This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2023.3305539

State-Dependent DMC with a Causal Helper

Amos Lapidoth, *Fellow, IEEE* and Ligong Wang, *Member, IEEE*

***Abstract*—A memoryless state sequence governing the behavior of a memoryless state-dependent channel is to be described causally to an encoder wishing to communicate over said channel. Given the maximal-allowed description rate, we seek the description that maximizes the Shannon capacity. It is shown that the maximum need not be achieved by a memoryless (symbol-by-symbol) description. Such descriptions are, however, optimal when the receiver is cognizant of the state sequence or when the description is allowed to depend on the message. For other cases, a block-Markov scheme with backward decoding is proposed.**

***Index Terms*—Block-Markov coding, channel capacity, causal state information, helper, Shannon strategy, state-dependent channel.**

I. Introduction and Problem Setup

The impact of state information on the capacity of a state-dependent discrete memoryless channel (SD-DMC) is well understood. State information at the receiver can be accounted for by appending it to the output, and the impact of state information at the transmitter depends on its timing: if it is provided strictly causally, it has no impact on capacity; if causally, then the capacity is as given by Shannon [1]; and if noncausally, then as given by Gel'fand and Pinsker [2]. Less studied is how the state information should be conveyed to the encoder when rate restrictions preclude its precise description. We address this issue here by studying the design and impact of rate-limited causal state descriptions to the encoder.

We account for the rate constraint by requiring that the time-i assistance provided to the encoder take value in some fixed set $\mathcal{T}$, whose cardinality $|\mathcal{T}|$ is typically smaller than that of the state alphabet $\mathcal{S}$. (When $|\mathcal{T}| \geq |\mathcal{S}|$ we are back to Shannon's causal state information, because the helper can then describe the state precisely.) We assume throughout that

$$|\mathcal{T}| \geq 2 \tag{1}$$

because, otherwise, the description is of no help. We refer to $\log|\mathcal{T}|$ as the *description rate*.

This way of accounting for rate restrictions is quite rigid: it allows for neither variable-length state descriptions nor for time sharing between fine and coarse quantizations. Precluding these techniques sharpens some of our conclusions.

It should be emphasized that causality does not imply that the helper must describe the state sequence "symbol-by-symbol," i.e., that the time-i assistance $T_i \in \mathcal{T}$ be determined by the time-i state S_i: the time-i assistance may depend on the entire state sequence up to time i, namely, S^i.

The fact that we only consider memoryless channels with independent and identically ditributed (IID) states might lead one to suspect that symbol-by-symbol helpers are optimal. Lending credence to this suspicion might be that, when the causal state information is perfect (i.e., when $|\mathcal{T}| \geq |\mathcal{S}|$), Shannon strategies achieve capacity, and those ignore the past states and set the time-i channel input $X_i(m, S^i)$ to be a function of the message m and the time-i state S_i only. But this is not the case. As we shall see in Example 9 ahead, causal helpers can outperform the best symbol-by-symbol helpers. This example will motivate us to propose a block-Markov communication scheme that can outperform all symbol-by-symbol schemes.

Symbol-by-symbol descriptions are, however, optimal in some special cases, e.g., when the state is known perfectly to the receiver (Theorem 6). They are also optimal when the helper is cognizant of the message (Theorem 4). They are effective in the sense that if some positive rate is achievable with a general rate-limited causal helper, then a positive rate is also achievable with a symbol-by-symbol helper. In fact, subject to (1), symbol-by-symbol helpers can achieve positive communication rates whenever the SD-DMC is of positive capacity when its state is revealed perfectly to both encoder and decoder (Theorem 3).

We only present results for point-to-point channels. State information is, of course, of great importance also in multi-user settings, but those are more complicated and are not even fully understood when the helper's rate suffices for perfect state description. In such settings strictly-causal state information may be helpful, and Shannon strategies are sub-optimal [3]–[5]. For much more on state information in communication systems, see [6]. Related to our work is [7] which—subject to the assumption of perfect receiver state information—analyzes a system comprising a symbol-by-symbol and a noncausal helper to the transmitter.

A. The channel model

We consider a state-dependent discrete memoryless channel with finite input, output, state, and description alphabets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{S}$, and $\mathcal{T}$, respectively. The states $S_1, S_2, \ldots$ are IID $\sim P_S$, where P_S is some given probability mass function (PMF) on $\mathcal{S}$. The channel is memoryless, and we denote its law $W(y|x,s)$: given the channel input $x \in \mathcal{X}$ and the channel state $s \in \mathcal{S}$, the probability of the channel output being $y \in \mathcal{Y}$ is $W(y|x,s)$ irrespectively of past inputs, states, or outputs. A fortiori, for every $x \in \mathcal{X}$ and $s \in \mathcal{S}$, the transition law $W(y|x,s)$ is nonnegative and $\sum_{y\in\mathcal{Y}} W(y|x,s) = 1$.

When communicating with blocklength n, the transmitted message m is chosen from the set of messages $\mathcal{M} = \{1, \ldots, 2^{nR}\}$, where R is the communication rate.

When the encoder is provided with perfect causal state information—a setting to which we refer as the "Shannon set-

The authors are with the Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland (e-mail: lapidoth@isi.ee.ethz.ch; ligwang@isi.ee.ethz.ch).

ting," because this is the setting studied in [1]—a blocklength-n encoder consists of n mappings

$$f_i\colon \mathcal{M}\times\mathcal{S}^i\to\mathcal{X},\quad (m,s^i)\mapsto x_i,\quad i=1,2,\ldots,n \tag{2}$$

with the understanding that the time-i channel input X_i that the encoder produces in order to convey the message m when the present and past states are S^i is $X_i=X_i(m,S^i)=f_i(m,S^i)$.

When the encoder is provided with *perfect strictly-causal* state information, these mappings take the form

$$f_i\colon \mathcal{M}\times\mathcal{S}^{i-1}\to\mathcal{X},\quad (m,s^{i-1})\mapsto x_i,\quad i=1,2,\ldots,n \tag{3}$$

with X_i now being a function of the message and past (and not present) states, i.e., $X_i=X_i(m,S^{i-1})=f_i(m,S^{i-1})$.

The focus of our work, however, is the *causal helper setting*, where the helper consists of a sequence of mappings

$$h_i\colon \mathcal{S}^i\to\mathcal{T},\quad s^i\mapsto t_i,\quad i=1,2,\ldots,n \tag{4}$$

with the understanding that the time-i help[1] T_i that is provided to the encoder is $T_i=T_i(S^i)=h_i(S^i)$. The time-i channel input produced by the encoder to convey the message m is then a function of m and the present-and-past help T^i. The encoder thus consists of a sequence of mappings

$$f_i\colon \mathcal{M}\times\mathcal{T}^i\to\mathcal{X},\quad (m,t^i)\mapsto x_i,\quad i=1,2,\ldots,n \tag{5}$$

with the understanding that the time-i channel input X_i produced by the encoder to convey the message m, after having received the present-and-past help T^i, is $X_i=X_i(m,T^i)=f_i(m,T^i)$. A special kind of helper is the *symbol-by-symbol* helper whose time-i help T_i is a function only of the time-i state S_i, i.e., it is a helper where the mappings (4) have the form

$$h_i\colon \mathcal{S}\to\mathcal{T},\quad s_i\mapsto t_i. \tag{6}$$

In all the above cases, a decoder is a mapping

$$g\colon \mathcal{Y}^n\to\mathcal{M},\quad y^n\mapsto\hat{m} \tag{7}$$

with the understanding that, upon receiving the channel output sequence Y^n, the decoder produces the message $g(Y^n)$. A decoder with perfect state information is a mapping

$$g\colon \mathcal{Y}^n\times\mathcal{S}^n\to\mathcal{M},\quad (y^n,s^n)\mapsto\hat{m} \tag{8}$$

with the decoded message now being $g(Y^n,S^n)$.

A rate R is said to be achievable if there exists a sequence of helpers, encoders, and decoders of said rate of average probability of error tending to zero, where the average is over the uniformly drawn message and the random state sequence. The capacity is the supremum of the achievable rates.

We recall that the capacity of an SD-DMC with perfect strictly-causal state information at the encoder equals the capacity without state information; i.e., strictly causal state information does not increase capacity. The capacity with perfect causal state information is given by [1]

$$\max I(U;Y), \tag{9a}$$

where the maximum is over the choice of the set $\mathcal{U}$ and over the joint distributions of the form

$$P_S(s)\,P_U(u)\,P_{X|US}(x|u,s)\,W(y|x,s). \tag{9b}$$

Without reducing the maximum, the conditional PMF $P_{X|US}$ above can be chosen to be deterministic. Thus, $\mathcal{U}$ can be taken as the set of all mappings from $\mathcal{S}$ to $\mathcal{X}$. We sometimes refer to these mappings as "Shannon strategies."

The rest of this paper is organized as follows: Section II presents our capacity results for various settings; Section III provides three counterexamples demonstrating, respectively, that revealing the message to the helper increases capacity, that symbol-by-symbol helpers need not be optimal, and that they need not be optimal even when the help is provided to both encoder and decoder; Section IV presents a block-Markov scheme that can outperform all symbol-by-symbol helpers; and Section V offers some intuition for some of the results and concludes the paper.

II. Capacity Results

A. Symbol-by-symbol helper

The following result on symbol-by-symbol helpers is a small variation on known results. Given a fixed symbol-by-symbol helper, we can view $h(S_i)$ as a new state and then invoke Shannon's result to obtain the best achievable rate that can be achieved with said helper [8], [9]. We include this theorem for completeness and because our definition of a symbol-by-symbol helper does not require that the functions in (6) be time invariant, i.e., not depend on i. Also, we allow for the past states (unquantized) to be revealed to the encoder. The converse must thus be slightly modified.

