Probing Capacity

Himanshu Asnani, *Student Member, IEEE*, Haim Permuter, *Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We consider the problem of optimal probing of states of a channel by transmitter and receiver for maximizing rate of reliable communication. The channel is discrete memoryless (DMC) with i.i.d. states. The encoder takes probing actions dependent on the message. It then uses the state information obtained from probing causally or noncausally to generate channel input symbols. The decoder may also take channel probing actions as a function of the observed channel output and use the channel state information thus acquired, along with the channel output, to estimate the message. We refer to the maximum achievable rate for reliable communication for such systems as the "*Probing Capacity*". We characterize this capacity when the encoder and decoder actions are cost constrained. To motivate the problem, we begin by characterizing the trade-off between the capacity and fraction of channel states the encoder is allowed to observe, while the decoder is aware of channel states. In this setting of 'to observe or not to observe' state at the encoder, we compute certain numerical examples which exhibit a pleasing phenomenon, where encoder can observe a relatively small fraction of states and yet communicate at maximum rate, i.e., rate when observing states at encoder is not cost constrained.**

***Index Terms*—Actions, channel with states, cost constraints, Gel'fand-Pinsker channel, probing capacity, Shannon channel.**

I. Introduction

SHANNON showed the importance of availability of channel state at the encoder for communication system in his seminal paper [1], where he computed capacity of DMC with i.i.d. states available causally to the encoder. This spawned an active research in the area of channel coding and was extended to various scenarios, notably for storage in computer memory. Kuznetsov and Tsybakov in [2] constructed defect-correcting codes for coding in computer memory with defective cells. Gel'fand and Pinsker in [3], extended work in [1] to the case where channel states are available noncausally to the encoder, again with applications for computer memories, which was further researched by Heegard and El Gamal in [4]. Keshet et al. presented a detailed survey in [5] on channel coding in the presence of state information, where the channel state information (CSI) signal is available at the transmitter (CSIT) or at the receiver (CSIR), or both.

The notion of *actions* in source coding context is introduced in [6]. Their setting is a generalization of the Wyner-Ziv source coding with decoder side information ([7]), where now the decoder can take actions based on the index obtained from the encoder to affect the formation or availability of side information. Authors in [8], studied the channel coding dual where the transmitter takes actions that affect the formation of channel states. This framework captures various new coding scenarios which include two stage recording on a memory with defects, motivated by similar problems in magnetic recording and computer memories. Kittichokechai *et al.* in [9] studied a variant of the problem in [6] and [8], where encoder and decoder both have action dependent partial side information. However, in the source coding formulation of [6], they restricted the actions to be taken by decoder while in the channel coding scenario of [8] and [9], actions were taken only by the encoder.

In this paper, we revisit channel coding scenarios but now cost constrained actions are taken to acquire any partial or complete channel state information by the encoder, the decoder or both. Our framework is aimed at capturing and understanding the tradeoffs involved in natural scenarios where the acquisition of channel state information is associated with expenditure of costly system resources. The encoder and decoder actions are cost constrained creating tension between achievable rate and the cost of acquisition of the channel state (or the defect) information. Note that our framework differs from those of [8] and [9] where actions affect the channel, followed by channel encoding. In our scenario channel statistics are not affected, i.e., nature generates the state sequence i.i.d $\sim P_S$. Our work is novel in the sense that not only the encoder but the decoder also takes actions to acquire channel state information. Encoder takes actions (A_e) depending on messages. Decoder also takes actions (A_d) depending upon observed channel output. Using their respective actions, encoder and decoder observe partial states, S_e and S_d through discrete memoryless channel (DMC), $P_{S_e,S_d|S,A_e,A_d}$. The encoder can causally or noncausally use its partial state information to generate the channel input symbols. In this paper, we characterize the fundamental limit of such a framework and call it **Probing Capacity**. When the actions are not taken by the decoder, there is an equivalence between our setting and that of channels with action dependent states as in [8], which we make explicit in Section III.

The rest of the paper is organized as follows. We begin with a motivating scenario in Section II, where decoder knows the complete state and the encoder takes message dependent binary actions to observe or not to observe the channel state. This is generalized in Section III, when only encoder takes actions. This section also establishes the equivalence between our framework

Manuscript received October 04, 2010; revised May 24, 2011; accepted June 11, 2011. Date of current version November 11, 2011. The material in this paper was presented in part at the 48th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 2010.

H. Asnani and T. Weissman are with the Information Systems Lab, Electrical Engineering Department, Stanford University, CA 94305 USA (e-mail: asnani@stanford.edu; sachy@stanford.edu).

H. Permuter is with the Department of Electrical and Computer Engineering, Ben Gurion University of the Negev, Beer-Sheva, Israel (e-mail: haimp@bgu.ac.il).

Communicated by D. Guo, 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.2011.2162089

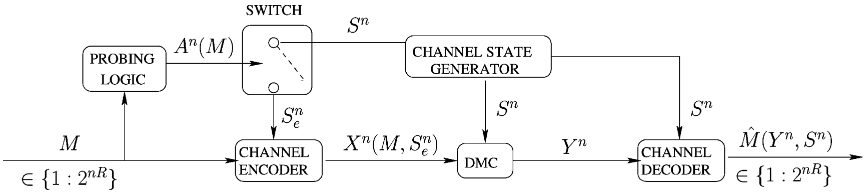

Fig. 1. Encoder takes message dependent actions to observe state, encodes using available partial state information noncausally while decoder knows the complete channel state sequence.

of optimal probing and that of channels with action dependent states in [8]. Motivated by the framework of communication over slow fading channels, where the information of channel states is to be exploited on the fly, we have in Section IV characterization of the *probing capacity* where encoder takes actions to get channel states and use them causally to construct channel inputs and decoder takes actions strictly causally dependent on channel outputs. Note that in this section, we characterize a novel and a generalized setting, where both encoder and decoder take costly actions to get channel state information. Later in this section, inspired by coding on computer memory with defects, we explain the noncausal case, i.e., when channel states are used noncausally by the encoder to generate channel input symbols and decoder waits for the entire channel output before taking actions to get channel states. This in general is a hard problem and we show its equivalence to a relay channel problem with infinite lookahead at the relay. In Section V, we work out several examples, with some surprising implications. The paper is concluded in Section VI with directions of future research.

II. To Observe or Not to Observe Channel States at Encoder

We begin by explaining the notation to be used throughout this paper. Let upper case, lower case, and calligraphic letter denote, respectively, random variables, specific or deterministic values which random variables may assume, and their alphabets. For two jointly distributed random variables, X and Y, let P_X, P_{XY} and $P_{X|Y}$, respectively, denote the marginal of X, joint distribution of (X, Y) and conditional distribution of X given Y. X_m^n is a shorthand for $n-m+1$ tuple $\{X_m, X_{m+1}, \ldots, X_{n-1}, X_n\}$. We impose the assumption of finiteness of cardinality on all alphabets, unless otherwise indicated.

In this section, we consider the problem of optimal probing where encoder takes a ‘costly’ action depending upon message and use it to probe the channel and observe or not the channel state. The actions are binary, hence while action, $A = 1$ corresponds to the case when encoder observes the channel state, action, $A = 0$ implies no acquired state information. We further assume decoder knows the complete state information and that the encoder uses partial state information noncausally to generate channel input symbol.

A. Problem Setup

The setting is depicted in Fig. 1: Message M is selected uniformly from a uniform distribution on the message set $\mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$. Nature generates states sequence $S^n \in \mathcal{S}^n$ i.i.d $\sim P_S$, independent of message. A $(2^{nR}, n)$ code consists of:

- *Probing Logic*: $f_A : M \to A^n \in \{0,1\}^n$ such that the action sequence A^n satisfies the cost constraints
$$\Lambda(A^n) = \frac{1}{n}\sum_{i=1}^{n}\Lambda(A_i) \leq \Gamma \tag{1}$$
where $\Lambda(\cdot)$ is the cost function while Γ is the cost constraint. Given nature generated state sequence S^n and message dependent action sequence A^n, encoder receives partial state information $S_e^n \in \{\{*\} \cup \mathcal{S}\}^n$ through a deterministic channel characterized by
$$S_e = h(S, A) = S \text{ if } A = 1 \tag{2}$$
$$S_e = h(S, A) = * \text{ if } A = 0 \tag{3}$$
where $*$ stands for erasure or no information of state symbol. Thus, $A = 1$ corresponds to an observation of the channel state while $A = 0$ to a lack of an observation. Without loss of generality we can assume, $\Lambda(0) = 0$.
- *Encoding*: $f_e : (M, S_e^n) \to X^n \in \mathcal{X}^n$, i.e., encoder uses the partial state information noncausally to generate channel input symbols.
- *Decoding*: $f_d : (Y^n, S^n) \to \hat{M} \in \{1, 2, \ldots, |\mathcal{M}|\}$, where the channel output $Y^n \in \mathcal{Y}^n$.

The joint PMF on $(M, A^n, S^n, S_e^n, X^n, Y^n, \hat{M})$ induced by a given scheme is
$$\begin{aligned} &P_{M,A^n,S^n,S_e^n,X^n,Y^n,\hat{M}}(m, a^n, s^n, s_e^n, x^n, y^n, \hat{m}) \\ &= \frac{\mathbf{1}_{\{f_A(m)=a^n, f_e(m,s_e^n)=x^n, f_d(y^n,s^n)=\hat{m}\}}}{|\mathcal{M}|}\prod_{i=1}^{n}P_S(s_i) \\ &\times \prod_{i=1}^{n}\mathbf{1}_{\{s_{e,i}=h(s_i,a_i)\}}P_{Y|X,S}(y_i|x_i, s_i). \end{aligned} \tag{4}$$

The probability of error is calculated as $P_e = P(M \neq \hat{M}(Y^n, S^n))$. The rate R is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes for increasing block lengths satisfying the cost constraints (1) with $\frac{1}{n}\log|\mathcal{M}| \leq R$ and $P_e^n \overset{n\to\infty}{\longrightarrow} 0$.

B. Probing Capacity

Theorem 1: The cost constrained *‘probing capacity’* of the system in Fig. 1 with channel inputs constructed using the observed state sequence noncausally while decoder has complete information of the state is given by
$$C(\Gamma) = \max[I(X; Y|S)], \tag{5}$$

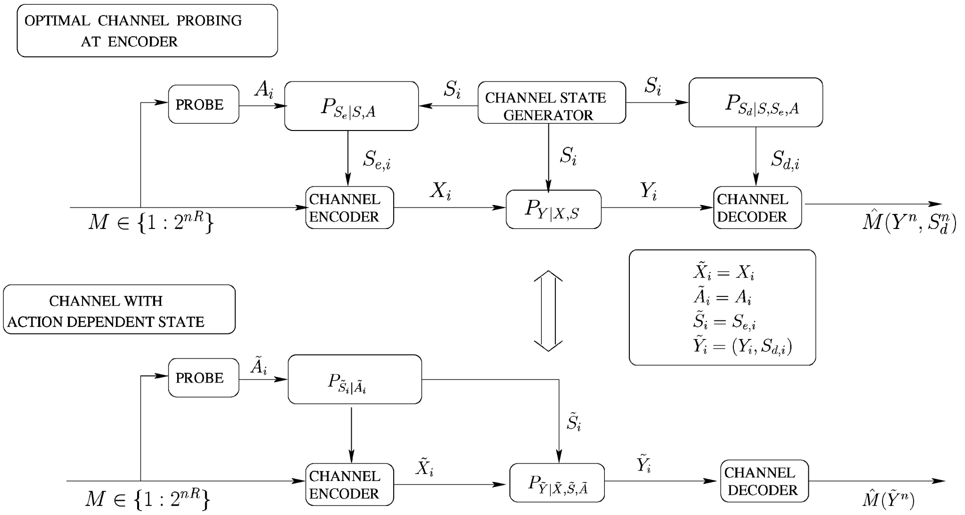

Fig. 2. Equivalence of our setting of probing the channel state at the encoder to that of channels with action dependent states in [8].

where maximization is over all joint distributions of the form

$$P_{A,S,S_e,X,Y} = P_A P_S \mathbf{1}_{\{S_e=h(S,A)\}} P_{X|S_e,A} P_{Y|X,S}, \qquad (6)$$

for some $P_A, P_{X|S_e,A}$ such that $E[\Lambda(A)] \leq \Gamma$.

Proof: We state theorems for generalized settings in Section III by drawing equivalence from [8, Theorems 1 and 2] and show how they can be used to prove this theorem. However for a standalone proof with a simpler achievability, see Appendix A. ■

Note 1 (Causal Probing): Note that the capacity is the same if we now consider the setting where the encoder generates channel input sequences using observed state causally. The converse for the noncausal setting provides the converse for the causal setting. As in the achievability of fading channels in [10], here also the achievability for noncausal probing uses the channel state symbols in an i.i.d manner, i.e., channel input symbol depends on the state only through the current state symbol. Hence the achievability remains same for the causal case. This establishes that causal probing capacity equal the noncausal probing capacity. Note that in general capacity for causal and noncausal probing might not be the same, it holds for our specific setting considered in Fig. 1.

