Capacity of Markov Channels with Partial State Feedback[1]

Serdar Yüksel
Dept. of Electrical Engineering
Yale University
New Haven, CT 06511
serdar.yuksel@yale.edu

Sekhar Tatikonda
Dept. of Electrical Engineering
Yale University
New Haven, CT 06511
sekhar.tatikonda@yale.edu

***Abstract*— We study the capacity of Markov channels with causal deterministic partial (quantized) state feedback. We assume the feedback channel to be memoryless, the channel state process to be Markovian, belong to a finite set, and the state and observation transitions to satisfy some general mixing conditions. For such channels, we obtain a single-letter characterization for the capacity with feedback. We further show that for every $\epsilon > 0$, there exists a finite length memory (sliding) encoder structure that leads to an $\epsilon-$optimal capacity; hence practically optimal performance can be achieved. We show that the non-linear filter generating the conditional state density provides the sufficient statistic for the optimal coding scheme.**

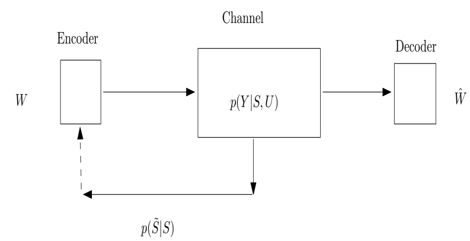

Fig. 1: Communication with Partial State Feedback.

I. Introduction and Literature Review

This paper studies the capacity problem of Markovian channels with partial state feedback. It is known that state feedback can increase the capacity of channels with memory. In practical situations, though, one may not be able to feedback complete state information. In this paper we examine the capacity of such Markovian channels with partial state information at the transmitter. In addition a finite memory encoder is presented that is ϵ-optimal.

Capacity with partial information at the transmitter is related to the problem of coding with unequal side information at the encoder and the decoder. The capacity of memoryless channels with various cases of side information being available at neither, either or both the transmitter and receiver have been studied in [3] and [4]. Tatikonda and Mitter [1] studied the capacity of channels with memory and complete noiseless output feedback and introduced a stochastic control formulation for the computation via the properties of the directed mutual information.

The case where there is some deterministic information available at the transmitter has been studied in [5], which also studied the case of a non-i.i.d. channel, however, the analysis has limiting assumptions on the available partial information in that the encoder is only allowed to use the most recent observation symbol for encoding. Reference GoldsmithVaraiya considered fading channels with side information at the transmitter. Viswanathan [15] studied the capacity of Markovian channels with delayed feedback and reference [9] studied the capacity of Markov channels with noiseless causal state feedback. Reference [7] considered the feedback capacity of finite state machine channels. See also [11], [6] and [12] for partial side information at the transmitter under different settings.

In sections 5 and 8 of [1], partial feedback is discussed, along with a discussion when sufficient statistics for encoding is computable by the output feedback. This paper builds on this work by examining the case of partial state feedback.

Below we state the main contributions of the paper:

- We obtain the capacity for Markovian channels with quantized state feedback.
- We show that, under some general mixing conditions, for any target ϵ, there exists a finite length buffer encoder and decoder achieving an $\epsilon-$optimal capacity. Thus, a finite length sliding block coder is almost optimal.
- We show that, building on the findings in [1], under some mixing assumptions, the sufficient statistic for the coder is the conditional state estimate generated by a filter.

