Conveying Data and State with Feedback

Shraga I. Bross
Faculty of Engineering, Bar-Ilan University
Ramat-Gan 52900, Israel
brosss@biu.ac.il

Amos Lapidoth
ETH Zurich
CH-8092 Zurich, Switzerland
lapidoth@isi.ee.ethz.ch

***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.**

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 is typically increased 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 cases depending on whether the state-information (SI) is revealed to the encoder strictly-causally or 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, in addition to fresh data, the encoder also transmits 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 description pertaining to Block-$(b-1)$ that was transmitted in Block-b. 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 allows 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 [2]. Particularly relevant to our setting is [4], which deals with the Gaussian channel with noncausal SI and mean squared-error state-reconstruction fidelity. Also relevant to us is [3], where 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 [3] is thus more of a "guessing" nature than an "estimation" nature. For this problem [3] characterizes the tension between the data rate and the exponential growth of the list-size in the blocklength. The converse in [3] is based on the extension of Fano's inequality to lists and is hence inapplicable to our setting.

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 $\big(P_{\mathrm{c}}(y|x,s), P_S\big)$, 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 is a pair $(\tilde{Y}, \tilde{S})$, the state is S of PMF P_S, and the transition law factorizes as

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

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 that observes $\tilde{S}$. 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 [5] 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 [5, Section 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. (For more on the no-feedback case see [1].)

The work of S. Bross was supported by the Israel Science Foundation under Grant 455/14.

978-1-5090-1806-2/16/$31.00 ©2016 IEEE

II. The Set-Up

We are given a SD-DMC $\big(P_{\rm c}(y|x,s), P_S\big)$ and a nonnegative distortion function $d\colon \mathcal{S}\times\hat{\mathcal{S}} \to [0,\infty)$, where $\hat{\mathcal{S}}$ is the reconstruction alphabet, and where the alphabets $\mathcal{X}$, $\mathcal{S}$, $\mathcal{Y}$, $\hat{\mathcal{S}}$ are all finite. The maximum of d is finite and is denoted $d_{\max}$:

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

The distortion between the n-length sequences $\mathbf{s} \in \mathcal{S}^n$ and $\tilde{\mathbf{s}} \in \hat{\mathcal{S}}^n$ is defined as the average distortion between the corresponding components

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

We are also given some allowed distortion D. Let n denote the blocklength, R the data rate, and $\mathcal{W} = \{1,\dots e^{nR}\}$ the set of messages. In all our settings the decoder consists of two mapping. The first mapping

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

is used to decode the message, and we denote by $\hat{W}$ the result of applying it to the received sequence $\mathbf{Y}$. 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 ϕ_S to $\mathbf{Y}$. (Throughout we denote n-length sequences with bold letters, so $\mathbf{Y}$ stands for $Y_1,\dots,Y_n$. We use Y_i^j for $Y_i,\dots,Y_j$ and suppress i when it is 1.)

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

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

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

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

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 $P_{\rm e}^{(n)}$.

A pair (R,D) is achievable if for every $\varepsilon>0$ there exists some positive integer $n_0(\varepsilon)$ such that for every blocklength n exceeding $n_0(\varepsilon)$ there exists 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

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

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 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 the maximum exists since $\mathscr{R}$ is closed. The different settings have different RnS capacities,

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

all of which are denoted by 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 allowed distortion D is zero, and the distortion function is the Hamming distortion function $(s,\tilde{s}) \mapsto \mathbb{1}\{\tilde{s}\neq s\}$, which is zero when $\hat{s}$ equals s, and is one otherwise. (Here and throughout $\mathbb{1}\{\text{statement}\}$ is one or zero depending on whether or not the statement holds).

Remark 1: This case is reminiscent of the case where Δ in [3] is $H(S)$. It is not identical because the latter case corresponds to a subexponential list and our case corresponds to an ε-ball.