Theorem 1: If only symbol-by-symbol helpers are permitted, then the capacity is

$$\max I(U;Y), \tag{10a}$$

where the maximum is over the choice of a set $\mathcal{U}$ and over the joint distributions of the form

$$P_S(s)\,P_U(u)\,P_{T|S}(t|s)\,P_{X|UT}(x|u,t)\,W(y|x,s), \tag{10b}$$

where—without reducing the maximum—the cardinality of $\mathcal{U}$ may be restricted to $|\mathcal{X}|^{|\mathcal{T}|}$, and the conditional PMFs $P_{T|S}$ and $P_{X|UT}$ can be chosen to be deterministic.

This is also the capacity if the time-i channel input produced by the encoder may depend not only on m and T^i but also on the past states S^{i-1}, i.e., when the encoder's time-i mapping is of the form

$$f_i\colon \mathcal{M}\times\mathcal{T}^i\times\mathcal{S}^{i-1}\to\mathcal{X},\quad (m,t^i,s^{i-1})\mapsto x_i. \tag{11}$$

Proof: In the direct part, we shall prove that (10) is achievable also in the absence of perfect past state information; in the converse part, we shall prove that one cannot achieve a higher rate even in its presence. That $P_{T|S}$ and $P_{X|UT}$ can be chosen to be deterministic follows because, for any fixed P_U, the mutual information $I(U;Y)$ is convex in $P_{Y|U}$, and because $P_{Y|U}$ is linear in both $P_{T|S}$ and $P_{X|UT}$. Once $P_{X|UT}$ is chosen to be deterministic, $\mathcal{U}$ can be restricted to comprise

[1] We use "help," "assistance," and "description" interchangeably.

the mappings from $\mathcal{T}$ to $\mathcal{X}$, hence its cardinality need not exceed $|\mathcal{X}|^{|\mathcal{T}|}$.

Direct part. Choose $P_{T|S}$ to be a deterministic mapping that achieves the maximum in (10), so $T = h(S)$ with probability one for some $h(\cdot)$. The helper produces $t_i = h(s_i)$ for every i. This implies that the sequence T^n is IID. The encoder and the decoder treat T as the channel state governing the SD-DMC

$$\tilde{W}(y|x,t) = \sum_{s\in\mathcal{S}} P_{S|T}(s|t)\, W(y|x,s) \tag{12}$$

and can thus achieve (10) using Shannon strategies [1]; cf. (9).

Converse part. Given mappings $\{f_i\}$ as in (11), we can use Fano's inequality to infer the existence of some sequence $\{\epsilon_n\}$ tending to zero for which

$$\begin{aligned} n(R-\epsilon_n) &\le I(M;Y^n) && (13)\\ &= \sum_{i=1}^{n} I(M;Y_i|Y^{i-1}) && (14)\\ &\le \sum_{i=1}^{n} I(M,Y^{i-1};Y_i) && (15)\\ &\le \sum_{i=1}^{n} I(M,Y^{i-1},S^{i-1};Y_i) && (16)\\ &= \sum_{i=1}^{n} I(U_i;Y_i), && (17)\end{aligned}$$

where in the last equality we defined, for every $i \in \{1,\ldots,n\}$,

$$U_i \triangleq (M, Y^{i-1}, S^{i-1}). \tag{18}$$

It remains to check that, for every i, the joint distribution of $(S_i, T_i, U_i, X_i, Y_i)$ has the form (10b). To this end, it suffices to verify that the following three conditions are satisfied:

$$(U_i, T_i) \multimap (X_i, S_i) \multimap Y_i \tag{19a}$$

$$S_i \multimap (U_i, T_i) \multimap X_i \tag{19b}$$

$$U_i \perp\!\!\!\perp (S_i, T_i). \tag{19c}$$

Indeed, (19a) is satisfied because the channel is memoryless; (19b) because, given $U_i = (M, Y^{i-1}, S^{i-1})$, one can compute T^{i-1} (as a function of S^{i-1}), and X_i is determined by (M, T^i, S^{i-1}); and (19c) because the state sequence is IID, and because, with a symbol-by-symbol helper, T_i depends on S_i alone and not on S^{i-1}. This completes the proof of the converse part and hence of the theorem. ∎

The rate (10) is a lower bound to the causal-helper capacity, because every symbol-by-symbol helper is causal. This bound, however, is not tight:

Remark 2: For some channels, there exist causal helpers that outperform all symbol-by-symbol helpers; see Example 9 ahead.

This bound does, however, characterize the SD-DMCs having positive causal-helper capacity:

Theorem 3: Subject to (1), the following statements are equivalent:

1) The causal-helper capacity is positive.
2) The symbol-by-symbol helper capacity is positive.
3) The capacity of the SD-DMC when the state is revealed perfectly to both encoder and decoder is positive. Equivalently, there exists some state $s^\star \in \mathcal{S}$ with $P_S(s^\star) > 0$ and some $x_1, x_2 \in \mathcal{X}$ such that

$$W(\cdot|x_1, s^\star) \neq W(\cdot|x_2, s^\star), \tag{20}$$

indicating that the output PMFs induced by $(x_1, s^\star)$ and by $(x_2, s^\star)$ differ.

Proof: We first verify the equivalence of the two conditions in 3). The capacity of an SD-DMC with perfect state information at both encoder and decoder is given by (see [10] and references therein)

$$\max_{P_{X|S}} I(X;Y|S), \tag{21a}$$

where the conditional mutual information is computed with respect to the joint PMF

$$P_S(s)\, P_{X|S}(x|s)\, W(y|x,s). \tag{21b}$$

This capacity is positive if, and only if, there exists some $s^\star \in \mathcal{S}$ such that $P_S(s^\star) > 0$ and

$$\max_{P_{X|S=s^\star}} I(X;Y|S=s^\star) > 0. \tag{22}$$

Inequality (22) holds if, and only if, there exist $x_1, x_2 \in \mathcal{X}$ that satisfy (20).

We next show the equivalence of 1), 2), and 3). Of the three capacities, the symbol-by-symbol helper capacity is smallest, and the perfect-encoder-and-decoder-state-information capacity is largest. It thus suffices to prove, as we proceed to do, that if the latter is positive, then so is the former.

Assume that 3) holds. By Assumption (1), $\mathcal{T}$ has at least two distinct elements. Call them 0 and 1. Now consider the time-invariant symbol-by-symbol mapping, $h_i(\cdot) = h(\cdot)$ where

$$h(s) = \begin{cases} 0 & \text{if } s = s^\star \\ 1 & \text{if } s \in \mathcal{S} \setminus \{s^\star\}. \end{cases} \tag{23}$$

The helper thus only tells the encoder whether or not the current state is $s^\star$. Let P_U be the uniform distribution over the two mappings u_1, u_2 from $\mathcal{T}$ to $\mathcal{X}$, where

$$u_1(0) = x_1 \tag{24a}$$

$$u_2(0) = x_2 \tag{24b}$$

$$u_1(1) = u_2(1) = x_0, \tag{24c}$$

where x_0 can be chosen to be any element of $\mathcal{X}$ (not necessarily different from x_1 or x_2). More formally, we choose $\mathcal{U} = \{1, 2\}$; the PMF P_U to be uniform over $\mathcal{U}$; and we choose $P_{X|UT}$ to be deterministic, so $x = x(u,t)$, where

$$x(u=1,t) = \begin{cases} x_1 & \text{if } t = 0 \\ x_0 & \text{if } t = 1 \end{cases} \tag{25}$$

and

$$x(u=2,t) = \begin{cases} x_2 & \text{if } t = 0 \\ x_0 & \text{if } t = 1. \end{cases} \tag{26}$$

This choice of P_U and $P_{X|UT}$ implies that, or every $y \in \mathcal{Y}$,

$$P_{Y|U}(y|u_1) = P_S(s^\star)\, W(y|x_1, s^\star) + \sum_{s \neq s^\star} P_S(s)\, W(y|x_0, s) \tag{27}$$

$$P_{Y|U}(y|u_2) = P_S(s^\star)\, W(y|x_2, s^\star) + \sum_{s \neq s^\star} P_S(s)\, W(y|x_0, s). \tag{28}$$

By (20),

$$P_{Y|U}(\cdot|u_1) \neq P_{Y|U}(\cdot|u_2), \tag{29}$$

which implies that

$$I(U;Y) > 0. \tag{30}$$

It then follows from Theorem 1 that the capacity of this channel with a symbol-by-symbol helper is positive. ■

B. Helper cognizant of message

When a causal helper is cognizant of the message that the encoder wishes to send, its time-i help is characterized by a mapping of the form

$$h_i \colon \mathcal{M} \times \mathcal{S}^i \to \mathcal{T}, \quad (m, s^i) \mapsto t_i. \tag{31}$$

Said helper is a *symbol-by-symbol message-cognizant helper* if this function has the form

$$h_i \colon \mathcal{M} \times \mathcal{S} \to \mathcal{T}, \quad (m, s_i) \mapsto t_i. \tag{32}$$

In both cases the encoder is as before, i.e., characterized by mappings of the form (5). For this setting, symbol-by-symbol message-cognizant helpers achieve capacity:

Theorem 4: The capacity of an SD-DMC with a message-cognizant causal helper is achieved by a message-cognizant symbol-by-symbol helper and is given by

$$\max I(U;Y), \tag{33a}$$

where the maximum is over the choice of a set $\mathcal{U}$ and over the joint distributions of the form

$$P_S(s)\, P_U(u)\, P_{T|US}(t|u,s)\, P_{X|UT}(x|u,t)\, W(y|x,s), \tag{33b}$$

where, without loss of optimality, $P_{T|US}$ and $P_{X|UT}$ can be chosen to be deterministic.

Proof: That $P_{T|US}$ and $P_{X|UT}$ can be chosen to be deterministic follows because, for any fixed P_U, $I(U;Y)$ is convex in $P_{Y|U}$, and because $P_{Y|U}$ is linear in both $P_{T|US}$ and $P_{X|UT}$. (See (34) ahead.)

