Capacity of Channels With Action-Dependent States

Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We consider channels with action-dependent states: Given the message to be communicated, the transmitter chooses an action sequence that affects the formation of the channel states, and then creates the channel input sequence based on the state sequence. We characterize the capacity of such a channel both for the case where the channel inputs are allowed to depend noncausally on the state sequence and the case where they are restricted to causal dependence. Our setting covers previously considered scenarios involving transmission over channels with states known at the encoder, as well as various new coding scenarios for channels with a "rewrite" option that may arise naturally in storage for computer memories with defects or in magnetic recoding. A few examples are worked out in detail.**

***Index Terms*—Actions, channel with a rewrite option, channel with states, cost constraints, dirty paper coding, Gel'fand–Pinsker Channel, Shannon Channel.**

I. Introduction

COMMUNICATION through state-dependent channels, with states known at the transmitter, is a problem that has received much attention since the work of Shannon [11], Kusnetsov and Tsybakov [7], Gel'fand and Pinsker [4], and Heegard and El Gamal [5]. The assumption in these seminal papers, as well as in the work on communication with state-dependent channels that followed (cf. [6] and references therein), is that the channel states are generated by nature, and cannot be affected or controlled by the communication system.

In this paper, we revisit this problem setting for the case where the transmitter can take actions that affect the formation of the states. Specifically, we consider a communication system where encoding is in two parts: given the message, an action sequence is created. The actions affect the formation of the channel states, which are accessible to the transmitter when producing the channel input sequence. A channel with action-dependent states then is characterized by two ingredients: the distribution of state given an action $P_{S|A}$ (in lieu of the distribution of the state P_S in the original setting) and, as in the original setting, the distribution of the channel output given the input and state $P_{Y|X,S}$. We characterize the capacity of such a channel both for the case where the channel inputs are allowed to depend noncausally on the state sequence, and that where they are restricted to causal dependence.

Manuscript received January 20, 2009; revised December 25, 2009. Date of current version October 20, 2010. The material in this paper was presented at the International Symposium on Information Theory, Seoul, Korea, June 2009.

The author is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA, on leave from the Department of Electrical Engineering, Technion, Haifa 32000, Israel (e-mail: tsachy@stanford.edu).

Communicated by M. C. Gastpar, Associate Editor for Shannon Theory.

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

Digital Object Identifier 10.1109/TIT.2010.2068991

Our problem setting, depicted in Fig. 1, is one of 2-stage coding, where what we refer to as the "actions" is nothing but the channel input sequence in the first stage of the encoding. Following this first encoding stage, the remaining problem, as would be depicted by deleting the two boxes in the upper left part of Fig. 1, is the standard one of coding for channels with states known at the transmitter. Thus, inspired by such fields as control, dynamic programming, and Markov decision processes, we have elected to refer to that part of the encoding that affects the formation of the states as "actions."

Beyond merely generalizing previously considered problems involving coding with states known at the transmitter, including some recent ones considered in [12] pertaining to multiple access channels with states that we expand on in Section IV-B-I, our framework captures various new channel coding scenarios that may arise naturally in recording for magnetic storage devices or coding for computer memories with defects. Concretely, consider a 2-stage procedure for recording on a memory with defects. After writing into the memory for the first time, the encoder observes a noisy version of what the decoder will see when it tries to read from the memory. The encoder is now allowed to rewrite at whichever memory locations it chooses before the decoder attempts to decipher the information. How much information can be reliably communicated in this process? Suppose that in the first use of the memory neither encoder nor decoder know where the defects are. Then, in the second use (the "rewrite" stage), the encoder will have some idea on these defects according to the signal it input in the first stage and the noisy measurement of the channel output for that stage. In general, there is a tension between the amount of information the encoder can convey in the first pass and its ability to learn about the channel state to better communicate in the second pass. Our framework quantifies this tension and yields a characterization of the fundamental limits on communication for such 2-stage coding systems.

Our problem can be thought of as the channel coding dual of source coding, with decoder side information, where the decoder is allowed to choose actions that affect the nature and quality of the side information, as considered in [9].

The remainder of the paper is organized as follows: Sections II and III are dedicated, respectively, to characterizing the fundamental limits of communication over channels with action-dependent states available at the encoder, when the states are available noncausally and causally. In Section IV we extend our results to the case of cost constraints, point out equivalent representations of our capacity formulas, and discuss some special cases. In Section V we apply our results to characterize the capacity for various coding scenarios involving a channel "rewrite" option. In Section VI we look at an extension of Costa's dirty paper problem [1] to our setting, which we call the "writing-on-clean-paper-and-then-writing-on-its-corrupted

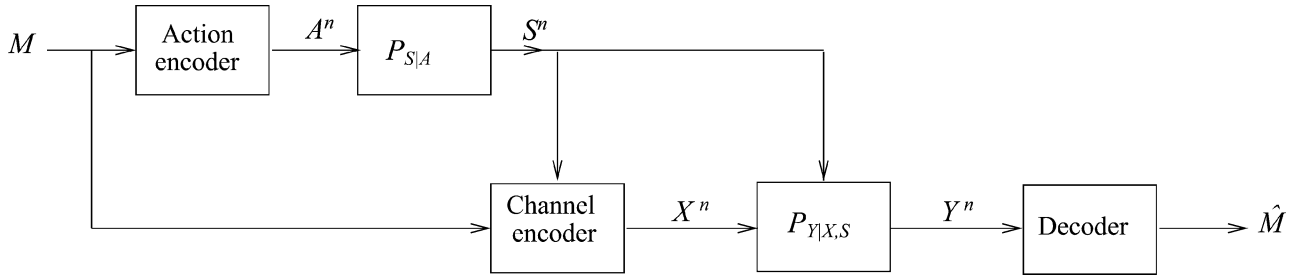

Fig. 1. Channel with action-dependent states.

version" channel. We conclude in Section VII with a summary of our work and some future directions.

II. Noncausally Available States

Let upper case, lower case, and calligraphic letters denote, respectively, random variables, specific or deterministic values they may assume, and their alphabets. For two jointly distributed random objects X and Y, let P_X, $P_{X,Y}$, and $P_{X|Y}$ respectively denote the distribution of X, the joint distribution of X, Y, and the conditional distribution of X given Y. In particular, when X and Y are discrete, $P_{X|Y}$ represents the stochastic matrix whose elements are $P_{X|Y}(x|y) = P(X = x|Y = y)$. X_m^n denotes the $n-m+1$-tuple $(X_m, \ldots, X_n)$ when $m \leq n$ and the empty set otherwise. X^n is shorthand for X_1^n.

We dedicate this section to characterizing the fundamental limits on reliable communication for schemes of the form depicted in Fig. 1: given the message M, selected uniformly at random from the message set $\mathcal{M} = \{1, \ldots, |\mathcal{M}|\}$, an action sequence $A^n = A^n(M)$ is selected. Nature now generates the state sequence S^n as the output of the memoryless channel $P_{S|A}$ whose input is A^n. A channel input sequence is now selected on the basis of the message and the whole state sequence $X^n = X^n(M, S^n)$. The joint PMF of M, A^n, S^n, X^n, Y^n, induced by a given scheme, is thus

$$P_{M,A^n,S^n,X^n,Y^n}(m, a^n, s^n, x^n, y^n) = \frac{1_{\{A^n(m)=a^n, X^n(m,s^n)=x^n\}}}{|\mathcal{M}|} \times \prod_{i=1}^{n} P_{S|A}(s_i|a_i) P_{Y|X,S}(y_i|x_i, s_i). \quad (1)$$

The associated probability of error is $P_e = P(M \neq \hat{M}_{ML}(Y^n))$, where $\hat{M}_{ML}(Y^n)$ is the best (maximum likelihood) estimate of M based on Y^n under the joint distribution in (1). The rate R is said to be achievable if there exists a sequence of schemes for increasing block lengths with $\frac{1}{n}\log|\mathcal{M}| \geq R$ and $P_e \xrightarrow{n\to\infty} 0$. The capacity of the channel with action-dependent states known noncausally to the transmitter is the supremum over all achievable rates.

Theorem 1: The capacity of the channel with action-dependent states known noncausally to the transmitter is given by

$$C = \max [I(U;Y) - I(U;S|A)] \quad (2)$$
$$= \max [I(A,U;Y) - I(U;S|A)], \quad (3)$$

where the maxima are over all joint distributions of the form

$$P_{A,S,U,X,Y}(a, s, u, x, y) = P_A(a) P_{S|A}(s|a) \times P_{U|S,A}(u|s, a) 1_{\{x=f(u,s)\}} P_{Y|X,S}(y|x, s) \quad (4)$$

for some P_A, $P_{U|S,A}$, f and $|\mathcal{U}| \leq |\mathcal{A}||\mathcal{S}||\mathcal{X}| + 1$.

Comments:

- As in the classical Gel'fand–Pinsker setting [4], the maximum in (2) does not increase when allowing more general distributions of the form

$$P_{A,S,U,X,Y}(a, s, u, x, y) = P_A(a) P_{S|A}(s|a) \times P_{U|S,A}(u|s, a) P_{X|U,S}(x|u, s) P_{Y|X,S}(y|x, s) \quad (5)$$

  i.e., allowing a general conditional distribution $P_{X|U,S}$ rather than restricting X to be a deterministic function of (U, S) as in (4). The reason is that $I(U;Y) - I(U;S|A)$ is a convex functional of $P_{X|U,S}$ [when all other distributions and conditional distributions in (4) are fixed], and thus maximized at a corner point of the simplex. To see why this convexity holds note that $I(U;S|A)$ depends only on $P_{U,S,A}$, so we need only establish convexity of $I(U;Y)$ in $P_{X|U,S}$. To this end, take two conditional distributions $P^{(1)}_{X|U,S}$, $P^{(2)}_{X|U,S}$, and let $P_{X|U,S} = \alpha P^{(1)}_{X|U,S} + (1-\alpha) P^{(2)}_{X|U,S}$ for $\alpha \in [0,1]$. The convexity of $I(U;Y)$ in $P_{X|U,S}$ follows by the convexity of $I(U;Y)$ in $P_{Y|U}$ (cf., e.g., [2]) upon noting that $P_{Y|U} = \alpha P^{(1)}_{Y|U} + (1-\alpha) P^{(2)}_{Y|U}$ when $P_{Y|U}$, $P^{(1)}_{Y|U}$, $P^{(2)}_{Y|U}$ denote the conditional distributions induced, respectively, by $P_{X|U,S}$, $P^{(1)}_{X|U,S}$, $P^{(2)}_{X|U,S}$.
- Let $C_{GP}(P_S, P_{Y|X,S})$ denote the capacity of the channel with states known noncausally at the transmitter, as considered in [4], [5]. It is natural to wonder how the capacity in our setting, as characterized in Theorem 1, compares with $\max_{a\in\mathcal{A}} C_{GP}(P_{S|A=a}, P_{Y|X,S})$, the rate that would be achieved by greedily selecting the action leading to the best Gel'fand–Pinsker (GP) channel at all time points, and proceeding with an optimal GP code for that channel. Under such a strategy, no information is conveyed by the actions which are only used to set up the channel, and communication is performed only at the second stage. For an extreme example of how suboptimal such a strategy can be, consider a channel for which $P_{Y|X,S} = P_{Y|S}$. Clearly, $C_{GP}(P_S, P_{Y|X,S}) = 0$ for any such channel and, therefore, $\max_{a\in\mathcal{A}} C_{GP}(P_{S|A=a}, P_{Y|X,S}) = 0$. On the other hand, as is readily seen to follow by applying Theorem 1 or from first principles, the capacity of this channel

is $\max_{P_A} I(A;Y)$, which may be positive. More generally, the capacity achieving scheme finds the optimal balance between conveying information through the choice of actions and the tendency to take actions that will result in states conducive for the communication in the second stage.