Finally, although not of finite alphabet, the Gaussian channel is also of interest to us. This is a memoryless channel, where $Y = x+S+Z$, 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 $\sum_{k=1}^n \mathbb{E}\big[X_k^2\big] \le nP$ for some given maximal-allowed average power P.

III. Main Results

Except when we discuss the Gaussian channel, we assume throughout a SD-DMC $\big(P_{\rm c}(y|x,s), P_S\big)$ with finite alphabets and a (finite) nonnegative distortion function $d\colon \mathcal{S}\times\hat{\mathcal{S}} \to [0,\infty)$. 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): For every $D \ge d_{\min}$,

$$C^{\text{s-c}}_{\text{FB}}(D) = \max_{P_X,\, P_{U|XSY},\, g} \Big\{ I(X;Y) - I(S;U|XY) \Big\}, \tag{10}$$

where U is an auxiliary chance variable taking values in a set $\mathcal{U}$; the mutual informations are computed w.r.t. the joint PMF

$$P_{SXYU}(s,x,y,u) = P_S(s)\,P_X(x)\,P_{\rm c}(y|x,s) \cdot P_{U|XSY}(u|x,s,y); \tag{11}$$

the mapping g is of the form

$$g\colon \mathcal{U}\times\mathcal{X}\times\mathcal{Y} \to \hat{\mathcal{S}}; \tag{12}$$

and we require that

$$\mathbb{E}\big[d\big(S, g(U,X,Y)\big)\big] \le D, \tag{13}$$

where $d_{\min} = \max\{D \colon C^{\text{s-c}}_{\text{FB}}(D) = 0\}$. The cardinality of the set $\mathcal{U}$ in which U takes values can be bounded by $|\mathcal{U}| \leq |\hat{\mathcal{S}}|$. Alternatively, $C^{\text{s-c}}_{\text{FB}}(D)$ can be expressed for $D \geq d_{\min}$ as

$$C^{\text{s-c}}_{\text{FB}}(D) = \max_{P_X} \Big\{ I(X;Y) - R_{S|XY}(D) \Big\}, \tag{14}$$

where $R_{S|XY}(\cdot)$ is the rate-distortion function of the source S when both encoder and reconstructor are cognizant of (X,Y) and the joint law of X, Y, S is $P_X(x)\, P_S(s)\, P_{\text{c}}(y|x,s)$.

In general, $C^{\text{s-c}}_{\text{FB}}(D)$ can exceed $C^{\text{s-c}}(D)$, but in the lossless case they are equal:

Proposition 1 (Lossless Reconstruction: Strictly-Causal SI): Suppose $\tilde{\mathcal{S}}$ equals $\mathcal{S}$, and $d(s,\tilde{s})$ is the Hamming distortion function. When the maximal allowed distortion D is zero and the state is revealed to the encoder strictly causally,

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

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

Another example where feedback does not increase the RnS capacity is the Gaussian channel:

Proposition 2 (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 be revealed to the encoder strictly-causally, and let the distortion measure be $(s,\hat{s}) \mapsto (s-\hat{s})^2$. In this setting $C^{\text{s-c}}(D) = C^{\text{s-c}}_{\text{FB}}(D)$ for every $D > 0$. Moreover, if we define for every $0 \leq \gamma \leq 1$ the quantities

$$R_\gamma = \frac{1-\gamma}{2} \log\left(\frac{P + \sigma_s^2 + N}{\sigma_s^2 + N} \right) \tag{15a}$$

$$D_\gamma = \sigma_s^2 \frac{N}{\sigma_s^2 + N} \left(\frac{\sigma_s^2 + N}{P + \sigma_s^2 + N} \right)^{\gamma}, \tag{15b}$$

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 $C^{\text{s-c}}_{\text{FB}}$ (and hence also $C^{\text{s-c}}$) is given parametrically by

$$C^{\text{s-c}}_{\text{FB}}(D_\gamma) = R_\gamma, \quad 0 \leq \gamma \leq 1. \tag{16}$$