Direct part. Fix a PMF P_U, a (deterministic) mapping $h \colon \mathcal{U} \times \mathcal{S} \to \mathcal{T}$, and a (deterministic) mapping $f \colon \mathcal{U} \times \mathcal{T} \to \mathcal{X}$ that achieve the maximum in (33). Consider the "super channel" with input alphabet $\mathcal{U}$, output alphabet $\mathcal{Y}$, and of the conditional law

$$\tilde{W}(y|u) = \sum_{\substack{s \in \mathcal{S} \\ t \in \mathcal{T}}} P_S(s)\, P_{T|US}(t|u,s)\, P_{X|UT}(x|u,t)\, W(y|x,s) \tag{34}$$

induced by (33b).

We will show that, given any codebook $\{u^n(m)\}_{m \in \mathcal{M}}$ for this super channel, there exists a scheme with a message-dependent symbol-by-symbol helper that achieves the same error probability on the original channel. To this end, suppose that $m \in \mathcal{M}$ is the message to be transmitted, and $u^n(m)$ is the corresponding codeword for the super channel. Since the helper knows the message m, it also knows $u^n(m)$. In the proposed scheme for the original channel, the helper produces

$$t_i = h\big(u_i(m), s_i\big), \qquad i = 1, \ldots, n. \tag{35}$$

The encoder—that knows $u^n(m)$ (because it is cognizant of m) and that obtains t_i from the helper—sends

$$x_i = f\big(u_i(m), t_i\big). \tag{36}$$

The conditional distribution $P_{Y^n|M}(y^n|m)$ of Y^n given m that this scheme induces is

$$P_{Y^n|M}(y^n|m) = \prod_{i=1}^{n} \tilde{W}\big(y_i \big| u_i(m)\big), \tag{37}$$

so the probability of error of this scheme is identical to that of the code on the super channel. The proposed scheme can thus achieve the capacity of the super channel, which equals (33).

Converse part. Fix mappings $\{h_i\}$ as in (31), and define

$$U_i \triangleq (M, T^{i-1}, Y^{i-1}), \qquad i = 1, \ldots, n. \tag{38}$$

From Fano's inequality we infer that, if a uniformly drawn message M is transmitted using a helping scheme with vanishing probability of error, then, for some ϵ_n that tends to zero as $n \to \infty$,

$$n(R - \epsilon_n) \leq I(M; Y^n) \tag{39}$$

$$= \sum_{i=1}^{n} I(M; Y_i | Y^{i-1}) \tag{40}$$

$$\leq \sum_{i=1}^{n} I(M, Y^{i-1}; Y_i) \tag{41}$$

$$\leq \sum_{i=1}^{n} I(U_i; Y_i). \tag{42}$$

It remains to check that the joint distribution of $(S_i, T_i, U_i, X_i, Y_i)$ has the form (33b), i.e., that the following Markov and independence conditions are satisfied:

$$(U_i, T_i) \multimap (X_i, S_i) \multimap Y_i \tag{43a}$$

$$S_i \multimap (U_i, T_i) \multimap X_i \tag{43b}$$

$$U_i \perp\!\!\!\perp S_i. \tag{43c}$$

Here, (43a) is satisfied because the channel is an SD-DMC; (43b) because X_i can be determined from (M, T^i) (and hence from (U_i, T_i)); and (43c) because the state sequence is memoryless. ■

Note that the difference in (10) from (33) is that $P_{T|US}$ replaces $P_{T|S}$. This can also be seen in the difference between (19c) and (43c).

Having the helper be cognizant of the transmitted message is advantageous. The best message-cognizant causal helper can outperform all message-oblivious causal helpers:

Remark 5: The capacity of an SD-DMC with a causal message-cognizant helper can exceed that with the best causal message-oblivious helper; see Example 7.

C. Channel state known to decoder

Theorem 6: When the decoder has perfect state information, i.e., when it is of the form (8), the causal-helper capacity is given by

$$\max I(X;Y|S), \tag{44a}$$

where the maximum is over all the joint distributions of the form

$$P_S(s)\, P_{T|S}(t|s)\, P_{X|T}(x|t)\, W(y|x,s). \tag{44b}$$

Moreover, said capacity can be achieved with a symbol-by-symbol helper.

Proof: Direct part. We use a symbol-by-symbol helper and apply Theorem 1. Since the decoder is cognizant of the state, the state can be viewed as part of the channel output, so we can replace $\max I(U;Y)$ in Theorem 1 with $\max I(U;Y,S)$, which can be simplified as follows:

$$\begin{aligned}\max I(U;Y,S) &= \max I(U;Y|S) &(45)\\ &= \max I(X;Y|S), &(46)\end{aligned}$$

where (45) holds because U is independent of S in (10b); and (46) holds because, when both $P_{T|S}$ and $P_{X|UT}$ are chosen to be deterministic, X is a function of U and S, and because $U \multimap (X,S) \multimap Y$ forms a Markov chain. Integrating U out reduces the joint PMF (10b) to (44b).

Converse part. Fix mappings $\{h_i\}$ as in (4), and define for every $i \in \{1,\ldots,n\}$

$$\begin{aligned}U_i &\triangleq (M, Y^{i-1}) &(47)\\ V_i &\triangleq S^{i-1}. &(48)\end{aligned}$$

From Fano's inequality we infer that, if the decoder is cognizant of the state, and if a uniformly drawn message M is transmitted using a helping scheme with vanishing probability of error, then, for some ϵ_n that tends to zero as $n \to \infty$,

$$\begin{aligned}n(R-\epsilon_n) &\le I(M;Y^n,S^n) &(49)\\ &= \sum_{i=1}^{n} I(M;Y_i,S_i|Y^{i-1},S^{i-1}) &(50)\\ &\le \sum_{i=1}^{n} I(M,Y^{i-1};Y_i,S_i|S^{i-1}) &(51)\\ &= \sum_{i=1}^{n} I(M,Y^{i-1};Y_i|S_i,S^{i-1}) &(52)\\ &= \sum_{i=1}^{n} I(U_i;Y_i|S_i,V_i). &(53)\end{aligned}$$

The capacity is thus upper-bounded by the maximum of

$$I(U;Y|S,V) \tag{54}$$

over the auxilliary sets $\mathcal{U}$ and $\mathcal{V}$, and over the joint distributions of the form

$$\begin{aligned}&P_V(v)\, P_S(s)\, P_{U|V}(u|v)\, P_{T|SV}(t|s,v)\\ &\qquad \cdot P_{X|TUV}(x|t,u,v)\, W(y|x,s). &(55)\end{aligned}$$

We further upper-bound (54) by its maximum over $V=v$:

$$I(U;Y|S,V) \le \max_{v\in\mathcal{V}} I(U;Y|S,V=v). \tag{56}$$

Since $V \perp\!\!\!\perp S$, we can remove V by fixing $V=v$ that achieves the above maximum, so the upper bound becomes $\max I(U;Y|S)$ over the joint distribution (10b). As in the direct part, $I(U;Y|S) = I(X;Y|S)$, which completes the proof. ■

III. Counterexamples

A. Revealing the message to the helper increases capacity

Example 7: The channel input X is a binary tuple

$$X = (A,B) \tag{57}$$

with A,B both taking values in $\{0,1\}$. The state S too is a binary tuple

$$S = (S^{(0)}, S^{(1)}), \tag{58}$$

where $S^{(0)}$ and $S^{(1)}$ are independent and both uniform over $\{0,1\}$. The channel output is

$$Y = (A, B \oplus S^{(A)}). \tag{59}$$

The helper's description rate is 1 bit:

$$\mathcal{T} = \{0,1\}, \tag{60}$$

so it cannot fully describe the state.

Claim 8: For the SD-DMC of Example 7:

1) The capacity with a message-cognizant causal helper is 2 bits per channel use.
2) The capacity with a message-oblivious causal helper is $\log 3$, which can be achieved with a symbol-by-symbol helper. Furthermore, $\log 3$ is the capacity also when the message-oblivious helper is noncausal and the help T is provided also to the decoder.

Proof: Message-cognizant helper. The capacity cannot exceed $\log|\mathcal{X}| = \log|\mathcal{Y}| = 2$ bits, so we focus on achievability and describe a scheme that can convey 2 bits error-free in a single channel use. Let α and β denote the information bits to be conveyed. Since the helper is cognizant not only of the state but also of (α,β), it can assist the encoder by providing it with

$$T = S^{(\alpha)}. \tag{61}$$

The encoder can then produce the channel input

$$X = (\alpha, \beta \oplus T) = (\alpha, \beta \oplus S^{(\alpha)}), \tag{62}$$

where $\oplus$ denotes modular-2 addition. The output is then

$$Y = (\alpha, \beta \oplus S^{(\alpha)} \oplus S^{(\alpha)}) = (\alpha,\beta) \tag{63}$$

and both bits are correctly conveyed without the need for decoding. (The achievability of 2 bits could also be deduced from Theorem 4 by choosing $U = (A,V)$ uniform on $\{0,1\}\times\{0,1\}$, $T = S^{(A)}$, and $X = (A, V \oplus T)$.)

We next turn to the message-oblivious helper. But first we recall a result on the "sum channel" [11, Problem 4.18], [12, Problem 7.28].

Sum channel. Consider a discrete memoryless channel that is the "sum" of ℓ disjoint sub-channels in the following sense:

$$\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_\ell, \qquad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, i \neq j \tag{64}$$

$$\mathcal{Y} = \mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_\ell, \qquad \mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, i \neq j \tag{65}$$

$$\Pr(Y \in \mathcal{Y}_i | X = x) = 1, \quad x \in \mathcal{X}_i. \tag{66}$$

Let C_i denote the capacity of the i-th sub-channel, i.e., the channel with input alphabet $\mathcal{X}_i$ and output alphabet $\mathcal{Y}_i$. Then the capacity of the sum channel is

$$C = \log \sum_{i=1}^{\ell} 2^{C_i}. \tag{67}$$

Message-oblivious helper, direct part. Let the (symbol-by-symbol) helper produce

$$T = S^{(0)}. \tag{68}$$

Since the first component of the output Y is equal to the first component of the input X (see (63)), we can view the channel as the sum of two channels: the first where $A = 0$ so $\mathcal{X}_1 = \mathcal{Y}_1 = \{(0,0),(0,1)\}$ and the second where $A = 1$ so $\mathcal{X}_2 = \mathcal{Y}_2 = \{(1,0),(1,1)\}$.

