The Capacity of Channels With Feedback

Sekhar Tatikonda, *Member, IEEE*, and Sanjoy Mitter, *Life Fellow, IEEE*

***Abstract*—In this paper, we introduce a general framework for treating channels with memory and feedback. First, we prove a general feedback channel coding theorem based on Massey's concept of *directed information*. Second, we present coding results for Markov channels. This requires determining appropriate sufficient statistics at the encoder and decoder. We give a recursive characterization of these sufficient statistics. Third, a dynamic programming framework for computing the capacity of Markov channels is presented. Fourth, it is shown that the average cost optimality equation (ACOE) can be viewed as an implicit single-letter characterization of the capacity. Fifth, scenarios with simple sufficient statistics are described. Sixth, error exponents for channels with feedback are presented.**

***Index Terms*—Capacity, directed information, dynamic programming, feedback, Markov channels, sufficient statistics.**

I. Introduction

THIS paper presents a general framework for proving coding theorems for channels with memory and feedback. Because of increased demand for wireless communication and networked systems there is a renewed interest in this problem. Feedback can increase the capacity of a noisy channel, decrease the complexity of the encoder and decoder, and reduce latency.

Recently, Verdú and Han presented a very general formulation of the channel coding problem without feedback [34]. Specifically, they provided a coding theorem for finite-alphabet channels with arbitrary memory. They worked directly with the information density and provided a Feinstein-like lemma for the converse result. Here we generalize that formulation to the case of channels with feedback. In this case, we require the use of code functions as opposed to codewords. A code function maps a message and the channel feedback information into a channel input symbol. Shannon introduced the use of code functions, which he called strategies, in his work on transmitter side information [28]. Code functions are also sometimes called codetrees [21].

We convert the channel coding problem with feedback into a new channel coding problem without feedback. The channel inputs in this new channel are code functions. Unfortunately, the space of code functions can be quite complicated to work with. We show that we can work directly with the original space of channel inputs by making explicit the relationship between code-function distributions and channel input distributions. This relationship allows us to convert a mutual information optimization problem over code-function distributions into a *directed information* optimization problem over channel input distributions.

Directed information was introduced by Massey [23] who attributes it to Marko [22]. Directed information can be viewed as a causal version of mutual information. Kramer [20], [21] used directed information to prove capacity theorems for general discrete memoryless networks. These networks include the memoryless two-way channel and the memoryless multiple-access channel. In this paper, we examine single-user channels with memory and feedback. One of the main difficulties in this problem has to do with the fact that the transmitter and the receiver may have different information about the state of the channel. We show how to choose appropriate sufficient statistics at both the transmitter and the receiver.

The problem of optimal channel coding goes back to the original work of Shannon [26]. The channel coding problem with feedback goes back to early work by Shannon, Dobrushin, Wolfowitz, and others [27], [12], [38]. In particular, Shannon introduced the feedback problem. Both Shannon and Dobrushin examined the case of memoryless channels with feedback. Wolfowitz, in his book, describes a variety of finite-state channels with state calculable by the sender or the receiver. We generalize these results to general Markov channels with output feedback. We do not assume that the state is known to either the transmitter or the receiver.

There is a long history of work regarding Markov channels and feedback. Here we describe a few connections to that literature. Mushkin and Bar-David [24] determined the capacity of the Gilbert-Elliot channel without feedback. Blackwell, Breiman, and Thomasian [4] examine finite-state indecomposable channels without feedback. Goldsmith and Varaiya [16] examine nonintersymbol interference (ISI) Markov channels without feedback. For the case of independent and identically distributed (i.i.d.) inputs and symmetric channels, they introduce sufficient statistics that lead to a single-letter formula. In this paper, we identify the appropriate statistics when feedback is available. In a certain sense, to be explained in this paper, the Markov channel with feedback problem is easier than the Markov channel without feedback problem. This is because in the full channel output feedback case the decoder's information pattern is nested in the encoder's information pattern [36]. In this paper, we do not treat noisy feedback.

Viswanathan [35], Caire and Shamai [6], and Das and Narayan [10] all examine different classes of channels with memory and side information at the transmitter and the receiver. Chen and Berger [7] examine Markov channels when the state is known to both the transmitter and the receiver. In this paper,

Manuscript received December 05, 2001; revised January 15, 2007. Current version published December 24, 2008. This work was supported by the National Science Foundation under Awards CCF-0430922, CCF-0325774, and ECCS-0801549.

S. Tatikonda is with the Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA (e-mail: sekhar.tatikonda@yale.edu).

S. Mitter is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02319 USA (e-mail: mitter@mit.edu).

Communicated by H. Yamamoto, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2008.2008147

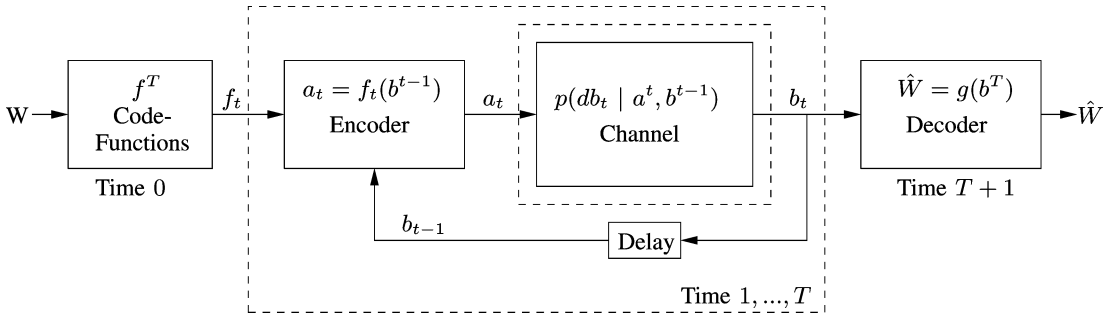

Fig. 1. Interconnection.

we present a general framework for treating Markov channels with ISI and feedback.

Many authors consider conditions that ensure the Markov channel is *information stable* [25]. For example, Cover and Pombra [8] show that Gaussian channels with feedback are always information stable. Shannon [29] introduced the notion of "recoverable state property" by which a channel can be reset to a known state by using a fixed finite sequence of channel inputs. In addition, some authors consider conditions that ensure the Markov channel is *indecomposable* [15], [4]. In our work, it is shown that solutions to the associated average cost optimality equation (ACOE) imply information stability. In addition, the sufficient condition provided here for the existence of a solution to the ACOE implies a strong mixing property of the underlying Markov channel in the same way that indecomposability does. The ACOE can be viewed as an implicit single-letter characterization of the channel capacity.

We consider Markov channels with finite-state, channel input, and channel output alphabets. But with the introduction of appropriate sufficient statistics, we quickly find ourselves working with Markov channels over general alphabets and states. As shown by Csiszár [9], for example, treating general alphabets involve many technical issues that do not arise in the finite-alphabet case.

Tatikonda first introduced the dynamic programming approach to the directed information optimization problem [31]. Yang, Kavcic, and Tatikonda have examined the case of finite-state machine Markov channels [39]. Here we present a stochastic control framework that treats many Markov channels including finite-state machine Markov channels.

In general, it is difficult to solve the ACOE. This is because the sufficient statistics can be quite complicated. Hence, it will be difficult to get an explicit formula for the feedback channel capacity. There are, though, many scenarios when the sufficient statistics become much simpler and hence the ACOE becomes simpler. We discuss these scenarios in Section VIII. In these cases, one can apply exact or approximate dynamic programming techniques to solve the ACOE.

In summary, the main contributions of this paper are as follows. 1) We prove a general feedback channel coding theorem based on Massey's concept of *directed information* [23]. 2) We present coding results for Markov channels. This requires determining appropriate sufficient statistics at the encoder and decoder. We give a recursive characterization of these sufficient statistics. 3) A dynamic programming framework for computing the capacity of Markov channels is presented. 4) It is shown that the ACOE can be viewed as an implicit single-letter characterization of the capacity. 5) Scenarios with simple sufficient statistics are described. 6) Error exponents for channels with feedback are presented. Preliminary versions of this work have appeared in [31]–[33].

II. Feedback and Causality

Here we discuss some of the subtleties (some more so than others) of feedback and causality inherent in the feedback capacity problem. We give a high-level discussion here and give specific definitions in the next section. The channel at time t is modeled as a stochastic kernel $p(db_t \,|\, a^t, b^{t-1})$, where $a^t = (a_1, \ldots, a_t)$ and $a^0 = \emptyset$. See Fig. 1. The channel output is fed back to the encoder with delay one. At time t, the encoder takes the message and the past channel output symbols $B_1, \ldots, B_{t-1}$ and produces a channel input symbol A_t. At time T, the decoder takes all the channel output symbols $B_1, \ldots, B_T$ and produces the decoded message. Hence, the time ordering of the variables is

$$\text{message}, A_1, B_1, A_2, B_2, \ldots, A_T, B_T, \text{decoded message}. \tag{1}$$

When there is no feedback, under suitable conditions, $\sup_{P(da^T)} I(A^T; B^T)$ characterizes the maximum number of messages one can send with small probability of decoding error. Our goal in this paper is to generalize this to the case of feedback. To that end, we now mention some subtleties that will guide our approach. See also Massey [23] and Marko [22].

One should not supremize the mutual information $I(A^T, B^T)$ over the stochastic kernel $p(da^T \,|\, b^T)$. We can factor $p(da^T \,|\, b^T) = \otimes_{t=1}^T p(da_t \,|\, a^{t-1}, b^T)$. This states that, at time t, the channel input symbol A_t depends on the future channel output symbols B_t^T. This violates the causality implicit in our encoder description. In fact, $p(da^T \,|\, b^T)$ is the posterior probability used by the decoder to decode the message at time T. Instead, as we will show, one should supremize the mutual information over the *directed stochastic kernel*: $\vec{p}(da^T \,|\, b^T) = \otimes_{t=1}^T p(da_t \,|\, a^{t-1}, b^{t-1})$. See Definition 4.1.

One should not use the stochastic kernel $p(db^T \,|\, a^T)$ as a model of the channel when there is feedback. To compute the mutual information, we need to work with the joint measure $P(da^T, db^T)$. In general, it is not possible to find a joint measure consistent with the stochastic kernels: $\vec{p}(da^T \,|\, b^T)$ and $p(db^T \,|\, a^T)$. Instead, as we will show, the appropriate model for

the channel when there is feedback is a sequence of stochastic kernels: $\{p(db_t \,|\, a^t,\ b^{t-1})\}_{t=1}^T$. See Section III.

One should not use the mutual information $I(A^T;B^T)$ when there is feedback. When there is feedback the conditional probabilities $P(db_t \,|\, a^T, b^{t-1}) \neq P(db_t \,|\, a^t, b^{t-1})$ almost surely under $P(da^T, db^T)$. Even though A_{t+1} occurs after B_t, it still has a probabilistic influence on it. This is because under feedback A_{t+1} is influenced by the past channel output B_t. To quote Massey [23], "statistical dependence, unlike causality, has no inherent directivity." The mutual information factors as $I(A^T;B^T) = \sum_{t=1}^T I(A^T;B_t \,|\, B^{t-1})$. The information transmitted to the receiver at time t, given by $I(A^T;B_t \,|\, B^{t-1})$, depends on the future A_{t+1}^T. Instead, as we will show, we should use the *directed information*: $I(A^T \to B^T)$. See Definition 4.2.

III. Channels With Feedback

In this section, we formulate the feedback channel coding problem. We first introduce some notation. Let $p(dy \,|\, x)$ represent a stochastic kernel from the measurable spaces $\mathcal{X}$ to $\mathcal{Y}$. See the Appendix for definitions and properties of stochastic kernels.

Given a joint measure $P(dx, dy)$, we use $P(Y = y \,|\, X = x)$ (or just $P(y \,|\, x)$) to represent the conditional probability (when it exists.) In general, lower case letters $p, q, r, \ldots$ will be used for stochastic kernels and upper case letter $P, Q, R, \ldots$ will be used for joint measures or conditional probabilities. Let $\mathcal{P}(\mathcal{X})$ represent the space of all probability measures on $\mathcal{X}$ endowed with the topology of weak convergence.

Capital letters $A, B, X, Y, Z, \ldots$ will represent random variables and lower case letters $a, b, x, y, z, \ldots$ will represent particular realizations. For the stochastic kernel $p(dy \,|\, x)$, we have $p(y \,|\, x)$ being a number. Given a joint measure $P_{X,Y}(dx, dy) = P_X(dx) \otimes p(dy \,|\, x)$, we have $p(y \,|\, X)$ being a random variable taking value $p(y \,|\, x)$ with probability $P_X(x)$, $p(Y \,|\, X)$ being a random variable taking value $p(y \,|\, x)$ with probability $P_{X,Y}(x, y)$, and $p(dy \,|\, X)$ being a random measure-valued element taking value $p(dy \,|\, x)$ with probability $P_X(x)$. Finally, let the notation $X - Y - Z$ denote that the random elements X, Y, Z form a Markov chain.

We are now ready to formulate the feedback channel coding problem. Let $\{A_t\}_{t=1}^T$ be random elements in the finite[1] set $\mathcal{A}$ with the power set σ-algebra. These represent the channel inputs. Similarly, let $\{B_t\}_{t=1}^T$ be random elements in the finite set $\mathcal{B}$ with the power set σ-algebra. These represent the channel outputs. Let $\mathcal{A}^T$ and $\mathcal{B}^T$ represent the T-fold product spaces with the product σ-algebras (where T may be infinity). We use "log" to represent logarithm base 2.

A *channel* is a family of stochastic kernels $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$. These channels are nonanticipative with respect to the time-ordering (1) because the conditioning includes only a^t, b^{t-1}.

We now define a code function. This is an extension of the usual concept of codeword. Let $\mathcal{F}_t$ be the set of all measurable maps $f_t : \mathcal{B}^{t-1} \to \mathcal{A}$ taking $b^{t-1} \mapsto a_t$. Endow $\mathcal{F}_t$ with the power set σ-algebra. Let $\mathcal{F}^T = \prod_{t=1}^T \mathcal{F}_t$ denote the Cartesian product endowed with the product σ-algebra. Note that since $\mathcal{A}$ and $\mathcal{B}$ are finite, the space $\mathcal{F}^T$ is at most countable. A *channel code function* is an element $f^T = (f_1, \ldots, f_T) \in \mathcal{F}^T$. A distribution on $\mathcal{F}^T$ is given by a specification of a sequence of *code-function stochastic kernels* $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$. Specifically, $P_{F^T}(df^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1})$. We will use the notation $f^t(b^{t-1}) = (f_1, f_2(b_1), \ldots, f_t(b^{t-1}))$.

A *message set* is a set $\mathcal{W} = \{1, \ldots, M\}$. Let the distribution P_W on the message set $\mathcal{W}$ be the uniform distribution. A *channel code* is a list of M channel code functions denoted by $f^T[w],\ w \in \mathcal{W}$. For message w at time t with channel feedback b^{t-1}, the channel encoder outputs $f_t[w](b^{t-1})$. A *channel code without feedback* is a list of M channel codewords denoted by $a^T[w], w \in \mathcal{W}$. For message w at time t, the channel encoder outputs $a_t[w]$ independent of the past channel outputs b^{t-1}.

A *channel decoder* is a map $g : \mathcal{B}^T \to \mathcal{W}$ taking $b^T \mapsto w$. The decoder waits till it observes all the channel outputs before reconstructing the input message. The order of events is shown in Fig. 1.

Definition 3.1: A (T, M, ϵ) *channel code* over time horizon T consists of M code functions, a channel decoder g, and an error probability satisfying $\frac{1}{M}\sum_{w=1}^M \Pr(w \neq g(b^T) \,|\, w) \leq \epsilon$. A (T, M, ϵ) *channel code without feedback* is defined similarly with the restriction that we use M codewords.

In the following, the superscripts "o" and "nfb" represent the words "operational" and "no feedback." Following [34], we define the following.

Definition 3.2: R is an *ϵ-achievable rate* if, for all $\delta > 0$, there exists, for all sufficiently large T, (T, M, ϵ) channel codes with rate $\frac{\log M}{T} > R - \delta$. The maximum *$\epsilon$-achievable* rate is called the *ϵ-capacity* and denoted C_ϵ^{o}. The *operational channel capacity* is defined as the maximal rate that is ϵ-achievable for all $0 < \epsilon < 1$ and is denoted C^{o}. Analogous definitions for $C_\epsilon^{\mathrm{o,nfb}}$ and $C^{\mathrm{o,nfb}}$ hold in the case of no feedback.

Before continuing we quickly remark on some other formulations in the literature. Some authors work with different sets of channels for each blocklength T. See, for example, [13] and [34]. In our context, this would correspond to a different sequence of channels for each T: $\{p_T(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$. Theorem 5.1 will continue to hold if we use channels of this form. It is hard, though, to imagine a context with feedback where nature will provide a different set of channels depending on the time horizon. Hence, the formulation in this paper is natural when feedback is available and opens the way to treating Markov channels.

Note that in Definition 3.2 we are seeking a single number C^{o} that is an achievable capacity for all sufficiently large T. Some authors instead, see [8], for example, seek a sequence of numbers $\{C_T^{\mathrm{o}}\}$ such that there exists a sequence of channel codes $\{(T, 2^{TC_T^{\mathrm{o}}}, \epsilon_T)\}$ with $\epsilon_T \to 0$. It will turn out that for the time-invariant Markov channels described in Section VI the notion of capacity described in Definition 3.2 is the appropriate one. We will further elaborate on this point in Section IV-A after we have reviewed the concept of information stability.

[1]The methods in this paper can be generalized to channels with abstract alphabets.

A. Interconnection of Code Functions to the Channel

Now we are ready to interconnect the pieces: channel, channel code, and decoder. We follow Dobrushin's program and define a joint measure over the variables of interest that is consistent with the different components [13]. We will define a new channel without feedback that connects the code functions to the channel outputs. Corollary 3.1 shows that we can connect the messages directly to the channel output symbols.

Let $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ be a sequence of code-function stochastic kernels with joint measure $P_{F^T}(df^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1})$ on $\mathcal{F}^T$. For example, P_{F^T} may be a distribution that places mass $1/M$ on each of M different code functions. Given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ and a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, we want to construct a new channel that interconnects the random variables F^T to the random variables B^T. We use "Q" to denote the new joint measure $Q(df^T, da^T, db^T)$ that we will construct. The following three reasonable properties should hold for our new channel.

Definition 3.3: A measure $Q(df^T, da^T, db^T)$ is said to be *consistent* with the code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ and the channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$ if for each t, the following hold.

i) *There is no feedback to the code functions in the new channel*: The measure on $\mathcal{F}^T$ is chosen at time 0. Thus, it cannot causally depend on the A_t's and B_t's. Thus, for each t and all f_t, we have
$$Q(f_t \,|\, f^{t-1}, a^{t-1}, b^{t-1}) = p(f_t \,|\, f^{t-1})$$
for Q almost all $f^{t-1}, a^{t-1}, b^{t-1}$.

ii) *The channel input is a function of the past outputs*: For each t, $A_t = F_t(B^{t-1})$ $Q-a.s.$ In other words, for each t and all a_t, we have
$$Q(a_t \,|\, f^t, a^{t-1}, b^{t-1}) = \delta_{\{f_t(b^{t-1})\}}(a_t)$$
for Q almost all f^t, a^{t-1}, b^{t-1}. Here δ is the Dirac measure: $\delta_{\{\alpha\}}(a) = 1$ if $a = \alpha$ and 0 else.

iii) *The new channel preserves the properties of the underlying channel*: For each t, and all b_t, we have
$$Q(b_t \,|\, f^t, a^t, b^{t-1}) = p(b_t \,|\, a^t, b^{t-1})$$
for Q almost all f^t, a^t, b^{t-1}.

Note that in ii) we have assumed that the channel input is a function of the past outputs. One could consider more general stochastic encoders (as is often done for compound channels.) In our case, the channel is assumed to be known to the transmitter and the receiver.

The next lemma shows that there exists a unique *consistent* measure Q and provides the channel from $\mathcal{F}^T$ to $\mathcal{B}^T$.

Lemma 3.1: Given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ and a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, there exists a unique consistent measure $Q(df^T, da^T, db^T)$ on $\mathcal{F}^T \times \mathcal{A}^T \times \mathcal{B}^T$. Furthermore, the channel from $\mathcal{F}^T$ to $\mathcal{B}^T$ for each t and all b_t is given by
$$Q(b_t \,|\, F^t = f^t,\ B^{t-1} = b^{t-1}) = p(b_t \,|\, f^t(b^{t-1}), b^{t-1}) \quad (2)$$
for Q almost all f^t, b^{t-1}.

Proof: Let $Q(df^T, da^T, db^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1}) \otimes \delta_{\{f_t(b^{t-1})\}}(da_t) \otimes p(db_t \,|\, a^t, b^{t-1})$. For finite T, this measure exists (see the Appendix). By the Ionescu–Tulcea theorem, this measure exists for the $T = \infty$ case. Clearly, this Q is consistent and by construction it is unique.

For each (f^t, b^t), the joint measure can be decomposed as
$$\begin{aligned} &Q(f^t, b^t) \\ &\quad = \sum_{a^t} Q(f^t, a^t, b^t) \\ &\quad = \sum_{a^t} \prod_{i=1}^t p(f_i \,|\, f^{i-1}) \delta_{\{f_i(b^{i-1})\}}(a_i) p(b_i \,|\, a^i, b^{i-1}) \\ &\quad = p(b_t \,|\, f^t(b^{t-1}), b^{t-1}) p(f_t \,|\, f^{t-1}) \\ &\qquad \times \sum_{a^{t-1}} \prod_{i=1}^{t-1} p(f_i \,|\, f^{i-1}) \delta_{\{f_i(b^{i-1})\}}(a_i) p(b_i \,|\, a^i, b^{i-1}) \\ &\quad = p(b_t \,|\, f^t(b^{t-1}), b^{t-1}) Q(f^t, b^{t-1}). \end{aligned}$$
Thus, we have shown (2). □

Hence, for any sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$, the stochastic kernel $p(db_t \,|\, f^t(b^{t-1}),\, b^{t-1})$ can be chosen as a version of the regular conditional distribution $Q(db_t \,|\, F^t = f^t, B^{t-1} = b^{t-1})$. Thus, the stochastic kernels $\{p(db_t \,|\, f^t(b^{t-1}),\, b^{t-1})\}_{t=1}^T$ can be viewed as the channel from $\mathcal{F}^T$ to $\mathcal{B}^T$. Note that the dependence is on $f^t(b^{t-1})$ and not f^t. We will see in Section V that this observation will greatly simplify computation.

The almost sure qualifier in (2) comes from the fact that $Q(f^t, b^{t-1})$ may equal zero for some f^t, b^{t-1}. This can happen, for example, if either f^t has zero probability of appearing under $P_{F^T}(df^T)$ or b^{t-1} has zero probability of appearing under the channel $\{p(db_t \,|\, a^t, b^{t-1})\}$.

A distribution P_W on $\mathcal{W}$ induces a measure P_{F^T} on $\mathcal{F}^T$.

Corollary 3.1: A distribution P_W on $\mathcal{W}$, a channel code $\{f^T[w]\}_{w=1}^M$, and the channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$ uniquely define a measure $Q(dw, da^T, db^T)$ on $\mathcal{W} \times \mathcal{A}^T \times \mathcal{B}^T$. Furthermore, the channel from $\mathcal{W}$ to $\mathcal{B}^T$ for each t and all b_t is given for Q almost all w, b^{t-1} by
$$Q(b_t \,|\, W = w, B^{t-1} = b^{t-1}) = p(b_t \,|\, f^t[w](b^{t-1}), b^{t-1}).$$

IV. Directed Information

As discussed in Section II, the traditional mutual information is insufficient for dealing with channels with feedback. Here we generalize Massey's [23] and Marko's [22] notion of *directed information* to take into account any time ordering of the random variables of interest. But first we generalize Kramer's [20] notion of causal conditioning to arbitrary time orderings.

