mathbf{S}^{(b-1)}$. The resulting expected distortion is then

$$\mathbb{E}\left[\left(\mathbf{S}^{(b-1)} - \mathbb{E}\big[\mathbf{S}^{(b-1)}\big|W_s^{(b)}, \mathbf{S}^{(b-1)} + \mathbf{Z}^{(b-1)}\big]\right)^2\right] = \frac{\sigma_s^2 N}{\sigma_s^2+N}\, e^{-2R_s}. \tag{78}$$

The achievability of (R_γ, D_γ) now follows by choosing

$$R = \frac{1-\gamma}{2}\log\left(\frac{P+\sigma_s^2+N}{\sigma_s^2+N}\right) - \varepsilon \tag{79}$$

and

$$R_s = \frac{\gamma}{2}\log\left(\frac{P+\sigma_s^2+N}{\sigma_s^2+N}\right) + \frac{\varepsilon}{2}. \tag{80}$$

■

D. Converse for Strictly-Causal SI without Feedback

We next show how the technique that we employed to prove the converse part of Theorem 1 can be used in order to provide an alternative proof for the converse in the absence of feedback, i.e., the converse part of Theorem 2. Define the auxiliary random variable $V_k \triangleq (W, S^{k-1})$ as in (60a), and define

$$U_k \triangleq (W, S^{k-1}, Y_{k+1}^n). \tag{81}$$

Note that, for every $k \in [1:n]$, the time-k state S_k is independent of V_k and that in the absence of feedback $U_k \multimap (X_k, S_k) \multimap Y_k$ forms a Markov chain. Furthermore, because $X_k = X_k(W, S^{k-1})$ (no-feedback), it follows—as noted in [3, Sec. II.B]—that

$$S_k \multimap (W, S^{k-1}, Y_k, Y_{k+1}^n) \multimap Y^{k-1} \tag{82}$$

forms a Markov chain. Indeed, X^{k-1} is a deterministic function of (W, S^{k-2}), and conditioned on (X^{k-1}, S^{k-1}) the random vector Y^{k-1} is independent of the other variables including Y_{k+1}^n.

As a result,

$$\begin{aligned}\hat{S}_k &\overset{(a)}{=} \hat{S}_k(Y^n)\\ &\overset{(b)}{=} \hat{S}_k\big(W, S^{k-1}, Y_{k+1}^n, Y_k, Y^{k-1}\big)\\ &= \hat{S}_k(U_k, Y_k, Y^{k-1}).\end{aligned}$$

Here

(*a*) follows since the reconstruction function is defined by (5); and

(*b*) follows since one can ignore (W, S^{k-1}) and take $\hat{S}_k$ to be a function of Y^n.

Moreover the Markov chain (82) and Lemma 1 in [3, Sec. II.B] ensure the existence of a reconstruction $\hat{S}_k^*(U_k, Y_k)$ which dominates $\hat{S}_k$ in the sense that

$$\mathbb{E}\big[d\big(S_k, \hat{S}_k^*(U_k, Y_k)\big)\big] \le \mathbb{E}\big[d\big(S_k, \hat{S}_k(U_k, Y_k, Y^{k-1})\big)\big]. \tag{83}$$

Thus, replacing $\hat{S}_k$ by $\hat{S}_k^*$, which is a deterministic function of (U_k, Y_k), does not increase the expected distortion. This observation interpreted as the "data processing inequality" for estimation has already been made in [3, Lemma 1]. Furthermore, the identity (59) continues to hold when we replace $\hat{S}_k$ by $\hat{S}_k^*$ because it builds on (58) (which is just the chain rule) and on (57), which continues to hold when we replace $\hat{S}_k$ by $\hat{S}_k^*$ because

$$\begin{aligned}&I(Y^n\hat{S}_k^*; S_k|WS^{k-1})\\ &\quad= I(Y^n; S_k|WS^{k-1}) + I(\hat{S}_k^*; S_k|WY^nS^{k-1})\\ &\quad\overset{(a)}{=} I(Y^n; S_k|WS^{k-1}),\end{aligned} \tag{84}$$

where (a) follows since $\hat{S}_k^*$ is a function of $(W, S^{k-1}, Y_k, Y_{k+1}^n)$.

Consequently, by (54) and (59) (with $\hat{S}_k$ replaced by $\hat{S}_k^*$),

$$\begin{aligned}&n(R'-\eta_n) + \sum_{k=1}^n I(\hat{S}_k^*; S_k|WS^{k-1})\\ &\le I(W;Y^n)\\ &\quad+\sum_{k=1}^n\Big[I\big(Y^n; S_k\big|WS^{k-1}\big) - I\big(Y^n; S_k\big|W\hat{S}_k^*S^{k-1}\big)\Big]\\ &= I(W;Y^n) + I(Y^n;S^n|W) - \sum_{k=1}^n I\big(Y^n; S_k\big|W\hat{S}_k^*S^{k-1}\big)\\ &= I(WS^n;Y^n) - \sum_{k=1}^n I\big(Y^n; S_k\big|W\hat{S}_k^*S^{k-1}\big)\\ &= \sum_{k=1}^n\Big[I(Y_k; WS^n|Y^{k-1}) - I(S_k; Y^n|W\hat{S}_k^*S^{k-1})\Big]\\ &\overset{(a)}{=} \sum_{k=1}^n\Big[I\big(Y_k; WX_kS^n\big|Y^{k-1}\big) - I\big(S_k; Y^n\big|W\hat{S}_k^*S^{k-1}\big)\Big]\\ &= \sum_{k=1}^n\Big[H(Y_k|Y^{k-1}) - H(Y_k|X_kWS^nY^{k-1})\\ &\qquad - I\big(S_k; Y^n\big|W\hat{S}_k^*S^{k-1}\big)\Big]\\ &\overset{(b)}{\le} \sum_{k=1}^n\Big[H(Y_k) - H(Y_k|X_kS_kU_k) - I\big(S_k; Y^n\big|W\hat{S}_k^*S^{k-1}\big)\Big]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) - I\big(S_k; Y^n \big| W \hat{S}_k^* S^{k-1}\big)\Big] \\
&= \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) - H\big(S_k \big| W \hat{S}_k^* S^{k-1}\big) \\
&\qquad + H\big(S_k \big| W Y^n \hat{S}_k^* S^{k-1}\big)\Big] \\
&\overset{(c)}{=} \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) - H\big(S_k \big| W \hat{S}_k^* S^{k-1}\big) \\
&\qquad + H\big(S_k \big| W X_k Y^n S^{k-1}\big)\Big] \\
&\overset{(d)}{=} \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) - H\big(S_k \big| W \hat{S}_k^* S^{k-1}\big) \\
&\qquad + H\big(S_k \big| W X_k Y_k Y_{k+1}^n S^{k-1}\big)\Big] \\
&= \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) - H\big(S_k \big| V_k \hat{S}_k^*\big) + H(S_k | X_k Y_k U_k)\Big] \\
&= \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) + I\big(S_k; V_k \hat{S}_k^*\big) - I(S_k; X_k Y_k U_k)\Big] \\
&\overset{(e)}{=} \sum_{k=1}^{n}\Big[I(Y_k; X_k S_k U_k) + I\big(S_k; \hat{S}_k^* \big| V_k\big) - I(S_k; X_k Y_k U_k)\Big] \\
&= \sum_{k=1}^{n}\Big[I(Y_k; X_k U_k) + I\big(S_k; \hat{S}_k^* \big| V_k\big) - I(S_k; X_k U_k)\Big] \\
&\overset{(f)}{=} \sum_{k=1}^{n}\Big[I(Y_k; X_k U_k) + I\big(S_k; \hat{S}_k^* \big| V_k\big) - I(S_k; U_k | X_k)\Big]. \qquad (85)
\end{aligned}
$$

Here

- (a) follows since X_k is a function of (W, S^{k-1});
- (b) follows since $(W S^{n\setminus k} Y^{n\setminus k})$ —○— $(X_k S_k)$ —○— Y_k forms a Markov chain hence

$$
\begin{aligned}
H(Y_k | W X_k S^n Y^{k-1}) &= H(Y_k | W S^{k-1} Y_{k+1}^n X_k S_k) \\
&= H(Y_k | X_k S_k U_k),
\end{aligned}
$$

  and since conditioning cannot increase entropy;
- (c) follows since $\hat{S}_k^*$ is a deterministic function of $(W, S^{k-1}, Y_k, Y_{k+1}^n)$, and X_k is a function of (W, S^{k-1});
- (d) follows since S_k —○— $(W, S^{k-1}, X_k, Y_k, Y_{k+1}^n)$ —○— Y^{k-1} forms a Markov chain as per (82);
- (e) follows because S_k is independent of (W, S^{k-1}); and
- (f) follows because X_k is independent of S_k.

Subtracting the sum that appears on both sides of (85), we obtain

$$
n(R' - \eta_n) \le \sum_{k=1}^{n}\Big[I(X_k U_k; Y_k) - I(S_k; U_k | X_k)\Big]. \qquad (86)
$$

Define now $\hat{S}^* \triangleq \hat{S}_J^*$, and $S \triangleq S_J$, and consider the expected distortion:

$$
\begin{aligned}
D + \varepsilon &\ge \mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big] \\
&= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d\big(S_k, \hat{S}_k\big)\big] \\
&\ge \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d\big(S_k, \hat{S}_k^*\big)\big] \\
&= \mathbb{E}\big[d\big(S_J, \hat{S}_J^*\big)\big] \\
&= \mathbb{E}\big[d\big(S, \hat{S}^*\big)\big] \\
&= \mathbb{E}\big[d\big(S, g(U, X, Y)\big)\big],
\end{aligned}
$$

where the second line follows from (3); the third by (83), the fourth by conditioning on J and then averaging over it; the fifth from the definition of the chance variables S and $\hat{S}^*$; and the last by the fact that $\hat{S}_k^*$ is computable from (U_k, X_k, Y_k). The translation of (86) into a single-letter form completes the proof.

V. Proofs—Causal State Information

This section provides the proofs for the results related to causal state information. We begin with Theorem 5.

A. Proof of Theorem 5

Before proving Theorem 5, we denote the RHS of (25) by $\tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D)$ and record some of its properties. Recall that the RHS of (25) is not altered if, as we do, we impose the cardinality bounds (29). We denote the capacity of the channel when the state is revealed causally to the encoder by $C_{X\to Y}^{\mathrm{c}}$, so [17]

$$
C_{X\to Y}^{\mathrm{c}} \triangleq \max_{P_T,\, f} I(T; Y), \qquad (87)
$$

where T is an auxiliary chance variable taking values in $\mathcal{T}$, the mapping f is from $\mathcal{T} \times \mathcal{S}$ to $\mathcal{X}$, and the mutual information is computed with respect to the joint distribution

$$
P_{STXY}(s, t, x, y) = P_S(s)\, P_T(t)\, \mathbb{1}\{x = f(t, s)\}\, P_{\mathrm{c}}(y|x, s).
$$

Proposition 10: The function $\tilde{C}_{\mathrm{FB}}^{\mathrm{c}} : \mathbb{R}_+ \to \mathbb{R}_+$ is monotonically nondecreasing and upper bounded by $C_{X\to Y}^{\mathrm{c}}$,

$$
\tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D) \le C_{X\to Y}^{\mathrm{c}}, \quad D \in \mathbb{R}_+, \qquad (88)
$$

with equality whenever $D \ge d_{\max}$. Moreover, it is concave and continuous.

Proof: The proof is similar to the proof of Proposition 9 and is thus omitted. The main difference is in replacing (46) with

$$
\begin{aligned}
\tilde{T} &= \big(T^{(Q)}, Q\big) && (89a)\\
\tilde{U} &= U^{(Q)} && (89b)\\
P_{\tilde{T}}\big(t^{(q)}, q\big) &= P_Q(q)\, P_{T^{(q)}}\big(t^{(q)}\big) && (89c)\\
P_{X|\tilde{T}S}\big(x \big| (t^{(q)}, q), s\big) &= \mathbb{1}\Big\{x = f^{(q)}\big(t^{(q)}, s\big)\Big\} && (89d)\\
P_{\tilde{U}|\tilde{T}SY}\big(u^{(q)}\big|(t^{(q)}, q), s, y\big) &= P_{U^{(q)}|T^{(q)}SY}\big(u^{(q)}|t^{(q)}, s, y\big), && (89e)\\
g\big(u^{(q)}, (t^{(q)}, q), y\big) &= g^{(q)}\big(u^{(q)}, t^{(q)}, y\big). && (89f)\\
& && (89g)
\end{aligned}
$$

■

We are now ready to prove the converse part of Theorem 5.

Proof of the converse part of Theorem 5: Consider any achievable pair (R, D), and let $\varepsilon > 0$ be arbitrarily small but

for now fixed. We will show that the achievability of (R, D) implies that

$$R - \varepsilon \le \tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D + \varepsilon). \tag{90}$$

Since $\tilde{C}_{\mathrm{FB}}^{\mathrm{c}}$ is continuous (Proposition 10), this implies (upon letting ε tend to zero from above) that

$$R \le \tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D) \tag{91}$$

and thus establishes the converse.

