The Capacity of Finite-State Markov Channels with Feedback

Jun Chen, *Student Member, IEEE*, Toby Berger[†], *Fellow, IEEE*

Abstract

We consider a class of finite-state Markov channels with feedback. After introducing a troduce a simplified equivalent channel model, we construct the optimal stationary and nonstationary input processes that maximize the long-term directed mutual information. Furthermore, we give a sufficient condition under which the channel's Shannon capacity can be achieved by a stationary input process. The corresponding converse coding theorem and direct coding theorem are proved.

Index Terms—Channel capacity, Markov channel, feedback, typicality.

I. INTRODUCTION

We study the capacity of a feedback channel whose state process can be affected by its input and whose state information is available at both the transmitter and the receiver. Our channel model is illustrated in Fig. 1. Were it not for the feedback, our channel would belong to the family of finite-state channels. The FSC literature is vast; see, for example, [1]–[3]. In [4], Verdu and Han gave a general capacity formula for channels without feedback. If in our model the state process were not affected by the input, i.e., $p(s_{k+1}|x_k, s_k) = p(s_{k+1}|s_k)$, the model would reduce to a special case in the general framework of [5] and [6].

The feedback channel coding problem goes back to early work by Shannon [7], Dobrushin [8] and Wolfowitz [9]. Tatikonda [10] introduced a model of feedback channels which can be viewed as a generalization of the formulation in [4], derived a general formula for the capacity of
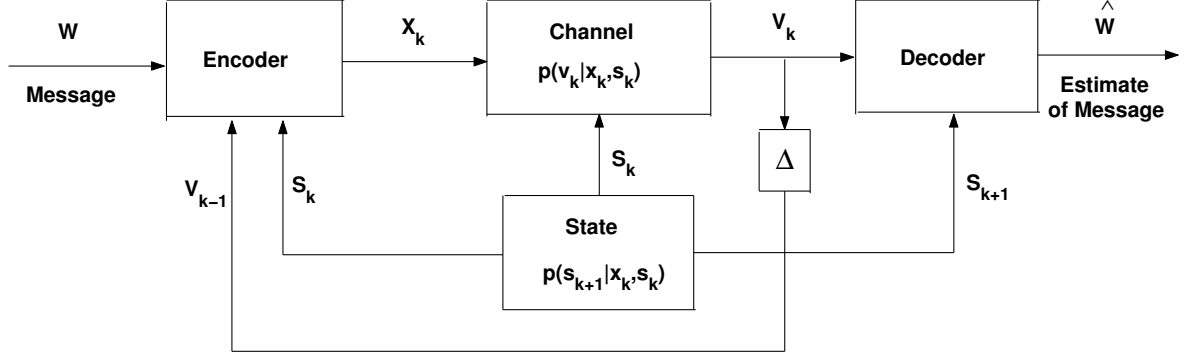
Fig. 1 Our Model

channels in this class, and used dynamic programming to compute the optimal input distribution. We show that channels descirbed by our model, which is called the Markov channel in [10], possess under certain conditions a relatively simple capacity formula and that the corresponding optimal input distribution can be computed with markedly less complexity than in the general case.

It's easy to see that if we let $Y_k = (V_k, S_{k+1})$ and $Q(y_k|x_k, y_{k-1}) = p(v_k|x_k, s_k)p(s_{k+1}|x_k, s_k)$, then Fig. 1 can be simplified to the model shown in Fig. 2. Therefore, we henceforth consider only this simplified channel model. The model in Fig. 2 was perhaps first introduced in [11],
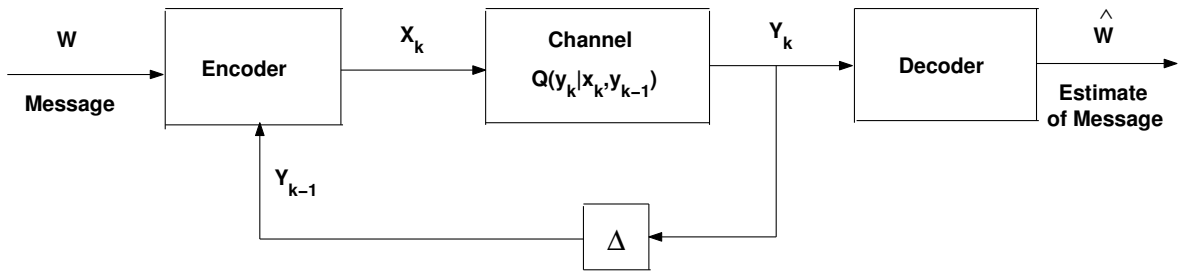
Fig. 2 An Equivalent Model

[12]. Ying and Berger [13] analyzed the capacity of this channel model when the output is binary. In this paper we will give a more general treatment.

The rest of this paper is divided into eight sections. In Section II we introduce several basic notations and definitions. In Section III we prove the converse channel coding theorem for our model, which provides an upper bound on the achievable rate of information transmission

through the channel. Then we give a recursive formula to calculate the maximal directed mutual information in Section IV. In Section V we analyze the optimum stationary input distribution that maximizes the long-term directed mutual information. We generalize in Section VI to analyze the optimum not-necessarily-stationary input. A sufficient condition under which the optimum stationary input is actually optimum among all the input distributions is given in Section VII. We prove the direct channel coding theorem and suggest a coding scheme in Section VIII. Finally, several directions to extend our results are discussed in Section IX which serves as a conclusion.

II. PRELIMINARIES

A. Notation

We assume throughout the paper that the channel input and output alphabets both are finite . Without loss of generality, we let $x \in \{1, 2, \cdots, N_x\}$ and $y \in \{1, 2, \cdots, N_y\}$.

B. Code Description

An (n, M, ε, y_0) feedback code for our channel consists of

1) An encoding function f that maps the set of messages $\mathcal{W} = \{1, \cdots, M\}$ to channel input words of blocklength n through a sequence of functions $\{f_{y_0,k}\}_{k=1}^{n}$ that depend only on the message W and the channel outputs up to time $k - 1$, i.e.,

$$X_k = f_{y_0,k}(W, Y_1^{k-1}). \tag{1}$$

Although it might seem to be more general to let

$$X_k = \tilde{f}_{y_0,k}(W, X_1^{k-1}, Y_1^{k-1}),$$

this actually is equivalent to (1), as shown by the following argument:

$$
\begin{aligned}
X_1 &= \tilde{f}_{y_0,1}(W) = f_{y_0,1}(W), \\
X_2 &= \tilde{f}_{y_0,2}(W, X_1, Y_1) = \tilde{f}_{y_0,2}(W, \tilde{f}_{y_0,1}(W), Y_1) = f_{y_0,2}(W, Y_1).
\end{aligned}
$$

It follows easily by induction that $X_k = \tilde{f}_{y_0,k}(W, X_1^{k-1}, Y_1^{k-1}) = f_{y_0,k}(W, Y_1^{k-1})$, so (1) is of full generality.

2) A decoding function g_{y_0} that maps a received sequence of n channel outputs to the message set $g_{y_0} : \mathcal{Y}_1^n \rightarrow \mathcal{W}$ such that the average probability of decoding error satisfies

$$P_e \triangleq \frac{1}{M} \sum_{w=1}^{M} P(\hat{W} \neq w | W = w, Y_0 = y_0) \leq \varepsilon, \text{ where } \hat{W} = g_{y_0}(Y_1^n). \qquad (2)$$

Note: The encoding function f_{y_0} and decoding function g_{y_0} both depend on the initial channel state y_0.

Definition 1: R_{y_0} is an ε-achievable rate given the initial state y_0 if for every $\delta > 0$ there exists, for all sufficiently large n, an (n, M, ε, y_0) code such that $\frac{1}{n} \log M \geq R_{y_0} - \delta$. R_{y_0} is achievable if it is ε-achievable for all $\varepsilon > 0$. The supremum of all achievable rates R_{y_0} is defined as the feedback capacity $C_{y_0}^{fd}$ given the initial state y_0.

III. CONVERSE CHANNEL CODING THEOREM

This section is devoted to the proof of the converse channel coding theorem.

Theorem 1(converse channel coding theorem): Given the initial state y_0, information transmission with an arbitrary small expected frequency of errors is not possible if $R > \limsup\limits_{n \rightarrow \infty} \frac{C_{y_0,n}}{n}$.

Here

$$C_{y_0,n} = \max_{p(X_1^n) \in \mathcal{P}^*(X_1^n)} [I(X_1; Y_1 | Y_0 = y_0) + \sum_{k=2}^{n} I(X_k; Y_k | Y_{k-1})],$$

and $\mathcal{P}^*(X_1^n)$ is the set of input distributions on X_1^n which consists of all the probability mass functions that satisfy

$$p(X_k | X_1^{k-1}, Y_0^{k-1}) = p(X_k | Y_{k-1}), \ k = 1, 2, \cdots, n.$$

Proof:

In the proof we implicitly assume that $P(Y_0 = y_0) = 1$ and thus use Y_0 instead of y_0.

Let W be the message random variable. For any (n, M, ε, y_0) code, by Fano's inequality

$$H(W | Y_0^n) \leq h(P_e) + P_e \log M. \qquad (3)$$

Since

$$H(W | Y_0^n) = H(W) - I(W; Y_0^n) = \log M - I(W; Y_0^n),$$

we have

$$(1 - P_e) \log M \leq h(P_e) + I(W; Y_0^n),$$

which we rewrite as

$$\frac{1}{n}\log M \leq \frac{h(P_e) + I(W; Y_0^n)}{n(1 - P_e)}.$$ (4)

As $n \to \infty$, $P_e \to 0$. Hence the channel capacity

$$\begin{aligned}
C_{y_0}^{fd} &= \limsup_{n\to\infty} \frac{1}{n}\log M \\
&\leq \limsup_{n\to\infty} \max_{p(X_1^n)} \frac{1}{n} I(W; Y_0^n).
\end{aligned}$$ (5)

We have

$$\begin{aligned}
I(W; Y_0^n) &= H(Y_0^n) - H(Y_0^n|W) \\
&= \sum_{k=1}^{n}[H(Y_k|Y_0^{k-1}) - H(Y_k|Y_0^{k-1}, W)] \\
&\leq \sum_{k=1}^{n}[H(Y_k|Y_{k-1}) - H(Y_k|Y_0^{k-1}, X_k, W)] \\
&\stackrel{(a)}{=} \sum_{k=1}^{n}[H(Y_k|Y_{k-1}) - H(Y_k|Y_{k-1}, X_k)] \\
&= \sum_{k=1}^{n} I(X_k; Y_k|Y_{k-1}),
\end{aligned}$$ (6)

where (a) holds because, when conditioned on the input X_k and the previous output Y_{k-1} (i.e., the current channel state), the channel output Y_k becomes independent of both the message W and the earlier outputs Y_0^{k-2} .

We call $\sum_{k=1}^{n} I(X_k; Y_k|Y_{k-1})$ the directed mutual information. The concept of directed mutual information was introduced by Massey [14] who attributes it to Marko [15]. See [10] for a detailed discussion of this concept. It has been shown in [13], that the maximum directed mutual information for our channel model is attained inside $\mathcal{P}^*(X_1^n)$; i.e., no loss of generality results from restricting $p(X_1^n)$ to $\mathcal{P}^*(X_1^n)$ when maximizing the directed information.

So we have

$$
\begin{aligned}
C_{y_0}^{fd} & \leq \quad \limsup_{n\to\infty} \max_{p(X_1^n)} \frac{1}{n} I(W; Y_0^n) \\
& \leq \quad \limsup_{n\to\infty} \max_{p(X_1^n)} \frac{1}{n} \sum_{k=1}^{n} I(X_k; Y_k | Y_{k-1}) \\
& = \quad \limsup_{n\to\infty} \max_{p(X_1^n)\in\mathcal{P}^*(X_1^n)} \frac{1}{n} \sum_{k=1}^{n} I(X_k; Y_k | Y_{k-1}) \\
& = \quad \limsup_{n\to\infty} \frac{C_{y_0,n}}{n} \qquad\qquad\qquad \blacksquare
\end{aligned}
$$

We remark that, although $\limsup\limits_{n\to\infty} \frac{C_{y_0,n}}{n}$ is an upper bound on the achievable rate, it is not always tight. Consider, for example, Fig. 3 in which the transition probability associated with
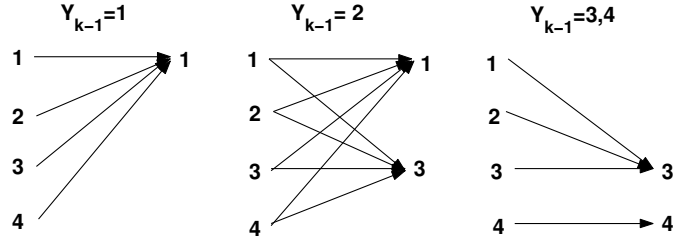

Fig. 3 Example 1

every arrow in the middle figure is $\frac{1}{2}$. It is apparent that

1) If $Y_0 = 1$, we can transmit no information through this channel.

2) If $Y_0 = 3$ or 4, we can transmit 1 bit of information per channel use.

3) If $Y_0 = 2$, then $Y_1 = 1$ with probability $\frac{1}{2}$, whereupon $Y_k = 1$ for all k. Then half of the time $Y_1 = 1$, in which case $Y_k = 1$ for all k. The other half time $Y_1 = 3$, in which case we can transmit 1 bit of information per channel use after that.

According to Definition 1, the capacity of the channel of Figure 3 is 0 if the initial channel state $Y_0 = 2$. However, one readily can compute that $\limsup\limits_{n\to\infty} \frac{C_{2,n}}{n} = \frac{1}{2}$, an equal mixture of the channel capacity for $Y_0 = 1$ and that for $Y_0 \in \{3, 4\}$.

IV. RECURSIVE FORMULA FOR THE MAXIMUM DIRECTED MUTUAL INFORMATION

We mentioned in Section III that the maximum directed mutual information of our channel model is attained inside $\mathcal{P}^*(X_1^n)$. This not only greatly simplifies the structure of the input

distribution that maximizes the directed mutual information, but also makes the joint (input, output) process possess a Markov structure, as described by the following lemma.