Definition 4.1: We are given a sequence of stochastic kernels $\{p(da_i \,|\, a^{i-1})\}_{i=1}^N$. Let $I = \{i_1, \ldots, i_K\} \subseteq \{1, \ldots, N\}$ where $1 \le i_1 < i_2 < \cdots < i_K \le N$. Let $I^c = \{1, \ldots, N\} \setminus I$. Let $A^I = (A_{i_1}, \ldots, A_{i_K})$. Define A^{I^c} similarly. Then, the *directed stochastic kernel of A^I with respect to A^{I^c}* is
$$\vec{p}_{A^I \,|\, A^{I^c}}\left(da^I \,|\, a^{I^c}\right) = \otimes_{k=1}^K p_{A_{i_k} \,|\, A^{i_k-1}}\left(da_{i_k} \,|\, a^{i_k-1}\right).$$

For each a^{I^c}, the directed stochastic kernel $\vec{p}_{A^I \mid A^{I^c}}(dA^I \mid a^{I^c})$ is a well-defined measure (see the Appendix). For example

$$\int f(a^N)\vec{p}_{A^I \mid A^{I^c}}\left(da^I \mid a^{I^c}\right) = \int p(da_{i_1} \mid a^{i_1-1}) \int p(da_{i_2} \mid a^{i_2-1}) \cdots \int p\left(da_{i_k} \mid a^{i_k-1}\right) f(a^N)$$

for all bounded functions f measurable with respect to the product σ-algebra on $\mathcal{A}^N$. Note that this integral is a measurable function of a^{I^c}.

One needs to be careful when computing the marginals of a directed stochastic kernel. For example, if given $p(da_1)$, $p(da_2 \mid a_1)$, and $p(da_3 \mid a_1, a_2)$ with the resulting joint measure $P(da_1, da_2, da_3)$, then with the obvious time ordering

$$\sum_{a_1 \in \mathcal{A}} \vec{p}(a_1, a_3 \mid a_2) = \sum_{a_1 \in \mathcal{A}} (p(a_1)p(a_3 \mid a_1, a_2)) \neq P(a_3 \mid a_2)$$

for P almost all a_1, a_2, a_3 unless $A_1 - A_2 - A_3$ forms a Markov chain under P. Here $P(a_3 \mid a_2)$ represents the conditional probability under P.

Definition 4.2: Given a sequence of stochastic kernels $\{p(da_i \mid a^{i-1})\}_{i=1}^N$ and $I \subseteq \{1, \ldots, N\}$, the *directed information* is defined as

$$I\left(A^I \to A^{I^c}\right) = D\left(P_{A^I, A^{I^c}} \,\|\, \vec{P}_{A^I \mid A^{I^c}} P_{A^{I^c}}\right) \quad (3)$$

where $D(\cdot \,\|\, \cdot)$ is the divergence, $P_{A^I, A^{I^c}}(da^I, da^{I^c}) = \vec{p}_{A^{I^c} \mid A^I}(da^{I^c} \mid a^I) \otimes \vec{p}_{A^I \mid A^{I^c}}(da^I \mid a^{I^c})$, and $\vec{P}_{A^I \mid A^{I^c}} P_{A^{I^c}}(da^I, da^{I^c}) = \vec{p}_{A^I \mid A^{I^c}}(da^I \mid a^{I^c}) \otimes P_{A^{I^c}}(da^{I^c})$ [here $P_{A^{I^c}}(da^{I^c})$ is the marginal of $P_{A^I, A^{I^c}}(da^I, da^{I^c})$].

Note that this definition is more general than the one given by Massey [23]. We can recover Massey's definition of directed information by applying Definition 4.2 to $A^I = A^T$ and $A^{I^c} = B^T$ with the time ordering given in (1): $I(A^T \to B^T) = \sum_{t=1}^T I(A^t;\ B_t \mid B^{t-1})$. Unlike the chain rule for mutual information, the superscript on A in the summation is "t" and not "T." From Definition 4.2, one can easily show

$$I(A^T \to B^T) = E\left[\log \frac{\vec{p}_{B^T \mid A^T}(B^T \mid A^T)}{P_{B^T}(B^T)}\right] = E\left[\log \frac{p_{A^T \mid B^T}(A^T \mid B^T)}{\vec{p}_{A^T \mid B^T}(A^T \mid B^T)}\right]$$

where the stochastic kernel $p_{A^T \mid B^T}(da^T \mid b^T)$ is a version of the conditional distribution $P(da^T \mid b^T)$. The second equality shows that the directed information is the ratio between the posterior distribution and a "causal" prior distribution.

Note that $I(A^T; B^T) = E[\log \frac{\vec{p}(B^T \mid A^T)\vec{p}(A^T \mid B^T)}{P(B^T)P(A^T)}] = I(A^T \to B^T) + I(B^T \to A^T)$. By Definition 4.2 and time ordering (1), we have $I(B^T \to A^T) = \sum_{t=1}^T I(A_t; B^{t-1} \mid A^{t-1})$. Now $I(B^T \to A^T) = 0$ if and only if for each t the following $A_t - A^{t-1} - B^{t-1}$ forms a Markov chain under P. This Markov chain can be interpreted as there being no "information" flowing from the receiver to the transmitter. Because divergence is nonnegative, we can conclude that $I(A^T; B^T) \geq I(A^T \to B^T)$ with equality if and only if there is no feedback [23], [20].

A. Information Density, Directed Information, and Capacity

When computing the capacity of a channel it will turn out that we will need to know the convergence properties of the random variables $\frac{1}{T} \log \frac{P_{A^T, B^T}(A^T, B^T)}{\vec{P}_{A^T \mid B^T} P_{B^T}(A^T, B^T)}$. This is the *normalized information density* discussed in [34] suitably generalized to treat feedback. If there are reasonable regularity properties, like information stability (see below), then these random variables will converge in probability to a deterministic limit. In the absence of any such structure, we are forced to follow Verdú and Han's lead and define the following "floor" and "ceiling" limits [34].

The *limsup in probability* of a sequence of random variables $\{X_t\}$ is defined as the smallest extended real number α such that $\forall \epsilon > 0 \lim_{t \to \infty} \Pr[X_t \geq \alpha + \epsilon] = 0$. The *liminf in probability* of a sequence of random variables $\{X_t\}$ is defined as the largest extended real number α such that $\forall \epsilon > 0 \lim_{t \to \infty} \Pr[X_t \leq \alpha - \epsilon] = 0$.

Let $\vec{i}(a^T; b^T) = \log \frac{P_{A^T, B^T}(a^T, b^T)}{\vec{P}_{A^T \mid B^T} P_{B^T}(a^T, b^T)}$. For a sequence of joint measures $\{P_{A^T, B^T}\}_{T=1}^\infty$, let

$$\underline{I}(A \to B) = \liminf_{\text{in prob}} \frac{1}{T} \vec{i}(A^T; B^T)$$

and

$$\overline{I}(A \to B) = \limsup_{\text{in prob}} \frac{1}{T} \vec{i}(A^T; B^T).$$

Lemma 4.1: For any sequence of joint measures $\{P_{A^T, B^T}\}_{T=1}^\infty$, the following holds: $\underline{I}(A \to B) \leq \liminf_{T \to \infty} \frac{1}{T} I(A^T \to B^T) \leq \limsup_{T \to \infty} \frac{1}{T} I(A^T \to B^T) \leq \overline{I}(A \to B)$.

Proof: See the Appendix. □

We extend Pinsker's [25] notion of information stability. A given sequence of joint measures $\{P_{A^T, B^T}\}_{T=1}^\infty$ is *directed information stable* if $\lim_{T \to \infty} P(|\frac{\vec{i}(A^T; B^T)}{I(A^T \to B^T)} - 1| > \epsilon) = 0 \, \forall \epsilon > 0$. The following lemma shows that directed information stability implies $\frac{1}{T} \vec{i}(a^T; b^T)$ and concentrates around its mean $\frac{1}{T} I(A^T \to B^T)$. This mean needs not necessarily converge.

Lemma 4.2: If the sequence of joint measures $\{P_{A^T, B^T}\}_{T=1}^\infty$ is *directed information stable*, then $\underline{I}(A \to B) = \liminf_{T \to \infty} \frac{1}{T} I(A^T \to B^T) \leq \limsup_{T \to \infty} \frac{1}{T} I(A^T \to B^T) = \overline{I}(A \to B)$.

Proof: Directed information stability implies for all $\epsilon > 0$

$$\lim_{T \to \infty} P\left(\left|\frac{1}{T} \vec{i}(A^T; B^T) - \frac{1}{T} I(A^T \to B^T)\right| > \frac{1}{T} I(A^T \to B^T)\epsilon\right) = 0.$$

Because $\mathcal{B}$ is finite, we know $\frac{1}{T} I(A^T \to B^T) \leq \log |\mathcal{B}|$, hence for all $\epsilon > 0$

$$\lim_{T \to \infty} P\left(\left|\frac{1}{T} \vec{i}(A^T; B^T) - \frac{1}{T} I(A^T \to B^T)\right| > \epsilon\right) = 0.$$

This observation along with Lemma 4.1 proves the lemma. □

To compute the different "information" measures, we need to determine the joint measure $P_{A^T,B^T}(da^T, db^T)$. This can be done if we are given a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$ and we specify a sequence of kernels $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$.

Definition 4.3: A *channel input distribution* is a sequence of kernels $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$. A *channel input distribution without feedback* is a channel input distribution with the further condition that for each t the kernel $p(da_t \,|\, a^{t-1}, b^{t-1})$ is independent of b^{t-1}. (Specifically, $p(da_t \,|\, a^{t-1}, b^{t-1}) = p(da_t \,|\, a^{t-1}, \tilde{b}^{t-1})\ \forall b^{t-1}, \tilde{b}^{t-1}$.)

Let $\mathcal{D}_T = \{\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\}$ be the set of all channel input distributions. Let $\mathcal{D}_T^{\text{nfb}} \subset \mathcal{D}_T$ be the set of channel input distributions without feedback. We now define the directed information optimization problems. Fix a channel $\{p(db_t \,|\, a^t, b^{t-1})\}$. For finite T, let

$$C_T = \sup_{\mathcal{D}_T} \frac{1}{T} I(A^T \to B^T)$$

and

$$C_T^{\text{nfb}} = \sup_{\mathcal{D}_T^{\text{nfb}}} \frac{1}{T} I(A^T \to B^T) = \sup_{\mathcal{D}_T^{\text{nfb}}} \frac{1}{T} I(A^T; B^T).$$

For the infinite horizon case, let

$$C = \sup_{\{\mathcal{D}_T\}_{T=1}^\infty} \underline{I}(A \to B) \tag{4}$$

and

$$C^{\text{nfb}} = \sup_{\{\mathcal{D}_T^{\text{nfb}}\}_{T=1}^\infty} \underline{I}(A \to B) = \sup_{\{\mathcal{D}_T^{\text{nfb}}\}_{T=1}^\infty} \underline{I}(A; B).$$

Verdú and Han proved the following theorem for the case without feedback [34].

Theorem 4.1: For channels without feedback, $C^{\text{o,nfb}} = C^{\text{nfb}}$.

In a certain sense, we already have the solution to the coding problem for channels with feedback. Specifically, Lemma 3.1 tells us that the feedback channel problem is equivalent to a new channel coding problem without feedback. This new channel is from $\mathcal{F}^T$ to $\mathcal{B}^T$ and has channel kernels defined by (2). Thus, we can directly apply Theorem 4.1 to this new channel.

This can be a very complicated problem to solve. We would have to optimize the mutual information over distributions on code functions. The directed information optimization problem can often be simpler. One reason is that we can work directly on the original $\mathcal{A}^T \times \mathcal{B}^T$ space and not on the $\mathcal{F}^T \times \mathcal{B}^T$ space. The second half of this paper describes a stochastic control approach to solving this optimization. In the next section, though, we present the feedback coding theorem.

V. Coding Theorem for Channels With Feedback

In this section, we prove the following theorem.

Theorem 5.1: For channels with feedback, $C^{\text{o}} = C$.

We first give a high-level summary of the issues involved. The converse part is straightforward. For any channel code and channel, we know by Lemma 3.1 that there exists a unique consistent measure $Q(df^T, da^T, db^T)$. From this measure, we can compute the induced channel input distribution $\{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$. (These stochastic kernels are a version of the appropriate conditional probabilities.) Now $\{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T \in \mathcal{D}_T$ but it needs not be the supremizing channel input distribution. Thus, the directed information under the induced channel input distribution may be less than the directed information under the supremizing channel input distribution. This is how we will show $C^{\text{o}} \le C$.

The direct part is the interesting part of the Theorem 5.1. Here, we take the optimizing channel input distribution $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$ and construct a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$. We then prove the direct part of the coding theorem for the channel from $\mathcal{F}^T$ to $\mathcal{B}^T$ by the usual techniques for channels without feedback. By a suitable construction of P_{F^T}, it can be shown that the induced channel input distribution equals the original channel input distribution.

A. Main Technical Lemmas

We first discuss the channel input distribution induced by a given code-function distribution. Define the $\text{graph}(f_t) = \{(b^{t-1}, a_t) : f_t(b^{t-1}) = a_t\} \subset \mathcal{B}^{t-1} \times \mathcal{A}$. Let $\Upsilon_t(b^{t-1}, a_t) = \{f_t : (b^{t-1}, a_t) \in \text{graph}(f_t)\}$ and $\Upsilon^t(b^{t-1}, a^t) = \{f^t : (b^{j-1}, a_j) \in \text{graph}(f_j), j = 1, \ldots, t\}$.

In Lemma 3.1, we showed that the channel from $\mathcal{F}^T$ to $\mathcal{B}^T$ depends only on the channel from $\mathcal{A}^T$ to $\mathcal{B}^T$. Hence, for each t and all b_t, we have $Q(b_t \,|\, F^t = f^t, B^{t-1} = b^{t-1}) = p(b_t \,|\, f^t(b^{t-1}), b^{t-1}) = p(b_t \,|\, \tilde{f}^t(b^{t-1}), b^{t-1})$ where the first equality holds Q almost all f^t, b^{t-1} and the second equality holds $\forall \tilde{f}^t \in \Upsilon^t(b^{t-1}, f^t(b^{t-1}))$.

We now show that the induced channel input distribution only depends on the sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$.

Lemma 5.1: We are given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$, a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, and a consistent joint measure $Q(df^T, da^T, db^T)$. Then, the induced channel input distribution, for each t and all a_t, is given by

$$Q(a_t \,|\, a^{t-1}, b^{t-1}) = P_{F^T}\left(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1})\right) \tag{5}$$

for Q almost all a^{t-1}, b^{t-1}. Here $P_{F^T}(df^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1})$.

Proof: Note that $P_{F^T}(\Upsilon^{t-1}(b^{t-2}, a^{t-1})) = Q(\Upsilon^{t-1}(b^{t-2}, a^{t-1})) \ge Q(\Upsilon^{t-1}(b^{t-2}, a^{t-1}), a^{t-1}, b^{t-1}) = Q(a^{t-1}, b^{t-1})$. Thus, $Q(a^{t-1}, b^{t-1}) > 0$ implies $P_{F^T}(\Upsilon^{t-1}(b^{t-2}, a^{t-1})) > 0$. Hence, the right-hand side of (5) exists Q-almost surely. Now for each t and (a^{t-1}, b^{t-1}) such that $Q(a^{t-1}, b^{t-1}) > 0$, we have

$$\begin{aligned}
&Q(a^t, b^{t-1}) \\
&\quad = \sum_{f^t} \left(\prod_{i=1}^t p(f_i \,|\, f^{i-1}) \delta_{\{f_i(b^{i-1})\}}(a_i)\right) \vec{p}(b^{t-1} \,|\, a^{t-1}) \\
&\quad \overset{(a)}{=} \sum_{f^t \in \Upsilon^t(b^{t-1}, a^t)} \left(\prod_{i=1}^{t-1} p(f_i \,|\, f^{i-1})\right) \vec{p}(b^{t-1} \,|\, a^{t-1}) \\
&\quad \overset{(b)}{=} P_{F^T}(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1})) \\
&\qquad \times P_{F^T}(\Upsilon^{t-1}(b^{t-2}, a^{t-1}))\ \vec{p}(b^{t-1} \,|\, a^{t-1})
\end{aligned}$$

$$= P_{F^T}(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1})) \times \sum_{f^{t-1}} \left(\prod_{i=1}^{t-1} p(f_i \,|\, f^{i-1}) \delta_{\{f_i(b^{i-1})\}}(a_i) \right) \vec{p}(b^{t-1} \,|\, a^{t-1})$$
$$= P_{F^T}(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1})) Q(a^{t-1}, b^{t-1})$$

where (a) follows because $\vec{p}(b^{t-1} \,|\, a^{t-1})$ does not depend on f^{t-1} and the delta functions $\{\delta_{f_i(b^{i-1})}(a_i)\}$ restrict the sum over f^{t-1}. Line (b) follows because $Q(a^{t-1}, b^{t-1}) > 0$ and hence the conditional probability exists. □

The almost sure qualifier in (5) comes from the fact that $Q(a^{t-1}, b^{t-1})$ may equal zero for some a^{t-1}, b^{t-1}. This can happen, for example, if $P_{F^T}(df^T)$ puts zero mass on those f^{t-1} that produce a^{t-1} from b^{t-1} or if b^{t-1} has zero probability of appearing under the channel $\{p(db_t \,|\, a^t, b^{t-1})\}$.

We now show the equivalence of the directed information measures for both the "$\mathcal{F}^T - \mathcal{B}^T$" and the "$\mathcal{A}^T - \mathcal{B}^T$" channels.

Lemma 5.2: For each finite T and every consistent joint measure $Q(df^T, da^T, db^T)$, we have

$$\frac{Q_{F^T,B^T}(F^T, B^T)}{Q_{F^T} Q_{B^T}(F^T, B^T)} = \frac{Q_{A^T,B^T}(A^T, B^T)}{\vec{Q}_{A^T \,|\, B^T} Q_{B^T}(A^T, B^T)} \; Q - \text{a.s.} \tag{6}$$

hence $I(F^T; B^T) = I(A^T \to B^T)$. Furthermore, if given a sequence of consistent measures $\{Q(df^T, da^T, db^T)\}_{T=1}^{\infty}$, then $\underline{I}(F; B) = \underline{I}(A \to B)$.

Proof: Fix T finite. Then, for every (f^T, a^T, b^T) such that $Q(f^T, a^T, b^T) > 0$, we have

$$\frac{Q_{F^T,B^T}(f^T, b^T)}{Q_{F^T} Q_{B^T}(f^T, b^T)}$$
$$= \frac{\sum_{\tilde{a}^T} Q(f^T, \tilde{a}^T, b^T)}{Q_{B^T}(b^T) Q_{F^T}(f^T)}$$
$$= \frac{\sum_{\tilde{a}^T} \prod_{t=1}^{T} p(f_t \,|\, f^{t-1}) \delta_{\{f_t(b^{t-1})\}}(\tilde{a}_t) p(b_t \,|\, \tilde{a}^t, b^{t-1})}{Q_{B^T}(b^T) Q_{F^T}(f^T)}$$
$$= \frac{P_{F^T}(f^T) \vec{p}_{B^T \,|\, A^T}(b^T \,|\, a^T)}{Q_{B^T}(b^T) Q_{F^T}(f^T)}$$
$$\overset{(a)}{=} \frac{\vec{p}_{B^T \,|\, A^T}(b^T \,|\, a^T) P_{F^T}(\Upsilon(b^{T-1}, a^T))}{Q_{B^T}(b^T) \vec{q}_{A^T \,|\, B^T}(a^T \,|\, b^T)}$$
$$= \frac{\vec{p}_{B^T \,|\, A^T}(b^T \,|\, a^T) \sum_{\tilde{f}^T} \prod_{t=1}^{T} p(\tilde{f}_t \,|\, \tilde{f}^{t-1}) \delta_{\{\tilde{f}_t(b^{t-1})\}}(a_t)}{Q_{B^T}(b^T) \vec{q}_{A^T \,|\, B^T}(a^T \,|\, b^T)}$$
$$= \frac{\sum_{\tilde{f}^T} Q(\tilde{f}^T, a^T, b^T)}{Q_{B^T}(b^T) \vec{q}_{A^T \,|\, B^T}(a^T \,|\, b^T)}$$
$$= \frac{Q_{A^T,B^T}(a^T, b^T)}{\vec{Q}_{A^T \,|\, B^T} Q_{B^T}(a^T, b^T)}$$

where (a) follows because the Q marginal $Q(df^T) = P_{F^T}(df^T)$ and for $Q(f^T, a^T, b^T) > 0$ Lemma 5.1 shows $P_{F^T}(\Upsilon(b^{T-1}, a^T)) = \vec{q}_{A^T \,|\, B^T}(a^T \,|\, b^T)$.

Furthermore, if given a sequence of consistent measures $\{Q(df^T, da^T, db^T)\}_{T=1}^{\infty}$, (6) states that for each T the random variables on the left-hand side and right-hand side are almost surely equal. Hence, $\underline{I}(F; B) = \underline{I}(A \to B)$. □

We have shown how a code-function distribution induces a channel input distribution. As we discussed in the introduction to this section, we would like to choose a channel input distribution $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$, and construct a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^{T}$, such that the resulting induced channel input distribution $\{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$ equals the chosen channel input distribution. This is shown pictorially

$$\{p(da_t | a^{t-1}, b^{t-1})\} \to \{p(df_t | f^{t-1})\} \to \{q(da_t | a^{t-1}, b^{t-1})\}.$$

The first arrow represents the construction of the code-function distribution from the chosen channel input distribution. The second arrow is described by the result in Lemma 5.1. Lemma 5.2 states that $\underline{I}_Q(F; B) = \underline{I}_Q(A \to B)$. Let $P(da^T, db^T)$ correspond to the joint measure determined by the left channel input distribution in the diagram and the channel. If we can find conditions such that the induced channel input distribution $\{q(da_t \,|\, a^{t-1}, b^{t-1}\}$ equals the chosen channel input distribution $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}$, then $\underline{I}_Q(A \to B) = \underline{I}_P(A \to B)$. Consequently, $\underline{I}_Q(F; B) = \underline{I}_P(A \to B)$.

Definition 5.1: We call a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^{T}$, with resulting joint measure $P_{F^T}(df^T)$, *good with respect to the channel input distribution* $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$ if for each t and all a^t, b^{t-1}, we have

$$P_{F^T}(\Upsilon^t(b^{t-1}, a^t)) = \vec{p}(a^t \,|\, b^{t-1}).$$

Lemma 5.4 shows that good code-function distributions exist. But first, we show the equivalence of the chosen and induced channel input distributions when a good code-function distribution is used.

Lemma 5.3: We are given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^{T}$, a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^{T}$, and a consistent joint measure $Q(df^T, da^T, db^T)$. We are also given a channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$. The induced channel input distribution satisfies for each t and all a_t

$$Q(a_t \,|\, a^{t-1}, b^{t-1}) = r(a_t \,|\, a^{t-1}, b^{t-1}) \tag{7}$$

for Q almost all a^{t-1}, b^{t-1} if and only if the sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^{T}$ is good with respect to $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$.

Proof: First, assume that $\{p(df_t \,|\, f^{t-1})\}$ is good with respect to $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}$. Then, for each t and all a_t

$$Q(a_t \,|\, a^{t-1}, b^{t-1}) \overset{(a)}{=} P_{F^T}(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1}))$$
$$= \frac{\vec{r}(a^t \,|\, b^{t-1})}{\vec{r}(a^{t-1} \,|\, b^{t-2})}$$
$$= r(a_t \,|\, a^{t-1}, b^{t-1})$$

where each equality holds Q almost all a^{t-1}, b^{t-1}. Line (a) follows from Lemma 5.1. Now assume that (7) holds. Then, $\forall t$ and $\forall a_t$, we have $P_{F^T}(\Upsilon_t(b^{t-1}, a_t) \,|\, \Upsilon^{t-1}(b^{t-2}, a^{t-1})) = Q(a_t \,|\, a^{t-1}, b^{t-1}) = r(a_t \,|\, a^{t-1}, b^{t-1})$ Q almost all a^{t-1}, b^{t-1}, where the first equality follows from Lemma 5.1. □

Lemma 5.4: For any channel input distribution $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^{T}$, there exists a sequence of code-functions stochastic kernels that are good with respect to it.