To establish (90), let $n_0 = n_0(\varepsilon)$ be sufficiently large so that for all $n \ge n_0$ there exists a blocklength-n code $\big(\{f_k\}_{k=1}^n, \phi_W, \phi_S\big)$ of rate

$$\frac{1}{n} \log |\mathcal{W}| \ge R - \varepsilon, \tag{92a}$$

fidelity

$$\mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big] \le D + \varepsilon, \tag{92b}$$

and average probability of error $P_{\mathrm{e}}^{(n)}$ satisfying

$$\lim_{n\to\infty} P_{\mathrm{e}}^{(n)} = 0. \tag{92c}$$

We will show that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{W}| \le \tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D + \varepsilon), \tag{93}$$

from which (90) will follow using (92a). It thus remains to establish (93).

Denote the code rate by R', so

$$R' \triangleq \frac{1}{n} \log |\mathcal{W}|. \tag{94}$$

Draw W uniformly over $\mathcal{W}$, and let the random n-tuples S^n, X^n, Y^n, and $\hat{S}^n$ be the result of transmitting W over the channel using the encoder $X_k = f_k(W, S^k, Y^{k-1})$ and of estimating the state sequence using ϕ_S. Using Fano's inequality and (92c),

$$n(R' - \eta_n) \le I(W; Y^n), \tag{95}$$

where

$$\lim_{n\to\infty} \eta_n = 0. \tag{96}$$

We now turn to the state estimation. Once again we use (59) instead of the data processing inequality. We also use the definitions in (60) and additionally define

$$T_k \triangleq (W, Y^{k-1}, S^{k-1}). \tag{97}$$

As before, V_k and S_k are independent, and $\hat{S}_k$ is a deterministic function of (U_k, Y_k).

By (95) and (59)

$$\begin{aligned}
&n(R' - \eta_n) + \sum_{k=1}^n I\big(\hat{S}_k; S_k \big| W S^{k-1}\big) \\
&\le I(W; Y^n) + \sum_{k=1}^n \Big[I(Y^n; S_k | W S^{k-1}) \\
&\quad - I\big(Y^n; S_k \big| W \hat{S}_k S^{k-1}\big)\Big] \\
&= I(W; Y^n) + I(S^n; Y^n | W) - \sum_{k=1}^n I\big(Y^n; S_k \big| W \hat{S}_k S^{k-1}\big) \\
&= I(W S^n; Y^n) - \sum_{k=1}^n I\big(Y^n; S_k \big| W \hat{S}_k S^{k-1}\big) \\
&= \sum_{k=1}^n \Big[I(Y_k; W S^n | Y^{k-1}) - I\big(S_k; Y^n \big| W \hat{S}_k S^{k-1}\big)\Big] \qquad (98) \\
&\overset{(a)}{=} \sum_{k=1}^n \Big[I(Y_k; T_k W S^n | Y^{k-1}) - I\big(S_k; Y^n | W \hat{S}_k S^{k-1}\big)\Big] \\
&= \sum_{k=1}^n \Big[H(Y_k | Y^{k-1}) - H(Y_k | T_k W S^n Y^{k-1}) \\
&\qquad - I\big(S_k; Y^n \big| W \hat{S}_k S^{k-1}\big)\Big] \\
&\overset{(b)}{\le} \sum_{k=1}^n \Big[H(Y_k) - H(Y_k | T_k S_k) - I\big(S_k; Y^n \big| W \hat{S}_k S^{k-1}\big)\Big] \\
&= \sum_{k=1}^n \Big[I(Y_k; T_k S_k) - I\big(S_k; Y_k U_k \big| V_k \hat{S}_k\big)\Big] \\
&= \sum_{k=1}^n \Big[I(T_k; Y_k) + I(S_k; Y_k | T_k) - I\big(S_k; Y_k U_k \big| V_k \hat{S}_k\big)\Big] \\
&\overset{(c)}{=} \sum_{k=1}^n \Big[I(T_k; Y_k) + H(S_k) - H(S_k | Y_k T_k) - H\big(S_k \big| V_k \hat{S}_k\big) \\
&\qquad + H\big(S_k \big| Y_k U_k T_k V_k \hat{S}_k\big)\Big] \\
&\le \sum_{k=1}^n \Big[I(T_k; Y_k) + H(S_k) - H(S_k | Y_k T_k) - H\big(S_k \big| V_k \hat{S}_k\big) \\
&\qquad + H(S_k | Y_k U_k T_k)\Big] \\
&= \sum_{k=1}^n \Big[I(T_k; Y_k) - I(S_k; U_k | Y_k T_k) + I\big(S_k; V_k \hat{S}_k\big)\Big] \\
&\overset{(d)}{=} \sum_{k=1}^n \Big[I(T_k; Y_k) - I(S_k; U_k | Y_k T_k) + I\big(S_k; \hat{S}_k \big| V_k\big)\Big]. \qquad (99)
\end{aligned}$$

Here

- (a) holds because T_k is a function of (W, S^{k-1}, Y^{k-1});
- (b) holds because $(W S^{n\setminus k} Y^{k-1})$ —○— $(X_k S_k)$ —○— Y_k forms a Markov chain and hence, since X_k is a function of (T_k, S_k), also $(W S^{n\setminus k} Y^{k-1})$ —○— $(T_k S_k)$ —○— Y_k forms a Markov chain (and because conditioning cannot increase entropy);
- (c) holds because S_k and T_k are independent, and because T_k is a function of (W, S^{k-1}, Y^{k-1}); and
- (d) holds because S_k is independent of V_k.

Subtracting the sum that appears on both sides of (99), we obtain

$$n(R' - \eta_n) \le \sum_{k=1}^n \Big[I(T_k; Y_k) - I(S_k; U_k | Y_k T_k)\Big]. \tag{100}$$

It can be verified that

1) T_k is independent of S_k;
2) $\hat{S}_k$ is a deterministic function of (U_k, Y_k); and
3) X_k is a deterministic function of (T_k, S_k).

Consequently, the joint law of $(S_k, X_k, T_k, Y_k, U_k, \hat{S}_k)$ can be expressed as

$$\begin{aligned} &P_{S_kT_kX_kY_kU_k\hat{S}_k}(s,t,x,y,u,\hat{s}) \\ &\quad = P_S(s)\, P_{T_k}(t)\, \mathbb{1}\{x = f_k(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_k|S_kT_kY_k}(u|s,t,y)\, \mathbb{1}\{\hat{s} = g_k(u,y)\}. \end{aligned}$$

Here $g_k(u,y)$ is the k-component of the result of applying ϕ_S to the n-vector obtained upon inserting y after the first $k-1$ components of the $(n-1)$-vector u.

Draw J uniformly from $\{1,\ldots,n\}$ independently of $\{(X_k, Y_k, S_k, T_k, U_k, \hat{S}_k),\ k = 1,\ldots,n\}$, and define the chance variables $U = (U_J, J)$, $T = T_J$, $S = S_J$, $Y = Y_J$, $X = X_J$, and $\hat{S} = \hat{S}_J$. The chance variable T_j, which is defined in (97), specifies not only W, Y^{j-1}, and S^{j-1} but also (implicitly) j (via the length of Y^{j-1} and S^{j-1}). Consequently, we can define the function $f(t_j, s)$ as

$$f(t_j, s) = f_j(t_j, s),$$

i.e., as the time-j channel input X_j, so

$$X = f(T, S).$$

We further define the function

$$g\big((u_j, j), t_j, y\big) = \phi_S^{(j)}(y^n)$$

so

$$\hat{S} = g(U, T, Y), \tag{101}$$

where $\phi_S^{(j)}(y^n)$—being the j-th component of the result of applying ϕ_S to y^n—is computable from ϕ_S and the tuple $\big((u_j, j), t_j, y_j\big)$, because the tuple specifies both j and y^n. The value of $g\big((u_j, j), t_j, y\big)$ does not depend on t_j, but we have added t_j as an argument so that the mapping have the form (27) that appears in the direct part.

Using J we may express (100) as

$$\begin{aligned} R' - \eta_n &\le \frac{1}{n}\sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(S_k; U_k|Y_kT_k)\Big] \\ &= I(T_J; Y_J|J) - I(S_J; U_J|Y_J, T_J, J) \\ &= H(T_J|J) - H(T_J|J, Y_J) - H(S_J|Y_J, T_J, J) \\ &\quad + H(S_J|U_J, Y_J, T_J, J) \\ &\overset{(e)}{\le} H(T_J) - H(T_J|Y_J) - H(S_J|Y_J, T_J) \\ &\quad + H(S_J|U_J, Y_J, T_J, J) \\ &= I(T_J; Y_J) - I(S_J; U_J, J|Y_J, T_J) \\ &= I(T; Y) - I(S; U|TY). \end{aligned} \tag{102}$$

Here (e) holds because the factorization

$$\begin{aligned} &P_{JS_JT_JX_JY_JU_J}(j,s,t,x,y,u) \\ &\quad = P_J(j)\, P_{S_JT_JX_JY_J|J}(s,t,x,y|j) \\ &\qquad \cdot P_{U_J|S_JT_JX_JY_JJ}(u|s,t,x,y,j) \\ &\quad = P_J(j)\, P_{T_j}(t)\, P_S(s)\, \mathbb{1}\{x = f_j(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_j|S_jT_jY_jj}(u|s,t,y,j) \\ &\quad = P_{T_J}(t)\, P_S(s)\, \mathbb{1}\{x = f_J(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_JJ|S_JT_JY_J}(u,j|s,t,y) \end{aligned} \tag{103}$$

shows that $J \multimap Y_J \multimap T_J$ and $J \multimap (T_J, Y_J) \multimap S_J$ are Markov chains. Furthermore, based on the definition of the chance variables (S, T, X, Y, U) the factorization (103) is in accordance with (26).

As to the expected distortion, starting from (92b),

$$\begin{aligned} D + \varepsilon &\ge \mathbb{E}\big[d(S^n, \hat{S}^n)\big] \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d(S_k, \hat{S}_k)\big] \\ &= \mathbb{E}\big[d(S_J, \hat{S}_J)\big] \\ &= \mathbb{E}\big[d(S, \hat{S})\big] \\ &= \mathbb{E}\big[d(S, g(U, T, Y))\big], \end{aligned} \tag{104}$$

where the last line follows by (101).

It now follows from (102), (104), and the fact that the joint law of S, T, X, Y, U factorizes as in (26) that

$$R' - \eta_n \le \tilde{C}_{\mathrm{FB}}^{\mathrm{c}}(D + \varepsilon), \tag{105}$$

which, in view of (96), establishes (93) and hence concludes the proof of (91).

Having established the converse part of Theorem 5, we now prove its direct part.

Proof of the direct part of Theorem 5: The proof is very similar to the proof of the direct part of Theorem 1. The main difference is that the codebook (which we use in Block-b to send the Block-b fresh information and the description information related to the Block-$(b-1)$ state sequence) must utilize the state information. To this end we follow [17] and, instead of having the components of the codewords be input symbols, we use Shannon strategies.

B. Proof of Proposition 6

For the direct part we use Shannon's code construction for a SD-DMC with causal receiver SI [17] in combination with Slepian and Wolf [18] lossless source coding of the state sequence conditioned on the decoded codeword T and with reconstructor-SI Y. Since the Slepian-Wolf encoding rate is $R_{\mathrm{SW}} = H(S|TY)$ the rate $I(T;Y) - R_{\mathrm{SW}} = I(T;Y) - H(S|TY)$ is achievable.

The converse, which we prove with feedback, follows from Theorem 5: when the maximal-allowed Hamming distortion is zero, the distortion constraint (28) translates to $H(S|UTY)$ being zero and hence to $I(S; U|TY)$ being $H(S|TY)$, in which case (25) yields $R \le I(T;Y) - H(S|TY)$.

C. Converse for Causal SI without Feedback

We next show how the technique that we employed to prove the converse part of Theorem 5 can be used in order to provide an alternative proof of the converse in the absence of feedback, namely, the converse part of [3, Th. 3], a theorem that we recall below.

Theorem 11 (Causal SI and No Feedback [3, Th. 3]):

$$C^{\mathrm{c}}(D) = \max_{P_T,\, P_{U|TS},\, f}\big\{I(UT; Y) - I(UT; S)\big\}, \tag{106}$$

where the maximum is over all joint PMFs of the form

$$P_{STXYU}(s,t,x,y,u) = P_S(s)\, P_T(t)\, \mathbb{1}\{x = f(t,s)\}$$

$$\cdot P_{\mathrm{c}}(y|x,s)P_{U|TS}(u|t,s) \quad (107)$$

for which there exists a mapping

$$g: \mathcal{U} \times \mathcal{T} \times \mathcal{Y} \to \hat{\mathcal{S}} \quad (108)$$

satisfying

$$\mathbb{E}\big[d\big(S, g(U,T,Y)\big)\big] \le D, \quad (109)$$

where the expectation and the mutual informations are computed with respect to the above P_{STXYU}.