- Writing out the expression in (3) as

$$I(A,U;Y)-I(U;S|A)=I(A;Y)+[I(U;Y|A)-I(U;S|A)] \tag{6}$$

shows that capacity achieving schemes are of a 2-stage form: The first stage involves the choice and communication of a state sequence, through which $I(A;Y)$ bits can be communicated per channel use, while the second stage consists of coding over a Gel'fand–Pinsker channel with states whose distribution is the conditional one given the action sequence, which has been deciphered following the first stage.

- It might sometimes be natural to consider channels of the form $P_{Y|X,S,A}$. The capacity expressions for this seemingly more general channel remain almost unchanged, the only difference being that X need be taken of the form $X(U,S,A)$ rather than $X(U,S)$. This follows directly by defining a new state $S'=(S,A)$ and applying the above characterization.

Proof of Theorem 1: We first establish the equality in (3), i.e., that the maximization over $I(U;Y)-I(U;S|A)$ in (2) can be replaced by one over $I(A,U;Y)-I(U;S|A)$, As in (3). That the latter upper bounds the former trivially follows from the data processing inequality. For the reverse inequality, note that $I(A,U;Y)-I(U;S|A)=I(A,U;Y)-I(A,U;S|A)=I(U';Y)-I(U';S|A)$, where $U'=(U,A)$ and the joint distribution (A,S,U',X,Y) satisfies the same conditional independence relations as are required of (A,S,U,X,Y) [cf. (5)]. It thus remains to establish equality (2), and that the associated maximization is unaffected by the bound on the cardinality of $\mathcal{U}$, to which we now turn.

Proof of Achievability: We use arguments that have become standard since the work of Gel'fand and Pinsker [4] and Heegard and El Gamal [5]. Fix P_A, $P_{U|S,A}$, f and consider A, S, U, X, Y jointly distributed as in (4).

- Generate[1] $\{A^n(m)\}_{m=1}^{2^{nR}}$ n-tuples iid $\sim P_A$
- For each $1\le m\le 2^{nR}$ generate $\{U^n(j,m)\}_{j=1}^{2^{nR'}}$ iid $\sim \prod_{i=1}^n P_{U|A}(\cdot|A_i(m))$
- Encoding:
  - Choose $A^n(M)$ as the action sequence.
  - Let S^n be the state sequence generated in response to the action sequence.
  - Let J be the smallest value of j such that $(U^n(j,M),\ S^n,\ A^n(M))\ \in\ T_P$,[2] and take $J=1$ if no such j exists.
  - Let the channel input sequence be given by $X^n=f(U^n(J,M),S^n)$, where $f(U^n,S^n)$ denotes the n-tuple whose ith component is $f(U_i,S_i)$.
- Decoding:
  - Seeing the channel output Y^n, let $\hat{M}$ be the smallest value of $\hat{m}$ for which there exists a $\hat{j}$ such that $(A^n(\hat{m}),U^n(\hat{j},\hat{m}),Y^n)\ \in\ T_P$, and take $\hat{M}\ =\ 1$ if no such $\hat{m}$ exists.

The event $M\neq\hat{M}$ is contained in the union of the following three error events:

- At the encoding stage, there exists no j such that $(U^n(j,M),S^n,A^n(M))\ \in\ T_P$. The probability of this event is vanishing in n so long as $R'>I(U;S|A)$. To see this note that, for a particular $1\le j\le 2^{nR'}$,[3]

$$P((U^n(j,M),S^n,A^n(M))\in T_P|A^n(M)\in T_P)\doteq 2^{-nI(U;S|A)}$$

so, for $R'>I(U;S|A)$, the probability

$$P\Big(\cap_{1\le j\le 2^{nR'}}\{(U^n(j,M),S^n,A^n(M))\notin T_P\}\Big|\,A^n(M)\in T_P\Big)$$

is vanishing (in fact, double-exponentially rapidly). Since $A^n(M)\in T_P$ with probability overwhelmingly close to 1, so is the unconditioned probability.[4]

- There exists $\hat{m}\neq M$ such that $(A^n(\hat{m}),U^n(\hat{j},\hat{m}),Y^n)\in T_P$ for some $\hat{j}$. To bound the probability of this event note first that, for all $\hat{m}$ and $\hat{j}$

$$P\Big((A^n(\hat{m}),U^n(\hat{j},\hat{m}),Y^n)\in T_P|\hat{m}\neq M\Big)\doteq 2^{-nI(A,U;Y)}\le 2^{-nI(U;Y)}.$$

Therefore, by the union bound

$$P\Big(\cup_{\hat{j}}\Big\{(A^n(\hat{m}),U^n(\hat{j},\hat{m}),Y^n)\in T_P\Big\}\Big|\,\hat{m}\neq M\Big)\dot{\le} 2^{-n(I(U;Y)-R')}.$$

It follows that the probability of the existence of $\hat{m}\ \neq\ M$ for which there exists a $\hat{j}$ such that $(A^n(\hat{m}),U^n(\hat{j},\hat{m}),Y^n)\ \in\ T_P$ is vanishing provided $R<I(U;Y)-R'$.

- $(A^n(M),U^n(j,M),Y^n)\notin T_P$ for all j. The probability of this event is vanishing so long as $R'>I(U;S|A)$ since, as argued for the first event, $(U^n(J,M),S^n,A^n(M))\in T_P$ with probability approaching one and, hence, by the Markov lemma (cf., e.g., [2]), so is the probability that $(U^n(J,M),S^n,A^n(M),X^n,Y^n)\in T_P$.

Thus we have established the existence of a sequence of schemes with $R'\ =\ I(U;S|A)+\varepsilon$, $R\ =\ I(U;Y)-I(U;S|A)-2\varepsilon$, and vanishing probability of decoding error in the blocklength n.

[1]Here and throughout we ignore integer constraints, writing 2^{nR} in lieu of the more precise $\lfloor 2^{nR}\rfloor$.

[2]To avoid cumbersome notation, we let T_P generically denote sets that are typical in the sense of [3] (cf., in particular, the δ-convention therein), with respect to (joint) distributions that are clear from the context.

[3]Here and throughout we use the notation $\doteq$ to denote equality in the exponential order. Specifically, $a_n\doteq b_n$ is shorthand for $\frac{1}{n}|\log(a_n/b_n)|\le\varepsilon_n$, where $\{\varepsilon_n\}$ is a universal sequence, independent of the particularities of $\{a_n\}$ and $\{b_n\}$, and satisfying $\varepsilon_n\to 0$ as $n\to\infty$. Similarly, $a_n\ \dot{\le}\ b_n$ (respectively, $a_n\ \dot{\ge}\ b_n$) is shorthand for $\frac{1}{n}\log(a_n/b_n)\le\varepsilon_n$ (respectively, $\frac{1}{n}\log(b_n/a_n)\le\varepsilon_n$).

[4]Note the argument given here for why the probability that there exists no j such that $(U^n(j,M),S^n,A^n(M))\ \in\ T_P$ is vanishing so long as $R'\ >\ I(U;S|A)$ is based on joint typicality arguments that are by now standard. Henceforth we state facts that are due to such standard arguments without spelling these arguments out. We refer to [3] for a detailed exposition of the use of joint typicality arguments to establish facts similar to those we state here, and to [6] for a demonstration of their use specifically in the context of channel coding with state information.

Proof of Converse: Fix a scheme and consider (7)–(15), shown at the bottom of the page, where

- Equation (8) is due to the Markov relation $M - A^n - S^n$
- Equation (10) is due to the identities $I(M;Y_i|Y^{i-1}) = I(M,S_{i+1}^n,A^n;\ Y_i|Y^{i-1}) \ -I(S_{i+1}^n,A^n;Y_i|M,Y^{i-1})$ and $I(M;S_i\ |S_{i+1}^n,A^n) = I(M,Y^{i-1};\ S_i|S_{i+1}^n,A^n) -I(Y^{i-1};S_i|M,\ S_{i+1}^n,A^n) = I(M,Y^{i-1};S_i\ |S_{i+1}^n,A^n) -I(Y^{i-1};S_i,\ A^n|M,S_{i+1}^n,A^n)$
- Equation (11) follows from the identity $\sum_{i=1}^n I(S_{i+1}^n,A^n;Y_i|M,Y^{i-1}) = \sum_{i=1}^n I(Y^{i-1};S_i,A^n|M,S_{i+1}^n,A^n)$, which can be seen as shown in the equation at the bottom of the page.
- Equation (13) is due to the Markov relation $S_i - A_i - (S_{i+1}^n,A^{n\setminus i})$ (where $A^{n\setminus i} = (A^{i-1},A_{i+1}^n)$) and the definition $U_i = (M,Y^{i-1},S_{i+1}^n,A^n)$
- The maximization in (15) is over distributions of the form in (4) for some P_A, $P_{U|S,A}$, f. That this maximum upper bounds each summand in (14) is due to the equivalence, noted following the statement of the theorem, between maximization over distributions of the form in (4) and the

$$
\begin{aligned}
&I(M;Y^n) && (7)\\
&\quad = I(M;Y^n) - I(M;S^n|A^n) && (8)\\
&\quad = \sum_{i=1}^n I(M;Y_i|Y^{i-1}) - I(M;S_i|S_{i+1}^n,A^n) && (9)\\
&\quad = \sum_{i=1}^n I(M,S_{i+1}^n,A^n;Y_i|Y^{i-1}) - I(S_{i+1}^n,A^n;Y_i|M,Y^{i-1})\\
&\qquad - I(M,Y^{i-1};S_i|S_{i+1}^n,A^n) + I(Y^{i-1};S_i,A^n|M,S_{i+1}^n,A^n) && (10)\\
&\quad = \sum_{i=1}^n I(M,S_{i+1}^n,A^n;Y_i|Y^{i-1}) - I(M,Y^{i-1};S_i|S_{i+1}^n,A^n) && (11)\\
&\quad \le \sum_{i=1}^n H(Y_i) - H(Y_i|Y^{i-1},M,S_{i+1}^n,A^n) - [H(S_i|S_{i+1}^n,A^n) - H(S_i|Y^{i-1},M,S_{i+1}^n,A^n)] && (12)\\
&\quad = \sum_{i=1}^n H(Y_i) - H(Y_i|U_i) - [H(S_i|A_i) - H(S_i|U_i,A_i)] && (13)\\
&\quad = \sum_{i=1}^n I(U_i;Y_i) - I(U_i;S_i|A_i) && (14)\\
&\quad \le n\max[I(U;Y) - I(U;S|A)] && (15)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^n I(S_{i+1}^n,A^n;Y_i|M,Y^{i-1}) &= \sum_{i=1}^{n-1} I(S_{i+1}^n,A^n;Y_i|M,Y^{i-1})\\
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I(S_j,A^n;Y_i|M,Y^{i-1},S_{j+1}^n,A^n)\\
&= \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} I(S_i,A^n;Y_j|M,Y^{j-1},S_{i+1}^n,A^n)\\
&= \sum_{i=2}^{n}\sum_{j=1}^{i-1} I(S_i,A^n;Y_j|M,Y^{j-1},S_{i+1}^n,A^n)\\
&= \sum_{i=2}^{n} I(S_i,A^n;Y^{i-1}|M,S_{i+1}^n,A^n)\\
&= \sum_{i=1}^{n} I(S_i,A^n;Y^{i-1}|M,S_{i+1}^n,A^n)
\end{aligned}
$$

form in (5), and the facts that for each $1 \le i \le n$ we have $P_{S_i|A_i} = P_{S|A}$, $P_{Y_i|X_i,S_i} = P_{Y|X,S}$ and the Markov relations $X_i - (U_i, S_i) - A_i$ and $(U_i, A_i) - (X_i, S_i) - Y_i$. The bound on the cardinality of $\mathcal{U}$ follows in a standard way via the support lemma of [3]: $\mathcal{U}$ should have $|\mathcal{A}||\mathcal{S}||\mathcal{X}| - 1$ elements to preserve $P_{A,S,X}$ (which in turn preserves also $P_{A,S,X,Y}$, $H(Y)$ and $H(S|A)$), plus one element to preserve $H(Y|U)$ and $H(S|U, A)$, and one for preserving the relation $X - (U, S) - A$.