B. Causal State Information

Theorem 2 (Causal SI and Feedback): For $D \geq d_{\min}$,

$$C^{\text{c}}_{\text{FB}}(D) = \max_{P_T,\, P_{U|TSY},\, f,\, g} \Big\{ I(T;Y) - I(S;U|TY) \Big\} \tag{17}$$

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 w.r.t. the joint PMF

$$\begin{aligned} P_{STXYU}(s,t,x,y,u) &= P_S(s)\, P_T(t)\, \mathbb{1}\{x = f(t,s)\} \\ &\quad \cdot P_{\text{c}}(y|x,s)\, P_{U|TSY}(u|t,s,y); \end{aligned} \tag{18}$$

the mapping g is of the form

$$g \colon \mathcal{U} \times \mathcal{T} \times \mathcal{Y} \to \hat{\mathcal{S}}; \tag{19}$$

and we require that

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

where $d_{\min} = \min\{D \colon C^{\text{c}}_{\text{FB}}(D) = 0\}$. Moreover, in the above maximization we may restrict the cardinalities to $|\mathcal{U}| \leq |\hat{\mathcal{S}}|$ and $|\mathcal{T}| \leq \min\{|\mathcal{X}| \cdot |\mathcal{S}|, |\mathcal{Y}|\} + 1$.

Alternatively, $C^{\text{c}}_{\text{FB}}(D)$ can be expressed for $D \geq d_{\min}$ as

$$C^{\text{c}}_{\text{FB}}(D) = \max_{P_T,\, f} \Big\{ I(T;Y) - R_{S|TY}(D) \Big\}, \tag{21}$$

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

$$P_{STY}(s,t,y) = P_S(s)\, P_T(t)\, P_{\text{c}}\big(y|f(t,s), s\big). \tag{22}$$

As in the strictly-causal case, when the reconstruction of the state must be lossless, feedback is not needed:

Proposition 3 (Lossless Reconstruction: Causal SI): If the state is revealed to the encoder causally and we require lossless reconstruction, then

$$\begin{aligned} C^{\text{c}}(0) &= C^{\text{c}}_{\text{FB}}(0) \\ &= \max_{P_T,\, f} \Big\{ I(T;Y) - H(S|TY) \Big\}, \end{aligned}$$

whenever the RHS of the above is nonnegative . Here I and H are computed with respect to the joint PMF

$$P_S(s)\, P_T(t)\, \mathbb{1}\big\{x = f(t,s)\big\}\, P_{\text{c}}(y|x,s), \tag{23}$$

and the mapping f is as in Theorem 2.

C. Noncausal State Information

For the noncausal case we only provide bounds. Define

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

where U and T are auxiliary chance variables taking values in $\mathcal{U}$ and $\mathcal{T}$ respectively; where the mutual informations are computed w.r.t. the joint PMF

$$\begin{aligned} P_{STXUY} &= P_S(s)\, P_{TX|S}(t,x|s)\, P_{\text{c}}(y|x,s) \\ &\quad \cdot P_{U|STXY}(u|s,t,x,y); \end{aligned} \tag{24}$$

the mapping g is from $\mathcal{U} \times \mathcal{T} \times \mathcal{Y}$ to $\hat{\mathcal{S}}$; and we require that

$$\mathbb{E}\big[d\big(S, g(U,T,Y)\big) \big] \leq D. \tag{25}$$

Define

$$\begin{aligned} R^{(\text{u})} = \max_{P_{TX|S},\, P_{U|STXY},\, g} \min \, \{ & I(T;Y) - I(T;S), \\ & I(XS;Y) - I(S;UTY) \}, \end{aligned}$$

where the mutual informations are computed w.r.t. (24) under the constraint (25).