To prove the converse, we use the definitions in (60) and additionally define $T_k \triangleq V_k = (W, S^{k-1})$. As before, we note that V_k (and hence also T_k) is independent of S_k, and that $\hat{S}_k$ is a deterministic function of (U_k, Y_k). Furthermore, since $X_k = X_k(T_k, S_k)$ and there is no feedback,

$$U_k \multimap (T_k, S_k) \multimap Y_k \quad (110)$$

forms a Markov chain. By (98),

$$\begin{aligned}
&n(R'-\eta_n) + \sum_{k=1}^{n} I\big(\hat{S}_k; S_k \big| W S^{k-1}\big) \\
&\le \sum_{k=1}^{n}\Big[I(Y_k; WS^n|Y^{k-1}) - I\big(S_k; Y^n\big|W\hat{S}_k S^{k-1}\big)\Big] \\
&\overset{(a)}{=} \sum_{k=1}^{n}\Big[I(Y_k; T_k WS^n|Y^{k-1}) - I\big(S_k; Y^n|W\hat{S}_k S^{k-1}\big)\Big] \\
&= \sum_{k=1}^{n}\Big[H(Y_k|Y^{k-1}) - H(Y_k|T_k WS^n Y^{k-1}) \\
&\qquad - I\big(S_k; Y^n\big|W\hat{S}_k S^{k-1}\big)\Big] \\
&\overset{(b)}{\le} \sum_{k=1}^{n}\Big[H(Y_k) - H(Y_k|T_k S_k) - I\big(S_k; Y^n\big|W\hat{S}_k S^{k-1}\big)\Big] \\
&= \sum_{k=1}^{n}\Big[I(Y_k; T_k S_k) - I\big(S_k; Y_k U_k\big|V_k\hat{S}_k\big)\Big] \\
&= \sum_{k=1}^{n}\Big[I(T_k; Y_k) + I(S_k; Y_k|T_k) - I\big(S_k; Y_k U_k\big|T_k\hat{S}_k\big)\Big] \\
&\overset{(c)}{=} \sum_{k=1}^{n}\Big[I(T_k; Y_k) + H(S_k) - H(S_k|Y_k T_k) - H\big(S_k\big|T_k\hat{S}_k\big) \\
&\qquad + H\big(S_k\big|Y_k U_k T_k\hat{S}_k\big)\Big] \\
&= \sum_{k=1}^{n}\Big[I(T_k; Y_k) + H(S_k) - H(S_k|Y_k T_k) - H\big(S_k|T_k\hat{S}_k\big) \\
&\qquad + H(S_k|Y_k U_k T_k)\Big] \\
&= \sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(S_k; U_k|Y_k T_k) + I\big(S_k; T_k\hat{S}_k\big)\Big] \\
&\overset{(d)}{=} \sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(S_k; U_k|Y_k T_k) + I\big(S_k; \hat{S}_k\big|T_k\big)\Big]. \quad (111)
\end{aligned}$$

Here

(*a*) holds because $T_k = (W, S^{k-1})$;
(*b*) holds because $(W, S_{k+1}^n, Y^{k-1}) \multimap (T_k, S_k) \multimap Y_k$ forms a Markov chain (and because conditioning cannot increase entropy);
(*c*) holds because S_k and T_k are independent; and
(*d*) holds because S_k is independent of T_k.

Subtracting the sum that appears on both sides of (111), we obtain

$$n(R'-\eta_n) \le \sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(S_k; U_k|Y_k T_k)\Big]. \quad (112)$$

The joint law of $(S_k, X_k, T_k, Y_k, U_k, \hat{S}_k)$ factorizes as

$$\begin{aligned}
&P_{S_k T_k X_k Y_k U_k \hat{S}_k}(s,t,x,y,u,\hat{s}) \\
&\quad = P_S(s)\, P_{T_k}(t)\, \mathbb{1}\{x = f_k(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\
&\qquad \cdot P_{U_k|S_k T_k}(u|s,t)\, \mathbb{1}\{\hat{s} = g_k(u,y)\}, \quad (113)
\end{aligned}$$

where $g_k(u,y)$—as in the case with feedback—is the k-th component of the reconstruction mapping when applied to the n-vector obtained upon inserting y after the first $k-1$ components of the $(n-1)$-vector u.

Draw J uniformly from $\{1,\ldots,n\}$ independently of $\{(X_k, Y_k, S_k, T_k, U_k, \hat{S}_k),\ k = 1,\ldots,n\}$, and define the chance variables $U = (U_J, J)$, $T = T_J$, $S = S_J$, $Y = Y_J$, $X = X_J$, and $\hat{S} = \hat{S}_J$. As before, the chance variable T_j, which is defined as (W, S^{j-1}), specifies also j, and we can thus define the function $f\big(t_j, s\big)$ as

$$f\big(t_j, s\big) = f_j(t_j, s),$$

i.e., as the time-j channel input X_j, so

$$X = f(T, S).$$

We further define the function

$$g\big((u_j, j), t_j, y\big) = \phi_S^{(j)}\big(y^n\big)$$

so

$$\hat{S} = g(U, T, Y), \quad (114)$$

where $\phi_S^{(j)}\big(y^n\big)$—being the j-th component of the result of applying the state-reconstruction mapping ϕ_S to y^n—is computable from ϕ_S and the tuple $\big((u_j, j), t_j, y_j\big)$, because the tuple specifies both j and y^n.

Using J, we express (112) as

$$\begin{aligned}
R'-\eta_n &\le \frac{1}{n}\sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(S_k; U_k|Y_k T_k)\Big] \\
&= I(T_J; Y_J|J) - I(S_J; U_J|Y_J, T_J, J) \\
&= H(T_J|J) - H(T_J|J, Y_J) - H(S_J|Y_J, T_J, J) \\
&\quad + H(S_J|U_J, Y_J, T_J, J) \\
&\overset{(e)}{\le} H(T_J) - H(T_J|Y_J) - H(S_J|Y_J, T_J) \\
&\quad + H(S_J|U_J, Y_J, T_J, J) \\
&= I(T_J; Y_J) - I(S_J; U_J, J|Y_J, T_J) \\
&= I(T; Y) - I(S; U|TY). \quad (115)
\end{aligned}$$

Here (e) holds because the factorization

$$\begin{aligned} &P_{JS_JT_JX_JY_JU_J}(j,s,t,x,y,u) \\ &\quad = P_J(j)\, P_{S_JT_JX_JY_J|J}(s,t,x,y|j) \\ &\qquad \cdot P_{U_J|S_JT_JX_JY_JJ}(u|s,t,x,y,j) \\ &\quad = P_J(j)\, P_{T_j}(t)\, P_S(s)\, \mathbb{1}\{x = f_j(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_j|S_jT_jj}(u|s,t,j) \\ &\quad = P_{T_J}(t)\, P_S(s)\, \mathbb{1}\{x = f_J(t,s)\}\, P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_JJ|S_JT_J}(u,j|s,t) \end{aligned} \tag{116}$$

shows that $J \multimap Y_J \multimap T_J$ and $J \multimap (T_J, Y_J) \multimap S_J$ are Markov chains. Furthermore, based on the definition of the chance variables (S, T, X, Y, U) the factorization (116) is in accordance with (107).

Finally, to establish the equivalence of (115) and (106) note that, under the law (107), the functional on the RHS of (106) may be expressed as follows

$$\begin{aligned} &I(UT;Y) - I(UT;S) \\ &\quad \overset{(a)}{=} I(UT;Y) - I(U;S|T) \\ &\quad = I(T;Y) + I(U;Y|T) - I(U;S|T) \\ &\quad = I(T;Y) - H(U|TY) + H(U|ST) \\ &\quad \overset{(b)}{=} I(T;Y) - H(U|TY) + H(U|STY) \\ &\quad = I(T;Y) - I(S;U|TY). \end{aligned}$$

Here (a) holds because T is independent of S, and (b) holds because $U \multimap (S,T) \multimap Y$ forms a Markov chain.

VI. Proofs—Noncausal State Information

In this section we provide the proofs for the results related to strictly-causal state information. We begin with a proof of Theorem 7.

A. Proof of Theorem 7

Before proving Theorem 7, we denote the RHS of (41) by $\tilde{R}^{(\mathrm{u})}(D)$ and record some of its properties. Recall that the RHS of (41) is not altered if, as we do, we impose the cardinality bounds (39).

We denote the Gel'fand-Pinsker capacity of the channel when the state is revealed to the encoder noncausally by $C^{\text{G-P}}_{X\to Y}$, so [7]

$$C^{\text{G-P}}_{X\to Y} \triangleq \max_{P_{T|S},\, f} \big\{ I(T;Y) - I(T;S) \big\}, \tag{117}$$

where T is auxiliary chance variable taking values in $\mathcal{T}$; the mapping f is from $\mathcal{T} \times \mathcal{S}$ to $\mathcal{X}$, and the mutual information is computed with respect to the joint distribution P_{STXY} that is given by

$$\begin{aligned} P_{STXY}(s,t,x,y) = P_S(s)\, P_{T|S}(t|s)\, \mathbb{1}\big\{x = f(t,s)\big\} \\ \cdot P_{\mathrm{c}}(y|x,s). \end{aligned}$$

Proposition 12: The function $\tilde{R}^{(\mathrm{u})} \colon \mathbb{R}_+ \to \mathbb{R}_+$ is monotonically nondecreasing and upper bounded by $C^{\text{G-P}}_{X\to Y}$, so

$$\tilde{R}^{(\mathrm{u})}(D) \le C^{\text{G-P}}_{X\to Y}, \quad D \in \mathbb{R}_+, \tag{118}$$

with equality whenever $D \ge d_{\max}$. Moreover, it is concave and continuous.

Proof: The proof is similar to the proof of Proposition 9 and is thus omitted. The main difference is in replacing (46) with

$$\tilde{T} = \big(T^{(Q)}, Q\big) \tag{119a}$$

$$\tilde{U} = U^{(Q)} \tag{119b}$$

$$P_{\tilde{T}X|S}\big((t^{(q)},q),x\big|s\big) = P_Q(q)\, P_{T^{(q)}X|S}\big(t^{(q)},x|s\big) \tag{119c}$$

$$g\big(u^{(q)},(t^{(q)},q),y\big) = g^{(q)}\big(u^{(q)},t^{(q)},y\big), \tag{119d}$$

and

$$\begin{aligned} &P_{\tilde{U}|\tilde{T}SXY}\big(u^{(q)}\big|(t^{(q)},q),s,x,y\big) \\ &\qquad = P_{U^{(q)}|T^{(q)}SXY}\big(u^{(q)}|t^{(q)},s,x,y\big). \end{aligned} \tag{119e}$$

■

Proof of the upper bound in Theorem 7: Consider any achievable pair (R, D), and let $\varepsilon > 0$ be arbitrarily small but for now fixed. We will show that the achievability of (R, D) implies that

$$R - \varepsilon \le \tilde{R}^{(\mathrm{u})}(D + \varepsilon). \tag{120}$$

Since $\tilde{R}^{(\mathrm{u})}$ is continuous (Proposition 12), this implies (upon letting ε tend to zero from above) that

$$R \le \tilde{R}^{(\mathrm{u})}(D) \tag{121}$$

and thus establishes the upper bound.

To establish (120), let $n_0 = n_0(\varepsilon)$ be sufficiently large so that for all $n \ge n_0$ there exists a blocklength-n code $\big(\{f_k\}_{k=1}^n, \phi_W, \phi_S\big)$ of rate

$$\frac{1}{n} \log |\mathcal{W}| \ge R - \varepsilon; \tag{122a}$$

fidelity

$$\mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big] \le D + \varepsilon; \tag{122b}$$

and average probability of error $P_{\mathrm{e}}^{(n)}$ satisfying

$$\lim_{n\to\infty} P_{\mathrm{e}}^{(n)} = 0. \tag{122c}$$

We will show that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{W}| \le \tilde{R}^{(\mathrm{u})}(D + \varepsilon), \tag{123}$$

from which (120) will follow using (122a).

Define R' as in (53), namely, as $n^{-1} \log |\mathcal{W}|$; draw W uniformly over $\mathcal{W}$; and let the random n-tuples S^n, X^n, Y^n, and $\hat{S}^n$ be the result of transmitting W over the channel using the encoder $X_k = f_k(W, S^n, Y^{k-1})$ and of estimating the state sequence using ϕ_S. Using Fano's inequality and (122c),

$$n(R' - \eta_n) \le I(W; Y^n), \tag{124}$$

where

$$\lim_{n\to\infty} \eta_n = 0. \tag{125}$$

We next turn to the estimation. Instead of (59), we shall use the identity

$$\begin{aligned} &I(Y^n; S_k | W S_{k+1}^n) \\ &\qquad = I(\hat{S}_k; S_k | W S_{k+1}^n) + I(Y^n; S_k | W \hat{S}_k S_{k+1}^n), \end{aligned} \tag{126}$$

which can be proved as follows. Using the chain rule and the fact $\hat{S}_k$ is a function of Y^n,

$$\begin{aligned} &I(Y^n\hat{S}_k; S_k|WS_{k+1}^n) \\ &\quad = I(Y^n; S_k|WS_{k+1}^n) + I(\hat{S}_k; S_k|WY^nS_{k+1}^n) \\ &\quad = I(Y^n; S_k|WS_{k+1}^n). \end{aligned} \tag{127}$$

And expanding in the other order,

$$\begin{aligned} &I(Y^n\hat{S}_k; S_k|WS_{k+1}^n) \\ &\quad = I(\hat{S}_k; S_k|WS_{k+1}^n) + I(Y^n; S_k|W\hat{S}_kS_{k+1}^n). \end{aligned} \tag{128}$$

The identity (126) now follows from (127) and (128).

