State Amplification

Young-Han Kim, *Member, IEEE*, Arak Sutivong, and Thomas M. Cover, *Fellow, IEEE*

***Abstract*—We consider the problem of transmitting data at rate R over a state-dependent channel $p(y|x,s)$ with state information available at the sender and at the same time conveying the information about the channel state itself to the receiver. The amount of state information that can be learned at the receiver is captured by the mutual information $I(S^n;Y^n)$ between the state sequence S^n and the channel output Y^n. The optimal tradeoff is characterized between the information transmission rate R and the state uncertainty reduction rate Δ, when the state information is either causally or noncausally available at the sender. In particular, when state transmission is the only goal, the maximum uncertainty reduction rate is given by $\Delta^* = \max_{p(x|s)} I(X,S;Y)$. This result is closely related and in a sense dual to a recent study by Merhav and Shamai, which solves the problem of *masking* the state information from the receiver rather than conveying it.**

***Index Terms*—Capacity, causal state information, channels with state information, joint source–channel coding, noncausal state information, state amplification, state uncertainty reduction, writing on dirty paper.**

I. Introduction

A CHANNEL $p(y|x,s)$ with noncausal state information at the sender has capacity

$$C = \max_{p(u,x|s)} (I(U;Y) - I(U;S)) \tag{1}$$

as shown by Gelfand and Pinsker [13]. Transmitting at capacity, however, obscures the state information S^n as received by the receiver Y^n. In some instances we wish to convey the state information S^n itself, which could be time-varying fading parameters or an original image that we wish to enhance. For example, a stage actor with face S uses makeup X to communicate to the back row audience Y. Here X is used to enhance and exaggerate S rather than to communicate new information. Another motivation comes from cognitive radio systems [12], [22], [8], [17] with the additional assumption that the secondary user X^n communicates its own message and at the same time facilitates the transmission of the primary user's signal S^n. How should the transmitter communicate over the channel to "amplify" his knowledge of the state information to the receiver? What is the optimal tradeoff between state amplification and independent information transmission?

To answer these questions, we study the communication problem depicted in Fig. 1. Here the sender has access to the channel state sequence $S^n = (S_1, S_2, \ldots, S_n)$, independent and identically distributed (i.i.d.) according to $p(s)$, and wishes to transmit a message index $W \in [2^{nR}] := \{1, 2, \ldots, 2^{nR}\}$, independent of S^n, as well as to help the receiver reduce the uncertainty about the channel state in n uses of a state-dependent channel $(\mathcal{X} \times \mathcal{S}, p(y|x,s), \mathcal{Y})$. Based on the message W and the channel state S^n, the sender chooses $X^n(W, S^n)$ and transmits it across the channel. Upon observing the channel output Y^n, the receiver guesses $\hat{W} \in [2^{nR}]$ and forms a list $L_n(Y^n) \subseteq \mathcal{S}^n$ that contains likely candidates of the actual state sequence S^n.

Without any observation Y^n, the receiver would know only that the channel state S^n is one of $2^{nH(S)}$ typical sequences (with almost certainty) and we can say the uncertainty about S^n is $H(S^n)$. Now upon observing Y^n and forming a list $L_n(Y^n)$ of likely candidates for S^n, the receiver's list size is reduced from $nH(S)$ to $\log |L_n|$. Thus, we define the channel state uncertainty reduction rate to be

$$\Delta = \frac{1}{n}(H(S^n) - \log |L_n|) = H(S) - \frac{1}{n}\log |L_n|$$

as a natural measure for the amount of information the receiver learns about the channel state. In other words, the uncertainty reduction rate $\Delta \in [0, H(S)]$ captures the difference between the original channel state uncertainty and the residual state uncertainty after observing the channel output. Later, in Section III, we will draw a connection between the list size reduction and the conventional information measure $I(S^n;Y^n)$ that also captures the amount of information Y^n learns about S^n.

More formally, we define a $(2^{nR}, 2^{n\Delta}, n)$ code as the encoder map

$$X^n : [2^{nR}] \times \mathcal{S}^n \to \mathcal{X}^n$$

and decoder maps

$$\hat{W} : \mathcal{Y}^n \to [2^{nR}]$$
$$L_n : \mathcal{Y}^n \to 2^{\mathcal{S}^n}$$

with list size

$$|L_n| = 2^{n(H(S)-\Delta)}.$$

The probability of a message decoding error $P_{e,w}^{(n)}$ and the probability of a list decoding error $P_{e,s}^{(n)}$ are defined, respectively, as

$$P_{e,w}^{(n)} = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \Pr(\hat{W} \neq w | W = w)$$
$$P_{e,s}^{(n)} = \Pr(S^n \notin L_n(Y^n))$$

Manuscript received March 2, 2007; revised January 20, 2008. This work was supported in part by the National Science Foundation under Grants CCR-0311633 and CCF-0515303. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, June 2007.

Y.-H. Kim was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: yhk@ucsd.edu).

A. Sutivong was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with McKinsey & Company, Bangkok 10330, Thailand (e-mail: arak_sutivong@mckinsey.com).

T. M. Cover is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: cover@stanford.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2008.920242

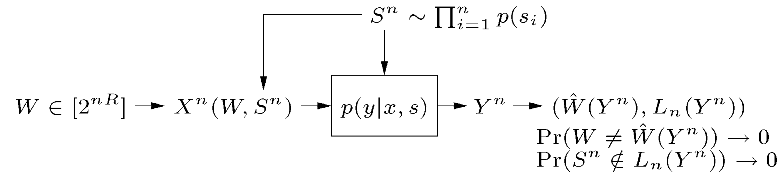

$S^n \sim \prod_{i=1}^n p(s_i)$
$W \in [2^{nR}]$
$X^n(W, S^n)$
$p(y|x,s)$
Y^n
$(\hat{W}(Y^n), L_n(Y^n))$
$\Pr(W \neq \hat{W}(Y^n)) \to 0$
$\Pr(S^n \notin L_n(Y^n)) \to 0$

Fig. 1. Pure information transmission versus state uncertainty reduction.

where the message index W is chosen uniformly over $[2^{nR}]$ and the state sequence S^n is drawn i.i.d. $\sim p(s)$, independent of W. A pair (R, Δ) is said to be achievable if there exists a sequence of $(2^{nR}, 2^{n\Delta}, n)$ codes with $P_{e,w}^{(n)} \to 0$ and $P_{e,s}^{(n)} \to 0$ as $n \to \infty$. Finally, we define the optimal (R, Δ) tradeoff region, or the *tradeoff region* in short, to be the closure of all achievable (R, Δ) pairs, and denote it by $\mathcal{R}^*$.

This paper shows that the tradeoff region $\mathcal{R}^*$ can be characterized as the union of all (R, Δ) pairs satisfying

$$\begin{aligned} R &\leq I(U;Y) - I(U;S) \\ \Delta &\leq H(S) \\ R + \Delta &\leq I(X,S;Y) \end{aligned}$$

for some joint distribution of the form $p(s)p(u,x|s)p(y|x,s)$.