Lemma 1 [11], [13]: If we restrict the distribution of input X_1^n to $\mathcal{P}^*(X_1^n)$, then we have

1) $\{Y_k, k = 0, 1, \cdots, n\}$ is a first-order Markov chain,
2) $\{(X_k, Y_k), k = 1, 2, \cdots, n\}$ also is a first-order Markov chain.

Lemma 1 evidences how the underlying Markov structure in our channel model allows us to bring to bear on the problem at hand powerful techniques from Markov theory and dynamic programming. This is partially reflected in the following theorem.

Theorem 2:

$$C_{i,n} = \max_{\vec{p}_{i,n}} \{ I(\vec{p}_{i,n}, Q_i) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}) C_{j,n-1} \} \quad \text{for any} \quad i = 1, 2, \cdots, N_y. \tag{7}$$

where

1) $\vec{p}_{i,n}$ is the distribution of the first input when i is the initial state. (The inclusion of n in the subscript is intended to stress that this distribution generally depends on n; we emphasize that $\vec{p}_{i,n}$ is not the input distribution at time n.)
2) $C_{i,n} = \max\limits_{p(X_1^n) \in \mathcal{P}^*(X_1^n)} \{ \sum\limits_{k=2}^{n} I(X_k; Y_k | Y_{k-1}) + I(X_1; Y_1 | Y_0 = i) \}$;
3) Q_i is the channel transition probability matrix for state i, i.e., $Q_i(j, l) = Q(Y_k = j | X_k = l, Y_{k-1} = i)$;
4) $T_{i,j}(\vec{p}_{i,n}) = \sum\limits_{l=1}^{N_x} \vec{p}_{i,n}(l) Q_i(j, l)$, where $\vec{p}_{i,n}(l)$ is the lth component of the probability vector $\vec{p}_{i,n}$;
5) $C_{i,0} = 0$ for $i = 1, 2, \cdots, N_y$.

Proof: We prove this theorem by induction.

It's obviously true when $n = 1$. $\tag{8}$

Suppose it's true when $n = 1, \cdots, m - 1$. $\tag{9}$

Then

$$C_{i,m} = \max_{p(X_1^m) \in \mathcal{P}^*(X_1^m)} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m} I(X_k; Y_k|Y_{k-1})\}$$

$$= \max_{p(X_1^m) \in \mathcal{P}^*(X_1^m)} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-1} I(X_k; Y_k|Y_{k-1}) + I(X_m; Y_m|Y_{m-1})\}$$

$$= \max_{p(X_1^m) \in \mathcal{P}^*(X_1^m)} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-1} I(X_k; Y_k|Y_{k-1})$$
$$+ \sum_{j=1}^{N_y} P(Y_{m-1} = j)I(X_m; Y_m|Y_{m-1} = j)\}$$

$$\overset{(b)}{=} \max_{p(X_1^{m-1}) \in \mathcal{P}^*(X_1^{m-1})} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-1} I(X_k; Y_k|Y_{k-1}) + \sum_{j=1}^{N_y} P(Y_{m-1} = j)C_{j,1}\} \quad (10)$$

$$= \max_{p(X_1^{m-1}) \in \mathcal{P}^*(X_1^{m-1})} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-2} I(X_k; Y_k|Y_{k-1})$$
$$+ I(X_{m-1}; Y_{m-1}|Y_{m-2}) + \sum_{j=1}^{N_y} P(Y_{m-1} = j)C_{j,1}\}$$

$$= \max_{p(X_1^{m-1}) \in \mathcal{P}^*(X_1^{m-1})} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-2} I(X_k; Y_k|Y_{k-1})$$
$$+ \sum_{l=1}^{N_y} P(Y_{m-2} = l)[I(X_{m-1}; Y_{m-1}|Y_{m-2} = l) + \sum_{j=1}^{N_y} P(Y_{m-1} = j|Y_{m-2} = l)C_{j,1}]\} \quad (11)$$

$$\overset{(c)}{=} \max_{p(X_1^{m-2}) \in \mathcal{P}^*(X_1^{m-2})} \{I(X_1; Y_1|Y_0 = i) + \sum_{k=2}^{m-2} I(X_k; Y_k|Y_{k-1}) + \sum_{l=1}^{N_y} P(Y_{m-2} = l)C_{l,2}\} \quad (12)$$

$$= \cdots$$

$$\overset{(d)}{=} \max_{p(X_1) \in \mathcal{P}^*(X_1)} \{I(X_1; Y_1|Y_0 = i) + \sum_{j=1}^{N_y} P(Y_1 = j)C_{j,m-1}\}$$

$$= \max_{\vec{p}_{i,m}} \{I(\vec{p}_{i,m}, Q_i) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,m})C_{j,m-1}\},$$

where

(b) holds because $I(X_m; Y_m|Y_{m-1} = j)$ depends only on $p(X_m|Y_{m-1} = j)$ and its value does not affect the remaining part of (10), so we can maximize it greedily. The result of maximization follows from (8).

(c) holds because the quantity in the square brackets in (11) depends only on $p(X_{m-1}|Y_{m-2} = l)$ and its value does not affect the remaining part of (11), so we can maximize it greedily. The result of said maximization follows from the induction hypothesis (9), and

(d) holds via the same line of reasoning as in (11) and (12). ∎

Theorem 2 shows that we can compute $\{C_{i,n}, i = 1, 2, \cdots, N_y\}$ in the following recursive way:

Use (7) to find $\{C_{i,1}, i = 1, 2, \cdots, N_y\}$ and their corresponding $\{\vec{p}_{i,1}^{\,*}, i = 1, 2, \cdots, N_y\}$. Put $\{C_{i,1}, i = 1, 2, \cdots, N_y\}$ into (7) to find $\{C_{i,2}, i = 1, 2, \cdots, N_y\}$ and their corresponding $\{\vec{p}_{i,2}^{\,*}, i = 1, 2, \cdots, N_y\}$. Repeat until we get $\{C_{i,n}, i = 1, 2, \cdots, N_y\}$ and their corresponding $\{\vec{p}_{i,n}^{\,*}, i = 1, 2, \cdots, N_y\}$ (Note: The maximum value of a continuous function over a compact set always exists, though $\vec{p}_{i,k}^{\,*}, (i = 1, 2, \cdots, N_y; k = 1, 2, \cdots, n)$ need not be unique in general).

Since $Q_i(i = 1, 2, \cdots, N_y)$ are fixed, for simplicity we henceforth abbreviate $I(\vec{p}, Q_i)$ as $I_i(\vec{p})$.

V. Optimum Stationary Input Distribution

From Theorem 2 we see that the optimum input distribution generally depends on time, which significantly complicates the problem. In this section we restrict our attention to input distributions that depend only on the current channel state (i.e., the previous channel output) but do not depend on time. We call these the stationary [1] input distributions, and we let $\mathcal{P}^{**}$ denote the set of such distributions. Correspondingly, $\mathcal{P}^{*}$ defined in Section III is the set of nonstationary input distributions. Let $\vec{p}_i$ denote the input distribution when the channel state is i. Then no loss in generality results from writing an element of P^{**} as $(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})$. It's clear that, if we restrict the input distribution into P^{**}, then $\{Y_k, k = 0, 1, \cdots\}$ is a homogeneous Markov chain.

When the input distribution is stationary, we can easily find the following recursive formula:

$$I_{i,n} = I_i(\vec{p}_i) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_i) I_{j,n-1}, \tag{13}$$

---

[1] Here, the term "stationary" does not have its usual connotation in the theory of random processes. Specifically, since we do not rule out initial conditions that cause the state process's marginals to undergo a transient, similar transient behavior may well be exhibited by the marginals of the input process. What motivates our use of the term "stationary" is that, if the channel satisfies certain conditions (which will be made clear in Section VI) and the input distribution is "stationary" under our definition, then the joint input and output process (i.e., the joint input and state process) forms an irreducible and aperiodic homogeneous Markov chain and thus is asymptotically mean stationary in the sense of [16], [17].

where

1) $T_{i,j}(\vec{p_i}) = \sum_{l=1}^{N_x} \vec{p_i}(l)Q_i(j,l);$

2) $I_{i,m} = \sum_{k=2}^{m} I(X_k;Y_k|Y_{k-1}) + I(X_1;Y_1|Y_0=i)$ with $P(X_k=j|Y_{k-1}=i) = \vec{p_i}(j).$

It follows from (13) that

$$
\begin{aligned}
I_{i,n+1} - I_{i,n} &= \sum_{j=1}^{N_y} T_{i,j}(\vec{p_i})I_{j,n} - \sum_{j=1}^{N_y} T_{i,j}(\vec{p_i})I_{j,n-1} \\
&= \sum_{j=1}^{N_y} T_{i,j}(\vec{p_i})(I_{j,n} - I_{j,n-1}).
\end{aligned}
\tag{14}
$$

Using matrix representation, we can write (14) as

$$
\begin{bmatrix}
I_{1,n+1} - I_{1,n} \\
I_{2,n+1} - I_{2,n} \\
\vdots \\
I_{N_y,n+1} - I_{N_y,n}
\end{bmatrix} = T
\begin{bmatrix}
I_{1,n} - I_{1,n-1} \\
I_{2,n} - I_{2,n-1} \\
\vdots \\
I_{N_y,n} - I_{N_y,n-1}
\end{bmatrix},
\tag{15}
$$

where

$$
T =
\begin{bmatrix}
T_{1,1}(\vec{p_1}) & T_{1,2}(\vec{p_1}) & \cdots & T_{1,N_y}(\vec{p_1}) \\
T_{2,1}(\vec{p_2}) & T_{2,2}(\vec{p_2}) & \cdots & T_{2,N_y}(\vec{p_2}) \\
\vdots & \vdots & \ddots & \vdots \\
T_{N_y,1}(\vec{p}_{N_y}) & T_{N_y,2}(\vec{p}_{N_y}) & \cdots & T_{N_y,N_y}(\vec{p}_{N_y})
\end{bmatrix}
$$

is a transition matrix for the homogeneous Markov chain $\{Y_k, k=0,1,\cdots\}$. It follows easily from (15) that

$$
\begin{bmatrix}
I_{1,n+1} - I_{1,n} \\
I_{2,n+1} - I_{2,n} \\
\vdots \\
I_{N_y,n+1} - I_{N_y,n}
\end{bmatrix} = T^n
\begin{bmatrix}
I_{1,1} \\
I_{2,1} \\
\vdots \\
I_{N_y,1}
\end{bmatrix}.
\tag{16}
$$

Since $\{Y_k, k=0,1,\cdots\}$ is a homogenous Markov chain, we can divide its states into two categories: the transient states and the recurrent states. Recurrent states can be decomposed into disjoint irreducible closed sets. Furthermore, there are two kinds of irreducible closed sets: aperiodic and periodic. Now we discuss them separately.

a) Aperiodic irreducible closed set (Suppose it contains states $a_1, a_2, \cdots, a_i$).

Now consider the principal submatrix of T with respect to the $a_1^{th}, a_2^{th}, \cdots, a_i^{th}$ columns and rows, we denote it as T_a. Clearly, we can derive from (16) that

$$
\begin{bmatrix}
I_{a_1,n+1} - I_{a_1,n} \\
I_{a_2,n+1} - I_{a_2,n} \\
\vdots \\
I_{a_i,n+1} - I_{a_i,n}
\end{bmatrix}
= T_a^n
\begin{bmatrix}
I_{a_1,1} \\
I_{a_2,1} \\
\vdots \\
I_{a_i,1}
\end{bmatrix},
\tag{17}
$$

where

$$
T_a =
\begin{bmatrix}
T_{a_1,a_1}(\vec{p}_{a_1}) & T_{a_1,a_2}(\vec{p}_{a_1}) & \cdots & T_{a_1,a_i}(\vec{p}_{a_1}) \\
T_{a_2,a_1}(\vec{p}_{a_2}) & T_{a_2,a_2}(\vec{p}_{a_2}) & \cdots & T_{a_2,a_i}(\vec{p}_{a_2}) \\
\vdots & \vdots & \ddots & \vdots \\
T_{a_i,a_1}(\vec{p}_{a_i}) & T_{a_i,a_2}(\vec{p}_{a_i}) & \cdots & T_{a_i,a_i}(\vec{p}_{a_i})
\end{bmatrix}.
$$

By the Markov convergence theorem for an aperiodic irreducible closed set (see e.g. [18]), we have

$$
T_a^n \to
\begin{bmatrix}
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i} \\
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i}
\end{bmatrix}
\text{ as } n \to \infty,
\tag{18}
$$

where $\{\pi_{a_1}, \cdots, \pi_{a_i}\}$ is the unique stationary (or equilibrium) distribution for the aperiodic irreducible closed set $\{a_1, \cdots, a_i\}$. So by (17) and (18),

$$
\begin{bmatrix}
I_{a_1,n+1} - I_{a_1,n} \\
I_{a_2,n+1} - I_{a_2,n} \\
\vdots \\
I_{a_i,n+1} - I_{a_i,n}
\end{bmatrix}
\to
\begin{bmatrix}
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i} \\
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i} \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{a_1} & \pi_{a_2} & \cdots & \pi_{a_i}
\end{bmatrix}
\begin{bmatrix}
I_{a_1,1} \\
I_{a_2,1} \\
\vdots \\
I_{a_i,1}
\end{bmatrix}
$$

as $n \to \infty$. Thus we have

$$
\begin{aligned}
\lim_{n \to \infty} \frac{I_{a_k,n}}{n} &= \lim_{n \to \infty} \frac{\sum\limits_{l=1}^{n}(I_{a_k,l} - I_{a_k,l-1})}{n} \\
&= \lim_{n \to \infty} (I_{a_k,n} - I_{a_k,n-1}) \\
&= \sum_{l=1}^{i} \pi_{a_l} I_{a_l,1},
\end{aligned}
\tag{19}
$$

which is independent of $k(k = 1, 2, \cdots, i)$; that is, starting from any state in a given aperiodic irreducible closed set, the limiting average directed mutual information is identical.

b) Periodic irreducible closed set (Suppose it contains states $b_1, b_2, \cdots, b_j$ and suppose the period is d. Let $S_0, S_1, \cdots, S_{d-1}$ be the cyclic decomposition of the state space).