Theorem 3 (Noncausal SI and Feedback):

$$R^{(\text{l})} \leq C^{\text{nc}}_{\text{FB}}(D) \leq R^{(\text{u})}. \tag{26}$$

If $R^{(\text{u})}$ is attained by a law under which X is a deterministic function of (S,T), then $C^{\text{nc}}_{\text{FB}}(D) = R^{(\text{l})} = R^{(\text{u})}$.

Remark 2 (Gaussian Channel: Noncausal SI): If the channel is Gaussian and fidelity is measured in terms of mean squared-error, the (R,D) tradeoff is characterized in [4, Theorem 2]. The converse proof in [4, Section III.B] can be modified to account for output feedback and to yield the same bound. Consequently, output feedback does not increase the (R,D) tradeoff region also in this case.

IV. Proof of Theorem 1: Sketch

The proofs can be found in [1]. Here we only outline the proof of Theorem 1.

1) Converse: Before proving the converse we denote the r.h.s. of (10) by $\tilde{C}_{\text{FB}}^{\text{s-c}}(D)$ and study some of its properties. In the following we denote the nonnegative reals by $\mathbb{R}_+$.

Proposition 4: The function $\tilde{C}_{\text{FB}}^{\text{s-c}}\colon \mathbb{R}_+ \to \mathbb{R}_+$ is monotonically nondecreasing and upper bounded by the channel capacity $\mathcal{C}_{X\to Y}$

$$\tilde{C}_{\text{FB}}^{\text{s-c}}(D) \le \mathcal{C}_{X\to Y}, \tag{27}$$

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

Proof: See [1]. ■

We can now prove the converse. Let $\left(\{f_k\}_{k=1}^n, \phi_W^{(n)}, \phi_S^{(n)}\right)$ define an achievable (R, D) pair, so $n^{-1}\log|\mathcal{W}| \ge R-\varepsilon$ and $\mathbb{E}[d(S^n, \hat{S}^n)] \le D+\varepsilon$. We will show that

$$\frac{1}{n}\log|\mathcal{W}| \le \tilde{C}_{\text{FB}}^{\text{s-c}}(D+\varepsilon), \tag{28}$$

which will thus imply that

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

The inequality $R \le \tilde{C}_{\text{FB}}^{\text{s-c}}(D)$ will then follow from the continuity of $\tilde{C}_{\text{FB}}^{\text{s-c}}$ by letting ε tend to zero.

Expanding the following conditional mutual information in two ways and noting that $\hat{S}_k$ is a function of Y^n, we obtain

$$\begin{aligned}
&I(Y^n\hat{S}_k; S_k|WS^{k-1}) \\
&\quad= I(Y^n; S_k|WS^{k-1}) + I(\hat{S}_k; S_k|WY^nS^{k-1}) \\
&\quad= I(Y^n; S_k|WS^{k-1}) \\
&\quad= I(\hat{S}_k; S_k|WS^{k-1}) + I(Y^n; S_k|W\hat{S}_kS^{k-1}).
\end{aligned} \tag{29}$$

Since S_k is independent of (W, S^{k-1}),

$$\sum_{k=1}^n I(\hat{S}_k; S_k|WS^{k-1}) = \sum_{k=1}^n I(W\hat{S}_kS^{k-1}; S_k). \tag{30}$$

Define the auxiliary random variables

$$V_k \triangleq (W, S^{k-1}),\ U_k \triangleq Y^{n\setminus k}, \tag{31}$$

and note that V_k is independent of S_k, and that $\hat{S}_k$ is a deterministic function of (U_k, Y_k). By Fano's inequality and (29)