II. Preliminaries and Problem Formulation

A. *Filtering of Markovian Processes*

Consider a Markov chain $\{S_t, t = 1, 2, \dots\}$ which lives in state space $\mathcal{S}$ with a stationary transition matrix A such that $P(\tilde{S}_{t+1} = j | S_t = i) = A_{i,j}$. Suppose this Markovian process is observed through an observation process $\{\tilde{S}_t\}, \tilde{S}_t \in \tilde{\mathcal{S}}$. Consider $\{\tilde{S}_t\}$ is generated through a quantization process such that $\tilde{S}_t = Q(S_t), \quad t = 1, 2, \dots$. Such a process is called a *partially observed Markov process*. Let the conditional densities be defined as

$$\pi_t(i) := P(S_t = i | \tilde{S}_1^t), \quad i = 1, 2, \dots, |\mathcal{S}|$$

[1]This work was partially supported by the National Science Foundation under Grant ECCS-0547199.

and $\pi_t = \{\pi_t(1), \pi_t(2), \ldots, \pi_t(|\mathcal{S}|)\}$. Hence $\pi_t \in \mathcal{P}_{\mathcal{S}}$ denote the conditional density of the process. The evolution of the conditional density equation satisfies a Markov recursion:

$$\pi_{t+1}(\pi_t, \tilde{s}_{t+1})(s_{t+1}) = \frac{\sum \pi_t(s_t) p(s_{t+1}|s_t) \delta(\tilde{s}_{t+1}|s_{t+1})}{\sum \sum \pi_t(s_t) p(s_{t+1}|s_t) \delta(\tilde{s}_{t+1}|s_{t+1})}.$$

The process $\{S_t, \pi_t\} \in \mathcal{S} \times \mathcal{P}_{\mathcal{S}}$ is a jointly Markovian system. We first have a definition.

Definition 2.1: A finite state Markov chain with transition matrix A, is *strongly mixing* if all the off-diagonal entries of A are strictly positive.

The conditions for the weak convergence of $\{S_t, \pi_t\}$ to an invariant distribution has not been precisely understood, in particular for the cases where the observation sequence is a deterministic function of the state sequence and the observation transition matrix involves zeros, as in a quantization operation. If the state dynamics are strongly mixing then convergence to an invariant distribution is guaranteed [2]. However, the strong mixing condition can be too conservative. The positive recurrence of the Markov chain, though, is not strong enough for the existence of a unique invariant filter. For a counter-example see [13]. Towards the characterization of conditions leading to a convergent filter, we make a definition.

Definition 2.2: A matrix P is subrectangular if $P_{i,j} > 0, P_{m,n} > 0$ implies $P_{i,n}, P_{m,j} > 0$.

$$M_{i,j}(\tilde{s}) = \begin{cases} p_{ij} & \text{if } g(j) = \tilde{s}, \\ 0 & \text{otherwise.} \end{cases}$$

The following is basically due to Kaijser [13].

Lemma 2.1: Suppose the Markov chain $\{S_n\}$ is ergodic. Further, suppose that there exists a finite sequence $\tilde{s}_1^l$ such that the matrix product $\prod_{t=1}^{l} M(\tilde{s}_t)$, is a non-zero subrectangular matrix. Then π_t converges to some π weakly for all initial π_0. The convergence to the invariant filter is geometric in that, there exists a T_0 such that for every $m, d > T_0$ and sufficiently large t:

$$||p(s_t|\tilde{s}_{t-d}^{t-1}) - p(s_t|\tilde{s}_{t-m}^{t-1})||_{TV} \leq K\gamma^{|d-m|},$$

for some $K < \infty, |\gamma| < 1$.

B. Problem Formulation

We consider channels with Markovian state dynamics. Such a model is applicable in fading scenarios in wireless systems. Under the causality restriction on the partial side information available at the encoder and full side information available at the receiver, we aim to compute the capacity.

We now present some notation. For a vector v, $v_i^m = \{v_i, v_{i+1}, \ldots, v_m\}$ denotes a sequence. Following the usual convention, for a random variable, small letters denote a particular realization of the random variable, whereas capital letters denote the random variable.