Proof: For all f^t, define $p(f_t \,|\, f^{t-1})$ as follows:

$$p(f_t \,|\, f^{t-1}) = \prod_{b^{t-1}} p(f_t(b^{t-1}) \,|\, f^{t-1}(b^{t-2}), b^{t-1}). \quad (8)$$

We first show that $p(f_t \,|\, f^{t-1})$ defined in (8) is a stochastic kernel. Note that for each t and all f^{t-1}, we have

$$\begin{aligned}
&\sum_{f_t} p(f_t \,|\, f^{t-1}) \\
&\quad = \sum_{f_t} \prod_{b^{t-1}} p(f_t(b^{t-1}) \,|\, f^{t-1}(b^{t-2}), b^{t-1}) \\
&\quad \overset{(a)}{=} \prod_{b^{t-1}} \sum_{a_t} p(a_t \,|\, f^{t-1}(b^{t-2}), b^{t-1}) \\
&\quad = 1
\end{aligned}$$

where (a) follows because the sum is over *all* functions $f_t : \mathcal{B}^{t-1} \to \mathcal{A}$. The sum over f_t can be viewed as a $|\mathcal{B}|^{t-1}$-fold summation over the alphabet $\mathcal{A}$, one for each element in the domain of f_t. Thus, the sum of products can be written as a product of sums.

We now show by induction that for each t and all a^t, b^{t-1}, we have $P_{F^T}(\Upsilon^t(b^{t-1}, a^t)) = \vec{p}(a^t \,|\, b^{t-1})$. For $t = 1$ and (8), we have $P_{F^T}(\Upsilon_1(a_1)) = \sum_{f_1 \in \Upsilon_1(a_1)} p(f_1) = \vec{p}(a_1)$. For $t+1$, we have

$$\begin{aligned}
&P_{F^T}(\Upsilon^{t+1}(b^t, a^{t+1})) \\
&\quad = \sum_{f^t \in \Upsilon^t(b^{t-1}, a^t)} \prod_{i=1}^{t} p(f_i \,|\, f^{i-1}) \\
&\qquad \times \sum_{f_{t+1} \in \Upsilon_{t+1}(b^t, a_{t+1})} \prod_{\tilde{b}^t} p(f_{t+1}(\tilde{b}^t) \,|\, f^t(\tilde{b}^{t-1}), \tilde{b}^t) \\
&\quad \overset{(a)}{=} \sum_{f^t \in \Upsilon^t(b^{t-1}, a^t)} \prod_{i=1}^{t} p(f_i \,|\, f^{i-1})\, p(a_{t+1} \,|\, a^t, b^t) \\
&\qquad \times \sum_{f_{t+1} \in \Upsilon_{t+1}(b^t, a_{t+1})} \prod_{\tilde{b}^t \neq b^t} p(f_{t+1}(\tilde{b}^t) \,|\, f^t(\tilde{b}^{t-1}), \tilde{b}^t) \\
&\quad \overset{(b)}{=} \sum_{f^t \in \Upsilon^t(b^{t-1}, a^t)} \prod_{i=1}^{t} p(f_i \,|\, f^{i-1})\, p(a_{t+1} \,|\, a^t, b^t) \\
&\qquad \times \prod_{\tilde{b}^t \neq b^t} \sum_{\tilde{a}_{t+1}} p(\tilde{a}_{t+1} \,|\, f^t(\tilde{b}^{t-1}), \tilde{b}^t) \\
&\quad = P_{F^T}(\Upsilon^t(b^{t-1}, a^t)) p(a_{t+1} \,|\, a^t, b^t) \\
&\quad \overset{(c)}{=} \vec{p}(a^t \,|\, b^{t-1})\, p(a_{t+1} \,|\, a^t, b^t) \\
&\quad = \vec{p}(a^{t+1} \,|\, b^t)
\end{aligned}$$

where (a) follows because $f^{t+1} \in \Upsilon^{t+1}(b^t, a^{t+1})$. Line (b) follows from an argument similar to that given above. Specifically, the sum over $f_{t+1} \in \Upsilon_{t+1}(b^t, a_{t+1})$ can be viewed as a $|\mathcal{B}|^t - 1$-fold summation over the alphabet $\mathcal{A}$ (the -1 comes from removing the b^t term). Line (c) follows from the induction hypothesis. □

In the above construction (8), we have enforced independence across the different b^{t-1}. Specifically, for $b^{t-1} \neq \bar{b}^{t-1}$, we have

$$\begin{aligned}
&P_{F^T}(\Upsilon(b^{t-1}, a_t) \cap \Upsilon(\bar{b}^{t-1}, \bar{a}_t) \,|\, f^{t-1}) \\
&\quad = P_{F^T}(\Upsilon(b^{t-1}, a_t) \,|\, f^{t-1}) P_{F^T}(\Upsilon(\bar{b}^{t-1}, \bar{a}_t) \,|\, f^{t-1}).
\end{aligned}$$

We do not need to assume this independence in order to find a sequence of code-function stochastic kernels good with respect to a given channel input distribution. For example, it is known that Gaussian (linear) channel input distributions are optimal for Gaussian channels. For more details, see [8], [31], and [40]. When dealing with more complicated alphabets, one may want the functions f_t to be continuous with respect to the topologies of $\mathcal{A}$ and $\mathcal{B}$. Continuity is trivially satisfied in the finite-alphabet case. See [39] for an example of a finite-alphabet, finite-state Markov channel.

Note that it is possible for distinct code-function stochastic kernels to induce the same channel input distribution (almost surely.) Similarly, there may be many code-functions stochastic kernels that are good with respect to a given channel input distribution (and hence, via Lemma 5.3, induce the same channel input distribution). As an example consider the case when the channel input distribution does not depend on the channel output: $\{p(da_t \,|\, a^{t-1})\}_{t=1}^{T}$. One choice of P_{F^T} is given by (8)

$$p(f_t \,|\, f^{t-1}) = \prod_{b^{t-1}} p(f_t(b^{t-1}) \,|\, f^{t-1}(b^{t-2})).$$

By Lemma 5.4, this P_{F^T} is good with respect to $\{p(da_t \,|\, a^{t-1})\}_{t=1}^{T}$. Another choice puts zero mass on code functions that depend on feedback (i.e., only use codewords)

$$P_{\bar{F}^T}(f^T) = \begin{cases} \prod_{t=1}^{T} p(a_t \,|\, a^{t-1}), & \text{if } f_t(b^{t-1}) = a_t, \forall b^{t-1}, \forall t \\ 0, & \text{else.} \end{cases}$$

One can show that this $P_{\bar{F}^T}(df^T)$ is good with respect to $\{p(da_t \,|\, a^{t-1})\}$ by checking for each t, $P_{\bar{F}^T}(\Upsilon^t(b^{t-1}, a^t)) = \prod_{i=1}^{t} p(a_t \,|\, a^{i-1})$.

For memoryless channels, we know the optimal channel input distribution is $\{p(da_t)\}_{t=1}^{T}$. Feedback in this case cannot increase capacity but that does not preclude us from using feedback. For example, feedback is known to increase the error exponent and hence decrease latency.

B. Feedback Channel Coding Theorem

Now we prove the feedback channel coding Theorem 5.1. We start with the converse part and then prove the direct part.

a) Converse Theorem: Choose a (T, M, ϵ) channel code $\{f^T[w]\}_{w=1}^{M}$. Place a prior probability $\frac{1}{M}$ on each code function $f^T[w]$. By Lemma 3.1 and Corollary 3.1, this defines consistent measures $Q(df^T, da^T, db^T)$ and $Q(dw, da^T, db^T)$. The following is a generalization of the Verdú–Han converse [34] for bounding the block error probability.

Lemma 5.5: Every (T, M, ϵ) channel code satisfies $\forall \gamma > 0$

$$\epsilon \geq Q\left(\frac{1}{T} \log \frac{Q_{A^T, B^T}(A^T, B^T)}{\vec{Q}_{A^T \,|\, B^T} Q_{B^T}(A^T, B^T)} \leq \frac{1}{T} \log M - \gamma\right) - 2^{-\gamma T}.$$

Proof: Choose a $\gamma > 0$. Let $D_w \subset \mathcal{B}^T$ be the decoding region for message w. The only restriction we place on the decoding regions is that they do not intersect: $D_w \cap D_{\tilde{w}} = \emptyset \; \forall \tilde{w} \neq w$. (This is always true when using a channel decoder: $D_w = \{b^T : g(b^T) = w\}$.)

Under this restriction on the decoder, Verdú and Han show [34, Th. 4] that any (T, M, ϵ) channel code for the channel $\{p(db_t \,|\, f^t, b^{t-1})\}$ without feedback [see (2)] satisfies for all $\gamma > 0$

$$\epsilon \geq Q\left(\frac{1}{T}\log\frac{Q_{F^T,B^T}(F^T,B^T)}{Q_{F^T}Q_{B^T}(F^T,B^T)} \leq \frac{1}{T}\log M - \gamma\right) - 2^{-\gamma T}.$$

By Lemma 5.2, we know that $\frac{Q_{F^T,B^T}(F^T,B^T)}{Q_{F^T}Q_{B^T}(F^T,B^T)} = \frac{Q_{A^T,B^T}(A^T,B^T)}{\vec{Q}_{A^T\,|\,B^T}Q_{B^T}(A^T,B^T)}$ holds Q-a.s. □

Note that for any channel decoder g the decoding regions $D_w = \{b^T : g(b^T) = w\}$ do not overlap and hence we are able to apply the Verdú and Han converse. Thus, the lemma holds independently of the decoder g that one uses.

Theorem 5.2: The channel capacity $C^{\mathrm{o}} \leq C$.

Proof: Assume there exists a sequence of (T, M_T, ϵ_T) channel codes with $\lim_{T\to\infty}\epsilon_T = 0$ and with rate $R = \liminf_{T\to\infty}\frac{1}{T}\log M_T$. By Lemma 5.5, we know that for all $\gamma > 0$

$$\epsilon_T \geq Q\left(\frac{1}{T}\log\frac{Q_{A^T,B^T}(A^T,B^T)}{\vec{Q}_{A^T\,|\,B^T}Q_{B^T}(A^T,B^T)} \leq \frac{1}{T}\log M_T - \gamma\right) - 2^{\gamma T}.$$

The first term on the right-hand side must go to zero as $T \to \infty$ because the error is going to zero. By the definition of C, we know $\limsup_{T\to\infty}\frac{1}{T}\log M_T - \gamma < C$. Since γ can be chosen to be arbitrarily small, we see $R \leq C$. Thus, $C^{\mathrm{o}} \leq C$. □

b) Direct Theorem: We will prove the direct theorem via a random coding argument. The following is a generalization of Feinstein's lemma [14], [34].

Lemma 5.6: Fix a time T, a $0 < \epsilon < 1$, and a channel $\{p(db_t \,|\, b^{t-1}, a^t)\}_{t=1}^T$. Then, for all $\gamma > 0$ and any channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$, there exists a (T, M, ϵ) channel code that satisfies

$$\epsilon \leq R\left(\frac{1}{T}\log\frac{R_{A^T,B^T}(A^T,B^T)}{\vec{R}_{A^T\,|\,B^T}R_{B^T}(A^T,B^T)} \leq \frac{1}{T}\log M + \gamma\right) + 2^{-\gamma T}$$

where $R_{A^T,B^T}(da^T, db^T) = \otimes_{t=1}^T p(db_t \,|\, b^{t-1}, a^t) \otimes r(da_t \,|\, a^{t-1}, b^{t-1})$.

Proof: Let $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ be any sequence of code-function stochastic kernels good with respect to the channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$. Let $Q(df^T, da^T, db^T)$ be the consistent joint measure determined by this $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ and the channel.

Verdú and Han show in [34, Th. 2] that for the channel $\{p(db_t \,|\, f^t, b^{t-1})\}_{t=1}^T$ without feedback and for every $\gamma > 0$, there exists a channel code (T, M, ϵ) that satisfies

$$\epsilon \leq Q\left(\frac{1}{T}\log\frac{Q_{F^T,B^T}(F^T,B^T)}{Q_{F^T}Q_{B^T}(F^T,B^T)} \leq \frac{1}{T}\log M + \gamma\right) + 2^{-\gamma T}.$$

Lemma 5.2 shows $\frac{Q_{F^T,B^T}(F^T,B^T)}{Q_{F^T}Q_{B^T}(F^T,B^T)} = \frac{Q_{A^T,B^T}(A^T,B^T)}{\vec{Q}_{A^T\,|\,B^T}Q_{B^T}(A^T,B^T)}$ holds Q almost surely. Lemma 5.3 shows $Q(a_t \,|\, a^{t-1}, b^{t-1}) = r(a_t \,|\, a^{t-1}, b^{t-1})$ Q almost surely. Hence, $Q(da^T, db^T) = R_{A^T,B^T}(da^T, db^T)$. □

Recall that the random coding argument underlying this result requires a distribution on channel codes given by randomly drawing M code functions uniformly from $Q(df^T)$.

Theorem 5.3: The channel capacity $C^{\mathrm{o}} \geq C$.

Proof: We follow [34]. Fix an $\epsilon > 0$. We will show that C is an ϵ-achievable rate by demonstrating for every $\delta > 0$ and all sufficiently large T that there exists a sequence of $(T, M, 2^{-\frac{\delta}{4}T} + \frac{\epsilon}{2})$ codes with rate $C - \delta \leq \frac{\log M}{T} \leq C - \frac{\delta}{2}$. If in the previous lemma we choose $\gamma = \frac{\delta}{4}$,, then we get

$$\begin{aligned} R\left(\frac{1}{T}\log\frac{R_{A^T,B^T}(A^T,B^T)}{\vec{R}_{A^T\,|\,B^T}R_{B^T}(A^T,B^T)} \leq \frac{1}{T}\log M + \frac{\delta}{4}\right) + 2^{-\frac{\delta}{4}T} \\ \leq \left(\frac{1}{T}\log\frac{R_{A^T,B^T}(A^T,B^T)}{\vec{R}_{A^T\,|\,B^T}R_{B^T}(A^T,B^T)} \leq C - \frac{\delta}{4}\right) + 2^{-\frac{\delta}{4}T} \\ \leq \frac{\epsilon}{2} + 2^{-\frac{\delta}{4}T} \end{aligned}$$

where the second inequality holds for all sufficiently large T. To see this, note that by the definition of C and T large enough, the mass below $C - \frac{\delta}{4}$ has probability zero. □

Proof of Theorem 5.1: By combining Theorems 5.2 and 5.3, we have $C^{\mathrm{o}} = C$. □

We have shown that C is the feedback channel capacity. It should be clear that if we restrict ourselves to channels without feedback then we recover the original coding theorem by Verdú and Han [34].

We end this section with a discussion of the strong converse.

Definition 5.2: A channel with feedback capacity C^{o} has a *strong converse* if for all $\delta > 0$ and every sequence of channel codes, $\{(T, M_T, \epsilon_T)\}$, for which $\liminf\frac{\log M_T}{T} > C^{\mathrm{o}} + \delta$ satisfies $\lim_{T\to\infty}\epsilon_T = 1$.

Proposition 5.1: A channel with feedback capacity C^{o} has a strong converse if and only if $\sup_{\{\mathcal{D}_T\}_{T=1}^\infty}\underline{I}(A \to B) = \sup_{\{\mathcal{D}_T\}_{T=1}^\infty}\overline{I}(A \to B)$ and thus $C^{\mathrm{o}} = \lim_{T\to\infty}\frac{1}{T}I(A^T \to B^T)$.

Proof: The first part follows from [34, Th. 7]. The latter part follows from Theorem 5.1, Lemma 4.1, and the finiteness of $\mathcal{B}$. □

C. General Information Pattern

So far we have assumed that the encoder has access to all the channel outputs B^{t-1}. There are many situations, though, where the information pattern [36] at the encoder may be restricted. Let $\mathcal{E}$ be a finite set and let $E_t = \psi_t(B^t)$. Here the measurable functions $\psi_t : \mathcal{B}^t \to \mathcal{E}$ determine the information fed back from the decoder to the encoder. Let $\Psi^T = \{\psi_t\}_{t=1}^T$. In the case of Δ-delayed feedback, we have $E_t = \psi_t(B^t) = B_{t-\Delta+1}$. If $\Delta = 1$, then $E_t = \psi_t(B^t) = B_t$ and we are in the situation discussed above. Quantized channel output feedback can be handled by letting the $\{\psi_t\}$ be quantizers. The time ordering is $A_1, B_1, E_1, A_2, B_2, E_2, \ldots, A_T, B_T, E_T$.

A *channel code function with information pattern* Ψ^T is a sequence of T measurable maps $\{f_t\}_{t=1}^T$ such that $f_t : \mathcal{E}^{t-1} \to \mathcal{A}$ taking $e^{t-1} \mapsto a_t$. Denote the set of all code functions with re-

stricted information pattern Ψ^T by $\mathcal{F}^{T,\Psi} \subseteq \mathcal{F}^T$. The *operational capacity with information pattern* Ψ^∞, denoted by $C^{\mathrm{o},\Psi}$, is defined similarly to Definition 3.2.

Just as in Section III–A, we can define a joint measure $P(df^T, da^T, db^T, de^T)$ as the interconnection of the code functions and the channel $\{p(db_t \,|\, a^t, b^{t-1})\}$. Lemma 3.1 follows as before except that now condition two of consistency requires both $A_t = F^t(E^{t-1}), E_t = \Psi(B^t)\ \ Q-a.s.$

Define the *channel input distribution with information pattern* Ψ to be a sequence of stochastic kernels $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}$ with the further condition that for each t the kernel $p(da_t \,|\, a^{t-1}, b^{t-1}) = p(da_t \,|\, a^{t-1}, \psi^{t-1}(b^{t-1}))$. Let $\mathcal{D}_T^\Psi = \{\{p(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\}$ be the set of all channel input distributions with information pattern Ψ. Let

$$C_T^\Psi = \sup_{\mathcal{D}_T^\Psi} \frac{1}{T} I(A^T \to B^T), \qquad \text{for finite } T$$

and

$$C^\Psi = \sup_{\{\mathcal{D}_t^\Psi\}_{t=1}^\infty} \underline{I}(A \to B).$$

For the general information pattern, Lemmas 5.1–5.4 and Theorems 5.1–5.6 continue to hold with obvious modifications.

Theorem 5.4: For channels with information pattern Ψ, we have $C^{\mathrm{o},\Psi} = C^\Psi$.

Intuitively, the reason this result holds is because the feedback is a causal, deterministic function of the channel outputs. It would be interesting to examine the case with noisy feedback. Unfortunately, this is a much more complicated problem. It is related to the problem of channel coding with side information at the encoder.

D. Error Exponents

We can generalize Gallager's random coding error exponent [15] to feedback channels. Specifically, Propositions 5.2 and 5.3 show that the error exponent can be computed directly in terms of the "$\mathcal{A}^T - \mathcal{B}^T$" channel. See also [31].

Definition 5.3: We are given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$, a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, and a consistent joint measure $Q(df^T, da^T, db^T)$. The *random coding error exponent* is

$$E_T\left(R, \{p(df_t \,|\, f^{t-1}\}_{t=1}^T\right) = \max_{0\le\rho\le 1}\left(-\rho R - \frac{1}{T}\ln \sum_{b^T}\left[\sum_{f^T} Q(f^T)\left[Q(b^T|f^T)\right]^{\frac{1}{1+\rho}}\right]^{1+\rho}\right)$$

where $Q(db^T \,|\, f^T) = \otimes_{t=1}^T Q(db_t \,|\, f^t, b^{t-1})$ is defined as in Lemma 3.1.

From Section III, we know that we can view the channel with feedback as a channel without feedback from $\mathcal{F}^T$ to $\mathcal{B}^T$. Thus, we can directly apply [15, Th. 5.6.1] to see that $E_T\left(R, \{p(df_t \,|\, f^{t-1}\}_{t=1}^T\right)$ is the random coding error exponent for channel codes drawn from $P_{F^T}(df^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1})$.

We now show that we can simplify the form of the error exponent by writing it directly in terms of the channel input distribution defined on $\mathcal{A}^T \times \mathcal{B}^T$. To that end, we define the directed random coding error exponent.

Definition 5.4: Given a channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$ and a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, define the *directed random coding error exponent* to be

$$\vec{E}_T\left(R, \{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\right) = \max_{0\le\rho\le 1}\left(-\rho R - \frac{1}{T}\ln \sum_{b^T}\left[\sum_{a^T} \vec{r}(a^T|b^T)\left[\vec{p}(b^T|a^T)\right]^{\frac{1}{1+\rho}}\right]^{1+\rho}\right).$$

Proposition 5.2: We are given a sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$, a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$, and a consistent joint measure $Q(df^T, da^T, db^T)$. Let $\{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$ be a version of the induced channel input distribution given in (5) of Lemma 5.1. Then

$$\vec{E}_T\left(R, \{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\right) = E_T\left(R, \{p(df_t \,|\, f^{t-1}\}_{t=1}^T\right).$$

Proof: Under the consistent measure Q and for $0 \le \rho \le 1$ and all b^T, we have

$$\begin{aligned}
&\sum_{f^T} Q(f^T)\left[Q(b^T \,|\, f^T)\right]^{\frac{1}{1+\rho}} \\
&\quad= \sum_{f^T, a^T} Q(f^T, a^T)\left[Q(b^T \,|\, f^T)\right]^{\frac{1}{1+\rho}} \\
&\quad\overset{(a)}{=} \sum_{f^T, a^T} Q(f^T, a^T, b^T)\left[Q(b^T \,|\, f^T)\right]^{\frac{-\rho}{1+\rho}}\{Q(b^T \,|\, f^T) > 0\} \\
&\quad\overset{(b)}{=} \sum_{f^T, a^T} Q(f^T, a^T, b^T)\left[\vec{p}(b^T \,|\, a^T)\right]^{\frac{-\rho}{1+\rho}}\{\vec{p}(b^T \,|\, a^T) > 0\} \\
&\quad= \sum_{f^T, a^T}\left(\prod_{t=1}^T p(f_t \,|\, f^{t-1})\delta_{\{f_t(b^{t-1})\}}(a_t) p(b_t \,|\, a^t, b^{t-1})\right) \\
&\qquad\times\left[\vec{p}(b^T \,|\, a^T)\right]^{\frac{-\rho}{1+\rho}}\{\vec{p}(b^T \,|\, a^T) > 0\} \\
&\quad= \sum_{a^T}\left(\sum_{f^T}\prod_{t=1}^T p(f_t \,|\, f^{t-1})\delta_{\{f_t(b^{t-1})\}}(a_t)\right) \\
&\qquad\times \vec{p}(b^T \,|\, a^T)\left[\vec{p}(b^T \,|\, a^T)\right]^{\frac{-\rho}{1+\rho}}\{\vec{p}(b^T \,|\, a^T) > 0\} \\
&\quad= \sum_{a^T} P_{F^T}(\Upsilon^T(b^{T-1}, a^T))\left[\vec{p}(b^T \,|\, a^T)\right]^{\frac{1}{1+\rho}} \\
&\quad\overset{(c)}{=} \sum_{a^T} \vec{q}(a^T \,|\, b^T)\left[\vec{p}(b^T \,|\, a^T)\right]^{\frac{1}{1+\rho}}.
\end{aligned}$$

In line (a), $\{\cdot\}$ is the indicator function, line (b) follows from Lemma 3.1, and line (c) follows from Lemma 5.1. □

Proposition 5.3: We are given a channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1}\}_{t=1}^T$ and a channel $\{p(db_t \,|\, a^t, b^{t-1})\}_{t=1}^T$. Let the sequence of code-function stochastic kernels $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$ be good with respect to the channel input distribution $\{r(da_t \,|\, a^{t-1}, b^{t-1}\}_{t=1}^T$. Then

$$E_T\left(R, \{p(df_t|f^{t-1}\}_{t=1}^T\right) = \vec{E}_T\left(R, \{r(da_t|a^{t-1}, b^{t-1})\}_{t=1}^T\right).$$

Proof: We are given a channel input distribution $\{r(a_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$. In addition, $\{p(f_t \,|\, f^{t-1}\}_{t=1}^T$ is good with respect to this channel input distribution.

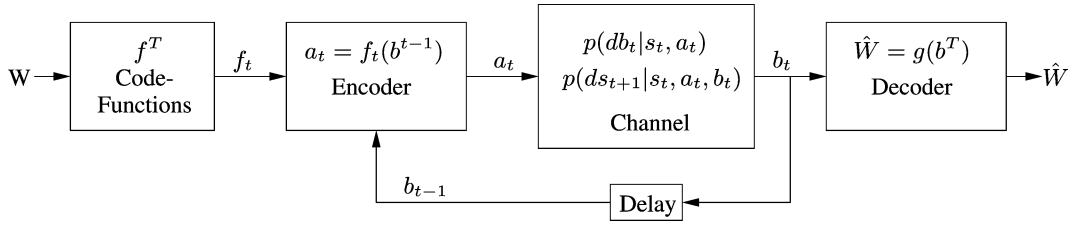

Fig. 2. Markov channel.