As a special case, if the encoder's sole goal is to "amplify" the state information ($R = 0$), then the maximum uncertainty reduction rate

$$\Delta^* = \sup\{\Delta : (R, \Delta) \text{ is achievable for some } R \geq 0\}$$

is given by

$$\Delta^* = \min\{H(S), \max_{p(x|s)} I(X,S;Y)\}. \tag{2}$$

The maximum uncertainty reduction rate Δ^* is achieved by designing the signal X^n to enhance the receiver's estimation of the state S^n while using the remaining pure information-bearing freedom in X^n to provide more information about the state. More specifically, there are three different components involved in reducing the receiver's uncertainty about the state.

1) The transmitter uses the channel capacity to convey the state information. In Section II, we study the classical setup [19], [15] of coding for memory with defective cells (Example 1) and show that this "source–channel separation" scheme is optimal when the memory defects are symmetric.
2) The transmitter gets out of the way of the receiver's view of the state. For instance, the maximum uncertainty reduction for the binary multiplying channel $Y = X \cdot S$ (Example 2 in Section II) with binary input $X \in \{0,1\}$ and binary state $S \in \{0,1\}$ is achieved by sending $X \equiv 1$.
3) The transmitter actively amplifies the state. In Example 3 in Section III, we consider the Gaussian channel $Y = X + S + Z$ with Gaussian state S and Gaussian noise Z. Here the optimal transmitter amplifies the state as $X = \alpha S$ under the given power constraint $EX^2 \leq P$.

It is interesting to note that the maximum uncertainty reduction rate Δ^* is the information rate $I(X,S;Y)$ that could be achieved if both the state S and the signal X could be freely designed, instead of the state S being generated by nature. This rate also appears in the sum rate of the capacity region expression for the cooperative multiple-access channel [7, Problem 15.1] and the multiple-access channel with cribbing encoders by Willems and van der Meulen [32].

When the state information is only *causally* available at the transmitter, that is, when the channel input X_i depends on only the past and current channel state S^i, we will show that the tradeoff region $\mathcal{R}^*$ is given as the union of all (R, Δ) pairs satisfying

$$\begin{aligned} R &\leq I(U;Y) \\ \Delta &\leq H(S) \\ R + \Delta &\leq I(X,S;Y) \end{aligned}$$

for some joint distribution $p(s)p(u)p(x|u,s)p(y|x,s)$. Interestingly, the maximum uncertainty reduction rate Δ^* stays the same as in the noncausal case (2). Thus, causality incurs no cost on the (sum) rate which is again reminiscent of the multiple-access channel with cribbing encoders [32].

The problem of communication over state-dependent channels with state information known at the sender has attracted a great deal of attention. This research area was first pioneered by Shannon [27], Kuznetsov and Tsybakov [19], and Gel'fand and Pinsker [13]. Several advancements in both theory and practice have been made over the years. For instance, Heegard and El Gamal [15], [14] characterized the channel capacity and devised practical coding techniques for a computer memory with defective cells. Costa [5] studied the now famous "writing on dirty paper" problem and showed that the capacity of an additive white Gaussian noise channel is not affected by additional interference, as long as the entire interference sequence is available at the sender prior to the transmission. This fascinating result has been further extended with strong motivations from applications in digital watermarking (see, for example, Moulin and O'Sullivan [24], Chen and Wornell [3], and Cohen and Lapidoth [4]) and multiple-antenna broadcast channels (see, for example, Caire and Shamai [2], Weingarten, Steinberg, and Shamai [31], and Mohseni and Cioffi [23]). Readers are referred to Caire and Shamai [1], Lapidoth and Narayan [20], and Jafar [16] for more complete reviews on the theoretical development of the field. On the practical side, Erez, Shamai, and Zamir [10], [34] proposed efficient coding schemes based on lattice strategies for binning. More recently, Erez and ten Brink [11] report efficient coding techniques that almost achieve the capacity of Costa's dirty paper channel.

In [29], [30], we formulated the problem of simultaneously transmitting pure information and helping the receiver estimate the channel state under a distortion measure. Although the characterization of the optimal rate–distortion tradeoff is still open in general (cf. [28]), a complete solution is given for the Gaussian case (the writing on dirty paper channel) under quadratic distortion [29]. In this particular case, optimality was shown for a simple power-sharing scheme between pure information transmission via Costa's original coding scheme and state amplification via simple scaling.

Recently, Merhav and Shamai [21] considered a related problem of transmitting pure information, but this time under the additional requirement of *minimizing* the amount of information the receiver can learn about the channel state. In this interesting work, the optimal tradeoff between pure information rate R and the amount of state information E is characterized for both causal and noncausal setups. Furthermore, for the Gaussian noncausal case (writing on dirty paper), the optimal rate–distortion tradeoff is given under quadratic distortion. (This may well be called "writing dirty on paper.")

The current paper thus complements [21] in a dual manner. It is refreshing to note that our notion of uncertainty reduction rate Δ is essentially equivalent to Merhav and Shamai's notion of E; both notions capture the normalized mutual information $I(S^n; Y^n)$. (See the discussion in Section III.) The crucial difference is that Δ is to be maximized while E is to be minimized. Both problems admit single-letter optimal solutions.

The rest of this paper is organized as follows. In the next section, we establish the optimal (R, Δ) tradeoff region for the case in which the state information S^n is noncausally available at the transmitter before the actual communication. Section III extends the notion of state uncertainty reduction to continuous alphabets, by identifying the list decoding requirement $S^n \in L_n(Y^n)$ with the mutual information rate $\frac{1}{n}I(S^n; Y^n)$. In particular, we characterize the optimal (R, Δ) tradeoff region for Costa's "writing on dirty paper" channel. Since the intuition gained from the study of the noncausal setup carries over when the transmitter has causal knowledge of the state sequence, the causal case is treated only briefly in Section IV, followed by concluding remarks in Section V.

II. Optimal (R, Δ) Tradeoff: Noncausal Case

In this section, we characterize the optimal tradeoff region between the pure information rate R and the state uncertainty reduction rate Δ with state information noncausally available at the transmitter, as formulated in Section I.

Theorem 1: The tradeoff region $\mathcal{R}^*$ for a state-dependent channel $(\mathcal{X} \times \mathcal{S}, p(y|x,s), \mathcal{Y})$ with state information S^n noncausally known at the transmitter is the union of all (R, Δ) pairs satisfying

$$R \le I(U;Y) - I(U;S) \tag{3}$$

$$\Delta \le H(S) \tag{4}$$

$$R + \Delta \le I(X, S; Y) \tag{5}$$

for some joint distribution of the form $p(s)p(u,x|s)p(y|x,s)$, where the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \le |\mathcal{X}| \cdot |\mathcal{S}|$.

As will be clear from the proof of the converse, the region given by (3)–(5) is convex. (We can merge the time-sharing random variable into U.) Since the auxiliary random variable U affects the first inequality (3) only, the cardinality bound on $\mathcal{U}$ follows directly from the usual technique; see Gel'fand and Pinsker [13] or a general treatment by Salehi [26]. Finally, we can take X as a deterministic function of (U, S) without reducing the region, but at the cost of increasing the cardinality bound of U; refer to the proof of Lemma 2 below.

It is easy to see that we can recover the Gel'fand–Pinsker capacity formula

$$\begin{aligned} C &= \max\{R : (R, \Delta) \in \mathcal{R}^* \text{ for some } \Delta \ge 0\} \\ &= \max_{p(x,u|s)} (I(U;Y) - I(U;S)). \end{aligned}$$

For the other extreme case of pure state amplification, we have the following result.