Now consider the principal submatrix of T with respect to the $b_1^{th}, b_2^{th}, \cdots, b_j^{th}$ columns and rows, we denote it as T_b. Clearly, we can derive from (16) that

$$\begin{bmatrix} I_{b_1,n+1} - I_{b_1,n} \\ I_{b_2,n+1} - I_{b_2,n} \\ \vdots \\ I_{b_j,n+1} - I_{b_j,n} \end{bmatrix} = T_b^n \begin{bmatrix} I_{b_1,1} \\ I_{b_2,1} \\ \vdots \\ I_{b_j,1} \end{bmatrix}, \tag{20}$$

where

$$T_b = \begin{bmatrix} T_{b_1,b_1}(\vec{p}_{b_1}) & T_{b_1,b_2}(\vec{p}_{b_1}) & \cdots & T_{b_1,b_j}(\vec{p}_{b_1}) \\ T_{b_2,b_1}(\vec{p}_{b_2}) & T_{b_2,b_2}(\vec{p}_{b_2}) & \cdots & T_{b_2,b_j}(\vec{p}_{b_2}) \\ \vdots & \vdots & \ddots & \vdots \\ T_{b_j,b_1}(\vec{p}_{b_j}) & T_{b_j,b_2}(\vec{p}_{b_j}) & \cdots & T_{b_j,b_j}(\vec{p}_{b_j}) \end{bmatrix}.$$

By the Markov convergence theorem for a periodic irreducible closed set (see e.g. [18]), we have

$$\frac{1}{n} \sum_{m=1}^{n} (T_b^m)_{k,l} \to \pi_{b_l}$$

where $(T_b^m)_{k,l}$ is the component on the kth row and l^{th} column of matrix $\underbrace{T_b \cdot T_b \cdots T_b}_{m}$, $\{\pi_{b_1}, \cdots, \pi_{b_j}\}$ is the unique stationary (or equilibrium) distribution for the aperiodic irreducible closed set $\{b_1, b_2, \cdots, b_j\}$. More specifically, if $b_k \in \mathcal{S}_0$ and $b_l \in \mathcal{S}_r$ (actually we just need the subscript of the cyclic state space that they belong to differ by $r(r = 0, 1, \cdots, d-1)$), then we have

$$\lim_{m \to \infty} (T_b^{md+r})_{k,l} = \pi_{b_l} d; \tag{21}$$

$$(T_b^m)_{k,l} = 0 \text{ for } (m \mod d) \neq r. \tag{22}$$

It follows by (20), (21) and (22) that

$$
\begin{aligned}
\lim_{n \to \infty} \frac{I_{b_k,n}}{n} &= \lim_{n \to \infty} \frac{\sum\limits_{m=1}^{n} (I_{b_k,m} - I_{b_k,m-1})}{n} \\
&= \lim_{n \to \infty} \frac{\sum\limits_{m=1}^{n} \sum\limits_{l=1}^{j} [(T_b^{m-1})_{k,l} I_{b_l,1}]}{n} \\
&= \sum_{l=1}^{j} I_{b_l,1} \lim_{n \to \infty} \frac{\sum\limits_{m=1}^{n} (T_b^{m-1})_{k,l}}{n} \\
&= \sum_{l=1}^{j} \pi_{b_l} I_{b_l,1},
\end{aligned}
\tag{23}
$$

which is independent of $k(k = 1, 2, \cdots, j)$; that is, starting from any state in a given periodic irreducible closed set, the limiting average directed mutual information is identical.

Since the limiting average directed mutual information is seen to have the same expression for an aperiodic irreducible closed set as for a periodic irreducible closed set, in order to find the stationary input that maximizes the average directed mutual information given the initial state $Y_0 = k$, we can proceed as follows, where without loss of generality we suppose $k = N_y$:

i) Let $\mathcal{P}$ be the space containing all the $N_x \times 1$ probability vectors. Now consider the product space $\Lambda = \mathcal{P}_1 \times \mathcal{P}_2 \times \cdots \times \mathcal{P}_{N_y}$, where $\mathcal{P}_i = \mathcal{P}$ for all $i = 1, 2, \cdots, N_y$. Decompose Λ into 2^{N_y-1} disjoint subsets such that $\Lambda = \bigcup\limits_{i=0}^{2^{N_y-1}-1} \Lambda_i$ and the Λ_i have the property that for any $\vec{p}_1 \times \vec{p}_2 \times \cdots \times \vec{p}_{N_y} \in \Lambda_i$, the indices of the positions that are 1 in the binary expansion of i correspond to the states that form an irreducible closed set with state $k(= N_y)$. For example, let $N_y = 4$ and $i = 3$; then the binary expansion of i is 011. For any $\vec{p}_1 \times \vec{p}_2 \times \vec{p}_3 \times \vec{p}_4 \in \Lambda_3$, states 1, 2, 4 form an irreducible closed set. For any $\vec{p}_1 \times \vec{p}_2 \times \cdots \times \vec{p}_{N_y} \in \Lambda_0$, state $k(= N_y)$ is either a transient state or an irreducible closed set formed by itself.

ii) The maximum average directed mutual information for a stationary input distribution can be obtained by

$$
\max_{i \in \{1,2,\cdots,2^{N_y-1}\}} \left\{ \max_{(\vec{p}_1 \times \cdots \times \vec{p}_{N_y}) \in \Lambda_i} \left(\sum_{l \in \Delta_i} \pi_l I_{l,1} \right) \right\}
\tag{24}
$$

if the above maximization operation is feasible. Here Δ_i is the set containing the indices of the positions that are 1 in the binary expansion of $i + 2^{N_y-1}$. For the previous example, if $i = 3$, then $\Delta_i = \{1, 2, 4\}$.

iii) In ii), we did not consider Λ_0 in the maximization. Clearly, under some stationary input distribution $(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})$, if state k forms an irreducible closed set by itself, then

$$\lim_{n \to \infty} \frac{I_{k,n}}{n} = 0;$$

if state k is transient, then we have

$$\lim_{n \to \infty} \frac{I_{k,n}}{n} = \sum_i P_i I_i^*,$$

where P_i is the probability that the Markov chain will end in the irreducible closed set i and I_i^* is the limiting average directed mutual information for this irreducible closed set as was discussed in a) and b). The example discussed in Section III is a special case of iii).

When the initial state is transient under any input distribution, it may seem to be a good choice to maximize the probability that the Markov chain will be absorbed in the irreducible closed set that has the largest limiting average directed mutual information. To see that the problem actually is much more complicated, consider the example shown in Fig. 4. If $Y_0 = 2$, then we
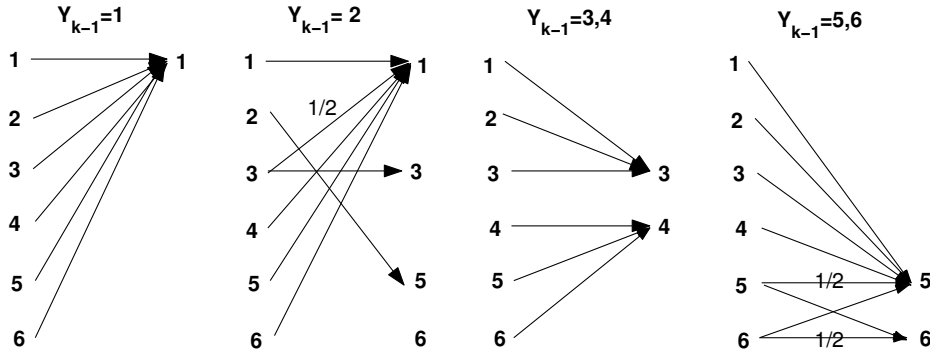

Fig. 4 Example 2

can let $X_1 = 2$ or 3 (other inputs will drive the Markov chain into state 1 which is a dead end). But if we choose $X_1 = 3$, then with probability $\frac{1}{2}$, the Markov chain will be driven to state 1 and stuck there forever; also, with probability $\frac{1}{2}$, the Markov chain will be driven to state 3 and we can transmit 1 bit of information per channel use after that. If we choose $X_1 = 2$, then the Markov chain will be driven to state 5 and we can transmit $\log_2 5 - 2$ bits of information per channel use after that. Clearly, for this channel model, if we want to drive the Markov chain into an irreducible closed set with highest limiting average directed mutual information – namely

$\{3, 4\}$ – then we need to take the risk that we may actually end in the bad irreducible closed set $\{1\}$. The feedback channel capacity introduced in Definition 1 can be roughly interpreted as the maximal reliable communication rate in the worst case scenario. So for this channel model, it's easy to check that $C_2^{fd} = \log_2 5 - 2$. By Example 1 and 2, we can see that, if there does not exist a input distribution under which all the channel states form a single irreducible set, the feedback channel capacity given in Definition 1 may not reveal the intrinsic structure of the channel. Outage capacity seems to be a more proper concept in this context.

VI. OPTIMAL NONSTATIONARY INPUT DISTRIBUTION

Next, we study the nonstationary input distribution that maximizes the limiting average directed mutual information. Here, "nonstationary" means the input depends both on the current channel state (i.e., the previous channel output) and on time, whereas "stationary" (as was discussed in Section V) means the input depends only on current channel state. As we saw in Section V, the transition matrix of the Markov process $\{Y_k, k = 0, 1, \cdots\}$ depends on the input distribution. If the input is stationary, then $\{Y_k, k = 0, 1, \cdots\}$ is a homogeneous Markov chain and we have a simple way to determine the unique decomposition of the state space into disjoint irreducible closed sets and transient state sets. But when the input distribution is not stationary, then $\{Y_k, k = 0, 1, \cdots\}$ becomes an inhomogeneous Markov chain and there is no simple method to determine whether a state is transient or recurrent. Roughly speaking, if we view a Markov chain as a random walk on a directed graph, then the connectivity of this graph (which is determined by the transition matrix) is fixed for a homogeneous Markov chain, while it changes with time for an inhomogeneous Markov chain. In our case the connectivity of the graph is determined by the input distribution, so it will change with time if the input is nonstationary. In order to make the analysis tractable, we need to impose some restrictions on our model.

First we introduce two concepts: strong irreducibility and strong aperiodicity. Here we imitate the definitions of irreducibility and aperiodicity in the classic Markov theory.

Definition 2 (Strong irreducibility): Let

$$\tilde{T}(i, j) = \min_{l \in \{1, \cdots, N_x\}} \{Q(Y_k = j | X_k = l, Y_{k-1} = i)\}$$

We say there exists a directed edge from state i to state j if $\tilde{T}(i, j) > 0$. We say a Markov chain $\{Y_k, k = 0, 1, \cdots\}$ is strongly irreducible if for any two states i and j (i can be equal to j), there

exists a directed path from i to j. For simplicity, we just say $\tilde{T}$, the $N_y \times N_y$ matrix whose (i,j) is $\tilde{T}(i,j)$, is strongly irreducible, since $\tilde{T}$ contains all the information that determines whether the Markov chain $\{Y_k, k = 0, 1, \cdots\}$ is strongly irreducible or not.

Definition 3: Where the "length" of a path is the number of edges comprising the path, let $\mathcal{D}_i$ be the set of lengths of all the possible closed paths from state i to state i. Let d_i be the greatest common divisor of $\mathcal{D}_i$. d_i is called the period of state i.

The following result says that period is a class property.

Lemma 2: If the Markov chain $\{Y_k, k = 0, 1, \cdots\}$ is strongly irreducible, then $d_i = d_j$ for any i and j.

Proof: Let K and L be the integers such that there exist a directed path of length K from state i to state j and a directed path of length L from state j to state i. So there exists a directed path of length $K + L$ from state j to state j. Hence $d_j | (K + L)$.

Let $m \in \mathcal{D}_i$, by Definition 3, there exists a directed path of length $K + m + L$ from state j to state j. So $d_j | (K + m + L)$, and hence $d_j | m$. Since $m \in \mathcal{D}_i$ is arbitrary, $d_j | d_i$.

Interchanging the roles of j and i gives $d_i | d_j$, and hence $d_i = d_j$. ∎

So for a strongly irreducible Markov chain $\{Y_k, k = 0, 1, \cdots\}$, all the states have the same period, which we shall denote by d.

Definition 4 (Strong aperiodicity): We say a strongly irreducible Markov chain $\{Y_k, k = 0, 1, \cdots\}$ is strongly aperiodic if $d = 1$.

For simplicity, we just say that $\tilde{T}$ is strongly irreducible and strongly aperiodic. Clearly, our definitions of irreducibility and aperiodicity are stronger than those in the usual sense; i.e., if the Markov chain $\{Y_k, k = 0, 1, \cdots\}$ is stongly irreducible and strongly aperiodic, then it's irreducible and aperiodic in the usual sense for any input distribution $\{\vec{p}_{1,k}, \vec{p}_{2,k}, \cdots, \vec{p}_{N_y,k}\}_{k=1}^{\infty}$. Note, however, such a Markov chain is in general inhomogeneous since $\{\vec{p}_{1,k}, \vec{p}_{2,k}, \cdots, \vec{p}_{N_y,k}\}$ may depend on k. Again, if we view the Markov process $\{Y_k, k = 0, 1, \cdots\}$ as a random walk on a directed graph, then strong irreducibility and strong aperiodicity assure us that this directed graph is always strongly connected and all the states of it are of period 1 no matter what input distribution is generating $\{Y_k\}$ via the channel.

We now proceed to prove a lemma which will be useful in the proof of the main theorem in this section.

Lemma 3: If $\tilde{T}$ is strongly irreducible and strongly aperiodic, then there is a K such that

$(\tilde{T})^K(i,j) > 0$ for all i, j.

Proof: Since $\tilde{T}$ is strongly irreducible, every row of $\tilde{T}$ should have at least one positive element. So we can scale every row of $\tilde{T}$ to make it to be a transition matrix T in which the summation of the elements on every row is 1. Clearly, T is irreducible and aperiodic in the usual sense and we have

$$T(i,j) = 0 \Leftrightarrow \tilde{T}(i,j) = 0.$$

By [16, Ex 5.9], there exists an K such that $T^K(i,j) > 0$ for all i, j. Since

$$T(i,j) = 0 \Leftrightarrow \tilde{T}(i,j) = 0,$$

it follows that

$$T^n(i,j) = 0 \Leftrightarrow \tilde{T}^n(i,j) = 0$$

for any positive integer n. So we can conclude that

$$\tilde{T}^K(i,j) > 0$$

for all i, j. ∎

Theorem 3: If $\tilde{T}$ is strongly irreducible and strongly aperiodic, then $\lim\limits_{n \to \infty} \frac{C_{i,n}}{n}$ exists and is independent of i.