The encoders of both channels observe T, but the second ignores it. The first encoder, cognizant of $S^{(A)} = S^{(0)}$, can perfectly control the second output bit $B \oplus S^{(0)}$, so $C_1 = 1$ bit. The encoder for the second sub-channel, where $A = 1$, is incognizant of $S^{(A)} = S^{(1)}$, so the sub-channel's output bit $B \oplus S^{(1)}$ is random and independent of the input, so $C_2 = 0$. The capacity of the sum channel is thus

$$\log\bigl(2^1 + 2^0\bigr) = \log 3. \tag{69}$$

Message-oblivious helper, converse part. We first consider the special case where the helper is symbol-by-symbol. This step is unnecessary but sheds light on the proof of the general case. Assuming that the help is provided also to the decoder and treating $T = h(S)$ as the channel meta state, we can express the capacity as

$$\max I(X;Y|T), \tag{70}$$

where the maximization is over the conditional law $P_{X|T}$. When we fix any $T = t$, the channel becomes a sum channel. The first sub-channel is where $A = 0$. With $A = 0$ fixed, the maximum value of $H(Y|T = t)$ is 1 bit, which is achieved when B is uniform, whereas

$$H(Y|X, T = t) = H(B \oplus S^{(0)}|B, T = t) = H(S^{(0)}|T = t). \tag{71}$$

We conclude that, conditional on $T = t$, the capacity of the first sub-channel is

$$C_1 = 1 - H(S^{(0)}|T = t). \tag{72}$$

Similarly, under the same conditioning, the capacity of the second sub-channel is

$$C_2 = 1 - H(S^{(1)}|T = t). \tag{73}$$

Conditional on $T = t$, the capacity of the sum channel is

$$\max I(X;Y|T = t) = \log\Bigl(2^{1-H(S^{(0)}|T=t)} + 2^{1-H(S^{(1)}|T=t)}\Bigr). \tag{74}$$

This and the inequality

$$2^{1-a} \leq 2 - a, \qquad a \in [0,1], \tag{75}$$

imply that

$$\max I(X;Y|T = t) \leq \log\Bigl(2 - H(S^{(0)}|T = t) + 2 - H(S^{(1)}|T = t)\Bigr) \tag{76}$$

$$\leq \log\Bigl(4 - H(S^{(0)}, S^{(1)}|T = t)\Bigr). \tag{77}$$

Averaging over t and employing Jensen's inequality, we obtain an upper bound on the capacity with a message-oblivious symbol-by-symbol helper when the help is provided also to the decoder:

$$C \leq \max \sum_t P_T(t) \log\Bigl(4 - H(S^{(0)}, S^{(1)}|T = t)\Bigr) \tag{78}$$

$$\leq \max \log\Bigl(4 - H(S^{(0)}, S^{(1)}|T)\Bigr) \tag{79}$$

$$\leq \log(4 - 1) \tag{80}$$

$$= \log 3, \tag{81}$$

where (79) follows because $\log$ is concave; and (80) because

$$H(S^{(0)}, S^{(1)}|T) = \underbrace{H(S^{(0)}, S^{(1)})}_{=2} - \underbrace{I(S^{(0)}, S^{(1)}; T)}_{\leq H(T) \leq 1} \tag{82}$$

$$\geq 1. \tag{83}$$

We next show that one cannot achieve any rate larger than $\log 3$ even with a noncausal helper, and when the help is provided also to the decoder. In the rest of this proof, we shall slightly abuse notation to use T to denote the n-letter assistance, i.e., T is a function of S^n and takes values in $\{0,1\}^n$. Using Fano's inequality we can infer that, if there exists a coding scheme of rate R that has vanishing error probability, then, for some ϵ_n that tends to zero as $n \to \infty$,

$$n(R - \epsilon_n) \leq I(X^n; Y^n|T), \tag{84}$$

so we shall bound the maximum of $I(X^n; Y^n|T)$ over all $P_{X^n|T}$. Given $T = t$, the n-letter channel can be viewed as a sum channel containing 2^n sub-channels, each corresponding to a different choice of the binary n-tuple $(A_1, \ldots, A_n)$. Conditional on $T = t$, the capacity of the sub-channel corresponding to $A_1 = a_1, \ldots, A_n = a_n$ is

$$\exp_2\Bigl\{n - H\Bigl(S_1^{(a_1)}, \ldots, S_n^{(a_n)}\Bigr|T = t\Bigr)\Bigr\}, \tag{85}$$

where $\exp_2\{\xi\}$ denotes 2^ξ. Hence,

$$\max_{P_{X^n|T=t}} I(X^n; Y^n|T = t) = \log \sum_{a^n \in \{0,1\}^n} \exp_2\Bigl\{n - H\Bigl(S_1^{(a_1)}, \ldots, S_n^{(a_n)}\Bigr|T = t\Bigr)\Bigr\} \tag{86}$$

$$= \log \sum_{a^n \in \{0,1\}^n} \prod_{i=1}^{n} \exp_2 \Bigl\{1 - H\Bigl(S_i^{(a_i)}\Bigr|S_1^{(a_1)}, \ldots, S_{i-1}^{(a_{i-1})}, T = t\Bigr)\Bigr\} \tag{87}$$

$$\leq \log \sum_{a^n \in \{0,1\}^n} \prod_{i=1}^{n} \exp_2\Bigl\{1 - H\Bigl(S_i^{(a_i)}\Bigr|S^{i-1}, T = t)\Bigr)\Bigr\} \tag{88}$$

$$= \log \prod_{i=1}^{n} \Bigg(\exp_2\Big\{1 - H\Big(S_i^{(0)}\Big|S^{i-1}, T=t\Big)\Big\} + \exp_2\Big\{1 - H\Big(S_i^{(1)}\Big|S^{i-1}, T=t\Big)\Big\}\Bigg) \tag{89}$$

$$= \sum_{i=1}^{n} \log \Bigg(\exp_2\Big\{1 - H\Big(S_i^{(0)}\Big|S^{i-1}, T=t\Big)\Big\} + \exp_2\Big\{1 - H\Big(S_i^{(1)}\Big|S^{i-1}, T=t\Big)\Big\}\Bigg) \tag{90}$$

$$\leq \sum_{i=1}^{n} \log \Bigg(2 - H\Big(S_i^{(0)}\Big|S^{i-1}, T=t\Big) + 2 - H\Big(S_i^{(1)}\Big|S^{i-1}, T=t\Big)\Bigg) \tag{91}$$

$$\leq \sum_{i=1}^{n} \log \Big(4 - H\big(S_i\big|S^{i-1}, T=t\big)\Big) \tag{92}$$

$$\leq n \log\Bigg(4 - \frac{1}{n}\sum_{i=1}^{n} H\big(S_i\big|S^{i-1}, T=t\big)\Bigg) \tag{93}$$

$$= n \log\Bigg(4 - \frac{H(S^n|T=t)}{n}\Bigg), \tag{94}$$

where (91) follows from (75), and (93) follows from the concavity of the logarithm. Averaging (94) over T and again using the concavity of the logarithm, we obtain the bound

$$I(X^n; Y^n|T) \leq n \log\Bigg(4 - \frac{H(S^n|T)}{n}\Bigg). \tag{95}$$

Note that

$$H(S^n|T) = H(S^n) - I(S^n; T) \tag{96}$$
$$\geq H(S^n) - H(T) \tag{97}$$
$$\geq 2n - n \tag{98}$$
$$= n. \tag{99}$$

This and (95) imply that

$$I(X^n; Y^n|T) \leq n \log 3, \tag{100}$$

which completes the proof. ■

B. A causal helper that outperforms all symbol-by-symbol helpers

Example 9: The channel input is

$$X = (A, B, C), \tag{101}$$

where A, B take values in $\{0,1\}$, and C in $\{0,1\}^\eta$ for some integer η (larger than 10). The state is a pair

$$S = (S^{(0)}, S^{(1)}), \tag{102}$$

where $S^{(0)}$ and $S^{(1)}$ are independent, both uniform on $\{0,1\}$. The output is

$$Y = (A', D^{(0)}, D^{(1)}), \tag{103}$$

with A' taking values in $\{0,1\}$, and with $D^{(0)}$ and $D^{(1)}$ in $\{0,1\}^\eta$.

The channel law is the following:

- Conditional on X and S with $B \neq S^{(A)}$, the output Y is uniformly distributed over its alphabet.
- Conditional on X and S with $B = S^{(A)}$,

$$A' = A \tag{104a}$$
$$D^{(B)} = C \tag{104b}$$

deterministically, and

$$D^{(B\oplus 1)} \sim \text{equiprobable over } \{0,1\}^\eta. \tag{104c}$$

The (message-oblivious) helper's description rate is 1 bit:

$$\mathcal{T} = \{0,1\}. \tag{105}$$

Claim 10: In Example 9:

1) There exists a (non symbol-by-symbol) causal helper that allows for the reliable transmission of η bits per channel use.
2) When restricted to symbol-by-symbol helpers, the capacity is strictly less than η.

Proof: General causal helper. To prove the achievability of η bits per channel use with a causal (non symbol-by-symbol) helper, consider the following coding scheme. Represent the message that is to be transmitted in n channel uses as a sequence $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ of binary η-tuples, so $\alpha_i \in \{0,1\}^\eta$. Set α_n to be some arbitrary η-tuple, say all-zero. The transmission rate is thus $(n-1)\eta/n$, which approaches η as $n \to \infty$.