Corollary 1: Under the condition of Theorem 1, the maximum uncertainty reduction rate

$$\Delta^* = \max\{\Delta : (R, \Delta) \in \mathcal{R}^* \text{ for some } R \ge 0\}$$

is given by

$$\Delta^* = \min\{H(S), \max_{p(x|s)} I(X, S; Y)\}. \tag{6}$$

Thus, the receiver can learn about the state S^n essentially at the maximal cut-set rate $I(X, S; Y)$.

Before we prove Theorem 1, we need the following two lemmas. The first one extends Fano's inequality [7, Lemma 7.9.1] to list decoding.

Lemma 1: For a sequence of list decoders $L_n : \mathcal{Y}^n \to 2^{\mathcal{S}^n}$, $Y^n \mapsto L_n(Y^n)$ with list size $|L_n|$ fixed for each n, let $P_{e,s}^{(n)} = \Pr(S^n \notin L_n(Y^n))$ be the sequence of corresponding probabilities of list decoding error. If $P_{e,s}^{(n)} \to 0$, then

$$H(S^n|Y^n) \le \log|L_n| + n\epsilon_n$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Proof: Define an error random variable E as

$$E = \begin{cases} 0, & \text{if } S^n \in L_n \\ 1, & \text{if } S^n \notin L_n. \end{cases}$$

We can then expand

$$\begin{aligned} H(E, S^n|Y^n) &= H(S^n|Y^n) + H(E|Y^n, S^n) \\ &= H(E|Y^n) + H(S^n|Y^n, E). \end{aligned}$$

Note that $H(E|Y^n) \le 1$ and $H(E|Y^n, S^n) = 0$. We can also bound $H(S^n|Y^n, E)$ as

$$\begin{aligned} H(S^n|E, Y^n) &= H(S^n|Y^n, E=0)\Pr(E=0) \\ &\quad + H(S^n|Y^n, E=1)\Pr(E=1) \\ &\le \log|L_n|(1 - P_{e,s}^{(n)}) + n\log|\mathcal{S}|P_{e,s}^{(n)} \end{aligned}$$

where the inequality follows because when there is no error, the remaining uncertainty is at most $\log |L_n|$, and when there is an error, the uncertainty is at most $n \log |\mathcal{S}|$. This implies that

$$\begin{aligned} H(S^n|Y^n) &\le 1 + \log |L_n|(1 - P_{e,s}^{(n)}) + n \log |\mathcal{S}| P_{e,s}^{(n)} \\ &= \log |L_n| + 1 + (n \log |\mathcal{S}| - \log |L_n|) P_{e,s}^{(n)}. \end{aligned}$$

Taking $\epsilon_n = \frac{1}{n} + (\log |\mathcal{S}| - \frac{1}{n} \log |L_n|) P_{e,s}^{(n)}$ proves the desired result. □

The second lemma is crucial to the proof of Theorem 1 and contains a more interesting technique than Lemma 1. This lemma shows that the third inequality (5) can be replaced by a tighter inequality (7) below (recall that $I(U,S;Y) \le I(X,S;Y)$ since $U \to (X,S) \to Y$), which becomes crucial for the achievability proof of Theorem 1.

Lemma 2: Let $\mathcal{R}$ be the union of all (R, Δ) pairs satisfying (3)–(5). Let $\mathcal{R}_0$ be the closure of the union of all (R, Δ) pairs satisfying

$$R \le I(U;Y) - I(U;S) \quad (3)$$

$$\Delta \le H(S) \quad (4)$$

$$R + \Delta \le I(U,S;Y) \quad (7)$$

for some joint distribution $p(s)p(x,u|s)p(y|x,s)$, where the auxiliary random variable U has finite cardinality. Then

$$\mathcal{R} = \mathcal{R}_0.$$

Proof: Since $U \to (X,S) \to Y$ forms a Markov chain, it is trivial to check that

$$\mathcal{R}_0 \subseteq \mathcal{R}. \quad (8)$$

For the other direction of inclusion, we need some notation. Let $\mathcal{P}$ be the set of all distributions of the form $p(s)p(x,u|s)p(y|x,s)$ consistent with the given $p(s)$ and $p(y|x,s)$, where the auxiliary random variable U is defined on an arbitrary finite set. Further, let $\mathcal{P}'$ be the restriction of $\mathcal{P}$ such that $X = f(U,S)$ for some function f, i.e., $p(x|u,s)$ takes values 0 or 1 only.

If we define $\mathcal{R}_1$ to denote the closure of all (R, Δ) pairs satisfying (3), (4), and (7) over $\mathcal{P}'$, or equivalently, if $\mathcal{R}_1$ is defined to be the restriction of $\mathcal{R}_0$ over a smaller set of distributions $\mathcal{P}'$, then clearly

$$\mathcal{R}_1 \subseteq \mathcal{R}_0. \quad (9)$$

Let $\mathcal{R}_2$ be defined as the closure of (R, Δ) pairs satisfying (3)–(5). Since $X \to (U,S) \to Y$ forms a Markov chain on $\mathcal{P}'$, we have

$$\mathcal{R}_2 \subseteq \mathcal{R}_1. \quad (10)$$

To complete the proof, it now suffices to show that

$$\mathcal{R} \subseteq \mathcal{R}_2. \quad (11)$$

To see this, we restrict $\mathcal{R}_2$ to the distributions of the form $U = (V, \tilde{U})$ with V independent of $(\tilde{U}, S)$, namely

$$p(x,u|s) = p(x,v,\tilde{u}|s) = p(v)p(\tilde{u}|s)p(x|v,\tilde{u},s) \quad (12)$$

with deterministic $p(x|v,\tilde{u},s)$, i.e., x is a function of $(v,\tilde{u},s)$, and call this restriction $\mathcal{R}_3$. Since X is a deterministic function of $(V,\tilde{U},S)$ and at the same time $(V,\tilde{U}) \to (X,S) \to Y$ form a Markov chain, $\mathcal{R}_3$ can be written as the closure of all (R, Δ) pairs satisfying

$$\begin{aligned} R &\le I(V,\tilde{U};Y) - I(V,\tilde{U};S) \\ \Delta &\le H(S) \\ R + \Delta &\le I(V,\tilde{U},S;Y) = I(X,S;Y) \end{aligned}$$

for some distribution of the form $p(s)p(x,v,\tilde{u}|s)p(y|x,s)$ satisfying (12). But we have

$$\begin{aligned} I(V,\tilde{U};Y) - I(V,\tilde{U};S) &\ge I(\tilde{U};Y) - I(V,\tilde{U};S) \\ &= I(\tilde{U};Y) - I(\tilde{U};S) \end{aligned}$$

and the set of conditional distributions on $(\tilde{U}, X)$ given S satisfying (12) is as rich as any $p(\tilde{u},x|s)$. (Indeed, any conditional distribution $p(a|b)$ can be represented as $\sum_c p(c)p(a|b,c)$ for appropriately chosen $p(c)$ and *deterministic* distribution $p(a|b,c)$ with cardinality of C upper-bounded by $(|\mathcal{A}| - 1)|\mathcal{B}| + 1$; see also [32, Eq. (44)].) Therefore, we have