For this sequence of code-function stochastic kernels $\{p(f_t \,|\, f^{t-1}\}_{t=1}^T$, let $Q(df^T, da^T, db^T)$ be the associated consistent measure and let $\{q(a_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T$ be a version of the induced channel input distribution. By Lemma 5.3, we know for each t and all a_t that $q(a_t \,|\, a^{t-1}, b^{t-1}) = r(a_t \,|\, a^{t-1}, b^{t-1})$ for Q almost all a^{t-1}, b^{t-1}. Hence

$$\begin{aligned} &E_T\left(R, \{p(df_t \,|\, f^{t-1}\}_{t=1}^T\right) \\ &\quad \stackrel{(a)}{=} \vec{E}_T\left(R, \{q(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\right) \\ &\quad = \vec{E}_T\left(R, \{r(da_t \,|\, a^{t-1}, b^{t-1})\}_{t=1}^T\right) \end{aligned}$$

where (a) follows from Proposition 5.2. □

In summary, Propositions 5.2 and 5.3 show the equivalence of the random coding error exponent and the directed random coding error exponent. The latter is defined over $\mathcal{A}^T \times \mathcal{B}^T$. In Section VIII, we describe some cases where one can compute the directed error exponent.

VI. Markov Channels

In this section, we formulate the Markov channel feedback capacity problem. As before, let $\mathcal{A}, \mathcal{B}$ be spaces with a finite number of elements representing the channel input and channel output, respectively. Furthermore, let $\mathcal{S}$ be a state space with a finite number of elements with the counting σ-algebra. Let S_t, A_t, B_t be measurable random elements taking values in $\mathcal{S}, \mathcal{A}, \mathcal{B}$, respectively. See Fig. 2.

There is a natural time ordering on the random variables of interest

$$W, S_1, A_1, B_1, S_2, \ldots, \overbrace{A_t, B_t, S_{t+1}}^{t\text{th epoch}}, \ldots, S_T, A_T, B_T, \hat{W}. \tag{9}$$

First, at time 0, a message W is produced and the initial state S_1 is drawn. The order of events in each of the T epochs is described in (9). At the beginning of tth epoch, the channel input symbol A_t is placed on the channel by the transmitter, then B_t is observed by the receiver, then the state of the system evolves to S_{t+1}, and then, finally the receiver feeds back information to the transmitter. At the beginning of the $t+1$ epoch, the transmitter uses the feedback information to produce the next channel input symbol A_{t+1}. Finally, at time T, after observing B_T, the decoder outputs the reconstructed message $\hat{W}$.

Definition 6.1: A *Markov channel* consists of an initial state distribution $p(ds_1)$, the state transition stochastic kernels $\{p(ds_{t+1} \,|\, s_t, a_t, b_t)\}_{t=1}^{T-1}$, and the channel output stochastic kernels $\{p(db_t \,|\, s_t, a_t)\}_{t=1}^T$. If the stochastic kernel $p(ds_{t+1} \,|\, s_t, a_t, b_t)$ is independent of a_t, b_t for each $t = 1, \ldots, T-1$, then we say the channel is a *Markov channel without ISI*. Note that we are assuming that the kernels $\{p(ds_{t+1} \,|\, s_t, a_t, b_t)\}$ and $\{p(db_t \,|\, s_t, a_t)\}$ are stationary (independent of time).

As before, a *channel code function* is a sequence of T deterministic measurable maps $\{f_t\}_{t=1}^T$ such that $f_t : \mathcal{B}^{t-1} \to \mathcal{A}$, which takes $b^{t-1} \mapsto a_t$. We do not assume, for now, that the state of the channel is observable to the encoder or decoder. This will have the effect of restricting ourselves to channel input distributions of the form $\{p(da_t \,|\, a^{t-1}, b^{t-1})\}$ as opposed to $\{p(da_t \,|\, s^t, a^{t-1}, b^{t-1})\}$. We assume that we have *full* output feedback. This ensures that the information pattern at the receiver is nested in the information pattern at the transmitter. As we will see, this nesting allows us to use the dynamic programming methodology to compute capacity. Computing the capacity of Markov channels under *partial* output feedback, as described in Section V-C, turns out to be quite difficult and will not be treated here. Finally, we assume that *both* the encoder and the decoder know $p(ds_1)$. In Section VIII-A, we show how to introduce state feedback.

A. The Sufficient Statistic $\{\Pi_t\}$

Given a sequence of code-function distributions $\{p(df_t \,|\, f^{t-1})\}_{t=1}^T$, we can interconnect the Markov channel to the source. Via a straightforward generalization of Definition 3.3 and Lemma 3.1, one can show there exists a unique consistent measure: $Q(df^T, ds^T, da^T, db^T) = \otimes_{t=1}^T p(df_t \,|\, f^{t-1}) \otimes p(ds_t \,|\, s_{t-1}, a_{t-1}, b_{t-1}) \otimes \delta_{\{f_t(b^{t-1})\}}(da_t) \otimes p(db_t \,|\, s_t, a_t)$. Unlike in Lemma 3.1, determining the channel without feedback from $\mathcal{F}^T$ to $\mathcal{B}^T$ takes a bit more work. To that end, we introduce the sufficient statistics $\{\Pi_t\}$.

Let $\Pi(ds) \in \mathcal{P}(\mathcal{S})$ be an element in the space of probability measures on $\mathcal{S}$. Define a stochastic kernel from $\mathcal{P}(\mathcal{S}) \times \mathcal{A}$ to $\mathcal{S} \times \mathcal{B}$

$$r(ds, db \,|\, \pi, a) = p(db \,|\, s, a) \otimes \pi(ds). \tag{10}$$

The next lemma follows from Theorem A.3 in the Appendix.

Lemma 6.1: There exists a stochastic kernel $r(ds \,|\, \pi, a, b)$ from $\mathcal{P}(\mathcal{S}) \times \mathcal{A} \times \mathcal{B}$ to $\mathcal{S}$ such that

$$r(ds, db \,|\, \pi, a) = r(ds \,|\, \pi, a, b) \otimes r(db \,|\, \pi, a)$$

where $r(db \,|\, \pi, a)$ is the marginal of $r(ds, db \,|\, \pi, a)$. Specifically, for each b

$$r(b \,|\, \pi, a) = \sum_{\tilde{s}} p(b \,|\, \tilde{s}, a)\pi(\tilde{s}). \tag{11}$$

The statistic $\pi(ds)$ is often called the *a priori* distribution of the state and $r(ds \mid \pi, a, b)$ the *a posteriori* distribution of the state after observing a, b. We recursively define the sufficient statistics $\{\Pi_t\}_{t=1}^T$. Specifically, $\Pi_t : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \to \mathcal{P}(\mathcal{S})$ is defined as follows:

$$\pi_1(ds_1) = p(ds_1) \tag{12}$$

(where $p(ds_1)$ is given in Definition 6.1), and for each a^t, b^t and all s_{t+1}, let

$$\pi_{t+1}[a^t, b^t](s_{t+1}) = \sum_{s_t} p(s_{t+1} \mid s_t, a_t, b_t) r\left(s_t \mid \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t, b_t\right). \tag{13}$$

Equations (12) and (13) are the so-called filtering equations. Equation (13) implies there exists a stationary, measurable function Φ_Π such that $\pi_{t+1} = \Phi_\Pi(\pi_t, a_t, b_t)$ for all $t = 1, \ldots, T-1$. Note the statistic Π_t depends on information from *both* the transmitter and the receiver. It can be viewed as the combined transmitter and receiver estimate of the state.

The next lemma shows that the $\{\Pi_t\}$ are consistent.

Lemma 6.2: We are given a sequence of code-function stochastic kernels $\{p(df_t \mid f^{t-1})\}$, a Markov channel $p(ds_1), \{p(ds_{t+1} \mid s_t, a_t)\}, \{p(db_t \mid s_t, a_t)\}$, and a consistent joint measure $Q(df^T, ds^T, da^T, db^T)$. Then, for each t and all s_t, we have

$$Q(s_t \mid f^t, a^t, b^{t-1}) = \pi_t[a^{t-1}, b^{t-1}](s_t) \tag{14}$$

for Q almost all f^t, a^t, b^{t-1}.

Proof: We will prove (14) by induction. For $t = 1$ and all s_1, we have $\pi_1(s_1)Q(f_1, a_1) = p(s_1)p(f_1)\delta_{\{f_1\}}(a_1) = Q(f_1, s_1, a_1)$. Now for $t+1$ and all s_{t+1}, we have

$$\begin{aligned}
&\pi_{t+1}[a^t, b^t](s_{t+1})Q(f^{t+1}, a^{t+1}, b^t) \\
&\quad = \pi_{t+1}[a^t, b^t](s_{t+1}) \sum_{s_t} Q(f^{t+1}, s_t, a^{t+1}, b^t) \\
&\quad \overset{(a)}{=} \left(\sum_{\tilde{s}_t} p(s_{t+1} | \tilde{s}_t, a_t, b_t) r\left(\tilde{s}_t | \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t, b_t\right) \right) \\
&\qquad \times \left(\sum_{s_t} \delta_{\{f_{t+1}(b^t)\}}(a_{t+1}) p(f_{t+1} \mid f^t) p(b_t \mid s_t, a_t) \right. \\
&\qquad\qquad \left. \times \delta_{\{f_t(b^{t-1})\}}(a_t) \pi[a^{t-1}, b^{t-1}](s_t) Q(f^t, a^{t-1}, b^{t-1}) \right) \\
&\quad = \sum_{\tilde{s}_t} p(s_{t+1} \mid \tilde{s}_t, a_t, b_t) r(\tilde{s}_t \mid \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t, b_t) \\
&\qquad \times \sum_{s_t} p(b_t \mid s_t, a_t) \pi_t[a^{t-1}, b^{t-1}](s_t) \delta_{\{f_{t+1}(b^t)\}}(a_{t+1}) \\
&\qquad \times p(f_{t+1} \mid f^t) \delta_{\{f_t(b^{t-1})\}}(a_t) Q(f^t, a^{t-1}, b^{t-1}) \\
&\quad \overset{(b)}{=} \sum_{\tilde{s}_t} p(s_{t+1} \mid \tilde{s}_t, a_t, b_t) p(b_t \mid \tilde{s}_t, a_t) \pi[a^{t-1}, b^{t-1}](\tilde{s}_t) \\
&\qquad \times \delta_{\{f_{t+1}(b^t)\}}(a_{t+1}) p(f_{t+1} \mid f^t) \delta_{\{f_t(b^{t-1})\}} \\
&\qquad \times (a_t) Q(f^t, a^{t-1}, b^{t-1}) \\
&\quad \overset{(c)}{=} \sum_{\tilde{s}_t} \delta_{\{f_{t+1}(b^t)\}}(a_{t+1}) p(s_{t+1} \mid \tilde{s}_t, a_t, b_t) p(f_{t+1} \mid f^t) \\
&\qquad \times p(b_t \mid \tilde{s}_t, a_t) \delta_{\{f_t(b^{t-1})\}}(a_t) Q(f^t, \tilde{s}_t, a^{t-1}, b^{t-1}) \\
&\quad = \sum_{\tilde{s}_t} Q(f^{t+1}, \tilde{s}_t, s_{t+1}, a^{t+1}, b^t) \\
&\quad = Q(f^{t+1}, s_{t+1}, a^{t+1}, b^t)
\end{aligned}$$

where (a) follows from the definition of Π_t and the induction hypothesis. Line (b) follows from Lemma 6.1 and (c) is another application of the induction hypothesis. □

Note that (14) states that the conditional probability $Q(ds_t \mid f^t, a^t, b^{t-1})$ does not depend on f^t almost surely. In addition, the filtering (12) and (13) are defined independently of the code-function distributions (or equivalently, the channel input distributions). This is an example of Witsenhausen's [37] observation that there is policy independence of the filter. Finally, observe that (14) and the fact that Π_t is a function of A^{t-1}, B^{t-1} imply that $S_t - \Pi_t - (F^t, A^t, B^{t-1})$ forms a Markov chain under any consistent measure Q.

B. Markov Channel Coding Theorem

We are now in a position to describe the "$\mathcal{F}^T - \mathcal{B}^T$" channel in terms of the underlying Markov channel. We then prove the Markov channel coding theorem.

Lemma 6.3: We are given a sequence of code-function stochastic kernels $\{p(df_t \mid f^{t-1})\}$, a Markov channel $p(ds_1), \{p(ds_{t+1} \mid s_t, a_t, b_t)\}, \{p(db_t \mid s_t, a_t)\}$, and a consistent joint measure $Q(df^T, ds^T, da^T, db^T)$. Then, for each t and all b_t, we have

$$Q(b_t \mid f^t, a^t, b^{t-1}) = r\left(b_t \mid \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t\right) \tag{15}$$

for Q almost all f^t, a^t, b^{t-1}, where $r(db \mid \pi, a)$ was defined in (11).

Proof: For each t, note that

$$\begin{aligned}
Q(f^t, a^t, b^t) &= \sum_{s_t} Q(f^t, s_t, a^t, b^t) \\
&= \sum_{s_t} p(b_t \mid s_t, a_t) Q(f^t, s_t, a^t, b^{t-1}) \\
&\overset{(a)}{=} \sum_{s_t} p(b_t \mid s_t, a_t) \pi_t[a^{t-1}, b^{t-1}](s_t) Q(f^t, a^t, b^{t-1}) \\
&= r(b_t \mid \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t) Q(f^t, a^t, b^{t-1})
\end{aligned}$$

where (a) follows from Lemma 6.2. □

The previous lemma shows that $B - (\Pi_t, A_t) - (F^t, A^{t-1}, B^{t-1})$ forms a Markov chain under Q.

Corollary 6.1: We are given a sequence of code-function stochastic kernels $\{p(df_t \mid f^{t-1})\}$, a Markov channel $p(ds_1), \{p(ds_{t+1} \mid s_t, a_t, b_t)\}, \{p(db_t \mid s_t, a_t)\}$, and a consistent joint measure $Q(df^T, ds^T, da^T, db^T)$. Then, for each t and all b_t, we have

$$Q(b_t \mid f^t, b^{t-1}) = r(b_t \mid \pi_t[f^{t-1}(b^{t-2}), b^{t-1}](ds_t), f_t(b^{t-1})) \tag{16}$$

for Q almost all f^t, b^{t-1}.

Proof: For each t, note that

$$\begin{aligned} &Q(f^t, b^t) \\ &\quad = \sum_{a^t} Q(f^t, a^t, b^t) \\ &\quad = \sum_{a^t} r(b_t \,|\, \pi_t[a^{t-1}, b^{t-1}](ds_t), a_t) Q(f^t, a^t, b^{t-1}) \\ &\quad = r(b_t \,|\, \pi_t[f^{t-1}(b^{t-2}), b^{t-1}](ds_t), f_t(b^{t-1})) Q(f^t, b^{t-1}) \end{aligned}$$

where the second line follows from Lemma 6.3. □

The corollary shows that we can convert a Markov channel into a channel of the general form considered in Sections III–V. Hence, we can define the operational channel capacity C^{o} for the Markov channel with feedback in exactly the same way we did in Definition 4.3. We can also use the same definitions of capacity C as before. Thus, we can directly apply Theorem 5.1 and its generalization Theorem 5.4 to prove the following.

Theorem 6.1: $C^{\mathrm{o}} = C$ for Markov channels and $C^{\mathrm{o},\Psi} = C^{\Psi}$ for Markov channels with information pattern Ψ.

We end this section by noting that the use of $\{\Pi_t\}$ can simplify the form of the directed information and the choice of the channel input distribution.

Lemma 6.4: For Markov channels, we have $I(F^T; B^T) = I(A^T \to B^T) = \sum_{t=1}^{T} I(A_t, \Pi_t; B_t \,|\, B^{t-1})$.

Proof: The first equality follows from Lemma 5.2. The second equality follows from noting that $I(A^T \to B^T) = \sum_{t=1}^{T} I(A^t; B_t \,|\, B^{t-1})$. For $t = 1$, we know $\Pi_1(ds_1) = p(ds_1)$ is a fixed, nonrandom, measure known to both the transmitter and the receiver. Hence, $I(A_1; B_1) = I(A_1, \Pi_1; B_1)$. For $t > 1$, we have $I(A^t; B_t \,|\, B^{t-1}) = I(A_t, \Pi_t; B_t \,|\, B^{t-1}) + I(A^{t-1}; B_t \,|\, \Pi_t, A_t, B^{t-1}) - I(\Pi_t; B_t \,|\, A^t, B^{t-1})$. Now $I(\Pi_t; B_t \,|\, A^t, B^{t-1}) = 0$ because Π_t is a function of A^{t-1}, B^{t-1}. Lemma 6.3 implies that $(A^{t-1}, B^{t-1}) - (\Pi_t, A_t) - B_t$ is a Markov chain hence $I(A^{t-1}; B_t \,|\, \Pi_t, A_t, B^{t-1}) = 0$. □

We view the pair (A_t, Π_t) as an input to the channel. Intuitively, the encoder needs to send information about its state estimate Π_t so that the decoder can decode the message.

Lemma 6.5: Given a Markov channel $p(ds_1)$, $\{p(ds_{t+1} \,|\, s_t, a_t, b_t)\}, \{p(db_t \,|\, s_t, a_t)\}$, and a channel input distribution $\{q(da_t \,|\, a^{t-1}, b^{t-1})\}$ with resulting joint measure $Q(ds^T, da^T, db^T)$, there exists another channel input distribution of the form $\{r(da_t \,|\, \pi_t, b^{t-1})\}$ with resulting joint measure $R(ds^T, da^T, db^T)$ such that for each t[2]

$$R(d\pi_t, da_t, db^t) = Q(d\pi_t, da_t, db^t)$$

and hence $I_R(A_t, \Pi_t; B_t \,|\, B^{t-1}) = I_Q(A_t, \Pi_t; B_t \,|\, B^{t-1})$.

Proof: From Lemmas 6.2 and 6.3 and (13), we know

$$\begin{aligned} Q(d\pi^T, da^T, db^T) = \bigotimes_{t=1}^{T} & r(db_t \,|\, \pi_t, a_t) \otimes q(da_t \,|\, a^{t-1}, b^{t-1}) \\ & \otimes \delta_{\{\Phi_\Pi(\pi_{t-1}, a_{t-1}, b_{t-1})\}}(d\pi_t). \end{aligned} \tag{17}$$

[2]For any Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$, let $Q(\pi_t \in \Omega, a_t = \bar{a}_t, b^t = \bar{b}^t) = Q\left(\{(a^t, b^t) \colon a_t = \bar{a}_t,\ b^t = \bar{b}^t, \pi_t[a^{t-1}, b^{t-1}] \in \Omega\}\right)$.

By abusing notation, $\delta_{\{\Phi_\Pi(\pi_0, a_0, b_0)\}}(d\pi_1) = \delta_{\{p(ds_1)\}}(d\pi_1)$. For each t, define the stochastic kernel $r(da_t \,|\, \pi_t, b^{t-1})$ to be a version of the conditional distribution $Q(da_t \,|\, \pi_t, b^{t-1})$ (see Theorem A.3 in the Appendix).

Proceed by induction. For $t = 1$, we know $\pi_1(ds_1) = p(ds_1)$. For any Borel measurable set, $\Omega \subset \mathcal{P}(\mathcal{S}), a_1, b_1$

$$\begin{aligned} R(\Omega, a_1, b_1) &= \int_\Omega r(b_1 \,|\, \pi_1, a_1) r(a_1 \,|\, \pi_1) \delta_{\{p(ds_1)\}}(d\pi_1) \\ &= \int_\Omega r(b_1 \,|\, \pi_1, a_1) Q(d\pi_1, a_1) \\ &= Q(\Omega, a_1, b_1). \end{aligned}$$

Now for $t+1$ and any Borel measurable set $\Omega \subset \mathcal{P}(\mathcal{S})$, a_{t+1}, b^{t+1}

$$\begin{aligned} &R(\pi_{t+1} \in \Omega, a_{t+1}, b^{t+1}) \\ &\quad = \sum_{a_t} \int_\Omega \int_{\mathcal{P}(\mathcal{S})} R(d\pi_t, d\pi_{t+1}, a_t, a_{t+1}, b^{t+1}) \\ &\quad = \sum_{a_t} \int_\Omega \int_{\mathcal{P}(\mathcal{S})} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) r(a_{t+1} \,|\, \pi_{t+1}, b^t) \\ &\qquad \times \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) R(d\pi_t, a_t, b^t) \\ &\quad \overset{(a)}{=} \sum_{a_t} \int_\Omega \int_{\mathcal{P}(\mathcal{S})} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) r(a_{t+1} \,|\, \pi_{t+1}, b^t) \\ &\qquad \times \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) Q(d\pi_t, a_t, b^t) \\ &\quad = \sum_{a_t} \int_\Omega \int_{\mathcal{P}(\mathcal{S})} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) r(a_{t+1} \,|\, \pi_{t+1}, b^t) \\ &\qquad \times Q(d\pi_t, d\pi_{t+1}, a_t, b^t) \\ &\quad = \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) r(a_{t+1} \,|\, \pi_{t+1}, b^t) Q(d\pi_{t+1}, b^t) \\ &\quad \overset{(b)}{=} \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) Q(d\pi_{t+1}, a_{t+1}, b^t) \\ &\quad = Q(\pi_{t+1} \in \Omega, a_{t+1}, b^{t+1}) \end{aligned}$$

where (a) follows from the induction hypothesis and (b) follows by the construction of $r(da_{t+1} \,|\, \pi_{t+1}, b^t)$. □

Hence, we can without loss of generality restrict ourselves to channel input distributions of the form $\{q(da_t \,|\, \pi_t, b^{t-1})\}$. Note that the dependence on a^{t-1} appears only through $\pi_t[a^{t-1}, b^{t-1}](ds_t)$. If $\pi_t[a^{t-1}, b^{t-1}]$ is not a function of a^{t-1}, then the distribution of a_t will depend only on the feedback b^{t-1}. We discuss when this happens in Section VIII.

In summary, we have shown that any Markov channel $p(ds_1)$ $\{p(ds_{t+1} \,|\, s_t, a_t, b_t)\}, \{p(db_t \,|\, s_t, a_t)\}$ can be converted into another Markov channel with initial state $\Pi_1(d\pi_1) = \delta_{\{p(ds_1)\}}(d\pi_1)$, deterministic state transitions $\Pi_{t+1} = \Phi_\Pi(\Pi_t, A_t, B_t)$,, and channel output stochastic kernels $\{r(db_t \,|\, \pi_t, a_t)\}$. We call this the *canonical* Markov channel associated with the original Markov channel. Thus, the problem of determining the capacity of a Markov channel with state space $\mathcal{S}$ has been reduced to determining the capacity of the canonical Markov channel. This latter Markov channel has state space $\mathcal{P}(\mathcal{S})$ and state computable from the channel inputs and outputs.

Note that even if the original Markov channel does not have ISI, it is typically the case that the canonical Markov channel $\Pi_{t+1} = \Phi_\Pi(\Pi_t, A_t, B_t)$ will have ISI. This is because the choice of channel input can help the decoder identify the channel. This property is called *dual control* in the stochastic control literature [2].

VII. The MDP Formulation

Our goal in this section is to formulate the following optimization problem for Markov channels with feedback as an infinite horizon average cost problem.

Problem A

$$\sup_{\mathcal{D}_\infty} \liminf_{T\to\infty} \frac{1}{T} I(A^T \to B^T). \tag{18}$$

By Lemma 6.4, we have $\sup_{\mathcal{D}_\infty} \liminf_{T\to\infty} \frac{1}{T} I(A^T \to B^T) = \sup_{\mathcal{D}_\infty} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(A_t, \Pi_t; B_t \,|\, B^{t-1})$. Before proceeding, the reader may notice that the optimization in Problem A is different than the one given in (4): $C = \sup_{\{\mathcal{D}_T\}_{T=1}^\infty} \underline{I}(A \to B)$. In the course of this section, it will be shown that the optimization in (4) is equivalent to Problem A. That one can without loss of generality restrict the optimization to D_∞ instead of $\{\mathcal{D}_T\}_{T=1}^\infty$ shown to be a consequence of Bellman's *principle of optimality*. In addition, conditions will be given such that under the optimal channel input distribution, we have $\liminf_{T\to\infty} \frac{1}{T} I(A^T \to B^T) = \underline{I}(A \to B)$.