The channel state is a Markov chain $\{S_t, t = 1, 2, \ldots\}$ in $\mathcal{S}$. The observation process $\tilde{S}$ is generated through a memoryless deterministic function (as in a quantizer) such that $\tilde{S}_t = Q(S_t), \quad t = 0, 1, \ldots$. The channel input, U lives in $\mathcal{U}$ and output Y takes valuns in $\mathcal{Y}$. The messages, W, belong to $\mathcal{W}$, and are uniformly distributed. The channel satisfies, $\forall t \geq 0$,

$$p(Y_t|U_1^t, S_0^t) = p(Y_t|U_t, S_t).$$

Finally, the receiver has access to full state information.

We now present the coder class:

An (N, M, ϵ) channel code over time horizon N consists of a sequence of code-functions $\{f_t, k \geq 1\}$, and a decoding function, g, such that

- The sequence of code-functions $\{f_t, k \geq 1\}$ are measurable with respect to $(W, \tilde{S}_0^{t-1}, U_1^{t-1})$ (with $\tilde{S}_0 = \tilde{s}_0$) and have range space in $\mathcal{U}$. Let $\mathcal{F}^T$ be the set of all such measurable maps. A channel-code function, f^T is an element of $\mathcal{F}^T$.
- A channel code is a sequence of code-functions denoted by $f^T[w], w \in \mathcal{W}$. At time t, with state feedback $\tilde{s}_0^{t-1}$, the channel encoder has output $f_t[w](\tilde{s}_0^{t-1}, u_1^{t-1})$. We note that there is no loss in restricting the domain to $\mathcal{W} \times \mathcal{S}^t$, since the past codefunction outputs can be reconstructed through this domain. Hence, the channel code output is $f_t[w](\tilde{s}_0^{t-1})$.
- A channel decoder function g which maps $(\mathcal{Y}^N, \mathcal{S}^N) \to \mathcal{W}$, such that the (average) probability of error satisfies:

$$P_e := \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} p(\hat{w} \neq w | W = w) \leq \epsilon,$$

where $|\mathcal{W}| = M$.

Definition 2.3: R is an $\epsilon-$achievable rate if for every $\delta > 0$, there exists, for all sufficiently large N, an (N, M, ϵ) code such that $\frac{1}{N} \log_2(M) \geq R - \delta$. R is achievable if it is $\epsilon-$achievable for all $\epsilon > 0$. The supremal value of all achievable rates is defined as the capacity.

Upon the description of the code and the channel we now proceed with the computation of the capacity. We take the following approach: We first provide a converse bound, which will lead to an achievable rate, through the information stability of the channel.

III. Capacity of Markov Channels with Quantized Feedback

A. Converse to the Channel Theorem

Consider a coding scheme such that $\epsilon \to 0$. We have via Fano's inequality,

$$H(W|Y^N; S^N) \leq h(p_e) + p_e \log_2(M),$$

where p_e is the probability of error. And,

$$H(W|Y^N; S^N) = \log_2(M) - I(W; Y^N, S^N)$$

Thus,

$$\log_2(M)/N \leq \frac{I(W; Y^N, S^N) + h(p_e)}{N(1 - p_e)}$$

Since $p_e \to 0$ as $N \to \infty$, we have

$$\limsup_{N \to \infty} \log_2(M)/N \leq \limsup_{N \to \infty} (1/N) I(W; Y^N, S^N),$$