Note 2 (Probing Independent of Messages): If action sequence is taken independent of message, *time sharing* is optimal. This is because when action sequence is independent of message, the setting is equivalent to the case when decoder knows the action. Let $C(0)$ and $C(1)$ denote the capacity at cost $\Gamma = 0$ and $\Gamma = 1$, respectively. The capacity in this case is

$$\begin{aligned} C(\Gamma) &= \max[I(X;Y|S,A)] & (7)\\ &= \max[p(A=0)I(X;Y|S,A=0) \\ &\quad + p(A=1)I(X;Y|S,A=1)] & (8)\\ &= p(A=0)C(0) + p(A=1)C(1). & (9)\end{aligned}$$

TABLE I
EQUIVALENCE OF SETTING IN [8] TO OUR FORMULATION OF OPTIMAL PROBING AT ENCODER

Action Dependent State Channels ([8])	Optimal Encoder State Probing
$\tilde{X}$	X
$\tilde{A}$	A
$\tilde{S}$	S_e
$\tilde{Y}$	(Y, S_d)

III. Equivalence Between Encoder Probing and Channels With Action-Dependent States

In the previous section, we motivated the basic problem of characterizing the capacity when observation of the channel state at the encoder comes at a price. We had further assumed that the decoder knew the complete state information. In this section, we point out the equivalence of general setting of action dependent channel probing at the encoder with the setting of channels with action dependent states considered in [8]. In our generalized setting, actions are taken in an alphabet $\mathcal{A}$ and encoder observes S_e through a DMC $P_{S_e|S,A}$. The setting in [8] and [9] is as follows. Given a message M, encoder takes actions $A^n = A^n(M)$, which affect the formation of channel states. These states are then used by the encoder causally or noncausally to generate channel input.

First consider the case when decoder does not know the channel states. Now in our setting we are given from nature $P_S, P_{S_e|S,A}, P_{Y|X,S}$, but this is equivalent to $P_{S_e|A}, P_{Y|X,S_e,A}$ since S^n is not available at encoder or decoder and hence can be averaged out. This establishes the equivalence as depicted in Table I and Fig. 2. If the decoder now knows the channel state S_d through DMC $P_{S_d|S,S_e,A}$ we can replace $\tilde{Y}$ in Fig. 2 with (Y, S_d) to compute capacity.

Hence, using the proven equivalence, we invoke and list theorems from [8] transformed for our setting.

Theorem 2 (Equivalent to Theorem 1 in [8].): The cost constrained *'probing capacity'* when the encoder generates channel inputs using partial state information *noncausally* as in Fig. 2 with cost constraint Γ (as in (1)), is given by

$$\begin{aligned} C_{nc}(\Gamma) &= \max[I(U;Y,S_d) - I(U;S_e|A)] & (10)\\ &= \max[I(A,U;Y,S_d) - I(U;S_e|A)] & (11)\end{aligned}$$

where maximization is over all joint distributions of the form

$$P_{A,S,S_e,U,S_d,X,Y} = P_A P_S P_{S_e|S,A} P_{U|S_e,A} P_{S_d|S,S_e,A} \times \mathbf{1}_{\{X=f(U,S_e)\}} P_{Y|X,S} \tag{12}$$

for some $P_A, P_{U|S_e,A}, f$ such that $E[\Lambda(A)] \leq \Gamma$ and $|\mathcal{U}| \leq |\mathcal{A}|\,|\mathcal{S}|\,|\mathcal{S}_e|\,|\mathcal{S}_d|\,|\mathcal{X}| + 3$.

Theorem 3 (Equivalent to Theorem 2 in [8].): The cost constrained *"probing capacity"* when the encoder generates channel inputs using partial state information *causally* as in Fig. 2 with cost constraint Γ (as in (1)), is given by

$$C_c(\Gamma) = \max[I(U; Y, S_d)] \tag{13}$$

where maximization is over all joint distributions of the form

$$P_{U,A,S,S_e,S_d,X,Y} = P_U \mathbf{1}_{\{A=g(U)\}} P_S P_{S_e|S,A} P_{S_d|S,S_e,A} \times \mathbf{1}_{\{X=f(U,S_e)\}} P_{Y|X,S} \tag{14}$$

for some P_U, g, f such that $E[\Lambda(A)] \leq \Gamma$ and $|\mathcal{U}| \leq \min\{|\mathcal{Y}|\,|\mathcal{S}_d|\,, |\mathcal{A}|\,|\mathcal{S}|\,|\mathcal{S}_e|\,|\mathcal{S}_d|\,|\mathcal{X}| + 3\}$.

Note 3: Note that auxiliary variable U has an increased cardinality as compared to equivalent setting in [8]. This stems from the following:

- Output Y is replaced with (Y, S_d), hence in causal setting we have $|\mathcal{U}| \leq |\mathcal{Y}|\,|\mathcal{S}_d|$ following the arguments in [8].
- To preserve $P_{A,S,S_e,S_d,X}$, in both causal and noncausal setting we have $|\mathcal{U}| \leq |\mathcal{A}|\,|\mathcal{S}|\,|\mathcal{S}_e|\,|\mathcal{S}_d|\,|\mathcal{X}| - 1$. In causal setting, four more elements are needed, one to preserve $H(Y, S_d|U)$, one to preserve independence of S with (A, U) and two more each to preserve markov chains $(S_e, S_d) - (S, A) - U$ and $X - (U, S_e) - (A, S, S_d)$. In non causal setting, four more elements (other than $|\mathcal{U}| \leq |\mathcal{A}|\,|\mathcal{S}|\,|\mathcal{S}_e|\,|\mathcal{S}_d|\,|\mathcal{X}| - 1$) are needed, one to preserve $H(S_e|A, U) - H(Y, S_d|U)$, one to preserve independence of S with A and two more to preserve markov chains, $U - (S_e, A) - (S, S_d)$ and $X - (U, S_e) - (A, S, S_d)$.

Deriving Theorem 1 using Theorems 2 and 3

Theorems 2 and 3 generalize the setting in Theorem 1, hence here we would like to derive the capacity results in Theorem 1 from Theorems 2 and 3. We have already pointed out that capacity of the setting in Fig. 1 is the same whether encoder encodes using partial information causally or noncausally (call it $C(\Gamma) = C_c(\Gamma) = C_{nc}(\Gamma)$). (Subscripts 'c' and 'nc' stand for capacity for causal and noncausal encoding of partial state information). We claim to prove the result $C(\Gamma) = C_c(\Gamma) = C_{nc}(\Gamma)$ using Theorems 2 and 3.

For noncausal encoding (using Theorem 2)

$$\begin{aligned} C_{nc}(\Gamma) &= \max[I(A, U; Y, S) - I(U; S_e|A)] && (15)\\ &= \max[I(A, U; Y|S) + I(A, U; S) \\ &\quad - I(U; S_e|A)] && (16)\\ &\overset{(a)}{=} \max[H(Y|S) - H(Y|S, A, U, S_e, X) \\ &\quad + I(U; S|A) - I(U; S_e|A)] && (17)\\ &= \max[H(Y|S) - H(Y|S, A, U, S_e, X) \\ &\quad - H(U|S, A, S_e) + H(U|S_e, A)] && (18)\\ &\overset{(b)}{=} \max[H(Y|S) - H(Y|S, X) - H(U|A, S_e) \\ &\quad + H(U|S_e, A)] && (19)\\ &= \max[I(X; Y|S)] && (20) \end{aligned}$$

where

- (a) follows from the fact that $S_e = h(S, A)$ and $X = f(U, S_e)$ and that A is independent of S.
- (b) follows from the DMC ($P_{Y|X,S}$) assumption and that $U - (S_e, A) - S$ is a Markov Chain.

This maximization is over joint distribution

$$\begin{aligned} &P_{A,S,S_e,X,Y} \\ &= \sum_U P_{A,S,S_e,U,X,Y} && (21)\\ &= \sum_U P_A P_S P_{S_e|S,A} P_{U|S_e,A} \mathbf{1}_{\{X=f(U,S_e)\}} P_{Y|X,S} && (22)\\ &\overset{(c)}{=} \sum_U P_A P_S P_{S_e|S,A} P_{U|S_e} \mathbf{1}_{\{X=f(U,S_e)\}} P_{Y|X,S} && (23)\\ &= P_A P_S P_{S_e|S,A} P_{X|S_e} P_{Y|X,S} && (24) \end{aligned}$$

where (c) follows from the fact that knowing S_e implies knowing A. Hence, we have from (20) and (24). $C_{nc}(\Gamma) = C(\Gamma)$.

Now for causal encoding (using Theorem 3)

$$\begin{aligned} C_c(\Gamma) &= \max[I(U; Y, S)] && (25)\\ &\overset{(d)}{=} \max[I(U; Y|S)] && (26)\\ &\overset{(e)}{=} \max[I(A, U; Y|S)] && (27)\\ &= \max[H(Y|S) - H(Y|S, A, U, S_e, X)] && (28)\\ &= \max[H(Y|S) - H(Y|S, X)] && (29)\\ &= \max[I(X; Y|S)] && (30) \end{aligned}$$

where (d) follows from the fact that U and S are independent and (e) follows from the relation $A = g(U)$. This maximization is over joint distribution

$$P_{U,A,S,S_e,X,Y} = P_U \mathbf{1}_{\{A=g(U)\}} P_S P_{S_e|S,A} \times \mathbf{1}_{\{X=f(U,S_e)\}} P_{Y|X,S}. \tag{31}$$

We will now show that joint distribution of the form in Theorem 1 is contained in (31). So the joint distribution in Theorem 1

$$P_{A,S,S_e,X,Y} = P_A P_S P_{S_e|S,A} P_{X|S_e,A} P_{Y|X,S}. \tag{32}$$

Now

$$\begin{aligned} &P_{A,Q,S,S_e,X,Y} \\ &\overset{(f)}{=} P_S P_{S_e|S,A} P_Q P_A \mathbf{1}_{\{X=F(S_e,A,Q)\}} P_{Y|X,S} && (33)\\ &\overset{(g)}{=} P_U \mathbf{1}_{\{A=g(U)\}} P_S P_{S_e|S,A} \mathbf{1}_{\{X=F(U,S_e)\}} P_{Y|X,S} && (34) \end{aligned}$$

where (f) follows from the Functional Representation Lemma ([11]), Q is independent of S_e, A and (g) follows from defining $U = (A, Q)$. Hence by (30) and (34) we have shown that $C_c(\Gamma) \geq C(\Gamma)$. But $C_c(\Gamma) \leq C_{nc}(\Gamma) = C(\Gamma)$. This completes the claim.

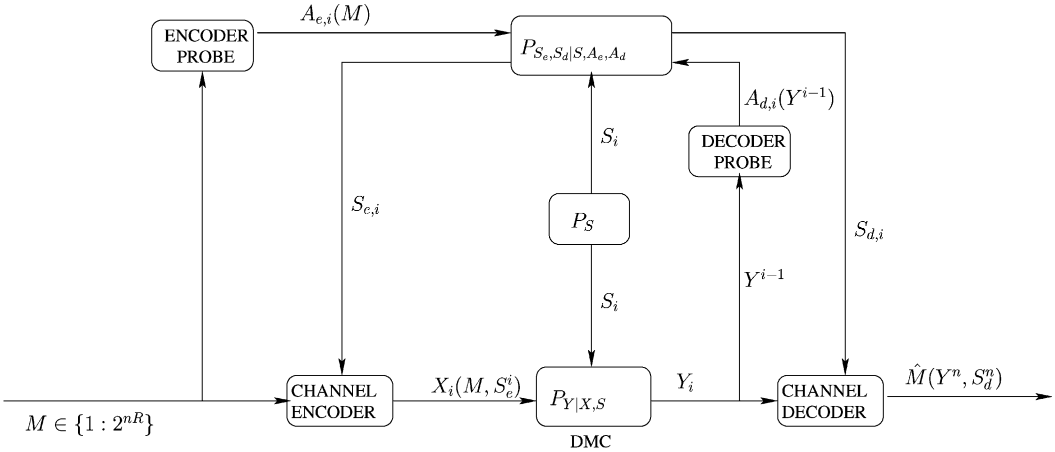

Fig. 3. Encoder and decoder both take actions to observe partial state information and use it for encoding and decoding.

IV. Optimal Probing at Both Encoder and Decoder

In earlier sections, we considered the framework where only encoder was allowed to take actions. In this section we further generalize the setting where decoder can also take actions based on the channel output and then obtain its own partial state information which is used to construct estimate of the transmitted message. We motivate this general setting in the framework of communication over slow fading Channels.

Consider a point to point communication system where in each time epoch channel state is i.i.d. $\sim P_S(s_i),\ s_i \in \mathcal{S}$. In the next epoch the information of this present state is lost, hence encoder and decoder have to exploit whatever information is available to them causally to get the best achievable rate. More precisely, consider the setup as depicted in Fig. 3: Message M is selected uniformly from a uniform distribution on the message set $\mathcal{M} = \{1, 2, \dots, |\mathcal{M}|\}$. Nature generates states sequence $S^n \in \mathcal{S}^n$ i.i.d $\sim P_S$, independent of message. A $(2^{nR}, n)$ code consists of:

- *Probing Logic*:
  - Encoder Probing Logic $f_{A_{e,i}} : M \to A_{e,i} \in \mathcal{A}_e$.
  - Decoder Probing Logic $f_{A_{d,i}} : Y^{i-1} \to A_{d,i} \in \mathcal{A}_d$, where channel output $Y \in \mathcal{Y}$.

  Further the encoder and decoder actions are cost constrained,

$$\Lambda(A_e^n, A_d^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(A_{e,i}, A_{d,i}) \le \Gamma \tag{35}$$

  where $\Lambda(\cdot,\cdot)$ is the cost function while Γ is the cost constraint. Given nature generated state sequence S^n, message dependent encoder action sequence A_e^n and channel output dependent decoder action sequence A_d^n, encoder acquires partial state information $S_e^n \in \mathcal{S}_e^n$ (which we will call CSIT, i.e., Channel State Information at Transmitter) and decoder $S_d^n \in \mathcal{S}_d^n$ (which we will call CSIR, i.e., Channel State Information at Receiver), through a DMC $P_{S_e,S_d|S,A_e,A_d}$.
- *Encoding*: $f_{e,i} : (M, S_e^i) \to X_i \in \mathcal{X}$.
- *Decoding*: $f_d : (Y^n, S_d^n) \to \hat{M} \in \{1, 2, \dots, |\mathcal{M}|\}$.