Having established (126), we now define the auxiliary chance variables

$$T_k \triangleq (W, S_{k+1}^n, Y^{k-1}), \quad U_k \triangleq Y_{k+1}^n, \tag{129}$$

and note that

$$T_k \multimap (X_k, S_k) \multimap Y_k \tag{130a}$$

is a Markov chain, and that

$$\hat{S}_k = \hat{S}_k(U_k, T_k, Y_k). \tag{130b}$$

By (124) and (126)

$$\begin{aligned} &n(R'-\eta_n) + \sum_{k=1}^n I\big(\hat{S}_k; S_k\big|WS_{k+1}^n\big) \\ &\le I(W; Y^n) \\ &\quad + \sum_{k=1}^n\Big[I\big(Y^n; S_k\big|WS_{k+1}^n\big) - I\big(Y^n; S_k\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &= I(W; Y^n) + I(S^n; Y^n|W) - \sum_{k=1}^n I\big(Y^n; S_k\big|W\hat{S}_kS_{k+1}^n\big) \\ &= I(WS^n; Y^n) - \sum_{k=1}^n I\big(Y^n; S_k\big|W\hat{S}_kS_{k+1}^n\big) \\ &= \sum_{k=1}^n\Big[I\big(Y_k; WS^n\big|Y^{k-1}\big) - I\big(S_k; Y^n\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &\overset{(a)}{=} \sum_{k=1}^n\Big[I\big(Y_k; WX_kS^n\big|Y^{k-1}\big) - I\big(S_k; Y^n\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &= \sum_{k=1}^n\Big[H\big(Y_k\big|Y^{k-1}\big) - H\big(Y_k\big|X_kWS^nY^{k-1}\big) \\ &\qquad - I\big(S_k; Y^n\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &\overset{(b)}{\le} \sum_{k=1}^n\Big[H(Y_k) - H\big(Y_k\big|X_kS_k\big) - I\big(S_k; Y^n\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &= \sum_{k=1}^n\Big[I(Y_k; X_kS_k) - I\big(S_k; Y^n\big|W\hat{S}_kS_{k+1}^n\big)\Big] \\ &= \sum_{k=1}^n\Big[I(Y_k; X_kS_k) + H(S_k) - H\big(S_k\big|W\hat{S}_kS_{k+1}^n\big) \\ &\qquad - H(S_k) + H\big(S_k\big|WY^n\hat{S}_kS_{k+1}^n\big)\Big] \\ &\overset{(c)}{=} \sum_{k=1}^n\Big[I(Y_k; X_kS_k) + I(S_k; W\hat{S}_kS_{k+1}^n) - I(S_k; T_kU_kY_k)\Big] \\ &\overset{(d)}{=} \sum_{k=1}^n\Big[I(Y_k; X_kS_k) + I\big(S_k; \hat{S}_k\big|WS_{k+1}^n\big) \\ &\qquad - I(S_k; T_kU_kY_k)\Big]. \end{aligned} \tag{131}$$

Here

(*a*) holds because X_k is a function of (W, S^n, Y^{k-1});
(*b*) holds because $(WS^{n\setminus k}Y^{k-1}) \multimap (X_kS_k) \multimap Y_k$ forms a Markov chain, and because conditioning cannot increase entropy;
(*c*) holds because $\hat{S}_k$ is a function of Y^n, and by (129); and
(*d*) follows because S_k is independent of (W, S_{k+1}^n).

Subtracting the sum that appears on both sides of (131), we obtain

$$n(R'-\eta_n) \le \sum_{k=1}^n\Big[I(X_kS_k; Y_k) - I(S_k; U_kT_kY_k)\Big]. \tag{132}$$

Note that, by (130), the joint law of $(S_k, X_k, T_k, Y_k, U_k, \hat{S}_k)$ factorizes as

$$\begin{aligned} &P_{S_kT_kX_kY_kU_k\hat{S}_k}(s,t,x,y,u,\hat{s}) \\ &\quad = P_S(s)\,P_{T_k|S_k}(t|s)\,P_{X_k|S_kT_k}(x|s,t)\,P_{\mathrm{c}}(y|x,s) \\ &\qquad \cdot P_{U_k|S_kT_kX_kY_k}(u|s,t,x,y)\,\mathbb{1}\{\hat{s} = g_k(u,t,y)\}. \end{aligned}$$

Here $g_k(u,t,y)$ is the k-th component of the result of applying ϕ_S to the n-vector y^n that is obtained from (u,t,y) by appending y followed by the $(n-k)$-vector u to the last $k-1$ components of the vector t.

Draw J uniformly from $\{1,\ldots,n\}$ independently of $\{(X_k, Y_k, S_k, T_k, U_k, \hat{S}_k),\ k = 1,\ldots,n\}$, and define the chance variables $U = (U_J, J)$, $T = T_J$, $S = S_J$, $Y = Y_J$, $X = X_J$, and $\hat{S} = \hat{S}_J$. Define also the function

$$g\big((u_j, j), t_j, y\big) = \phi_S^{(j)}\big(y^n\big)$$

so

$$\hat{S} = g(U, T, Y), \tag{133}$$

where $\phi_S^{(j)}\big(y^n\big)$, being the j-th component of the result of applying ϕ_S to y^n, is computable from ϕ_S and the tuple $\big((u_j, j), t_j, y_j\big)$ because the tuple fully specifies both j and y^n.

Using J we may express (132) as

$$\begin{aligned} R'-\eta_n &\le \frac{1}{n}\sum_{k=1}^n\Big[I(X_kS_k; Y_k) - I(S_k; U_kY_kT_k)\Big] \\ &= I(X_JS_J; Y_J|J) - I(S_J; U_JY_JT_J|J) \\ &= H(Y_J|J) - H(Y_J|X_JS_JJ) - H(S_J|J) \\ &\quad + H(S_J|U_JY_JT_JJ) \\ &\overset{(e)}{=} H(Y_J|J) - H(Y_J|X_JS_J) - H(S_J) \\ &\quad + H(S_J|U_JY_JT_JJ) \\ &\le H(Y_J) - H(Y_J|X_JS_J) - H(S_J) \\ &\quad + H(S_J|U_JY_JT_JJ) \\ &= I(X_JS_J; Y_J) - I(S_J; U_JY_JT_J, J) \\ &= I(XS; Y) - I(S; UTY). \end{aligned} \tag{134}$$

Here (e) holds because S_J is independent of J (since the channel states are drawn IID), and the factorization

$$\begin{aligned}
&P_{JS_JT_JX_JY_JU_J}(j,s,t,x,y,u)\\
&\quad= P_J(j)\,P_{S_JT_JX_JY_J|J}(s,t,x,y|j)\\
&\qquad\cdot P_{U_J|S_JT_JX_JY_JJ}(u|s,t,x,y,j)\\
&\quad= P_J(j)\,P_S(s)\,P_{T_j|S}(t|s)\,P_{X_j|ST_j}(x|s,t)\,P_{\mathrm{c}}(y|x,s)\\
&\qquad\cdot P_{U_j|ST_jX_jY_jj}(u|s,t,x,y,j)\\
&\quad= P_S(s)\,P_{T_J|S}(t|s)\,P_{X_J|ST_J}(x|s,t)\,P_{\mathrm{c}}(y|x,s)\\
&\qquad\cdot P_{U_JJ|ST_JX_JY_J}(u,j|s,t,x,y) \qquad (135)
\end{aligned}$$

shows that $J \multimap\!\!- (X_J, S_J) \multimap\!\!- Y_J$ forms a Markov chain and hence that $H(Y_J|S_J, X_J, J) = H(Y_J|S_J, X_J)$.

The factorization (135) also shows that the PMF of (S, T, X, Y, U) has the required form (40b).

The RHS of (134) is the second term in the minimum in (41). We next turn to the first term in that minimum, a term which is "Gel'fand-Pinsker like." The converse proof of the capacity formula for the ordinary Gel'fand-Pinsker channel, which also holds with feedback, establishes that [7]

$$n(R' - \eta_n) \le \sum_{k=1}^{n}\Big[I(T_k; Y_k) - I(T_k; S_k)\Big], \qquad (136)$$

where T_k is defined in (129). The single-letter form of (136) in terms of S, T, Y is

$$R' - \eta_n \le I(T; Y) - I(T; S). \qquad (137)$$

The combination of (134) and (137) yields

$$\begin{aligned}
R' - \eta_n \le \min\big\{&I(T; Y) - I(T; S),\\
&I(XS; Y) - I(S; UTY)\big\}. \qquad (138)
\end{aligned}$$

As to the expected distortion, we proceed from (122b) as in (104):

$$\begin{aligned}
D + \varepsilon &\ge \mathbb{E}\big[d\big(S^n, \hat{S}^n\big)\big]\\
&= \mathbb{E}\big[d\big(S, g(U, T, Y)\big)\big], \qquad (139)
\end{aligned}$$

where the last equality follows by (133). It now follows from (138), (139), and the fact that the joint law of S, X, T, Y, U factorizes as in (40b) that

$$R' - \eta_n \le \tilde{R}^{(\mathrm{u})}(D + \varepsilon), \qquad (140)$$

which, in view of (125), establishes (123) and hence concludes the proof of the upper bound (121).

To simplify the comparison between the upper and lower bounds, we note that since $T \multimap\!\!- (X, S) \multimap\!\!- Y$ forms a Markov chain

$$\begin{aligned}
&I(XS; Y) - I(S; TY) - I(S; U|TY)\\
&\quad= I(XST; Y) - I(S; TY) - I(S; U|TY)\\
&\quad= I(ST; Y) + I(X; Y|ST) - I(S; TY) - I(S; U|TY)\\
&\quad= I(T; Y) - I(T; S) + I(X; Y|ST) - I(S; U|TY)\\
&\quad= I(T; Y) - I(T; S) + H(X|ST) - H(X|STY)\\
&\qquad- I(S; U|TY)\\
&\quad= I(T; Y) - I(T; S) + H(X|ST) - H(X|STUY)\\
&\qquad+ H(X|STUY) - H(X|STY) - I(S; U|TY)\\
&\quad= I(T; Y) - I(T; S) + I(X; UY|ST) - I(X; U|STY)\\
&\qquad- I(S; U|TY)\\
&\quad= I(T; Y) - I(T; S) + I(X; UY|ST)\\
&\qquad- I(SX; U|TY). \qquad (141)
\end{aligned}$$

The upper bound (138) can thus also be expressed as follows

$$\begin{aligned}
R' - \eta_n \le \min\big\{&I(T; Y) - I(T; S), I(T; Y) - I(T; S)\\
&+ I(X; UY|ST) - I(SX; U|TY)\big\}. \qquad (142)
\end{aligned}$$

Consequently, $R^{(\mathrm{l})}$ and $R^{(\mathrm{u})}$ coincide whenever $R^{(\mathrm{u})}$ is attained by a law under which $X \multimap\!\!- (S, T) \multimap\!\!- UY$ forms a Markov chain. ■

Proof of the lower bound in Theorem 7: The proof of the lower bound as expressed in (35) is very similar to the proof in Section IV-A of the direct part of Theorem 1. The main difference is that the codebook we use to send the fresh information and the state-description information is based on the Gel'fand-Pinsker codebook (for sending fresh information only). As shown in the Appendix, this codebook can yield not only the message but also the transmitted codeword in the form of T. This and the output sequence then serve as side-information for the description of the state sequence.[5] ■

B. Proof of Remark 2

For our example of a scenario where the upper bound is not tight, we shall need the following proposition.

Proposition 13: Let the state S of a SD-DMC have the form $S = (S_{\mathrm{a}}, S_{\mathrm{b}})$, where S_{a} and S_{b} are independent; the decoder wishes to recover S_{b} losslessly; and only S_{a} influences the channel's behavior. Suppose the state S is revealed to the encoder noncausally, and denote the resulting Gel'fand-Pinsker capacity by $C_{\mathrm{a}}^{\mathrm{G\text{-}P}}$. Then the RnS capacity, with or without feedback, of the channel is given by

$$C^{\mathrm{nc}}(0) = C_{\mathrm{FB}}^{\mathrm{nc}}(0) = C_{\mathrm{a}}^{\mathrm{G\text{-}P}} - H(S_{\mathrm{b}}), \qquad (143)$$

whenever the RHS is positive. Moreover, for this scenario the lower bound $R^{(\mathrm{l})}$ in (34) is tight.