We have,

$$
\begin{aligned}
&\frac{1}{N}I(W;Y^N,S^N)\\
=\ &\frac{1}{N}\sum I(W;Y_t,S_t|Y_1^{t-1},S_1^{t-1})\\
=\ &\frac{1}{N}\sum\Big(I(W,\tilde{S}_0^{t-1};Y_t,S_t|Y_1^{t-1},S_1^{t-1})\\
&-I(\tilde{S}_0^{t-1};Y_t,S_t|Y_1^{t-1},S_1^{t-1},W)\Big)\\
=\ &\frac{1}{N}\sum I(W,\tilde{S}_0^{t-1};Y_t,S_t|Y_1^{t-1},S_1^{t-1}) \qquad \text{(III.1)}\\
=\ &\frac{1}{N}\sum_{t=1}^{N}\Big(H(Y_t,S_t|Y_1^{t-1},S_1^{t-1})\\
&-H(Y_t,S_t|Y_1^{t-1},S_1^{t-1},W,\tilde{S}_0^{t-1})\Big)\\
=\ &\frac{1}{N}\sum_{t=1}^{N}\Big(H(Y_t,S_t|Y_1^{t-1},S_1^{t-1})\\
&-H(Y_t,S_t|Y_1^{t-1},S_1^{t-1},W,\tilde{S}_0^{t-1},f_t[W](\tilde{S}_0^{t-1}))\Big)\\
=\ &\frac{1}{N}\sum_{t=1}^{N}\Big(H(Y_t,S_t|Y_1^{t-1},S_1^{t-1})\\
&-H(Y_t,S_t|Y_1^{t-1},S_1^{t-1},W,\tilde{S}_0^{t-1},U_t)\Big)\\
\leq\ &\frac{1}{N}\sum_{t=1}^{N}\Big(H(Y_t,S_t|Y_1^{t-1},S_1^{t-1}) \qquad \text{(III.2)}\\
&-H(Y_t,S_t|Y_{t-1},S_{t-1},U_t)\Big)\\
=\ &\frac{1}{N}\sum_{t=1}^{N}H(Y_t|S_t,Y_1^{t-1},S_1^{t-1})-H(Y_t|S_t,U_t)
\end{aligned}
$$

Equation (III.1) follows from $H(\tilde{S}_t|S_t)=0$. The inequality (III.2) follows from the fact that conditioned on U_t and S_t, Y_t is independent of W and $\{S_t\}$ is Markov. Hence, we have

$$
\begin{aligned}
C\ \leq\ &\frac{1}{N}\sum_{t=1}^{N}\sup_{p_t(U_t|\tilde{S}_0^{t-1})}\Big(H(Y_t|S_t,Y_1^{t-1},S_1^{t-1})\\
&-H(Y_t|S_t,U_t)\Big)\\
=\ &\frac{1}{N}\sum_{t=1}^{N}\sup_{p_t(U_t|\tilde{S}_0^{t-1})}I(U_t;Y_t|S_t,Y_1^{t-1},S_1^{t-1}).
\end{aligned}
$$

It should be noted that

$$
\begin{aligned}
&I(U_t;Y_t|S_t,Y_1^{t-1},S_1^{t})\\
&=E_{Y_1^t,S_1^t}\left[\log_2\Big(\frac{p(Y_t|S_t,U_t)p(U_t|Y_1^{t-1},S_1^{t-1})}{\int p(Y_t|S_t,U_t)p(U_t|Y_1^{t-1},S_1^{t-1})}\Big)\right]\\
&=E_{Y_1^t,S_1^t}\left[\log_2\Big(\frac{p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}{\int p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}\Big)\right]\\
&=E_{Y_t,S_1^{t-1},S_t}\left[\log_2\Big(\frac{p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}{\int p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}\Big)\right]\\
&=E_{Y_t,\tilde{S}_0^{t-1},S_t}\left[\log_2\Big(\frac{p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}{\int p(Y_t|S_t,U_t)p(U_t|\tilde{S}_1^{t-1})}\Big)\right]\\
&=:E_{\tilde{S}_0^{t-1},S_t}[c(p(U_t|\tilde{S}_1^{t-1}),S_t)],
\end{aligned}
$$

where the integration involves terms of the form $p(U_t|\tilde{S}_0^{t-1})p(\tilde{S}_1^{t-1}|Y_1^{t-1},S_1^{t-1})$, which has been omitted for convenience in presentation.

Hence, we have simplified the cost as a function of the control action. In the following we formulate the problem as a Markov Decision Process. In particular, we define the action space, state space, transition function and the cost function. The MDP falls within the so-called partially observed Markov chain classification.