To compute $I(A_t, \Pi_t; B_t \,|\, B^{t-1})$, we need to know the measure

$$\begin{aligned} Q(d\pi_t, da_t, db^t) = r(db_t \,|\, \pi_t, a_t) \otimes q(da_t \,|\, \pi_t, b^{t-1}) \\ \otimes Q(d\pi_t, db^{t-1}). \end{aligned} \tag{19}$$

By Lemma 6.5, we can restrict ourselves to channel input distributions of the form $\{q(da_t \,|\, \pi_t, b^{t-1})\}$.

To formulate the optimization in Problem A as a stochastic control problem, we need to specify the state space, the control actions, and the running cost. On the first glance, it may appear that the encoder should choose control actions of the form $u_t(da_t)$ based on the information $(\pi_t[a^{t-1}, b^{t-1}], b^{t-1})$. Unfortunately, one cannot write the running cost in terms of $u_t(da_t)$. To see this, observe that the argument under the expectation in $I(A_t, \Pi_t; B_t \,|\, B^{t-1}) = E[\log \frac{r(B_t \,|\, \Pi_t, A_t)}{Q(B_t \,|\, B^{t-1})}]$ can be written as

$$\log \frac{r(b_t \,|\, a_t, \pi_t)}{Q(b_t \,|\, b^{t-1})} = \log \frac{r(b_t \,|\, a_t, \pi_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) Q(d\tilde{\pi}_t, d\tilde{a}_t \,|\, b^{t-1})} \tag{20}$$

for Q almost all π_t, a_t, b^t. This depends on $Q(d\pi_t, da_t \,|\, b^{t-1})$ and not $Q(da_t \,|\, \pi_t, b^{t-1})$.

This suggests that the control actions should be stochastic kernels of the form $u_t(da_t \,|\, \pi_t)$. In many cases, the space that these kernels live in has a natural parameterization. For example, Yang *et al.* [39] present a natural parametrization for a class of finite state, Markov channels with state computable at the transmitter. As an another example, for Gaussian channels, it is known that the optimal input distribution is linear and can be parameterized by its coefficients [8], [31], [40]. In this paper, we will choose control actions of the form $u_t(d\pi_t, da_t)$. This is consistent with our view that the pair (A_t, Π_t) is an input to the channel. Of course, there are restrictions on the marginal of Π_t. The next section formalizes the stochastic control problem with this choice of control action.

A. Partially Observed Markov Decision Problem

Here we first describe the components of the partially observed Markov decision problem (POMDP) formulation. In the next section, we show the equivalence of this POMDP formulation to the optimization in Problem A.

Consider the control action $u(d\pi, da)$ in the control space $\mathcal{U} = \mathcal{P}(\mathcal{P}(\mathcal{S}) \times \mathcal{A})$. The space $\mathcal{U}$ is a Polish space (i.e., a complete, separable metric space) equipped with the topology of weak convergence.

The state at time $t > 1$ is $X_t = (\Pi_{t-1}, A_{t-1}, B_{t-1}) \in \mathcal{P}(\mathcal{S}) \times \mathcal{A} \times \mathcal{B}$ and $X_1 = \emptyset$. The dynamics are given as

$$r(dx_{t+1} \,|\, x_t, u_t) = r(db_t \,|\, \pi_t, a_t) \otimes u_t(d\pi_t, da_t). \tag{21}$$

Note that the dynamics depends on u_t but not on x_t. The observation at time $t > 1$ is given by $Y_t = B_{t-1}$ and $Y_1 = \emptyset$. Note that Y_t is a deterministic function of X_t.

As discussed, one of the main difficulties in formulating (18) as a POMDP has to do with the form of the cost (20). The cost at time t is given as shown in (22) at the bottom of the page. Note that the cost is just a function of u_t, x_{t+1}.

The information pattern at the controller at time t is $(Y^t, U^{t-1}) = (B^{t-1}, U^{t-1}) \in \mathcal{B}^{t-1} \times \mathcal{U}^{t-1}$. The policy at time t is a stochastic kernel $\mu_t(du_t \,|\, b^{t-1}, u^{t-1})$ from $\mathcal{B}^{t-1} \times \mathcal{U}^{t-1}$ to $\mathcal{U}$. A policy $\{\mu_t\}$ is said to be a *deterministic* policy if for each t and all (b^{t-1}, u^{t-1}), the stochastic kernel $\mu_t(du_t \,|\, b^{t-1}, u^{t-1})$ assigns mass one to only one point in $\mathcal{U}$. In this case, we will abuse notation and write $u_t = \mu_t[b^{t-1}]$. Technically, we should explicitly include $p(ds_1)$ and the other channel parameters in the information pattern. But because the channel parameters are fixed throughout and to simplify notation, we will not explicitly mention the control policy's dependence on them.

The time order of events is the usual one for POMDPs: $X_1, Y_1, U_1, X_2, Y_2, U_2 \ldots$. For a given policy $\{\mu_t\}$, the resulting joint measure is

$$\begin{aligned} &R^\mu(dx^T, dy^T, du^T) \\ &\quad = \bigotimes_{t=1}^{T} \mu_t(du_t \,|\, y^t, u^{t-1}) \otimes r(dy_t \,|\, x_t) \otimes r(dx_t \,|\, x^{t-1}, u^{t-1}) \end{aligned}$$

$$c(x_t, u_t, x_{t+1}) = \begin{cases} \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t)}, & \text{if } \int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t) > 0 \\ 0, & \text{else.} \end{cases} \tag{22}$$

where $r(dy_t \,|\, x_t)$ corresponds to the functional relationship between Y_t and X_t. In terms of the original channel variables, this can be written as

$$R^{\mu}(du^T, d\pi^T, da^T, db^T) = \bigotimes_{t=1}^{T} r(db_t \,|\, \pi_t, a_t) \otimes u_t(d\pi_t, da_t) \otimes \mu_t(du_t \,|\, u^{t-1}, b^{t-1}) \quad (23)$$

where we have used (21).

Note that this R^{μ} measure is not the same as the measure Q used in (17) of Lemma 6.5. Compare the differences between the R^{μ} and Q. In particular, notice that under the Q measure $\{\Pi_t\}$ is determined by the function Φ_{Π} given in (13), whereas under the R^{μ} measure $\{\Pi_t\}$ is determined by the choice of policy $\{\mu_t\}$. The next two sections discuss the relation between these two different measures.

B. The Sufficient Statistic $\{\Gamma_t\}$ and the Control Constraints

As described above, $\{\Pi_t\}$ is defined differently under the measure Q given in (17) and under R^{μ} defined in (23). We need to ensure that the $\{\Pi_t\}$ play similar roles in both cases. To this end, we will next define appropriate control constraints.

Equation (21) states $r(d\pi, da, db \,|\, u) = r(db \,|\, \pi, a) \otimes u(d\pi, da)$. The following lemma follows from Theorem A.3 in the Appendix.

Lemma 7.1: There exists a stochastic kernel $r(d\pi, da \,|\, u, b)$ from $\mathcal{U} \times \mathcal{B}$ to $\mathcal{P}(\mathcal{S}) \times \mathcal{A}$ such that $r(d\pi, da, db \,|\, u) = r(d\pi, da \,|\, u, b) \otimes r(db \,|\, u)$ where $r(db \,|\, u)$ is the marginal of $r(d\pi, da, db \,|\, u)$.

We now define the statistics $\{\Gamma_t\} \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$, where $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ is the space of probability measures on probability measures on $\mathcal{S}$. Specifically, $\Gamma_t : \mathcal{U}^{t-1} \times \mathcal{B}^{t-1} \to \mathcal{P}(\mathcal{P}(\mathcal{S}))$ is defined as follows. For $t = 1$, let

$$\gamma_1(d\pi_1) = \delta_{\{p(ds_1)\}}(d\pi_1) \quad (24)$$

and for $t > 1$ and each u^{t-1}, b^{t-1} and all Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$, let

$$\gamma_t[u^{t-1}, b^{t-1}](\Omega) = \int\int \{\Phi_{\Pi}(\pi_{t-1}, a_{t-1}, b_{t-1}) \in \Omega\} r \times (d\pi_{t-1}, da_{t-1} \,|\, u_{t-1}, b_{t-1}). \quad (25)$$

Here $\{\cdot\}$ corresponds to the indicator function. Note that for $t > 1, \gamma_t[u^{t-1}, b^{t-1}](d\pi_t)$ depends only on u_{t-1}, b_{t-1}. Also $p(ds_1)$ is fixed. Thus, we will abuse notation and just write $\gamma_t[u_{t-1}, b_{t-1}](d\pi_t)$ for all $t \geq 1$.

Equation (25) implies there exists a deterministic, stationary, measurable function Φ_{Γ} such that $\gamma_{t+1} = \Phi_{\Gamma}(u_t, b_t)$ for all $t = 1, \ldots, T-1$. Note that because of feedback the statistic Γ_t can be computed at both the transmitter and the receiver. It can be viewed as the receiver's estimate of the transmitter's estimate of the state of the channel.

We now define the control constraints. Let

$$\mathcal{U}(\gamma) = \{u(d\pi, da) : u(d\pi, da) \in \mathcal{U}, u(d\pi) = \gamma(d\pi)\}. \quad (26)$$

Note that for each $\gamma \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ the set $\mathcal{U}(\gamma)$ is compact. To see this, note that $\mathcal{U}$ is compact and the constraint defining $\mathcal{U}(\gamma)$ is linear.

For each t and (u^{t-1}, b^{t-1}), the control constraint $\mathcal{U}_t(\,\cdot\,) \subset \mathcal{U}$ is defined as

$$\mathcal{U}_t(u^{t-1}, b^{t-1}) = \mathcal{U}(\gamma_t[u_{t-1}, b_{t-1}]). \quad (27)$$

For each t, the policy μ_t will enforce the control constraint. Specifically, for all (u^{t-1}, b^{t-1})

$$\mu_t(\{u_t \in \mathcal{U}_t(\gamma_t[u_{t-1}, b_{t-1}])\} \,|\, u^{t-1}, b^{t-1}) = 1. \quad (28)$$

The next lemma shows that the $\{\Gamma_t\}$ are consistent with the conditional probabilities $R^{\mu}(d\pi_t \,|\, u^{t-1}, b^{t-1})$.

Lemma 7.2: We are given $p(ds_1)$, the dynamics (21), and a policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting measure $R^{\mu}(du^T, d\pi^T, da^T, db^T)$. Then, for each t, we have

$$R^{\mu}(d\pi_t \,|\, u^{t-1}, b^{t-1}) = \gamma_t[u_{t-1}, b_{t-1}](d\pi_t) \quad (29)$$

for R^{μ} almost all u^{t-1}, b^{t-1}.

Proof: Fix a Borel measurable set $\Omega \subset \mathcal{P}(\mathcal{S})$. For any t, any Borel measurable sets $\Theta_k \subset \mathcal{U}, k = 1, \ldots, t-1$ and any b^{t-1}, we have

$$\begin{aligned} &R^{\mu}(\Omega, \Theta^{t-1}, b^{t-1}) \\ &\quad = \int_{\mathcal{U}, \Theta^{t-1}, \Omega} R^{\mu}(d\pi_t, du^{t-1}, du_t, b^{t-1}) \\ &\quad = \int_{\mathcal{U}, \Theta^{t-1}} u_t(\Omega, \mathcal{A}) \mu_t(du_t \,|\, u^{t-1}, b^{t-1}) R^{\mu}(du^{t-1}, b^{t-1}) \\ &\quad = \int_{\Theta^{t-1}} \left(\int_{\mathcal{U}} u_t(\Omega, \mathcal{A}) \mu_t(du_t | u^{t-1}, b^{t-1}) \right) R^{\mu}(du^{t-1}, b^{t-1}) \\ &\quad = \int_{\Theta^{t-1}} \gamma_t[u_{t-1}, b_{t-1}](\Omega) R^{\mu}(du^{t-1}, b^{t-1}) \end{aligned}$$

where the last equality follows because the control policy μ_t satisfies the control constraint given in (28). □

Equations (29) and (25) show that the conditional probability $R^{\mu}(d\pi_t \,|\, u^{t-1}, b^{t-1})$ does not depend on the policy μ and u^{t-2}, b^{t-2} almost surely. See comments after Lemma 6.2.

We can simplify the form of the cost, in the standard way, by computing the expectation over the next state. For each t, define

$$\begin{aligned} \bar{c}(u_t) &= [c(X_t, U_t, X_{t+1}) \,|\, u_t] \\ &= \int r(db_t \,|\, \pi_t, a_t) u_t(d\pi_t, da_t) \\ &\quad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t)} \end{aligned} \quad (30)$$

which follows from (22) and the fact that c_t does not depend on x_t.

In summary, we have formulated an average cost, infinite horizon, POMDP.

Problem B

$$\sup_{\{\mu_t\}} \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E_{R^\mu}[\overline{c}(U_t)]. \tag{31}$$

Here the dynamics are given by (21) and the costs are given by (30). The supremization is over all policies that satisfy the control constraint (28).

C. Equivalence of Problems A and B

We now show the equivalence of the optimization problems posed in Problem A and Problem B. As discussed at the end of Section VII-A, the measures Q and R^μ can be different. By equivalence, we mean that for any choice of channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ with resulting joint measure $Q(ds^T, da^T, db^T)$, we can find a control policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting joint measure $R^\mu(du^T, d\pi^T, da^T, db^T)$ such that for each t

$$Q(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t). \tag{32}$$

Vice versa, given any policy $\{\mu_t\}$ satisfying the control constraint (28), we can find a channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ such that the above marginals are equal. This equivalence will imply that the optimal costs for the two problems are the same and the optimal channel input distribution for (18) is related to the optimal policy for (31).

Lemma 7.3: For every channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ with resulting joint measure $Q(ds^T, da^T, db^T)$, there exists a deterministic policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting joint measure $R^\mu(du^T, d\pi^T, da^T, db^T)$ such that $R^\mu(d\pi_t, da_t, db^t) = Q(d\pi_t, da_t, db^t)$ for each t.

Proof: For each t, choose a *deterministic policy* that satisfies

$$\mu_t[b^{t-1}](d\pi_t, da_t) = Q(d\pi_t, da_t \,|\, b^{t-1})$$

for Q almost all b^{t-1}. Proceed by induction. For $t = 1$, we have $R^\mu(d\pi_1, da_1, db_1) = r(db_1 \,|\, \pi_1, a_1) \otimes \mu_1(d\pi_1, da_1) = r(db_1 \,|\, \pi_1, a_1) \otimes Q(d\pi_1, da_1) = Q(d\pi_1, da_1, db_1)$. For $t+1$, we have for any Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$ and all a_{t+1}, b^{t+1}

$$\begin{aligned}
&R^\mu(\Omega, a_{t+1}, b^{t+1}) \\
&\quad = \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) \mu_{t+1}[b^t](d\pi_{t+1}, a_{t+1})\, R^\mu(b^t) \\
&\quad \overset{(a)}{=} \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) Q(d\pi_{t+1}, a_{t+1} \,|\, b^t)\, Q(b^t) \\
&\quad = Q(\Omega, a_{t+1}, b^{t+1}).
\end{aligned}$$

Line (a) follows by the induction hypothesis and our choice of μ_{t+1}.

We show that the policy $\{\mu_t\}$ satisfies the constraint (28). For $t = 1$, we have $\mu_1(d\pi_1) = Q(d\pi_1) = \gamma_1(d\pi_1)$. For $t > 1$, we have for any Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$ and all b^{t-1}

$$\begin{aligned}
&\mu_t[b^{t-1}](\Omega) R^\mu(b^{t-1}) \\
&\quad \overset{(a)}{=} Q(\Omega, b^{t-1}) \\
&\quad \overset{(b)}{=} \int\int \{\Phi_\Pi(\pi_{t-1}, a_{t-1}, b_{t-1}) \in \Omega\} r(b_{t-1} \,|\, \pi_{t-1}, a_{t-1}) \\
&\qquad \times \mu_t[b^{t-2}](d\pi_{t-1}, da_{t-1}) Q(b^{t-2}) \\
&\quad \overset{(c)}{=} \int\int \{\Phi_\Pi(\pi_{t-1}, a_{t-1}, b_{t-1}) \in \Omega\} \\
&\qquad \times r(d\pi_{t-1}, da_{t-1} \,|\, \mu_t[b^{t-2}], b_{t-1}) r \\
&\qquad \times (b_{t-1} \,|\, \mu_t[b^{t-2}]) R^\mu(b^{t-2}) \\
&\quad \overset{(d)}{=} \gamma_t[\mu_{t-1}[b^{t-2}], b_{t-1}](\Omega) R^\mu(b^{t-1})
\end{aligned}$$

where (a) follows from the first part and the choice of control, (b) follows from our choice of control, (c) follows from the first part and Lemma 7.1, and (d) follows from (25). Finally, altering μ_t on a set of measure zero if necessary, we can ensure that for each t the deterministic policy μ_t will enforce the control constraint. Specifically, for each b^{t-1}, we have $\mu_t[b^{t-1}] \in \mathcal{U}(\gamma_t[\mu_{t-1}[b^{t-2}], b_{t-1}])$. □

Lemma 7.4: For every policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting joint measure $R^\mu(du^T, d\pi^T, da^T, dbT)$, there exists a channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ with resulting joint measure $Q(ds^T, da^T, db^T)$ such that $Q(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$ for each t.

Proof: For each t, pick a channel input distribution such that

$$q(da_t \,|\, \pi_t, b^{t-1}) = R^\mu(da_t \,|\, \pi_t, b^{t-1})$$

for R^μ almost all π_t, b^{t-1}. We proceed by induction. For $t = 1$, we have $Q(d\pi_1, da_1, db_1) = r(db_1 \,|\, \pi_1, a_1) \otimes q(da_1 \,|\, \pi_1) \otimes \delta_{\{p(ds_1)\}}(d\pi_1) = R^\mu(d\pi_1, da_1, db_1)$. For $t+1$, we have for any Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$ and all a_{t+1}, b^{t+1}

$$\begin{aligned}
&Q(\Omega, a_{t+1}, b^{t+1}) \\
&\quad = \int_{\mathcal{A},\mathcal{P}(\mathcal{S}),\Omega} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) q(a_{t+1} \,|\, \pi_{t+1}, b^t) \\
&\qquad \times \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) Q(d\pi_t, da_t, b^t) \\
&\quad \overset{(a)}{=} \int_{\mathcal{A},\mathcal{P}(\mathcal{S}),\Omega} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) q(a_{t+1} \,|\, \pi_{t+1}, b^t) \\
&\qquad \times \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) R^\mu(d\pi_t, da_t, b^t) \\
&\quad = \int_{\mathcal{A},\mathcal{P}(\mathcal{S}),\Omega} r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) q(a_{t+1} \,|\, \pi_{t+1}, b^t) \\
&\qquad \times \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) \\
&\qquad \times \left(\int_{\mathcal{U}} r(b_t \,|\, \pi_t, a_t) u_t(d\pi_t, da_t) R^\mu(du_t, b^{t-1}) \right) \\
&\quad \overset{(b)}{=} \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1})\, q(a_{t+1} \,|\, \pi_{t+1}, b^t) \\
&\qquad \times \int_{\mathcal{U},\mathcal{A},\mathcal{P}(\mathcal{S})} \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) r(d\pi_t, da_t \,|\, u_t, b_t) \\
&\qquad \times r(b_t \,|\, u_t) R^\mu(du_t, b^{t-1}) \\
&\quad \overset{(c)}{=} \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) q(a_{t+1} \,|\, \pi_{t+1}, b^t) \\
&\qquad \times \int_{\mathcal{U}} \gamma_{t+1}[u_t, b_t](d\pi_{t+1}) R^\mu(du_t, b^t) \\
&\quad \overset{(d)}{=} \int_\Omega r(b_{t+1} \,|\, \pi_{t+1}, a_{t+1}) q(a_{t+1} \,|\, \pi_{t+1}, b^t) R^\mu(d\pi_{t+1}, b^t) \\
&\quad \overset{(e)}{=} R^\mu(\Omega, a_{t+1}, b^{t+1})
\end{aligned}$$

where (a) follows from the induction hypothesis, (b) follows from Lemma 7.1, (c) follows from (25), (d) follows from Lemma 7.2, and (e) follows from the choice of channel input distribution. □

The next lemma shows that the optimal policies for problem B can be restricted to deterministic policies.

Lemma 7.5: For every policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting joint measure R^μ, there exists a deterministic policy $\{\bar{\mu}_t\}$ satisfying the control constraint (28) with resulting joint measure $R^{\bar{\mu}}$ such that for each $tR^{\bar{\mu}}(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$ and $E_{R^\mu}[\bar{c}(U_t)] \leq E_{R^{\bar{\mu}}}[\bar{c}(U_t)]$.

Proof: Fix $\{\mu_t\}$. By Lemma 7.4, we know there is a channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ such that for each t, $Q(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$. By Lemma 7.3, we know there is a deterministic policy $\{\bar{\mu}_t\}$ such that for each t, $R^{\bar{\mu}}(d\pi_t, da_t, db^t) = Q(d\pi_t, da_t, db^t)$. Hence, for this $\{\bar{\mu}_t\}$, we have $R^{\bar{\mu}}(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$.

For each t, any Borel measurable $\Omega \in \mathcal{P}(\mathcal{S})$, and $\forall\, a_t, b^{t-1}$

$$\begin{aligned}
&\bar{\mu}[b^{t-1}](\Omega, a_t)\, R^\mu(b^{t-1}) \\
&\quad = \bar{\mu}[b^{t-1}](\Omega, a_t) R^{\bar{\mu}}(b^{t-1}) \\
&\quad = R^{\bar{\mu}}(\Omega, a_t, b^{t-1}) \\
&\quad = R^\mu(\Omega, a_t, b^{t-1}) \\
&\quad = \int_{\mathcal{U}} u_t(\Omega, a_t)\, R^\mu(du_t, b^{t-1}).
\end{aligned}$$

Hence, $\bar{\mu}[b^{t-1}](d\pi_t, da_t) = \int_{\mathcal{U}} u_t(d\pi_t, da_t) R^\mu(du_t \,|\, b^{t-1})$ for R^μ almost all b^{t-1}.

Now from (30) and for each t

$$\begin{aligned}
&E_{R^\mu}\,[\bar{c}(U_t)] \\
&\quad = \int_{\mathcal{U}\times\mathcal{B}^{t-1}} R^\mu(du_t, db^{t-1}) \int_{\mathcal{B},\mathcal{A},\mathcal{P}(\mathcal{S})} r(db_t \,|\, \pi_t, a_t) \\
&\qquad \times u_t(d\pi_t, da_t) \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int_{\mathcal{A},\mathcal{P}(\mathcal{S})} r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t)} \\
&\quad \overset{(a)}{\leq} \int_{\mathcal{B}^{t-1}} R^\mu(db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) \\
&\qquad \times \left(\int_{\mathcal{U}} u_t(d\pi_t, da_t) R^\mu(du_t \,|\, b^{t-1}) \right) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) \left(\int_{\mathcal{U}} u(d\tilde{\pi}_t, d\tilde{a}_t) R^\mu(du_t \,|\, b^{t-1}) \right)} \\
&\quad \overset{(b)}{=} \int_{\mathcal{B}^{t-1}} R^\mu(db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) \bar{\mu}[b^{t-1}](d\pi_t, da_t) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) \bar{\mu}_t[b^{t-1}](d\tilde{\pi}_t, d\tilde{a}_t)} \\
&\quad \overset{(c)}{=} \int R^{\bar{\mu}}(du_t, db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) u_t(d\pi_t, da_t) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t)} \\
&\quad = E_{R^{\bar{\mu}}}[\bar{c}(U_t)]
\end{aligned}$$

where (a) follows from the concavity of $-u \log u$ and the conditional Jensen's inequality, (b) follows from above, and (c) follows because $R^\mu(db^{t-1}) = R^{\bar{\mu}}(db^{t-1})$ and $\bar{\mu}_t$ is a deterministic policy. □

Theorem 7.1: Problems A and B have equal optimal costs.