The proof is completed via the usual appeal to Fano's inequality. □

Comments:

- It is natural to wonder whether "feedback" from the past states at the action stage might increase the capacity. In the proof of the converse part we have used the Markov relation $S_i - A_i - (S_{i+1}^n, A^{n\setminus i})$, which need not hold when allowing actions of the form $A_i(M, S^{i-1})$. Thus, our converse does not hold for that case and whether capacity could be increased when such dependence is allowed remains open.
- On the other hand, it is readily verified that all the Markov relations in the converse proof remain intact when the usual type of feedback is allowed, i.e., when X_i is allowed to be of the form $X_i(M, S^n, Y^{i-1})$. Evidently, as in the classical case of noncausal state dependence without actions [8], in the present setting feedback does not increase capacity, either.

III. Causally Available States

Consider now a setting similar to that of the previous section, with an action sequence, as before, of the form $A^n(M)$, but with a channel input restricted to causal dependence on the state sequence, i.e., of the form $X_i(M, S^i)$, $1 \le i \le n$. So the joint PMF of $M,\ A^n,\ S^n,\ X^n,\ Y^n$, induced by a given scheme, is

$$P_{M,A^n,S^n,X^n,Y^n}(m, a^n, s^n, x^n, y^n) = \frac{1_{\{A^n(m)=a^n\}}}{|\mathcal{M}|} \times \prod_{i=1}^{n} P_{S|A}(s_i|a_i) 1_{\{X_i(m,s^i)=x_i\}} P_{Y|X,S}(y_i|x_i, s_i). \quad (16)$$

The associated probability of error is $P_e = P(M \neq \hat{M}_{ML}(Y^n))$, where $\hat{M}_{ML}(Y^n)$ is the best (maximum likelihood) estimate of M based on Y^n under the joint distribution in (16). As before, the rate R is said to be achievable if there exists a sequence of schemes for increasing block lengths with $\frac{1}{n}\log|\mathcal{M}| \ge R$ and $P_e \xrightarrow{n\to\infty} 0$. The capacity of the channel with action-dependent states known causally to the decoder is the supremum over all achievable rates.

Theorem 2: The capacity of the channel with action-dependent states known causally to the encoder is given by

$$C = \max I(U; Y) \quad (17)$$

$$= \max I(U, A; Y) \quad (18)$$

where $U,\ A,\ S,\ X,\ Y$ are distributed according to

$$P_{U,A,S,X,Y}(u, a, s, x, y) = P_U(u) 1_{\{g(u)=a\}} \times P_{S|A}(s|a) 1_{\{f(u,s)=x\}} P_{Y|X,S}(y|x, s) \quad (19)$$

for some P_U, f, g and $|\mathcal{U}| \le \min\{|\mathcal{Y}|, |\mathcal{A}||\mathcal{S}||\mathcal{X}| + 1\}$.

Comments:

- Note the convexity of $I(U; Y)$ in $P_{A|U}$ (due to its convexity in $P_{Y|U}$), which implies that the maximum in (17) would be unaffected when allowing a general $P_{A|U}$ rather than A which is a function of U, as is implied by (19). Also, A being a function of U implies that the above maximization would remain unchanged when allowing an f of the form $f(u,\ s,\ a)$. Since the convexity of $I(U; Y)$ in $P_{A|U}$ also implies its convexity in $P_{X|U,\ S,\ A}$ for fixed $P_{U,\ S,\ A}$, it further follows that the maximum would be unaffected upon allowing a general $P_{X|U,S,A}$ in lieu of the $X = f(U,\ S)$ relationship implied in (20). Thus, C in (17) can also be expressed as $C = \max I(U; Y)$, where $U,\ A,\ S,\ X,\ Y$ are distributed according to

$$P_{U,A,S,X,Y}(u, a, s, x, y) = P_{U,A}(u, a) P_{S|A}(s|a) \times P_{X|U,S,A}(x|u, s, a) P_{Y|X,S}(y|x, s) \quad (20)$$

  for some distributions $P_{U,A}$ and $P_{X|U,S,A}$. It follows that, given jointly distributed $U_i,\ A_i,\ S_i,\ X_i,\ Y_i$, to check that $I(U_i; Y_i) \le \max I(U; Y)$ one need only verify that: $S_i - A_i - U_i$, that $P_{S_i|A_i} = P_{S|A}$, that $Y_i - (X_i, S_i) - (A_i, U_i)$, and that $P_{Y_i|X_i,S_i} = P_{Y|X,S}$. We will use this fact in the proof of Theorem 2.
- Since $I(U; S|A) = 0$ when $U - A - S$, the capacity expression for the causal case can be viewed as a maximization of the same functional as that for the noncausal case, but over a smaller set of distributions restricted to satisfy, in addition to the constraints from the noncausal case, the Markov relation $U - A - S$.
- Letting $C_S(P_S, P_{Y|X,S})$ denote the capacity of the channel with states known causally at the transmitter, as considered in [11] (the subscript standing for "Shannon"), a comment analogous to that made in the setting of noncausally available states, about the way that $\max_{a\in\mathcal{A}} C_S(P_{S|A=a}, P_{Y|X,S})$ compares with the capacity characterized in Theorem 2, is applicable here. Specifically, here too it is easy to think of channels where $\max_{a\in\mathcal{A}} C_S(P_{S|A=a}, P_{Y|X,S}) = 0$ while the capacity for the setting of Theorem 2 can be arbitrarily large.
- The identity $I(A, U; Y) = I(A; Y) + I(U; Y|A)$ endows the expression in (18) with a natural 2-stage coding interpretation, analogous to that discussed following (6), where the second part encoding is now played by a Shannon rather than a Gel'fand–Pinsker channel.
- As for the setting of the previous section, channels of the form $P_{Y|X,S,A}$ are accommodated simply by allowing X in the maximization (17) to be of the form $X(U, S, A)$ rather than $X(U, S)$, a fact that follows directly by noting

that such a scenario is embedded in the original one upon defining the new state $S' = (S, A)$.

Proof of Theorem 2: The equality between the maxima, (18), follows similarly as the equality between the maxima in (3). We, thus, turn to proving (17). The achievability part follows as in the original setting of [11], by constructing a good code for the standard problem of channel coding over the DMC from U to Y. For the converse part, fix an arbitrary scheme and consider

$$\begin{aligned} I(M;Y^n) &= H(Y^n) - H(Y^n|M) && (21)\\ &\le \sum_{i=1}^{n} H(Y_i) - H(Y_i|M, Y^{i-1}) && (22)\\ &\le \sum_{i=1}^{n} H(Y_i) - H(Y_i|M, Y^{i-1}, S^{i-1}) && (23)\\ &= \sum_{i=1}^{n} I(U_i; Y_i) && (24)\\ &\le nC && (25) \end{aligned}$$

where the equality before last follows from defining $U_i = (M, Y^{i-1}, S^{i-1})$. To see why the last inequality holds note the relations $S_i - A_i - U_i$, $P_{S_i|A_i} = P_{S|A}$, $Y_i - (X_i, S_i) - (U_i, A_i)$, and $P_{Y_i|X_i,S_i} = P_{Y|X,S}$, so it remains only to justify the bound on the cardinality of $\mathcal{U}$. That this cardinality need not be larger than that of the channel output alphabet follows from an argument identical to that given in [10] for why in the classical channel coding problem there is a capacity achieving distribution putting positive mass on a number of channel input symbols that is no larger than the number of channel output symbols. That it need not exceed $|\mathcal{A}||\mathcal{S}||\mathcal{X}| + 1$ is due to an argument similar to that given at the end of the proof of Theorem 1: the requirement to preserve $H(S|U, A)$ is replaced by the requirement to preserve the Markov relation $S - A - U$ so the overall cardinality bound remains unchanged. The proof is completed via a standard use of Fano's inequality. □

It is readily checked that all the Markov and marginal distribution relations verified in the proof continue to hold when A_i is allowed to depend on past states, i.e., to be of the form $A_i(M, S^{i-1})$ rather than just $A_i(M)$. Thus, unlike for the setting of the previous section, here we have proof that "feedback" from the past states at the action stage does not increase the capacity. In fact, said relations are readily verified to continue to hold for even more general action strategies where, in addition to the past channel states, there is feedback from the past channel outputs available, i.e., schemes of the form $A_i(M, S^{i-1}, Y^{i-1})$.[5] Finally, as in the setting of noncausal encoding of the previous section, and as in the classical case of causal state dependence without actions [8], here too it can be seen that feedback at the encoding stage does not increase the capacity by verifying that the above relations continue to hold for channel inputs of the form $X_i(M, S^i, Y^{i-1})$.

[5]Note that it is meaningless to consider schemes of this form in the setting of the previous section where the channel inputs (and hence outputs) are formed only after the whole state (and hence action) sequence has been formed.

IV. Extensions and Special Cases

A. Cost Constraints

As in the classical problems, where it is often natural to introduce cost constraints on the channel input sequence, in our present setting it is natural to consider constraints on the cost of actions, of channel inputs, and of combinations thereof. Indeed, the cost in our setting, in its most general form, should be a function of both the action and the channel input symbol. For example, when both actions and channel input symbols are real-valued, it is natural to constrain the power of the sum of those symbols, i.e., to consider the cost function $(a + x)^2$. Further, as in classical problems where one may be concerned say with both peak and average power constraints, it will make sense to accommodate the possibility of $d \ge 1$ cost functions. Equivalently, we may assume one cost function of the form $\Lambda : \mathcal{A} \times \mathcal{X} \to \mathbb{R}^d$ and refer to

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\Lambda(A_i, X_i)\right] \qquad (26)$$

as the *cost* (vector) associated with a coding scheme. Given a vector $\lambda \in \mathbb{R}^d$, we refer to a rate R as *achievable at cost* λ if there exists a sequence of schemes for increasing block lengths with $\frac{1}{n}\log|\mathcal{M}| \ge R$, $P_e \xrightarrow{n\to\infty} 0$, and $\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=1}^{n}\Lambda_k(A_i, X_i)\right] \le \lambda_k$ for $1 \le k \le d$ (where Λ_k and λ_k denote the kth coordinates of Λ and λ). The capacity $C(\lambda)$ is the supremum over all rates achievable at cost λ.