We have the state sequence as: $\{S_t, t\geq 0\}$. The control action is a measurable mapping from $\sigma(\tilde{S}_0^{t-1})$ to $\mathcal{P}_{\mathcal{U}}$, hence control actions take value in the space of probability distributions on $\mathcal{U}$.

The partially observed Markov chain can be transformed to a fully observed Markov chain as follows: Define a new (enlarged) state: $\bar{X}_t := p(S_t|\tilde{S}_0^{t-1}), \quad \bar{X}_t\in\mathcal{P}_{\mathcal{S}}$. It should be noted that, $\bar{X}_t$ is a controlled Markov chain, which is indeed control-free:

$$p(\bar{X}_t|\bar{X}_1^{t-1},S_1^{t-1},p(U_t|\tilde{S}_1^{t-1}))=p(\bar{X}_t|\bar{X}_{t-1},\tilde{S}_{t-1})$$

In particular, it follows the non-linear filter recursion provided in the introduction.

In the fully observed Markov chain formulation, the cost function can be expressed as:

$$c'(p(U_t|\tilde{s}_1^{t-1}),p(S_t|\tilde{s}_1^{t-1})).$$

We have now completed the MDP formulation.

Theorem 3.1: The optimal stationary solution, C^o, solves the following fixed-point equation:

$$
\begin{aligned}
C^o+w(\bar{x})\ =\ &\sup_{p(U|\bar{x})}\Big\{c'(\bar{x},p(U|\bar{x})\\
&+E_{\bar{x}'|p(U|\bar{x})}[w(\bar{x}')|\bar{x}]\Big\}
\end{aligned}
$$

In particular, there exists a solution to the above Bellman's equation.

We say a policy is stationary if the coder policy is measurable with respect to $\sigma(\bar{X}_t)$ (note that the past state values are not used). We say a policy is admissible if it is measurable with respect to $\sigma(\bar{X}_0^t)$. We have the following Lemma.

Lemma 3.1: There exists a stationary solution to the optimization problem. Furthermore, this solution is the optimal solution among all admissible coding policies.

Proof: Proof follows from the continuity properties of the transition matrix in the channel input distribution, and the fact that mutual information is non-negative and continuous. See [14] for general conditions for the existence of optimal stationary solutions. The fully observed state process has a unique invariant distribution. Hence the optimal solution over all admissible coding policies is achieved by a stationary policy [10]. ⋄

B. Direct Theorem

Our achievability proof follows from the information stability of the state process, via Feinstein's lemma [16], which in turn guarantees that the information density [1] concentrates around its mean, hence converges in probability. The receiver has perfect access to the information available at the transmitter, and thus has access to the codebook used at each time instant. Hence, a multiplexing-demultiplexing policy is possible. The new state process $p(S_t|\tilde{S}_0^{t-1})$ has a unique invariant distribution. Hence, under any stationary policy, the sample path costs

$$\frac{1}{N}\sum_{t=1}^{N} c'(p(U_t|\tilde{S}_1^{t-1}), p(S_t|\tilde{S}_1^{t-1}) \to E[c(p(U_t|\tilde{\pi}), \tilde{\pi}], \quad a.s.$$

where $\tilde{\pi}$ denotes the invariant distribution. The process $\tilde{\pi}_t$ lives in an infinite-dimensional state space. However, this space is Polish (complete, seperable, metric space), and hence is dense under some metric, such as the total variation norm. The elements in this space can be arbitrarily approximated by a countable subset in $\mathcal{P}_{\mathcal{S}}$.

Ideally, for every visit of the process to $\tilde{\pi}_t$, the encoder will use another distribution (multiplexing), and since this information is available at the decoder, the decoder could apply demultiplexing. The channel and the side information at the encoder will form a joint Markov process, for every realization of which, the encoder uses a separate random codebook. However, since the state space is non-atomic, that is the probability measure of any sample point is zero, we need to provide another construction. We have two options, one is to discretize the space of probability measures on $\mathcal{S}$ (which is possible, since the space has a countable dense subset), and the other one is to restrict the memory available at the encoder. In the former, the Markov recursion will not be satisfied. Hence, we pursue the second alternative, which is also more practical as it has a sliding block coder interpretation.