Define

$$T_0 \triangleq 0. \tag{106}$$

At each time i, the help is

$$T_i = S_i^{(T_{i-1})}, \tag{107}$$

and the channel input is

$$A_i = T_{i-1} \tag{108}$$
$$B_i = T_i \tag{109}$$
$$C_i = \alpha_i. \tag{110}$$

This guarantees that, at each i,

$$B_i = S^{(A_i)}, \tag{111}$$

so the channel behaves according to (104). Moreover, (108) and (104a) imply that from Y_{i+1}—specifically from A'_{i+1}—the decoder will learn B_i because $B_i = A_{i+1} = A'_{i+1}$. It will then be able to recover α_i without error by reading $D_i^{(B_i)}$ (which was received at time i). In this way, the decoder recovers $\alpha_1, \ldots, \alpha_{n-1}$ error free. This concludes the proof of the first part of the claim.

Symbol-by-symbol helper. There are ostensibly 2^4 symbol-by-symbol helpers. But T and $T \oplus 1$ give identical performance. Likewise swapping $S^{(0)}$ and $S^{(1)}$ or replacing either (or both of them) with the complement does not change performance. After accounting for these symmetries, we must only analyze three symbol-by-symbol helpers:

$$T = S^{(0)} \tag{112}$$
$$T = S^{(0)} \wedge S^{(1)} \tag{113}$$
$$T = S^{(0)} \oplus S^{(1)}. \tag{114}$$

To analyze $T = S^{(0)} \oplus S^{(1)}$, define $E = E(X, S)$ as

$$E = \begin{cases} 1, & \text{if } B = S^{(A)} \\ 0, & \text{otherwise.} \end{cases} \tag{115}$$

Recall that, conditional on (X, S) with $E(X, S)$ being zero, the output Y is equiprobable over its alphabet. We can upper-bound the capacity by assuming that T is revealed also to the decoder and then upper-bounding $\max I(X; Y|T)$. This we do as follows:

$$\begin{aligned} I(X;Y|T) &\le I(X;Y,E|T) && (116)\\ &\le 1 + I(X;Y|E,T) && (117)\\ &= 1 + \sum_t P_T(t)\, I(X;Y|E,T=t). && (118) \end{aligned}$$

For each $t \in \{0, 1\}$

$$\begin{aligned} I(X;Y|E,T=t) &= P_{E|T}(1|t)\, I(X;Y|E=1,T=t) && (119)\\ &\le P_{E|T}(1|t)\, \log|\mathcal{X}| && (120)\\ &= \frac{1}{2}(\eta+2), && (121) \end{aligned}$$

where the first equality holds because $I(X; Y|E = 0, T = t)$ is zero; and the last equality because, irrespectively of $P_{X|T}$, the probability of $E(X, S)$ being one is always $1/2$. The capacity with this symbol-by-symbol helper is thus upper-bounded by $2+\eta/2$, which is strictly smaller than η whenever η exceeds 4.

Consider now $T = S^{(0)} \wedge S^{(1)}$. Upper-bounding the mutual information conditional on $T = 1$ by $\log|\mathcal{X}|$,

$$\begin{aligned} I(X;Y|T) &\le P_T(1)\,(\eta+2) + P_T(0)\, I(X;Y|T=0) && (122)\\ &= \frac{1}{4}(\eta+2) + \frac{3}{4} I(X;Y|T=0). && (123) \end{aligned}$$

The mutual information term can then be bounded by (with E defined as in (115))

$$\begin{aligned} I(X;Y|T=0) &\le I(X;Y,E|T=0) && (124)\\ &\le 1 + I(X;Y|E,T=0) && (125)\\ &= 1 + P_{E|T}(1|0)\,(\eta+2) && (126)\\ &\le 1 + \frac{2}{3}(\eta+2), && (127) \end{aligned}$$

where the last inequality holds because, conditional on $S^{(0)} \wedge S^{(1)} = 0$, the states $(0, 0), (0, 1), (1, 0)$ are each of probability $1/3$, so $P_{E|T}(1|0)$ cannot exceed $2/3$. The above inequalities demonstrate that the capacity with this helper is upper-bounded by $9/4 + 3\eta/4$. This is strictly smaller than η whenever $\eta \ge 10$.

The final symbol-by-symbol helper $T = S^{(0)}$ can be analyzed using the duality-based upper bound [13, Thm. 5.1]; see the appendix. ∎

C. Help to both encoder and decoder

The following example shows that symbol-by-symbol helpers need not be optimal even when the help is provided to both encoder and decoder.

Example 11: The channel input is binary, and the state is quaternary

$$\mathcal{X} = \{0, 1\} \tag{128}$$

$$\mathcal{S} = \{0, 1, 2, 3\}. \tag{129}$$

The output is a binary 4-tuple

$$Y = (Y^{(0)}, Y^{(1)}, Y^{(2)}, Y^{(3)}) \in \{0, 1\}^4. \tag{130}$$

Conditional on the input $X = x$ and the state $S = s$, the component of Y that is indexed by s equals x deterministically

$$Y^{(s)} = x \tag{131a}$$

and the other components are IID Bernoulli $(1/2)$

$$\{Y^{(s')}\}_{s' \in \mathcal{S}\setminus\{s\}} \sim \text{IID Bern}(1/2). \tag{131b}$$

The helper's description rate is 1 bit:

$$\mathcal{T} = \{0, 1\}. \tag{132}$$

Claim 12: If the help in Example 11 is provided to the encoder causally and also to the decoder, then:

1) The highest capacity achievable with a symbol-by-symbol helper is 0.5 bit.
2) With some causal non symbol-by-symbol helper the capacity is at least $0.5 + 0.1875 \log 1.5 \approx 0.61$ bit.

Proof: Symbol-by-symbol helper. After accounting for symmetries and relabelings, only two different symbol-by-symbol helpers remain. The first is

$$T = \begin{cases} 0, & \text{if } S \in \{0, 1\} \\ 1, & \text{otherwise.} \end{cases} \tag{133}$$

To analyze it, suppose that both encoder and decoder know that $T = 0$. The output components $Y^{(2)}$ and $Y^{(3)}$ can then be discarded because they are known to be independent of the input. Conditional on $X = x$, the remaining components have the following distribution:

$$(Y^{(0)}, Y^{(1)}) = \begin{cases} (x, x) & \text{w.p. } 0.5 \\ (x, x \oplus 1) & \text{w.p. } 0.25 \\ (x \oplus 1, x) & \text{w.p. } 0.25. \end{cases} \tag{134}$$

It then follows that

$$H\Big(Y^{(0)}, Y^{(1)}\Big|X, T=0\Big) = 1.5. \tag{135}$$

Since

$$\max H\Big(Y^{(0)}, Y^{(1)}\Big|T=0\Big) = 2, \tag{136}$$

with the maximum achieved by a uniform X,

$$\max I(X;Y|T=0) = 0.5. \tag{137}$$

Both (135) and (136) continue to hold when, rather than $T = 0$, we consider $T = 1$. We thus conclude that the maximum achievable rate with the helper (133) is 0.5 bit.

The second symbol-by-symbol helper to be considered is

$$T = \begin{cases} 0, & S = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{138}$$

When $T = 0$, which happens with probability 0.25, the decoder knows that $Y^{(0)} = X$, so

$$\max I\big(X; Y^{(0)} \,\big|\, T = 0\big) = 1. \tag{139}$$

When $T = 1$, which happens with probability 0.75, $Y^{(0)}$ is independent of the input and can be discarded, whereas the conditional probabilities of the remaining components given $X = x$ can be written down explicitly as in (134) (details omitted). From these probabilities we can compute

$$\max I\Big(X; Y^{(1)}, Y^{(2)}, Y^{(3)}\Big|T = 1\Big) = 1.5 - 0.75 \log 3. \tag{140}$$

Using (139) and (140), we obtain that the maximum rate achievable by the helper (138) is

$$0.25 \cdot 1 + 0.75 \cdot (1.5 - 0.75 \log 3) \approx 0.483, \tag{141}$$

which is inferior to the capacity with the first helper. The capacity with the best symbol-by-symbol helper is thus 0.5 bit and is achieved by the helper (133).

General helper. Consider a two-letter helper: over two channel uses, the helper uses T_1, T_2 to describe S_1 perfectly and S_2 not at all. (More formally, the time-$(2i+1)$ help T_{2i+1} and the time-$(2i+2)$ help T_{2i+2} describe the time-$(2i+1)$ state S_{2i+1} ignoring the time-$(2i+2)$ state S_{2i+2}.) At time $2i+2$ the receiver is cognizant of S_{2i+1} and can hence recover X_{2i+1} (which equals the S_{2i+1}-th component of Y_{2i+1}). The bit X_{2i+1} is thus recovered error free. As for X_{2i+2}, it is transmitted with neither encoder nor decoder cognizant of the state, hence we can write the probabilities of the different values of Y_{2i+2} as in (134) (with the details again omitted). The channel from X_{2i+2} to Y_{2i+2} is of capacity

$$0.375 \log 1.5. \tag{142}$$

We can thus transmit $(1 + 0.375 \log 1.5)$ bits with two channel uses, from which the second part of the claim follows. ∎

IV. A Block-Markov Scheme

Inspired by Example 9, we propose a communication scheme employing block-Markov encoding and backward decoding. Choose three finite auxiliary sets $\mathcal{Z}, \mathcal{U}$, and $\mathcal{V}$, and fix a joint distribution of the form

$$P_S(s)\, P_Z(z)\, P_{T|SZ}(t|s,z)\, P_{U|Z}(u|z) \cdot P_{X|UT}(x|u,t) P_{V|T}(v|t)\, W(y|x,s), \tag{143}$$

where $P_{T|SZ}$ and $P_{X|UT}$ are deterministic, so there exist mappings $f\colon \mathcal{U} \times \mathcal{T} \to \mathcal{X}$ and $h\colon \mathcal{S} \times \mathcal{Z} \to \mathcal{T}$ such that, with probability one,

$$T = h(S, Z) \tag{144}$$

$$X = f(U, T). \tag{145}$$

In the following, we use boldface letters such as $\mathbf{x}$ and $\mathbf{t}$ to denote length-n vectors, and we extend h and f to apply to n-length vectors component-wise. For example, $\mathbf{t} = h(\mathbf{s}, \mathbf{z})$ indicates that each component of $\mathbf{t}$ is the result of applying h to the corresponding components of $\mathbf{s}$ and $\mathbf{z}$. Component-wise extension of conditional probabilities to length-n vectors are denoted using the superscript $\times n$ as in $P_{U|Z}^{\times n}(\mathbf{u}|\mathbf{z})$.