Proof: We begin with the converse, which we derive with feedback. By replacing the message in the standard converse to the Gel'fand-Pinsker problem with the pair $\big(W, S_{\mathrm{b}}^n\big)$, we obtain the upper bound

$$I\big(W, S_{\mathrm{b}}^n; Y^n\big) \le n\, C_{\mathrm{a}}^{\mathrm{G\text{-}P}}. \qquad (144)$$

Using the chain-rule we then obtain

$$\begin{aligned}
n\, C_{\mathrm{a}}^{\mathrm{G\text{-}P}} &\ge I\big(W, S_{\mathrm{b}}^n; Y^n\big) &(145)\\
&= I\big(W; Y^n\big) + I\big(S_{\mathrm{b}}^n; Y^n\big|W\big) &(146)\\
&= I\big(W; Y^n\big) + I\big(S_{\mathrm{b}}^n; W, Y^n\big) &(147)
\end{aligned}$$

[5] There is a minor technicality in analyzing the resulting lossy source-coding problem and in claiming that the required description rate can be arbitrarily close to $R_{S|TY}(D)$: by the nature of Gel'fand-Pinsker coding, even under random coding, the triple comprising the state sequence, the codeword, and the output sequence is not drawn IID [6, Remark 7.8]. This issue can be resolved by noting that, by the *Conditional Typicality Lemma* [6, Sec. 2.5], this triple is with high probability in $\mathcal{T}_{\varepsilon}^{(n)}$ and by then employing the *Type Covering Lemma* [5, Lemma 9.1].

$$\geq I\big(W; Y^n\big) + I\big(S_{\mathrm{b}}^n; Y^n\big) \tag{148}$$

$$\geq I\big(\big(W; Y^n\big) + I\big(S_{\mathrm{b}}^n; \hat{S}^n\big) \tag{149}$$

$$\geq n\,R - P_e^{(n)}\,n\,R + n\,R_{\mathrm{H}}(D), \tag{150}$$

where $P_e^{(n)}$ is the average probability of error (which tends to zero), and $R_{\mathrm{H}}(D)$ is the Hamming rate-distortion function of the source $\{S_{\mathrm{b}}\}$ (which tends to $H\big(S_{\mathrm{b}}\big)$ in the lossless limit). Dividing both sides by n and letting n tend to infinity, proves that the RnS capacity with feedback is upper-bounded by the RHS of (143).

As to the direct part, we note that to achieve the RHS of (143) without feedback, we can losslessly describe S_{b}^n and then send the binary description together with $n\,R$ message bits reliably over the Gel'fand-Pinsker channel.

The lower bound $R^{(\mathrm{l})}$ of (34) is tight because we can choose U as S_{b} and choose T to be the auxiliary chance variable that achieves the Gel'fand-Pinsker capacity and that is independent of S_{b}. ■

We can now construct an example where the upper bound is not tight as follows. Consider a SD-DMC with state $S = (S_{\mathrm{a}}, S_{\mathrm{b}})$, where S_{a} and S_{b} are independent; the decoder wishes to recover S_{b} losslessly; and only S_{a} influences the channel's behavior. The channel comprises two parallel subchannels. The first—whose input is x' and whose output y' is equal to x'—is a noiseless "bit pipe" of capacity C_{pipe}. The second, $W_{\mathrm{a}}(y''|x'', s_{\mathrm{a}})$, is a state-dependent channel whose Gel'fand-Pinsker capacity is denoted $C_{\mathrm{a}}^{\mathrm{G\text{-}P}}$ and whose capacity when S_{a} is revealed to both encoder and decoder is $C_{\mathrm{a}}^{|S_{\mathrm{a}}}$, with the latter being strictly larger than $C_{\mathrm{a}}^{\mathrm{G\text{-}P}}$

$$C_{\mathrm{a}}^{\mathrm{G\text{-}P}} < C_{\mathrm{a}}^{|S_{\mathrm{a}}}. \tag{151}$$

The Gel'fand-Pinsker capacity of the aggregate channel $C_{\mathrm{G\text{-}P}}$ is

$$C_{\mathrm{G\text{-}P}} = C_{\mathrm{pipe}} + C_{\mathrm{a}}^{\mathrm{G\text{-}P}}.$$

And since we must reconstruct S_{b} losslessly, the optimal choice of U in both the lower bound and the upper bound is S_{b}. The lower bound is thus

$$R^{(\mathrm{l})} = C_{\mathrm{pipe}} + C_{\mathrm{a}}^{\mathrm{G\text{-}P}} - H(S_{\mathrm{b}}).$$

For the upper bound, let $P_{X''|S_{\mathrm{a}}}$ achieve $C_{\mathrm{a}}^{|S_{\mathrm{a}}}$, and let $P_{X'}$ achieve C_{pipe} (i.e., be uniform on the input alphabet of the bit-pipe subchannel). Consider the joint distribution according to which $T = X' \sim P_{X'}$; the pair (T, X') is independent of (X'', S); and X'' is drawn conditionally on S according to $P_{X''|S_{\mathrm{a}}}$, i.e., the conditional input distribution achieving $C_{\mathrm{a}}^{|S_{\mathrm{a}}}$.

For this distribution

$$I(T; Y', Y'') - I(T; S) = C_{\mathrm{pipe}},$$

and (upon substituting S_{b} for U)

$$I(X, S; Y) - I(S; UTY) \geq C_{\mathrm{pipe}} + C_{\mathrm{a}}^{|S_{\mathrm{a}}} - H(S_{\mathrm{b}}).$$

The upper bound is thus (at least)

$$\min\big\{C_{\mathrm{pipe}}, C_{\mathrm{pipe}} + C_{\mathrm{a}}^{|S_{\mathrm{a}}} - H(S_{\mathrm{b}})\big\}.$$

If

$$C_{\mathrm{a}}^{\mathrm{G\text{-}P}} < H(S_{\mathrm{b}}) \leq C_{\mathrm{a}}^{|S_{\mathrm{a}}},$$

then the upper bound is (at least) C_{pipe}; the lower bound—which is tight by Proposition 13—is smaller than C_{pipe}; and the upper bound is thus loose.

C. Proof of Proposition 8

The converse is based on the upper bound in (42). We first note that if $P_{XT|S}$, $P_{U|STXY}$, and g are valid choices in the maximization (41) defining $R^{(\mathrm{u})}$, then—because D is zero and $d(\cdot,\cdot)$ is the Hamming distortion—$H(S|UTY)$ must be zero. In fact, in this maximization we can choose U to equal S and $g(u, t, y)$ to equal u. For zero Hamming distortion we can therefore rewrite (41) as

$$R^{(\mathrm{u})} = \max_{P_{TX|S}} \min\big\{I(T; Y) - I(T; S), I(XS; Y) - H(S)\big\}. \tag{152}$$

Consequently, if R is achievable with lossless reconstruction of the state sequence, then—for some joint law of the form $P_S P_{TX|S} P_{Y|X,S}$—R is upper bounded by the RHS of (152) or, equivalently, R must satisfy the following two inequalities:

$$R \leq I(T; Y) - I(T; S) \tag{153}$$

$$R + H(S) \leq I(XS; Y). \tag{154}$$

By [10, Lemma 2], these two inequalities can be replaced by the single inequality

$$R \leq \max_{P_{TX|S}} I(ST; Y) - H(S). \tag{155}$$

The proof of the converse is now completed by noting that

$$\begin{aligned} I(TS; Y) - H(S) &= I(T; Y) + I(S; Y|T) - H(S) \\ &= I(T; Y) - I(T; S) - H(S|TY). \end{aligned} \tag{156}$$

To show that $I(T; Y) - I(T; S) - H(S|TY)$ is maximized by a law in which X is a deterministic function of (S, T), note that

$$\begin{aligned} &I(T; Y) - I(T; S) - H(S|TY) \\ &\quad = [I(T; Y) + I(S; TY)] - I(T; S) - H(S). \end{aligned}$$

For a fixed P_{TS}, the functional inside the squared brackets is convex in $P_{X|ST}$ (and hence maximized when X is a deterministic function of (S, T)) while $I(T; S) + H(S)$ is fixed.

The direct part does not utilize the feedback link. We use the Gel'fand-Pinsker code construction to transmit two streams: a data stream and a state-description stream. The state description is at rate $H(S|TY) + \varepsilon$, and is based on Slepian-Wolf source coding [18] where the state sequence is described to a decoder that is furnished with Y and T. The decoder decodes the Gel'fand-Pinsker codeword to decode the data stream and the state description. Using the latter, the channel outputs, and T it then reconstructs the state sequence.[6]

[6]Here too we encounter a technicality similar to the one we addressed in Footnote 5: the triple comprising the state sequence, the codeword, and the output sequence is not drawn IID. But, as in that footnote, it is with high probability typical, and this suffices for the binning argument to go through.

D. On $R^{(\mathrm{u})}$ for the Gaussian Channel

We next determine the smallest mean squared-error with which we can estimate the state of a Gaussian channel while still communicating at a positive rate, i.e., the infimum of the D's for which $C_{\mathrm{FB}}^{\mathrm{nc}}(D)$ is positive. Later we shall study the distortions for which $C_{\mathrm{FB}}^{\mathrm{nc}}(D)$ exceeds some positive rate R. Our focus will be on our upper bound $R^{(\mathrm{u})}$ (41), which implies a lower bound on the reconstruction distortion. In treating the zero-rate case we shall, in fact, consider a slightly looser bound that results from dropping the term $I(T;Y) - I(T;S)$ from the RHS of (41) to obtain

$$C_{\mathrm{FB}}^{\mathrm{nc}}(D) \le \max_{P_{TX|S},\, P_{U|STXY}} I(XS;Y) - I(S;UTY). \quad (157)$$

For the LHS to be positive, the distortion D must be sufficiently large for the RHS to be positive. To derive a lower bound on D we shall thus upper-bound $I(XS;Y)$ and lower-bound $I(S;UTY)$. We begin with the latter:

$$\begin{aligned} I(S;UTY) &= h(S) - h(S|UTY) \\ &\overset{(a)}{=} h(S) - h(S|\hat{S}UTY) \\ &\overset{(b)}{\ge} h(S) - h(S|\hat{S}) \\ &= h(S) - h(S-\hat{S}|\hat{S}) \\ &\overset{(c)}{\ge} h(S) - h(S-\hat{S}) \\ &\overset{(d)}{\ge} h(S) - \frac{1}{2}\log 2\pi e \sigma^2_{S|UTY} \\ &= \frac{1}{2}\log \frac{\sigma_s^2}{\sigma^2_{S|UTY}} \\ &\ge \frac{1}{2}\log \frac{\sigma_s^2}{D}, \end{aligned} \quad (158)$$

where

- (a) holds because $\hat{S}$ is a function of (U,T,Y);
- (b) and (c) hold because conditioning cannot increase differential entropy; and
- (d) holds because the Gaussian distribution maximizes the differential entropy for a given variance.

As to the term $I(XS;Y)$,

$$\begin{aligned} I(XS;Y) &= h(Y) - h(Y|XS) \\ &= h(Y) - h(Z) \\ &\overset{(e)}{\le} \frac{1}{2}\log \frac{\mathbb{E}[Y^2]}{N} \\ &\overset{(f)}{\le} \frac{1}{2}\log \frac{\left(\sigma_s + \sqrt{P}\right)^2 + N}{N}, \end{aligned} \quad (159)$$

where

- (e) holds because the Gaussian distribution maximizes the entropy for a given variance; and
- (f) holds because $\mathbb{E}[Y^2] \le \left(\sigma_s + \sqrt{P}\right)^2 + N$ whenever $\mathbb{E}[X^2] \le P$, and Z is independent of (X,S).

It follows from (158) and (159) that, for the RHS of (157) to be positive, the distortion D must satisfy

$$D > \sigma_s^2 \frac{N}{\left(\sigma_s + \sqrt{P}\right)^2 + N} \quad (160)$$

as reported in [19, Sec. II] for the no-feedback case.

We next turn to the case where the required communication rate is some positive rate $R > 0$. For this to be possible, D must surely satisfy (160), which is what we now assume.

To derive necessary conditions on (R,D) to be achievable, we shall use the upper bound $R^{(\mathrm{u})}$ of (41) by replacing the maximization on the LHS of (41) by two maximizations: the first over $\rho \in [-1,1]$, and the second—as in (41)—but with the additional constraint that under $P_S P_{TX|S}$ the second moment of X be P and the correlation $\rho(X,S)$ between X and S be ρ.

In fact, in the first maximization we may exclude ρ from being ± 1, because, as we next argue, if ρ is ± 1, then $I(T;Y) - I(T;S)$ must be zero. Indeed, if $\rho(X,S)$ is ± 1, then X equals βS (almost surely) for some $\beta \in \mathbb{R}$, and—since $Y = X + S + Z$—we then have

$$\begin{aligned} I(T;Y) - I(T;S) &= I(T;(\beta+1)S + Z) - I(T;S) \\ &\le I\big(T;(\beta+1)S + Z, Z\big) - I(T;S) \\ &= I\big(T;(\beta+1)S, Z\big) - I(T;S) \\ &= I\big(T;(\beta+1)S\big) - I(T;S) \\ &= 0, \end{aligned} \quad (161)$$

where the fourth line follows from the independence between Z and (T,S).