IV. Restriction of the Encoder to use Finite-Memory

In this section, we restrict the encoder to use finite-memory. We define $\tilde{\pi}_t^d = \tilde{S}_{t-d}^{t-1}$ as the finite dimensional information available at the transmitter with regard to the state.

The side information process and the channel process $\tilde{\pi}_t^d, S_t$ will have an ergodic distribution. This process evolves in the set $\tilde{\mathcal{S}}^d \times \mathcal{S}$. For every visit of the process to $\{\tilde{\pi}_t^d\}$ to any point in $\tilde{\mathcal{S}}^d$, the encoder will use a separate distribution.

With the number of channel uses growing, the normalized number of visits to any state, the long term occupation measure,

$$p_0(\tilde{s}_{t-d}^{t-1}) := (\frac{1}{N})\sum_{t=1}^{N} 1_{(\tilde{\pi}_t^d = \tilde{s}_{t-d}^{t-1})},$$

where $1_{(.)}$ is the indicator function, will be the same as the statistical average under the invariant distribution. We have, under the stationarity assumption, information stability. Hence, the information density will concentrate at the average (in probability). We thus have the following:

Theorem 4.1: Suppose the state process is a finite-state, irreducible and recurrent Markov process.Then, the following rate

$$C^d := \sum_{s,\tilde{\pi}^d} p(\tilde{\pi}^d) \sup_{p(U|\tilde{\pi}^d)} \{p(s|\tilde{\pi}^d) I(U;Y|s,\tilde{\pi}^d)\},$$

is achievable. Furthermore, there exists no rate larger than this which is achievable.

In the following we will study the process $\{C^d\}$.

Theorem 4.2: Suppose the channel and the observation dynamics are such that the filter process has an invariant distribution. For some $|\gamma| < 1$ and some $K < \infty$, and $m > d$ and some t_0 and some t_1, only a function of the transition matrix and the observation process, such that for all $m, d > t_0, m - d > t_1$, the channel capacities satisfy the following: $|C^d - C^m| \leq K\gamma^{|m-d|}$.

Proof: We provide a sketch of the proof. Clearly we have that $C^m \geq C^d$, since the optimization is over a larger space for memory length m. In the following we show that, for some sufficiently small ϵ, $C^d \geq C_m - \epsilon$. We have that

$$I(U_t;Y_t|S_t,\tilde{S}_{t-m+1}^{t-1}) = \sum_{\tilde{s}_{t-m+1}^{t-1}, s_t} p(\tilde{s}_{t-m+1}^{t-1}, s_t) H(Y_t|s_t,\tilde{s}_{t-m+1}^{t-1}) - \sum_{s_t,u_t} p(s_t,u_t) H(Y_t|s_t,u_t)$$

Further,

$$\sum_{\tilde{s}_{t-m+1}^{t-1}, s_t} p(\tilde{s}_{t-m+1}^{t-1}, s_t) H(Y_t|s_t,\tilde{s}_{t-m+1}^{t-1}) = \sum_{\tilde{s}_{t-d}^{t-1}, s_t} p(\tilde{s}_{t-d}^{t-1}, s_t) \Bigg(\sum_{\tilde{s}_{t-m}^{t-d-1}, s_t} p(\tilde{s}_{t-m}^{t-d-1}|\tilde{s}_{t-d}^{t-1}, s_t) H(Y_t|s_t,\tilde{s}_{t-m+1}^{t-1}) \Bigg) \leq \sum_{\tilde{s}_{t-d}^{t-1}, s_t} p(\tilde{s}_{t-d}^{t-1}, s_t) \Bigg(H(p'(Y_t|s_t,\tilde{s}_{t-d+1}^{t-1})) \Bigg), \qquad \text{(IV.1)}$$

where the inequality follows from the fact that averaging increases the entropy. Here, by an abuse of notation, $H(p'(.))$

denotes the conditional entropy with

$$p'(U_t|\tilde{s}_{t-d}^{t-1}) = \sum_{\tilde{s}_{t-m}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1}|\tilde{s}_{t-d}^{t-1})\delta(\tilde{s}_{t-d}^{t-1}) \quad \text{(IV.2)}$$