Theorem 3: The capacities of the channel with action-dependent states known noncausally and causally to the transmitter, under a cost constraint λ, are given by the respective maximizations in Theorem 1 and Theorem 2, with the same cardinality bounds and an additional cost constraint

$$E\left[\Lambda(A, X)\right] \le \lambda. \qquad (27)$$

Note that both sides of inequality (27) are d-dimensional vectors, and inequality between vectors is to be understood componentwise.

Proof of Theorem 3: We prove the Theorem for the case where states are known noncausally. Proof for the case of causally available states is similar (and simpler). That the introduction of cost constraints entails no increase in the cardinality of $\mathcal{U}$ follows from the fact that the preservation of $P_{A,S,X}$, which was argued in the absence of a cost constraint, automatically implies the preservation of $E\left[\Lambda(A, X)\right]$. Let now $C^{(I)}(\lambda)$ denote the maximum specified in Theorem 1, with the additional constraint $E\left[\Lambda(A, X)\right] \le \lambda$. That $C(\lambda) \ge C^{(I)}(\lambda)$ follows from essentially the same achievability arguments as in the case without a cost constraint. The converse, namely that $C(\lambda) \le C^{(I)}(\lambda)$, follows similarly as in the unconstrained case once concavity of $C^{(I)}(\lambda)$ is established. To this end, define

$$C_Q^{(I)}(\lambda) = \max\left[I(U;Y|Q) - I(U;S|A,Q)\right] \qquad (28)$$

where the maximization is over all joint distributions of the form

$$P_{Q,A,S,U,X,Y}(q,a,s,u,x,y) = P_Q(q)P_{A|Q}(a|q)P_{S|A}(s|a) \times P_{U|S,A,Q}(u|s,a,q)1_{\{x=f(u,q,s)\}}P_{Y|X,S}(y|x,s) \quad (29)$$

for some P_Q, $P_{A|Q}$, $P_{U|S,A,Q}$, f and such that $E[\Lambda(A,X)] \leq \lambda$. Thus, $C_Q^{(I)}(\lambda)$ is the "concavification" of $C^{(I)}(\lambda)$ via the "time-sharing" random variable Q. Since the maximum defining $C_Q^{(I)}(\lambda)$ is over a larger set than that in the definition of $C^{(I)}(\lambda)$, we obviously have $C_Q^{(I)}(\lambda) \geq C^{(I)}(\lambda)$. It remains to argue why $C_Q^{(I)}(\lambda) \leq C^{(I)}(\lambda)$. To this end note that under any distribution of the form in (29), [see (30)–(33) at the bottom of the page], where the inequality follows since $H(Y|Q) \leq H(Y)$ and $H(S|A,Q) = H(S|A)$ due to the Markov relation $S - A - Q$, and the last equality follows by letting $U' = (Q,U)$. The proof is completed by noting that the joint distribution of (A,S,U',X,Y) is of the form in the feasible set for the maximization defining $C^{(I)}(\lambda)$. □

B. Special Cases

Common Message Capacity of MAC With States at One Transmitter: A setting considered in [12] is that of communicating a common message over a memoryless state-dependent multiple access channel characterized by $P_{Y|S,X_1,X_2}$, where the state sequence is known (noncausally) to the second encoder, but unknown at the first encoder and at the receiver. This problem, motivated in [12] by multiterminal communication scenarios involving transmitters with different degrees of channel state information, can be seen as a special case of our setting via the following associations:

- $A \rightarrow X_1$
- $P_{S|A} \rightarrow P_S$
- $X \rightarrow X_2$
- $P_{Y|S,A,X} \rightarrow P_{Y|S,X_1,X_2}$

Applying Theorem 1 to this case, keeping in mind the last comment following the statement of that theorem, about channels of the form $P_{Y|S,A,X}$, and noting that $I(U;S|X_1) = I(U,X_1;S)$ when X_1 and S are independent, we get that the capacity is given by

$$\max[I(U;Y) - I(U,X_1;S)] \quad (34)$$

under joint distributions of the form

$$P_{X_1}(x_1)P_S(s)P_{U|S,X_1}(u|s,x_1)1_{\{x_2=f(u,s,x_1)\}} \times P_{Y|S,X_1,X_2}(y|s,x_1,x_2) \quad (35)$$

where the maximization is over P_{X_1}, $P_{U|S,X_1}$, and f. This recovers Corollary 2 of [12].

Actions Seen by Decoder: Noncausal Knowledge of States at Transmitter: Consider the case where the decoder has access to the actions taken. Noting that this is a special case of our setting by taking the pair (Y,A) as the new channel output, that $U - (X,S,A) - Y$ if and only if $U - (X,S,A) - (Y,A)$, and the identity (see the equation at the bottom of the page) we obtain that the capacity for this case is given by

$$\max[H(A) + I(U;Y|A) - I(U;S|A)] \quad (36)$$

where the maximization is over the same set of distributions as in Theorem 1. This expression is quite intuitive: The amount of information per symbol that can be conveyed through the actions in the first stage is represented by the term $H(A)$. In the second stage, both encoder and decoder know the action sequence, so can condition on it and proceed with ordinary Gel'fand–Pinsker coding on each subsequence associated with each action symbol, achieving a rate $I(U;Y|A) - I(U;S|A)$. The maximization is a search for the optimal tradeoff between the amount of information that can be conveyed by the actions, and the quality of the Gel'fand–Pinsker channel that they induce.

Causal Knowledge of States at Transmitter: Consider now the case where the states are known causally to the transmitter. Noting the same things as above, the capacity representation in (18), and the identity

$$\begin{aligned} I(U,A;Y,A) &= H(A) + H(U|A) - H(U,A|Y,A) \\ &= H(A) + H(U|A) - H(U|Y,A) \\ &= H(A) + I(U;Y|A) \end{aligned}$$

$$\begin{aligned} I(U;Y|Q) - I(U;S|A,Q) &= H(Y|Q) - H(Y|Q,U) - H(S|A,Q) + H(S|A,Q,U) && (30)\\ &\leq H(Y) - H(Y|Q,U) - H(S|A) + H(S|A,Q,U) && (31)\\ &= I(Q,U;Y) - I(Q,U;S|A) && (32)\\ &= I(U';Y) - I(U';S|A) && (33) \end{aligned}$$

$$\begin{aligned} I(A,U;Y,A) - I(U;S|A) &= H(A) + H(U|A) - H(A,U|Y,A) - I(U;S|A) \\ &= H(A) + H(U|A) - H(U|Y,A) - I(U;S|A) \\ &= H(A) + I(U;Y|A) - I(U;S|A) \end{aligned}$$

we obtain that the capacity for this case is given by

$$\max[H(A)+I(U;Y|A)] \tag{37}$$

where the maximization is over the same set of distributions as in Theorem 2. Analogously as in the preceding case, here the expression has a similar interpretation of conveying information in the first part via the selection of actions and then proceeding with coding for the ordinary channel with causally available states at the transmitter a la Shannon [11], on each subsequence associated with each action symbol.

V. Channels With a Rewrite Option

The generic framework considered thus far can be specialized to various scenarios involving coding for channels with a "rewrite" option. Such channels are natural to study in our current 2-part coding scenario when viewed as computer memories with defects, as formalized and motivated in [5]. We now detail some such scenarios.

A. Noise-Free Feedback

Consider a DMC characterized by $P_{Y|X}$. After using the channel once and observing its output with no additional noise, the encoder makes another pass where it may rewrite at whichever locations it chooses, and the channel output after a rewrite will be an independent realization of the same channel $P_{Y|X}$. What is the capacity of such a coding scenario?

This scenario can be cast into our framework via the following associations: The role of the action sequence is played by the first channel input, which we denote here by X, which takes values in the alphabet of possible inputs to the channel $P_{Y|X}$, namely in $\mathcal{X}$. The role of the channel state information is played by the channel output after the first pass, which we denote by $Y^{(1)}$ (the superscript (1) indicating that this is the channel output after the *first* pass), and takes values in the alphabet of the output of the channel $P_{Y|X}$, namely in $\mathcal{Y}$. The channel input in the second stage we denote here by $\tilde{X}$ (playing the role of X in our generic framework), which takes values in the alphabet $\tilde{\mathcal{X}} = \{\text{norewrite}\} \cup \mathcal{X}$. We denote the final channel output by $Y^{(2)}$, the superscript (2) pertaining to it being the channel output after the second pass, it too takes values in the alphabet of the output of the channel $P_{Y|X}$, namely in $\mathcal{Y}$. Note that $Y^{(2)}$ plays the role of Y from our generic framework.

Thus, the role of $P_{S|A}$ from our generic framework is played here by $P_{Y^{(1)}|X} = P_{Y|X}$, while that of $P_{Y|X,S,A}$ is played by $P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}$, where [see (38) at the bottom of the page]. Applying Theorem 1, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass may depend noncausally on the channel output from the first pass is given by

$$\max[I(X,U;Y^{(2)}) - I(U;Y^{(1)}|X)] \tag{39}$$

where the maximization is over all joint distributions of the form

$$\begin{aligned} &P_{X,Y^{(1)},U,\tilde{X},Y^{(2)}}(x,y^{(1)},u,\tilde{x},y^{(2)}) \\ &\quad = P_X(x)P_{Y|X}(y^{(1)}|x)P_{U|Y^{(1)},X}(u|y^{(1)},x) \\ &\qquad \times 1_{\{\tilde{x}=f(u,y^{(1)},x)\}}P_{Y^{(2)}|Y^{(1)},\tilde{X}}(y^{(2)}|y^{(1)},\tilde{x}) \end{aligned} \tag{40}$$

with $P_{Y^{(2)}|Y^{(1)},\tilde{X}}$ given in (38), and where the maximization is over $P_X, P_{U|Y^{(1)},X}, f$, with $|\mathcal{U}| \le |\mathcal{X}||\mathcal{Y}|(|\mathcal{X}|+1)+1$.

Applying Theorem 2, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass depend causally on the channel output components from the first pass is given by

$$\max I(U;Y^{(2)}) \tag{41}$$

where $U,\ X,\ Y^{(1)},\ \tilde{X},\ Y^{(2)}$ are distributed according to

$$\begin{aligned} &P_{U,X,Y^{(1)},\tilde{X},Y^{(2)}}(u,x,y^{(1)},\tilde{x},y^{(2)}) \\ &\quad = P_U(u)1_{\{x=g(u)\}}P_{Y|X}(y^{(1)}|x)1_{\{\tilde{x}=f(u,y^{(1)})\}} \\ &\qquad \times P_{Y^{(2)}|Y^{(1)},\tilde{X}}(y^{(2)}|y^{(1)},\tilde{x}) \end{aligned} \tag{42}$$

with $P_{Y^{(2)}|Y^{(1)},\tilde{X}}$ given in (38), and where the maximization is over $P_U,\ g,\ f$, with $|\mathcal{U}| \le |\mathcal{Y}|$. In other words, perhaps not surprisingly given [11], capacity is achieved by coding for a memoryless channel whose input alphabet is $\mathcal{X} \times \{f : \mathcal{Y} \to \tilde{\mathcal{X}}\}$, output alphabet is $\mathcal{Y}$, and where the probability of a channel output symbol y given input "symbol" $(x, f(\cdot))$ is the probability that the symbol y is eventually observed at the output of the channel when the symbol x is first input into it and then $f(\cdot)$, evaluated at the output symbol in response to the first input, is used to determine whether and what will be inserted for the rewrite.