The block-Markov scheme we propose divides the transmission time into λ blocks, with each of the first $\lambda - 1$ blocks being of length n, and with the last being possibly longer. Since λ will be very large, the last block will have negligible effect on the transmission rate. The scheme employs binning and superposition coding. In the superposition coding, the Block j cloud center $\mathbf{z}^{(j)}$—being determined at the end of Block $(j-1)$—is known to both encoder and helper before the block begins. The satellite $\mathbf{u}^{(j)}$ is determined by the message m_j that is transmitted in Block j. The assistance produced in Block j is $\mathbf{t}^{(j)} = h(\mathbf{s}^{(j)}, \mathbf{z}^{(j)})$. A v-sequence, $\mathbf{v}^{(j)}$, is then chosen based on $\mathbf{t}^{(j)}$, and is binned as in Wyner-Ziv coding [14], treating the outputs $\mathbf{y}^{(j)}$ as side information that is available to the decoder. The bin index determines the cloud center in Block $(j+1)$. We elaborate below.

Fix three positive rates R, R_v, and $\tilde{R}$ that will be specified later. In the following, "typical" is short for strongly typical with respect to corresponding marginal of the joint distribution (143) [15]. We shall not explicitly write "ϵ-typical," but it shall be understood that the implicit parameter ϵ does not depend on n and can be chosen arbitrarily close to zero. However, the choice of ϵ may need to be different in each case below, a technicality that, to simplify the exposition, we ignore. Furthermore, the bounds we shall derive on the rates should involve slacks that are related to ϵ. This too we ignore.

Codebook construction. Generate $2^{n(R_v+\tilde{R})}$ length-n sequences IID $\sim P_V$

$$\mathbf{v}(\ell, k), \quad \ell \in \{1, \dots, 2^{nR_v}\}, k \in \{1, \dots, 2^{n\tilde{R}}\}. \tag{146}$$

Independently of the above, generate 2^{nR_v} sequences IID $\sim P_Z$

$$\mathbf{z}(\ell), \quad \ell \in \{1, \dots, 2^{nR_v}\}. \tag{147}$$

For each $\ell \in \{1, \dots, 2^{nR_v}\}$, independently generate 2^{nR} sequences $\sim P_{U|Z}^{\times n}\big(\,\cdot\,\big|\mathbf{z}(\ell)\big)$

$$\mathbf{u}(\ell, m), \quad m \in \{1, \dots, 2^{nR}\}. \tag{148}$$

Encoding. Let $\ell_0 \triangleq 1$. When encoding for Block j, where $j \in \{1, \dots, \lambda - 1\}$, the encoder and the helper have already identified ℓ_{j-1} (which they compute at the end of Block $(j-1)$). They pick the cloud center

$$\mathbf{z}^{(j)} = \mathbf{z}(\ell_{j-1}). \tag{149}$$

The helper produces the help

$$\mathbf{t}^{(j)} = h(\mathbf{s}^{(j)}, \mathbf{z}^{(j)}). \tag{150}$$

Denoting the message to be sent over all the blocks $(m_1, \dots, m_{\lambda-1})$, the encoder selects the satellite according to the message $m_j \in \{1, \dots, 2^{nR}\}$:

$$\mathbf{u}^{(j)} = \mathbf{u}(\ell_{j-1}, m_j). \tag{151}$$

It sends

$$\mathbf{x}^{(j)} = f(\mathbf{u}^{(j)}, \mathbf{t}^{(j)}). \tag{152}$$

(Note that the mappings (150) and (152) are indeed causal.) The helper and the encoder look for the first pair of indices $(\ell_j, k_j) \in \{1, \dots, 2^{nR_v}\} \times \{1, \dots, 2^{n\tilde{R}}\}$ such that

$$\mathbf{v}^{(j)} = \mathbf{v}(\ell_j, k_j) \tag{153}$$

satisfies

$$\Big(\mathbf{v}^{(j)}, \mathbf{t}^{(j)}\Big) \text{ are jointly typical.} \tag{154}$$

They discard k_j and use ℓ_j to pick the cloud center in Block $(j+1)$. If no such pair of indices can be found, then they choose $\ell_j = 1$. We think of ℓ_j as a bin index.

In the last block, Block λ, no message bits are sent, and we only convey the index $\ell_{\lambda-1}$. To this end, we employ a symbol-by-symbol helper of positive capacity. For now, we assume that such a helper exists, and proceed to derive an achievable rate by analyzing Blocks 1 through $(\lambda-1)$. Later we will assert that this assumption is unnecessary, because it is satisfied whenever the rate whose achievability we seek to prove using the block-Markov scheme—namely, (161) ahead—is positive. To prove the assertion, we shall recall that, by Theorem 3, a symbol-by-symbol helper of positive capacity exists whenever the capacity with perfect state information at both encoder and decoder is positive. To prove the assertion, we shall thus only need to show that this latter capacity is positive whenever (161) is positive. This will follow once we show that said capacity is greater than or equal to (161).

Decoding. Based on the output sequence $\mathbf{y}^{(\lambda)}$ received in the last block, Block λ, the decoder recovers $\ell_{\lambda-1}$. It then proceeds with backward decoding from Block $(\lambda-1)$ to Block 1. By the time the decoder decodes Block j, it will have already recovered ℓ_j (from its decoding Block $(j+1)$). Denote the recovered value by $\hat{\ell}_j$. To decode Block j, the decoder looks for indices $\hat{k}_j$, $\hat{\ell}_{j-1}$, and $\hat{m}_j$ such that

$$\Big(\mathbf{z}(\hat{\ell}_{j-1}), \mathbf{u}(\hat{\ell}_{j-1}, \hat{m}_j), \mathbf{v}(\hat{\ell}_j, \hat{k}_j), \mathbf{y}^{(j)}\Big) \text{ are jointly typical.} \tag{155}$$

If such indices can be found, then it picks the first triple of such indices, outputs $\hat{m}_j$ as its guess for m_j, and keeps $\hat{\ell}_{j-1}$ to be used when decoding Block $(j-1)$.

Analysis. A number of failure modes must be addressed.

- Decoding of $\ell_{\lambda-1}$ fails. Since the capacity with the aforementioned symbol-by-symbol helper is positive, there exists a $\gamma > 0$ such that, provided Block λ has length at least γn, the probability of a decoding error in this last block can be made to tend to zero as $n \to \infty$. Since we shall later choose λ very large, the exact value of γ will not affect the overall rate.
- When encoding for Block j, the encoder and the helper cannot find indices satisfying (154). Notice that the vectors (146) are generated IID according to P_V and independently of $\mathbf{s}^{(j)}$ and $\mathbf{z}^{(j)} = \mathbf{z}(\ell_{j-1})$. It follows that these vectors are also independent of $\mathbf{t}^{(j)}$, whereas the latter, when averaged over the randomly generated codebook, is IID according to P_T. Hence the probability of this event can be made to vanish as $n \to \infty$ as long as

$$R_v + \tilde{R} > I(V;T). \tag{156}$$

The following error events all concern the decoding task for Block $j \in \{1, \ldots, \lambda-1\}$. We assume that ℓ_j has been correctly decoded in Block $(j+1)$ (so $\hat{\ell}_j = \ell_j$).

- The chosen indices satisfy (154), but $\big(\mathbf{z}^{(j)}, \mathbf{u}^{(j)}, \mathbf{v}^{(j)}, \mathbf{y}^{(j)}\big)$ are not jointly typical. By our construction, over the randomly generated codebook, $\big(\mathbf{S}^{(j)}, \mathbf{Z}^{(j)}, \mathbf{T}^{(j)}, \mathbf{U}^{(j)}, \mathbf{X}^{(j)}, \mathbf{Y}^{(j)}\big)$ are IID according to P_{SZTUXY}; in particular, conditional on any $\mathbf{t}^{(j)}$, the probability of the tuple $\big(\mathbf{z}^{(j)}, \mathbf{u}^{(j)}, \mathbf{y}^{(j)}\big)$ is $P_{ZUY|T}^{\times n}\big(\mathbf{z}^{(j)}, \mathbf{u}^{(j)}, \mathbf{y}^{(j)} \big| \mathbf{t}^{(j)}\big)$. It then follows by the Markov Lemma [10, Lemma 12.1] that, given jointly typical $\big(\mathbf{v}^{(j)}, \mathbf{t}^{(j)}\big)$, the probability that $\big(\mathbf{Z}^{(j)}, \mathbf{U}^{(j)}, \mathbf{Y}^{(j)}\big)$ are jointly typical with $\mathbf{v}^{(j)}$ tends to one as $n \to \infty$.
- There exists some $k' \neq k_j$ such that $\big(\mathbf{v}(\ell_j, k'), \mathbf{y}^{(j)}\big)$ are jointly typical.[2] By our construction, $\mathbf{v}(\ell_j, k')$ is generated IID according to P_V and independently of $\mathbf{y}^{(j)}$. It then follows that the probability of this error event can be made to approach zero as $n \to \infty$ as long as

$$\tilde{R} < I(V;Y). \tag{157}$$

  Note that, in the joint distribution (143), $V \multimap T \multimap Y$ forms a Markov chain. Therefore (156) and (157) together imply that