We now consider the second term:

$$\sum_{\tilde{s}_{t-m+1}^{t-1}, s_t} p(\tilde{s}_{t-m+1}^{t-1}, s_t)H(Y_t|s_t, u_t) \quad \text{(IV.3)}$$

We consider the change in the conditional density $p(u_t|s_t)$. For this we show that the joint density

$$p(U_t|s_t) = \sum_{\tilde{\pi}^d} p(U_t|\tilde{\pi}_t^d)p(\tilde{\pi}_t^d|s_t)$$

has a negligible variation in increasing memory length. Let $p^*(U_t|s_t)$ be an induced distribution generated via

$$\begin{aligned}
p^*(U_t|s_t) &= \sum_{\tilde{s}_{t-m}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1}|s_t) \\
&= \sum_{\tilde{s}_{t-m}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})p(s_t|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1})/p(s_t) \\
&= \sum_{\tilde{s}_{t-m}^{t-1}} \sum_{\tilde{s}_{t-d}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})p(s_t, \tilde{s}_{t-d}^{t-1}|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1})/p(s_t) \\
&= \sum_{\tilde{s}_{t-m}^{t-1}} \sum_{\tilde{s}_{t-d}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})p(s_t|\tilde{s}_{t-d}^{t-1}, \tilde{s}_{t-m}^{t-1}) \\
&\quad .p(\tilde{s}_{t-d}^{t-1}|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1})/p(s_t) \\
&\leq \sum_{\tilde{s}_{t-m}^{t-1}} \sum_{\tilde{s}_{t-d}^{t-1}} p(U_t|\tilde{s}_{t-m}^{t-1})(p(s_t|\tilde{s}_{t-d}^{t-1}) + \eta) \\
&\quad .p(\tilde{s}_{t-d}^{t-1}|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1})/p(s_t) \\
&= \sum_{\tilde{s}_{t-d}^{t-1}} p'(U|\tilde{s}_{t-d}^{t-1})p(\tilde{s}_{t-d}^{t-1}|s_t) + \eta \qquad \text{(IV.4)}
\end{aligned}$$

with

$$p'(U_t|\tilde{s}_{t-d}^{t-1}) = \sum_{\tilde{s}_{t-m}^{t-1}} p^*(U|\tilde{s}_{t-m}^{t-1})p(\tilde{s}_{t-m}^{t-1}|\tilde{s}_{t-d}^{t-1})\delta(\tilde{s}_{t-d}^{t-1}) \quad \text{(IV.5)}$$

Here, following Lemma 2.1, $\eta < K\gamma^{|m-d|}$. Likewise, one shows that there exist input distributions such that

$$p^*(U_t|s_t) + \eta > p'(U_t|s_t) \geq p^*(U_t|s_t) - \eta,$$

for some $\eta > 0$, which is a function of d, m, and the joint Markov system dynamics. Since entropy is continuous in $p(u|s)$, this ensures a geometrically diminishing variation. Equations (IV.1, IV.4) thus lead to the result. ⋄

V. General Theorem

Theorem 4.1 shows that the capacity sequence $\{C^n\}$, which takes values in R forms a Cauchy sequence. Hence there exists a limit. This limit value is arbitrarily closely achieved via a large but finite memory. We now show that this limiting distribution is related to the sufficient statistic. As shown in [1], the problem becomes that of finding a consistent measure induced by the entire data history, generated only through the sufficient statistic. We now present the general theorem.