$$\mathcal{R} \subseteq \mathcal{R}_3 \subseteq \mathcal{R}_2 \quad (13)$$

which completes the proof. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1: For the proof of achievability, in the light of Lemma 2, it suffices to prove that any pair (R, Δ) satisfying (3), (4), (7) for some $p(u,x|s)$ is achievable. Since the coding technique is quite standard, we only sketch the proof here. For fixed $p(u,x|s)$, the result of Gel'fand–Pinsker [13] shows that the transmitter can send $I(U;Y) - I(U;S)$ bits reliably across the channel. Now we allocate $0 \le R \le I(U;Y) - I(U;S)$ bits for sending the pure information and use the remaining $\Gamma = I(U;Y) - I(U;S) - R$ bits for sending the state information by random binning. More specifically, we assign typical S^n sequences to $2^{n\Gamma}$ bins at random and send the bin index of the observed S^n using $n\Gamma$ bits. At the receiving end, the receiver is able to decode the codeword U^n from Y^n with high probability. Using joint typicality of (Y^n, U^n, S^n), the state uncertainty can be first reduced from $H(S)$ to $H(S|Y,U)$. Indeed, the number of typical S^n sequences jointly typical with (Y^n, U^n) is bounded by $2^{n(H(S|Y,U)+\epsilon)}$. In addition, using $\Gamma = I(U;Y) - I(U;S) - R$ bits of independent refinement information from the hash index of S^n, we can further reduce the state uncertainty by Γ. Hence, by taking the list of all S^n sequences jointly typical with (Y^n, U^n) satisfying the hash check, we have the total state uncertainty reduction rate

$$\begin{aligned} \Delta &= I(U,Y;S) + \Gamma \\ &= I(U,Y;S) + I(U;Y) - I(U;S) - R \\ &= I(U,S;Y) - R. \end{aligned}$$

By varying $0 \le R \le I(U;Y) - I(U;S)$, it can be readily seen that all (R, Δ) pairs satisfying

$$\begin{aligned} R &\le I(U;Y) - I(U;S) \\ \Delta &\le H(S) \\ R + \Delta &\le I(U,S;Y) \end{aligned}$$

for any fixed $p(x,u|s)$ are achievable.

For the proof of converse, we have to show that given any sequence of $(2^{nR}, 2^{n\Delta}, n)$ codes with $P_{e,w}^{(n)}, P_{e,s}^{(n)} \to 0$, the (R, Δ) pairs must satisfy

$$\begin{aligned} R &\le I(U;Y) - I(U;S) \\ \Delta &\le H(S) \\ R + \Delta &\le I(X,S;Y) \end{aligned}$$

for some joint distribution $p(s)p(x,u|s)p(y|x,s)$.

The pure information rate R can be readily bounded from the previous work by Gel'fand and Pinsker [13, Proposition 3]. Here we repeat a simpler proof given in Heegard [14, Appendix 2] for completeness; see also [9, Lecture 13]. Starting with Fano's inequality, we have the following chain of inequalities:

$$\begin{aligned} &nR \\ &\le I(W;Y^n) + n\epsilon_n \\ &= \sum_{i=1}^{n} I(W;Y_i|Y^{i-1}) + n\epsilon_n \\ &\le \sum_{i=1}^{n} I(W,Y^{i-1};Y_i) + n\epsilon_n \\ &= \sum_{i=1}^{n} I(W,Y^{i-1},S_{i+1}^n;Y_i) - \sum_{i=1}^{n} I(Y_i;S_{i+1}^n|W,Y^{i-1}) + n\epsilon_n \\ &\overset{\text{(a)}}{=} \sum_{i=1}^{n} I(W,Y^{i-1},S_{i+1}^n;Y_i) - \sum_{i=1}^{n} I(Y^{i-1};S_i|W,S_{i+1}^n) + n\epsilon_n \\ &\overset{\text{(b)}}{=} \sum_{i=1}^{n} I(W,Y^{i-1},S_{i+1}^n;Y_i) - \sum_{i=1}^{n} I(W,Y^{i-1},S_{i+1}^n;S_i) + n\epsilon_n \end{aligned}$$

where (a) follows from the Csiszár sum formula

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

and (b) follows because (W, S_{i+1}^n) is independent of S_i. Recognizing the auxiliary random variable $U_i = (W, Y^{i-1}, S_{i+1}^n)$ and noting that $U_i \to (X_i, S_i) \to Y_i$ form a Markov chain, we have

$$nR \le \sum_{i=1}^{n} (I(U_i;Y_i) - I(U_i;S_i)) + n\epsilon_n. \tag{14}$$

On the other hand, since $\log|L_n| = n(H(S) - \Delta)$, we can trivially bound Δ by Lemma 1 as

$$\begin{aligned} n\Delta &\le nH(S) - H(S^n|Y^n) + n\epsilon_n' \\ &\le nH(S) + n\epsilon_n'. \end{aligned}$$

Similarly, we can bound $R + \Delta$ as

$$\begin{aligned} n(R+\Delta) &\le I(W;Y^n) + I(S^n;Y^n) + n\epsilon_n'' \\ &\overset{\text{(a)}}{\le} I(W;Y^n|S^n) + I(S^n;Y^n) + n\epsilon_n'' \\ &\le I(W,S^n;Y^n) + n\epsilon_n'' \\ &\overset{\text{(b)}}{=} I(X^n,S^n;Y^n) + n\epsilon_n'' \\ &\overset{\text{(c)}}{\le} \sum_{i=1}^{n} I(X_i,S_i;Y_i) + \epsilon_n'' \end{aligned} \tag{15}$$

where (a) follows since W is independent of S^n and conditioning reduces entropy, (b) follows from the data processing inequality (both directions), and (c) follows from the memorylessness of the channel.

We now introduce the usual time-sharing random variable Q uniform over $\{1, \ldots, n\}$, independent of everything else. Then (14) implies

$$\begin{aligned} R &\le I(U_Q;Y_Q|Q) - I(U_Q;S_Q|Q) + \epsilon_n \\ &= I(U_Q,Q;Y_Q) - I(U_Q,Q;S_Q) + \epsilon_n. \end{aligned}$$

On the other hand, (15) implies

$$\begin{aligned} R + \Delta &\le I(X_Q,S_Q;Y_Q|Q) + \epsilon_n'' \\ &\le I(X_Q,S_Q,Q;Y_Q) + \epsilon_n'' \\ &= I(X_Q,S_Q;Y_Q) + \epsilon_n'' \end{aligned}$$

where the last equality follows since $Q \to (X_Q, S_Q) \to Y_Q$ form a Markov chain.

Finally, we recognize $U = (U_Q, Q), X = X_Q, S = S_Q$, $Y = Y_Q$, and note that $S \sim p(s)$, $\Pr(Y = y|X = x, S = s) = p(y|x,s)$, and $U \to (X,S) \to Y$, which completes the proof of the converse. □

Roughly speaking, the optimal coding scheme is equivalent to sending the codeword U^n reliably at the Gel'fand–Pinsker rate $R' = I(U;Y) - I(U;S)$ and reducing the receiver's uncertainty by $\Delta' = I(S;U,Y)$ from Y^n and the decoded codeword U^n. It should be noted that (R', Δ') has the same form as the achievable region for the dual tradeoff problem between pure information rate R and (minimum) normalized mutual information rate $E = \frac{1}{n}I(S^n;Y^n)$ studied in [21]. But we can reduce the uncertainty about S^n further by allocating part Γ of the pure information rate R' to convey independent refinement information (hash index of S^n). By varying $\Gamma \in [0, R']$ we can trace the entire tradeoff region $(R' - \Gamma, \Delta' + \Gamma)$.