Proof: By Theorem 2, we have

$$C_{i,n} = I_i(\vec{p}_{i,n}^{\,*}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}^{\,*}) C_{j,n-1}, \tag{25}$$

where $\vec{p}_{i,n}^{\,*} \in \arg\max\limits_{\vec{p}_{i,n}} \{ I_i(\vec{p}_{i,n}) + \sum\limits_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}) C_{j,n-1} \}$. Similarly,

$$C_{i,n+1} = I_i(\vec{p}_{i,n+1}^{\,*}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n+1}^{\,*}) C_{j,n}. \tag{26}$$

By the definition of $\vec{p}_{i,n}^{\,*}$ and $\vec{p}_{i,n+1}^{\,*}$, we have

$$C_{i,n} \geq I_i(\vec{p}_{i,n+1}^{\,*}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n+1}^{\,*}) C_{j,n-1}, \tag{27}$$

$$C_{i,n+1} \geq I_i(\vec{p}_{i,n}^{\,*}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}^{\,*}) C_{j,n}. \tag{28}$$

It follows from (26) and (27) that

$$C_{i,n+1} - C_{i,n} \leq \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n+1}^{\,*})(C_{j,n} - C_{j,n-1}). \tag{29}$$

Similarly, by (25) and (28) we have

$$C_{i,n+1} - C_{i,n} \geq \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}^{\,*})(C_{j,n} - C_{j,n-1}). \tag{30}$$

Using matrix representation, we can write (29) and (30) jointly as

$$T_n^* \begin{bmatrix} C_{1,n} - C_{1,n-1} \\ C_{2,n} - C_{2,n-1} \\ \vdots \\ C_{N_y,n} - C_{N_y,n-1} \end{bmatrix} \leq \begin{bmatrix} C_{1,n+1} - C_{1,n} \\ C_{2,n+1} - C_{2,n} \\ \vdots \\ C_{N_y,n+1} - C_{N_y,n} \end{bmatrix} \leq T_{n+1}^* \begin{bmatrix} C_{1,n} - C_{1,n-1} \\ C_{2,n} - C_{2,n-1} \\ \vdots \\ C_{N_y,n} - C_{N_y,n-1} \end{bmatrix}, \tag{31}$$

where

$$T_n^* = \begin{bmatrix} T_{1,1}(\vec{p}_{1,n}^{\,*}) & T_{1,2}(\vec{p}_{1,n}^{\,*}) & \cdots & T_{1,N_y}(\vec{p}_{1,n}^{\,*}) \\ T_{2,1}(\vec{p}_{2,n}^{\,*}) & T_{2,2}(\vec{p}_{2,n}^{\,*}) & \cdots & T_{2,N_y}(\vec{p}_{2,n}^{\,*}) \\ \vdots & \vdots & \ddots & \vdots \\ T_{N_y,1}(\vec{p}_{N_y,n}^{\,*}) & T_{N_y,2}(\vec{p}_{N_y,n}^{\,*}) & \cdots & T_{N_y,N_y}(\vec{p}_{N_y,n}^{\,*}) \end{bmatrix}.$$

Since $\tilde{T}$ is strongly irreducible and strongly aperiodic, by Lemma 3, there is a K such that

$$\tilde{T}^K(i,j) > 0$$

for all i, j. From (31), we can get

$$(\prod_{i=0}^{K-1} T_{n+i}^*) \begin{bmatrix} C_{1,n} - C_{1,n-1} \\ C_{2,n} - C_{2,n-1} \\ \vdots \\ C_{N_y,n} - C_{N_y,n-1} \end{bmatrix} \leq \begin{bmatrix} C_{1,n+K} - C_{1,n+K-1} \\ C_{2,n+K} - C_{2,n+K-1} \\ \vdots \\ C_{N_y,n+K} - C_{N_y,n+K-1} \end{bmatrix} \leq (\prod_{i=1}^{K} T_{n+i}^*) \begin{bmatrix} C_{1,n} - C_{1,n-1} \\ C_{2,n} - C_{2,n-1} \\ \vdots \\ C_{N_y,n} - C_{N_y,n-1} \end{bmatrix}. \tag{32}$$

It's easy to see that

$$(\prod_{l=0}^{K-1} T_{m+l}^*)(i,j) \geq \tilde{T}^K(i,j)$$

for all positive integers i, j, m. Let

$$\alpha = \min_{i,j} \{\tilde{T}^K(i,j)\}.$$

Clearly, we have $\frac{1}{2} \geq \alpha > 0$. Let

$$
\begin{aligned}
M_n &= \max_{i \in \{1,2,\cdots,N_y\}} (C_{i,n} - C_{i,n-1}), \\
m_n &= \min_{i \in \{1,2,\cdots,N_y\}} (C_{i,n} - C_{i,n-1}).
\end{aligned}
$$

By (32), we can get

$$
\begin{aligned}
M_{n+K} &\leq (1-\alpha)M_n + \alpha m_n, \\
m_{n+K} &\geq \alpha M_n + (1-\alpha)m_n,
\end{aligned}
$$

and thus we have

$$
0 \leq (M_{n+K} - m_{n+K}) \leq (1-2\alpha)(M_n - m_n).
$$

It follows by recursion that

$$
0 \leq (M_{nK+1} - m_{nK+1}) \leq (1-2\alpha)^n(M_1 - m_1), \tag{33}
$$

and thus

$$
\lim_{n\to\infty} (M_{nK+1} - m_{nK+1}) = 0.
$$

By (31), it's easy to show that M_n is monotonically decreasing, m_n is monotonically increasing and both of them are bounded, so their limits exist. Hence

$$
\lim_{n\to\infty} (M_{nK+1} - m_{nK+1}) = 0
$$

implies

$$
\lim_{n\to\infty} M_n = \lim_{n\to\infty} m_n.
$$

That is, $\lim_{n\to\infty} (C_{i,n+1} - C_{i,n})$ exists and is independent of i. So we can conclude that

$$
\lim_{n\to\infty} \frac{C_{i,n}}{n} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n}(C_{i,k} - C_{i,k-1})}{n} = \lim_{n\to\infty} (C_{i,n} - C_{i,n-1})
$$

exists and is independent of i. ∎

Now we begin to analyze the convergence rate of $(C_{j,n+1} - C_{j,n}) - (C_{i,n+1} - C_{i,n})$, which is useful for the next section.

By (33), we have, for any i and j ($i = 1, 2, \cdots, N_y; j = 1, 2, \cdots, N_y$),

$$
\begin{aligned}
|(C_{j,n} - C_{j,n-1}) - (C_{i,n} - C_{i,n-1})| &\leq |M_n - m_n| \leq |M_{\lfloor \frac{n-1}{K} \rfloor K+1} - m_{\lfloor \frac{n-1}{K} \rfloor K+1}| \\
&\leq (1 - 2\alpha)^{\lfloor \frac{n-1}{K} \rfloor} (M_1 - m_1) \\
&\leq (1 - 2\alpha)^{\frac{n-1}{K} - 1} (M_1 - m_1) \\
&= \beta \gamma^n,
\end{aligned}
\tag{34}
$$

where $\beta = (1 - 2\alpha)^{-\frac{K+1}{K}} (M_1 - m_1)$ and $\gamma = (1 - 2\alpha)^{\frac{1}{K}}$. Note: $0 \leq \gamma < 1$.

VII. Convergence of Nonstationary Input to Stationary Input

We now show that under certain conditions the limiting maximum average directed mutual information actually can be achieved by a stationary input. Before proving the main theorem in this section, we need to introduce several definitions.

Definition 4 (the vector p-norms): For $\vec{x} \in \mathcal{R}^{n \times 1}$,

$$
||\vec{x}||_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}, \ p \geq 1.
$$

Specifically, for $p = 1, 2, \infty$, we have

$$
\begin{aligned}
||\vec{x}||_1 &= |\vec{x}(1)| + \cdots + |\vec{x}(n)|, \\
||\vec{x}||_2 &= (|\vec{x}(1)|^2 + \cdots + |\vec{x}(n)|^2)^{\frac{1}{2}}, \\
||\vec{x}||_\infty &= \max_{1 \leq i \leq n} |\vec{x}(i)|.
\end{aligned}
$$

Definition 5 (the matrix p-norms): For $A \in \mathcal{R}^{m \times n}$,

$$
||A||_p = \sup_{\vec{x} \neq 0} \frac{||A\vec{x}||_p}{||\vec{x}||_p}, \ p \geq 1.
$$

Specifically, for $p = 1, 2, \infty$, we have

$$
\begin{aligned}
||A||_1 &= \max_{1 \leq j \leq n} \sum_{i=1}^{m} |A(i,j)|, \\
||A||_2 &\text{ is the square root of the largest eigenvalue of } A^T A, \\
||A||_\infty &= \max_{1 \leq i \leq m} \sum_{j=1}^{n} |A(i,j)|.
\end{aligned}
$$

For the detailed discussion of the properties of the vector p-norms and matrix p-norms, see [19].

Definition 6: Let $\mathcal{P}$ denote the set of all $N_x \times 1$ probability vectors. We say an $N_y \times N_x$ channel transition probability matrix $Q \in \Psi_{N_y \times N_x}$ if there exists a complete subset $\tilde{\mathcal{P}} \subset \mathcal{P}$ such that the following three conditions are satisfied:

(i) $\{Q\vec{p} : \vec{p} \in \mathcal{P}\} = \{Q\vec{p} : \vec{p} \in \tilde{\mathcal{P}}\}$.

(ii) For any $\vec{q} \in \{Q\vec{p} : \vec{p} \in \mathcal{P}\}$,

$$\{\arg \max_{\vec{p}:Q\vec{p}=\vec{q} \ and \ \vec{p}\in\mathcal{P}} I(\vec{p},Q)\} \cap \{\arg \max_{\vec{p}:Q\vec{p}=\vec{q} \ and \ \vec{p}\in\tilde{\mathcal{P}}} I(\vec{p},Q)\} \neq \emptyset.$$

(iii) There exists a positive constant λ such that

$$\frac{\partial I(\vec{p}_2,Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p}_1,Q)}{\partial \vec{l}} \leq -\lambda||\vec{p}_2 - \vec{p}_1||_2$$

for any nonidentical $\vec{p}_1, \vec{p}_2 \in \tilde{\mathcal{P}}$ and $\vec{l}$ with the direction from $\vec{p}_1$ to $\vec{p}_2$.

$\tilde{\mathcal{P}}$ is called the complete subset with respect to Q.

See the appendix for the detailed discussion of Definition 6.

We also need the following lemma concerning the backward product of matrices.

Lemma 4 [13]: If $\sum_{k=1}^{\infty} ||A - A_k||_\infty < \infty$, then for any $\varepsilon > 0$, there exists a positive number r_1 such that, for all $n > m > r_1$,

$$\left|\left|\prod_{k=0}^{n-m} A_{n-k} - A^{n-m+1}\right|\right|_\infty \leq \varepsilon,$$

where $A, A_k(k = 1, 2, \cdots)$ are stochastic matrices.

Now we are ready to prove the main theorem in this section.

Theorem 4: If $\tilde{T}$ is strongly irreducible and strongly aperiodic and $Q_i \in \Psi_{N_y \times N_x}(i = 1, 2, \cdots, N_y)$, then

$$\lim_{n\to\infty} \frac{C_{i,n}}{n} = \max_{(\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_{N_y})} \{\sum_{l=1}^{N_y} \pi_l I_{l,1}\},$$

where $(\pi_1, \pi_2, \cdots, \pi_{N_y})$ is the equilibrium distribution of the channel output process induced by the stationary input distribution $(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})$.

Proof: Let $\tilde{\mathcal{P}}_i$ be the complete subset with respect to Q_i. Since

$$\sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n})C_{j,n-1} = \vec{C}_{n-1}^T Q\vec{p}_{i,n},$$

where $\vec{C}_{n-1} = [C_{1,n-1}, \cdots, C_{N_y,n-1}]^T$, Definition 6 and the discussion in the appendix assure us that

$$C_{i,n} = \max_{\vec{p}_{i,n}}\{I_i(\vec{p}_{i,n}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n})C_{j,n-1}\}$$

is attained inside $\tilde{\mathcal{P}}_i$ and there exists a unique $\vec{p}_{i,n}^* \in \tilde{\mathcal{P}}_i$ such that

$$C_{i,n} = I_i(\vec{p}_{i,n}^*) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p}_{i,n}^*)C_{j,n-1}.$$

First we derive a bound on the convergence rate of $\vec{p}_{i,n}^*$. Let

$$f_{i,n}(\vec{p}) = I_i(\vec{p}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p})(C_{j,n-1} - C_{i,n-1}).$$

We have

$$\begin{aligned}
f_{i,n+1}(\vec{p}) &= I_i(\vec{p}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p})(C_{j,n} - C_{i,n}) \\
&= f_{i,n}(\vec{p}) + \sum_{j=1}^{N_y} T_{i,j}(\vec{p})[(C_{j,n} - C_{i,n}) - (C_{j,n-1} - C_{i,n-1})],
\end{aligned}$$

and thus

$$\frac{\partial f_{i,n+1}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) = \frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) + \sum_{j=1}^{N_y} [\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)][(C_{j,n} - C_{i,n}) - (C_{j,n-1} - C_{i,n-1})], \quad (35)$$

where the direction of $\vec{l}$ is from $\vec{p}_{i,n}^*$ to $\vec{p}_{i,n+1}^*$ (We temporarily suppose that $\vec{p}_{i,n}^* \neq \vec{p}_{i,n+1}^*$). Since $f_{i,n+1}(\vec{p}_{i,n+1}^*)$ is the maximum value of $f_{i,n+1}(\vec{p})$ and the direction of $\vec{l}$ is from $\vec{p}_{i,n}^*$ to $\vec{p}_{i,n+1}^*$, it follows that

$$\frac{\partial f_{i,n+1}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) \geq 0. \quad (36)$$