Proof: For any deterministic policy $\{\mu_t\}$ satisfying the control constraint (28) with resulting joint measure R^μ and, as given in Lemma 7.4, an associated channel input distribution $\{q(da_t \,|\, \pi_t, b^{t-1})\}$ with associated joint measure Q, the following holds for each t:

$$\begin{aligned}
&E_{R^\mu}[\bar{c}(U_t)] \\
&\quad = \int_{\mathcal{U}\times\mathcal{B}^{t-1}} R^\mu(du_t, db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) u_t(d\pi_t, da_t) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) u_t(d\tilde{\pi}_t, d\tilde{a}_t)} \\
&\quad \overset{(a)}{=} \int_{\mathcal{B}^{t-1}} R^\mu(db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) \mu_t[b^{t-1}](d\pi_t, da_t) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) \mu_t[b^{t-1}](d\tilde{\pi}_t, d\tilde{a}_t)} \\
&\quad \overset{(b)}{=} \int_{\mathcal{B}^{t-1}} R^\mu(db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) R^\mu(d\pi_t, da_t \,|\, b^{t-1}) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) R^\mu(d\tilde{\pi}_t, d\tilde{a}_t \,|\, b^{t-1})} \\
&\quad \overset{(c)}{=} \int_{\mathcal{B}^{t-1}} Q(db^{t-1}) \int r(db_t \,|\, \pi_t, a_t) Q(d\pi_t, da_t \,|\, b^{t-1}) \\
&\qquad \times \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) Q(d\tilde{\pi}_t, d\tilde{a}_t \,|\, b^{t-1})} \\
&\quad = I_Q(A_t, \Pi_t; B_t \,|\, B^{t-1}) \\
&\quad \overset{(d)}{=} I_{R^\mu}(A_t, \Pi_t; B_t \,|\, B^{t-1})
\end{aligned}$$

where (a) and (b) follow because μ_t is a deterministic policy, and hence, $R^\mu(d\pi_t, da_t \,|\, b^{t-1}) = \mu_t[b^{t-1}](d\pi_t, da_t)$. Lines (c) and (d) follow because $Q(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$. The theorem then follows from this observation and Lemmas 7.3–7.5. □

D. Fully Observed Markov Decision Problem

In this section, we make one final simplification. We will convert the POMDP in Problem B into a fully observed MDP on a suitably defined state space.

Note that the cost $\bar{c}(u)$ given in (30) at time t only depends on u_t. The control constraints $U(\gamma)$ given in (26) at time t only depends on γ_t. The statistics $\gamma_t[u^{t-1}, b^{t-1}]$ only depend on $p(ds_1)$ in the case $t = 1$ and only depends on u_{t-1}, b_{t-1} in the case $t > 1$.

This suggests that $\Gamma_t \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ could be a suitable fully observed state. The dynamics are given as $\gamma_1(d\pi_1) = \delta_{\{p(ds_1)\}}(d\pi_1)$ and for $t > 1$ that

$$\begin{aligned}
&r(d\gamma_{t+1} \,|\, \gamma_t, u_t) \\
&\quad = \int_{\mathcal{P}(\mathcal{S}),\mathcal{A},\mathcal{B}} \delta_{\{\Phi_\Gamma(u_t, b_t)\}}(d\gamma_{t+1})\, r(db_t \,|\, \pi_t, a_t)\, u_t(d\pi_t, da_t)
\end{aligned} \tag{33}$$

Lemma 7.6: For every policy $\{\mu_t\}$ satisfying (28) with resulting joint measure R^μ, we have for each $t > 1$

$$R^\mu(d\gamma_t \,|\, \gamma_{t-1}, u_{t-1}) = r(d\gamma_t \,|\, \gamma_{t-1}, u_{t-1}) \tag{34}$$

for R^μ almost all γ_{t-1}, u_{t-1}.

Proof: For each $t > 1$ and for any Borel measurable sets $\Omega_{t-1}, \Omega_t \subset \mathcal{P}(\mathcal{P}(\mathcal{S}))$ and any Borel measurable set $\Theta \subset \mathcal{U}$, we have

$$\begin{aligned} &R^\mu(\Omega_t, \Omega_{t-1}, \Theta) \\ &\quad = \int_{\Theta,\Omega_{t-1},\mathcal{B},\mathcal{A},\mathcal{P}(\mathcal{S})} R^\mu(\Omega_t, d\gamma_{t-1}, \\ &\qquad du_{t-1}, d\pi_{t-1}, da_{t-1}, db_{t-1}) \\ &\quad = \int \{\Phi_\Gamma(u_{t-1}, b_{t-1}) \in \Omega_t\} r(db_{t-1} \,|\, \pi_{t-1}, a_{t-1}) \\ &\qquad \times u_{t-1}(d\pi_{t-1}, da_{t-1}) R^\mu(d\gamma_{t-1}, du_{t-1}) \\ &\quad = \int_{\Theta,\Omega_{t-1}} r(\Omega_t \,|\, \gamma_{t-1}, u_{t-1}) R^\mu(d\gamma_{t-1}, du_{t-1}) \end{aligned}$$

where the last line follows from (33). □

Note that the dynamics given in (33), $r(d\gamma_{t+1} \,|\, \gamma_t, u_t)$ depends only on u_t. This along with the fact that the cost at time t only depends on u_t and the control constraint at time t only depends on γ_t suggests that we can simplify the form of the control policy from $\mu_t(du_t \,|\, u^{t-1}, b^{t-1})$ to $\mu_t(du_t \,|\, \gamma_t)$.

Theorem 7.2: Without loss of generality, the optimization given in Problem B can be modeled as a fully observed MDP with:

1) state space $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ and dynamics given by (33);
2) compact control constraints $U(\gamma)$ given by (26);
3) running cost $\bar{c}(u)$ given by (30).

Proof: See Section 10.2 in [3], in particular, Proposition 10.5. □

Lemmas 7.3–7.5 and Theorem 7.1 show that for any deterministic policy $\{\mu_t[b^{t-1}]\}$ with resulting joint measure R^μ, there is a corresponding channel input distribution $q(da_t \,|\, \pi_t, b^{t-1})$ with resulting joint measure Q such that for all t, $Q(d\pi_t, da_t, db^t) = R^\mu(d\pi_t, da_t, db^t)$. By Theorem 7.2, we know we can, without loss of generality, restrict ourselves to deterministic policies of the form: $\{\mu_t[\gamma_t]\}$. Under such a policy, we have

$$Q(da_t \,|\, \pi_t, b^{t-1}) = R^\mu(da_t \,|\, \pi_t, b^{t-1}) = R^\mu(da_t \,|\, \pi_t, \gamma_t)$$

for R^μ almost surely all π_t, b^{t-1}, γ_t. For a fixed deterministic policy, we can view γ_t as a function of b^{t-1}. Thus, the optimal channel input distribution takes the form $\{q(da_t \,|\, \pi_t, \gamma_t)\}$ and

$$\bar{c}(\mu_t[\gamma_t]) = I_Q(A_t, \Pi_t; B_t \,|\, \gamma_t) \quad R^\mu - \text{almost all } \gamma_t. \tag{35}$$

Recall that in (18) of Problem A, we started with terms of the form $I(A^t; B_t \,|\, B^{t-1})$. We have now simplified it to terms of the form $I(A_t, \Pi_t; B_t \,|\, \Gamma_t)$. This is a significant simplification because the size of $\mathcal{P}(\mathcal{P}(\mathcal{S})) \times \mathcal{P}(\mathcal{S}) \times \mathcal{A} \times \mathcal{B}$ is not growing in time whereas the size of $\mathcal{A}^t \times \mathcal{B}^t$ is growing in time. In review, Π_t can be viewed as the encoder's estimate of the state and Γ_t can be viewed as the decoder's estimate of the encoder's estimate of the state. In addition, Γ_t is known to the encoder.

E. ACOE and Information Stability

We present the ACOE for the fully observed MDP corresponding to the equivalent optimizations in Problems A and B. We then show that the process is information stable under the optimal input distribution. Finally, we relate the equivalent optimizations in (18) and (31) to the optimization given in (4): $\sup_{\{\mathcal{D}_T\}_{T=1}^\infty} \underline{I}(A \to B)$.

The following technical lemma is required to ensure the existence of a *measurable selector* in the ACOE given in (36). The proof is straightforward but tedious and can be found in the Appendix.

Lemma 7.7: For $|\mathcal{B}|$ finite, we have:

1) the cost is bounded and continuous; specifically, $0 \le \bar{c}(u) \le \log|\mathcal{B}|, \forall u \in \mathcal{U}$;
2) the control constraint function $\mathcal{U}(\gamma)$ is a continuous set-valued map between $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ and $\mathcal{U}$;
3) the dynamics $r(d\gamma_{t+1} \,|\, \gamma_t, u_t)$ are continuous.

We now present the average cost verification theorem.

Theorem 7.3: If there exists a $V^* \in \mathbb{R}$, a bounded function $w : \gamma \mapsto w(\gamma) \in \mathbb{R}$, and a policy μ^* achieving the supremum for each γ in the following ACOE:

$$V^* + w(\gamma) = \sup_{u \in \mathcal{U}(\gamma)} \left(\bar{c}(u) + \int w(\tilde{\gamma}) r(d\tilde{\gamma} \,|\, \gamma, u) \right) \tag{36}$$

then:

1) V^* is the optimal value of the optimization in Problem B; the optimal policy is the stationary, deterministic policy given by μ^*;
2) under this μ^*, we have

$$\begin{aligned} V^* &= \liminf_{T\to\infty} \frac{1}{T} E_{R^{\mu^*}} \left[\sum_{t=1}^T \bar{c}(U_t) \right] \\ &= \limsup_{T\to\infty} \frac{1}{T} E_{R^{\mu^*}} \left[\sum_{t=1}^T \bar{c}(U_t) \right] \end{aligned}$$

and

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T \bar{c}(U_t) = V^* \quad R^{\mu^*} - \text{a.s.}$$

Proof: See Lemma 7.7 and Theorems 6.2 and 6.3 of [1]. □

There exist many sufficient conditions for the existence of a solution. See [1] and [18] for a representative sample. Most of these conditions require the process be recurrent under the optimal policy. The following theorem describes one such sufficient condition.

Theorem 7.4: If there exists an $\alpha < 1$ such that

$$\begin{aligned} \sup_{\gamma_t, \tilde{\gamma}_t, u_t \in \mathcal{U}(\gamma_t), \tilde{u}_t \in \mathcal{U}(\tilde{\gamma}_t)} &\| r(d\gamma_{t+1} \,|\, \gamma_t, u_t) \\ &- r(d\gamma_{t+1} \,|\, \tilde{\gamma}_t, \tilde{u}_t) \|_{TV} \le \alpha \end{aligned} \tag{37}$$

then the ACOE (36) has a bounded solution. Here $\|\cdot\|_{TV}$ denotes the total variation norm.

Proof: See Corollary 6.1 of [1]. □

Condition (37) insures that for any stationary policy there exists a stationary distribution. Specifically, the following holds.

Proposition 7.1: If (37) holds then for all stationary policies of the form $\mu : \gamma \to u(d\pi, da)$ that satisfy the control constraint (26) there exists a probably measure λ_μ on $\mathcal{P}(\mathcal{S})$ such that for any $\epsilon > 0$ there exists a T large enough such that $\forall t > T$

$$\|r_\mu^t(d\gamma_t \,|\, \gamma_1) - \lambda_\mu(d\gamma_t)\|_{TV} \leq \epsilon \tag{38}$$

where $r_\mu^t(d\gamma_t \,|\, \gamma_1)$ is the t-step transition stochastic kernel under the stationary policy μ. Furthermore, $\lim_{T\to\infty} \frac{1}{T} E_{R^\mu}[\sum_{t=1}^T \bar{c}(\mu(\Gamma_t))] = \int \bar{c}(\mu(\gamma))\lambda_\mu(d\gamma) = \int I(A, \Pi; B \,|\, \gamma)\lambda_\mu(d\gamma)$ independent of the choice of $p(ds_1)$.

Proof: See Lemma 3.3 of [18]. □

For a channel input distribution of the form $\{q(da \,|\, \pi, \gamma)\}$, or equivalent deterministic policy $\mu[\gamma](da, d\pi) = q(da \,|\, \pi, \gamma) \otimes \gamma(d\pi)$, define for each π_t, γ_t the kernel

$$\begin{aligned} &r(d\pi_{t+1}, d\gamma_{t+1} \,|\, \pi_t, \gamma_t) \\ &\quad = \sum_{a_t, b_t} \delta_{\{\Phi_\Gamma(\mu[\gamma_t](da_t, d\pi_t), b_t)\}}(d\gamma_{t+1}) \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) \\ &\qquad \times r(b_t \,|\, \pi_t, a_t) q(a_t \,|\, \pi_t, \gamma_t). \end{aligned}$$

Lemma 7.8: We are given a channel input distribution $\{q(da \,|\, \pi, \gamma)\}$ and a Markov channel with resulting joint measure Q. Then, for each t, we have for Q-almost all π^t, γ^t

$$Q(d\pi_{t+1}, d\gamma_{t+1} \,|\, \pi^t, \gamma^t) = r(d\pi_{t+1}, d\gamma_{t+1} \,|\, \pi_t, \gamma_t).$$

Proof: Fix Borel measurable sets $\Omega_k \subset \mathcal{P}(\mathcal{S})$ and $\Theta_k \subset \mathcal{P}(\mathcal{P}(\mathcal{S}))$ for $k = 1, \ldots, t+1$. Then

$$\begin{aligned} &Q(\Omega^{t+1}, \Theta^{t+1}) \\ &\quad = \int_{\Omega^{t+1}, \Theta^{t+1}, \mathcal{A}^t, \mathcal{B}^t} Q(d\pi^{t+1}, d\gamma^{t+1}, da^t, db^t) \\ &\quad = \int \delta_{\{\Phi_\Pi(\pi_t, a_t, b_t)\}}(d\pi_{t+1}) \\ &\qquad \times \delta_{\{\Phi_\Gamma(q(da_t \,|\, \pi_t, \gamma_t) \otimes \gamma_t(d\pi_t), b_t)\}}(d\gamma_{t+1}) \\ &\qquad \times r(db_t \,|\, \pi_t, a_t) q(da_t \,|\, \pi_t, \gamma_t) Q \\ &\qquad \times (d\pi^t, d\gamma^t, da^{t-1}, db^{t-1}) \\ &\quad = \int_{\Omega^{t+1}, \Theta^{t+1}} r(d\pi_{t+1}, d\gamma_{t+1} \,|\, \pi_t, \gamma_t) Q(d\pi^t, d\gamma^t). \end{aligned}$$

Thus, the lemma is proved. □

To prove the next result we need a stronger mixing condition than that given in Theorem 7.4. Specifically, assume that there exists an $\alpha < 1$ such that for all channel input distributions of the form $\{q(da \,|\, \pi, \gamma)\}$ and all $\pi_t, \gamma_t, \tilde{\pi}_t, \tilde{\gamma}_t$

$$\|r(d\pi_{t+1}, d\gamma_{t+1} \,|\, \pi_t, \gamma_t) - r(d\pi_{t+1}, d\gamma_{t+1} \,|\, \tilde{\pi}_t, \tilde{\gamma}_t)\|_{TV} \leq \alpha. \tag{39}$$

For a consistent measure Q, define for Q almost all a_t, π_t, b^t

$$i_Q(a_t, \pi_t; b_t \,|\, b^{t-1}) = \log \frac{r(b_t \,|\, \pi_t, a_t)}{\int r(b_t \,|\, \tilde{\pi}_t, \tilde{a}_t) Q(d\tilde{\pi}_t, d\tilde{a}_t \,|\, b^{t-1})}.$$

The following theorem will allow us to view the ACOE (36) as an implicit single-letter characterization of the capacity of the Markov channel.

Theorem 7.5: Assume there exists a $V^* \in \mathbb{R}$, a bounded function $w : \gamma \mapsto w(\gamma) \in \mathbb{R}$, and a policy μ^* achieving the supremum for each γ in ACOE (36). Assume condition (39) holds. Then, for μ^* and resulting joint measure R^{μ^*}, let $\{q^*(da \,|\, \pi, \gamma)\}$ be the corresponding optimal channel input distribution (as in Lemma 7.4) and Q^* be the corresponding measure.

1) $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^t i_{Q^*}(A_t, \Pi_t; B_t \,|\, B^{t-1}) = V^*, Q^*-\text{a.s.}$
2) The channel is directed information stable and has a strong converse under the optimal channel input distribution $q^*(da \,|\, \pi, \gamma_t)$.
3) $V^* = C$ is the capacity of the channel.

Proof: We first prove part 2) and 3) assuming part 1) is true. Part 2) follows from part 1) and Proposition 5.1. To prove part 3), note

$$\begin{aligned} C &= \sup_{\{\mathcal{D}_T\}_{T=1}^\infty} \underline{I}_Q(A \to B) \\ &\overset{(a)}{\leq} \sup_{\{\mathcal{D}_T\}_{T=1}^\infty} \liminf_{T\to\infty} \tfrac{1}{T} I_Q(A^T \to B^T) \\ &\overset{(b)}{=} \sup_{\{\{\mu_t\}_{t=1}^T\}_{T=1}^\infty} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^T E_{R^\mu} \bar{c}(U_t) \\ &\overset{(c)}{=} \sup_{\{\mu_t\}_{t=1}^\infty} \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^T E_{R^\mu} \bar{c}(U_t) \\ &= V^* \end{aligned}$$

where (a) follows from Lemma 4.1, (b) follows from theorems 7.1 and 7.2, and (c) follows from Bellman's principle of optimality. Note the supremizations in (b) and (c) are over policies that satisfy the control constraint (28). Now by part 1), we see that (a) holds with equality. Hence, part 3) follows.

We need only to prove part 1). Note that Theorem 7.3 2) implies

$$\begin{aligned} V^* &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T \bar{c}(U_t)\ R^{\mu^*}-\text{a.s.} \\ &\overset{(a)}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T I_{Q^*}(A_t, \Pi_t; B_t \,|\, b^{t-1}),\ R^{\mu^*}-\text{almost all } b^\infty \\ &\overset{(b)}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T I_{Q^*}(A_t, \Pi_t; B_t \,|\, b^{t-1}),\ Q^*-\text{almost all } b^\infty \end{aligned}$$

where (a) follows from (35) and (b) follows because for each t, $Q^*(db^t) = R^{\mu^*}(db^t)$. Hence, $Q^*(db^\infty) = R^{\mu^*}(db^\infty)$. Define the nested family of sigma fields: $\mathbb{F}_t = \sigma(\Pi^{t+1}, A^t, B^t)$. Let

$$\begin{aligned} Z_t(\pi_t, a_t, b^t) = i_{Q^*}(a_t, \pi_t; b_t \,|\, b^{t-1}) \\ - E_{Q^*}\left[i_{Q^*}(A_t, \Pi_t; B_t \,|\, B^{t-1}) \,|\, \mathbb{F}_{t-1} \right]. \end{aligned}$$

The second term can be seen to be equal to $I_{Q^*}(A_t, \pi_t; B_t \,|\, b^{t-1})$. Now

$$\frac{1}{T}\sum_{t=1}^{T} i_{Q^*}(a_t, \pi_t; b_t \,|\, b^{t-1}) = \frac{1}{T}\sum_{t=1}^{T} I_{Q^*}(A_t, \pi_t; B_t \,|\, b^{t-1}) + \frac{1}{T}\sum_{t=1}^{T} Z_t(\pi_t, a_t, b^t). \quad (40)$$

We first show that the second term converges to zero. Clearly, Z_t is $\mathbb{F}_t$-measurable and $E_{Q^*}(Z_t \,|\, \mathbb{F}_{t-1}) = 0$. Hence, Z_t is a martingale difference sequence. The martingale stability theorem [30] states if $\lim_{T\to\infty}\sum_{t=1}^{T}\frac{1}{t^2}E_{Q^*}[Z_t^2 \,|\, \mathbb{F}_{t-1}] < \infty, Q^* - \text{a.s.}$, then $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} Z_t = 0, Q^* - \text{a.s.}$

To show bounded second moments, note that for any t and Q^*-almost all π^t, a^{t-1}, b^{t-1}, we have

$$\begin{aligned} &E_{Q^*}(Z_t^2 \,|\, \mathbb{F}_{t-1}) \\ &\overset{(a)}{\le} E_{Q^*}[i_{Q^*}^2(A_t, \Pi_t; B_t \,|\, B^{t-1}) \,|\, \mathbb{F}_{t-1}] \\ &\overset{(b)}{\le} E_{Q^*}\left[\log^2 r(B_t \,|\, \Pi_t, A_t) \,|\, \mathbb{F}_{t-1}\right] \\ &\quad + E_{Q^*}\left[\log^2\left(\int r(B_t|\tilde{\pi}_t, \tilde{a}_t) Q^*(d\tilde{\pi}_t, d\tilde{a}_t|B^{t-1})\right) |\mathbb{F}_{t-1}\right] \\ &\overset{(c)}{\le} 2\,|\mathcal{B}| \end{aligned}$$

where line (a) follows because the variance is always less than or equal to the second moment and line (b) follows because the cross term is always less than or equal to zero. To see line (c), note that the function $x\log^2 x$ achieves a maximum value of 1 over the domain $0 \le x \le 1$. Hence

$$\begin{aligned} &E_{Q^*}\left[\log^2 r(B_t \,|\, \Pi_t, A_t) \,|\, \mathbb{F}_{t-1}\right] \\ &\quad = E_{Q^*}\left[\sum_{b_t} r(b_t \,|\, \Pi_t, A_t)\log^2 r(b_t \,|\, \Pi_t, A_t) \,|\, \mathbb{F}_{t-1}\right] \\ &\quad \le |\mathcal{B}|. \end{aligned}$$

A similar argument holds for the other addend. Thus, $\sum_t \frac{2|\mathcal{B}|}{t^2}$ is summable, and hence, $\sum_t \frac{E_{Q^*}[Z_t^2 \,|\, \mathbb{F}_{t-1}]}{t^2}$ is summable.

Now we show that the first term in (40) converges to $V^*, Q^* - \text{a.s.}$. Under the optimal channel input distribution, we have $I_{Q^*}(A_t, \pi_t; B_t \,|\, b^{t-1}) = I_{Q^*}(A_t, \pi_t; B_t \,|\, \gamma_t)$. This latter term can be viewed as a bounded function of (π_t, γ_t).

Under the mixing condition (39), we know there exists a unique stationary distribution $\lambda(d\pi, d\gamma)$ for the Markov chain $\{(\Pi_t, \Gamma_t)\}$ such that $\lim_{t\to\infty} \|r^t(d\pi_t, d\gamma_t \,|\, \pi_1, \gamma_1) - \lambda(d\pi_t, d\gamma_t)\|_{TV} = 0$. This follows analogously to Proposition 7.1. By the mixing condition (39) and the strong law of large numbers for Markov chains [19, Th. 4.3.2], we know $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} I_{Q^*}(A_t, \pi_t; B_t \,|\, \gamma_t) = \int\int \lambda(d\pi, d\gamma) I_{Q^*}(A, \pi; B \,|\, \gamma)$ $Q^* -$ a.s.. Finally, note that since $R^{\mu^*}(db^\infty) = Q^*(db^\infty)$, we have by Proposition 7.1 and (35) that $\int\int \lambda(d\pi, d\gamma) I_{Q^*}(A, \pi; B \,|\, \gamma) = \int \lambda(d\gamma) I_{Q^*}(A, \Pi; B \,|\, \gamma) = \int \bar{c}(\mu^*[\gamma])\lambda_{\mu^*}(d\gamma) = V^*$. □

VIII. Cases With Simple Sufficient Statistics

As we have already seen, the sufficient statistics $\Pi_t \in \mathcal{P}(\mathcal{S})$ and $\Gamma_t \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$ can be quite complicated in general. This in turn implies that solving the ACOE equation (36) can be quite difficult. There are, though, many scenarios when the sufficient statistics become much simpler and hence the ACOE becomes simpler. In these cases, one can apply exact or approximate dynamic programming techniques to solve the ACOE. The ACOE is an implicit single-letter characterization of the capacity. In general, it will be difficult to get an explicit formula for the capacity.

A. S Computable From the Channel Input and Output

In many scenarios, the state S_t is computable from (A^{t-1}, B^{t-1}). One example of such a channel would be $\{p(db_t \,|\, a_t, a_{t-1}, b_{t-1})\}$. Here one could choose the state to be $S_t = (A_{t-1}, B_{t-1})$. We discuss other examples below.