$$\begin{aligned}
&n(R-\eta_n) + \sum_{k=1}^n I(\hat{S}_k; S_k|WS^{k-1}) \\
&\le I(W;Y^n) + \sum_{k=1}^n [I(Y^n; S_k|WS^{k-1}) \\
&\qquad\qquad -I(Y^n; S_k|W\hat{S}_kS^{k-1})] \\
&= I(W;Y^n) + I(Y^n; S^n|W) - \sum_{k=1}^n I(Y^n; S_k|W\hat{S}_kS^{k-1}) \\
&= I(WS^n; Y^n) - \sum_{k=1}^n I(Y^n; S_k|W\hat{S}_kS^{k-1}) \\
&= \sum_{k=1}^n [I(Y_k; WS^n|Y^{k-1}) - I(S_k; Y^n|W\hat{S}_kS^{k-1})] \\
&\overset{(a)}{=} \sum_{k=1}^n [I(Y_k; WX_kS^n|Y^{k-1}) - I(S_k; Y^n|W\hat{S}_kS^{k-1})] \\
&= \sum_{k=1}^n [H(Y_k|Y^{k-1}) - H(Y_k|X_kWS^nY^{k-1}) \\
&\qquad\qquad -I(S_k; Y^n|W\hat{S}_kS^{k-1})] \\
&\overset{(b)}{\le} \sum_{k=1}^n [H(Y_k) - H(Y_k|X_kS_k) - I(S_k; Y^n|W\hat{S}_kS^{k-1})] \\
&= \sum_{k=1}^n [I(Y_k; X_kS_k) - I(S_k; Y_kU_k|V_k\hat{S}_k)] \\
&= \sum_{k=1}^n [I(X_k; Y_k) + I(S_k; Y_k|X_k) - I(S_k; Y_kU_k|V_k\hat{S}_k)] \\
&\overset{(c)}{=} \sum_{k=1}^n [I(X_k; Y_k) + H(S_k) - H(S_k|Y_kX_k) \\
&\qquad\qquad -H(S_k|V_k\hat{S}_k) + H(S_k|X_kY_kU_kV_k\hat{S}_k)] \\
&\le \sum_{k=1}^n [I(X_k; Y_k) + H(S_k) - H(S_k|Y_kX_k) \\
&\qquad\qquad -H(S_k|V_k\hat{S}_k) + H(S_k|X_kY_kU_k)] \\
&= \sum_{k=1}^n [I(X_k; Y_k) - I(S_k; U_k|Y_kX_k) + I(S_k; V_k\hat{S}_k)].
\end{aligned} \tag{32}$$

Here

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

Substituting (30) into the l.h.s. of (32) we obtain

$$n(R-\eta_n) \le \sum_{k=1}^n [I(X_k; Y_k) - I(S_k; U_k|Y_kX_k)]. \tag{33}$$

Let J be a r.v. uniformly distributed over $\{1, \ldots, n\}$ and independent of $\{(X_k, Y_k, S_k, U_k, \hat{S}_k)\}, k = 1, \ldots, n$, and define $U = (U_J, J), S = S_J, Y = Y_J, X = X_J$, and $\hat{S} = \hat{S}_J$. Using J we may express (33) as

$$\begin{aligned}
&R-\eta_n \le \frac{1}{n}\sum_{k=1}^n [I(X_k; Y_k) - I(S_k; U_k|Y_kX_k)] \\
&= I(X_J; Y_J|J) - I(S_J; U_J|Y_J, X_J, J) \\
&= I(X_J; Y_J|J) - I(S_J; U_J, J|Y_J, X_J) + I(S_J; J|Y_J, X_J) \\
&\overset{(d)}{\le} I(X;Y) - I(S;U|XY).
\end{aligned} \tag{34}$$

Here, step (d) follows since $J \multimap (X_J, Y_J) \multimap S_J$ is a Markov chain hence $I(S_J; J|Y_J, X_J) = 0$, and since $I(X;Y) = I(P_X; W_{Y|X})$ is concave in P_X.

We consider now the expected distortion, for a given distortion constraint D,

$$D+\varepsilon \geq \mathbb{E}[d(S^n,\hat{S}^n)] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d(S_k,\hat{S}_k)\big]$$
$$= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d(S_k,\hat{S}_k(Y^n))\big] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[d(S_k,\hat{S}_k(Y^{n\setminus k},Y_k)\big]$$
$$= \mathbb{E}\big[d(S_J,\hat{S}_J(U_J,Y_J,J)\big] = \mathbb{E}\big[d(S,\hat{S}(U,Y)\big]. \quad (35)$$

Finally, it can be verified that (34) and (35) hold for a law of the form (11).