The joint PMF on $(m, a_e^n, a_d^n, s^n, s_e^n, s_d^n, x^n, y^n, \hat{m})$ induced by a given scheme is

$$\begin{aligned}
&P_{M,A_e^n,A_d^n,S^n,S_e^n,S_d^n,X^n,Y^n,\hat{M}}(\cdot)\\
&= \frac{1}{|\mathcal{M}|}\prod_{i=1}^{n} \mathbf{1}_{\{a_{d,i}=f_{A_{d,i}}(y^{i-1}\}}\mathbf{1}_{\{a_{e,i}=f_{A_{e,i}}(m)\}}\\
&\times \prod_{i=1}^{n} P_S(s_i) P_{S_e,S_d|S,A_e,A_d}(s_{e,i}, s_{d,i}|s, a_{e,i}, a_{d,i})\\
&\times \prod_{i=1}^{n} \mathbf{1}_{\{x_i=f_{e,i}(m,s_e^i)\}} P_{Y|X,S}(y_i|x_i, s_i) \times \mathbf{1}_{\{\hat{m}=f_d(y^n,s_d^n)\}}.
\end{aligned} \tag{36}$$

1) Probing Capacity:

Theorem 4: The cost constrained *"probing capacity"* for the scenario depicted in Fig. 3 is given by

$$C(\Gamma) = \max[I(U; Y, S_d|A_d)] \tag{37}$$

where maximization is over all joint distributions of the form

$$\begin{aligned}
&P_{S,A_d,U,A_e,S_e,X,Y,S_d}(s, a_d, u, a_e, s_e, x, y, s_d)\\
&= P_S(s)P_{A_d}(a_d)P_{U|A_d}(u|a_d)\mathbf{1}_{\{a_e=g(u,a_d)\}}\\
&\quad\times P_{S_e,S_d|S,A_e,A_d}(s_e, s_d|s, a_e, a_d)\mathbf{1}_{\{x=f(u,s_e,a_d)\}}P_{Y|X,S}
\end{aligned} \tag{38}$$

for some $P_{A_d}, P_{U|A_d}, g, f$ such that $\mathsf{E}[\Lambda(A_e, A_d)] \le \Gamma$ and $|\mathcal{U}| \le \min\{|\mathcal{Y}|\,|\mathcal{S}_d|\,|\mathcal{A}_d|, |\mathcal{S}|\,|\mathcal{A}_d|\,|\mathcal{A}_e|\,|\mathcal{S}_e|\,|\mathcal{S}_d|\,|\mathcal{X}| + 4\}$.

Proof:

Achievability: Fix $P_{A_d}, P_{U|A_d}, g, f$ which achieve $C(\frac{\Gamma}{1+\epsilon})$. Encoder and decoder decide on a sequence A_d^n, i.i.d $\sim P_{A_d}$. By similar arguments as in achievability of previous theorems using typical average lemma, constraints are satisfied. Now using Theorem 3 if $A_{d,i} = a\,\forall\, i$, error free communication is achieved if $R < I(U; Y, S_d|A_d = a)$. Hence since encoder and decoder both know A_d^n, we achieve $R < I(U; Y, S_d|A_d)$.

Converse: Suppose rate R is achievable. Now consider a sequence of $(2^{nR}, n)$ codes for which we have $P_e^n \overset{n\to\infty}{\longrightarrow}$. Consider

$$\begin{aligned} nR &= H(M) && (39)\\ &= I(M;Y^n,S_d^n) + H(M|Y^n,S_d^n). && (40)\end{aligned}$$

By Fano's Inequality ([12])

$$H(M|Y^n,S_d^n) \leq 1 + P_e^n R \leq n\epsilon_n, \qquad (41)$$

where $\epsilon_n \overset{n\to\infty}{\longrightarrow} 0$. Now consider

$$\begin{aligned} &I(M;Y^n,S_d^n)\\ &= H(Y^n,S_d^n) - H(Y^n,S_d^n|M) && (42)\\ &\overset{(a)}{=} \sum_{i=1}^{n} H(Y_i,S_{d,i}|Y^{i-1},S_d^{i-1},A_d^i)\\ &\quad - \sum_{i=1}^{n} H(Y_i,S_{d,i}|Y^{i-1},S_d^{i-1},M,A_d^i,A_e^n) && (43)\\ &\leq \sum_{i=1}^{n} H(Y_i,S_{d,i}|A_{d,i})\\ &\qquad - \sum_{i=1}^{n} H(Y_i,S_{d,i}|Y^{i-1},S_d^{i-1},S_e^{i-1},M,A_d^i,,A_e^n)\\ &&& (44)\\ &\overset{(b)}{=} \sum_{i=1}^{n} H(Y_i,S_{d,i}|A_{d,i}) - \sum_{i=1}^{n} H(Y_i,S_{d,i}|U_i,A_{d,i}) && (45)\\ &= \sum_{i=1}^{n} I(U_i;Y_i,S_{d,i}|A_{d,i}) && (46)\\ &\leq \sum_{i=1}^{n} C(\mathsf{E}[\Lambda(A_{e,i},A_{d,i})]) && (47)\\ &\overset{(c)}{\leq} nC(E[\Lambda(A_e^n,A_d^n)]) && (48)\\ &\overset{(d)}{\leq} nC(\Gamma) && (49)\end{aligned}$$

where

- (a) follows from the fact that $A_{d,i} = A_{d,i}(Y^{i-1})$ and $A_e^n = A_e^n(M)$.
- (b) follows by defining $U_i = (M,Y^{i-1},S_d^{i-1},S_e^{i-1},A_d^{i-1},A_e^n)$.
- (c) follows from the fact that $C(\Gamma)$ is concave in Γ. This is proved in Appendix B.
- (d) follows from the fact that $C(\Gamma)$ is non decreasing in Γ, which can be argued easily as larger Γ implies a larger feasible region and hence larger capacity.

We note the following relations:

- $A_{d,i} = A_{d,i}(Y^{i-1})$ is independent of S_i, it follows from proof of markov chain MC1 in Appendix C.
- We have the Markov Chains
  —$U_i - A_{d,i} - S_i$.
  —$A_{e,i} - (U_i,A_{d,i}) - S_i$.
  —$(S_{e,i},S_{d,i}) - (S_i,A_{e,i},A_{d,i}) - U_i$.
  —$X_i - (U_i,S_{e,i},A_{d,i}) - (A_{e,i},S_i,S_{d,i})$.
  —$Y_i - (X_i,S_i) - (U_i,A_{d,i},A_{e,i},S_{e,i},S_{d,i})$.
  These are proved in Appendix C.
- As U_i contains A_e^n, maximization is unaffected if we replace $P_{A_e|U,A_d}$ with $\mathbf{1}_{\{A_e=g(U,A_d)\}}$. Note that $I(U;Y,S_d|A_d)$ is convex in $P_{Y,S_d|U,A_d}$. This is due to the following standard arguments. Note $I(U;Y,S_d|A_d=a_d)$ is convex in $P(Y,S_d|U,A_d = a_d)$ for a fixed a_d. This implies $I(U;Y,S_d|A_d) = \sum_{a_d} p(A_d = a_d) f_{convex}(P(Y,S_d|U,A_d=a_d)) = \tilde{f}(P(Y,S_d|U,A_d))$, as $P(A_d)$ is fixed. But a convex combination of convex functions is indeed convex, this proves the claim or $\tilde{f}$ is convex. Convexity of $I(U;Y,S_d|A_d)$ in $P_{Y,S_d|U,A_d}$ implies convexity in $P_{X|U,S_e,A_d}$. Hence again, maximum would be unaffected if general $P_{X|U,S_e,A_d}$ is replaced with $X = f(U,S_e,A_d)$.
- Cardinality Bounds on U That set $\mathcal{U}$ needs no more than $|\mathcal{Y}|\,|\mathcal{S}_d|\,|\mathcal{A}_d|$ follows from arguments in [13]. Also, $\mathcal{U}$ needs $|\mathcal{S}|\,|\mathcal{S}_e|\,|\mathcal{A}_e|\,|\mathcal{A}_d|\,|\mathcal{S}_d|\,|\mathcal{X}| - 1$ to preserve $P_{S,A_e,A_d,S_e,S_d,X}$ (which preserves $H(Y_d,S_d|A_d)$), one element to preserve $H(Y,S_d|A_d,U)$, one element to preserve independence of S and A_d and three more to preserve the markov chains, $(U,A_e) - A_d - S$, $(S_e,S_d) - (S,A_e,A_d) - U$ and $X - (U,S_e,A_d) - (S,S_d,A_e)$.

The proof is then completed by using (40), (41) and (49) and letting $n \to \infty$. ■

Note 4: The result of Theorem 4 indicates that the A_d sequence is acting like a time sharing sequence, on which A_e will be embedded, highlighting the asymmetry between A_e and A_d. This also suggests that A_d need not depend on the channel outputs at all as it is acting like a time sharing sequence.

Note 5: Note that our analysis easily carries over to the case where there are multiple constraints, say with k cost functions $(\Lambda_1(\cdot),\ldots,\Lambda_k(\cdot))$ with cost constraints $(\Gamma_1,\ldots,\Gamma_k)$. A special case then would be when $k=2$, and $\Lambda_1(A_e,A_d) = \Lambda_1(A_e)$ and $\Lambda_2(A_e,A_d) = \Lambda_2(A_d)$, which is the setting with separate cost constraints on encoder and decoder actions.

Note 6: We can consider a more general setting where encoder and decoder feedback logic depend upon the respective past state observations, i.e., encoder takes actions, $A_{e,i}(M,S_e^{i-1})$, while decoder takes actions, $A_{d,i}(Y^{i-1},S_d^{i-1})$. While the achievability remains unchanged as in Theorem 4, it is easy to see the converse also holds with $U_i = (M,Y^{i-1},S_d^{i-1},S_e^{i-1},A_d^{i-1},A_e^i)$.

Note 7: (Computer Memory with Defects: Non-causal Probing at both Encoder and Decoder): Consider a computer memory with defects, as in what the encoder writes, X and what the decoder reads, Y are related to each other through a discrete memoryless channel, $P_{Y|X,S}$, where state S models defects. If there are no cost constraints to acquire the information about defects, encoder and decoder are better-off by coding and decoding using this entire state sequence S^n as it is available before writing and reading on the memory. Note that we assume neither the writing nor the reading operation changes the state. However when acquisition of this state information by the encoder as well as the decoder is cost constrained, encoder can take actions, $A_{e,i}(M)$ to get partial state information S_e^n

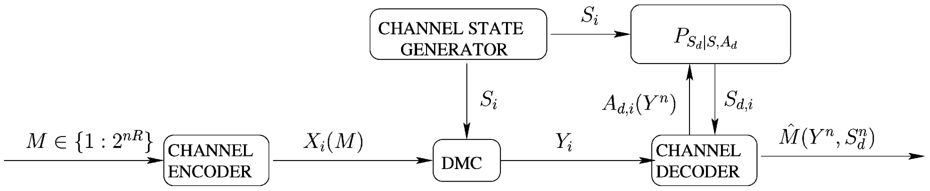

Fig. 4. Decoder takes actions dependent upon the entire observed channel output sequence and uses the actions to aquire partial channel state information. Encoder has no knowledge of channel states.

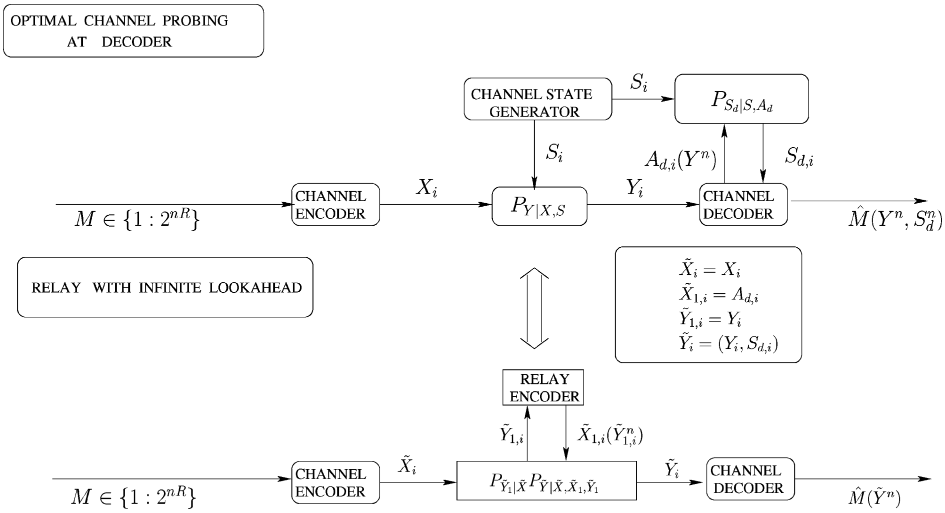

Fig. 5. Equivalence of setting in Fig. 4 with Relay with Infinite Lookahead.

and then write $X_i(M, S_e^n)$ while decoder can wait for entire memory to be written and then take actions, $A_{d,i}(Y^n)$. It will then obtain its side information S_d^n. Hence the setup remains similar as depicted in Fig. 3, the only difference from the setup in Section IV is that encoder now uses the partial state information, CSIT, noncausally to generate input symbols, i.e., $f_e : (M, S_e^n) \to X_i \in \mathcal{X}$, while decoder takes action based on entire channel output sequence, i.e., $f_{A_d} : Y^n \to A_{d,i} \in \mathcal{A}_d$. Also in order to avoid issues of instantaneous dependency, we must have

$$P_{S_e,S_d|S,A_e,A_d} = P_{S_e|S,A_e} \times P_{S_d|S,S_e,A_e,A_d}. \qquad (50)$$

Equivalence to Relay Problem: The above problem is in general a hard one. In fact even most of the special cases are open. For instance, consider a special case where A_e is binary, with cost function $\Lambda(A_e, A_d) = \Lambda(A_e)$, so there are no constraints on the action taken by the decoder A_d. Note $\Lambda(a) = a$, $a \in \{0, 1\}$ and we are interested in computing capacity as a function of cost constraint, $\Gamma = [0, 1]$. Under this special case too, the corner cases of zero cost and unity cost are open.