Since $\sum\limits_{j=1}^{N_y} T_{i,j}(\vec{p})(C_{j,n-1} - C_{i,n-1})$ is linear with respect to $\vec{p}$, it follows that

$$\frac{\partial^2 f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) - \frac{\partial^2 f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n}^*) = \frac{\partial^2 I_i}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) - \frac{\partial^2 I_i}{\partial \vec{l}}(\vec{p}_{i,n}^*).$$

Then by Definition 6, we have

$$\frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) - \frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n}^*) \leq -\lambda_i ||\vec{p}_{i,n}^* - \vec{p}_{i,n+1}^*||_2. \quad (37)$$

Since $f_{i,n}(\vec{p}_{i,n}^*)$ is the maximum value of $f_{i,n}(\vec{p})$ and the direction of $\vec{l}$ is from $\vec{p}_{i,n}^*$ to $\vec{p}_{i,n+1}^*$, it follows that

$$\frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n}^*) \leq 0. \tag{38}$$

Putting (38) into (37), we get

$$\frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) \leq -\lambda_i ||\vec{p}_{i,n}^* - \vec{p}_{i,n+1}^*||_2. \tag{39}$$

By (35), (36) and (39), we have

$$-\lambda_i ||\vec{p}_{i,n}^* - \vec{p}_{i,n+1}^*||_2 + \sum_{j=1}^{N_y} [\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)][(C_{j,n} - C_{i,n}) - (C_{j,n-1} - C_{i,n-1})]$$

$$\geq \quad \frac{\partial f_{i,n}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) + \sum_{j=1}^{N_y} [\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)][(C_{j,n} - C_{i,n}) - (C_{j,n-1} - C_{i,n-1})]$$

$$= \quad \frac{\partial f_{i,n+1}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*) \geq 0,$$

or equivalently

$$||\vec{p}_{i,n}^* - \vec{p}_{i,n+1}^*||_2 \leq \frac{1}{\lambda_i} \sum_{j=1}^{N_y} [\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)][(C_{j,n} - C_{j,n-1}) - (C_{i,n} - C_{i,n-1})]. \tag{40}$$

It follows from (34) that

$$|(C_{j,n} - C_{j,n-1}) - (C_{i,n} - C_{i,n-1})| \leq \beta \gamma^n,$$

and it's easy to check that

$$\sum_{j=1}^{N_y} |\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)| \leq N_x N_y.$$

So for any $n \geq 1$ we have

$$||\vec{p}_{i,n}^* - \vec{p}_{i,n+1}^*||_2 \leq \frac{1}{\lambda_i} \sum_{j=1}^{N_y} |\frac{\partial T_{i,j}}{\partial \vec{l}}(\vec{p}_{i,n+1}^*)||(C_{j,n} - C_{j,n-1}) - (C_{i,n} - C_{i,n-1})| \leq \frac{1}{\lambda_i} \beta \gamma^n N_x N_y. \tag{41}$$

It's obvious that (41) also holds when $\vec{p}_{i,n}^* = \vec{p}_{i,n+1}^*$. Thus

$$||\vec{p}_{i,m+n}^* - \vec{p}_{i,m}^*||_2 \leq \sum_{j=1}^{n} ||\vec{p}_{i,m+j}^* - \vec{p}_{i,m+j-1}^*||_2 \leq \frac{\beta N_x N_y}{\lambda_i} \sum_{i=1}^{n} \gamma^{m+i-1} \leq \frac{\beta N_x N_y \gamma^m}{\lambda_i(1 - \gamma)} \to 0, \tag{42}$$

for any $n \geq 1$ as $m \to \infty$. Hence $\vec{p}_{i,n}^*$ is a Cauchy sequence. Since $\tilde{\mathcal{P}}_i$ is complete, there exists $\vec{p}_i^* \in \tilde{\mathcal{P}}_i$ such that $\vec{p}_{i,n}^* \to \vec{p}_i^*$. Clearly, this result holds for all $i (i = 1, 2, \cdots, N_y)$.

Let $E_n = T_n^* - T^*$, where

$$T_n^* = \begin{bmatrix} T_{1,1}(\vec{p}_{1,n}^*) & T_{1,2}(\vec{p}_{1,n}^*) & \cdots & T_{1,N_y}(\vec{p}_{1,n}^*) \\ T_{2,1}(\vec{p}_{2,n}^*) & T_{2,2}(\vec{p}_{2,n}^*) & \cdots & T_{2,N_y}(\vec{p}_{2,n}^*) \\ \vdots & \vdots & \ddots & \vdots \\ T_{N_y,1}(\vec{p}_{N_y,n}^*) & T_{N_y,2}(\vec{p}_{N_y,n}^*) & \cdots & T_{N_y,N_y}(\vec{p}_{N_y,n}^*) \end{bmatrix}$$

and

$$T^* = \begin{bmatrix} T_{1,1}(\vec{p}_1^{\,*}) & T_{1,2}(\vec{p}_1^{\,*}) & \cdots & T_{1,N_y}(\vec{p}_1^{\,*}) \\ T_{2,1}(\vec{p}_2^{\,*}) & T_{2,2}(\vec{p}_2^{\,*}) & \cdots & T_{2,N_y}(\vec{p}_2^{\,*}) \\ \vdots & \vdots & \ddots & \vdots \\ T_{N_y,1}(\vec{p}_{N_y}^{\,*}) & T_{N_y,2}(\vec{p}_{N_y}^{\,*}) & \cdots & T_{N_y,N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix}.$$

By Holder's inequality and (42),

$$||\vec{p}_{i,n}^* - \vec{p}_i^{\,*}||_1 \leq \sqrt{N_x}||\vec{p}_{i,n}^* - \vec{p}_i^{\,*}||_2 \leq \frac{\beta N_x^{\frac{3}{2}} N_y \gamma^n}{\lambda_i(1-\gamma)}.$$

Clearly, we have

$$||E_n||_\infty = ||T_n^* - T^*||_\infty \leq N_x N_y \max_{i \in \{1,2,\cdots,N_y\}} \{||\vec{p}_{i,n}^* - \vec{p}_i^{\,*}||_1\} \leq \frac{\beta N_x^{\frac{5}{2}} N_y^2 \gamma^n}{\lambda(1-\gamma)},$$

where $\lambda = \min\{\lambda_1, \cdots, \lambda_{N_y}\}$. Hence

$$\sum_{n=1}^\infty ||E_n||_\infty \leq \sum_{n=1}^\infty \frac{\beta N_x^{\frac{5}{2}} N_y^2 \gamma^n}{\lambda(1-\gamma)} = \frac{\beta N_x^{\frac{5}{2}} N_y^2 \gamma}{\lambda(1-\gamma)^2} < \infty.$$

Applying Lemma 4, we deduce that for any $\varepsilon > 0$, there exists a positive number r_1 such that, for all $n > m > r_1$,

$$\left\| \prod_{k=0}^{n-m} T_{n-k}^* - (T^*)^{n-m+1} \right\|_\infty \leq \varepsilon. \tag{43}$$

Since $\vec{p}_{i,n}^* \to \vec{p}_i^{\,*}$ for every i, it follows that for any $\varepsilon > 0$, there exists a positive number r_2 such that, for all $n > r_2$,

$$\left\| \begin{bmatrix} I_1(\vec{p}_{1,n}^*) - I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n}^*) - I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \right\|_\infty \leq \varepsilon. \tag{44}$$

Let $r = \max(r_1, r_2) + 1$. In the remaining part of this proof, we implicitly assume $P(Y_0 = y_0) = 1$. Now we have

$$
\begin{aligned}
C_{y_0,n} &\overset{(e)}{=} \sum_{k=1}^{n} I(X_k; Y_k | Y_{k-1}) \\
&= \sum_{k=1}^{n} [P(Y_{k-1} = 1), \cdots, P(Y_{k-1} = N_y)] \begin{bmatrix} I_1(\vec{p}^*_{1,n+1-k}) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y,n+1-k}) \end{bmatrix} \\
&= [P(Y_0 = 1), \cdots, P(Y_0 = N_y)] \begin{bmatrix} I_1(\vec{p}^*_{1,n}) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y,n}) \end{bmatrix} \\
&\quad + \sum_{k=2}^{n} [P(Y_0 = 1), \cdots, P(Y_0 = N_y)] (\prod_{j=0}^{k-2} T^*_{n-j}) \begin{bmatrix} I_1(\vec{p}^*_{1,n+1-k}) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y,n+1-k}) \end{bmatrix}.
\end{aligned}
$$

(e) The input distributions of $X_1, X_2, \cdots, X_n$ are $\{\vec{p}^*_{1,n}, \cdots, \vec{p}^*_{N_y,n}\}$, $\{\vec{p}^*_{1,n-1}, \cdots, \vec{p}^*_{N_y,n-1}\}$, $\cdots$, $\{\vec{p}^*_{1,1}, \cdots, \vec{p}^*_{N_y,1}\}$ respectively.

Similarly, we have

$$
C^*_{y_0,n} \overset{(f)}{=} \sum_{k=1}^{n} I(X^*_k; Y^*_k | Y^*_{k-1})
$$

$$
= \sum_{k=1}^{n} [P(Y^*_{k-1} = 1), \cdots, P(Y^*_{k-1} = N_y)] \begin{bmatrix} I_1(\vec{p}^*_1) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y}) \end{bmatrix}
$$

$$
\overset{(g)}{=} [P(Y_0 = 1), \cdots, P(Y_0 = N_y)] \begin{bmatrix} I_1(\vec{p}^*_1) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y}) \end{bmatrix} + \sum_{k=2}^{n} [P(Y_0 = 1), \cdots, P(Y_0 = N_y)] (T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}^*_1) \\ \vdots \\ I_{N_y}(\vec{p}^*_{N_y}) \end{bmatrix}.
$$

(f) All the input distributions of $X^*_1, X^*_2, \cdots, X^*_n$ are $\{\vec{p}^*_1, \cdots, \vec{p}^*_{N_y}\}$; (g) $Y_0 = Y^*_0 = y_0$. So we have

$$
\left| \frac{1}{n} C_n(Y_0) - \frac{1}{n} C^*_n(Y_0) \right| \leq A_1 + A_2 + A_3, \tag{45}
$$

where

$$
A_1 = \frac{1}{n} \left| [P(Y_0=1),\cdots,P(Y_0=N_y)] \begin{bmatrix} I_1(\vec{p}_{1,n}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n}^{\,*}) \end{bmatrix} - [P(Y_0=1),\cdots,P(Y_0=N_y)] \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \right|,
$$

$$
A_2 = \frac{1}{n} \left| \sum_{k=n+2-r}^{n} [P(Y_0=1),\cdots,P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^*) \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^{\,*}) \end{bmatrix} \right.
$$
$$
\left. - \sum_{k=n+2-r}^{n} [P(Y_0=1),\cdots,P(Y_0=N_y)] (T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \right|,
$$

$$
A_3 = \frac{1}{n} \left| \sum_{k=2}^{n+1-r} [P(Y_0=1),\cdots,P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^*) \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^{\,*}) \end{bmatrix} \right.
$$
$$
\left. - \sum_{k=2}^{n+1-r} [P(Y_0=1),\cdots,P(Y_0=N_y)] (T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \right|.
$$

Now we evaluate A_1, A_2 and A_3 respectively:

i)
$$
\lim_{n\to\infty} A_1 = \lim_{n\to\infty} \frac{1}{n} \Bigg| [P(Y_0=1), \cdots, P(Y_0=N_y)] \begin{bmatrix} I_1(\vec{p}_{1,n}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n}^{\,*}) \end{bmatrix}
$$

$$
- [P(Y_0=1), \cdots, P(Y_0=N_y)] \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \Bigg| = 0. \tag{46}
$$

ii)
$$
\lim_{n\to\infty} A_2 = \lim_{n\to\infty} \frac{1}{n} \Bigg| \sum_{k=n+2-r}^{n} [P(Y_0=1), \cdots, P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^{*}) \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^{\,*}) \end{bmatrix}
$$

$$
- \sum_{k=n+2-r}^{n} [P(Y_0=1), \cdots, P(Y_0=N_y)] (T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \Bigg| = 0. \tag{47}
$$

iii)
$$
A_3 = \frac{1}{n} \Bigg| \sum_{k=2}^{n+1-r} [P(Y_0=1), \cdots, P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^{*}) \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^{\,*}) \end{bmatrix}
$$

$$
- \sum_{k=2}^{n+1-r} [P(Y_0=1), \cdots, P(Y_0=N_y)] (T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \Bigg|
$$

$$
\leq \frac{1}{n} \Bigg| \sum_{k=2}^{n+1-r} [P(Y_0=1), \cdots, P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^{*}) \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^{\,*}) \end{bmatrix}
$$

$$
- \sum_{k=2}^{n+1-r} [P(Y_0=1), \cdots, P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^{*}) \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix} \Bigg|
$$

$$
+ \frac{1}{n} \Bigg| \sum_{k=2}^{n+1-r} [P(Y_0=1), \cdots, P(Y_0=N_y)] (\prod_{j=0}^{k-2} T_{n-j}^{*}) \begin{bmatrix} I_1(\vec{p}_1^{\,*}) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^{\,*}) \end{bmatrix}
$$

$$- \sum_{k=2}^{n+1-r} [P(Y_0 = 1), \cdots, P(Y_0 = N_y)](T^*)^{k-1} \begin{bmatrix} I_1(\vec{p}_1^*) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^*) \end{bmatrix}$$

$$\leq \frac{1}{n} \sum_{k=2}^{n+1-r} |||[P(Y_0 = 1), \cdots, P(Y_0 = N_y)]|||_\infty \left\| \prod_{j=0}^{k-2} T_{n-j}^* \right\|_\infty \left\| \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^*) - I_1(\vec{p}_1^*) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^*) - I_{N_y}(\vec{p}_{N_y}^*) \end{bmatrix} \right\|_\infty$$

$$+ \frac{1}{n} \sum_{k=2}^{n+1-r} |||[P(Y_0 = 1), \cdots, P(Y_0 = N_y)]|||_\infty \left\| \prod_{j=0}^{k-2} T_{n-j}^* - (T^*)^{k-1} \right\|_\infty \left\| \begin{bmatrix} I_1(\vec{p}_1^*) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^*) \end{bmatrix} \right\|_\infty. \quad (48)$$

Clearly, we have

$$|||[P(Y_0 = 1), \cdots, P(Y_0 = N_y)]|||_\infty = 1, \quad (49)$$

$$\left\| \prod_{j=0}^{k-2} T_{n-j}^* \right\|_\infty = 1, \quad (50)$$

$$\left\| \begin{bmatrix} I_1(\vec{p}_1^*) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y}^*) \end{bmatrix} \right\|_\infty \leq \log N_y. \quad (51)$$

By (43)

$$\left\| \prod_{j=0}^{k-2} T_{n-j}^* - (T^*)^{k-1} \right\|_\infty \leq \varepsilon, \quad (52)$$

since $n + 2 - k \geq r + 1 = \max(r_1, r_2) + 2 > r_1$ if $k \leq n + 1 - r$. By (44)

$$\left\| \begin{bmatrix} I_1(\vec{p}_{1,n+1-k}^*) - I_1(\vec{p}_1^*) \\ \vdots \\ I_{N_y}(\vec{p}_{N_y,n+1-k}^*) - I_{N_y}(\vec{p}_{N_y}^*) \end{bmatrix} \right\|_\infty \leq \varepsilon, \quad (53)$$

since $n + 1 - k \geq r = \max(r_1, r_2) + 1 \geq r_2$ if $k \leq n + 1 - r$.