Example I: BSC: Consider the case where the channel $P_{Y|X}$ is a $\mathrm{BSC}(\delta)$, $0 \le \delta \le 1/2$. Under any joint distribution allowed in the maximization in (41)

$$\begin{aligned} I(U;Y^{(2)}) &= H(Y^{(2)}) - H(Y^{(2)}|U) && (43)\\ &\le \log|\mathcal{Y}| - H(Y^{(2)}|U) && (44)\\ &\le 1 - h_2(\delta^2). && (45) \end{aligned}$$

To see why the last inequality holds note that for all $u \in \mathcal{U}$, $H(Y^{(2)}|U=u)$ is the entropy of the channel output after the

$$\begin{aligned} P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}(y^{(2)}|x,y^{(1)},\tilde{x}) &= P_{Y^{(2)}|Y^{(1)},\tilde{X}}(y^{(2)}|y^{(1)},\tilde{x}) \\ &= \begin{cases} 1_{\{y^{(1)}=y^{(2)}\}}, & \text{if } \tilde{x} = \text{no rewrite} \\ P_{Y|X}(y^{(2)}|\tilde{x}), & \text{otherwise.} \end{cases} \end{aligned} \tag{38}$$

(option of a) rewrite operation when the first input is some deterministic symbol (which equals $g(u)$) and the rewrite operation is determined based on the channel output according to some deterministic function (which equals $f(u,\cdot)$). In the case of the BSC, this entropy is lower bounded by $h_2(\delta^2)$, δ^2 being the probability that a flip would occur twice in two consecutive uses of the channel, which is readily verified by exhaustive search to be the lowest value that can be induced, across all possible choices of $g(u)$ and $f(u,\cdot)$, for the probability $P(Y^{(2)}=1|U=u)$. That $1-h_2(\delta^2)$ is indeed the capacity for this case follows from the fact that equality can be achieved in (44) and in (45) by taking in (42)$U\sim\text{Bernoulli}(1/2)$, g the identity mapping, and f to be given by

$$f(u,y^{(1)})=\begin{cases}\text{no rewrite}, & \text{if } u=y^{(1)}\\ u, & \text{otherwise.}\end{cases} \tag{46}$$

Moving to a derivation of a lower bound on capacity in the case where the rewrite operations may depend noncausally on the channel output from the first pass, consider a joint distribution of the type allowed in (40), as follows: $X\sim\text{Bernoulli}(1/2)$, $Y^{(1)}$ is the output of $P_{Y|X}$ (the $\text{BSC}(\delta)$ in this case), now generate $U\in\{0,1\}$ according to

$$U=\begin{cases}0, & \text{if } Y^{(1)}=X\\ \text{Bernoulli}(\alpha), & \text{otherwise}\end{cases} \tag{47}$$

where $0\le\alpha\le 1$ is a parameter. Now let

$$\begin{aligned}\tilde{X}&=f(U,Y^{(1)},X)\\ &=\begin{cases}\text{no rewrite}, & \text{if } Y^{(1)}=X \text{ or } U=1\\ X, & \text{otherwise}\end{cases}\end{aligned} \tag{48}$$

and $Y^{(2)}$ is the output of the rewrite channel $P_{Y^{(2)}|Y^{(1)},\tilde{X}}$. Under this joint distribution, simple computations give

$$\begin{aligned}H(Y^{(2)})&=1,\\ H(Y^{(1)}|X)&=h_2(\delta),\\ H(Y^{(1)}|X,U)&=h_2\left(\delta\frac{1-\alpha}{1-\delta\alpha}\right)(1-\delta\alpha),\\ H(Y^{(2)}|X,U)&=h_2\left(\delta^2\frac{1-\alpha}{1-\delta\alpha}\right)(1-\delta\alpha)\end{aligned} \tag{49}$$

so that [see (50) and (51) at the bottom of the page]. As is to be expected, when $\alpha=0$, U is degenerate, so we recover a joint distribution of the type allowed and achieving the maximum in the causal case, and indeed the expression in (51) becomes $1-h_2(\delta^2)$ when $\alpha=0$. However, we may now optimize over α to obtain the following lower bound on the capacity of the BSC with a rewrite option for the noncausal case: See (52) at the bottom of the page. To see that this capacity can be strictly higher than its counterpart for the causality-constrained scenario, consider the case $\delta=1/2$. A straightforward calculation shows that in this case (52) assumes the value ≈ 0.207519 ($\alpha=1/3$ achieves the maximum), as compared to $1-h_2(1/4)\approx 0.188722$, which is the capacity under the causality constraint. Thus, for the BSC $(1/2)$, relaxation of the causality constraint boosts the rewrite capacity by at least 10%. We emphasize that the expression in (52) is merely a lower bound which we have no reason to believe is tight. Our choices of a binary U, in particular the one defined in (47), and then the particular mapping f in (48), were made for simplicity and are rather arbitrary, but are good enough to yield a lower bound strictly better (higher) than the capacity under the causality constraint.

B. Channels With a Rewrite Option Based on Noisy Feedback

We now generalize the setting of the previous subsection to the more realistic scenario where the rewrite decision is based on a *noisy* observation of the channel outputs from the first pass. The forward channel is, as before, a DMC $P_{Y|X}$, while the channel outputs from the first pass are observed by the rewrite encoder through a DMC $P_{Z|Y}$, the "backward" channel. What is the rewrite capacity in this noisy setting?

This we cast into the generic framework via the following associations: The role of the action sequence is played by the first channel input, which we denote here by X, taking values in the alphabet of possible inputs to the forward channel $P_{Y|X}$, namely in $\mathcal{X}$. The role of the channel state information is played by the channel output after the first pass (which we denote here by $Y^{(1)}$ as in the noise-free case) corrupted by the backward channel $P_{Z|Y}$, so we naturally denote it here by Z, taking values in the alphabet of the output of the channel $P_{Z|Y}$, namely in $\mathcal{Z}$. The channel input in the second stage we denote here by $\tilde{X}$ (playing the role of X in our generic framework), which takes values in the alphabet $\tilde{\mathcal{X}}=\{\text{norewrite}\}\cup\mathcal{X}$. We denote the

$$I(X,U;Y^{(2)})-I(U;Y^{(1)}|X)=H(Y^{(2)})-H(Y^{(2)}|X,U)-H(Y^{(1)}|X)+H(Y^{(1)}|X,U) \tag{50}$$

$$=1-h_2(\delta)+\left[h_2\left(\delta\frac{1-\alpha}{1-\delta\alpha}\right)-h_2\left(\delta^2\frac{1-\alpha}{1-\delta\alpha}\right)\right](1-\delta\alpha). \tag{51}$$

$$1-h_2(\delta)+\max_{0\le\alpha\le 1}\left\{\left[h_2\left(\delta\frac{1-\alpha}{1-\delta\alpha}\right)-h_2\left(\delta^2\frac{1-\alpha}{1-\delta\alpha}\right)\right](1-\delta\alpha)\right\}. \tag{52}$$

final channel output by $Y^{(2)}$, the superscript (2) pertaining to it being the channel output after the second pass, as in the case of noise-free feedback. It too takes values in the alphabet of the output of the channel $P_{Y|X}$, namely in $\mathcal{Y}$. As in the case of noise-free feedback, $Y^{(2)}$ plays the role of Y from our generic framework.

The role of $P_{S|A}$ is thus played here by $P_{Z|X}$, where $P_{Z|X}(z|x) = \sum_y P_{Y|X}(y|x)P_{Z|Y}(z|y)$ and $P_{Y|X,S,A}$ is played by $P_{Y^{(2)}|\tilde{X},Z,X}$, given as

$$P_{Y^{(2)}|\tilde{X},Z,X}(y^{(2)}|\tilde{x},z,x) = \begin{cases} P_{Y|X,Z}(y^{(2)}|x,z), & \text{if } \tilde{x} = \text{no rewrite} \\ P_{Y|X}(y^{(2)}|\tilde{x}), & \text{otherwise} \end{cases} \tag{53}$$

where $P_{Y|X,Z}$ is induced by $P_{Y|X}$ and $P_{Z|Y}$, namely

$$P_{Y|X,Z}(y|x,z) = \frac{P_{Y|X}(y|x)P_{Z|Y}(z|y)}{\sum_{y'} P_{Y|X}(y'|x)P_{Z|Y}(z|y')}. \tag{54}$$

Applying Theorem 2, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass depend causally on the channel output components from the first pass is given by

$$\max I(U;Y^{(2)}) \tag{55}$$

where $U,\ X,\ Y^{(1)},\ Z,\ \tilde{X},\ Y^{(2)}$ are distributed according to

$$\begin{aligned} &P_{U,X,Y^{(1)},Z,\tilde{X},Y^{(2)}}\left(u,x,y^{(1)},z,\tilde{x},y^{(2)}\right) \\ &\quad = P_U(u)1_{\{g(u)=x\}}P_{Y|X}\left(y^{(1)}|x\right)P_{Z|Y}\left(z|y^{(1)}\right) \\ &\qquad \times 1_{\{f(u,z)=\tilde{x}\}}P_{Y^{(2)}|\tilde{X},Z,X}\left(y^{(2)}|\tilde{x},z,x\right) \end{aligned} \tag{56}$$

with $P_{Y^{(2)}|\tilde{X},Z,X}$ given in (53), and where the maximization is over $P_U,\ g,\ f$, with $|\mathcal{U}| \le \min\{|\mathcal{Y}|, |\mathcal{X}|(|\mathcal{X}|+1)|\mathcal{Z}|+1\}$. In other words, capacity is achieved by coding for a memoryless channel whose input alphabet is $\mathcal{X} \times \{f : \mathcal{Y} \to \tilde{\mathcal{X}}\}$, output alphabet is $\mathcal{Y}$, and where the probability of a channel output symbol y given input "symbol" $(x, f(\cdot))$ is the probability that the symbol y is eventually observed at the output of the channel when the symbol x is first input into it and then $f(\cdot)$, evaluated at the noisy measurement of the channel output after the first input, is used to determine whether that location will be rewritten to and, if so, what the new input symbol will be.

Applying Theorem 1, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass may depend noncausally on the channel output from the first pass is given by

$$\max[I(X,U;Y^{(2)}) - I(U;Z|X)] \tag{57}$$

where the maximization is over all joint distributions of the form

$$\begin{aligned} &P_{X,Y^{(1)},Z,U,\tilde{X},Y^{(2)}}\left(x,y^{(1)},u,\tilde{x},y^{(2)}\right) \\ &\quad = P_X(x)P_{Y|X}\left(y^{(1)}|x\right)P_{Z|Y}\left(z|y^{(1)}\right) \\ &\qquad \times P_{U|Z,X}(u|z,x)1_{\{\tilde{x}=f(u,z,x)\}} \\ &\qquad \times P_{Y^{(2)}|\tilde{X},Z,X}\left(y^{(2)}|\tilde{x},z,x\right) \end{aligned} \tag{58}$$

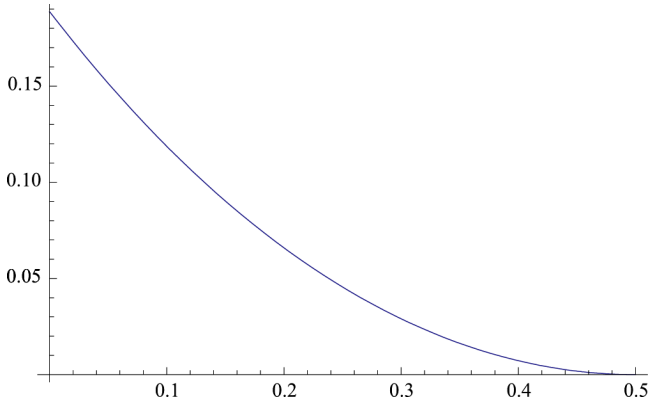

Fig. 2. Capacity of $\mathrm{BSC}(\delta)$ (in bits per channel use) with rewrite based on $\mathrm{BSC}(\varepsilon)$-corrupted feedback from the first pass, when the rewrite is restricted to causal dependence on the channel outputs from the first pass. Plotted here for $\delta = 1/2$, as a function of ε.

with $P_{Y^{(2)}|\tilde{X},Z,X}$ given in (53), and where the maximization is over $P_X, P_{U|Z,X}, f$, with $|\mathcal{U}| \le |\mathcal{X}|(|\mathcal{X}|+1)|\mathcal{Z}|+1$.