$$R_v > I(V;T|Y). \tag{158}$$

  Conversely, given R_v satisfying (158), there exists a choice of $\tilde{R}$ to satisfy both (156) and (157).
- There exists some $m' \neq m_j$ such that $\big(\mathbf{z}(\ell_{j-1}), \mathbf{u}(\ell_{j-1}, m'), \mathbf{v}(\ell_j, k_j), \mathbf{y}^{(j)}\big)$ are jointly typical. As discussed above, with high probability $\big(\mathbf{z}(\ell_{j-1}), \mathbf{v}(\ell_j, k_j), \mathbf{y}^{(j)}\big)$ are jointly typical. By our construction, $\mathbf{u}(\ell_{j-1}, m')$ is generated with probability $P_{U|Z}^{\times n}(\mathbf{u}(\ell_{j-1}, m')|\mathbf{z}(\ell_{j-1}))$, hence the probability of a so-generated sequence being jointly typical with $\big(\mathbf{z}(\ell_{j-1}), \mathbf{v}(\ell_j, k_j), \mathbf{y}^{(j)}\big)$ is approximately $2^{-nI(U;V,Y|Z)}$. It follows that the probability of this error can be made to vanish as $n \to \infty$ provided

$$R < I(U;V,Y|Z) = I(U;Y|V,Z), \tag{159}$$

  where the equality follows because, in the joint distribution (143), $U \multimap Z \multimap V$ forms a Markov chain.
- There exist some $\ell' \neq \ell_{j-1}$ and $m' \neq m_j$ such that $\big(\mathbf{z}(\ell'), \mathbf{u}(\ell', m'), \mathbf{v}(\ell_j, k_j), \mathbf{y}^{(j)}\big)$ are jointly typical. With high probability, $\big(\mathbf{v}(\ell_j, k_j), \mathbf{y}^{(j)}\big)$ are jointly typical, and, independently, $(\mathbf{z}(\ell'), \mathbf{u}(\ell', m'))$ are drawn IID according to P_{ZU}. Therefore the probability that these vectors are jointly typical for any pair (ℓ', m') is approximately $2^{-nI(U,Z;V,Y)}$. Consequently, the probability of this type of error can be made to vanish as $n \to \infty$ provided

$$R + R_v < I(U,Z;V,Y). \tag{160}$$

Summarizing the above analyses, we conclude that, for the block-Markov scheme to succeed with high probability, it suffices that the rate R be smaller than

$$\min\big\{I(U;Y|V,Z),\ I(U,Z;V,Y) - I(V;T|Y)\big\}. \tag{161}$$

We now return to our assertion regarding the existence of a symbol-by-symbol helper of positive capacity and show that (161) does not exceed the capacity with a message-cognizant helper (33), let alone the capacity when perfect state

[2]This will cause an error not in decoding m_j, but in decoding ℓ_{j-1}, which will (very likely) cause decoding errors in Blocks $j-1$ to 1.

information is available to both encoder and decoder. Recall that this will allow us to dispense with the assumption that there exists an effective symbol-by-symbol helper, which we need in Block λ.

To show the desired inequality, consider the second term in the minimizaiton in (161):

$$\begin{aligned} &I(U,Z;V,Y) - I(V;T|Y) \\ &= I(U,Z;Y) + I(U,Z;V|Y) - H(V|Y) + H(V|T) \quad (162) \\ &= I(U,Z;Y) - H(V|U,Z,Y) + H(V|T) \quad (163) \\ &\leq I(U,Z;Y), \quad (164) \end{aligned}$$

where (162) holds because $V \multimap T \multimap Y$ forms a Markov chain; and (164) because $V \multimap T \multimap (U,Z,Y)$ forms a Markov chain. If we define

$$U' \triangleq (U,Z), \quad (165)$$

then the joint distribution of (U',X,Y,S,T) has the form (33b) with U' replacing U, therefore the right-hand side of (164) cannot exceed (33).

We have now completed the proof of the following result:

Theorem 13: Given any joint distribution of the form (143), the described block-Markov coding scheme can achieve any rate up to (161).

Claim 14: Maximizing (161) over all joint distributions of the form (143) yields an achievable rate that is at least as high as the capacity with the best symbol-by-symbol helper. In some cases it is strictly higher.

Proof: Choosing Z and V null (deterministic) reduces (161) to $I(U;Y)$, so the optimization over P_U leads to the capacity with the symbol-by-symbol helper $P_{T|S}$.

As for an example where the maximum of (161) is higher than the capacity of the best symbol-by-symbol helper, consider Example 9 with the following choices (where equalities are with probability one):

$$\begin{aligned} Z &\sim \text{uniform on } \{0,1\} && (166a) \\ T &= S^{(Z)} && (166b) \\ U &= (Z,C) \text{ with } C \text{ uniform on } \{0,1\}^\eta && (166c) \\ X &= (Z,T,C) && (166d) \\ V &= T = S^{(Z)}. && (166e) \end{aligned}$$

The output is then

$$Y = \begin{cases} (Z,C,\tilde{C}), & S^{(Z)} = 0 \\ (Z,\tilde{C},C), & S^{(Z)} = 1, \end{cases} \quad (167)$$

where $\tilde{C}$ is independent of all other random variables. We can then compute

$$\begin{aligned} I(U;Y|V,Z) &= \eta && (168) \\ I(U,Z;V,Y) &= \eta + 1 && (169) \\ I(V;T|Y) &= 1. && (170) \end{aligned}$$

The choice (166) thus results in (161) being η, which, by Claim 10, is higher than the capacity of any symbol-by-symbol helper. ∎

V. Concluding Remarks

Revealing the state of an SD-DMC to the encoder strictly causally does not increase capacity. The intuitive explanation that is usually given for this is that, because the channel and state are memoryless, the past states tell the encoder nothing about the channel's present behavior, and—while possibly useful to the decoder—it is more efficient for the encoder to convey fresh information than past states. So why, as Example 9 shows, are symbol-by-symbol helpers suboptimal?

A clue to this might be offered by the coding scheme we proposed for this example and by the general block-Markov scheme that builds on it. The idea behind both is that at time i the encoder conveys to the decoder some information about the past states that it has learned via past assistance and *that is known to the helper* (who knows S^{i-1} and consequently all the past assistance). Since the helper knows this information about S^{i-1}, this information plays a role similar to that of a message that is known to the helper. Since revealing the message to the helper can increase capacity (Example 7), it can be more efficient for the encoder to send this information about S^{i-1} than to send fresh information, which the helper does not know.

Such a scheme—where the encoder tells the decoder about past states—makes no sense if the decoder has full state information. This is congruent with the optimality of symbol-by-symbol helpers when the receiver has perfect state information (Theorem 6).

Such a scheme also makes no sense when the helper is cognizant of the message. In this case conveying fresh information is preferable to conveying information about the past states, because, in comparing the two approaches, both are done with the helper's knowledge of what is being conveyed, so the playing field is level. This provides intuition for Theorem 4, which states that, when the helper is cognizant of the message, symbol-by-symbol helpers are optimal.

Finally, we note that, although the assistance provided in Example 9 and in the block-Markov scheme depends on past states, it does not provide the encoder with any new information about the past states; it only provides the encoder information about the current state. The past states, however, determine *which information about the current state* is provided to the encoder. Indeed, in this example, even if the past states were provided to the encoder perfectly, symbol-by-symbol helpers would still be suboptimal. (By Theorem 1, the capacity with a symbol-by-symbol helper is unchanged when past states are provided to the encoder, so revealing the past states perfectly to the encoder would not allow the symbol-by-symbol helpers to catch up with the block-Markov scheme.)

Appendix

In this appendix we complete the proof of Claim 10 by analyzing the helper $T = S^{(0)}$. To this end, we view $S^{(0)}$ as a meta state, which is known causally to the encoder but not to the decoder. The capacity is then the maximum of $I(U;Y)$, where U takes values in the set of Shannon strategies that map $S^{(0)}$ to X [1]. Denote a generic Shannon strategy u as

$$u = (a^{(0)}, b^{(0)}, c^{(0)}, a^{(1)}, b^{(1)}, c^{(1)}), \quad (171)$$

indicating that u maps $S^{(0)} = 0$ to the channel input $(a^{(0)}, b^{(0)}, c^{(0)})$ and $S^{(0)} = 1$ to $(a^{(1)}, b^{(1)}, c^{(1)})$. The conditional law of Y given U is then

$$P_{Y|U}(y|u) = \sum_{s^{(1)}\in\{0,1\}} \frac{1}{2} W\big(y\big|x = (a_0, b_0, c_0), s = (0, s^{(1)})\big) + \sum_{s^{(1)}\in\{0,1\}} \frac{1}{2} W\big(y\big|x = (a_1, b_1, c_1), s = (1, s^{(1)})\big). \quad (172)$$

The duality-based upper bound [13, Thm. 5.1] states that any choice of a distribution Q on $\mathcal{Y}$ leads to an upper bound on the capacity

$$C \le \max_u D\big(P_{Y|U}(\cdot|u)\big\|Q\big). \quad (173)$$

Our choice of Q is one under which A', $D^{(0)}$, and $D^{(1)}$ are independent, with A' being Bernoulli (δ) (with $\delta < 1/2$ specified later) and with both $D^{(0)}$ and $D^{(1)}$ uniform:

$$Q(a', d^{(0)}, d^{(1)}) = \begin{cases} (1-\delta)\, 2^{-2\eta}, & a' = 0 \\ \delta\, 2^{-2\eta}, & a' = 1. \end{cases} \quad (174)$$

We next analyze $D\big(P_{Y|U}(\cdot|u)\big\|Q\big)$ for different strategies u.