Putting (49)-(53) into (48) yields

$$\limsup_{n \to \infty} A_3 \leq (\log N_y + 1)\varepsilon. \quad (54)$$

Now put (46), (47) and (54) back into (45), obtaining

$$\limsup_{n \to \infty} \frac{1}{n} |C_{y_0,n} - C_{y_0,n}^*| \leq (\log N_y + 1)\varepsilon.$$

Since here ε is arbitrary, it follows that

$$\lim_{n\to\infty} \frac{1}{n}|C_{y_0,n} - C^*_{y_0,n}| = 0. \tag{55}$$

The strong irreducibility of $\tilde{T}$ guarantees that the output process $\{Y_k, k = 0, 1, \cdots, n\}$ induced by any stationary input distribution form an irreducible Markov chain on the state space $\{1, 2, \cdots, N_y\}$. So from the analysis in Section V, we have

$$\lim_{n\to\infty} \frac{C^*_{y_0,n}}{n} = \sum_{l=1}^{N_y} \pi_l^* I_{l,1}(\vec{p_l}^*),$$

where $(\pi_1^*, \pi_2^*, \cdots, \pi_{N_y}^*)$ is the equilibrium distribution of the channel output process induced by the stationary input distribution $(\vec{p_1}^*, \vec{p_2}^*, \cdots, \vec{p_{N_y}}^*)$. Hence

$$\lim_{n\to\infty} \frac{C_{y_0,n}}{n} = \sum_{l=1}^{N_y} \pi_l^* I_{l,1}(\vec{p_l}^*). \tag{56}$$

It is obvious that the limiting average directed mutual information induced by the optimal input distribution is always greater than or equal to the limiting average directed mutual information induced by any stationary input distribution, so we have

$$\lim_{n\to\infty} \frac{C_{y_0,n}}{n} \geq \max_{(\vec{p_1}, \vec{p_2}, \cdots, \vec{p_{N_y}})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\}. \tag{57}$$

Combining (56) and (57) and noticing the arbitrariness of y_0, we can conclude that

$$\lim_{n\to\infty} \frac{C_{i,n}}{n} = \max_{(\vec{p_1}, \vec{p_2}, \cdots, \vec{p_{N_y}})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\} = \sum_{l=1}^{N_y} \pi_l^* I_{l,1}(\vec{p_l}^*), \tag{58}$$

for all i $(i = 1, 2, \cdots, N_y)$.   ∎

Remark: The condition $Q_i \in \Psi_{N_y \times N_x} (i = 1, 2, \cdots, N_y)$ is introduced for purely technical reasons. It enables us to prove that $\vec{p}_{i,n}^*$ converges to $\vec{p}_i^*$ exponentially fast by exploiting the strict concavity property of mutual information function. Theorem 4 still holds when this condition is removed. However, the proof will be less direct compared with the current one.

VIII. DIRECT CHANNEL CODING THEOREM

In this section we prove the direct channel coding theorem and suggest a coding scheme for our channel model.

Theorem 5 (direct channel coding theorem): If $\tilde{T}$ is strongly irreducible and strongly aperiodic, then all rates less than $\max\limits_{(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})} \left\{ \sum\limits_{l=1}^{N_y} \pi_l I_{l,1} \right\}$ are achievable, where $(\pi_1, \pi_2, \cdots, \pi_{N_y})$ is the equilibrium distribution of the channel output process induced by the stationary input distribution $(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})$.

Proof: We shall implicitly assume that $P(Y_0 = y_0) = 1$.

It has been shown in [10] that the general formula for the capacity of feedback channels is

$$C_{y_0}^{fd} = \sup_{p(X_1^\infty) \in \mathcal{P}(X_1^\infty)} \underline{I}(X \rightarrow Y) = \sup_{p(X_1^\infty) \in \mathcal{P}(X_1^\infty)} \underset{\text{in probability}}{\lim\inf} \frac{1}{n} \vec{i}(X^n; Y^n),$$

where

1) $\mathcal{P}(X_1^\infty) = \{ \{ P(X_k | x_1^{k-1}, y_0^{k-1}) \}_{k=1}^\infty \}$ is the set of all channel input distributions;

2) $\vec{i}(X^n; Y^n) = \log \dfrac{P(X_1^n, Y_1^n | Y_0)}{P(X_1 | Y_0) P(Y_1^n | Y_0) \prod\limits_{k=2}^{n} P(X_k | X_1^{k-1}, Y_0^{k-1})}$;

3) The $\lim\inf$ in probability of a sequence of random variables $\{Z_n\}_{n=1}^\infty$, denoted by $\underset{\text{in probability}}{\lim\inf} Z_n$, is defined as the largest extended real number c such that $\forall \varepsilon > 0$, $\lim\limits_{n \rightarrow \infty} Pr[Z_n \leq c - \varepsilon] = 0$.

Let $(\vec{p}_1^*, \vec{p}_2^*, \cdots, \vec{p}_{N_y}^*) \in \arg \max\limits_{(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})} \left\{ \sum\limits_{l=1}^{N_y} \pi_l I_{l,1} \right\}$. Let $P(X_k = i | X_1^{k-1}, Y_0^{k-2}, Y_{k-1} = j) = \vec{p}_j^*(i)$, for $k = 1, 2, \cdots$; $i = 1, 2, \cdots, N_x$; $j = 1, 2, \cdots, N_y$. Since $\tilde{T}$ is strongly irreducible and strongly aperiodic, we can check that the joint process $\{X_k, Y_k, Y_{k-1}\}_{k=1}^\infty$ induced by the stationary input distribution $(\vec{p}_1^*, \vec{p}_2^*, \cdots, \vec{p}_{N_y}^*)$ constitutes an irreducible Markov chain with $P(X_k = i, Y_k = j, Y_{k-1} = l)$ converging to $\pi_l^* \vec{p}_l^*(i) Q_l(j|i)$, where $(\pi_1^*, \pi_2^*, \cdots, \pi_{N_y}^*)$ is the equilibrium distribution of the channel output process induced by the stationary input distribution $(\vec{p}_1^*, \vec{p}_2^*, \cdots, \vec{p}_{N_y}^*)$. Note:

1) The recurrent state space of this Markov chain maybe smaller than $\{1, 2, \cdots, N_x\} \times \{1, 2, \cdots, N_y\} \times \{1, 2, \cdots, N_y\}$. The strong irreducibility and strong aperiodicity of $\tilde{T}$ only guarantees that $\{Y_k, k = 0, 1, \cdots\}$ induced by the stationary input distribution $(\vec{p}_1^*, \vec{p}_2^*, \cdots, \vec{p}_{N_y}^*)$ form an irreducible and aperiodic Markov chain on the state space $\{1, 2, \cdots, N_y\}$.

2) We assume the strong irreducibility and strong aperiodicity of $\tilde{T}$ only for simplicity. More generally, we would need to decompose the state space into disjoint irreducible closed sets, whereupon the proof would proceed along the same lines.

Now we have

$$
\begin{aligned}
\vec{i}^*(X^n; Y^n) &= \log \frac{P(X_1^n, Y_1^n | Y_0)}{P(X_1 | Y_0) P(Y_1^n | Y_0) \prod\limits_{k=2}^{n} P(X_k | X_1^{k-1}, Y_0^{k-1})} \\
&= \log \frac{\prod\limits_{k=1}^{n} P(X_k, Y_k | Y_{k-1})}{\prod\limits_{k=1}^{n} P(X_k | Y_{k-1}) P(Y_k | Y_{k-1})} \\
&= \sum_{k=1}^{n} \log \frac{P(X_k, Y_k | Y_{k-1})}{P(X_k | Y_{k-1}) P(Y_k | Y_{k-1})}.
\end{aligned}
$$

Since $\log \frac{P(X_k, Y_k | Y_{k-1})}{P(X_k | Y_{k-1}) P(Y_k | Y_{k-1})}$ is a function of (X_k, Y_k, Y_{k-1}), and $\{(X_k, Y_k, Y_{k-1})\}$ is an ergodic process, it follows that $\{\log \frac{P(X_k, Y_k | Y_{k-1})}{P(X_k | Y_{k-1}) P(Y_k | Y_{k-1})}\}$ is also an ergodic process. So we have

$$
\begin{aligned}
\frac{1}{n} \vec{i}^*(X^n; Y^n) &= \frac{1}{n} \sum_{k=1}^{n} \log \frac{P(X_k, Y_k | Y_{k-1})}{P(X_k | Y_{k-1}) P(Y_k | Y_{k-1})} \\
&\to \sum_{i,j,l} \pi_l^* \vec{p}_l^{\,*}(i) Q_l(j|i) \log \frac{Q_l(j|i)}{\sum\limits_{i=1}^{N_x} \vec{p}_l^{\,*}(i) Q_l(j|i)} \\
&= \sum_{l=1}^{N_y} \pi_l^* \sum_{i,j} \vec{p}_l^{\,*}(i) Q_l(j|i) \log \frac{Q_l(j|i)}{\sum\limits_{i=1}^{N_x} \vec{p}_l^{\,*}(i) Q_l(j|i)} \\
&= \sum_{l=1}^{N_y} \pi_l^* I_{l,1}^* \\
&= \max_{(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\} \text{ in probability,}
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\underline{I}^*(X \to Y) &= \liminf_{\text{in probability}} \frac{1}{n} \vec{i}^*(X^n; Y^n) \\
&= \max_{(\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_{N_y})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\}.
\end{aligned}
$$

Now the proof is complete since

$$
\begin{aligned}
C_{y_0}^{fd} &= \sup_{p(X_1^\infty)\in\mathcal{P}(X_1^\infty)} \underline{I}(X \to Y) \\
&\geq \underline{I}^*(X \to Y) \\
&= \max_{(\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_{N_y})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\}.
\end{aligned}
$$

$\blacksquare$

Theorem 6: If $\tilde{T}$ is strongly irreducible and strongly aperiodic and $Q_i \in \Psi_{N_y \times N_x}(i = 1, 2, \cdots, N_y)$, then

$$
C_{y_0}^{fd} = \max_{(\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_{N_y})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\}
$$

and is independent of the initial state y_0.

Proof: By Theorem 1,

$$
C_{y_0}^{fd} \leq \limsup_{n\to\infty} \frac{C_{y_0,n}}{n}.
$$

If $\tilde{T}$ is strongly irreducible and strongly aperiodic, $Q_i \in \Psi_{N_y \times N_x}(i = 1, 2, \cdots, N_y)$, by Theorem 4,

$$
\lim_{n\to\infty} \frac{C_{y_0,n}}{n} = \max_{(\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_{N_y})} \left\{ \sum_{l=1}^{N_y} \pi_l I_{l,1} \right\}.
$$

Theorem 5 shows the tightness of this upper bound and completes the proof. $\blacksquare$

In the remainder part of this section, we suggest a coding scheme for our channel model, which also makes the meaning of the capacity formula transparent.

We first consider the channel model shown in Fig. 5, in which the state information is
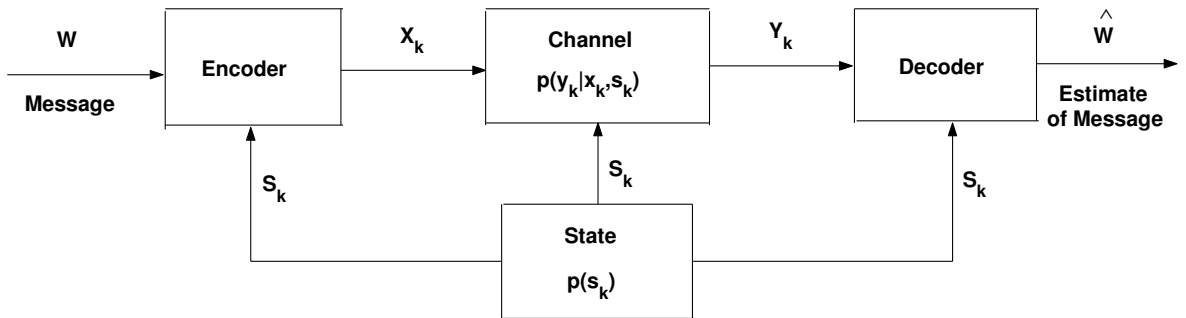

Fig. 5 Finite State Channel with State Information Available at Transmitter and Receiver

simultaneously available to both the transmitter T and the receiver R. If the state process is stationary and ergodic, then it's well known [9] [20] [21] that the channel capacity is

$$C^{TR} = \sum_{j=1}^{K} \pi(s_j) \max_{p(X|S=s_j)} I(X;Y|S=s_j),$$

where K is the number of channel states and π is the stationary distribution of the ergodic state process $\{S_k\}_{k=1}^{\infty}$. The following is the outline of the coding scheme for this channel [21]:

Let

$$C_j^{TR} = \max_{p(X|S=s_j)} I(X;Y|S=s_j).$$

Fix the blocklength N. Let N_j be the number of times during the N symbols for which the channel state is s_j, i.e.,

$$N_j = |\{k|S_k = s_j, 1 \le k \le N\}|.$$

Let $\overline{N}_j = E[N_j]$. Since the state process is stationary and ergodic, we have $\overline{N}_j = \pi(s_j)N$ and

$$\lim_{N \to \infty} \frac{N_j}{N} = \pi(s_j)$$

in probability.