- For zero cost, the system is a special case of "Relay Channel with Infinite Lookahead", which is an open problem with only bounds known as in [11, Chapter 17]. We show the equivalence of this problem at zero cost to that of Relay with Infinite Lookahead, as depicted in Fig. 5 and Table II. In a standard relay, the relay encoder generates symbols strictly causally, i.e., $X_{1,i} = X_{1,i}(Y^{i-1})$ (See Fig. 5), in case of a relay with lookahead, relay encoding is in general $X_{1,i} = X_{1,i}(Y^{i+d})$ for a lookahead d. While $d = 0$ corresponds to the case of causal relay or relay without delay, in the case of relay with infinite lookahead or noncausal relay, relay encoding can depend on the entire sequence, Y_1^n.
- When cost is unity, similar to that of zero cost, the setting can be shown equivalent to the case of relay channel with states known noncausally to the encoder and relay has infinite lookahead. This problem too is in general open. When states are also available noncasually to the relay which instead of infinite lookahead has zero lookahead (the case of standard relay), authors in [14](cf Theorem 2.1) derive a lower bound on the capacity.

TABLE II
EQUIVALENCE OF SETTING IN FIG. 4 WITH RELAY WITH INFINITE LOOKAHEAD [11]

Relay with Infinite Lookahead ([11])	Decoder Probing in Fig. 4
$\tilde{X}$	X
$\tilde{X}_1$	A_d
$\tilde{Y}_1$	Y
$\tilde{Y}$	(Y, S_d)

V. NUMERICAL EXAMPLES

A. Discrete Channels

1) (Noncausal Probing): To Observe or Not to Observe Channel State at Encoder:

Example 1: (Binary States, $S(\alpha)$ channel and $Z(\beta)$, Decoder observes complete channel state): Consider the communication system shown in Fig. 6 with binary input and output. Decoder knows the state completely. Actions are binary which correspond to observe or not to observe state at encoder. Also the cost

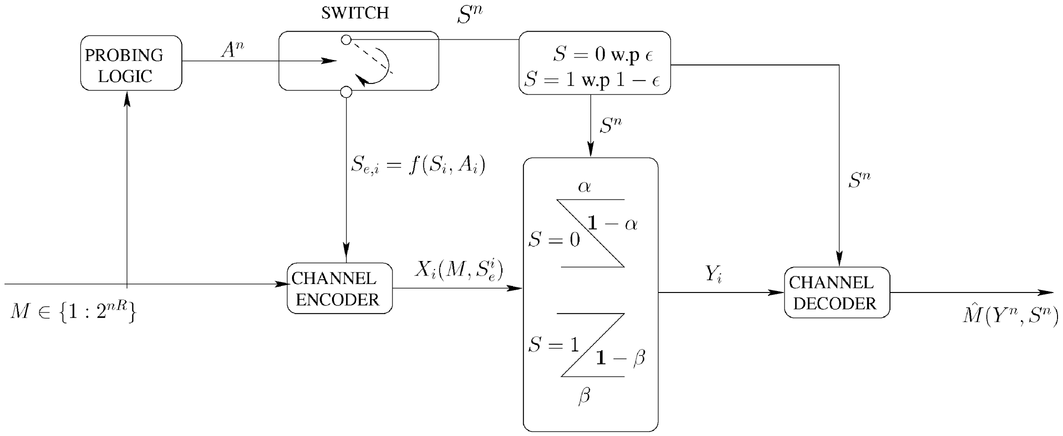

Fig. 6. Example 1.

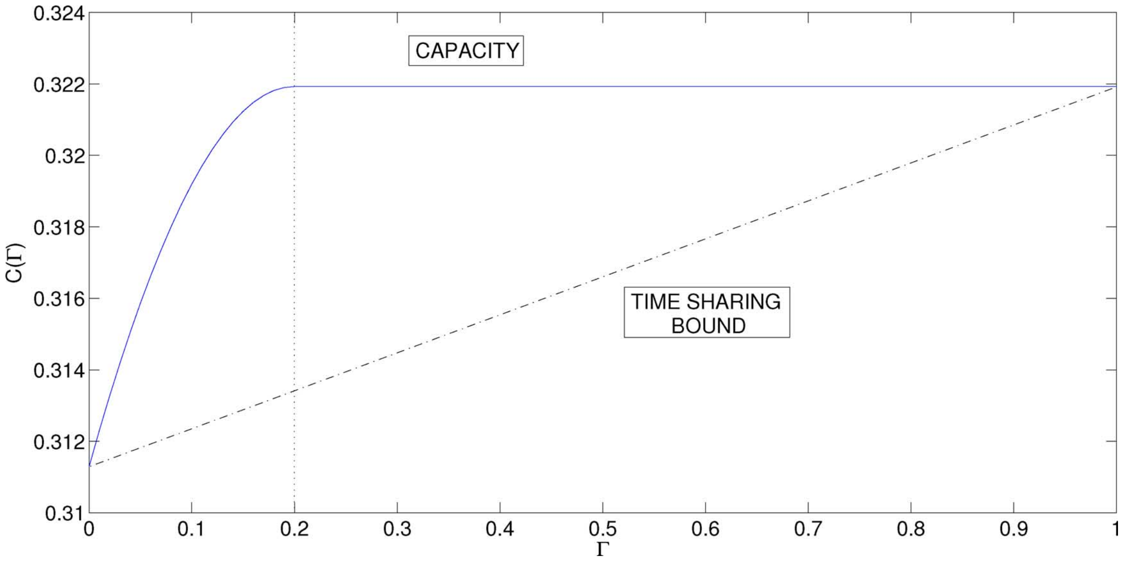

Fig. 7. Cost-capacity tradeoff for Example 1. Time sharing is strictly suboptimal.

function, $\Lambda(a) = a$, for actions, $a \in \{0, 1\}$. We compute the capacity using Theorem 1. $S_e \in \{*, 0, 1\}$ and $\alpha = \beta = \epsilon = 0.5$. We assume the following:

$$P(X = 0|S_e = *) = p_1 \quad (51)$$

$$P(X = 0|S_e = 0) = p_2 \quad (52)$$

$$P(X = 0|S_e = 1) = p_3. \quad (53)$$

As $C(\Gamma)$ is non decreasing in Γ. $P(A = 1) = \Gamma$. We obtain for $\Gamma \in [0, 1]$

$$\begin{aligned} &C(\Gamma) \\ &= \max_{p_1, p_2, p_3 \in [0,1]} [\epsilon h_2 (\alpha((1 - \Gamma)p_1 + \Gamma p_2)) \\ &- \epsilon((1 - \Gamma)p_1 + \Gamma p_2) h_2(\alpha) \\ &+ (1 - \epsilon) h_2 (\beta((1 - \Gamma)(1 - p_1) + \Gamma(1 - p_3))) \\ &- (1 - \epsilon)((1 - \Gamma)(1 - p_1) + \Gamma(1 - p_3)) h_2(\beta)]. \quad (54) \end{aligned}$$

We compute the above expression numerically (Fig. 7). Note here that decoder knows the complete state, hence by the note at the end of Theorem 1, the capacity remains the same even if there is causal probing.

*Note 8 (*Cut-Off Point ≈ 0.2 *in Fig. 7):* An observation from this example which is perhaps somewhat surprising is that in order to achieve the maximum capacity (which is at $\Gamma = 1$) one needs to only observe a fraction of states ≈ 0.2. This threshold however can also be theoretically derived. Essentially, we find out the range of $\Gamma \in [0, 1]$ for which the capacity achieving joint distribution in $C(\Gamma)$ induces exactly the same marginals, $P_{X|S}$ as when the cost is unity. Let p_1^*, p_2^*, and p_3^* be optimal distributions for cost Γ as in (54). The marginals are equal to

$$P(X = 0|S = 0) = (1 - \Gamma)p_1^* + \Gamma p_2^* \quad (55)$$

$$P(X = 0|S = 1) = (1 - \Gamma)p_1^* + \Gamma p_3^*. \quad (56)$$

For $\Gamma = 1$, we can easily compute $P(X = 0|S = 0) = 0.4$ and $P(X = 0|S = 1) = 0.6$. Therefore, for the marginals to be same

$$(1 - \Gamma)p_1^* + \Gamma p_2^* = 0.4 \quad (57)$$

$$(1 - \Gamma)p_1^* + \Gamma p_3^* = 0.6 \quad (58)$$

or

$$\Gamma(p_3^* - p_2^*) = 0.2. \quad (59)$$

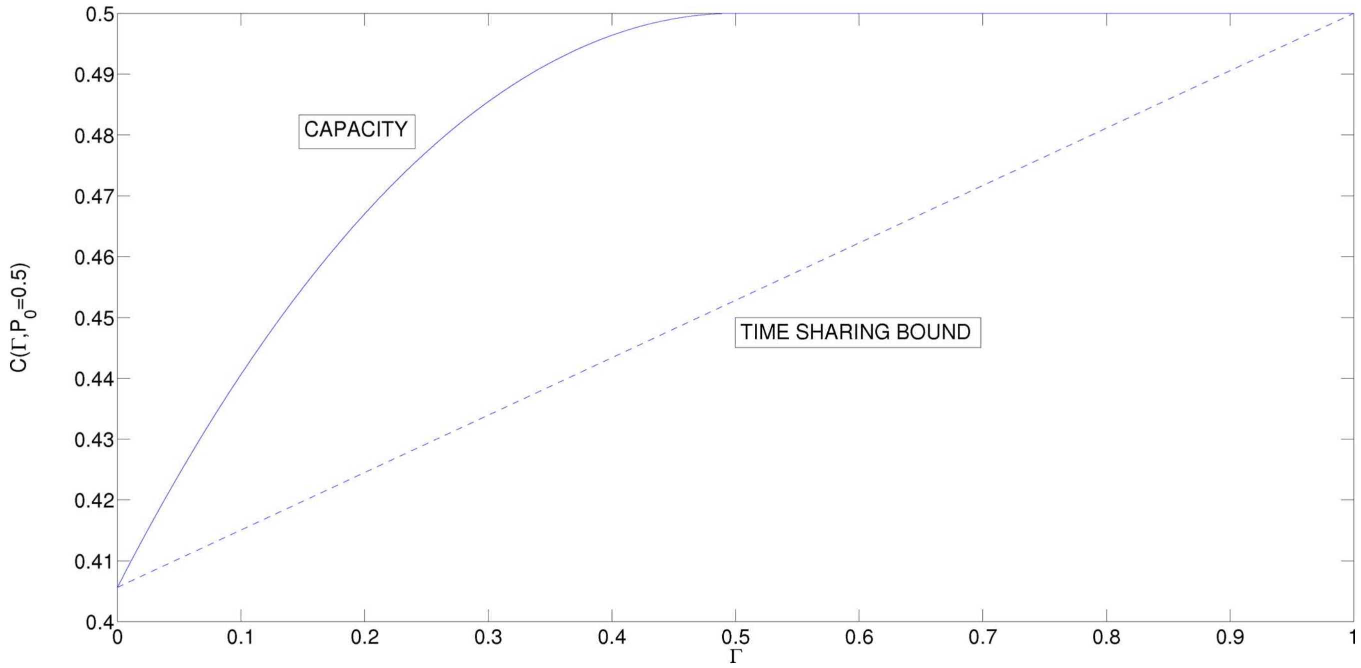

Fig. 8. Cost-capacity tradeoff for Example 2 for $P_0 = 0.25$. The dotted straight line is obtained by time sharing between zero cost and unit cost capacity.

Since $p_2^*, p_3^* \in [0,1]$, it is easy to see that if the cost $\Gamma \gtrsim 0.2$, we can find (p_1^*, p_2^*, p_3^*) such that $C(\Gamma) = C(1)$. At $\Gamma = 0.2$, optimal scheme is $X = S_e \oplus 1$ if $S_e \neq *$, and $Bern(0.5)$ otherwise. Note that this kind of phenomenon is particular to the example we consider here, in general it would be dependent on the channel parameters of the problem.