- Consider any u with $a^{(0)} = a^{(1)} = 0$, $b^{(0)} = 0$, and $b^{(1)} = 1$. Such a u guarantees that $B = S^{(A)} = S^{(0)}$, and
$$Y = \begin{cases} (0, c^{(0)}, \tilde{D}) & \text{w.p. } 1/2 \\ (0, \tilde{D}, c^{(1)}) & \text{w.p. } 1/2, \end{cases} \quad (175)$$
where $\tilde{D}$ is uniform over $\{0,1\}^\eta$ and independent of everything else. Hence $P_{Y|U}(y|u)$ equals $2^{-\eta}$ for $y = (0, c^{(0)}, c^{(1)})$, and equals $2^{-\eta-1}$ for the other $(2^{\eta+1} - 2)$ outcomes of positive probability. (Outputs of the form $(0, \kappa, \kappa')$ where $\kappa \ne c^{(0)}$ and $\kappa' \ne c^{(1)}$ have zero probability.) It then follows that
$$D\big(P_{Y|U}(\cdot|u)\big\|Q\big) = 2^{-\eta} \log \frac{2^{-\eta}}{(1-\delta)2^{-2\eta}} + (1 - 2^{-\eta}) \log \frac{2^{-\eta-1}}{(1-\delta)2^{-2\eta}} \quad (176)$$
$$= \eta + 2^{-\eta} - 1 + \log \frac{1}{1-\delta}. \quad (177)$$
- Now consider u with $a^{(0)} = a^{(1)} = 0$, $b^{(0)} = 1$, and $b^{(1)} = 0$. For this u, we always have $B \ne S^{(A)}$, so Y conditional on u is always uniform, and
$$D\big(P_{Y|U}(\cdot|u)\big\|Q\big) = D\big(P_{A'|U}(\cdot|u)\big\|Q_{A'}\big) \quad (178)$$
$$= D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big) \quad (179)$$
$$\le \log \frac{1}{\delta}. \quad (180)$$
- Consider u with $a^{(0)} = a^{(1)} = 0$ and $b^{(0)} = b^{(1)} = 0$ (so the meta state is nearly ignored). Conditional on such a u, $B = S^{(A)}$ when $S^{(0)} = 0$ and $B \ne S^{(A)}$ when $S^{(0)} = 1$, so
$$Y = \begin{cases} (0, c^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/2 \\ (\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/2, \end{cases} \quad (181)$$
where $\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}$ are all uniform and independent of everything else. In the first case above, we have a probability of $2^{-\eta}$ on each of the 2^η realizations of Y of positive probability; and in the second case, we have a probability of $2^{-2\eta-1}$ on each outcome of positive probability. By convexity of relative entropy, we can upper-bound $D\big(P_{Y|U}(\cdot|u)\big\|Q\big)$ by the weighted sum of the relative entropies resulting from these two cases (in the second case it is $D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big)$):
$$D\big(P_{Y|U}(\cdot|u)\big\|Q\big) \le \frac{1}{2} \log \frac{2^{-\eta}}{(1-\delta)2^{-2\eta}} + \frac{1}{2} D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big) \quad (182)$$
$$\le \frac{\eta}{2} + \log \frac{1}{\delta}. \quad (183)$$
One can easily verify that the same bound holds for $a^{(0)} = a^{(1)} = 0$ and $b^{(0)} = b^{(1)} = 1$.
- Consider u with $a^{(0)} = a^{(1)} = 1$ and $b^{(0)} = b^{(1)} = 0$ (so the meta state is again nearly ignored). Conditional such a u, there is a probability of $1/2$ that $B \ne S^{(A)} = S^{(1)}$, and
$$Y = \begin{cases} (\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/2 \\ (1, c^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/4 \\ (1, c^{(1)}, \tilde{D}^{(1)}) & \text{w.p. } 1/4, \end{cases} \quad (184)$$
where $\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}$ are all uniform and independent of everything else. Again using convexity of relative entropy we have
$$D\big(P_{Y|U}(\cdot|u)\big\|Q\big) \le \frac{1}{2} D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big) + \frac{1}{4} \log \frac{2^{-\eta}}{\delta 2^{-2\eta}} + \frac{1}{4} \log \frac{2^{-\eta}}{\delta 2^{-2\eta}} \quad (185)$$
$$\le \frac{\eta}{2} + \log \frac{1}{\delta}. \quad (186)$$
By similar analysis, the bound (186) can be shown to hold for all u with $a^{(0)} = a^{(1)} = 1$.
- Next consider u with $a^{(0)} = 0$, $a^{(1)} = 1$, $b^{(0)} = b^{(1)} = 0$. When $S^{(0)} = 0$ we always have $B = S^{(A)} = S^{(0)}$; but, when $S^{(0)} = 1$, B is independent of $S^{(A)} = S^{(1)}$. The output is thus as follows:
$$Y = \begin{cases} (0, c^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/2 \\ (1, c^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/4 \\ (\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}) & \text{w.p. } 1/4, \end{cases} \quad (187)$$
where $\tilde{A}, \tilde{D}^{(0)}, \tilde{D}^{(1)}$ are all uniform and independent of everything else. Using convexity of relative entropy we have
$$D\big(P_{Y|U}(\cdot|u)\big\|Q\big) \le \frac{1}{2} \log \frac{2^{-\eta}}{(1-\delta)2^{-2\eta}} + \frac{1}{4} \log \frac{2^{-\eta}}{\delta 2^{-2\eta}} + \frac{1}{4} D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big) \quad (188)$$
$$\le \frac{3\eta}{4} + \log \frac{1}{\delta}. \quad (189)$$

The same bound holds for $a^{(0)}=0$, $a^{(1)}=1$, $b^{(0)}=0$, $b^{(1)}=1$, as well as for $a^{(0)}=1$, $a^{(1)}=0$, $b^{(1)}=1$ (and $b^{(0)}=0$ or 1).

- Finally, consider $a^{(0)}=0$, $a^{(1)}=1$, $b^{(0)}=1$, and $b^{(1)}=0$. (The bound will also apply to $a^{(0)}=0$, $a^{(1)}=1$, $b^{(0)}=b^{(1)}=1$, and to $a^{(0)}=1$, $a^{(1)}=0$, $b^{(1)}=0$, $b^{(0)}=0$ or 1.) In this case, when $S^{(0)}=0$, we always have $B\neq S^{(A)}=S^{(0)}$; and when $S^{(0)}=1$, we have that B is independent of $S^{(A)}=S^{(1)}$. Consequently,

$$Y=\begin{cases}(1,c^{(1)},\tilde{D}^{(1)}) & \text{w.p. } 1/4\\ (\tilde{A},\tilde{D}^{(0)},\tilde{D}^{(1)}) & \text{w.p. } 3/4,\end{cases} \tag{190}$$

where $\tilde{A},\tilde{D}^{(0)},\tilde{D}^{(1)}$ are all uniform and independent of everything else. We then have

$$\begin{aligned} &D\big(P_{Y|U}(\cdot|u)\big\|Q\big)\\ &\quad\le \frac{1}{4}\log\frac{2^{-\eta}}{\delta 2^{-2\eta}}+\frac{3}{4}D\big(\text{Bern}(1/2)\big\|\text{Bern}(\delta)\big) &(191)\\ &\quad\le \frac{\eta}{4}+\log\frac{1}{\delta}. &(192)\end{aligned}$$

Summarizing all above cases, the capacity with the helper $T=S^{(0)}$ is upper-bounded as

$$C\le\max\left\{\eta+2^{-\eta}-1+\log\frac{1}{1-\delta},\ \frac{3\eta}{4}+\log\frac{1}{\delta}\right\}. \tag{193}$$

If we choose $\delta=1/4$, then the bound becomes

$$C\le\max\left\{\eta+2^{-\eta}-\log\frac{3}{2},\ \frac{3\eta}{4}+2\right\}, \tag{194}$$

which is less than η if $\eta>8$. This completes the proof.

When η is very large, we can choose δ to be close to zero, and the best bound obtained from (193) will be approximately $\eta-1$. In fact, $\eta-1$ is indeed approximately the best performance with a symbol-by-symbol helper when η is very large.

REFERENCES

[1] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Research and Development*, vol. 2, pp. 289–293, 1958.

[2] S. I. Gel'fand and M. S. Pinsker, "Coding for channels with random parameters," *Prob. Contr. and Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[3] A. Lapidoth and Y. Steinberg, "The multiple-access channel with causal side information: Common state," *IEEE Trans. Inform. Theory*, vol. 59, no. 1, pp. 32–50, 2013.

[4] A. Lapidoth and Y. Steinberg, "The multiple-access channel with causal side information: Double state," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1379–1393, 2013.

[5] M. Li, O. Simeone, and A. Yener, "Multiple access channels with states causally known at transmitters," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1394–1404, 2013.

[6] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," *Foundations and Trends® in Communications and Information Theory*, vol. 4, no. 6, pp. 445–586, 2008.

[7] A. Rosenzweig, Y. Steinberg, and S. Shamai (Shitz), "On channels with partial channel state information at the transmitter," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1817–1830, May 2005.

[8] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2007–2019, 1999.

[9] S. Jafar, "Capacity with causal and noncausal side information: A unified view," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5468–5474, 2006.

[10] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[11] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, second ed., 2006.

[13] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat fading channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2426–2467, Oct. 2003.

[14] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, second ed., 2011.

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 ETH Zurich, Switzerland. He served in the years 2003–2004 and 2009 as Associate Editor for Shannon Theory for the IEEE Transactions on Information Theory.

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.

Ligong Wang (S'08, M'12) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2004, and the M.Sc. and Dr.Sc. degrees in electrical engineering from ETH Zurich, Switzerland, in 2006 and 2011, respectively. In the years 2011–2014 he was a Postdoctoral Associate with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. He joined the CNRS of France as a Researcher (chargé de recherche) in 2014; he is on leave of absence from the CNRS since 2023. He is now a Senior Researcher with the Signal and Information Processing Laboratory, ETH Zurich, Switzerland, and also a Lecturer with the Lucerne University of Applied Sciences and Arts (HSLU), Switzerland. Since 2019 he has been serving as Associate Editor for Shannon Theory for the IEEE Transactions on Information Theory.