It turns out an alternative coding scheme based on Wyner–Ziv source coding with side information [33], instead of random binning, also achieves the tradeoff region $\mathcal{R}^*$. To see this, fix any $p(u,x|s)$ and $p(v|s)$ satisfying

$$\Gamma := I(V;S|U,Y) \le I(U;Y) - I(U;S)$$

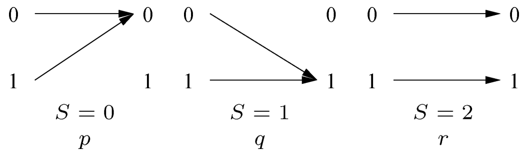


Fig. 2. Memory with defective cells.

and consider the Wyner–Ziv encoding of S^n with covering codeword V^n and side information (U^n, Y^n) at the decoder. More specifically, we can generate $2^{nI(V;S)}$ V^n codewords and assign them into $2^{n\Gamma}$ bins. As before, we use the Gel'fand–Pinsker coding to convey a message of rate $I(U;Y) - I(U;S)$ reliably over the channel. Since the rate $\Gamma = I(V;S|U,Y)$ is sufficient to reconstruct V^n at the receiver with side information Y^n and U^n, we can allocate the rate Γ for conveying V^n and use the remaining rate $R = I(U;Y) - I(U;S) - \Gamma$ for extra pure information. Forming a list of S^n jointly typical with (Y^n, U^n, V^n) results in the uncertainty reduction rate Δ given by

$$\begin{aligned}\Delta &= I(S;Y,U,V)\\ &= I(S;Y,U) + \Gamma\\ &= I(S;U,Y) + I(U;Y) - I(U;S) - R\\ &= I(U,S;Y) - R.\end{aligned}$$

Thus the tradeoff region $\mathcal{R}^*$ can be achieved via the combination of two fundamental results in communication with side information: channel coding with side information by Gel'fand and Pinsker [13] and rate distortion with side information by Wyner and Ziv [33]. It is also interesting to note that the information about S^n can be transmitted in a manner completely independent of geometry (random binning) or completely dependent on geometry (random covering); refer to [6] for a similar phenomenon in a relay channel problem.

When Y is a function of (X,S), it is optimal to identify $U = Y$, and Theorem 1 simplifies to the following corollary.

Corollary 2: The tradeoff region $\mathcal{R}^*$ for a deterministic state-dependent channel $Y = f(X,S)$ with state information S^n noncausally known at the transmitter is the union of all (R,Δ) pairs satisfying

$$R \le H(Y|S) \tag{16}$$

$$\Delta \le H(S) \tag{17}$$

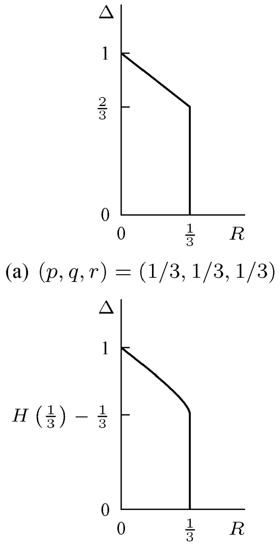
$$R + \Delta \le H(Y) \tag{18}$$

for some joint distribution of the form $p(s)p(x|s)p(y|x,s)$. In particular, the maximum uncertainty reduction rate is given by

$$\Delta^* = \min\{H(S), \max_{p(x|s)} H(Y)\}. \tag{19}$$

The next two examples show different flavors of optimal state uncertainty reduction.

Example 1: Consider the problem of conveying information using a write-once memory device with stuck-at defective cells [19], [15] as depicted in Fig. 2. Here each memory cell has probability p of being stuck at 0, probability q of being stuck at 1, and probability r of being a good cell, with $p + q + r = 1$. It is easy to see that the channel output Y is a simple deterministic function of the channel input X and the state S.



(a) $(p,q,r) = (1/3, 1/3, 1/3)$

(b) $(p,q,r) = (1/2, 1/6, 1/3)$

Fig. 3. The optimal (R,Δ) tradeoff for memory with defective cells.

Now it is easy to verify that the tradeoff region $\mathcal{R}^*$ is given by

$$R \le rH(\alpha) \tag{20}$$

$$\Delta \le H(p,q,r) \tag{21}$$

$$R + \Delta \le H(p + \alpha r, q + (1-\alpha) r) \tag{22}$$

where α can be chosen arbitrarily ($0 \le \alpha \le 1$). This region is achieved by choosing $p(x) \sim \mathrm{Bern}(\alpha)$. Without loss of generality, we can choose $X \sim \mathrm{Bern}(\alpha)$ independent of S, because the input X affects Y only when $S = 2$. There are two cases to consider.

(a) If $p = q$, then the choice of $\alpha^* = 1/2$ maximizes both (20) and (22), and hence achieves the entire tradeoff region $\mathcal{R}^*$. The optimal transmitter splits the full channel capacity $C = rH(\alpha^*) = r$ to send both the pure information and the state information. (See Fig. 3(a) for the case $(p,q,r) = (1/3, 1/3, 1/3)$.)

(b) On the other hand, when $p \ne q$, there is a clear tradeoff in our choice of α. For example, consider the case $(p,q,r) = (1/2, 1/6, 1/3)$. If the goal is to communicate pure information over the channel, we should take $\alpha^* = 1/2$ to maximize the number of distinguishable input preparations. This gives the channel capacity $C = rH(\alpha) = 1/3$. If the goal is, however, to help the receiver reduce the state uncertainty, we take $\alpha^* = 0$, i.e., we transmit a fixed signal $X \equiv 0$. This way, the transmitter can minimize his interference with the receiver's view of the state S. The entire tradeoff region is given in Fig. 3(b).

Example 2: Consider the binary multiplying channel $Y = X \cdot S$, where the output Y is the product of the input $X \in \{0,1\}$ and the state $S \in \{0,1\}$. We assume that the state sequence S^n

is drawn i.i.d. according to $\mathrm{Bern}(\gamma)$. It can be easily shown that the optimal tradeoff region is given by

$$R \le \gamma H(\alpha) \tag{23}$$

$$\Delta \le H(\gamma) \tag{24}$$

$$R + \Delta \le H(\alpha\gamma). \tag{25}$$

This is achieved by $p(x) \sim \mathrm{Bern}(\alpha)$, independent of S.

As in Example 1(b), there is a tension between the pure information transmission and the state amplification. When the goal is to maximize the pure information rate, we should choose $\alpha^* = 1/2$ to achieve the capacity $C = \gamma$. But when the goal is to maximize the state uncertainty reduction rate, we should choose $\alpha^* = 1$ ($X \equiv 1$) to achieve $\Delta^* = H(\gamma)$. In words, to maximize the state uncertainty reduction rate, the transmitter simply clears the receiver's view of the state.

III. Extension to Continuous State Space

The previous section characterized the tradeoff region $\mathcal{R}^*$ between the pure information rate R and the state uncertainty reduction rate $\Delta = H(S) - \frac{1}{n}\log|L_n(Y^n)|$. Apparently, the notion of uncertainty reduction rate Δ is meaningful only when the channel state S has finite cardinality (i.e., $|\mathcal{S}| < \infty$), or at least when $H(S) < \infty$.