Example 2: Binary States, Multiplier Channel with Power Constraints. Decoder has complete state information: Consider a multiplier channel with binary inputs, outputs and states, $Y = S \cdot X$ where $S \sim Bern(0.5)$. Again note that actions are binary with $\Lambda(a) = a$ and $A = 1$ corresponds to an observation of the channel state while $A = 0$ to a lack of an observation. Let

$$p_* = p(x=1|s_e=*) \quad (60)$$
$$p_0 = p(x=1|s_e=0) \quad (61)$$
$$p_1 = p(x=1|s_e=1). \quad (62)$$

We see that capacity under the power constraint

$$p(x=1) \leq P_0 \in [0,1] \quad (63)$$

is

$$C(\Gamma, P_0) = \max \frac{1}{2} h_2 \left[(1-\Gamma)p_* + \Gamma p_1\right] \\ \text{subject to} \\ (1-\Gamma)p_* + \frac{\Gamma}{2}(p_0 + p_1) = P_0. \quad (64)$$

For $P_0 = 0.25$, we have

$$C(\Gamma, P_0 = 0.25) = 0.5 h_2 \left[\frac{1+2\Gamma}{4}\right] \text{ if } \Gamma \in [0, 0.5] \quad (65)$$
$$C(\Gamma, P_0 = 0.25) = 0.5 \text{ if } \Gamma \in [0.5, 1]. \quad (66)$$

The plot for $P_0 = 0.25$ is shown in Fig. 8.

Note that in both the above examples, the decoder knows complete state, hence by Note 1 capacity remains the same when there is causal probing.

2) (Causal Probing): To Observe or Not to Observe Channel State at Encoder:

Example 3: (Binary States, $S(\alpha)$ channel and $BSC(\delta)$, Decoder has no access to the state): Consider the communication system shown in Fig. 9 with binary input and output with $\epsilon = 0.5$, $\alpha = 0.1$ and $\delta = 0.3$. Here, states are not known to the decoder and encoder uses partial state information causally to generate channel input symbols. Actions are binary with cost, $\Lambda(a) = a$. $A = 1$ corresponds to an observation of the channel state while $A = 0$ to a lack of an observation. The evaluation of capacity expression involves an auxiliary random variable. We compute its lower bound on capacity numerically using Theorem 3 as shown in Fig. 10. Here also, time sharing is clearly not optimal.

Note 9: Note the interesting phenomenon in this example too (as in Example 1), where we just need to observe roughly a fraction of $\text{state} \sim 0.5$ to obtain the capacity at unit cost. This can be reasoned in a similar manner as reasoned for Example 1.

B. Continuous Channels

1) "Learning" to Write on a Dirty Paper: Using standard arguments, it can be shown that the capacity results carry over to the case of continuous channels with power constraints on input symbols. Let us recall the setting in Dirty Paper Coding. Costa in [15] considered the communication system as in Fig. 11.

The output of the channel is given as $Y^n = X(M, S^n) + S^n + Z^n$, where

- Channel state or Interference S^n is i.i.d. $S^n \sim \mathcal{N}(0, QI)$ independent of i.i.d. noise, $Z^n \sim \mathcal{N}(0, NI)$.
- Channel state or interference is known to the encoder noncausally. Encoder hence generates channel

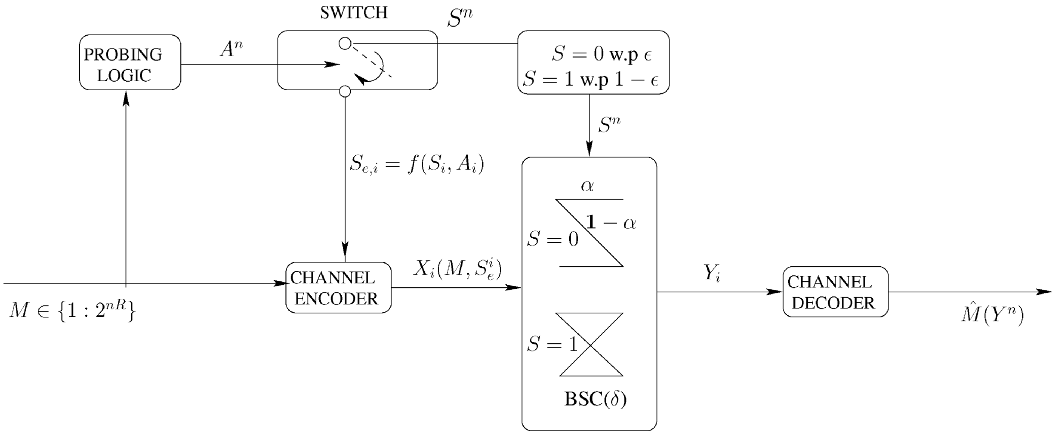

Fig. 9. Example 3.

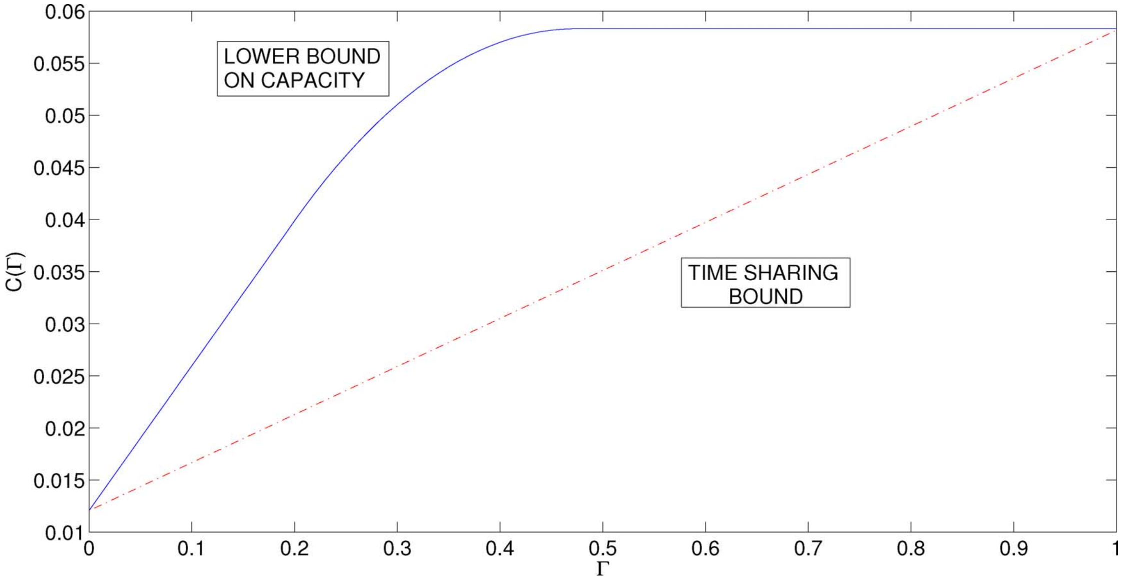

Fig. 10. Cost-capacity tradeoff for Example 3. The dotted straight line is obtained by time sharing between zero cost and unit cost capacity (Scheme 1). Time sharing between a scheme for which $A=g(U)=U$ in Theorem 3 (call it Scheme 2) and Scheme 1 gives a lower bound on the capacity indicated by solid line. It is evident that naive Scheme 1 (time sharing scheme between extreme capacities at zero and unit cost) is strictly sub-optimal.

inputs $X^n(M,S^n)$ which are cost constrained, i.e., $\frac{1}{n}\sum_{i=1}^{n}X_i^2\le P$.

- Decoder has no knowledge of channel state or interference.

It was shown that the capacity of this channel is $C(P/N)=\frac{1}{2}\log_2(1+P/N)$ which is equal to the capacity of a standard Gaussian channel with signal to noise ratio P/N. This is strictly larger than the capacity when S^n is unknown to both encoder and decoder, i.e., $\frac{1}{2}\log_2(1+P/(Q+N))$.

We now consider the setting as in Fig. 12. While in Writing on Dirty Paper, it was assumed that interference or channel state was completely available, but this might not be true in real systems one might have to pay a price to acquire this information. Hence, in contrast to writing on a paper where intensity and positions of all dirt spots are known, we have to take action to learn where the paper is most dirty, hence the name Learning to Write on a Dirty Paper. Actions are binary, with cost function, $\Lambda(a)=a$. Here also $A=1$ corresponds to an observation of the channel state while $A=0$ to a lack of an observation. Also,

$$S_e=h(S,A)=*\ \text{ if } A=0 \tag{67}$$

$$S_e=h(S,A)=S\ \text{ if } A=1 \tag{68}$$

where $*$ stands for erasure or no information.

Invoking Theorem 2, we have the capacity

$$\begin{aligned} C(\Gamma,P)&=\max[I(U;Y)-I(U;S_e|A)] &(69)\\ &=\max[I(A,U;Y)-I(U;S_e|A)] &(70)\end{aligned}$$

where maximization is over joint distribution

$$\begin{aligned} f_{A,U,S,S_e,X,Y}=&P_Af_Sf_{U|S_e,A}\mathbf{1}_{\{S_e=h(S,A)\}}\\ &\times\mathbf{1}_{\{X=f(U,S_e)\}}f_{Y|X,S} \end{aligned} \tag{71}$$

such that $p(A=1)\le\Gamma$ and $\mathsf{E}[X^2]\le P$. We give a lower bound on this capacity by considering a simple *power splitting* achievable scheme. Let us assume $X|(A=0)\sim\mathcal{N}(0,P_1)$ and $X|(A=1)\sim\mathcal{N}(0,P_2)$. Clearly $C(\Gamma,P)$ is maximized when $p(A=1)=\Gamma$. Therefore, we have from power constraints

$$(1-\Gamma)P_1+\Gamma P_2\le P. \tag{72}$$

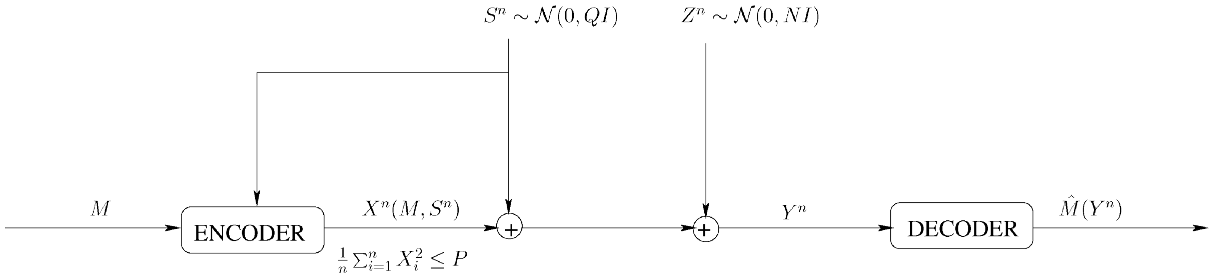
$S^n \sim \mathcal{N}(0, QI)$
$Z^n \sim \mathcal{N}(0, NI)$
M
ENCODER
$X^n(M, S^n)$
$\frac{1}{n}\Sigma_{i=1}^n X_i^2 \leq P$
Y^n
DECODER
$\hat{M}(Y^n)$

Fig. 11. Dirty Paper Coding as in [15].

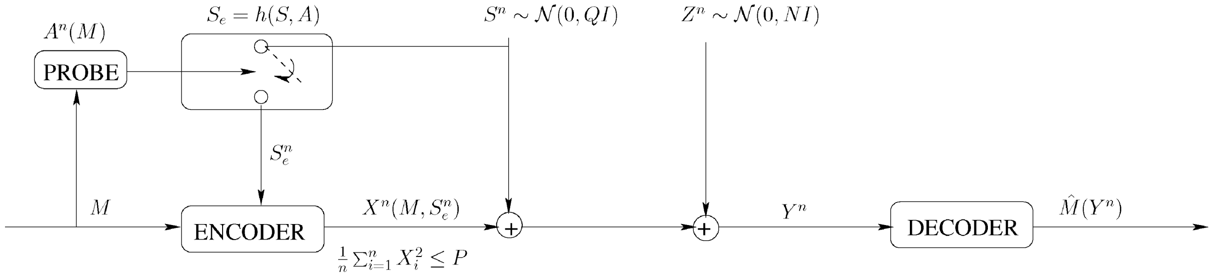
$A^n(M)$
PROBE
$S_e = h(S, A)$
$S^n \sim \mathcal{N}(0, QI)$
$Z^n \sim \mathcal{N}(0, NI)$
S_e^n
M
ENCODER
$X^n(M, S_e^n)$
$\frac{1}{n}\Sigma_{i=1}^n X_i^2 \leq P$
Y^n
DECODER
$\hat{M}(Y^n)$

Fig. 12. Learning to write on a Dirty Paper.