In this section, we assume that $p(ds_1) = \delta_{\{s_1\}}(ds_1)$ for some fixed state s_1 and for $t > 1$, we have $\Pi_t(ds_t) = \delta_{\{S_t\}}(ds_t)\, Q-\text{a.s.}$. Recall that Π_t is a function of (A^{t-1}, B^{t-1}) and satisfies the recursion $\pi_{t+1} = \Phi_\Pi(\pi_t, a_t, b_t)$. This in turn implies that there exists a function Φ_S such that $s_{t+1} = \Phi_S(s_t, a_t, b_t)$. To see this, recall (13). Because Π_t, Π_{t+1} are Diracs Q-almost surely, it must be the case that S_{t+1} is a function of $S_t, A_t, B_t\ Q - \text{a.s.}$.

Because $\Pi_t = \delta_{S_t}$, we can, in an abuse of notation, identify them together: $\Pi_t = S_t$. Hence, again in an abuse of notation, Γ can be viewed as the conditional probability of the state S as opposed to the conditional probability of Π. Specifically, $\Gamma \in \mathcal{P}(\mathcal{S})$ as opposed to $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{S}))$. Then, we can restrict ourselves to control policies of the form: $\mu : \mathcal{P}(\mathcal{S}) \to \mathcal{U} = \mathcal{P}(\mathcal{S}\times\mathcal{A})$ taking $\gamma \mapsto u(ds, da)$. Now the control constraints take the form $\mathcal{U}(\gamma) = \{u(ds, da) : u(ds, da) \in \mathcal{U}, u(ds) = \gamma(ds)\}$. Then, the channel input distribution has the form $\{q(da_t \,|\, s_t, \gamma_t)\}$. The dynamics of Γ_t given in (24) and (25) simplify to $\gamma_1(ds_1) = \delta_{\{s_1\}}(ds_1)$, and for $t > 1$ and all s_t, we have

$$\gamma_t[u^{t-1}, b^{t-1}](s_t) = \sum_{s_{t-1}, a_{t-1}} \delta_{\{\Phi_S(s_{t-1}, a_{t-1}, b_{t-1})\}}(s_t) r \times (ds_{t-1}, a_{t-1} \,|\, u_{t-1}, b_{t-1}) \quad (41)$$

Hence, (33) simplifies to

$$r(d\gamma_{t+1} \,|\, \gamma_t, u_t) = \sum_{s_t, a_t, b_t} \delta_{\{\Phi_\Gamma(u_t, b_t)\}}(d\gamma_{t+1}) \times p(b_t \,|\, s_t, a_t) u_t(s_t, a_t) \quad (42)$$

where $\Phi_\Gamma(u, b)$ comes from (41). The cost in (30) simplifies as well

$$\bar{c}(u) = \sum_{s,a,b} p(b \,|\, s, a) u(s, a) \log \frac{p(b \,|\, s, a)}{\sum_{\tilde{s},\tilde{a}} p(b \,|\, \tilde{s}, \tilde{a}) u(\tilde{s}, \tilde{a})}. \quad (43)$$

In addition, $I(A_t, \Pi_t; B_t \,|\, \Gamma_t) = I(A_t; B_t \,|\, S_t, \Gamma_t) + I(S_t; B_t \,|\, \Gamma_t)$. Note that the second term shows that S_t can also convey information to the decoder. Finally, the ACOE (36) in

Theorem 7.3 simplifies to an equation where $w(\gamma)$ is now a function over $\mathcal{P}(\mathcal{S})$

$$V^* + w(\gamma) = \sup_{u \in \mathcal{U}(\gamma)} \left(\bar{c}(u) + \int w(\tilde{\gamma}) r(d\tilde{\gamma} \,|\, \gamma, u) \right). \quad (44)$$

We now examine two cases where the computations simplify further: S is either computable from the channel input only or the channel output only.

Case 1—S Computable From the Channel Input Only: Here we assume S_t is computable from only A^{t-1} and hence S is known to the transmitter. Specifically, we assume that Π_t is a function of A^{t-1} and satisfies the recursion $\pi_{t+1} = \Phi_\Pi(\pi_t, a_t)$. This in turn implies there exists a function Φ_S such that $s_{t+1} = \Phi_S(s_t, a_t)$. These channels are often called *finite-state machine Markov channels*. Note that any general channel of the form $\{p(db_t \,|\, a_t, a_{t-\Delta}^{t-1})\}$, for a finite Δ, can be converted into a Markov channel with state $S_t = A_{t-\Delta}^{t-1}$ computable from the channel input.

As before, in an abuse of notation, we can identify $\Pi = S$ and Γ can be viewed as a conditional probability of the state S. Equations (41)–(44) continue to hold with obvious modifications. See [39] for more details. For Gaussian finite-state machine Markov channels, the estimate Γ_t can be easily computed by using a Kalman filter [40].

Case 2—S Computable From the Channel Output Only: Here we assume S_t is computable from only B^{t-1}. Specifically, we assume that S is known to the receiver, and via feedback, is known to the transmitter. Then, Π_t is a function of B^{t-1} and satisfies the recursion $\pi_{t+1} = \Phi_\Pi(\pi_t, b_t)$. This in turn implies there exists a function Φ_S such that $s_{t+1} = \Phi_S(s_t, b_t)$. Note that any general channel of the form $\{p(db_t \,|\, a_t, b_{t-\Delta}^{t-1})\}$, for a finite Δ, can be converted into a Markov channel with state $S_t = B_{t-\Delta}^{t-1}$ computable from the channel output.

As before, in an abuse of notation, we can identify $\Pi = S$. In addition, because Π is computable from the channel outputs, we can, again in an abuse of notation, identify $\Gamma = \Pi$ and hence identify $\Gamma = S$.

The control constraints simplify $\mathcal{U}(\gamma) = \mathcal{U}(s) = \{u(ds, da) : u \in \mathcal{U}, u(ds) = \delta_s(ds)\}$. Because the state S is known to both the transmitter and the receiver, we see that the control constraints become trivial. Hence, we can use control actions of the form $u(da)$ as opposed to $u(ds, da)$. We can then restrict ourselves to control policies of the form $\mu : \mathcal{S} \to \mathcal{U} = \mathcal{P}(\mathcal{A})$ taking $s \mapsto u(da)$. This implies that the channel input distribution has the form $\{q(da_t \,|\, s_t)\}$.

The dynamics in (33) simplify quite a bit: $\gamma_1 = s_1$ and for $t > 1$, we have $\gamma_t = s_t$ and

$$\begin{aligned} r(d\gamma_{t+1} \,|\, \gamma_t, u_t) &= r(ds_{t+1} \,|\, s_t, u_t) \\ &= \sum_{a_t, b_t} \delta_{\{\Phi_S(s_t, b_t)\}}(ds_{t+1}) p(b_t \,|\, s_t, a_t) u_t(a_t). \end{aligned} \quad (45)$$

The cost in (30) simplifies to

$$\begin{aligned} \bar{c}(s, u) &= \sum_{a,b} p(b \,|\, s, a) u(a) \log \frac{p(b \,|\, s, a)}{\sum_{\tilde{a}} p(b \,|\, s, \tilde{a}) u(\tilde{a})} \\ &= I(A; B \,|\, s). \end{aligned}$$

The last equality also follows because $I(A_t, \Pi_t; B_t \,|\, \Gamma_t) = I(A_t, S_t; B_t \,|\, S_t) = I(A_t; B_t \,|\, S_t)$. Finally, the ACOE (36) in Theorem 7.3 simplifies to an equation where $w(\gamma)$ is now a function over $\mathcal{S}$

$$\begin{aligned} &V^* + w(s) \\ &\quad = \sup_{u \in \mathcal{U}} \left(\bar{c}(s, u) + \sum_{\tilde{s}} w(\tilde{s}) r(\tilde{s} \,|\, s, u) \right) \\ &\quad = \sup_{q(da \,|\, s)} \left(I(A; B \,|\, s) + \sum_{\tilde{s}} w(\tilde{s}) r\left(\tilde{s} \,|\, s, q(da \,|\, s)\right) \right). \end{aligned} \quad (46)$$

Here the ACOE is defined over a finite-state space $\mathcal{S}$ and straightforward value and policy iteration techniques can be used to compute the solution (when it exists) [2]. In this, the sufficient condition (37) reduces to $\sup_{s_t, \tilde{s}_t, u_t, \tilde{u}_t} \| r(ds_{t+1} \,|\, s_t, u_t) - r(ds_{t+1} \,|\, \tilde{s}_t, \tilde{u}_t) \|_{TV} < 1$.

a) Markov Channels With State Observable to the Receiver: An important scenario that falls under the case just described is that of a Markov channel $p(ds_1), \{p(ds_{t+1} \,|\, s_t, a_t, b_t)\}, \{p(db_t \,|\, s_t, a_t)\}$, with state observable to the receiver. Specifically, at time t, we assume that along with B_t, the state S_{t+1} is observable to the receiver. The standard technique for dealing with this setting is to define a new channel output as follows: $\bar{B}_t = (B_t, S_{t+1})$. The new Markov channel has the same state transition kernel but the channel output is $p(d\bar{b}_t \,|\, s_t, a_t) = p(ds_{t+1} \,|\, s_t, a_t, b_t) \otimes p(db_t \,|\, s_t, a_t)$. We also assume that S_1 is observable to the transmitter. (This can be achieved by assuming that $\bar{B}_0 = S_1$ is transmitted during epoch 0.) Thus, the dynamics in (45) can be written as $\gamma_1 = s_1$, and for $t > 1$, we have $\gamma_t = s_t$ and

$$\begin{aligned} r(d\gamma_{t+1} \,|\, \gamma_t, u_t) &= r(ds_{t+1} \,|\, s_t, u_t) \\ &= \sum_{a_t, b_t} p(ds_{t+1} \,|\, s_t, a_t, b_t) p(b_t \,|\, s_t, a_t) u_t(a_t) \end{aligned} \quad (47)$$

Also, $I(A_t, \Pi_t; \bar{B}_t \,|\, \Gamma_t) = I(A_t; B_t \,|\, S_t) + I(A_t; S_{t+1} \,|\, S_t, B_t)$. The second addend is zero if there is no ISI. If there is no ISI, then (47) reduces to $r(d\gamma_{t+1} \,|\, \gamma_t, u_t) = p(d\gamma_{t+1} \,|\, \gamma_t) = p(ds_{t+1} \,|\, s_t)$. If $p(ds_{t+1} \,|\, s_t)$ is an ergodic transition kernel with stationary distribution ν, then there exists a bounded solution to the ACOE [1]. In this case, the ACOE (46) simplifies

$$\begin{aligned} V^* + w(s) &= \sup_{u \in \mathcal{U}} \left(\bar{c}(s, u) + \sum_{\tilde{s}} w(\tilde{s}) p(d\tilde{s} \,|\, s) \right) \\ &= \sup_{q(da \,|\, s)} I(A; B \,|\, s) + \sum_{\tilde{s}} w(\tilde{s}) p(d\tilde{s} \,|\, s). \end{aligned}$$

Note that $q(da\,|\,s)$ only enters the first term. Now integrate each term with respect to ν. This leads to $V^* = \sum_s \nu(s) \sup_u \bar{c}(s,u) = \sum_s \nu(s) \sup_{q(da\,|\,s)} I(A;B\,|\,s)$. Thus, we recover the well-known formula for the capacity of a non-ISI ergodic Markov channel with state available to both the transmitter and the receiver.

B. Π Computable From the Channel Output

Here we assume that Π_t is a function of B^{t-1} only and satisfies the recursion $\pi_{t+1} = \Phi_\Pi(\pi_t, b_t)$. Hence, $\Gamma_t(d\pi_t) = \delta_{\{\Pi_t\}}(d\pi_t)\, Q-\text{a.s.}$. We can then, in an abuse of notation, identify $\Gamma = \Pi$. Now Γ can be viewed as a conditional probability of the state S. Recall the discussion of the canonical Markov channel at the end of Section VI-B. Here we can view the associated canonical Markov channel as a Markov channel with state Π computable from the channel output only (as discussed in the previous section).

We can then restrict ourselves to control policies of the form $\mu : \mathcal{P}(\mathcal{S}) \to \mathcal{U} = \mathcal{P}(\mathcal{A})$ taking $\pi \mapsto u(da)$. To see this, note that the control constraints become trivial, and hence, we can use control actions of the form $u(da)$ as opposed to $u(d\pi, da)$. This implies that the channel input distribution has the form $\{q(da_t\,|\,\pi_t)\}$.

The dynamics in (33) simplifies to $\gamma_1 = \pi_1$ and for $t > 1$, we have $\gamma_t = \pi_t$ and

$$\begin{aligned} r(d\gamma_{t+1}\,|\,\gamma_t, u_t) &= r(d\pi_t\,|\,\pi_t, u_t) \\ &= \sum_{s_t, a_t, b_t} \delta_{\{\Phi_\Pi(\pi_t, b_t)\}}(d\pi_{t+1}) \\ &\quad \times p(b_t\,|\,s_t, a_t)\, \pi_t(s_t)\, u_t(a_t). \end{aligned} \tag{48}$$

The cost in (30) simplifies as well

$$\begin{aligned} \bar{c}(\pi, u) &= \sum_{s,a,b} p(b\,|\,s,a)\pi(s)u(a) \log \frac{\sum_{\tilde{s}} p(b\,|\,\tilde{s},a)\pi(\tilde{s})}{\sum_{\tilde{s},\tilde{a}} p(b\,|\,\tilde{s},\tilde{a})\pi(\tilde{s})u(\tilde{a})} \\ &= I(A;B\,|\,\pi). \end{aligned}$$

The last equality also follows because $I(A_t, \Pi_t; B_t\,|\,\Gamma_t) = I(A_t; B_t\,|\,\Pi_t)$.

Finally, the ACOE equation (36) in Theorem 7.3 simplifies to an equation where $w(\gamma)$ is now a function over $\mathcal{P}(\mathcal{S})$

$$\begin{aligned} &V^* + w(s) \\ &\quad = \sup_{u \in \mathcal{U}} \left(\bar{c}(\pi, u) + \int w(\tilde{\pi}) r(d\tilde{\pi}\,|\,\pi, u) \right) \\ &\quad = \sup_{q(da\,|\,\pi)} \left(I(A;B\,|\,\pi) + \int w(\tilde{\pi}) r\left(d\tilde{\pi}\,|\,\pi, q(da\,|\,\pi)\right) \right). \end{aligned} \tag{49}$$

As discussed above, the optimal channel input distribution $q(da_t\,|\,\pi_t, \gamma_t)$ can be written in the form $q(da_t\,|\,\pi_t)$, or more generally, $q(da_t\,|\,b^{t-1})$. Furthermore, the code-function distribution, given by (8), simplifies to a product distribution. Note that there is no a^{t-1} dependence in $q(da_t\,|\,b^{t-1})$. Hence, for each t and f_t, the code function distribution is given by $p(f_t) = \prod_{b^{t-1}} q(f_t(b^{t-1})\,|\,b^{t-1})$. Then, $P_{F^T}(df^T) = \otimes_{t=1}^T p(df_t)$. One can easily verify for each t that $P_{F^T}(\Upsilon^t(b^{t-1}, a^t)) = \vec{q}(a^t\,|\,b^{t-1})$ and hence is good with respect to $\{q(da_t\,|\,b^{t-1})\}$.

In summary, if the sufficient statistic Π is computable from the channel output, then the optimal code-function distribution can be taken to be a product measure. If Π_t depends on A^{t-1}, then the optimal code function, in general, will not be a product measure.

C. Error Exponents for Markov Channels

We can specialize the results on error exponents presented in Section V-D to Markov channels. For a given Markov channel $p(ds_1), p(ds_{t+1}\,|\,s_t, a_t, b_t), p(db_t\,|\,s_t, a_t)$ with a stationary channel input distribution $q(da\,|\,\pi, \gamma)$, the random coding directed error exponent takes the form $\vec{E}_T(R, q(da\,|\,\pi, \gamma)) = \max_{0 \le \rho \le 1}$ of

$$-\rho R - \frac{1}{T} \ln \sum_{b^T} \left[\sum_{a^T} \prod_{t=1}^T q(a_t\,|\,\pi_t, \gamma_t) \times \left[\prod_{t=1}^T r(b_t\,|\,\pi_t, a_t) \right]^{\frac{1}{1+\rho}} \right]^{1+\rho}.$$

In general, this can be difficult to compute. There are cases though where the formula simplifies. We describe one such case now. Consider a Markov channel without ISI and with the state observable to the receiver. As discussed in Section VIII-A2, the optimal channel input distribution for maximizing the directed information is stationary and takes the form $q(da\,|\,s)$. Assume that $p(ds_{t+1}\,|\,s_t)$ is ergodic with stationary distribution ν. We know that the capacity in this case equals $\sum_s \nu(s) \max_u \bar{c}(s,u) = \sum_s \nu(s) \max_{q(a\,|\,s)} I(A;B\,|\,s)$. The error exponent for channel code functions drawn randomly from the channel input distribution $\{q(da\,|\,s)\}$ can be written

$$\begin{aligned} &\vec{E}_T\left(R, q(da\,|\,s)\right) \\ &\quad = \max_{0 \le \rho \le 1} -\rho R - \frac{1}{T} \ln \sum_{s^T, b^T} \\ &\qquad \times \left[\sum_{a^T} \prod_{t=1}^T q(a_t|s_t)(p(b_t|s_t,a_t)p(s_t|s_{t-1}))^{\frac{1}{1+\rho}} \right]^{1+\rho} \\ &\quad = \max_{0 \le \rho \le 1} -\rho R - \frac{1}{T} \ln \sum_{s^T} P(s^T) \\ &\qquad \times \prod_{t=1}^T \sum_{b_t} \left[\sum_{a_t} q(a_t\,|\,s_t) p(b_t\,|\,s_t, a_t)^{\frac{1}{1+\rho}} \right]^{1+\rho} \end{aligned}$$

where $P(ds^T) = \otimes_{t=1}^T p(ds_t\,|\,s_{t-1})$.

Define $\phi_\rho(s) = \sum_b \left[\sum_a q(a\,|\,s) p(b\,|\,s,a)^{\frac{1}{1+\rho}} \right]^{1+\rho}$. Then

$$\frac{1}{T} \ln \sum_{s^T} P(s^T) \prod_{t=1}^T \left[\sum_{b_t, a_t} q(a_t\,|\,s_t)(p(b_t\,|\,s_t, a_t)^{\frac{1}{1+\rho}} \right]^{1+\rho}$$

$$\stackrel{(a)}{=} \frac{1}{T}\ln\sum_{s^T} P(s^T)\exp\left(T\sum_s \mu(s;s^T)\ln\phi_\rho(s)\right)$$
$$\stackrel{(b)}{=} \frac{1}{T}\ln\int P_T(d\mu)\exp\left(T\sum_s \mu(s)\ln\phi_\rho(s)\right).$$

In (a), $\mu(s;s^T)=\frac{1}{T}\sum_{t=1}^T \delta_{\{s\}}(s_t)$ is the empirical occupation measure for the realization s^T. In (b), $P_T(d\mu)$ corresponds to the probability that μ is the empirical occupation measure under $p(ds^T)$. Specifically, for a Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$, we have $p_T(\Omega)=P(\{s^T : \mu(s;s^T)\in\Omega\})$.

Sanov's theorem for the empirical measure of a Markov chain $p_T(d\mu)$ shows that the associated large deviation rate function is [11, Th. 3.1.6]

$$\mathcal{I}(\mu)=\sup_w \sum_s \mu(s)\ln\frac{w(s)}{\sum_{\tilde{s}} w(\tilde{s})p(s\,|\,\tilde{s})}$$

where $w:\mathcal{S}\to(0,\infty)$. An application of Varadhan's integral lemma [11, Th. 4.3.1] shows

$$\lim_{T\to\infty}\frac{1}{T}\ln\int p_T(d\mu)\exp\left(T\sum_s \mu(s)\ln\phi_\rho(s)\right) = \sup_\mu\left(\sum_s \mu(s)\ln\phi_\rho(s)-\mathcal{I}(\mu)\right).$$

Hence, the error exponent takes the form

$$\lim_{T\to\infty}\vec{E}_T\left(R,q(da\,|\,s)\right) = \max_{0\le\rho\le 1}\left(-\rho R-\sup_\mu\left(\sum_s \mu(s)\ln\phi_\rho(s)-\mathcal{I}(\mu)\right)\right).$$

The right-hand side can be viewed as a single-letter characterization of the error exponent.

The error exponent simplifies even more if the state process $\{S_t\}$ is i.i.d. Specifically, if $p(ds_t\,|\,s_{t-1})=\nu(ds_t)\ \forall s_{t-1}$. Sanov's theorem for the empirical measure of the i.i.d. process S^T shows that the associated large deviation rate function is $\mathcal{I}(\mu)=D(\mu\,|\,\nu)$. Then, the error exponent takes the form

$$\lim_{T\to\infty}\vec{E}_T\left(R,\{q(da\,|\,s)\}_{t=1}^T\right) = \max_{0\le\rho\le 1}\left(-\rho R-\sup_\mu\left(\sum_s \mu(s)\ln\phi_\rho(s)-D(\mu\,|\,\nu)\right)\right) = \max_{0\le\rho\le 1}\left(-\rho R-\ln\sum_s \nu(s)\phi_\rho(s)\right).$$

The supremizing μ can be shown to be a tilted version of ν. Specifically, $\mu(s)=\frac{\phi_\rho(s)\nu(s)}{\sum_{\tilde{s}}\phi_\rho(\tilde{s})\nu(\tilde{s})}$. This shows the role that atypical state sequences can have on the error exponent.

Note that for the non-ISI, ergodic Markov chain, the optimal channel input distribution has the form $q(da\,|\,s)$. In this case, optimality refers to maximizing the directed information. This $q(da\,|\,s)$, though, may not be the channel input distribution that maximizes the error exponent given in Definition 5.4. Intuitively, we expect state feedback to help increase capacity while we expect channel output feedback to help increase the error exponent. Maximizing the error exponent over all channel input distributions is a challenging open problem.

We have computed the error exponent for fixed length channel codes. It is known that one can get better error exponents if one allows variable length channel codes [5].

IX. Maximum-Likelihood Decoding

We now consider the problem of maximum-likelihood decoding. For a given message set $\mathcal{W}$, fix a channel code $\{f^T[w]\ w\in\mathcal{W}\}$. Assume the messages are chosen uniformly. Hence, each channel code function is chosen with probability $P_{F^T}(f^T[w])=\frac{1}{|\mathcal{W}|}$. For a consistent joint measure $Q(df^T,da^T,db^T)$, our task is to simplify the computation of $\arg\max_{w\in\mathcal{W}} Q(f^T[w]\,|\,b^T)$.

First consider the general channels described in Section V. Note $Q(f^T,b^T)=Q(f^T,a^T=f^T(b^{T-1}),b^T)=Q(f^T\,|\,a^T=f^T(b^{T-1}),b^T)Q(a^T=f^T(b^{T-1}),b^T)$. Also, if $Q(f^T,a^T,b^T)>0$, then

$$\begin{aligned}
&Q(f^T\,|\,a^T,b^T)\\
&\quad=\frac{Q(f^T,a^T,b^T)}{Q(a^T,b^T)}\\
&\quad=\frac{P_{F^T}(f^T)\vec{p}(b^T\,|\,a^T)\prod_{t=1}^T\delta_{\{f_t(b^{t-1})\}}(a_t)}{\vec{p}(b^T\,|\,a^T)\prod_{t=1}^T Q(a_t\,|\,a^{t-1},b^{t-1})}\\
&\quad\stackrel{(a)}{=}\frac{P_{F^T}(f^T)}{P_{F^T}(\Upsilon^T(b^{T-1},a^T))}\\
&\quad=\frac{1}{|\Upsilon^T(b^{T-1},a^T)|}
\end{aligned}$$

where (a) follows by Lemma 5.1. Note that this implies that $F^T-A^T-B^T$ is *not* a Markov chain under Q.