However, from the proof of Theorem 1 (the generalized Fano's inequality in Lemma 1), along with the fact that the optimal region is single-letterizable, we can take an alternative look at the notion of state uncertainty reduction as reducing the list size from $2^{nH(S)}$ to $|L_n(Y^n)|$. We will show shortly in Proposition 1 that the difference $\Delta = H(S) - \frac{1}{n}\log|L_n|$ of the normalized list size is essentially equivalent to the normalized mutual information $\Delta_I = \frac{1}{n}I(S^n;Y^n)$, which is well-defined for an arbitrary state space $\mathcal{S}$ and captures the amount of information the receiver Y^n can learn about the state S^n (or lack thereof [21]). Hence, the physically motivated notion Δ of list size reduction is consistent with the mathematical information measure Δ_I, and both notions of state uncertainty reduction can be used interchangeably, especially when $\mathcal{S}$ is finite.

To be more precise, we define a $(2^{nR}, n)$ code by an encoding function

$$X^n : [2^{nR}] \times \mathcal{S}^n \to \mathcal{X}^n$$

and a decoding function

$$\hat{W} : \mathcal{Y}^n \to [2^{nR}].$$

Then, the associated state uncertainty reduction rate for the $(2^{nR}, n)$ code is defined as

$$\Delta_I = \frac{1}{n}I(S^n;Y^n)$$

where the mutual information is with respect to the joint distribution

$$p(x^n, s^n, y^n) = p(x^n|s^n)\prod_{i=1}^{n} p(s_i)p(y_i|x_i, s_i)$$

induced by $X^n(W, S^n)$ with message W distributed uniformly over $[2^{nR}]$, independent of S^n. Similarly, the probability of error is defined as

$$P_e^{(n)} = \Pr(W \ne \hat{W}(Y^n)).$$

A pair (R, Δ) is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ and

$$\lim_{n\to\infty} \frac{1}{n}I(S^n;Y^n) \ge \Delta.$$

The closure of all achievable (R, Δ) pairs is called the tradeoff region $\mathcal{R}_I^*$. (Here we use the notation $\mathcal{R}_I^*$ instead of $\mathcal{R}^*$ to temporarily distinguish this from the original problem formulated in terms of the list size reduction.)

We now show that the optimal tradeoff $\mathcal{R}_I^*$ between the information transmission rate R and the mutual information rate Δ has the same solution as the optimal tradeoff $\mathcal{R}^*$ between R and the list size reduction rate Δ.

Proposition 1: The tradeoff region $\mathcal{R}_I^*$ for a state-dependent channel $(\mathcal{X} \times \mathcal{S}, p(y|x,s), \mathcal{Y})$ with state information S^n noncausally known at the transmitter is the closure of all (R, Δ) pairs satisfying

$$R \le I(U;Y) - I(U;S) \tag{3}$$

$$\Delta \le H(S) \tag{4}$$

$$R + \Delta \le I(X, S;Y) \tag{5}$$

for some joint distribution of the form $p(s)p(u,x|s)p(y|x,s)$ with auxiliary random variable U. Hence, $\mathcal{R}_I^*$ has the identical characterization as $\mathcal{R}^*$ in Theorem 1.

Proof: Let $\mathcal{R}^{**}$ be the region described by (3)–(5). We provide a sandwich proof $\mathcal{R}^{**} = \mathcal{R}^* \subseteq \mathcal{R}_I^* \subseteq \mathcal{R}^{**}$, which is given implicitly in the proof of Theorem 1.

More specifically, consider a finite partition[1] to quantize the state random variable S into $[S]$. Under this partition, let $\mathcal{R}_{[S]}^{**}$ be the set of all (R, Δ) pairs satisfying

$$R \le I(U;Y) - I(U;[S])$$

$$\Delta \le H([S])$$

$$R + \Delta \le I(X, [S];Y)$$

for some joint distribution $p([s])p(u,x|[s])p(y|x,[s])$ with auxiliary random variable U. Consider the original list size reduction problem with state information $[S]$ and let $\mathcal{R}_{[S]}^*$ denote the tradeoff region. Then Theorem 1 shows that $\mathcal{R}_{[S]}^{**} = \mathcal{R}_{[S]}^*$. In particular, for any $\epsilon > 0$ and $(R, \Delta) \in \mathcal{R}_{[S]}^{**}$, there exists a sequence of $(2^{n(R-\epsilon)}, 2^{n(\Delta-\epsilon)}, n)$ codes $X^n(W), \hat{W}(Y^n), L_n(Y^n)$ such that $P_{e,w}^{(n)} = \Pr(W \ne \hat{W}) \to 0$ and $P_{e,s}^{(n)} = \Pr([S]^n \ne L_n(Y^n)) \to 0$.

Now from the generalized Fano's inequality (Lemma 1), the achievable list size reduction rate $\Delta - \epsilon$ should satisfy

$$n(\Delta - \epsilon) \le I([S]^n;Y^n) + n\epsilon_n \le I(S^n;Y^n) + n\epsilon_n$$

with $\epsilon_n \to 0$ as $n \to \infty$. Hence, by letting $n \to \infty$ and $\epsilon \to 0$, we have from the definition of $\mathcal{R}_I^*$ that

$$\mathcal{R}_{[S]}^{**} = \mathcal{R}_{[S]}^* \subseteq \mathcal{R}_I^*.$$

Also, it follows trivially from repeating the intermediate steps in the converse proof of Theorem 1 that $\mathcal{R}_I^* \subseteq \mathcal{R}^{**}$.

[1] Recall that the mutual information between arbitrary random variables X and Y is defined as $I(X;Y) = \sup I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$, where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$; see Kolmogorov [18] and Pinsker [25].

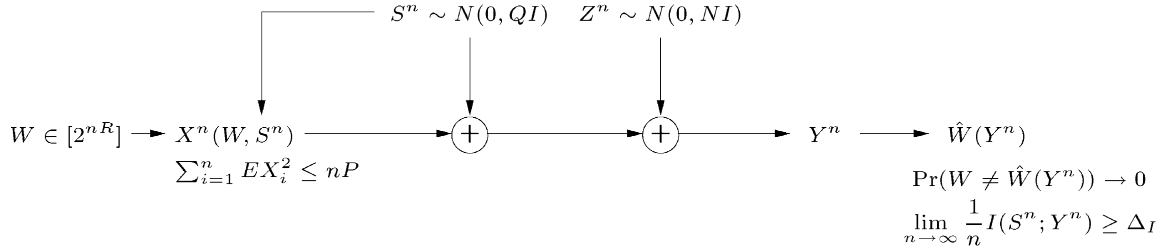

$S^n \sim N(0, QI)$
$Z^n \sim N(0, NI)$
$W \in [2^{nR}]$
$X^n(W, S^n)$
$\sum_{i=1}^n EX_i^2 \le nP$
+
+
Y^n
$\hat{W}(Y^n)$
$\Pr(W \ne \hat{W}(Y^n)) \to 0$
$\lim_{n\to\infty} \frac{1}{n} I(S^n; Y^n) \ge \Delta_I$

Fig. 4. Writing on dirty paper.