Further we assume, given action A, channel input X is independent of U, S, Z. Let

$$U|(A = 0) = X|(A = 0) \tag{73}$$
$$U|(A = 1) = X|(A = 1) + \alpha(P_2)S \tag{74}$$

where $\alpha(P_2) = P_2/(P_2 + 1)$. Since $Y = X + S + Z$, we have

$$Y|A = 0 \sim g_0 = \mathcal{N}(0, P_1 + Q + N) \tag{75}$$
$$Y|A = 1 \sim g_1 = \mathcal{N}(0, P_2 + Q + N). \tag{76}$$

Hence $Y \sim g = (1-\Gamma)\mathcal{N}(0, P_1+Q+N)+\Gamma\mathcal{N}(0, P_2+Q+N)$. Considering this distribution gives the following lower bound on capacity:

$$
\begin{aligned}
C_{lower} &= \max_{P_1,P_2}[I(A, U; Y) - I(U; S_e|A)] &(77)\\
&= \max_{P_1,P_2}[I(A; Y) + I(U; Y|A) - I(U; S_e|A)] &(78)\\
&\overset{(a)}{=} \max_{P_1,P_2}[h(g) - (1 - \Gamma)h(g_0) - \Gamma h(g_1)\\
&\quad + (1 - \Gamma)I(X; Y|A = 0)\\
&\quad + \Gamma(I(U; Y|A = 1) - I(U; S|A = 1))] &(79)\\
&\overset{(b)}{=} \max_{P_1,P_2}[h(g) - (1 - \Gamma)h(g_0) - \Gamma h(g_1)\\
&\quad + (1 - \Gamma)C(P_1/(Q + N)) + \Gamma C(P_2/N)] &(80)
\end{aligned}
$$

where

- (a) follows from the fact that S_e is just erasure for $A = 0$, while for $A = 1$ is equal to S. $h(g)$ denotes the differential entropy of a continuous random variable with distribution g.
- (b) follows from the fact that when $A = 0$

$$
\begin{aligned}
&I(X; Y|A = 0)\\
&= h(Y|A = 0) - h(Y|X, A = 0) &(81)\\
&= h(\mathcal{N}(0, P_1 + Q + N)) - h(\mathcal{N}(0, Q + N) &(82)\\
&= \frac{1}{2}\log_2(1 + P_1/(Q + N)) = C(P_1/(Q + N)) &(83)
\end{aligned}
$$

while for $A = 1$ following the similar steps as in [15, Eqs. 3–7] we obtain

$$
\begin{aligned}
&I(U; Y|A = 1) - I(U; S|A = 1)\\
&= \frac{1}{2}\log_2(1 + P_2/N) = C(P_2/N). &(84)
\end{aligned}
$$

Fig. 13 shows the plot of C_{lower} with Γ for $P = Q = N = 1$, which indeed performs better than naive time sharing between $C(P/N)$ and $C(P/(Q + N))$.

2) Fading Channels With Power Control: We revisit the setting of fading channels with encoder and decoder state information as in [10], but now the encoder takes actions to acquire the channel state from receiver state estimation, while decoder already knows the channel state. This is depicted in Fig. 14. Here $g[i]$ denotes the i.i.d. channel states which take value in a finite state, $\mathcal{S} = \{g_1, g_2\}$ with equal probability. $n[i]$ is i.i.d. Gaussian noise $\sim \mathcal{N}(0, N/2)$. Bandwidth for communication is B. $\gamma_1 = \frac{Pg_1}{NB}$ and $\gamma_2 = \frac{Pg_2}{NB}$ are signal to noise ratios, such that $\gamma_1 < \frac{\gamma_2}{1+2\gamma_2}$. Actions are binary which correspond to observe or not to observe state at encoder with cost functions $\Lambda(a) = a$ and cost constraint Γ. f is defined as in Theorem 1, i.e, $g_e[i] = f(g[i], a[i]) = g[i]$ if $a[i] = 1$ else $g_e[i]$ is an erasure, i.e., we do not know what is the channel state $g[i]$. From results in [10], we know that:

- Capacity when only decoder knows the state information

$$C(0) = \frac{B}{2}\log_2(1 + \gamma_1) + \frac{B}{2}\log_2(1 + \gamma_2). \tag{85}$$

- Capacity when encoder also knows the channel state (possibly through a noiseless feedback from decoder) in addition to decoder

$$C(1) = \frac{B}{2}\log_2(1 + 2\gamma_2). \tag{86}$$

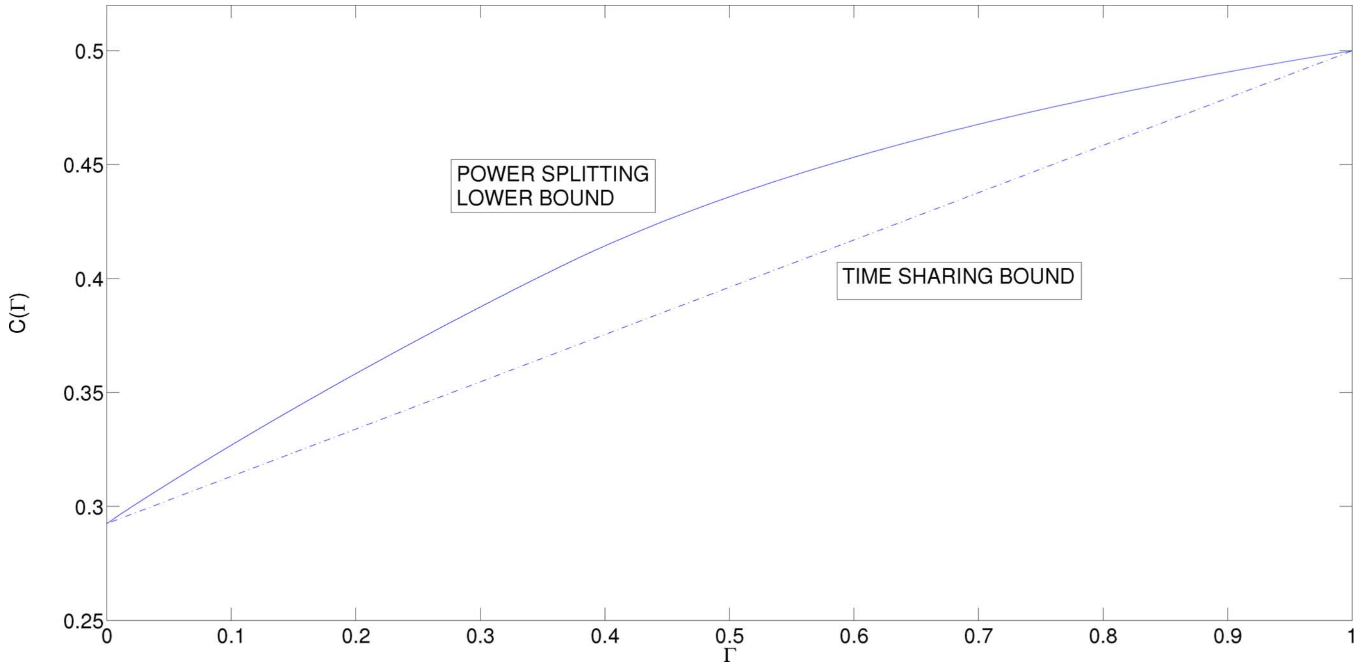

Fig. 13. Power Splitting lower bound on capacity for Learning to Write on Dirty Paper in Fig. 12.

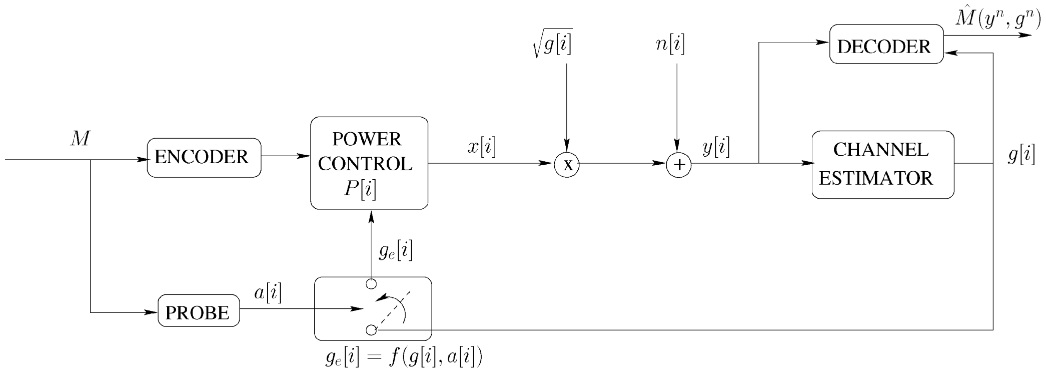

Fig. 14. Fading channels with encoder taking actions to acquire channel state for adaptive power control.

The above capacities form the extreme cases of zero and unit cost respectively for the communication system in Fig. 14. Using Theorem 1, we have the capacity for the communication system in Fig. 14 with bandwidth B as

$$\begin{aligned} C &= \max_{P_A f_{X|S_e}} 2BI(X;Y|S) && (87)\\ &= \max_{P_A, f_{X|S_e}} 2B[h(Y|S) - h(\mathcal{N}(0,NB))] && (88)\end{aligned}$$

such that $\mathsf{E}[\Gamma(A)] \leq \Gamma$ and $\mathsf{E}[X^2] \leq P$. Clearly maximum is attained for $p(A=1) = \Gamma$. To obtain a lower bound, we assume the following:

$$\begin{aligned} X|(S_e = *) &\sim \mathcal{N}(0, P_*) && (89)\\ X|(S_e = g_1) &\sim \mathcal{N}(0, P_1) && (90)\\ X|(S_e = g_2) &\sim \mathcal{N}(0, P_2). && (91)\end{aligned}$$

This implies $Y|(S = g_1) \sim (1-\Gamma)\mathcal{N}(0, NB + P_* g_1) + \Gamma\mathcal{N}(0, NB + P_1 g_1)$ and $Y|(S = g_2) \sim (1-\Gamma)\mathcal{N}(0, NB + P_* g_2) + \Gamma\mathcal{N}(0, NB + P_2 g_2)$, with power constraints

$$\mathsf{E}[X^2] = (1-\Gamma)P_* + \frac{\Gamma}{2}(P_1 + P_2) \leq P. \qquad (92)$$

Hence, a lower bound on capacity is

$$\begin{aligned} &C_{lower}(\Gamma, P)\\ &= 2B \max_{P_*, P_1, P_2} \left[\frac{h(f_{Y|S=g_1}) + h(f_{Y|S=g_2})}{2} - h(\mathcal{N}(0, NB)) \right],\\ &\text{subject to } (1-\Gamma)P_* + \frac{\Gamma}{2}(P_1 + P_2) \leq P.\end{aligned}$$

We plot $C_{lower}(\Gamma, P)$ as a function of Γ for $P = N = 1$, and $g_1 = 0.01$, $g_2 = 1$ in Fig. 15.

VI. Conclusion

In this work, we obtain "*Probing Capacity*" of systems which are characterized as follows:

- Channel is DMC with i.i.d states.
- Encoder takes *costly* actions and probes the channel for channel state information. This may be used causally or noncausally to generate channel input symbols.
- Decoder takes *costly* actions and probes the channel to obtain state information which is then used to construct message estimate.

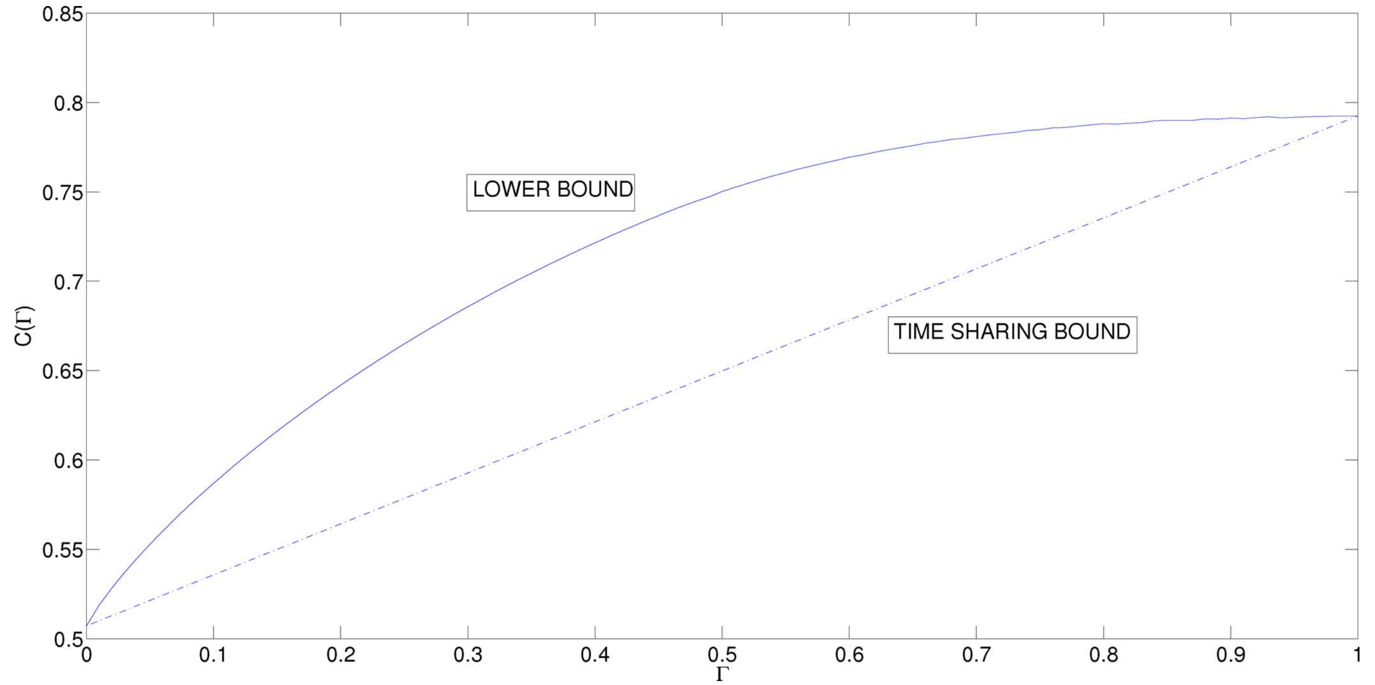

Fig. 15. Lower bound on fading channel communication system in Fig. 14. Time sharing is evidently highly suboptimal.

We also worked out examples of discrete and continuous channels in cases where only encoder probed the channel for states. We not only showed that a naive time sharing scheme is strictly suboptimal but also showed a pleasing phenomenon (see Example 1 in Section V) where one needs to observe only a fraction of states to obtain maximum rate of transmission i.e., rate when cost of state observation at encoder is not constrained.