2) Direct part: We propose a coding scheme that is based on Block-Markov superposition encoding and backward decoding.

We consider B blocks, each of n symbols. A sequence of $B-1$ messages $W^{(b)}$, for $b=1,\dots,B-1$, will be transmitted during B transmission blocks. Here the sequence $\{W^{(b)}\}$ is an i.i.d. sequence of uniform random variables over $\{1,\dots,e^{nR}\}$. As $B\to\infty$, for fixed n, the message rate, $\tilde{R}=R(B-1)/B$, is arbitrarily close to R.

We assume a tuple of random variables $S\in\mathcal{S}, U\in\mathcal{U}, X\in\mathcal{X}, Y\in\mathcal{Y}$, of joint law (11), and fix a sufficiently small $\varepsilon>0$. We will demonstrate that if $R\leq \tilde{C}^{\text{s-c}}_{\text{FB}}(D)$ for a reconstruction mapping of the form (12) such that $\mathbb{E}[d(S,g(U,X,Y)]\leq D$ then $(R,D)\in\mathscr{R}$.

Let X be a chance variable with PMF P_X on a finite set $\mathcal{X}$ then denote by $\mathcal{T}^{(n)}_{\delta}(X)$ the set of all $x^n\in\mathcal{X}^n$ which are strongly δ-typical w.r.t. X (for the formal definition see [6]).

Random coding: In each block $b, b=1,2,\dots,B$, we use the following construction.

- Generate $e^{n(R+R_s)}$ sequences $\boldsymbol{x}=(x_1,\dots,x_n)$, each with probability $\Pr(\boldsymbol{x})=\prod_{k=1}^{n}p_X(x_k)$. Label them $\boldsymbol{x}(w,\omega_0)$ where $w\in\{1,\dots,e^{nR}\}$ and $\omega_0\in\{1,\dots,e^{nR_s}\}$.
- For each typical sequence $\boldsymbol{y}$ generate e^{nR_s} sequences $\boldsymbol{u}=(u_1,u_2,\dots,u_n)$, each with probability $\Pr(\boldsymbol{u}|\boldsymbol{y})=\prod_{k=1}^{n}p_{U|Y}(u_k|y_k)$. Label them $\boldsymbol{u}(j)$ where $j\in\{1,\dots,e^{nR_s}\}$.

Encoding : We denote the realization of the messages sequence $\{W^{(b)}\}$ by $\{w^{(b)}\}$, and the realization of the state sequence $(S^{(b)}_1,S^{(b)}_2,\dots,S^{(b)}_n)$ by $\boldsymbol{s}^{(b)}$.

Let $j^{(b)}$ be the index such that, conditioned on $\boldsymbol{y}^{(b)}$, the sequence $\boldsymbol{u}(j^{(b)})$ is jointly typical with the pair $(\boldsymbol{s}^{(b)},\boldsymbol{x}(w^{(b)},\omega^{(b)}_0))$. The code builds upon a Block-Markov structure in which a quantized description of both the state sequence and the input sequence $(\boldsymbol{s}^{(b)},\boldsymbol{x}(w^{(b)},\omega^{(b)}_0))$ is encoded over the successive blocks b and $(b+1)$ such that $\omega^{(b+1)}_0=j^{(b)}$, for $b=1,\dots,B-1$.