Example II: BSC: Consider the case where both the forward and the backward channels are BSCs with respective parameters $0 \le \delta \le 1/2$ and $0 \le \varepsilon \le 1/2$. For the case where the rewrite operations in the second pass depend *causally* on the (noisy) observations of the channel output components from the first pass, the arguments in Section V-A-I carry over to this noisy scenario and imply that the capacity achieving triple (U, g, f) for the causal case is the same as that for the noise-free backward channel. Under this triple, a simple calculation shows that the channel from X to $Y^{(2)}$ is a BSC with crossover probability $\varepsilon\delta(2-\delta) + \delta^2(1-\varepsilon)$, so the resulting capacity is

$$1 - h_2\left(\varepsilon\delta(2-\delta) + \delta^2(1-\varepsilon)\right) \tag{59}$$

which is plotted in Fig. 2. Note, in particular, the two extremes: when $\varepsilon = 0$, we recover the $1 - h_2(\delta^2)$ of Section V-A-1 while for $\varepsilon = 1/2$, (59) becomes $1 - h_2(\delta)$, capacity of the vanilla $\mathrm{BSC}(\delta)$.

Moving to the case where the rewrite operations may depend noncausally on the channel output from the first pass, consider a joint distribution of the type allowed in (58), as follows: $X \sim \mathrm{Bernoulli}(1/2)$, $Y^{(1)}$ is the output of $P_{Y|X}$ (the $\mathrm{BSC}(\delta)$ in this case), Z is the output of $P_{Z|Y^{(1)}}$ (the $\mathrm{BSC}(\varepsilon)$ in this case), now generate $U \in \{0,1\}$ according to

$$U = \begin{cases} 0, & \text{if } Z = X \\ \mathrm{Bernoulli}(\alpha), & \text{otherwise} \end{cases} \tag{60}$$

where $0 \le \alpha \le 1$ is a parameter. Let now[6]

$$\begin{aligned} \tilde{X} &= f(U,Z,X) \\ &= \begin{cases} \text{no rewrite}, & \text{if } Z = X \text{ or } U = 1 \\ X, & \text{otherwise} \end{cases} \end{aligned} \tag{61}$$

and $Y^{(2)}$ is the output of the rewrite channel $P_{Y^{(2)}|\tilde{X},Z,X}$ in (53) which in this case is given by (62) at the bottom of the next page, where [see (63) at the bottom of the next page] and $*$ denotes binary convolution defined by $\varepsilon * \delta = \varepsilon(1-\delta) + \delta(1-\varepsilon)$. The above mappings and conditional distributions further induce the following joint and conditional PMFs: See (64) and (65) at the

[6]As in Section V-A, our choice of f here is rather arbitrary.

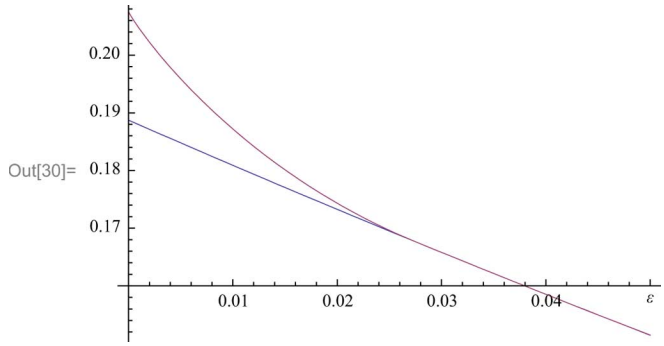

Fig. 3. The upper curve is the lower bound on the capacity (in bits per channel use) when there is no restriction to causality, as given in (68). The lower curve is the actual capacity of a $\mathrm{BSC}(1/2)$ with a rewrite option based on observation of the channel outputs from the first pass through a $\mathrm{BSC}(\varepsilon)$, when the rewrite is restricted to causality, as given in (59). The two curves touch the Y axis at the points mentioned in Section V-A, $\mathrm{namely}\ \approx 0.207519$ ($\alpha = 1/3$ achieves the maximum in (68) for this case), as compared to $1 - h_2(1/4) \approx 0.188722$.

bottom of the page. Equipped with the PMFs in (64) and (65), the entropies are readily obtained as (66), shown at the bottom of the next page, so that [see (67) at the bottom of the next page]. Note that when $\alpha = 0$, (67) becomes the third equation on the next page, i.e., it coincides with (59). On the other hand, when $\varepsilon = 0$, (67) is readily seen to coincide with (51).

Optimizing over α, we obtain the following lower bound on the capacity of the BSC with a rewrite option based on a noisy observation of the channel output, for the noncausal case (see (68) at the bottom of the next page). Fig. 3 presents a plot of the expression in (68), and one of the capacity when the rewrite is restricted to causality, when the forward channel is a $\mathrm{BSC}(1/2)$ and the backward one is $\mathrm{BSC}(\varepsilon)$.

C. Computer Memory With Defects and a Rewrite Option

Consider a computer memory with defects, characterized by the distribution of the state of each cell, P_S, and the channel $P_{Y|X,S}$. Consider the following two-pass coding scenario: the memory state is known neither to the encoder nor the decoder. After writing into the storage device and observing the channel output components, the encoder makes another encoding pass where it may rewrite at whichever memory locations it chooses. At each memory location, the state remains unchanged regardless of whether or not a rewrite was performed. What is the storage capacity of such a two-pass coding device?

$$P_{Y^{(2)}|\tilde{X},Z,X}\left(y^{(2)}|\tilde{x},z,x\right) = \begin{cases} P_{Y|X,Z}\left(y^{(2)}|x,z\right), & \text{if } \tilde{x} = \text{ no rewrite} \\ 1_{\{y^{(2)}=\tilde{x}\}}(1-\delta) + 1_{\{y^{(2)}\neq\tilde{x}\}}\delta, & \text{otherwise} \end{cases} \tag{62}$$

$$P_{Y|X,Z}(y|x,z) = \begin{cases} \frac{(1-\delta)(1-\varepsilon)}{1-\delta*\varepsilon}, & \text{if } (y,x,z) = (0,0,0) \text{ or } (y,x,z) = (1,1,1) \\ 1 - \frac{(1-\delta)(1-\varepsilon)}{1-\delta*\varepsilon}, & \text{if } (y,x,z) = (1,0,0) \text{ or } (y,x,z) = (0,1,1) \\ \frac{(1-\delta)\varepsilon}{\delta*\varepsilon}, & \text{if } (y,x,z) = (0,0,1) \text{ or } (y,x,z) = (1,1,0) \\ 1 - \frac{(1-\delta)\varepsilon}{\delta*\varepsilon}, & \text{if } (y,x,z) = (1,0,1) \text{ or } (y,x,z) = (0,1,0) \end{cases} \tag{63}$$

$$P_{X,Z,U}(x,z,u) = \begin{cases} \frac{1}{2}(1-\delta*\varepsilon), & \text{if } (x,z,u) = (0,0,0) \text{ or } (x,z,u) = (1,1,0) \\ \frac{1}{2}(\delta*\varepsilon)(1-\alpha), & \text{if } (x,z,u) = (0,1,0) \text{ or } (x,z,u) = (1,0,0) \\ \frac{1}{2}(\delta*\varepsilon)\alpha, & \text{if } (x,z,u) = (0,1,1) \text{ or } (x,z,u) = (1,0,1) \\ 0, & \text{if } (x,z,u) = (0,0,1) \text{ or } (x,z,u) = (1,1,1)\,. \end{cases} \tag{64}$$

$$\begin{aligned} P_{Y^{(2)}|X,U}\left(y^{(2)}|x,u\right) &= \sum_z P_{Y^{(2)},Z|X,U}\left(y^{(2)},z|x,u\right) \\ &= \sum_z P_{Z|X,U}(z|x,u) P_{Y^{(2)}|U,Z,X}\left(y^{(2)}|u,z,x\right) \\ &= \sum_z P_{Z|X,U}(z|x,u) P_{Y^{(2)}|\tilde{X},Z,X}\left(y^{(2)}|f(u,z,x),z,x\right) \\ &= \begin{cases} \frac{(1-\delta*\varepsilon)}{(1-\delta*\varepsilon)+(\delta*\varepsilon)(1-\alpha)} \frac{(1-\delta)(1-\varepsilon)}{(1-\delta*\varepsilon)} + \frac{(\delta*\varepsilon)(1-\alpha)}{(1-\delta*\varepsilon)+(\delta*\varepsilon)(1-\alpha)}(1-\delta), & \text{if } \left(y^{(2)},x,u\right) = (0,0,0) \\ \frac{(1-\delta)\varepsilon}{\delta*\varepsilon}, & \text{if } \left(y^{(2)},x,u\right) = (0,0,1)\,. \end{cases} \end{aligned} \tag{65}$$

This can be cast into our framework via the following associations: the role of the action is played by the first channel input, the role of the state by the channel output after the first pass $Y^{(1)}$, the role of the channel input played by the choice of encoding in the second pass (thus taking values in $\{\text{norewrite}\} \cup \mathcal{X}$) and the role of the channel output played by the channel output after the second pass, denoted here by $Y^{(2)}$. Thus $P_{S|A}$ is played here by $P_{Y^{(1)}|X}$, where $P_{Y^{(1)}|X}$ is the $P_{Y|X}$ induced by the original channel, i.e.