Having established that ± 1 can be excluded,

$$\begin{aligned} R^{(\mathrm{u})} = \max_{\rho\in(-1,1)} \max_{P_{TX|S},\, \rho(X,S)=\rho,\, P_{U|STXY}} \min\big\{ & I(T;Y) \\ & - I(T;S),\ I(XS;Y) - I(S;UTY)\big\}. \end{aligned} \quad (162)$$

For a fixed $\rho \in (-1,1)$, we now upper-bound $I(T;Y) - I(T;S)$ as follows:

$$\begin{aligned} &\max_{P_{TX|S},\, \rho(X,S)=\rho} I(T;Y) - I(T;S) \\ &\quad\le \max_{P_{TX|S},\, \rho(X,S)=\rho} I(T;YS) - I(T;S) \\ &\quad= \max_{P_{TX|S},\, \rho(X,S)=\rho} h(Y|S) - h(Y|T,S) \\ &\quad\le \max_{P_{TX|S},\, \rho(X,S)=\rho} h(Y|S) - h(Y|T,S,X) \\ &\quad= \max_{P_{X|S},\, \rho(X,S)=\rho} h(Y|S) - h(Z). \end{aligned} \quad (163)$$

Fixing $\rho(X,S)$ and the second moment of X fixes $\mathbb{E}[XS]$, and consequently [19, Appendix, Lemma 1],

$$\begin{aligned} &\max_{P_{X|S},\, \rho(X,S)=\rho} h(Y|S) \\ &\le \frac{1}{2}\log\left(2\pi e\left(\mathbb{E}[Y^2] - \frac{(\mathbb{E}[SY])^2}{\mathbb{E}[S^2]}\right)\right) \\ &= \frac{1}{2}\log\left(2\pi e\left(P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N - \left(\rho\sqrt{P} + \sigma_s\right)^2\right)\right) \\ &= \frac{1}{2}\log\left(2\pi e\left((1-\rho^2)P + N\right)\right). \end{aligned} \quad (164)$$

It now follows from (163) and (164) that for any fixed $\rho \in (-1,1)$

$$\max_{P_{TX|S},\, \rho(X,S)=\rho} I(T;Y) - I(T;S) \le \frac{1}{2}\log\frac{(1-\rho^2)P + N}{N}. \quad (165)$$

We next study the second term in the minimum in (162), namely $I(XS;Y) - I(S;UTY)$. Since,

$$\mathbb{E}\big[Y^2\big] = P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N, \tag{166}$$

it follows that

$$h(Y) \le \frac{1}{2}\log\Big(2\pi e\big(P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N\big)\Big), \tag{167}$$

and hence

$$I(XS;Y) \le \frac{1}{2}\log\big(P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N\big) - \frac{1}{2}\log N, \tag{168}$$

which combines with (158) to yield

$$\max_{P_{TX|S},\, \rho(X,S)=\rho,\, P_{U|STXY}} I(XS;Y) - I(S;UTY) \le \frac{1}{2}\log\frac{P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N}{N} - \frac{1}{2}\log\frac{\sigma_s^2}{D}. \tag{169}$$

From (169), (165), and our upper bound (41) we conclude that if a pair (R, D) is achievable, then for some $\rho \in (-1, 1)$

$$R \le \frac{1}{2}\log\frac{(1-\rho^2)P + N}{N} \tag{170a}$$

$$R \le \frac{1}{2}\log\frac{P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N}{N} - \frac{1}{2}\log\frac{\sigma_s^2}{D}, \tag{170b}$$

i.e.,

$$R \le \frac{1}{2}\log\frac{(1-\rho^2)P + N}{N} \tag{171a}$$

$$D \ge \frac{N\sigma_s^2}{P + \sigma_s^2 + 2\rho\sigma_s\sqrt{P} + N}2^{2R}. \tag{171b}$$

For a fixed R, the RHS of (171b) is monotonically decreasing in ρ. This RHS is thus minimal when ρ is chosen to be as large as possible. Choosing ρ to be too large would violate (171a). The RHS of (171b) is thus minimal when ρ is chosen to be the largest that still satisfies (171a), i.e., when ρ is chosen to be the nonnegative solution to the equation that results when the inequality sign in (171a) is replaced by an equality.

Given $\rho_\star \in [0, 1)$, define

$$R_{\rho_\star} = \frac{1}{2}\log\frac{(1-\rho_\star^2)P + N}{N}. \tag{172a}$$

If the required rate R is equal to $R_{\rho_\star}$, then the largest ρ for which (171a) holds is $\rho_\star$ and consequently the distortion achievable cannot be below $D_{\rho_\star}$, which results when we substitute $\rho_\star$ for ρ and $R_{\rho_\star}$ for R in (171b)

$$D_{\rho_\star} = \frac{N\sigma_s^2}{P + \sigma_s^2 + 2\rho_\star\sigma_s\sqrt{P} + N}2^{2R_{\rho_\star}}. \tag{172b}$$

For any $\rho_\star \in [0, 1)$, the least distortion that is achievable with communication rate $R_\star$ is thus lower bounded by $D_{\rho_\star}$. Conversely,

$$C_{\mathrm{FB}}^{\mathrm{nc}}(D_{\rho_\star}) \le R_{\rho_\star}, \quad 0 \le \rho_\star < 1. \tag{173}$$

An alternative form for (172) is obtained by defining $\gamma = 1 - \rho_\star^2 \in (0, 1]$. This leads to

$$R_\gamma = \frac{1}{2}\log\frac{\gamma P + N}{N}$$

$$D_\gamma = \frac{\sigma_s^2(\gamma P + N)}{\gamma P + \big(\sigma_s + \sqrt{(1-\gamma)P}\big)^2 + N}$$

as reported in [19, Th. 2, Sec. III] in the absence of feedback.

VII. Defective Memories

Consider binary memory cells whose state $S \in \{\mathrm{d}, \mathrm{w}\}$ indicates whether they are defective or working. When a cell is defective, it is "stuck-at-one" irrespective of what is written to it. Denoting the content of the cell by $x \in \{0, 1\}$; its output by $Y \in \{0, 1\}$; and its transition law when defective by $W^{(\mathrm{d})}(y|x)$,

$$W^{(\mathrm{d})}(y|x) = \mathbb{1}\{y = 1\}, \quad x, y \in \{0, 1\}. \tag{174}$$

We model a working cell as a Z-channel with a zero always being read as a zero, and with a one being read as a one with probability $1 - \varepsilon$, for some $0 \le \varepsilon < 1$:

$$W^{(\mathrm{w})}(0|0) = 1, \qquad W^{(\mathrm{w})}(1|1) = 1 - \varepsilon. \tag{175}$$

The different cells behave independently, with the probability that a cell is defective being p,

$$\Pr[S = \mathrm{d}] = p. \tag{176}$$

The writer, in addition to storing data, also wishes to describe the state of the cells to within some average Hamming distortion D. By not sending any data and simply writing zero to all the cells, the writer can convey the state of the cells perfectly: the reader can then declare that a cell is defective whenever the cell's output is one, and it can declare that the cell is working otherwise. We thus conclude that, by setting the transmission rate to zero, we can achieve zero state-reconstruction distortion, and hence

$$C_{\mathrm{FB}}^{\mathrm{nc}}(0) \ge C_{\mathrm{FB}}^{\mathrm{c}}(0) \ge C_{\mathrm{FB}}^{\mathrm{s\text{-}c}}(0) \ge 0. \tag{177}$$

(To achieve zero distortion it need not be necessary to communicate at zero rate. For example, if p is zero—and hence none of the memories defective—we can communicate at channel capacity while still maintaining zero state-reconstruction distortion.)

Denoting the capacity of the Z-channel corresponding to a working cell by $C^{(\mathrm{w})}$,

$$\begin{aligned} C^{(\mathrm{w})} &= \max_{0<q<1} H_{\mathrm{b}}(q(1-\varepsilon)) - qH_{\mathrm{b}}(\varepsilon) \\ &= -\ln\left[1 - q(1-\varepsilon)\right]\big|_{q=q^\star} \\ &= \ln\left(1 + (1-\varepsilon)\varepsilon^{\frac{\varepsilon}{1-\varepsilon}}\right) \text{ nats/cell}, \end{aligned} \tag{178}$$

where q stands for the probability of the input one;

$$q^\star = \left(1 - \varepsilon + \varepsilon^{-\frac{\varepsilon}{1-\varepsilon}}\right)^{-1};$$

and $H_{\mathrm{b}}(\xi) = -\xi\ln\xi - (1-\xi)\ln(1-\xi)$ is the binary entropy function of $\xi \in [0, 1]$.

In the absence of any state information, the channel behaves like the mixture channel

$$W_{Y|X} = pW^{(\mathrm{d})} + (1-p)W^{(\mathrm{w})} \tag{179}$$

whose law is

$$\begin{aligned} W_{Y|X}(0|0) &= 1-p \\ W_{Y|X}(1|1) &= p + (1-p)(1-\varepsilon). \end{aligned} \tag{180}$$

Denoting this channel's capacity $C_{\text{No-SI}}(p,\varepsilon)$,

$$\begin{aligned} C_{\text{No-SI}}(p,\varepsilon) &= \max_{0<q<1} H_{\mathrm{b}}\left[(1-q)(1-p)+q(1-p)\varepsilon\right] \\ &\qquad -(1-q)H_{\mathrm{b}}(p) - qH_{\mathrm{b}}((1-p)\varepsilon) \\ &= \frac{\varepsilon H_{\mathrm{b}}(p) - H_{\mathrm{b}}((1-p)\varepsilon)}{1-\varepsilon} + \ln\left(1+e^{t}\right) \text{ nats/cell}, \end{aligned} \tag{181}$$

where the capacity-achieving law assigns the input one the probability

$$q^{*}_{\text{No-SI}} = \frac{e^{-t}(1-p)-p}{(1-p)(1-\varepsilon)(1+e^{-t})}$$

and where we define

$$t \triangleq \frac{H_{\mathrm{b}}((1-p)\varepsilon) - H_{\mathrm{b}}(p)}{(1-p)(1-\varepsilon)}.$$

The capacity of this channel when the state is known to both encoder and decoder is the same as when it is only known to the decoder. We may therefore denote both capacities $C_{\text{PSI}}(p,\varepsilon)$, where

$$\begin{aligned} C_{\text{PSI}}(p,\varepsilon) &= \max_{p_{X|S}} I(X;Y|S) = \max_{p_X} I(X;Y|S) \\ &= (1-p)\ln\left(1+(1-\varepsilon)\varepsilon^{\frac{\varepsilon}{1-\varepsilon}}\right) \text{ nats/cell}. \end{aligned} \tag{182}$$

A. Strictly-Causal and Causal SI with Feedback

We next use (16) to compute the feedback RnS capacity of a memory cell with strictly-causal state-information under the Hamming distortion measure. We denote by $D \in [0,1]$ the maximal-allowed state-reconstruction distortion. We begin by computing the conditional rate-distortion function $R_{S|XY}(\cdot)$ when X is Bernoulli with probability of success q

$$X \sim \mathrm{Ber}(q). \tag{183}$$

The event to consider is $(X=1, Y=1)$, because all other outcomes fully determine the state S. We denote the probability of this event p_{11},

$$p_{11} = q\,p + q\,(1-p)\,(1-\varepsilon). \tag{184}$$

Since all other outcomes fully determine the state, $R_{S|XY}(D)$ is the product of p_{11} by the rate-distortion function evaluated at D/p_{11} of a source whose law is the conditional law of S given the event [1, eq. (6.1.21)]. Conditional on this event, the state is $\mathrm{Ber}(p_{\mathrm{d}|11})$, i.e., is defective with probability

$$\begin{aligned} p_{\mathrm{d}|11} &\triangleq \Pr[S=\mathrm{d}\,|\,X=1,Y=1] \\ &= \frac{p}{p+(1-p)(1-\varepsilon)}. \end{aligned} \tag{185}$$

Denoting the Rate-Distortion function of a $\mathrm{Ber}(\pi)$ source with maximal-allowed Hamming distortion δ_{H} by $R^{\mathrm{Ham}}_{\mathrm{Ber}}\left(\pi;\,\delta_{\mathrm{H}}\right)$ [4, eq. (10.23)],

$$\begin{aligned} &R^{\mathrm{Ham}}_{\mathrm{Ber}}\left(\pi;\,\delta_{\mathrm{H}}\right) \\ &\quad = \begin{cases} H_{\mathrm{b}}(\pi) - H_{\mathrm{b}}\left(\delta_{\mathrm{H}}\right) & \text{if } 0 \le \delta_{\mathrm{H}} \le \min\{\pi, 1-\pi\} \\ 0 & \text{if } \delta_{\mathrm{H}} > \min\{\pi, 1-\pi\}, \end{cases} \end{aligned} \tag{186}$$

we obtain

$$R_{S|XY}(D;q) = p_{11}\,R^{\mathrm{Ham}}_{\mathrm{Ber}}\left(p_{\mathrm{d}|11};\; D/p_{11}\right), \tag{187}$$

where we have added the parameter q to remind us that this conditional rate-distortion function depends on the probability q with which X equals one. For a fixed D, this function is monotonically non-decreasing in q.