Due to the feedback, we effectively have a different channel code without feedback for each b^{T-1}. For each b^{T-1}, define

$$\Lambda(b^{T-1})=\{a^T : a^T=f^T[w](b^{T-1}) \text{ for some } w\in\mathcal{W}\}.$$

Thus, computing $\arg\max_{w\in\mathcal{W}} Q(f^T[w]\,|\,b^T)$ is equivalent to

$$\arg\max_{a^T\in\Lambda(b^{T-1})}\frac{1}{|\Upsilon^T(b^{T-1},a^T)|}\times\prod_{t=1}^T p(b_t\,|\,a^t,b^{t-1})Q(a_t\,|\,a^{t-1},b^{t-1}) \quad (50)$$

where $\{Q(da_t\,|\,a^{t-1},b^{t-1})\}$ is the induced channel input distribution for $P_{F^T}(f^T[w])$.

For the Markov channel case, we may replace $p(db_t\,|\,a^t,b^{t-1})$ with $p(db_t\,|\,\pi_t,a_t)$ in (50). If, in addition, the channel code is chosen such that the induced channel input distribution has the form $\{q(da_t\,|\,\pi_t,\gamma_t)\}$, then

$$\arg\max_{a^T\in\Lambda(b^{T-1})}\frac{1}{|\Upsilon^T(b^{T-1},a^T)|}\times\prod_{t=1}^T p(b_t\,|\,\pi_t,a_t)q(a_t\,|\,\pi_t,\gamma_t). \quad (51)$$

In the case where $|\Upsilon^T(b^{T-1},a^T)|=1,\forall a^T\in\Lambda(b^{T-1})$, the optimization in (51) can be treated as a deterministic longest path problem.

X. Conclusion

We have presented a general framework for treating channels with memory and feedback. We first proved a general coding theorem based on Massey's concept of *directed information* and Dobrushin's program of communication as interconnection. We then specialized this result to the case of Markov channels. To compute the capacity of these Markov channels, we converted the directed information optimization problem into a partially observed MDP. This required identifying appropriate sufficient statistics at the encoder and the decoder. The ACOE verification theorem was presented and sufficient conditions for the existence of a solution were provided. The complexity of many feedback problems can now be understood by examining the complexity of the associated ACOE. Error exponents were presented.

The framework developed herein allows one to apply approximate dynamic programming techniques, such as value and policy iteration and reinforcement learning, for computing the capacity. Such dynamic programming techniques are presented in [39] for the case of finite-state machine Markov channels. Finally, the framework presented here allows one to compute the capacity under restricted classes of policies. This is useful if one is willing to sacrifice capacity for the benefit of a simpler policy.

Appendix

A. Review of Stochastic Kernels

The results here are standard and can be found in, for example, [3]. Let $(\mathcal{V},\mathcal{A})$ be a Borel space and let $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y},\mathcal{B}_{\mathcal{Y}})$ be Polish spaces equipped with their Borel σ-algebras.

Definition A.1: Let $\tau(dx\,|\,v)$ be a family of probability measures on $\mathcal{X}$ parameterized by $v\in\mathcal{V}$. We say that τ is a *stochastic kernel from $\mathcal{V}$ to $\mathcal{X}$* if for every Borel set $B\in\mathcal{B}_{\mathcal{X}}$, the function $v\mapsto\tau(B\,|\,v)\in[0,1]$ is measurable.

Lemma A.1: For $B\in\mathcal{B}_{\mathcal{X}}$, define $f_B:\mathcal{P}(\mathcal{X})\to[0,1]$ by $f_B:\mu\mapsto\mu(B)$ for $\mu\in\mathcal{P}(P\mathcal{X})$. Then

$$\mathcal{B}_{\mathcal{P}(\mathcal{X})}=\sigma[\cup_{B\in\mathcal{B}_{\mathcal{X}}}f_B^{-1}(\mathcal{B}_{I\!R})].$$

Theorem A.1: Let $\tau(dx\,|\,v)$ be a family of probability measures on $\mathcal{X}$ given $\mathcal{V}$. Then, $\tau(dx\,|\,v)$ is a stochastic kernel if and only if $v\in\mathcal{V}\mapsto\tau(dx\,|\,v)\in\mathcal{P}(\mathcal{X})$ is measurable. That is if and only if $\tau(\cdot\,|\,v)$ is a random variable from $\mathcal{V}$ into $\mathcal{P}(\mathcal{X})$.

Since $\tau(\cdot\,|\,v)$ is a random variable from $\mathcal{V}$ into $\mathcal{P}(\mathcal{X})$, it follows that the class of stochastic kernels is closed under weak limits (weak topology on the space of probability measures.)

We now discuss interconnections of stochastic kernels. Let $\tau_1(dx\,|\,v)$ be a stochastic kernel from $\mathcal{V}$ to $\mathcal{X}$ and $\tau_2(dy\,|\,v,x)$ be a stochastic kernel from $\mathcal{V}\times\mathcal{X}$ to $\mathcal{Y}$. Then, the joint stochastic kernel $\tau_1\otimes\tau_2$ from $\mathcal{V}$ to $\mathcal{X}\times\mathcal{Y}$ is, for all $v\in\mathcal{V}$, $A\in\mathcal{B}_{\mathcal{X}}$, and $B\in\mathcal{B}_{\mathcal{Y}}$, we have $\tau_1\otimes\tau_2(A\times B\,|\,v)=\int_{A\times B}\tau_2(dy\,|\,v,x)\tau_1(dx\,|\,v)=\int_A\tau_2(B\,|\,v,x)\tau_2(dx\,|\,v)$. Via the Ionescu–Tulcea theorem, this can be generalized to interconnections of countable number of stochastic kernels.

We now discuss the decompositions of measures.

Theorem A.2: Let $\lambda(dx\otimes dy)$ be a probability measure on $(\mathcal{X}\times\mathcal{Y},\mathcal{B}_{\mathcal{X}}\otimes\mathcal{B}_{\mathcal{Y}})$. Let $\lambda_1(A)=\lambda(A,\mathcal{Y}),A\in\mathcal{B}_{\mathcal{X}}$ be the first marginal. Then, there exists a stochastic kernel $\lambda(dy\,|\,x)$ on $\mathcal{Y}$ given $\mathcal{X}$ such that for all $A\in\mathcal{B}_{\mathcal{X}}$ and $B\in\mathcal{B}_{\mathcal{Y}}$, then we have

$$\lambda(A\times B)=\int_{A\times B}\lambda_1(dx)\lambda(dy\,|\,x)=\int_A\lambda(B\,|\,x)\lambda_1(dx).$$

This can be generalized to a parametric dependence.

Theorem A.3: Let $\lambda(dx\otimes dy\,|\,v)$ be a stochastic kernel on $\mathcal{X}\times\mathcal{Y}$ given $\mathcal{V}$. Let $\lambda_1(A\,|\,v)$ be the first marginal, which is a stochastic kernel on $\mathcal{X}$ given $\mathcal{V}$ defined by

$$\lambda_1(A\,|\,v)=\lambda(A,\mathcal{Y}\,|\,v),A\in\mathcal{B}_{\mathcal{X}},v\in\mathcal{V}.$$

Then, there exists a stochastic kernel $\lambda(dy\,|\,v,x)$ on $\mathcal{Y}$ given $\mathcal{V}\times\mathcal{X}$ such that $\forall v\in\mathcal{V}$, $A\in\mathcal{B}_{\mathcal{X}}$, and $B\in\mathcal{B}_{\mathcal{Y}}$, then we have

$$\begin{aligned}\lambda(A\times B\,|\,v)&=\int_{A\times B}\lambda_1(dx\,|\,v)\lambda(dy\,|\,v,x)\\&=\int_A\lambda(B\,|\,v,x)\lambda_1(dx\,|\,v).\end{aligned}$$

Let $\lambda(dx\otimes dy\,|\,v)$ be a stochastic kernel on $\mathcal{X}\times\mathcal{Y}$ given $\mathcal{V}$ and suppose the stochastic kernel $\tau(dy\,|\,v,x)$ on $\mathcal{Y}$ given $\mathcal{V}\times\mathcal{X}$ satisfies $\forall v\in\mathcal{V},\forall\,B\in\mathcal{B}_{\mathcal{Y}}$, then we have

$$\lambda(B\,|\,v,x)=\tau(B\,|\,v,x)\text{for}\lambda_1(dx\,|\,v)\text{ almost all }x.$$

Then, for any measurable function $g:\mathcal{V}\times\mathcal{X}\times\mathcal{Y}\to I\!R$ and all $v\in\mathcal{V}$, we have

$$E(g(V,X,Y)\,|\,v)=\int_{\mathcal{X}\times\mathcal{Y}}g(v,x,y)\tau(dy\,|\,v,x)\lambda_1(dx\,|\,v)$$

whenever the conditional expectation on the left-hand side exists.

Finally, recall that a stochastic kernel $\tau(dy\,|\,x)$ on $\mathcal{Y}$ given $\mathcal{X}$ is *continuous* if for all continuous bounded functions v on $\mathcal{Y}$, the function $\int v(y)p(dy\,|\,x)$ is a continuous and bounded function on $\mathcal{X}$.

Theorem A.4: If $\tau(dy\,|\,x)$ is a continuous stochastic kernel on $\mathcal{Y}$ given $\mathcal{X}$ and $v(x,y)$ is a continuous bounded function on $\mathcal{X}\times\mathcal{Y}$, then $\int v(x,y)\tau(dy\,|\,x)$ is a continuous bounded function on $\mathcal{X}$.

B. Lemma 4.1

We repeat the statement of Lemma 4.1 for convenience.

Lemma 4.1: For any sequence of joint measures $\{P_{A^T,B^T}\}_{T=1}^{\infty}$, we have $\underline{I}(A\to B)\le\liminf_{T\to\infty}\frac{1}{T}I(A^T\to B^T)\le\limsup_{T\to\infty}\frac{1}{T}I(A^T\to B^T)\le\overline{I}(A\to B)$.

The proof of Lemma 4.1 is adapted from [17, Lemma A1] and [34, Th. 8]. We need the following three lemmas. Combined they state that the mass of $\vec{i}(A^T;B^T)$ at the tails is small. Recall that $|\mathcal{B}|<\infty$.

Lemma A.2: Let $L > \log|\mathcal{B}|$. For any sequence of measures $\{P_{B^T}\}_{t=1}^T$, we have

$$\lim_{T\to\infty} E\left[\frac{1}{T}\log\frac{1}{P(B^T)}\,1_{\left\{\frac{1}{T}\log\frac{1}{P(B^T)}\geq L\right\}}\right]=0.$$

Proof: Let $\Omega=\{b^T : P(b^T)\leq 2^{-TL}\}$. Now

$$\begin{aligned}
&E\left[\frac{1}{T}\log\frac{1}{P(B^T)}1_{\left\{\frac{1}{T}\log\frac{1}{P(B^T)}\geq L\right\}}\right]\\
&=\frac{1}{T}\sum_{b^T\in\Omega}P(b^T)\log\frac{1}{P(b^T)}\\
&=\frac{1}{T}P(\Omega)\sum_{b^T\in\Omega}\frac{P(b^T)}{P(\Omega)}\log\frac{1}{\frac{P(b^T)}{P(\Omega)}}-\frac{1}{T}P(\Omega)\log P(\Omega)\\
&\leq\frac{1}{T}P(\Omega)\log|\mathcal{B}^T|-\frac{1}{T}P(\Omega)\log P(\Omega)\\
&\leq\frac{1}{T}P(\Omega)\log|\mathcal{B}^T|+\frac{1}{2T}
\end{aligned}$$

where the first inequality follows because entropy is maximized by the uniform distribution and the second inequality follows because $-x\log x\leq\frac{1}{2}$, $0\leq x\leq 1$. Now $P(\Omega)\leq|\Omega|2^{-TL}\leq|\mathcal{B}^T|2^{-TL}$. Thus

$$\begin{aligned}
&E\left[\frac{1}{T}\log\frac{1}{P(B^T)}\,1_{\left\{\frac{1}{T}\log\frac{1}{P(B^T)}\geq L\right\}}\right]\\
&\qquad\leq\log|\mathcal{B}|2^{-T(L-\log|\mathcal{B}|)}+\frac{1}{2T}.
\end{aligned}$$

This upper bound goes to zero as $T\to\infty$. □

Lemma A.4: For any sequence of joint measures $\{P_{A^T,B^T}\}_{T=1}^\infty$, we have

$$\lim_{T\to\infty} E\left[\frac{1}{T}\vec{i}(A^T;B^T)\,1_{\{\frac{1}{T}\vec{i}(A^T;B^T)\leq 0\}}\right]=0.$$

Proof: It follows from [25, p. 10].

Lemma A.4: Let $L>\log|\mathcal{A}|$. For any sequence of joint measures $\{P_{A^T,B^T}\}_{T=1}^\infty$, we have

$$\lim_{T\to\infty} E\left[\frac{1}{T}\vec{i}(A^T;B^T)\,1_{\{\frac{1}{T}\vec{i}(A^T;B^T)\geq L\}}\right]=0.$$

Proof: Let $\Omega=\{b^T : P(b^T)\leq 2^{-TL}\}$. Note that $\frac{1}{P(B^T)}\geq\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}$ $P_{A^T,B^T}-$a.s. Now

$$\begin{aligned}
&E\left[\frac{1}{T}\vec{i}(A^T,B^T)1_{\{\frac{1}{T}\vec{i}(A^T;B^T)\geq L\}}\right]\\
&\quad=E\left[\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}1_{\left\{\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\geq L\right\}}\right]\\
&\quad\leq E\left[\frac{1}{T}\log\frac{1}{P(B^T)}\,1_{\left\{\frac{1}{T}\log\frac{1}{P(B^T)}\geq L\right\}}\right].
\end{aligned}$$

By Lemma A.2, this last upper bound goes to zero as $T\to\infty$. □

Proof of Lemma 4.1: The second inequality is obvious. To prove the first inequality, note that $\forall\epsilon>0$, we have

$$\begin{aligned}
&\frac{1}{T}I(A^T\to B^T)\\
&\quad\geq E\left[\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}1_{\left\{\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\leq 0\right\}}\right]\\
&\qquad+0\times P\left[0\leq\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\leq\underline{I}(A\to B)-\epsilon\right]\\
&\qquad+(\underline{I}(A\to B))P\\
&\qquad\times\left[\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\geq\underline{I}(A\to B)-\epsilon\right].
\end{aligned}$$

The first addend goes to zero by Lemma A.3, the second addend equals zero, and the probability in the last addend goes to 1. Thus, for T large enough, $\frac{1}{T}I(A^T\to B^T)\geq\underline{I}-2\epsilon$. Since ϵ is arbitrary, we see that $\underline{I}(A\to B)\leq\liminf_{T\to\infty}\frac{1}{T}I(A^T\to B^T)$.

Now we treat the last inequality. For any $\epsilon>0$, we have

$$\begin{aligned}
&\frac{1}{T}I(A^T\to B^T)\\
&\quad\leq E\left[\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}1_{\left\{\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\geq L\right\}}\right]\\
&\qquad+LP\left[L\geq\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\geq\bar{I}(A\to B)+\epsilon\right]\\
&\qquad+(\bar{I}(A\to B)+\epsilon)P\\
&\qquad\times\left[\frac{1}{T}\log\frac{\vec{p}(B^T\,|\,A^T)}{P(B^T)}\leq\bar{I}(A\to B)+\epsilon\right].
\end{aligned}$$

The first addend goes to zero by Lemma A.4, the second addend goes to zero by definition of $\bar{I}$, and the probability in the last addend goes to 1. Thus, for T large enough, $\frac{1}{T}I(A^T\to B^T)\leq\bar{I}+2\epsilon$. Since ϵ is arbitrary, we see that $\limsup_{T\to\infty}\frac{1}{T}I(A^T\to B^T)\leq\underline{I}(A\to B)$. □

C. Lemma 7.7

We repeat the statement of Lemma 7.7 for convenience.

Lemma 7.7: For $|\mathcal{B}|$ finite, we have:

1) The cost is bounded and continuous; specifically, $0\leq\bar{c}(u)\leq\log|\mathcal{B}|,\forall u\in\mathcal{U}$;
2) the control constraint function $\mathcal{U}(\gamma)$ is a continuous set-valued map between $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ and $\mathcal{U}$;
3) the dynamics $r(d\gamma_{t+1}\,|\,\gamma_t,u_t)$ is continuous.

Proof: To prove part 1), recall $\bar{c}(u)=\int r(db|\pi,a)\,u(d\pi,da)\log\frac{r(b\,|\,\pi,a)}{\int r(b\,|\,\tilde{\pi},\tilde{a})u(d\tilde{\pi},d\tilde{a})}$. This $\bar{c}(u)$ corresponds to a mutual information with input distribution $u(d\pi,da)$ and an output B in a finite alphabet $\mathcal{B}$. Hence, $\bar{c}(u)\leq\log|\mathcal{B}|\,\forall u\in\mathcal{U}$. The cost is clearly continuous in $u\in\mathcal{U}$.

To prove part 2), recall $\mathcal{U}(\gamma)=\{u(d\pi,da) : u(d\pi,da)\in\mathcal{U},u(d\pi)=\gamma(d\pi)\}$. The set $U(\gamma)$ is compact for each $\gamma\in$

$\mathcal{P}(\mathcal{P}(\mathcal{S}))$. For any set $H \subset \mathcal{U}$, denote $\mathcal{U}^{-1}(H) = \{\gamma : U(\gamma) \cap H \neq \emptyset\}$. The set-valued map $\mathcal{U}(\gamma)$ is *continuous* if it is both:

1) upper semicontinuous (usc): $\mathcal{U}^{-1}(F)$ is closed in $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ for every closed set $F \subset \mathcal{U}$;
2) lower semicontinuous (usc): $\mathcal{U}^{-1}(G)$ is open in $\mathcal{P}(\mathcal{P}(\mathcal{S}))$ for every open set $G \subset \mathcal{U}$.

The control constraint $U(\gamma)$ is clearly both usc and lsc and hence is continuous.

To prove part 3), recall (33)

$$r(d\gamma_{t+1} \,|\, \gamma_t, u_t) = \int_{\mathcal{U},\mathcal{A},\mathcal{B}} \delta_{\{\Phi_\Gamma(u_t,b_t)\}}(d\gamma_{t+1})\, r(db_t \,|\, \pi_t, a_t)\, u_t(d\pi_t, da_t).$$

Since this stochastic kernel does not depend on γ_t, we only need to show that it is continuous in u_t. Specifically, let v be any continuous bounded function on $\mathcal{P}(\mathcal{P}(\mathcal{S}))$. We need to show

$$\int v\left(\Phi_\Gamma(u,b)\right) r(db \,|\, \pi, a)\, u(d\pi, da) \tag{A1}$$

is a continuous function of u_t.

By (25), we know for all Borel measurable $\Omega \subset \mathcal{P}(\mathcal{S})$

$$\gamma[u,b](\Omega) = \int \{\Phi_\Pi(\pi, a, b) \in \Omega\} r\left(d\pi, da \,|\, u, b\right). \tag{A2}$$

By Lemma 7.1, we know that for any Borel measurable $\Theta \subset \mathcal{P}(\mathcal{S})$, a, b, and u, we have

$$r(\Theta, a \,|\, u, b) = \frac{\int_\Theta r(b \,|\, \tilde{\pi}, a)\, u(d\tilde{\pi}, a)}{\int r(b \,|\, \tilde{\pi}, \tilde{a})\, u(d\tilde{\pi}, d\tilde{a})} \tag{A3}$$

when the denominator does not equal zero. Because $\mathcal{B}$ is finite and by repeated use of Theorem A.4, we see that (A3) is continuous in u, b for all Θ. This implies (A2) is continuous in u, b for all Ω, thus, implying (A1) is continuous in u. □

ACKNOWLEDGMENT

The authors would like to thank V. Borkar for many helpful discussions.

REFERENCES

[1] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. Ghost, and S. Marcels, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, Mar. 1993.

[2] D. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

[3] D. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete Time Case*. New York: Academic, 1978.

[4] D. Blackwell, L. Breiman, and A. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 18, pp. 1209–1220, 1958.

[5] M. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, 1976.

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

[7] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.

[8] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.

[9] I. Csiszár, "Arbitrarily varying channels with general alphabets and states," *IEEE Trans. Inf. Theory*, vol. 38, no. 6, pp. 1725–1742, Nov. 1992.

[10] A. Das and P. Narayan, "Capacities of time-varying multiple-access channels with side information," *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 4–25, Jan. 2002.

[11] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*. Boston, MA: Jones and Bartlett, 1993.

[12] R. Dobrushin, "Transmission of information in channels with feedback," *Theory Prob. Appl. 3*, vol. N4, pp. 395–412, 1958.

[13] R. Dobrushin, "A general formulation of the basic Shannon theorem in information theory," *Uspekhi Math. Nauk.*, vol. 14, no. 6, pp. 3–103, 1959.

[14] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory*, vol. 4, pp. 2–22, 1954.

[15] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[16] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.

[17] T. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[18] O. Hernandez-Lerma, *Adaptive Markov Control Processes*. New York: Springer-Verlag, 1989.

[19] O. Hernandez-Lerma and J. Lasserre, *Markov Chains and Invariant Probabilities*. Basel, Switzerland: Birkhauser, 2003.

[20] G. Kramer, *Directed Information for Channels with Feedback*, ser. ETH Series in Information Processing. Konstanz, Switzerland: Hartung-Gorre Verlag, 1998, vol. 11.

[21] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.

[22] H. Marko, "The bidirectional communication theory—A generalization of information theory," *IEEE Trans. Commun.*, vol. COM-21, no. 12, pp. 1345–1351, Dec. 1973.

[23] J. Massey, "Causality, feedback, and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, 1990, pp. 303–305.

[24] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.

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

[26] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, Jul./Oct. 1948.

[27] C. Shannon, "The zero error capacity of a noisy channel," *Inst. Radio Eng. Trans. Inf. Theory*, vol. IT-2, pp. S8–S19, Sep. 1956.

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

[29] C. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, J. Neyman, Ed., Berkeley, CA, 1961, vol. 1, pp. 611–644.

[30] W. Stout, *Almost Sure Convergence*. New York: Academic, 1974.

[31] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Electr. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, Aug. 2000.

[32] S. Tatikonda and S. Mitter, "Channel coding with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2001, p. 126.

[33] S. Tatikonda, "A Markov decision approach to feedback channel capacity," in *Proc. 44th IEEE Conf. Decision Control*, Dec. 2005, pp. 3213–3218.

[34] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[35] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 761–771, Mar. 1999.

[36] H. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proc. IEEE*, vol. 59, no. 11, pp. 1557–1566, Nov. 1971.

[37] H. Witsenhausen, "On policy independence of conditional expectations," *Inf. Control*, vol. 28, pp. 65–75, 1975.

[38] J. Wolfowitz, *Coding Theorems of Information Theory*. Berlin, Germany: Prentice-Hall, 1961.

[39] S. Yang, A. Kavcic, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.

[40] S. Yang, A. Kavcic, and S. Tatikonda, "On the feedback capacity of power constrained Gaussian channels with memory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 929–954, Mar. 2007.

Sekhar Tatikonda (S'92–M'00) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 2000.

From 2000 to 2002, he was a Postdoctoral Fellow in the Computer Science Department, University of California, Berkeley. He is currently an Associate Professor of Electrical Engineering, Yale University, New Haven, CT. His research interests include communication theory, information theory, stochastic control, distributed estimation and control, statistical machine learning, and inference.

Sanjoy Mitter (M'68–SM'77–F'79–LF'01) received the Ph.D. degree from the Imperial College of Science and Technology, London, U.K., in 1965.

He joined the Massachusetts Institute of Technology (MIT), Cambridge, in 1969, where he has been a Professor of Electrical Engineering since 1973. He was the Director of the MIT Laboratory for Information and Decision Systems from 1981 to 1999. He was also a Professor of Mathematics at the Scuola Normale, Pisa, Italy, from 1986 to 1996. He has held visiting positions at Imperial College, London, U.K.; University of Groningen, Groningen, The Netherlands; INRIA, Paris, France; Tata Institute of Fundamental Research, Tata, India; and ETH, Zürich, Switzerland. He was the McKay Professor at the University of California, Berkeley, in March 2000, and held the Russell-Severance-Springer Chair in fall 2003. His current research interests are communication and control in networked environments, the relationship of statistical and quantum physics to information theory, and control and autonomy and adaptiveness for integrative organization.

Dr. Mitter is a Member of the National Academy of Engineering, winner of the 2000 IEEE Control Systems Award, and winner of the Richard E. Bellman Control Heritage Award in 2007.