As directions of future work, the following are important questions/conjectures worth spending time and energy:

1. What if encoder actions depend on past sampled state, i.e., $A_{e,i} = A_{e,i}(M, S_e^{i-1})$ for the case when partial state information is to be used noncausally? Can capacity be increased?
2. What about probing capacity for channels with *memory*?
3. Does the Example 4 on 'Learning to write on a dirty paper' also support the pleasing phenomenon when we can observe only a fraction of states and still achieve Costa's dirty paper coding capacity, $C(P/N)$?
4. What if we take action to sample or not feedback at encoder or decoder for channels with memory?

Some of the results concerning sampling or not the feedback for finite state channels (FSC) have been characterized in [16], while the rest are under investigation.

Appendix A
Proof of Theorem 1

1) Achievability: We use *Rate-Splitting* and *Multiplexing* to achieve capacity (for a similar scheme refer to [10]). Note that in this problem while knowing S_e, we know A, hence we would show achievability with $P_{X|S_e,A}$ replaced by $P_{X|S_e}$. Without loss of generality we assume $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, hence $\mathcal{S}_e = \{*, 1, 2, \ldots, |\mathcal{S}|\}$. Fix $P_A, P_{X|S_e}$ which achieve $C(\frac{\Gamma}{1+\epsilon})$. We *split* message M of rate R into two messages M_1 and M_2 of rate R_1 and R_2, respectively.

- Generation of Codebooks:
  - —Generate codebook $\mathcal{C}_A$ of $\{A^n(m_1)\}_{m_1=1}^{2^{nR_1}}$ n-tuples i.i.d. $\sim$ P_A. To send message $M = (M_1, M_2)$, if $A^n(M_1) \in T_\epsilon^n(A)$ (T_ϵ^n are typical in the sense of [17]), then action $A^n(M_1)$ is taken, else $A_0^n = (0, 0, \ldots, 0)$ is taken. If $A^n(M_1) \in T_\epsilon^n(A)$, then by typical average lemma [11], constraints are satisfied as

$$\Lambda(A^n) = \frac{1}{n}\sum_{i=1}^{n}\Lambda(A_i) \leq (1+\epsilon)E(\Lambda(A)) = \Gamma. \qquad (93)$$

  - —For every $A^n(m_1)$, generate a codebook $\mathcal{C}_X(m_1)$ of $\{(X_e^n(m_1, m_2), X_1^n(m_1, m_2), \ldots, X_{|\mathcal{S}|}^n(m_1, m_2))\}_{m_2=1}^{2^{nR_2}}$ $(|\mathcal{S}| + 1)n$-tuples such that X_e^n, $X_1^n, \ldots, X_{|\mathcal{S}|}^n$ are i.i.d. $\sim$ $P_{X|S_e=*}, P_{X|S_e=1}, \ldots, P_{X|S_e=|\mathcal{S}|}$ respectively. Also generate a codebook $\mathcal{C}_X^0$ of codewords $\{(X_{0,e}^n(m_2)\}_{m_2=1}^{2^{nR_2}}$ i.i.d. $\sim P_{X|S_e=*}$.
- *Encoding:*
  - —Given a message $M = (M_1, M_2)$, encoder decides to take actions $A^n(M_1)$ or A_0^n depending whether $A^n(M_1)$ is in $T_\epsilon^n(A)$ or not. If $A^n(M_1) \in T_\epsilon^n(A)$ encoder finds $S_e^n = h(A^n(M_1), S^n)$, and then sends $X^n(M_1, M_2)$ using the following *multiplexing*:

$$X_i = X_{e,i}(M_1, M_2) \text{ if } S_{e,i} = *, \qquad (94)$$
$$X_i = X_{j,i}(M_1, M_2) \text{ if } S_{e,i} = j \in \{1, 2, \ldots, |\mathcal{S}|\}. \qquad (95)$$

    If $A^n(M_1) \notin T_\epsilon^n(A)$, encoder sends $X_0^n(M_2)$.
- *Decoding*: We perform *Successive Decoding* and *Demultiplexing*. By successive decoding we mean that actions are decoded first by decoder and then the actual codewords.
  - —On obtaining the channel output sequence Y^n and channel state sequence S^n decoder finds the smallest value of $\hat{M}_1$ for which $(A^n(\hat{M}_1), Y^n, S^n) \in$

$T_\epsilon^n(A, Y, S)$. If there is no such $\hat{M}_1$, decoder assumes $\hat{M}_1 = 1$.

- Once the decoder decodes the value of M_1, if $A^n(M_1) \in T_\epsilon^n(A)$, it knows $S_e^n = h(A^n(M_1), S^n)$ and hence, using the codebook $C_X(M_1)$, it *demultiplexes*

$$\{(X_e^n(M_1, m_2), X_1^n(M_1, m_2), \ldots, X_{|\mathcal{S}|}^n(M_1, m_2))\}_{m_2=1}^{2^{nR_2}}$$

to construct $X^n(M_1, m_2)_{m_2=1}^{2^{nR_2}}$ sequences as

$$X_i(M_1, m_2) = X_{e,i}(M_1, m_2) \text{ if } S_{e,i} = *, \tag{96}$$

$$X_i(M_1, m_2) = X_{j,i}(M_1, m_2) \quad \text{if } S_{e,i} = j \in \{1, 2, \ldots, |\mathcal{S}|\}. \tag{97}$$

- After demultiplexing, if $A^n(M_1) \in T_\epsilon^n(A)$, decoder finds the smallest value of $\hat{M}_2$ for which $(X^n(M_1, \hat{M}_2), Y^n | S^n, A^n(M_1)) \in T_\epsilon^n(X, Y | S^n, A^n(M_1)$. If there is no such $\hat{M}_2$, decoder assumes $\hat{M}_2 = 1$. If $A^n(M_1) \notin T_\epsilon^n(A)$, decoder finds the smallest value of $\hat{M}_2$ for which $(X_0^n(\hat{M}_2), Y^n | S^n) \in T_\epsilon^n(X, Y | S^n)$, else $\hat{M}_2 = 1$ is assumed.

- Analysis of Probability of Error: Without loss of generality, we can assume $M = (M_1, M_2) = (1, 1)$ was sent. We have the following error events:
  - $\mathcal{E}_{11} = \{A^n(1), Y^n, S^n\} \notin T_\epsilon^n(A, Y, S)$.
  - $\mathcal{E}_{12} = \{A^n(\hat{m}_1), Y^n, S^n\} \in T_\epsilon^n(A, Y, S)$ for $\hat{m}_1 \neq 1$.
  - $\mathcal{E}_{21} = \{X^n(1, 1), Y^n | S^n, A^n(1)\} \notin T_\epsilon^n(X, Y | S^n, A^n(1))$.
  - $\mathcal{E}_{22} = \{X^n(1, \hat{m}_2), Y^n | S^n, A^n(1)\} \in T_\epsilon^n(X, Y | S^n, A^n(1))$ for $\hat{m}_2 \neq 1$.

  Let $\mathcal{E}_0 = P((A^n(1), X^n(1, 1)) \in T_\epsilon^n(A, X))$. Hence

$$P(\mathcal{E}) = P(\mathcal{E}_0 \cap \mathcal{E}) + P(\mathcal{E}_0^c \cap \mathcal{E}) \tag{98}$$

$$\leq P(\mathcal{E}_0 \cap \mathcal{E}) + P(\mathcal{E}_0^c). \tag{99}$$

  Since $(1, 1)$ is the actual message being sent and action and channel input sequences are generated i.i.d., $\sim P_A P_{X|S_e}$, $A^n(1)$ and $X^n(1, 1)$ will be jointly typical as $n \to \infty$, to be more precise by the LLN (law of large numbers) arguments ([11]), $P(\mathcal{E}_0^c) \to 0$ as $n \to \infty$.

  Note in the following arguments the limit of the probabilities is taken as $n \to \infty$, this being implied we will omit using $n \to \infty$ repeatedly for the sake of brevity. We will now show that $P(\mathcal{E}_0 \cap \mathcal{E}) \to 0$. Let $\mathcal{E}_1 = \mathcal{E}_{11} \cup \mathcal{E}_{12}$ and $\mathcal{E}_2 = \mathcal{E}_{21} \cup \mathcal{E}_{22}$. By Law of Large Numbers, (LLN, ([11]), $P(\mathcal{E}_0 \cap \mathcal{E}_{11}) \to 0$. By Packing Lemma ([11]), $P(\mathcal{E}_0 \cap \mathcal{E}_{12}) \to$ if $R_1 < I(A; Y, S) = I(A; Y|S)$ which implies by union bound $P(\mathcal{E}_0 \cap \mathcal{E}_1) \leq P(\mathcal{E}_0 \cap \mathcal{E}_{11}) + P(\mathcal{E}_0 \cap \mathcal{E}_{12}) \to 0$.

  Similarly by LLN, $P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_{21}) \to 0$ and by Packing Lemma $P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_{22}) \to 0$ if $R_2 < I(X; Y|S, A)$ which implies by the union bound $P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_2) \leq P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_{21}) + P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_{22}) \to 0$. Hence, the total probability of error

$$P(\mathcal{E}_0 \cap \mathcal{E}) = P(\mathcal{E}_0 \cap (\mathcal{E}_1 \cup \mathcal{E}_2)) \tag{100}$$

$$\leq P(\mathcal{E}_0 \cap \mathcal{E}_1) + P(\mathcal{E}_0 \cap \mathcal{E}_1^c \cap \mathcal{E}_2) \to 0 \tag{101}$$

  if $R_1 < I(A; Y|S)$ and $R_2 < I(X; Y|S, A)$. Therefore, we obtain for vanishing probability of error that

$$R = R_1 + R_2 \tag{102}$$

$$< I(A; Y|S) + I(X; Y|A, S) \tag{103}$$

$$= I(X, A; Y|S) \tag{104}$$

$$= H(Y|S) - H(Y|X, S, A) \tag{105}$$

$$\overset{(*)}{=} I(X; Y|S) = C(\frac{\Gamma}{1 + \epsilon}), \tag{106}$$

  where $(*)$ is due to our channel assumption which is expressed in joint PMF of any induced scheme [cf (4)]. Proof of achievability is completed by taking $\epsilon \to 0$.

2) Converse: Suppose rate R is achievable. Now consider a sequence of $(2^{nR}, n)$ codes for which we have $P_e^n \overset{n\to\infty}{\longrightarrow} 0$. Consider

$$nR = H(M) \tag{107}$$

$$\overset{(a)}{=} H(M|S^n) \tag{108}$$

$$= I(M; Y^n|S^n) + H(M|Y^n, S^n). \tag{109}$$

By Fano's Inequality ([12])

$$H(M|Y^n, S^n) \leq 1 + P_e^n R \leq n\epsilon_n \tag{110}$$

where $\epsilon_n \overset{n\to\infty}{\longrightarrow} 0$. Now Consider

$$I(M; Y^n|S^n) = H(Y^n|S^n) - H(Y^n|M, S^n) \tag{111}$$

$$\overset{(b)}{=} \sum_{i=1}^n H(Y_i|S^n, Y^{i-1}) - \sum_{i=1}^n H(Y_i|Y^{i-1}, M, S^n, A^n, S_e^n, X^n) \tag{112}$$

$$\overset{(c)}{\leq} \sum_{i=1}^n H(Y_i|S_i) - \sum_{i=1}^n H(Y_i|X_i, S_i) \tag{113}$$

$$= \sum_{i=1}^n I(X_i; Y_i|S_i) \tag{114}$$

$$\leq \sum_{i=1}^n C(\Lambda(A_i)) \tag{115}$$

$$\overset{(d)}{\leq} nC(\frac{1}{n}\sum_{i=1}^n \Lambda(A_i)) \tag{116}$$

$$= nC(\Lambda(A^n)) \tag{117}$$

$$\overset{(e)}{\leq} nC(\Gamma) \tag{118}$$

where

- (a) follows from the fact that message is independent of state sequence.
- (b) follows from the fact that $A^n = A^n(M)$, $S_e = h(S, A)$ and $X^n = X^n(M, S_e^n)$.

- (c) follows from the fact that conditioning reduces entropy and from the Markov chain, $Y_i - (X_i, S_i) - (Y^{i-1}, M, S^{n\backslash i}, A^n, S_e^n, X^{n\backslash i})$ which is due to the induced joint probability distribution as in (4).
- (d) follows from the fact that $C(\Gamma)$ is concave in Γ. This is proved as follows. Let $C(\Gamma_1)$ and $C(\Gamma_2)$ be respectively achieved at joint $P_A^1 P_{X|S_e,A}^1$ and $P_A^2 P_{X|S_e,A}^2$. Let $P^1(\cdot)$ and $P^2(\cdot)$ be the corresponding joint distributions. Since $C(\Gamma)$ is nondecreasing in Γ (which can be argued easily as larger Γ implies a larger feasible region and hence larger capacity), therefore we have

$$\begin{aligned} \mathsf{E}_{P^1}[\Lambda(A)] &= \Gamma_1 &(119)\\ \mathsf{E}_{P^2}[\Lambda(A)] &= \Gamma_2. &(120)\end{aligned}$$

Now consider a joint distribution $P^\lambda = \lambda P^1 + (1-\lambda)P^2$. Clearly

$$\mathsf{E}_{P^\lambda}[\Lambda(A)] = \lambda\Gamma_1 + (1-\lambda)\Gamma_2. \tag{121}$$

Now observe that $I(X;Y|S)$ is concave in $P(Y|S)$ which is linear in $P_A P_{X|S_e,A}$. Hence, $I(X;Y|S)$ is concave in $P_A P_{X|S_e,A}$. Thus denoting R^λ as the value of $I(X;Y|S)$ at joint P^λ, we have

$$\lambda C(\Gamma_1) + (1-\lambda)C(\Gamma_2) \le R^\lambda \le C(\lambda\Gamma_1 + (1-\lambda)\Gamma_2)$$

which proves the concavity of $C(\Gamma)$ in Γ.
- (e) follows from the fact that $C(\Gamma)$ is non decreasing in Γ, as explained in (d) above.