A $(N, 2^{N(C^{TR}-\delta)}, \varepsilon)$ code for this channel is constructed by multiplexing K codes $(\overline{N}_j, 2^{\overline{N}_j(C_j^{TR}-\frac{\delta}{K})}, \frac{\varepsilon}{K})$ $(j = 1, 2, \cdots, K)$, where code $(\overline{N}_j, 2^{\overline{N}_j(C_j^{TR}-\frac{\delta}{K})}, \frac{\varepsilon}{K})$ corresponds to the channel state $s_j(j = 1, 2, \cdots, K)$. By doing this, we actually decompose the channel into K memoryless channels and the existence of these codes follows immediately from the direct coding theorem for memoryless channels. Since N_j is not necessarily equivalent to $\overline{N}_j$, the codes are truncated if $N_j < \overline{N}_j$ and zero-filled if $N_j > \overline{N}_j$. Represent each message W as a K-dimensional vector $[W_1, \cdots, W_K]$ with $W_j \in \{1, \cdots, 2^{\overline{N}_j(C_j^{TR}-\frac{\delta}{K})}\}(j = 1, 2, \cdots, K)$ and map the jth index (i.e. W_j) into a codeword form the jth code (i.e., the code with parameters $(\overline{N}_j, 2^{\overline{N}_j(C_j^{TR}-\frac{\delta}{K})}, \frac{\varepsilon}{K})$), for $1 \le j \le K$. If $S_k = s_j$, then the transmitter sends as the kth symbol the next unsent symbol of the codeword corresponding to the jth index of the message from the jth code $(k = 1, 2, \cdots, N; j = 1, 2, \cdots, K)$.

Since the receiver knows exactly the state information that was used at the transmitter, it can demultiplex the received stream into K separate codewords and decode them. Since the state process is stationary and ergodic, as $N \to \infty$, the rate

$$\sum_{j=1}^{K} \frac{N_j}{N} C_j^{TR} \to \sum_{j=1}^{K} \pi(s_j) C_j^{TR} = C^{TR}$$

is achievable.

It's easy to see that the capacity formula of this channel model closely resembles ours in that both of them can be represented as the average of mutual information over the stationary distribution of the channel state process. This suggests that the multiplexing coding scheme may work in our channel model as well. But we should note that in our model, the current channel state is the previous channel output, so it will be affected by the channel input. Although we have shown that if the input process is stationary, the output process (i.e., channel state process) is an irreducible homogeneous Markov chain and thus is ergodic, it does not imply that the output process induced by a specific codeword is still ergodic (or close to ergodic). The random coding argument based on strong typicality tells us that for a discrete memoryless channel with finite input and output alphabets, the statistics of the output process induced by a good codeword are close to those of the output process induced by the optimal input distribution that achieves the channel capacity. But strong typicality only guarantees that the statistics of a whole codeword are close to the optimal input distribution, while a truncated version may not have this property. In the multiplexing scheme, the output process is generated by several multiplexed codewords, each of which is designed for its corresponding decomposed memoryless channel. And the length of a multiplexed codeword is proportional to the stationary probability measure assigned on its corresponding state. We want the statistics of the output process induced by the multiplexed codewords to be close to the equilibrium distribution induced by the optimal stationary input. Clearly, this depends highly on the cooperation of the multiplexed codewords. Even a small fluctuation of the statistics in a portion of a multiplexed codeword may have a domino effect on the transmission of the other multiplexed codewords and finally make the output process deviate from the desired distribution. The result of the large deviation in the output process is that the symbols in some multiplexed codewords are totally sent while many symbols in some others of the multiplexed codewords are still unsent. So, in order to guarantee the stability of the multiplexing scheme, we need the multiplexed codewords to behave better than those codewords in the sense that their empirical distributions well-approximate the input distribution to which they correspond. (Theorem 7) below shows that for a discrete memoryless channel with finite input and output alphabets, the empirical statistics of each of the words of a good code can be made to closely approximate the optimum input distribution even despite their being subjected to truncation.

Before proving Theorem 7, we need to give a definition.

We consider an information source $\{X_k, k \geq 1\}$, where X_k are i.i.d. with distribution $p(x)$. Let $|\mathcal{X}|$ denote the cardinality of the set of values X_k may assume. Here we suppose $|\mathcal{X}| < \infty$ and $p(x) > 0$ for all $x \in \mathcal{X}$.

Definition 7: The super-typical set $\mathcal{T}_\delta^{m,n}$ with respect to $p(X)$ is the set of sequences $x_1^n = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ such that

$$\sum_{x \in \mathcal{X}} |\frac{1}{r} N(x, x_1^r) - p(x)| \leq \frac{\delta}{|\mathcal{X}|}$$

for all r ($m \leq r \leq n$), where $N(x, x_1^r)$ is the number of occurrences of x in the sequence x_1^r, and δ is an arbitrarily small positive real number. The members of $\mathcal{T}_\delta^{m,n}$ are called (δ, m, n) super-typical sequences.

It's clear that the super-typical set $\mathcal{T}_\delta^{m,n}$ is a subset of the strongly typical set $\mathcal{T}_\delta^n$. But the following lemma says that as m and n go to infinity, these two sets have no essential difference.

Lemma 5: For any $\varepsilon > 0$, $\delta > 0$, there exists a positive integer M such that when $n > m > M$, we have

$$P(X_1^n \in \mathcal{T}_\delta^{m,n}) > 1 - \varepsilon.$$

Proof: We write

$$N(x, X_1^r) = \sum_{i=1}^{r} B_i(x),$$

where $B_i(x) = 1_{\{X_i = x\}}$. Then $B_i(x), i = 1, 2, \cdots, r$, are i.i.d. random variables with $P\{B_i(x) = 1\} = p(x)$ and $P\{B_i(x) = 0\} = 1 - p(x)$. Note that

$$EB_i(x) = (1 - p(x)) \cdot 0 + p(x) \cdot 1 = p(x).$$

Let $U_i(x) = B_i(x) - p(x)$. Clearly

$$EU_i(x) = 0,$$

$$EU_i^2(x) = p(x) - p^2(x),$$

and

$$EU_i^4(x) = p(x) - 4p^2(x) + 6p^3(x) - 3p^4(x) \leq 7.$$

Let $S_r(x) = U_1(x) + \cdots + U_r(x)$. Now

$$ES_r^4(x) = E[\sum_{i=1}^{r} U_i(x)]^4 = E[\sum_{1 \le i,j,k,l \le r} U_i(x)U_j(x)U_k(x)U_l(x)].$$

Terms in the sum of the form $E[U_i^3(x)U_j(x)]$, $E[U_i^2(x)U_j(x)U_k(x)]$ and $E[U_i(x)U_j(x)U_k(x)U_l(x)]$ are 0 (if i, j, k, l are distinct) since the expectation of the product is the product of the expectations, and in each case one of the terms has expectation 0. The only terms that do not vanish are those of form $EU_i^4(x)$ and $E[U_i^2(x)U_j^2(x)]$. There are r and $3r(r-1)$ of these terms, respectively. The last observation implies

$$ES_r^4(x) = rEU_1^4(x) + 3(r^2 - r)[EU_1^2(x)]^2 \le 7r + 3r^2 - 3r \le 7r^2.$$

Chebyshev's inequality gives us

$$P\{|\frac{1}{r}N(x, X_1^r) - p(x)| > \frac{\delta}{|\mathcal{X}|^2}\} = P\{|S_r(x)| > \frac{r\delta}{|\mathcal{X}|^2}\} \le \frac{ES_r^4(x)}{(r\delta)^4}|\mathcal{X}|^8 \le \frac{7|\mathcal{X}|^8}{r^2\delta^4}.$$

Hence

$$
\begin{aligned}
P(X_1^n \notin \mathcal{T}_\delta^{m,n}) &\le \sum_{r=m}^{n} P\{\sum_x |\frac{1}{r}N(x, X_1^r) - p(x)| > \frac{\delta}{|\mathcal{X}|}\} \\
&\le \sum_{r=m}^{n} \sum_x P\{|\frac{1}{r}N(x, X_1^r) - p(x)| > \frac{\delta}{|\mathcal{X}|^2}\} \\
&\le \sum_{r=m}^{n} \sum_x \frac{7|\mathcal{X}|^8}{r^2\delta^4} \\
&\le \frac{7|\mathcal{X}|^9}{\delta^4} \sum_{r=m}^{\infty} \frac{1}{r^2}.
\end{aligned}
$$

Since $\sum_{r=1}^{\infty} \frac{1}{r^2} = \frac{\pi^2}{6} < \infty$, there exists a positive integer M, when $m > M$, we have $\frac{7|X|^9}{\delta^4} \sum_{r=m}^{\infty} \frac{1}{r^2} < \varepsilon$. Hence

$$P(X_1^n \in \mathcal{T}_\delta^{m,n}) = 1 - P(X_1^n \in \mathcal{T}_\delta^{m,n}) > 1 - \varepsilon.$$

when $n > m > M$. ∎

This lemma implies that no loss of generality results from restricting attention to the super-typical set. It is obvious that the random coding argument based on strong typicality can be translated to an argument based on super typicality without any change. So for any discrete memoryless channel with finite input and output alphabets, there exists a good codebook, all the codewords of which are super typical. This is summarized in Theorem 7.

Theorem 7: Let $Q(y|x)$ be any discrete memoryless channel with finite input and output alphabets. For any input distribution $p(x)$ that is consistent with $Q(y|x)$, for any $0 < \varepsilon_1, \varepsilon_2, \delta < 1$, there exists n_0 such that for all $n > m > n_0$, there exists a (n, M, ε_1) code with average probability of error less than ε_1 and $\frac{1}{n} \log M > I(p, Q) - \varepsilon_2$. Furthermore, each of these M codewords is (δ, m, n) super-typical with respect to $p(x)$.

Proof: The proof is omitted since it is almost the same as the standard proof of direct coding theorem for memoryless channel based on weak typicalicy or strong typicality; see [22]. The only difference is that we require that the randomly generated codewords satisfy super typicality. And Lemma 5 assures us that there is no essential difference between strong typicality and super typicality when the alphabets are finite.   ∎

We want to point out another difference between the multiplexing coding scheme for our feedback channel and that for the channel model shown in Fig. 5. Zero filling is not a good choice for our channel when $N_j > \overline{N}_j$, since this will make the statistics of the codeword deviate from the corresponding input distribution and thus induce a big deviation at the output. Instead, we will fill in the letters so as to ensure that the lengthened codewords still satisfy super typicality and/or we will try to drive the channel into those states that still have unsent symbols.

IX. CONCLUSION

We derived a simple formula for the capacity of finite state Markov channels with feedback when the channel transition probability satisfies certain conditions. Actually the same capacity formula holds under much weaker conditions. It will be shown elsewhere that, based on the classification of Markov decision processes [23], the capacity of Markov channels with feedback can be studied in full generality.

Finally, we mention the relationship between the channel whose state process cannot be affected by the input and the one whose state process can be affected. We assume in both cases the realization of the state process is available both at transmitter and receiver. For the channel whose state process cannot be affected by the input, the conventional multiplexing coding scheme [20] [21] can be viewed as a greedy algorithm which tries to maximize the immediate mutual information. For the channel whose state process can be affected by the input, this greedy algorithm is not optimal since we not only want to maximize the immediate mutual information but also want to visit the preferable states as often as possible. So the optimal coding scheme

is a tradeoff between these two goals. From this perspective, it seems appropriate to call this type of code an error correction and state control code in contrast with a conventional error correction code. Since the optimal coding scheme needs to exploit the ergodicity of the state process, however, it usually causes a long delay in decoding, especially when the state space is big and/or the probability measures assigned to some states are close to zero. Therefore, in some delay-limited applications, certain kinds of greedy schemes are more attractive. In this sense a channel whose state process can be affected by the input is considerably more flexible than one whose state process cannot be affected, since we can use "idle" periods to drive the channel into preferable states and thereby increase the efficiency when we really need to use the channel for information transmission. In a quite general manner of speaking, such a channel can be "matched" to an information source in the spirit of [11]. Perhaps this is one of the reasons why many real neural networks possess structure that subscribes to the channel models treated in this paper.

APPENDIX

Lemma 6: For any $N_y \times N_x$ channel transition probability matrix Q, if $rank(Q) = N_x$, then $Q \in \Psi_{N_y \times N_x}$.