$$P_{Y^{(1)}|X}(y|x) = \sum_s P_S(s) P_{Y|X,S}(y|x,s). \tag{69}$$

$P_{Y|X,S,A}$ is played by $P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}$, where [see (70), shown at the bottom of the next page] with $P_{S|X,Y}$ being the posterior

$$\begin{aligned}
H(Y^{(2)}) &= 1, \\
H(Z|X) &= h_2(\delta * \varepsilon), \\
H(Z|X,U) &= h_2\left(\frac{1-\delta*\varepsilon}{1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)}\right)[1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)], \\
H\left(Y^{(2)}|X,U\right) &= h_2\left(\frac{(1-\delta)(1-\varepsilon)+(\delta*\varepsilon)(1-\alpha)(1-\delta)}{1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)}\right) \\
&\quad \times [1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)] + h_2\left(\frac{(1-\delta)\varepsilon}{\delta*\varepsilon}\right)(\delta*\varepsilon)\alpha
\end{aligned} \tag{66}$$

$$\begin{aligned}
I\left(X,U;Y^{(2)}\right) - I(U;Z|X) &= H\left(Y^{(2)}\right) - H\left(Y^{(2)}|X,U\right) - H(Z|X) + H(Z|X,U) \\
&= 1 - h_2(\delta*\varepsilon) \\
&\quad + \left[h_2\left(\frac{1-\delta*\varepsilon}{1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)}\right) - h_2\left(\frac{(1-\delta)(1-\varepsilon)+(\delta*\varepsilon)(1-\alpha)(1-\delta)}{1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)}\right)\right] \\
&\quad \times [1-\delta*\varepsilon+(\delta*\varepsilon)(1-\alpha)] - h_2\left(\frac{(1-\delta)\varepsilon}{\delta*\varepsilon}\right)(\delta*\varepsilon)\alpha. \\
&= 1 - h_2(\delta*\varepsilon) + \left[h_2\left(\frac{1-\delta*\varepsilon}{1-(\delta*\varepsilon)\alpha}\right) - h_2\left(\frac{(1-\delta)(1-\varepsilon)+(\delta*\varepsilon)(1-\alpha)(1-\delta)}{1-(\delta*\varepsilon)\alpha}\right)\right] \\
&\quad \times [1-(\delta*\varepsilon)\alpha] - h_2\left(\frac{(1-\delta)\varepsilon}{\delta*\varepsilon}\right)(\delta*\varepsilon)\alpha.
\end{aligned} \tag{67}$$

$$\begin{aligned}
1 - h_2\left((1-\delta)(1-\varepsilon)+(\delta*\varepsilon)(1-\delta)\right) &= 1 - h_2\left(1-2\varepsilon\delta+2\varepsilon\delta^2-\delta^2\right) \\
&= 1 - h_2\left(1-2\varepsilon\delta+\varepsilon\delta^2-\delta^2+\varepsilon\delta^2\right) \\
&= 1 - h_2\left(1-[\varepsilon\delta(2-\delta)+\delta^2(1-\varepsilon)]\right) \\
&= 1 - h_2\left(\varepsilon\delta(2-\delta)+\delta^2(1-\varepsilon)\right)
\end{aligned}$$

$$\begin{aligned}
1 - h_2(\delta*\varepsilon) + \max_{0\le\alpha\le 1}\Bigg\{ &\left[h_2\left(\frac{1-\delta*\varepsilon}{1-(\delta*\varepsilon)\alpha}\right) - h_2\left(\frac{(1-\delta)(1-\varepsilon)+(\delta*\varepsilon)(1-\alpha)(1-\delta)}{1-(\delta*\varepsilon)\alpha}\right)\right] \\
&\times [1-(\delta*\varepsilon)\alpha] - h_2\left(\frac{(1-\delta)\varepsilon}{\delta*\varepsilon}\right)(\delta*\varepsilon)\alpha\Bigg\}.
\end{aligned} \tag{68}$$

distribution on the state given knowledge of the channel input *and* output, as induced by the original channel, namely

$$P_{S|X,Y}(s|x,y) = \frac{P_{S,Y|X}(s,y|x)}{P_{Y|X}(y|x)} = \frac{P_S(s)P_{Y|X,S}(y|x,s)}{\sum_{s'} P_S(s')P_{Y|X,S}(y|x,s')}. \quad (71)$$

Note that $Y^{(1)}$ is affected by the encoding "action" chosen for the first pass, and is then observed by the encoder before choosing its channel input symbol for the second pass. Further, knowledge of $Y^{(1)}$ conveys information about the state S, and thus affects the conditional distribution of the channel output if a rewrite operation is selected, so it is playing the role of the channel state when cast into our general setting.

Applying Theorem 2, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass depend causally on the channel output components from the first pass C_C^{CMDRW},[7] is given by

$$C_C^{CMDRW} = \max I(U;Y^{(2)}) \quad (72)$$

where $U,\ X,\ Y^{(1)},\ \tilde{X},\ Y^{(2)}$ are distributed according to

$$P_{U,X,Y^{(1)},\tilde{X},Y^{(2)}}\left(u,x,y^{(1)},\tilde{x},y^{(2)}\right) = P_U(u)1_{\{g(u)=x\}}P_{Y^{(1)}|X}\left(y^{(1)}|x\right) \times 1_{\left\{f\left(u,y^{(1)}\right)=\tilde{x}\right\}}P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}\left(y^{(2)}|x,y^{(1)},\tilde{x}\right) \quad (73)$$

with $P_{Y^{(1)}|X}$ and $P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}$ given in (69) and (70), and where the maximization is over P_U, g, f, with $|\mathcal{U}| \le |\mathcal{Y}|$.

Applying Theorem 1, with the above associations, we get that the capacity for the case where the rewrite operations in the second pass may depend noncausally on the channel output components from the first pass C_{NC}^{CMDRW},[8] is given by

$$C_{NC}^{CMDRW} = \max\left[I\left(X,U;Y^{(2)}\right) - I\left(U;Y^{(1)}|X\right)\right] \quad (74)$$

where the maximization is over all joint distributions of the form

$$P_{X,Y^{(1)},U,\tilde{X},Y^{(2)}}\left(x,y^{(1)},u,\tilde{x},y^{(2)}\right) = P_X(x)P_{Y^{(1)}|X}\left(y^{(1)}|x\right) \times P_{U|Y^{(1)},X}\left(u|y^{(1)},x\right)1_{\left\{\tilde{x}=f\left(u,y^{(1)},x\right)\right\}} \times P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}\left(y^{(2)}|x,y^{(1)},\tilde{x}\right) \quad (75)$$

[7]The superscript in C_C^{CMDRW} standing for "Computer Memory with Defects and a ReWrite option" while the subscript stands for "Causal."

[8]The subscript in C_{NC}^{CMDRW} standing for "Noncausal."

with $P_{Y^{(1)}|X}$ and $P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}$ given in (69) and (70), and where the maximization is over $P_X, P_{U|Y^{(1)},X}, f$, with $|\mathcal{U}| \le |\mathcal{X}|(|\mathcal{X}|+1)|\mathcal{Y}|+1$.

VI. The Gaussian Channel

Using standard arguments, the capacity results of the previous sections can be shown to carry over to continuous-alphabet channels, similarly as for the original problems of coding with transmitter state information, such as in [1]. In this section, we consider the "writing-on-clean-paper-and-then-writing-on-its-corrupted version" channel, which has the following relations between channel inputs, channel outputs, states and actions:

$$Y^n = S^n + X^n(M,S^n) + N^n = A^n(M) + W^n + X^n(M,S^n) + N^n \quad (76)$$

where

- $S^n = A^n(M) + W^n$
- W^n and N^n are independent, W^n is i.i.d. $\sim N(0,\sigma_W^2)$ and N^n is i.i.d. $\sim N(0,\sigma_N^2)$
- The actions are confined to

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(A_i)^2\right] \le P_A$$

- The subsequent channel inputs are confined to

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i)^2\right] \le P_X.$$

The continuous-alphabet extension of Theorem 3 implies that the capacity of this channel is given by

$$\max[I(A,U;Y) - I(U;S|A)] \quad (77)$$

where the maximization is over all jointly distributed variables obeying:

- $W \sim N(0,\sigma_W^2)$, $N \sim N(0,\sigma_N^2)$, $W \perp N$
- $S = A + W$, $A \perp W$
- X is a function of $U,\ S,\ A$
- $Y = S + X + N$, $N \perp (A,W,U,X)$
- $E[A^2] \le P_A$, $E[X^2] \le P_X$

Let $C_G = C_G(P_A, P_X, \sigma_W^2, \sigma_N^2)$ denote the maximum in (77) subject to the above constraints, and under the additional requirement that $(A,\ U,\ X,\ S,\ Y)$ be jointly Gaussian.[9] Specifically, consider the joint distribution formed by taking:

- $A \sim N(0,P_A)$

[9]Note that C_G is a lower bound on the capacity in this problem setting. The question of its tightness, as mentioned below, is left open.

$$P_{Y^{(2)}|X,Y^{(1)},\tilde{X}}\left(y^{(2)}|x,y^{(1)},\tilde{x}\right) = \begin{cases} 1_{\{y^{(1)}=y^{(2)}\}}, & \text{if } \tilde{x} = \text{no rewrite} \\ \sum_s P_{Y|X,S}\left(y^{(2)}|\tilde{x},s\right)P_{S|X,Y}\left(s|x,y^{(1)}\right), & \text{otherwise} \end{cases} \quad (70)$$

- $X = \alpha A + \gamma W + G$, where $\alpha^2 P_A + \gamma^2 \sigma_W^2 \leq P_X$, $G \sim N(0, P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2))$
- $U = \delta X + A + \beta W$
- W, N, A, G independent of each other

It is readily verified that this joint distribution obeys the required conditions (note that $U - (X, S, A) - Y$ holds since U is a function of X, S, A) and that, in fact, this is essentially (up to rescaling of variables that would not affect the expression whose maximum we seek) the most general form of the joint distribution subject to the variables being jointly Gaussian. We proceed to evaluate $I(A,U;Y) - I(U;S|A)$ for this case, and then to obtain C_G by optimizing over α, β, γ and δ. The variances are

$$\sigma_A^2 = P_A \tag{78}$$

$$\sigma_Y^2 = (1+\alpha)^2 P_A + (1+\gamma)^2 \sigma_W^2 + P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2) + \sigma_N^2 \tag{79}$$

$$\sigma_U^2 = (1+\alpha\delta)^2 P_A + (\gamma\delta + \beta)^2 \sigma_W^2 + \delta^2 (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)). \tag{80}$$

The covariances are

$$E[AY] = (1+\alpha) P_A \tag{81}$$

$$E[UA] = (1+\alpha\delta) P_A \tag{82}$$

$$E[UY] = (1+\alpha)(1+\alpha\delta) P_A + (1+\gamma)(\beta+\gamma\delta)\sigma_W^2 + \delta (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)). \tag{83}$$

The conditional variances can now be computed as (84) and (85) at the bottom of the page and

$$\sigma_{U|A,S}^2 = \delta^2 \sigma_G^2 = \delta^2 (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)). \tag{86}$$

Thus, see (87) and (88) at the bottom of the page, where σ_A^2, $\sigma_{U|A,S}^2$, $\sigma_{A|Y}^2$ and $\sigma_{U|A,Y}^2$ are given explicitly in terms of P_A, P_X,, σ_W^2, σ_N^2, α, β, γ and δ via (78) through (86).