The sequence of messages $\{w^{(b)}\}$, $b=1,2,\dots,B-1$ is encoded as follows:

In Block 1 the encoder sends $\boldsymbol{x}^{(1)}=\boldsymbol{x}(w^{(1)},1)$ –i.e. $\omega^{(1)}_0=1$. Upon observing $\boldsymbol{s}^{(b-1)}, b=2,3,\dots,B$, the encoder computes $j^{(b-1)}$ by finding the index $j^{(b-1)}$ such that, conditioned on $\boldsymbol{y}^{(b-1)}$, the sequence $\boldsymbol{u}(j^{(b-1)})$ is jointly typical with $(\boldsymbol{s}^{(b-1)},\boldsymbol{x}(w^{(b-1)},\omega^{(b-1)}_0))$. Then, in Block $b=2,3,\dots,B-1$, the encoder sends

$$\boldsymbol{x}^{(b)}=\boldsymbol{x}(w^{(b)},j^{(b-1)}),$$

and in Block B it sends $\boldsymbol{x}^{(B)}=\boldsymbol{x}(1,j^{(B-1)})$.

Decoding at the receiver: After the reception of Block B the receiver uses backward decoding starting from Block B to Block 1 and decodes the messages as well as the sequence $\{\boldsymbol{u}(j^{(b)})\}, b=1,\dots,B-1$, as follows.

In Block B the receiver looks for $\hat{\jmath}^{(B-1)}$ such that

$$\big(\boldsymbol{x}(1,\hat{\jmath}^{(B-1)}),\boldsymbol{y}^{(B)}\big)\in\mathcal{T}^{(n)}_{\varepsilon}(X,Y).$$

Next, assume that, decoding backwards up to (and including) Block $b+1$, the receiver decoded $\hat{\jmath}^{(B-1)},(\hat{w}^{(B-1)},\hat{\jmath}^{(B-2)}),\dots,(\hat{w}^{(b+1)},\hat{\jmath}^{(b)})$. To decode Block $b, b=B-1,\dots,2$ the receiver looks for $(\hat{w}^{(b)},\hat{\jmath}^{(b-1)})$ such that

$$\big(\boldsymbol{x}(\hat{w}^{(b)},\hat{\jmath}^{(b-1)}),\boldsymbol{u}(\hat{\jmath}^{(b)}),\boldsymbol{y}^{(b)}\big)\in\mathcal{T}^{(n)}_{\varepsilon}(X,U,Y),$$

while in Block 1 the receiver looks for $\hat{w}^{(1)}$ such that

$$\big(\boldsymbol{x}(\hat{w}^{(1)},1),\boldsymbol{u}(\hat{\jmath}^{(1)}),\boldsymbol{y}^{(1)}\big)\in\mathcal{T}^{(n)}_{\varepsilon}(X,U,Y).$$

State estimation at the receiver: The receiver forms its estimate of $\boldsymbol{s}^{(b)}$ symbol-wise as follows:

$$\hat{\boldsymbol{s}}^{(b)}=g\big(\boldsymbol{u}(\hat{\jmath}^{(b)}),\boldsymbol{x}(\hat{w}^{(b)},\hat{\jmath}^{(b-1)}),\boldsymbol{y}^{(b)}\big)$$

where g is defined by (12).

When a decoding step either fails to recover a unique index (or index pair) which satisfies the decoding rule, or there is more than one index (or index pair), then an index (or an index pair) is chosen at random.

The error probability analysis for our code construction establishes that, if

$$R+R_s<I(X;UY)$$
$$R_s>I(U;SX|Y), \quad (36)$$

there exists a code satisfying $P^{(n)}_{\text{e}}\to 0$ and $\mathbb{E}[d(S^n,\hat{S}^n)]\leq D+\varepsilon$, for sufficiently large n –i.e. if $R\leq\tilde{C}^{\text{s-c}}_{\text{FB}}(D)$ then $(R,D)\in\mathscr{R}$. The details can be found in [1].

References

[1] S. I. Bross and A. Lapidoth,"The rate-and-state capacity with feedback," In preparation.

[2] G. Keshet, Y. Steinberg and N. Merhav, "Channel coding in the presence of side information", *Foundations and Trends in Communications and Information Theory,* 4(6), pp. 445-586, 2007.

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

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

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

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.