Finally, taking a sequence of partitions with mesh → 0 and hence letting $\mathcal{R}^{**}_{[S]} \to \mathcal{R}^{**}$, we have the desired result. □

Since both notions of state uncertainty reduction, the list size reduction $nH(S) - \log|L_n|$ and the mutual information $I(S^n; Y^n)$ lead to the same answer, we will subsequently use them interchangeably and denote the tradeoff region by the same symbol $\mathcal{R}^*$.

Example 3: Consider Costa's writing on dirty paper model depicted in Fig. 4 as the canonical example of a continuous state-dependent channel. Here the channel output is given by $Y^n = X^n + S^n + Z^n$, where $X^n(W, S^n)$ is the channel input subject to a power constraint $\sum_{i=1}^n EX_i^2 \le nP$, $S^n \sim N(0, QI)$ is the additive white Gaussian state, and $Z^n \sim N(0, NI)$ is the white Gaussian noise. We assume that S^n and Z^n are independent.

For the writing on dirty paper model, we have the following tradeoff between the pure information transmission and the state uncertainty reduction.

Proposition 2: The tradeoff region $\mathcal{R}^*$ for the Gaussian channel depicted in Fig. 4 is characterized by the boundary points $(R(\gamma), \Delta(\gamma)),\ 0 \le \gamma \le 1$, where

$$R(\gamma) = \frac{1}{2}\log\left(1 + \frac{\gamma P}{N}\right) \tag{26}$$

$$\Delta(\gamma) = \frac{1}{2}\log\left(1 + \frac{\left(\sqrt{Q} + \sqrt{(1-\gamma)P}\right)^2}{\gamma P + N}\right). \tag{27}$$

Proof sketch: The achievability follows from Proposition 1 with trivial extension to the input power constraint. In particular, we use the simple power sharing scheme proposed in [29], where a fraction γ of the input power is used to transmit the pure information using Costa's writing on dirty paper coding technique, while the remaining $(1-\gamma)$ fraction of the power is used to amplify the state. In other words

$$X = V + \sqrt{(1-\gamma)\frac{P}{Q}}\,S \tag{28}$$

with $V \sim N(0, \gamma P)$ independent of S, and

$$U = V + \alpha S$$

with

$$\alpha = \frac{\gamma P}{\gamma P + N}\sqrt{\frac{(1-\gamma)P + Q}{Q}}.$$

Evaluating $R = I(U;Y) - I(U;S)$ and $\Delta = I(S;Y)$ for each γ, we recover (26) and (27).

The proof of converse is essentially the same as that of [29, Theorem 2], which we do not repeat here. □

As an extreme point of the (R, Δ), we recover Costa's writing on dirty paper result

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

by taking $\gamma = 1$. On the other hand, if state uncertainty reduction is the goal, then all of the power should be used for state amplification. The maximum uncertainty reduction rate

$$\Delta^* = \frac{1}{2}\log\left(1 + \frac{(\sqrt{P} + \sqrt{Q})^2}{N}\right)$$

is achieved with $X = \sqrt{\frac{P}{Q}}S$ and $\alpha = 0$.

In [29, Theorem 2], the optimal tradeoff was characterized between the pure information rate R and the receiver's state estimation error $D = \frac{1}{n}E\|S^n - \hat{S}^n(Y^n)\|^2$. Although the notion of state estimation error D in [29] and our notion of the uncertainty reduction rate Δ appear to be distinct objectives at first sight, the optimal solutions to both problems are identical, as shown in the proof of Proposition 2. There is no surprise here. Because of the quadratic Gaussian nature of both problems, minimizing the mean-squared error $E(S - \hat{S}(Y))^2$ can be recast into maximizing the mutual information $I(S;Y)$, and *vice versa*. Also, the optimal state uncertainty reduction rate Δ^* (or equivalently, the minimum state estimation error D^*) is achieved by the symbol-by-symbol amplification $X_i = \sqrt{\frac{P}{Q}}\,S_i$.

Finally, it is interesting to compare the optimal coding scheme (28) to the optimal coding scheme when the goal is to minimize (instead of maximizing) the uncertainty reduction [21], which is essentially based on coherent subtraction of X and S with possible randomization.

IV. Optimal (R, Δ) Tradeoff: Causal Case

The previous two sections considered the case in which the transmitter has complete knowledge of the state sequence S^n prior to the actual communication. In this section, we consider another model in which the transmitter learns the state sequence on the fly, i.e., the encoding function

$$X_i : [2^{nR}] \times \mathcal{S}^i \to \mathcal{X}, \qquad i = 1, 2, \ldots, n$$

depends causally on the state sequence.

We state our main theorem.

Theorem 2: The tradeoff region $\mathcal{R}^*$ for a state-dependent channel $(\mathcal{X} \times \mathcal{S}, p(y|x,s), \mathcal{Y})$ with state information S^n causally known at the transmitter is the union of all (R, Δ) pairs satisfying

$$R \le I(U;Y) \tag{29}$$

$$\Delta \le H(S) \tag{30}$$

$$R + \Delta \le I(X,S;Y) \tag{31}$$

for some joint distribution $p(s)p(u)p(x|u,s)p(y|x,s)$, where the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \le |\mathcal{X}| \cdot |\mathcal{S}|$.

As in the noncausal case, the region is convex. Since the auxiliary random variable U affects the first inequality (29) only, the cardinality bound $|\mathcal{U}| \le |\mathcal{X}| \cdot |\mathcal{S}|$ follows again from the standard argument. (A looser bound can be given by counting the number of functions $f : \mathcal{S} \to \mathcal{X}$; see Shannon [27].) Finally, we can take X as a deterministic function of (U, S) without decreasing the region.

Compared to the noncausal tradeoff region $\mathcal{R}^*_{\mathrm{nc}}$ in Theorem 1, the causal tradeoff region $\mathcal{R}^*_{\mathrm{c}}$ in Theorem 2 is smaller in general. More precisely, $\mathcal{R}^*_{\mathrm{c}}$ is characterized by the same set of inequalities (3)–(5) as in $\mathcal{R}^*_{\mathrm{nc}}$, but the set of joint distributions is restricted to those with auxiliary variable U independent of S. Indeed, from the independence between U and S, we can rewrite (29) as

$$R \le I(U;Y) = I(U;Y) - I(U;S) \tag{29$'$}$$

which is exactly the same as (3). Thus, the inability to use the future state sequence decreases the tradeoff region. However, only the inequality (29), or equivalently, the inequality (3), is affected by the causality, and the sum rate (31) does not change from (5).

Since the proof of Theorem 2 is essentially identical to that of Theorem 1, we skip most of the steps. The least straightforward part is the following lemma.

Lemma 3: Let $\mathcal{R}$ be the union of all (R, Δ) pairs satisfying (29)–(31). Let $\mathcal{R}_0$ be the closure of the union of all (R, Δ) pairs satisfying (29), (30), and

$$R + \Delta \le I(U,S;Y) \tag{32}$$

for some joint distribution $p(s)p(u)p(x|u,s)p(y|x,s)$ where the auxiliary random variable U has finite cardinality. Then

$$\mathcal{R} = \mathcal{R}_0.$$

Proof sketch: The proof is a verbatim copy of the proof of Lemma 2, except that here U is independent of S, i.e., $p(x,u|s) = p(u)p(x|u,s)$. The final step (13) follows since the set of conditional distributions on $X, U = (V, \tilde{U})$ given S of the form

$$p(x,u|s) = p(v)p(\tilde{u})p(x|v,\tilde{u},s) \tag{12$'$}$$

with deterministic $p(x|v,\tilde{u},s)$ is as rich as any $p(\tilde{u})p(x|\tilde{u},s)$, and

$$I(V,\tilde{U};Y) \ge I(\tilde{U};Y). \tag{13$'$}$$

With this replacement, the desired proof follows along the same lines as the proof of Lemma 2. □

As one extreme point of the tradeoff region $\mathcal{R}^*$, we recover the Shannon capacity formula [27] for channels with causal side information at the transmitter as follows:

$$C = \max_{p(u)p(x|u,s)} I(U;Y). \tag{33}$$

On the other hand, the maximum uncertainty reduction rate Δ^* for pure state amplification is identical to that for the noncausal case given in Corollary 1.