Proof: Let $\vec{\triangle p}$ be an arbitrary N_x-dimensional vector with the constraint $\sum_{i=1}^{N_x} \vec{\triangle p}(i) = 0$. Let $\vec{\triangle q} = Q\vec{\triangle p}$. We have

$$Q^T \vec{\triangle q} = Q^T Q \vec{\triangle p}.$$

If $rank(Q) = N_x$, then $Q^T Q$ is invertible, whereupon it follows that

$$\vec{\triangle p} = (Q^T Q)^{-1} Q^T \vec{\triangle q}.$$

So

$$\left\| \vec{\triangle p} \right\|_2 \leq \left\| (Q^T Q)^{-1} Q^T \right\|_2 \left\| \vec{\triangle q} \right\|_2,$$

which implies that

$$\frac{\partial^2 H(Q\vec{p})}{\partial(\vec{\triangle p})^2} \leq \frac{1}{||(Q^T Q)^{-1} Q^T||_2^2} \frac{\partial^2 H(\vec{q})}{\partial(\vec{\triangle q})^2}. \tag{59}$$

We have

$$
\begin{aligned}
\frac{\partial^2 H(\vec{q})}{\partial(\vec{\triangle q})^2} &= \sum_{i=1}^{N_y-1} \sum_{j=1}^{N_y-1} \frac{\partial^2 H(\vec{q})}{\partial \vec{q}(i) \partial \vec{q}(j)} \cos \alpha_i \cos \alpha_j \\
&= -\sum_{i=1}^{N_y-1} \frac{1}{\vec{q}(i)} \cos^2 \alpha_i - \sum_{i=1}^{N_y-1} \sum_{j=1}^{N_y-1} \frac{1}{\vec{q}(N_y)} \cos \alpha_i \cos \alpha_j \\
&= -\sum_{i=1}^{N_y-1} \frac{1}{\vec{q}(i)} \cos^2 \alpha_i - \frac{1}{\vec{q}(N_y)} \left(\sum_{j=1}^{N_y-1} \cos \alpha_j \right)^2 \leq -\sum_{i=1}^{N_y-1} \frac{1}{\vec{q}(i)} \cos^2 \alpha_i \\
&\leq -\sum_{i=1}^{N_y-1} \cos^2 \alpha_i = -1,
\end{aligned}
\tag{60}
$$

where

$$
\cos \alpha_i = \frac{\vec{\triangle q}(i)}{\sqrt{\sum_{j=1}^{N_y-1} [\vec{\triangle q}(j)]^2}} (i = 1, 2, \cdots, N_y - 1).
$$

Combining (59) and (60), we have

$$
\frac{\partial^2 H(Q\vec{p})}{\partial(\vec{\triangle p})^2} \leq -\frac{1}{||(Q^T Q)^{-1} Q^T||_2^2}.
\tag{61}
$$

Since

$$
I(\vec{p}, Q) = H(Q\vec{p}) - \sum_{i=1}^{N_x} \vec{p}(i) \sum_{j=1}^{N_y} Q(j,i) \log \frac{1}{Q(j,i)}.
$$

and $\sum_{i=1}^{N_x} \vec{p}(i) \sum_{j=1}^{N_y} Q(j,i) \log \frac{1}{Q(j,i)}$ is linear with respect to $\vec{p}$, it follows that

$$
\frac{\partial^2 I(\vec{p}, Q)}{\partial(\vec{\triangle p})^2} = \frac{\partial^2 H(Q\vec{p})}{\partial(\vec{\triangle p})^2} \leq -\frac{1}{||(Q^T Q)^{-1} Q^T||_2^2}.
\tag{62}
$$

Now let $\tilde{\mathcal{P}} = \mathcal{P}$. It's easy to verify the Condition (i) and (ii) in Definition 6 are trivially satisfied. Furthermore, for any nonidentical $\vec{p_1}, \vec{p_2} \in \tilde{\mathcal{P}}$ and $\vec{l}$ with the direction from $\vec{p_1}$ to $\vec{p_2}$, by (62), we have

$$
\begin{aligned}
\frac{\partial I(\vec{p_2}, Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p_1}, Q)}{\partial \vec{l}} &= \frac{\partial^2 I}{\partial \vec{l}^2}(\vec{p_1} + \theta(\vec{p_2} - \vec{p_1}), Q)||\vec{p_1} - \vec{p_2}||_2 \\
&\leq -\frac{1}{||(Q^T Q)^{-1} Q^T||_2^2}||\vec{p_1} - \vec{p_2}||_2,
\end{aligned}
$$

where $\theta \in (0, 1)$. So Condition (iii) in Definition 6 is also satisfied. ∎

Now we begin to discuss the implications of three conditions in Definition 6, which immediately suggests a way to construct $\tilde{\mathcal{P}}$ in the general case.

Consider the following equivalence relation:

$$\vec{p}_1 \sim \vec{p}_2 \quad \text{if and only if} \quad \vec{p}_1 - \vec{p}_2 \text{ is in the null space of } Q.$$

Let $\{\mathcal{P}_\alpha\}$ be a partition of $\mathcal{P}$ generated by the above equivalence relation such that for any α, $\vec{p}_1, \vec{p}_2 \in \mathcal{P}_\alpha$ if and only if $\vec{p}_1 \sim \vec{p}_2$.

Condition (i) implies that $\tilde{\mathcal{P}} \cap \mathcal{P}_\alpha \neq \emptyset$ for any α. Condition (iii) implies that for each α, $\tilde{\mathcal{P}} \cap \mathcal{P}_\alpha$ contains at most one element since if there exist two nonidentical $\vec{p}_1, \vec{p}_2 \in \tilde{\mathcal{P}} \cap \mathcal{P}_\alpha$, then

$$
\begin{aligned}
\frac{\partial I(\vec{p}_2, Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p}_1, Q)}{\partial \vec{l}} &= \frac{\partial^2 I}{\partial \vec{l}^2}(\vec{p}_1 + \theta(\vec{p}_2 - \vec{p}_1)) \|\vec{p}_1 - \vec{p}_2\|_2 \\
&= \frac{\partial^2 H}{\partial \vec{l}^2}(Q(\vec{p}_1 + \theta(\vec{p}_2 - \vec{p}_1))) \|\vec{p}_1 - \vec{p}_2\|_2 \\
&= 0,
\end{aligned}
$$

where the last equality follows from the fact that $\vec{l}$ is in the null space of Q. Hence $\tilde{\mathcal{P}} \cap \mathcal{P}_\alpha$ contains exactly one element.

Let $\vec{h} = [h_1, \cdots, h_{N_x}]^T$ with $h_i = \sum_{j=1}^{N_y} Q(j, i) \log \frac{1}{Q(j,i)}$ $(i = 1, 2, \cdots, N_x)$. Condition (ii) implies that

$$
\begin{aligned}
(\tilde{\mathcal{P}} \cap \mathcal{P}_\alpha) &\subseteq \arg\max_{\vec{p} \in \mathcal{P}_\alpha} I(\vec{p}, Q) \\
&= \arg\max_{\vec{p} \in \mathcal{P}_\alpha} [H(Q\vec{p}) - \vec{h}^T \vec{p}] \\
&= \arg\min_{\vec{p} \in \mathcal{P}_\alpha} \vec{h}^T \vec{p}
\end{aligned}
$$

where the last equality is because $H(Q\vec{p})$ is a constant for $\vec{p} \in \mathcal{P}_\alpha$.

Now we are ready to present the procedure of constructing $\tilde{\mathcal{P}}$.

1) Let $N_{x \wedge y} = \min(N_x, N_y)$. By singular value decomposition, there exist orthogonal matrices $U \in \mathcal{R}^{N_y \times N_y}$ and $V \in \mathcal{R}^{N_x \times N_x}$ such that

$$Q = U\Sigma V^T,$$

where

$$\Sigma = diag(\sigma_1, \cdots, \sigma_{N_{x \wedge y}}) \in \mathcal{R}^{N_y \times N_x}$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{N_{x \wedge y}} \geq 0$.

2) Let $r = rank(Q)$. Then we have $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{N_{x \wedge y}}$. Let $\Sigma_r = diag(\sigma_1, \cdots, \sigma_r, 1, \cdots, 1) \in \mathcal{R}^{N_y \times N_y}$. Partition V as $[\underbrace{V_r}_{r} \ \underbrace{\tilde{V}_r}_{N_x - r}]$ and let $\hat{V}_r = [\underbrace{V_r}_{r} \ \underbrace{\mathbf{0}}_{N_y - r}]$. Then

$$\vec{p} = \hat{V}_r \Sigma_r^{-1} U^T \vec{q} + \tilde{V}_r \vec{k}^*, \tag{63}$$

where $\vec{k}^* = [k_1^*, \cdots, k_{N_x - r}^*]^T$ is a solution to the following constrained optimization problem:

$$\arg\min_{\vec{k}} \vec{h}^T \tilde{V}_r \vec{k} \tag{64}$$

subject to $(\hat{V}_r \Sigma_r^{-1} U^T \vec{q} + \tilde{V}_r \vec{k}) \in \mathcal{P}$.

3)

$$\tilde{\mathcal{P}} \triangleq \{\hat{V}_r \Sigma_r^{-1} U^T \vec{q} + \tilde{V}_r \vec{k}^* : \vec{q} \in Q\mathcal{P}\},$$

where $Q\mathcal{P} \triangleq \{Q\vec{p} : \vec{p} \in \mathcal{P}\}$.

It is easy to check that the resulting $\tilde{\mathcal{P}}$ satisfies Conditions (i) and (ii) in Definition 6 and we have

$$\frac{\partial I(\vec{p}_2, Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p}_1, Q)}{\partial \vec{l}} < 0$$

for any nonidentical $\vec{p}_1, \vec{p}_2 \in \tilde{\mathcal{P}}$. We are unable to prove that $[\frac{\partial I(\vec{p}_2, Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p}_1, Q)}{\partial \vec{l}}]/||\vec{p}_2 - \vec{p}_1||_2$ is uniformly bounded away from 0 as required in Definition 6. However we believe it is true under fairly general conditions.

Now consider the following maximization problem:

$$\max_{\vec{p} \in \mathcal{P}} \{I(\vec{p}, Q) + \vec{c}^T Q\vec{p}\}, \tag{65}$$

where $\vec{c}$ is an arbitrary $N_y \times 1$ real vector. Suppose the maximum is attained inside $\mathcal{P}_{\alpha^*}$ for some α^*. Let $\vec{p}^*$ be that unique element of $\tilde{\mathcal{P}} \cap \mathcal{P}_{\alpha^*}$. Now we have

$$\begin{aligned}
\max_{\vec{p} \in \mathcal{P}} \{I(\vec{p}, Q) + \vec{c}^T Q\vec{p}\} &= \max_{\vec{p} \in \mathcal{P}_{\alpha^*}} \{I(\vec{p}, Q) + \vec{c}^T Q\vec{p}\} \\
&= \max_{\vec{p} \in \mathcal{P}_{\alpha^*}} [H(Q\vec{p}) - \vec{h}^T \vec{p} + \vec{c}^T Q\vec{p}] \\
&\stackrel{(h)}{=} H(Q\vec{p}^*) + \vec{c}^T Q\vec{p}^* - \min_{\vec{p} \in \mathcal{P}_{\alpha^*}} \vec{h}^T \vec{p} \\
&= H(Q\vec{p}^*) + \vec{c}^T Q\vec{p}^* - \vec{h}^T \vec{p}^*,
\end{aligned}$$

where (h) follows from the fact that $H(Q\vec{p}) + \vec{c}^{\,T}Q\vec{p}$ is a constant for $\vec{p} \in \mathcal{P}_{\alpha^*}$. So the maximum is attained inside $\tilde{\mathcal{P}}$. Furthermore, there exists a unique $\vec{p}^{\,*} \in \tilde{\mathcal{P}}$ that achieves this maximal value. Since if there are nonidentical $\vec{p_1}^{\,*}, \vec{p_2}^{\,*} \in \tilde{\mathcal{P}}$ such that

$$I(\vec{p_1}^{\,*}, Q) + \vec{c}^{\,T}Q\vec{p_1}^{\,*} = I(\vec{p_2}^{\,*}, Q) + \vec{c}^{\,T}Q\vec{p_2}^{\,*} = \max_{\vec{p} \in \mathcal{P}}\{I(\vec{p}, Q) + \vec{c}^{\,T}Q\vec{p}\,\},$$

by taking derivative with direction $\vec{l}$ from $\vec{p_1}^{\,*}$ to $\vec{p_2}^{\,*}$, we have

$$\frac{\partial}{\partial \vec{l}}[I(\vec{p_1}^{\,*}, Q) + \vec{c}^{\,T}Q\vec{p_1}^{\,*}] = \frac{\partial}{\partial \vec{l}}[I(\vec{p_2}^{\,*}, Q) + \vec{c}^{\,T}Q\vec{p_2}^{\,*}] = 0,$$

i.e.,

$$\frac{\partial I(\vec{p_2}^{\,*}, Q)}{\partial \vec{l}} - \frac{\partial I(\vec{p_1}^{\,*}, Q)}{\partial \vec{l}} = 0,$$

which is contradictory to Condition (iii) in Definition 6.

ACKNOWLEDGMENT

REFERENCES

[1] D. Blackwell, L. Breiman, and A. J. Thomansian, (1958), "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Stat.* 29, 1209-1220.

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

[3] R. Gray, M. O. Dunham, and R. L. Gobbi, "Ergodicity of Markov channels," *IEEE Tran. Inform. Theory*, vol. IT-33, Sept. 1987.

[4] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Tran. Inform. Theory*, vol. 40, pp. 1147-1157, July 1994.

[5] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Tran. Inform. Theory*, vol. 45, pp. 2007-2019, Sept. 1999.

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

[7] C.E.Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform Theory*, vol. IT-2, pp.S8-S19 Sept. 1956.

[8] R.L.Dobrushin, "Information transmission in a channel with feedback," *Theory Prob. Appl.*, pp.395-412, Dec. 1958.

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

[10] S.C. Tatikonda, "Control under communication constraints," PhD dissertation, MIT, 2000.

[11] T. Berger, "Living information theory", *Shannon Lecture*, Lausanne, Switzerland, 2002.

[12] T. Berger and Y. Ying, "Characterizing optimum (input,output) processes for finite-state channels with feedback", *Proc. International Symposium on Information Theory*, June 30 - July 4, 2003, Yokohama, Japan, pp. 117.

[13] Y. Ying and T. Berger, "Capacity of binary feedback channels with unit memory at the output," submitted to *IEEE Trans. Inform. Theory*, 2003.

[14] J. Massey, "Causality, feedback, and directed Information," *Proceeding of the 1990 International Symposium on Information Theory and its Applications (ISITA-90)*, pp.303-305.

[15] H. Marko, " The bidirectional communication theory - a generalization of information theory," *IEEE Tran. Comm.*, vol. COM-21, pp.1345-1351, Dec. 1973.

[16] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," Ann. Probab., 8: 962-973, 1980.

[17] R. M. Gray, *Entropy and Information Theory*. New York, Springer-Verlag, 1991.

[18] R. Durret, *Probability: Theory and Examples*, Second Edition. Duxbury Press, 1995.

[19] G. H. Gloum and C. F. Van Loan, *Matrix Computations*, Third Edition. The Johns Hopkins University Press, 1996.

[20] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Tran. Inform. Theory*, vol. 43, pp.1986-1992, Nov. 1997.

[21] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Tran. Inform. Theory*, vol. 45, pp. 761-771, Mar. 1999.

[22] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[23] M. L. Puterman,*Markov Decision Processes*. Wiley Inter-science, 1994.