As to the mutual information $I_q(X;Y)$ (again with the dependence on q made explicit), a direct computation yields,

$$\begin{aligned} I_q(X;Y) &= H_{\mathrm{b}}\left((1-q)(1-p)+q(1-p)\varepsilon\right) - (1-q)H_{\mathrm{b}}(p) \\ &\qquad -qH_{\mathrm{b}}\left((1-p)\varepsilon\right). \end{aligned} \tag{188}$$

It thus follows from (16) that

$$C^{\text{s-c}}_{\text{FB}}(D) = \max_{0\le q\le 1}\left\{I_q(X;Y) - R_{S|XY}(D;q)\right\}. \tag{189}$$

In fact, we can limit the optimization to values of q that are no larger than the value $q^{*}_{\text{No-SI}}$ of q that maximizes $I_q(X;Y)$, because $R_{S|XY}(D;q)$ is non-decreasing in q.

We can communicate at channel capacity provided that the allowed distortion D is such that $R_{S|XY}(D;q^{*}_{\text{No-SI}})$ is zero, i.e., as long as D is at least

$$\left(q^{*}_{\text{No-SI}}\,p + q^{*}_{\text{No-SI}}\,(1-p)\,(1-\varepsilon)\right)\min\{p_{\mathrm{d}|11}, 1-p_{\mathrm{d}|11}\}. \tag{190}$$

The case of *causal* state information requires hardly any extra work. In fact, in this example, causal SI affords no rate gains over the strictly-causal one, so $C^{\text{s-c}}_{\text{FB}}(D)$ and $C^{\text{c}}_{\text{FB}}(D)$ are the same. To show this, we shall consult (30) and argue that how a strategy maps a defective state influences neither $I(T;Y)$ nor $R_{S|TY}(D)$. More formally, since there are only two states, a strategy T is in a one-to-one relationship with the pair of random variables $T(\mathrm{d}), T(\mathrm{w}) \in \{0,1\}$. We will argue that only the distribution of $T(\mathrm{w})$ influences the above two terms, and that there is therefore no loss of optimality in setting $T(\mathrm{d})$ to equal $T(\mathrm{w})$ and in this way limiting ourselves to constant strategies, i.e., to schemes where the input to the channel does not depend on the state. Those can be implemented with strictly-causal SI by setting X to equal $T(\mathrm{w})$.

That $I\left(T(\mathrm{d}), T(\mathrm{w});\, Y\right) = I\left(T(\mathrm{w});\, Y\right)$ (and that hence only the distribution of $T(\mathrm{w})$ influences $I(T;Y)$) follows from

$$T(\mathrm{d}) \multimap T(\mathrm{w}) \multimap Y. \tag{191}$$

We next turn to $R_{S|TY}(D)$. When Y is zero, the state is guaranteed to be working. And when Y is one and $T(\mathrm{w})$ is zero, the state is guaranteed to be defective. The case to watch for is thus when Y is one and $T(\mathrm{w})$ is one. This corresponds to two events: $\left(T(\mathrm{w})=1,\; T(\mathrm{d})=0,\; Y=1\right)$ and

$\big(T(\mathrm{w}) = 1,\ T(\mathrm{d}) = 1,\ Y = 1\big)$. However, a straightforward calculation shows that the conditional distribution of S given these different events is the same. Consequently, for the purpose of calculating $R_{S|TY}(D)$, there is no need to distinguish between the two cases, and setting $T(\mathrm{d})$ to equal $T(\mathrm{w})$ does not influence $R_{S|TY}(D)$.

B. Noncausal SI

We next consider the noncausal RnS feedback capacity of our memory cells. Again we focus on the Hamming metric, and denote the maximally-allowed expected distortion $0 \le D \le 1$.

We begin by lower bounding $R^{(\mathrm{I})}$ of (34) by considering a specific joint distribution under which

$$\begin{aligned} P_{T|S}(0|\mathrm{d}) &= 1 \\ P_{T|S}(0|\mathrm{w}) &= \beta, \end{aligned} \tag{192}$$

with X being the deterministic function of (S, T) bellow:

$$x(\mathrm{d}, t) = 1, \quad t \in \{0, 1\} \tag{193a}$$

$$x(\mathrm{w}, 0) = 1 \tag{193b}$$

$$x(\mathrm{w}, 1) = 0. \tag{193c}$$

The salient features of this choice are:

$$T = 1 - X \tag{194a}$$

(any one-to-one relationship will do);

$$\Pr[X = 1 \,|\, Y = 1, S = \mathrm{d}] = 1 \tag{194b}$$

(compatible input [8]: when defective, store one);

$$\Pr[X = 1 \,|\, Y = 1, S = \mathrm{w}] = 1 \tag{194c}$$

(guaranteed because when the cell is working it behaves like a Z-channel that maps zero to zero); and

$$\Pr[X = 1 \,|\, S = \mathrm{w}] = \beta. \tag{194d}$$

As we shall see, this choice will result in $I(T; Y) - I(T; S)$ being equal to $I(X; Y|S)$. (Our approach is based on computation, but this result can be derived alternatively by writing $I(T; Y) - I(T; S)$ as $H(T|S) - H(T|Y)$ and using the above properties of the joint distribution.[7])

For this joint distribution

$$\begin{aligned} H(T) &= H_{\mathrm{b}}\big(p + (1-p)\beta\big) \\ H(T|S) &= (1-p)H_{\mathrm{b}}(\beta) \\ H(Y) &= H_{\mathrm{b}}\big((1-p)(1-\beta) + (1-p)\beta\varepsilon\big) \\ H(Y|T) &= \big(p + (1-p)\beta\big)H_{\mathrm{b}}\Big(\frac{(1-p)\beta\varepsilon}{p + (1-p)\beta}\Big). \end{aligned}$$

[7]From (194) it follows that $H(T|Y = 1)$ is zero and likewise $H(T|Y = 1, S = \mathrm{d})$. Consequently, $H(T|Y) = \Pr[Y = 0]\, H(T|Y = 0) = \Pr[S = \mathrm{w}]\, \Pr[Y = 0|S = \mathrm{w}]\, H(T|Y = 0, S = \mathrm{w}) = \Pr[S = \mathrm{w}]\, H(T|Y, S = \mathrm{w})$, where the second equality holds because the cell must be working if its output is not one. As to $H(T|S)$, we note that $H(T|S = \mathrm{d})$ must be zero (because (194b) implies that $\Pr[X = 1 \,|\, S = \mathrm{d}]$ is one, which combines with (194a)) to prove that $\Pr[H = 1 \,|\, S = \mathrm{d}]$ is zero), so $H(T|S) = \Pr[S = \mathrm{w}]\, H(T|S = \mathrm{w})$, or, in view of (194a), $H(T|S) = \Pr[S = \mathrm{w}]\, H(X|S = \mathrm{w})$. Thus, $H(T|S) - H(T|Y) = \Pr[S = \mathrm{w}]\, I(X; Y|S = \mathrm{w}) = I(X; Y|S)$.

Defining

$$R_{\mathrm{GP}}(p, \varepsilon, \beta) = I(T; Y) - I(T; S), \tag{195}$$

we have

$$\begin{aligned} &I(T; Y) - I(T; S) \\ &\quad = (1-p)\Big(\beta\varepsilon \ln \beta\varepsilon - \beta \ln \beta \\ &\qquad\qquad - \big(1 - \beta(1-\varepsilon)\big) \ln\big(1 - \beta(1-\varepsilon)\big)\Big) \\ &\quad = R_{\mathrm{GP}}(p, \varepsilon, \beta). \end{aligned} \tag{196}$$

It can be verified that

$$\begin{aligned} &\max_{0<\beta<1} R_{\mathrm{GP}}(p, \varepsilon, \beta) \\ &\quad = C_{\mathrm{GP}}(p, \varepsilon) = C_{\mathrm{PSI}}(p, \varepsilon) \\ &\quad = (1-p) \ln\Big(1 + (1-\varepsilon)\varepsilon^{\frac{\varepsilon}{1-\varepsilon}}\Big) \text{ nats/cell}, \end{aligned} \tag{197}$$

where, regardless of p, the Gel'fand-Pinsker capacity $C_{\mathrm{GP}}(p, \varepsilon)$ is achieved by

$$\beta^* = \big(1 - \varepsilon + \varepsilon^{-\frac{\varepsilon}{1-\varepsilon}}\big)^{-1}.$$

Thus, for this memory model, noncausal encoder SI is as beneficial as that when both encoder and decoder have SI. This is in agreement with the observations in [8, Sec. I] regarding similar models of defective memories.

From (34) we obtain

$$\begin{aligned} R^{(\mathrm{I})} &\ge I(T; Y) - I(T; S) - R_{S|TY}(D) \\ &= R_{\mathrm{GP}}(p, \varepsilon, \beta) - R_{S|TY}(D). \end{aligned} \tag{198}$$

We now calculate $R_{S|TY}(D)$. Our model implies that

$$(Y = 0) \implies (S = \mathrm{w}). \tag{199}$$

And it also implies, in view of (193), that

$$(T = 1,\ Y = 1) \implies (S = \mathrm{d}) \tag{200}$$

(because if the cell were working and T were one, we would write zero and hence read a zero).

The event to focus on is thus $(T = 0,\ Y = 1)$: in all other cases the decoder can recover the state from T and Y. Denoting the probability of this event by π_{01},

$$\pi_{01} = p + (1-p)(1-\varepsilon)\beta. \tag{201}$$

Conditional on this event, the state is defective with probability $\pi_{\mathrm{d}|01}$, where

$$\begin{aligned} \pi_{\mathrm{d}|01} &= \Pr[S = \mathrm{d} \,|\, T = 0, Y = 1] \\ &= \frac{p}{p + (1-p)(1-\varepsilon)\beta}. \end{aligned} \tag{202}$$

Since all other outcomes of (T, Y) yield zero reconstruction distortion,

$$R_{S|TY}(D; \beta) = \pi_{01}\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\pi_{\mathrm{d}|01}\,;\ D/\pi_{01}\big), \tag{203}$$

where we have made the dependence on β explicit. Combining (203) with (198) yields,

$$R^{(\mathrm{I})}(d) \ge \max_{0 \le \beta \le 1} \Big\{ R_{\mathrm{GP}}(p, \varepsilon, \beta) - \pi_{01}\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\pi_{\mathrm{d}|01};\ D/\pi_{01}\big) \Big\}. \tag{204}$$

We next upper-bound $R^{(\mathrm{u})}$ of (41) by $I(XS;Y) - I(S;UTY)$ and proceed to upper-bound the latter. We fix some joint distribution in the admissible set over which the maximum in (41) is taken, and we define

$$\tilde{\beta} = \Pr[X = 1 \,|\, S = \mathrm{w}] \tag{205a}$$
$$\tilde{D} = \mathbb{E}\big[d(S, \hat{S}) | Y = 1\big] \tag{205b}$$
$$\tilde{\pi}_1 = \Pr[Y = 1] \tag{205c}$$
$$\tilde{\pi}_{\mathrm{d}|1} = \Pr[S = \mathrm{d} \,|\, Y = 1], \tag{205d}$$

so

$$\tilde{\pi}_1 = p + (1-p)(1-\varepsilon)\tilde{\beta} \tag{206a}$$
$$\tilde{\pi}_{\mathrm{d}|1} = \frac{p}{p + (1-p)(1-\varepsilon)\,\tilde{\beta}} \tag{206b}$$

and

$$\tilde{D} \le \frac{D}{\tilde{\pi}_1}. \tag{207}$$

We next express $I(XS;Y) - I(S;UTY)$ as $I(X;Y|S) - I(S;UT|Y)$ and proceed to calculate $I(X;Y|S)$ and to lower-bound $I(S;UT|Y)$. Starting with $I(S;UT|Y)$,

$$\begin{aligned} I(S;UT|Y) &= \Pr[Y=1]\, I(S;UT|Y=1) \\ &= \Pr[Y=1]\, I\big(S;UT\hat{S}|Y=1\big) \\ &\ge \tilde{\pi}_1\, I\big(S;\hat{S}|Y=1\big) \\ &\ge \tilde{\pi}_1\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\tilde{\pi}_{\mathrm{d}|1};\ \tilde{D}\big) \\ &\ge \tilde{\pi}_1\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\tilde{\pi}_{\mathrm{d}|1};\ D/\tilde{\pi}_1\big), \end{aligned} \tag{208}$$

where the first line follows because, conditional on $Y = 0$, the state S is deterministic; the second because $\hat{S}$ is a deterministic function of (U, T, Y); the third by dropping (U, T) and recalling (205c); the fourth from (205b) and because the rate-distortion function minimizes mutual information subject to a distortion constraint; and the last line follows from (207).

As to $I(X;Y|S)$, since Y is deterministically one when the cell is defective,

$$\begin{aligned} &I(X;Y|S) \\ &\quad= \Pr[S=\mathrm{w}]\, I(X;Y|S=\mathrm{w}) \\ &\quad= \Pr[S=\mathrm{w}]\,\big(H(Y|S=\mathrm{w}) - H(Y|S=\mathrm{w},X)\big) \\ &\quad= \Pr[S=\mathrm{w}]\Big(H(Y|S=\mathrm{w}) \\ &\qquad\qquad - \Pr[X=0\,|\,S=\mathrm{w}]\, H(Y|S=\mathrm{w},X=0) \\ &\qquad\qquad - \Pr[X=1\,|\,S=\mathrm{w}]\, H(Y|S=\mathrm{w},X=1)\Big) \\ &\quad= \Pr[S=\mathrm{w}]\,\Big(H(Y|S=\mathrm{w}) \\ &\qquad\qquad - \Pr[X=1\,|\,S=\mathrm{w}]\, H(Y|S=\mathrm{w},X=1)\Big), \end{aligned}$$

where the last equality holds because when the state is working and the input is zero the output is deterministically zero.