Theorem 5.1: Suppose the state process is a finite-state, irreducible and recurrent Markov process, and the observation process $\tilde{S}$ is such that the joint process $(\tilde{\pi}_t, S_t) \in (\mathcal{P}_{\mathcal{S}} \times \mathcal{S})$ admits a unique invariant distribution. Then, the capacity of this channel is equal to $\lim_{d\to\infty} C^d = C$, where

$$C := \int_{s,\tilde{\pi}} p(\tilde{\pi}) \sup_{p(U|\tilde{\pi})} \{p(s|\tilde{\pi})I(U;Y|s,\tilde{\pi})\}.$$

VI. Concluding Remarks

We studied the capacity of channels with quantized (or deterministic partial) feedback. In particular, we showed that it is optimal to estimate the state and apply the coding based on this estimation (hence, seperation of estimation and control applies). The derivations here can also be applicable to Markov channels with input dependent dynamics. However, in those channels, the dual effect of the input will be present: Transmitting the message, and optimizing the stationary channel input distribution.

We note that the same capacity theorem applies to the case when the channel observations are noisy, but the receiver has access to this noise process. The case where the encoder information is not available at the decoder is currently under study, and it involves the understanding of a distributed agreement problem between the encoder and the decoder.

References

[1] S. Tatikonda and S. Mitter, "The Capacity of channels with feedback", submitted to *IEEE Trans. Information Theory*.

[2] F. LeGland and N. Oudjane, "Stability and uniform approximation of nonlinear filters using the Hilbert metric, and application to particle filters," *The Annals of Applied Probability*, 14: 144-187, Feb. 2004.

[3] C. Shannon, "Channels with side information at the transmitter," *IBM Journal Research and Development*, 2: 289-293, 1958.

[4] S. I. Gelfand and M. S. Pinsker, "Coding for Channel with Random Parameters," *Prob. of Cont. and Info. Theory*, 9 : 19-31, 1980.

[5] G. Caire and S. Shamai "On the capacity of some channels with channel state information", *IEEE Trans. Infomation Theory*, 45: 2007-2019, 1999.

[6] M. Salehi, "Capacity and coding for memories with real-time noisy defect information at encoder and decoder," *Proc. Inst. Elec. Eng. Pt.* 139:113-117, April 1992.

[7] S. Tatikonda, S. Yang and A. Kavcic, "Feedback capacity of finite-state machine channels," *IEEE Trans. Inform. Theory*, 799 - 810, 2005.

[8] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Info. Theory*, 43: 1986:1992, Nov. 1997.

[9] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Info. Theory*, 51: 780 - 789, 2005.

[10] V. S. Borkar, "Convex Analytic Methods in Markov Decision Processes," *Handbook of Markov Decision Processes: Methods and Applications*, Kluwer, Boston, MA.

[11] S. A. Jafar "Capacity with Causal and Non-Causal Side Information - A Unified View," *IEEE Trans. Info. Theory*, 52: 5468-5475, Dec. 2006.

[12] N. Jindal, "MIMO broadcast channels with finite rate feedback, IEEE Trans. on Inform. Theory, vol. 52, no. 11, pp. 50455059, Nov. 2006.

[13] T. Kaijser, "A limit theorem for partially observed Markov chains", *Annals of Probability* 4: 677: 696, 1975.

[14] O. Hernandez-Lerma, J. Gonzales-Hernandez, and R. R. Lopez-Martinez, "Constrained average cost Markov control processes in Borel spaces," *SIAM J. Con. and Opt.* vol. 42, pp. 442-468, 2003

[15] H. Viswanathan, "Capacity of Markov Channels with Receiver CSI and Delayed Feedback", *IEEE Trans. Inform. Theory* 45: 761-771, March 1999.

[16] S. Verdú and T. S. Han,"A general formula for channel capacity," *IEEE Trans. on Information Theory* 40: 1147-1157, July 1994.