We further note the following relations and Markov Chains:

- $A_i = A_i(M)$ is independent of S_i as state sequence is independent of message and actions are functions of message.
- $X_i - (S_{e,i}, A_i) - S_i$. This can be reasoned as follows. Since $X_i = X_i(M, S_e^n)$, it suffices to prove $(M, S_e^n) - (S_{e,i}, A_i) - S_i$. We observe the joint distribution can be factorized as,

$$\begin{aligned} &P(M, A^n, S^n, S_e^n)\\ &= P(M)\prod_{i=1}^n P(S_i)P(A_i|M)P(S_{e,i}|S_i, A_i) &(122)\\ &= \Phi_1(A^{n\backslash i}, M, S_e^{n\backslash i}, S^{n\backslash i}, A_i)\Phi_2(S_i, S_{e,i}, A_i) &(123)\\ &= \Phi_1'(A^{n\backslash i}, M, S_e^n, S^{n\backslash i}, A_i, S_{e,i})\Phi_2(S_i, S_{e,i}, A_i), &(124)\end{aligned}$$

which implies the Markov Chain $(A^{n\backslash i}, M, S_e^n, S^{n\backslash i}) - (S_{e,i}, A_i) - S_i$, which in turn implies $(M, S_e^n) - (S_{e,i}, A_i) - S_i$.
- $Y_i - (X_i, S_i) - (A_i, S_{e,i})$ follows from the DMC assumption on the channel which implies the induced joint probability distribution as in (4).

Hence by using (109), (110) and (118), and letting $n \to \infty$ we have $R \le C(\Gamma)$.

Appendix B
Concavity of Capacity in Cost

We prove the concavity of cost constrained capacity in Theorem 4 by concavification argument. Consider "concavification" of capacity in Theorem 4 as

$$C^Q(\Gamma) = \max[I(U; Y, S_d|A_d, Q)] \tag{125}$$

where maximization is over all joint distributions of the form

$$\begin{aligned} &P_{Q,S,A_d,U,A_e,S_e,X,Y,S_d}(s, a_d, u, a_e, s_e, x, y, s_d)\\ &= P_Q(q)P_S(s)P_{A_d|Q}(a_d|q)P_{U|A_d,Q}(u|a_d, q)\\ &\times \mathbf{1}_{\{a_e = g(u,a_d,q)\}} P_{S_e|S,A_e}(s_e|s, a_e)\mathbf{1}_{\{x=f(u,s_e,a_d,q)\}}\\ &\times P_{Y|X,S}P_{S_d|S,S_e,A_e,A_d}(s_d|s, s_e, a_e, a_d) &(126)\end{aligned}$$

for some $P_Q, P_{A_d|Q}, P_{U|A_d,Q}, g, f$ such that $\mathsf{E}[\Lambda(A_e, A_d)] \le \Gamma$. Clearly, $C^Q(\Gamma) \ge C(\Gamma)$. Left is to prove $C^Q(\Gamma) \le C(\Gamma)$

$$\begin{aligned} &I(U; Y, S_d|A_d, Q)\\ &= H(Y, S_d|A_d, Q) - H(Y, S_d|U, A_d, Q) &(127)\\ &\le H(Y, S_d|A_d) - H(Y, S_d|U, A_d, Q) &(128)\\ &= I(U'; Y, S_d|A_d) &(129)\end{aligned}$$

where last equality follows from the defining $U' = (U, Q)$. Proof is completed by noting that the joint distribution of $(S, A_d, U', A_e, S_e, X, Y, S_d)$ is same as that of $(S, A_d, U, A_e, S_e, X, Y, S_d)$.

Appendix C
Proof of Markov Chains in Theorem 4

We will prove the following Markov chains:

MC1 $U_i - A_{d,i} - S_i$.
MC2 $A_{e,i} - (U_i, A_{d,i}) - S_i$.
MC3 $(S_{e,i}, S_{d,i}) - (S_i, A_{e,i}, A_{d,i}) - U_i$.
MC4 $X_i - (U_i, S_{e,i}, A_{d,i}) - (A_{e,i}, S_i, S_{d,i})$.
MC5 $Y_i - (X_i, S_i) - (U_i, A_{d,i}, A_{e,i}, S_{e,i}, S_{d,i})$.

MC3 and MC5 follow from the DMC assumption in problem definition. Now for the rest consider the induced probability distribution by the given encoding and decoding scheme on $(m, a_e^n, s^n, s_e^n, x^n, y^n, a_d^n, s_d^n)$

$$\begin{aligned} &P_{M,A_e^n,S^n,S_e^n,X^n,Y^n,A_d^n,S_d^n}(\cdot)\\ &= \frac{1}{\mathcal{M}}\mathbf{1}_{\{a_e^n = A_e^n(m)\}}\prod_{i=1}^n \mathbf{1}_{\{a_{d,i} = A_{d,i}(y^{i-1})\}}P_S(s_i)\\ &\times \prod_{i=1}^n P_{S_e,S_d|S,A_e,A_d}(s_{e,i}, s_{d,i}|s_i, a_{e,i}, a_{d,i})\\ &\times \prod_{i=1}^n \mathbf{1}_{\{x_i = X_i(m, s_e^i)\}}P_{Y|X,S}(y_i|x_i, s_i). &(130)\end{aligned}$$

Averaging over $(S_{i+1}^n, S_{e,i}^n, X_i^n, Y_i^n, S_{d,i}^n, A_{d,i+1}^n)$, we get the induced joint probability distribution on $(m, a_e^n, s^i, s_e^{i-1}, x^{i-1}, y^{i-1}, a_d^{i-1}, s_d^{i-1})$

$$\begin{aligned}
&P_{M,A_e^n,S^i,S_e^{i-1},X^{i-1},Y^{i-1},A_d^i,S_d^{i-1}}(\cdot)\\
&= P_S(s_i) \times \frac{1}{\mathcal{M}} \mathbf{1}_{\{a_e^n = A_e^n(m)\}} \mathbf{1}_{\{a_{d,i} = A_{d,i}(y^{i-1})\}}\\
&\times \prod_{j=1}^{i-1} \mathbf{1}_{\{a_{d,j} = A_{d,j}(y^{j-1})\}} P_S(s_j)\\
&\times \prod_{j=1}^{i-1} P_{S_e,S_d|S,A_e,A_d}(s_{e,j} s_{d,j} | s_j, a_{e,j}, a_{d,j})\\
&\times \prod_{j=1}^{i-1} \mathbf{1}_{\{x_j = X_j(m,s_e^j)\}} P_{Y|X,S}(y_j|x_j,s_j)) \qquad (131)\\
&= \Phi_1(S_i)\Phi_2(M, A_e^n, S^{i-1}, S_e^{i-1}, X^{i-1}, Y^{i-1}, A_d^i, S_d^{i-1}) \qquad (132)\\
&= \Phi_1'(S_i, A_{d,i})\Phi_2(S^{i-1}, A_{d,i}, U_i, X^{i-1}). \qquad (133)
\end{aligned}$$

Equation (132) implies $A_{d,i}$ is independent of S_i while (133) implies markov chain $(U_i, X^{i-1}) - A_{d,i} - S_i$ which in turn implies MC1. MC2 is straightforward as U contains A_e^n.

Now averaging over $(S_{i+1}^n, S_{e,i+1}^n, X_{i+1}^n, Y_i^n, S_{d,i+1}^n, A_{d,i+1}^n)$ in (130), we obtain the joint probability distribution on $(m, a_e^n, s^i, s_e^i, x^i, y^{i-1}, a_d^{i-1}, s_d^{i-1})$

$$\begin{aligned}
&P_{M,A_e^n,S^i,S_e^i,X^i,Y^{i-1},A_d^i,S_d^i}(\cdot)\\
&= P_S(s_i) P_{S_e,S_d|S,A_e,A_d}(s_{e,i}, s_{d,i} | s_i, a_{e,i}, a_{d,i})\\
&\times \frac{1}{\mathcal{M}} \mathbf{1}_{\{a_e^n = A_e^n(m)\}} \mathbf{1}_{\{x_i = X_i(m,s_e^i)\}} \mathbf{1}_{\{a_{d,i} = A_{d,i}(y^{i-1})\}}\\
&\times \prod_{j=1}^{i-1} P_S(s_j) P_{S_e,S_d|S,A_e,A_d}(s_{e,j}, s_{d,j} | s_j, a_{e,j}, a_{d,j})\\
&\times \prod_{j=1}^{i-1} \mathbf{1}_{\{x_j = X_j(m,s_e^j)\}} P_{Y|X,S}(y_j|x_j,s_j)\\
&\times \prod_{j=1}^{i-1} \mathbf{1}_{\{a_{d,j} = A_{d,j}(y^{j-1})\}} \qquad (134)\\
&= \Phi_1(S_i, S_{e,i}, S_{d,i}, A_{e,i}, S_{e,i})\\
&\times \Phi_2(M, A_e^n, S^{i-1}, S_e^{i-1}, X^i, Y^{i-1}, A_d^i, S_d^{i-1}) \qquad (135)\\
&= \Phi_1'(S_i, A_{e,i}, S_{d,i}, U_i, S_{e,i}, A_{d,i})\\
&\times \Phi_2'(U_i, S_{e,i}, A_{d,i}, X^i, S^{i-1}). \qquad (136)
\end{aligned}$$

This implies the Markov Chain, $(S^{i-1}, X^i) - (U_i, S_{e,i}, A_{d,i}) - (S_i, A_{e,i}, S_{d,i})$, which implies MC4.

References

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

[2] 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.

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

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

[5] G. Keshet, Y. Steinberg, and N. Merhav, "Channel coding in the presence of side information," *Found. Trends Commun. Inf. Theory*, vol. 4, no. 6, pp. 445–586, 2007.

[6] H. H. Permuter and T. Weissman, "Source coding with a side information 'vending machine' at the decoder," in *Proc. 2009 IEEE Int. Conf. Information Theory (ISIT 09)*, Piscataway, NJ, 2009, pp. 1030–1034.

[7] 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.

[8] T. Weissman, "Capacity of channels with action-dependent states," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5396–5411, Nov. 2010.

[9] K. Kittichokechai, T. Oechtering, M. Skoglund, and R. Thobaben, "Source and channel coding with action-dependent partially known two-sided state information," in *Proc. 2010 IEEE Int. Conf. Information Theory (ISIT 10)*, Jun. 2010, pp. 629–633.

[10] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 11, pp. 1986–1992, Nov. 1997.

[11] A. E. Gamal and Y. H. Kim, "Lecture notes on network information theory," *CoRR*, vol. abs/1001.3404, 2010.

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

[13] M. Salehi, Cardinality Bounds on Auxiliary Variables in. Multiple-User Theory via the Method of Ahlswede and Korner Dept. Statistics, Stanford University, Stanford, CA, 1978, Tech. Rep. 33.

[14] A. Zaidi, L. Vandendorpe, and P. Duhamel, "Lower bounds on the capacity regions of the relay channel and the cooperative relay-broadcast channel with non-causal side information," in *Proc. IEEE Int. Conf. Communications (ICC 07)*, Jun. 2007, pp. 6005–6011.

[15] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[16] H. Asnani, H. H. Permuter, and T. Weissman, "To feed or not to feed back," *CoRR*, vol. abs/1011.1607, 2010.

[17] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Orlando, FL: Academic, 1982.

Himanshu Asnani (S'11) is currently a Ph.D. candidate in Information Systems Lab, Electrical Engineering Department at Stanford University. He is advised by Prof. Tsachy Weissman and co-advised by Prof. Balaji Prabhakar. His research interests include information theory, probability theory and statistical learning.

He received his B.Tech. from IIT Bombay and M.S. from Stanford University in 2009 and 2011, respectively. He is a Stanford Graduate Fellow (SGF) and recipient of Best Paper Award at MobiHoc 2009.

Haim Permuter (M'08) received his B.Sc. (summa cum laude) and M.Sc. (summa cum laude) degree in Electrical and Computer Engineering from the Ben-Gurion University, Israel, in 1997 and 2003, respectively, and Ph.D. degrees in Electrical Engineering from Stanford University, California in 2008.

Between 1997 and 2004, he was an officer at a research and development unit of the Israeli Defense Forces. He is currently a senior lecturer at Ben-Gurion University.

Dr. Permuter is a recipient of the Fulbright Fellowship, the Stanford Graduate Fellowship (SGF), Allon Fellowship, and the 2009 U.S.-Israel Binational Science Foundation Bergmann Memorial Award.

Tsachy Weissman (S'99–M'02–SM'07) graduated summa cum laude with a B.Sc. in electrical engineering from the Technion in 1997, and earned his Ph.D. from the same place in 2001. He then worked at Hewlett-Packard Laboratories with the information theory group until joining Stanford, where he has been on the faculty of the Electrical Engineering department since 2003, spending the two academic years 2007–2009 on leave at the Technion.

Tsachy's research is focused on information theory, statistical signal processing, the interplay between them, and their applications.

Among his recent awards and honors is an NSF CAREER award, 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 is on the editorial board of the IEEE TRANSACTIONS ON INFORMATION THEORY, serving as Associate Editor for Shannon Theory.