In terms of $\tilde{\beta}$,

$$H(Y|S=\mathrm{w}) = H_{\mathrm{b}}\big([1-\tilde{\beta}(1-\varepsilon)]\big) \tag{209}$$

and

$$\Pr[X=1\,|\,S=\mathrm{w}]\, H(Y|S=\mathrm{w},X=1) = \tilde{\beta} H_{\mathrm{b}}(\varepsilon), \tag{210}$$

so

$$\begin{aligned} I(X;Y|S) &= (1-p)\Big(H_{\mathrm{b}}\big([1-\tilde{\beta}(1-\varepsilon)]\big) - \tilde{\beta} H_{\mathrm{b}}(\varepsilon)\Big) \\ &= R_{\mathrm{GP}}(p,\varepsilon,\tilde{\beta}). \end{aligned} \tag{211}$$

The lower bound on $I(S;UT|Y)$ in (208) and the calculation of $I(X;Y|S)$ in (211) combine with the trivial bound $R^{(\mathrm{u})} \le I(XS;Y) - I(S;UTY)$ to yield

$$R^{(\mathrm{u})} \le \max_{0\le\tilde{\beta}\le 1} \Big\{ R_{\mathrm{GP}}(p,\varepsilon,\tilde{\beta}) - \tilde{\pi}_1\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\tilde{\pi}_{\mathrm{d}|1};\ D/\tilde{\pi}_1\big)\Big\}. \tag{212}$$

Since the RHS of (212) coincides with the RHS of (204) (because the functional dependence of $\tilde{\pi}_1$ and $\tilde{\pi}_{\mathrm{d}|1}$ on $\tilde{\beta}$ is that same as that of π_{01} and $\pi_{\mathrm{d}|01}$ on β), we conclude that $R^{(\mathrm{u})}$ and $R^{(\mathrm{l})}$ coincide, and consequently,

$$C_{\mathrm{FB}}^{\mathrm{nc}}(D) = \max_{0\le\beta\le 1} \Big\{ R_{\mathrm{GP}}(p,\varepsilon,\beta) - \pi_{01}\, R_{\mathrm{Ber}}^{\mathrm{Ham}}\big(\pi_{\mathrm{d}|01};\ D/\pi_{01}\big)\Big\}, \tag{213}$$

where π_{01} and $\pi_{\mathrm{d}|01}$ are given as functions of β in (201) and (202).

Appendix
Recovering the Codeword in Gel'fand-Pinsker Coding

We show here that the Gel'fand-Pinsker rate is also achievable if, in addition to the message, we also wish to recover the codeword. The notation we adopt in this appendix is that of [6, Sec. 7.6.1, pp. 180–181]. The auxiliary chance variable is thus denoted U and not—as in the body of the paper—T. The standard texts on this problem show that the rate $I(U;Y) - I(U;S)$ is achievable in the sense that it allows for the reliable recovery of the message. Here we show that this rate is achievable even under the more stringent requirement that the decoder recover the transmitted codeword (and not only the subcode to which it belongs).

The coding scheme is essentially that of [6, Sec. 7.6.1, pp. 180–181] but we are a bit more particular about how the index $l^\star$ of the transmitted codeword is chosen among the indices of the codewords in the subcode $\mathcal{C}(m)$ that are jointly typical with the state sequence s^n. Our encoder considers the indices of the codewords in $\mathcal{C}(m)$ in increasing order until it hits the first index of a codeword in $\mathcal{C}(m)$ that is jointly typical with the state sequence s^n. It does not look at the codewords in $\mathcal{C}(m)$ of higher index. If no such codeword exists in $\mathcal{C}(m)$, it sets $l^\star$ to be the largest index of a codeword in the subcode, i.e., $m\, 2^{n(\tilde{R}-R)}$.

We assume that $p(u|s)$ differs from $p(u)$, i.e., that U and S are not independent:

$$I(U;S) > 0. \tag{214}$$

Otherwise, $I(U;Y) - I(U;S)$ is achievable using Shannon strategies, which can yield the codeword without the need for subcodes (or bins).

We fix $\varepsilon > \varepsilon' > 0$ with ε' small enough so that

$$\lim_{n\to\infty} \max_{s^n} \Pr\big[U^n \in \mathcal{T}_{\varepsilon'}^{(n)}(U|s^n)\big] = 0, \tag{215}$$

i.e., so that drawing U^n IID $\sim p(u)$ will rarely result in (u^n, s^n) being in $\mathcal{T}_{\varepsilon'}^{(n)}$. (Every sufficiently small ε' will do thanks to (214) and the Joint Typicality Lemma [6, Sec. 2.5.1, p. 29] with $\varepsilon \leftarrow \varepsilon'$, $X \leftarrow \emptyset$, $\tilde{y}_n \leftarrow s^n$, $\tilde{Z}^n \leftarrow U^n$.)

After observing y^n, the receiver forms the list

$$\mathcal{L}(y^n) = \left\{l \in [1 : 2^{n\tilde{R}}] : (u^n(l), y^n) \in \mathcal{T}_{\varepsilon}^{(n)}\right\}. \quad (216)$$

If this list is empty or contains indices from two different subcodes, then the receiver in [6] declares an error. To the error events in [6] we add the event

$$\mathcal{E}_4 = \left\{|\mathcal{L}(y^n)| > 1\right\},$$

(which is implied by $\mathcal{E}_2^c \cap \mathcal{E}_3$) and we declare a failure if $\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_4$. (Recall that $\mathcal{E}_1$ is the event that no codeword in the subcode $\mathcal{C}(m)$ is jointly typical with the state sequence s^n, and $\mathcal{E}_2$ is the event that the transmitted codeword is not jointly typical with the received sequence y^n [6].) The conditions that guarantee that the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$ vanish are studied in [6]. Here we focus on the probability of $\mathcal{E}_4$ and study it when the codewords are generated at random.

We assume without loss of generality that $M = 1$, and we denote the set $\mathcal{L}(y^n)$, which is now random, by $\mathbb{L}(y^n)$. We begin by conditioning on

$$S^n = s^n, \quad Y^n = y^n, \quad L^\star = l^\star, \quad M = 1, \quad (217)$$

and study the conditional probability of $\mathcal{E}_4$

$$\Pr(\mathcal{E}_4 \,|\, S^n = s^n, Y^n = y^n, L^\star = l^\star, M = 1). \quad (218)$$

Under this conditioning, the codewords $\{U^n(l),\ 1 \le l < l^\star\}$ are drawn IID according to the conditional distribution of U^n given $(U^n, s^n) \notin \mathcal{T}_{\varepsilon'}^{(n)}$ (*c.f.* [13, Lemma 1]). For each $1 \le l < l^\star$, the probability that the corresponding codeword $U^n(l)$ is in $\mathbb{L}(y^n)$ is thus

$$\begin{aligned}\Pr\left[(U^n, y^n) \in \mathcal{T}_{\varepsilon}^{(n)} \,\middle|\, (U^n, s^n) \notin \mathcal{T}_{\varepsilon'}^{(n)}\right] &\le \frac{\Pr\left[(U^n, y^n) \in \mathcal{T}_{\varepsilon}^{(n)}\right]}{\Pr\left[(U^n, s^n) \notin \mathcal{T}_{\varepsilon'}^{(n)}\right]} \quad (219)\\ &\le \frac{2^{-n(I(U;Y)-\delta(\varepsilon))}}{\alpha_n}, \quad (220)\end{aligned}$$

where we have denoted the denominator by α_n, which tends to 1 by (215); and we have used the Joint Typicality Lemma with $X \leftarrow \emptyset$, $\tilde{y}^n \leftarrow y^n$, $\tilde{Z}^n \leftarrow U^n$. Since there are $l^\star - 1$ such indices, the conditional probability that at least one of them will be in $\mathbb{L}(y^n)$ is upper bounded by

$$(l^\star - 1)\, 2^{-n(I(U;Y)-\delta(\varepsilon))}\, \alpha_n^{-1}. \quad (221)$$

As to indices larger than $l^\star$, their corresponding codewords are drawn IID p_U also under the conditioning in (217) because, once $l^\star$ has been found, the codewords in the subcode $\mathcal{C}(m)$ of index larger than $l^\star$ are no longer considered by the encoder, and nor are codewords in other subcodes. Consequently, for each $l^\star < l \le 2^{n\tilde{R}}$, the (conditional as well as unconditional) probability that it is in $\mathbb{L}(y^n)$ is upper bounded by

$$2^{-n(I(U;Y)-\delta(\varepsilon))}. \quad (222)$$

Since there are $2^{n\tilde{R}} - l^\star$ such indices, the conditional probability that at least one of them will be in $\mathbb{L}(y^n)$ is upper bounded by

$$\left(2^{n\tilde{R}} - l^\star\right) 2^{-n(I(U;Y)-\delta(\varepsilon))}. \quad (223)$$

It now follows from (221) and (223) that the conditional probability of some $l \in [1 : 2^{n\tilde{R}}] \setminus \{l^\star\}$ being in $\mathbb{L}(y^n)$ is upper bounded by

$$2^{n\tilde{R}}\, 2^{-n(I(U;Y)-\delta(\varepsilon))}\, \alpha_n^{-1}, \quad (224)$$

so

$$\begin{aligned}&\Pr(\mathcal{E}_4 \,|\, S^n = s^n, Y^n = y^n, L^\star = l^\star, M = 1)\\ &\qquad \le 2^{n\tilde{R}}\, 2^{-n(I(U;Y)-\delta(\varepsilon))}\, \alpha_n^{-1}. \quad (225)\end{aligned}$$

When $\tilde{R}$ is smaller than $I(U;Y) - \delta(\varepsilon)$ (as it is chosen in [6]), this converges to zero. This establishes that $\Pr(\mathcal{E}_4)$ tends to zero as n tends to infinity.

ACKNOWLEDGMENT

The authors thank M. Wigger and Y.-H. Kim for fruitful discussions, and they thank the anonymous referees and the Associate Editor for their helpful comments.

REFERENCES

[1] T. Berger, *Rate Distortion Theory: Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[2] S. I. Bross, A. Lapidoth, and M. Wigger, "Dirty-paper coding for the Gaussian multiaccess channel with conferencing," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5640–5668, Sep. 2012.

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

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

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

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

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

[9] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," *Found. Trends Commun. Inf. Theory*, vol. 4, no. 6, pp. 445–586, Jun. 2008.

[10] Y. H. Kim, A. Sutivong, and T. M. Cover, "State amplification," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.

[11] B. Larrousse, S. Lasaulce, and M. Wigger, "Coordination in state-dependent distributed networks: The two-agent case," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 979–983.

[12] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 589–593.

[13] P. Minero, S. H. Lim, and Y.-H. Kim, "A unified approach to hybrid coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1509–1523, Apr. 2015.

[14] H. Permuter and T. Weissman, "Source coding with a side information 'vending machine,'" *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4530–4544, Jul. 2011.

[15] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[16] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.

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

[18] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.

[19] A. Sutivong, M. Chiang, T. M. Cover, and Y.-H. Kim, "Channel capacity and state estimation for state-dependent Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1486–1495, Apr. 2005.

[20] F. M. J. Willems and E. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 313–327, May 1985.

[21] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. Prospect Heights, IL, USA: Waveland Press, Inc., 1990.

[22] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 3, pp. 294–300, May 1975.

[23] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Inf. Control*, vol. 38, no. 1, pp. 60–80, 1978.

[24] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–11, Jan. 1976.

Shraga I. Bross (S'89–M'92–SM'09) received the B.Sc. and M.Sc. degrees from the Technion—Israel Institute of Technology, Haifa, in 1978 and 1983, and the Ph.D. degree from the University of Maryland, College Park in 1991, all in electrical engineering. During the 1991-1992 academic year he was a Post-doctoral Fellow in the ECE Department at the University of Waterloo, Canada. During 1992–1998 he was with Orckit Communications Ltd., Telaviv, Israel, in the capacity of a Senior Scientist. From 1998 to 2006 he was a Senior Research Fellow at the EE Department, Technion. Since 2007 he is with the Faculty of Engineering, Bar-Ilan University, Israel, where he is now an Associate Professor. His research interests are in Digital Communications and Information Theory.

Amos Lapidoth (S'89–M'95–SM'00–F'04) received the B.A. degree in Mathematics (*summa cum laude*, 1986), the B.Sc. degree in Electrical Engineering (*summa cum laude*, 1986), and the M.Sc. degree in Electrical Engineering (1990) all from the Technion—Israel Institute of Technology. His Ph.D. degree in Electrical Engineering is from Stanford University (1995).

In the years 1995–1999 he was an Assistant and Associate Professor at the department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT), and was the KDD Career Development Associate Professor in Communications and Technology. He is now Professor of Information Theory at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland

His research interests are in Digital Communications and Information Theory. He is the author of the textbook *A Foundation in Digital Communication*, second edition, Cambridge University Press, 2017.