Corollary 3: Under the condition of Theorem 2, the maximum uncertainty reduction rate Δ^* is given by

$$\Delta^* = \min\{H(S), \max_{p(x|s)} I(X,S;Y)\}. \tag{34}$$

Thus, the receiver can learn about the state essentially at the maximum cut-set rate, even under the causality constraint. For example, the symbol-by-symbol amplification strategy $X = \sqrt{\frac{P}{Q}}S$ is optimal for the Gaussian channel (Example 3) for both causal and noncausal cases.

Finally, we compare the tradeoff regions $\mathcal{R}^*_{\mathrm{c}}$ and $\mathcal{R}^*_{\mathrm{nc}}$ with a communication problem that has a totally different motivation, yet has a similar capacity expression. In [32, Situations 3 and 4], Willems and van der Meulen studied the multiple-access channel with cribbing encoders. In this communication problem, the multiple-access channel $(\mathcal{X} \times \mathcal{S}, p(y|x,s), \mathcal{Y})$ has two inputs and one output. The primary transmitter S and the secondary transmitter X wish to send independent messages $W_s \in [2^{n\Delta}]$ and $W_x \in [2^{nR}]$, respectively, to the common receiver Y. The difference from the classical multiple-access channel is that either the secondary transmitter X learns the primary transmitter's signal S on the fly $(X_i(W_x, S^i)$ [32, Situation 3]) or X knows the entire signal S^n ahead of time $(X_i(W_x, S^n)$ [32, Situation 4]). The capacity region $\mathcal{C}$ for both cases is given by all (R, Δ) pairs satisfying

$$R \le I(X;Y|S) \tag{35}$$

$$\Delta \le H(S) \tag{36}$$

$$R + \Delta \le I(X,S;Y) \tag{37}$$

for some joint distribution $p(x,s)p(y|x,s)$.

This capacity region $\mathcal{C}$ looks almost identical to the tradeoff regions $\mathcal{R}^*_{\mathrm{nc}}$ and $\mathcal{R}^*_{\mathrm{c}}$ in Theorems 1 and 2, except for the first inequality (35). Moreover, (35) has the same form as the capacity expression for channels with state information available at *both* the encoder and decoder, either causally or noncausally. (The causality has no cost when both the transmitter and the receiver share the same side information; see, for example, Caire and Shamai [1, Proposition 1].)

It should be stressed, however, that the problem of cribbing multiple-access channels and our state uncertainty reduction problem have a fundamentally different nature. The former deals with encoding and decoding of the signal S^n, while the latter deals with uncertainty reduction in an uncoded sequence S^n specified by nature. In a sense, the cribbing multiple-access channel is a detection problem, while the state uncertainty reduction is an estimation problem.

V. Concluding Remarks

Because the channel is state dependent, the receiver is able to learn something about the channel state from directly observing the channel output. Thus, to help the receiver narrow down the uncertainty about the channel state at the highest rate possible, the sender must jointly optimize between facilitating state estimation and transmitting refinement information, rather than merely using the channel capacity to send the state description. In particular, the transmitter should summarize the state information in such a way that the summary information results in the maximum uncertainty reduction when coupled with the receiver's initial estimate of the state. More generally, by taking away some resources used to help the receiver reduce the state uncertainty, the transmitter can send additional pure information to the receiver and trace the entire (R, Δ) tradeoff region.

There are three surprises here. First, the receiver can learn about the channel state and the independent message at a maximum cut-set rate $I(X, S; Y)$ over all joint distributions $p(x, s)$ consistent with the given state distribution $p(s)$. Second, to help the receiver reduce the uncertainty in the initial estimate of the state (namely, to increase the mutual information from $I(S; Y)$ to $I(X, S; Y)$), the transmitter can allocate the achievable information rate $I(U; Y) - I(U; S)$ in two alternative methods—random binning and its dual, random covering. Third, as far as the sum rate $R + \Delta$ and the maximum uncertainty reduction rate Δ^* are concerned, there is no cost associated with restricting the encoder to learn the state sequence on the fly.

Acknowledgment

The authors wish to thank the anonymous reviewer and the Associate Editor for their insightful comments, which helped to improve the quality of the paper.

References

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

[2] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[3] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, Jul. 2001.

[4] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1639–1667, Jun. 2002.

[5] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[6] T. M. Cover and Y.-H. Kim, "Capacity of a class of determinstic relay channels," in *Proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007, pp. 591–595.

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

[8] N. Devroye, P. Mitran, and V. Tarokh, "Achievable rates in cognitive radio channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1813–1827, May 2006.

[9] A. El Gamal, "Multiple User Information Theory," Course Notes, Stanford University, Stanford, CA, 2006, unpublished.

[10] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice strategies for canceling known interference," *IEEE Trans. Inf. Theory*, vol. I51, no. 11, pp. 3820–3833, Nov. 2005.

[11] U. Erez and S. ten Brink, "A close-to-capacity dirty paper coding scheme," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3417–3432, Oct. 2005.

[12] Federal Communications Commission, Cognitive Radio Technologies Proceeding (CRTP), ET Docket, no. 03-108.

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

[14] C. Heegard, "Capacity and Coding for Computer Memory with Defects," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1981.

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

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

[17] A. Jovičić and P. Viswanath, "Cognitive radio: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, submitted for publication.

[18] A. N. Kolmogorov, "Logical basis for information theory and probability theory," *IRE Trans. Inf. Theory*, vol. IT-2, no. 4, pp. 102–108, Dec. 1956.

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

[20] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[21] N. Merhav and S. Shamai (Shitz), "Information rates subject to state masking," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2254–2261, Jun. 2007.

[22] J. Mitolla, III, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Ph.D. dissertation, KTH Royal Inst. Techn., Stockholm, Sweden, 2000.

[23] M. Mohseni and J. M. Cioffi, "A proof of the converse for the capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, submitted for publication.

[24] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 563–593, Mar. 2003.

[25] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[26] M. Salehi, "Cardinality Bounds on Auxiliary Variables in Multiple-User Theory via the Method of Ahlswede and Körner," Dep. Statistics, Stanford Univ., Stanford, CA, 1978, Tech. Rep. 33.

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

[28] A. Sutivong, "Channel Capacity and State Estimation for State-Dependent Channels," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2003.

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

[30] A. Sutivong, T. M. Cover, M. Chiang, and Y.-H. Kim, "Rate vs. distortion trade-off for channels with state information," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 226.

[31] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

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

[33] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[34] R. Zamir, S. Shamai (Shitz), and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1276, Jun. 2002.