We proceed to maximize the expression in (88) with respect to α, β, γ and δ. To this end, note first that σ_A^2, $\sigma_{U|A,S}^2$, $\sigma_{A|Y}^2$ do not depend on β. As for $\sigma_{U|A,Y}^2$, substituting the relations in (78) through (83) into (85) gives a quadratic expression in β which is minimized by $\beta^* = \delta \frac{\alpha^2 P_A - P_X + \gamma\sigma_N^2 + \gamma^2 \sigma_W^2}{\alpha^2 P_A - P_X - \sigma_N^2 + \gamma^2 \sigma_W^2}$ with the value

$$\min_\beta \sigma_{U|A,Y}^2 = \delta^2 \frac{\sigma_N^2 (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2))}{\sigma_N^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2))}. \tag{89}$$

Substituting the above expressions we obtain (90)–(93) at the bottom of the page, where in the last line we have taken, without loss of generality, $\sigma_N^2 = 1$. We thus obtain (94) and (95) at the bottom of the next page.

$$\sigma_{A|Y}^2 = \sigma_A^2 - \frac{(E[AY])^2}{\sigma_Y^2} = P_A - \frac{(1+\alpha)^2 (P_A)^2}{(1+\alpha)^2 P_A + (1+\gamma)^2 \sigma_W^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)) + \sigma_N^2} \tag{84}$$

$$\sigma_{U|A,Y}^2 = \sigma_U^2 - \frac{-2E[AY]E[UA]E[UY] + (E[UY])^2 \sigma_A^2 + (E[UA])^2 \sigma_Y^2}{-(E[AY])^2 + \sigma_A^2 \sigma_Y^2} \tag{85}$$

$$I(A,U;Y) - I(U;S|A) = h(A) - h(A|Y) - h(U|A,Y) + h(U|A,S) \tag{87}$$

$$= \frac{1}{2} \log \left(\frac{\sigma_A^2 \sigma_{U|A,S}^2}{\sigma_{A|Y}^2 \sigma_{U|A,Y}^2} \right) \tag{88}$$

$$\max_\beta \frac{\sigma_A^2 \sigma_{U|A,S}^2}{\sigma_{A|Y}^2 \sigma_{U|A,Y}^2} = \frac{P_A (\sigma_N^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)))}{\sigma_N^2 \left(P_A - \frac{(1+\alpha)^2 (P_A)^2}{(1+\alpha)^2 P_A + (1+\gamma)^2 \sigma_W^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)) + \sigma_N^2} \right)} \tag{90}$$

$$= \frac{((1+2\alpha) P_A + P_X + \sigma_N^2 + (1+2\gamma)\sigma_W^2)(\sigma_N^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)))}{\sigma_N^2 (P_X - \alpha^2 P_A + \sigma_N^2 + \sigma_W^2 (1+2\gamma))} \tag{91}$$

$$= \frac{P_A (\sigma_N^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)))}{\sigma_N^2 \left(P_A - \frac{(1+\alpha)^2 (P_A)^2}{(1+\alpha)^2 P_A + (1+\gamma)^2 \sigma_W^2 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2)) + \sigma_N^2} \right)} \tag{92}$$

$$= \frac{[(1+2\alpha) P_A + P_X + 1 + (1+2\gamma)\sigma_W^2][1 + (P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2))]}{[P_X - \alpha^2 P_A + 1 + \sigma_W^2 (1+2\gamma)]} \tag{93}$$

Keeping the representation (6) in mind, we note that C_G is achievable by using the actions for communicating one message (using a Gaussian-generated code-book), and then communicating an independent message in the second-stage encoding using dirty paper coding. It is instructive to compare C_G to the following lower bounds on the capacity of this channel:

- Precancelation via X:
  - When $P_X \geq \sigma_W^2$, the encoder of the sequence X^n has access to W^n (as it knows S^n and M and hence also $A^n(M)$), and so can use part of its power to cancel W^n, and the remaining power to amplify the action sequence A^n. The rate achieved by this strategy is

$$\frac{1}{2}\log\left(1+\frac{\left(1+\sqrt{\frac{P_X-\sigma_W^2}{P_A}}\right)^2\cdot P_A}{\sigma_N^2}\right). \tag{96}$$

    Note that this corresponds to taking $\alpha=\sqrt{\frac{P_X-\sigma_W^2}{P_A}}$ and $\gamma=-1$ in lieu of the maximum in (95).
  - When $P_X < \sigma_W^2$, the encoder of the sequence X^n can use all of its power to cancel as much of W^n as it can. The encoding is done solely through the action sequence, and an effective noise power equal to σ_N^2 plus the part of W^n that has not been canceled. The rate achieved by this strategy is

$$\frac{1}{2}\log\left(1+\frac{P_A}{\left(1-\sqrt{\frac{P_X}{\sigma_W^2}}\right)^2\sigma_W^2+\sigma_N^2}\right). \tag{97}$$

    Note that this corresponds to taking $\alpha=0$ and $\gamma=-\sqrt{P_X/\sigma_W^2}$.
- Both A and X can work together to encode for a standard AWGN channel with noise power $\sigma_W^2+\sigma_N^2$. The rate achieved is

$$\frac{1}{2}\log\left(1+\frac{\left(1+\sqrt{\frac{P_X}{P_A}}\right)^2 P_A}{\sigma_W^2+\sigma_N^2}\right). \tag{98}$$

  Note that this corresponds to taking $\alpha=\sqrt{P_X/P_A}$ and $\gamma=0$.
- The rate

$$\frac{1}{2}\log\left(1+\frac{P_X}{\sigma_N^2}\right) \tag{99}$$

  is always achievable, regardless of σ_W^2, by standard dirty paper coding (i.e., treating S^n as interference).

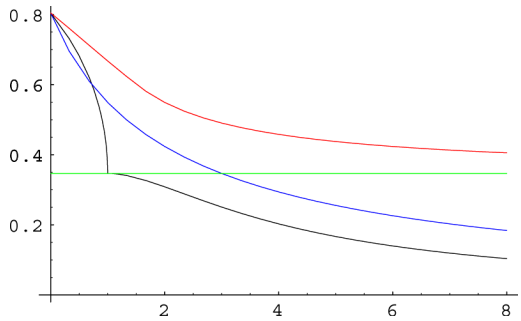

Fig. 4. Lower bounds on the capacity (in bits per channel use) of the Gaussian channel, for the case $P_A = P_X = \sigma_N^2 = 1$, as a function of σ_W^2. The upper (red) curve shows C_G. The blue curve (smooth one below that of C_G) shows the rate in (98), achieved when X and A cooperate to encode for a standard AWGN channel whose noise is $W+N$. The black curve (with the discontinuous derivative) shows the cancelation bounds in (96) and (97). The green line is the standard dirty paper rate in (99). As can be expected, the first three curves coincide at $\sigma_W^2 = 0$ and give the actual capacity for this case, which is the capacity of the AWGN channel at signal to noise ratio 4. At the other extreme, as should be expected, the red curve is approaching the green one for large σ_W^2 since the actual capacity approaches the dirty paper capacity in the limit of large σ_W^2. For the particular case where $P_X = P_A$, dirty paper coding achieves a better rate than precancelation when $P_X < \sigma_W^2$ as can be seen in the graph and by comparing (97) with (99). Whether the actual capacity is given by the red curve remains to be answered.

Fig. 4 displays a plot of C_G as a function of σ_W^2 for the case $P_A = P_X = \sigma_N^2 = 1$, as well as its lower bounds, the achievable rates in (96)–(99). Whether C_G is the capacity of this channel remains to be determined.

VII. Conclusions and Future Directions

We have extended the study of channels with states known at the transmitter to the case where the formation of the states is affected by actions taken at the encoder. The fundamental limits on reliable communication for such channels were characterized. It was seen that such a framework covers channels with a "rewrite" option based on (noiseless or noisy) feedback from the channel output in the first writing. Examples of such channels were explored in detail. Another practically motivated scenario covered by our framework involves an encoder that can choose to observe or not to observe the channel state. This scenario is studied in detail in [13].

At first glance it might seem like our problem setting can and should be extended in two ways. The first is in considering scenarios where the second stage encoding is based on a noisy rather than a noise-free version of the states. The second is in allowing for channels with outputs that depend not only on the state and the second-stage encoding, but also on the first-stage encoding, i.e., on the action sequence. Such "extensions," however, remain in the realm of our original problem formulation. The case where encoding in the second stage is based on a noisy observation S' of a state S through a channel $P_{S'|S}$ is covered

$$C_G = C_G(P_A, P_X, \sigma_W^2, 1) \tag{94}$$

$$= \frac{1}{2}\log\left(\max_{(\alpha,\gamma):\alpha^2 P_A+\gamma^2\sigma_W^2\leq P_X}\frac{[(1+2\alpha)P_A+P_X+1+(1+2\gamma)\sigma_W^2][1+(P_X-(\alpha^2 P_A+\gamma^2\sigma_W^2))]}{[P_X-\alpha^2 P_A+1+\sigma_W^2(1+2\gamma)]}\right). \tag{95}$$

by simply noting that the noisy state S' is itself a state and so the problem coincides with our original one upon considering the $P_{S'|A}$ and $P_{Y|X,S'}$ induced by $P_{S|A}$, $P_{Y|X,S}$, and $P_{S'|S}$. To see why our formulation accommodates channels of the form $P_{Y|X,S,A}$ note that the latter are reduced to $P_{Y|X,S'}$ upon letting $S' = (S, A)$ play the role of the state.

Some questions and directions that our work leaves open for future exploration are:

- Can actions of the form $A_i(M, S^{i-1})$ increase the capacity relative to actions that are not allowed to depend on past states, in the setting of Section II, where in the encoding of the channel input X^n noncausal dependence on the state sequence is allowed?
- For the Gaussian channel of Section VI, does a U jointly Gaussian with the remaining variables achieve the maximum in (77), i.e., is $C_G = C$?
- Extension of the channel model to more than two encoding stages: In some of the motivating examples given, it makes sense to consider problems where each memory location can be rewritten into more than once. In fact, for some of the scenarios considered, particularly the one involving noisy feedback from the channel output after the first writing stage, it makes sense to consider a setting where the encoder is allowed as many "rewrite" attempts as it desires.

ACKNOWLEDGMENT

The author is grateful for helpful discussions with H. Permuter, Y. Steinberg, and S. Verdú.

REFERENCES

[1] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, pp. 439–441, May 1983.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

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

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

[5] C. Heegard and A. E. Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inf. Theory*, vol. IT-29, pp. 731–739, Sep. 1983.

[6] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," *Found. Trends in Commun. Inf. Theory*, vol. 4, no. 6, 2007.

[7] A. V. Kuznetsov and B. S. Tsybakov, "Coding in a memory with defective cells," *Probl. Contr. and Inf. Theory*, vol. 10, no. 2, pp. 52–60, 1974.

[8] N. Merhav and T. Weissman, "Coding for the feedback Gel'fand-Pinsker channel and the feedforward Wyner-Ziv source," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4207–4211, Sep. 2006.

[9] H. Permuter and T. Weissman, "Source coding with a side information 'Vending machine' at the decoder," in *Proc. ISIT 09*, 2009.

[10] M. Salehi and , Cardinality Bounds on Auxiliary Variables in Multiple-User Theory via The Method of Ahlswede and Körner Dep. Statistics, Stanford Univ., Stanford, CA.

[11] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Dev.*, vol. 2, pp. 289–293, 1958.

[12] A. Somekh-Baruch, S. Shamai, and S. Verdú, "Cooperative multiple-access encoding with states available at one transmitter," *IEEE Trans. Inf. Theory*, vol. IT-54, pp. 4448–4469, Oct. 2008.

[13] H. Asnani, H. Permuter, and T. Weissman, "To observe or not to observe the channel state," in *Proc. 48th Annu. Allerton Conf. Commun.*, Sep. 2010.

Tsachy Weissman (S'99–M'02–SM'07) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Technion Israel Institute of Technology, Haifa, Israel.

During 2002 to 2003, he was with the Information Theory Research Group, HP Labs. He has been on the faculty of the Electrical Engineering Department, Stanford University, Stanford, CA, since 2003, spending the two academic years 2007 and 2009 with the Department of Electrical Engineering, Technion. He is the inventor of several patents and involved in a number of high-tech companies as a researcher or member of the technical board. His research is focused on information theory, statistical signal processing, the interplay between them, and their applications.

Dr. Weissman received a joint IT/COM Societies Best Paper award, a Horev Fellowship for Leaders in Science and Technology, and a Henry Taub Prize for Excellence in Research. He currently serves as Associate Editor for Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY.