Feedback Capacity of a Class of Symmetric Finite-State Markov Channels

Nevroz Şen, Student Member, IEEE, Fady Alajaji, Senior Member, IEEE, and Serdar Yüksel, Member, IEEE

Abstract

We consider the feedback capacity of a class of symmetric finite-state Markov channels. Here, symmetry is defined as a generalized version of the symmetry defined for discrete memoryless channels. We show that feedback does not increase capacity for such a class of finite-state channels and that both their non-feedback and feedback capacities are achieved by a uniform independent and identically distributed (i.i.d.) input.

Index Terms

Channel capacity, finite-state Markov channels, dynamic programming, feedback capacity.

I. INTRODUCTION AND LITERATURE REVIEW

Although feedback does not increase the capacity of discrete memoryless channels (DMCs) [1], it generally increases the capacity of channels with memory. In this work, we study the feedback capacity of a class of channels with memory and show that feedback does not increase their capacity. More explicitly, we consider finite-state Markov (FSM) channels [2], [3], [4] which encompass symmetry in their channel transition matrices.

FSM channels have been widely used to effectively model wireless fading channels (e.g., cf. [5], [6], [7], [8]). A definition of symmetric finite-state Markov channels is given in [9] and

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).
The authors are with the Department of Mathematics and Statistics, Queen’s University, Kingston, ON K7L 3N6, Canada {email: nsen, fady, yuksel@mast.queensu.ca}
and capacity without feedback is calculated where it is shown that the capacity-achieving distribution is uniform and that this distribution yields a uniform output distribution. In [11], it is shown that feedback does not increase the capacity of discrete channels with modulo additive noise. It is also shown that for any channel with memory satisfying the symmetry conditions defined in [12], feedback does not increase its capacity. Recently, it has been shown that feedback does not increase the capacity of the compound Gilbert-Elliot channel [13], which is a family of FSM channels, where the capacity is achieved by applying a uniform input. In a related work, the capacity of finite-state indecomposable channels with side information at the transmitter is investigated [14]. In particular, it is shown that the capacity of finite-state indecomposable Markovian channels with (modulo) additive noise, where the noise is a deterministic function of the state, is not increased with the availability of side information at the transmitter. In a more recent work, it has been shown that it is possible to formulate the computation of feedback capacity as a dynamic programming problem and therefore it can be solved by using the value iteration algorithm under information stability conditions [15], [16]. In [17], finite-state channels with feedback, where feedback is a time-invariant deterministic function of the output samples, is considered. It is shown that if the state of the channel is known both at the encoder and the decoder then feedback does not increase capacity. In [18] and [19], directed information is used to calculate the feedback capacity of some classes of FSM channels. In particular, the channel state is assumed in [18] to be a deterministic function of the previous state and input; whereas in [19] the channel state is assumed to be a deterministic function of the output. In [20], time varying channels are modeled as FSM channels and their capacity is studied as a function of the feedback delay assuming perfect channel state information at the receiver. In addition to these results, it has also been shown that feedback does not increase the capacity for a binary erasure channel with Markovian state [21]. Although not closely related with our result, an important insight into the use of feedback in a real time casual coding context is presented in [22]. In particular, it is shown that feedback is useful in general casual coding problems of a Markov source over a noisy channel; however it is not useful if the channel is symmetric (as defined in [22]) and memoryless.

Considering the noise structure in typical communication channels and the results in the literature that we presented above, it is worth to look for the most general notion of symmetry for channels with memory under which feedback does not increase capacity. With this motivation,
we study the feedback capacity of a class of symmetric FSM channels, which we call “quasi-symmetric” FSM channels, and prove that feedback does not help increase their capacity. This result is shown by demonstrating that for an FSM channel satisfying the symmetry conditions defined in the paper, its feedback capacity is achieved by uniform input policies which implies that its non feedback capacity is also achieved by uniform input policies. Along this way, we first show the existence of a noise process, due to the symmetry characteristics of the channel, which is independent of input. With this fact, the feedback capacity problem reduces to the maximization of entropy of the output process. In the second step, we show that this entropy is maximized by uniform input policies. It should be noted that in quasi-symmetric FSM channels with more than one partitions, uniform inputs do not yield uniform outputs which is a key symmetry feature used in previous works for showing that feedback does not increase capacity for symmetric channels with memory (e.g., [11], [12]). This second step is solved via a dynamic programming approach which shows that it is possible to learn the channel via past feedback policies that affect the future input policies by modifying the induced channel that the receiver observes. We demonstrate that, when the FSM channel satisfies the condition that the column sums of its channel transition matrices are invariant with respect to the state process, it is still possible to learn the channel via past input policies however, the optimal policy is still the same even with this learning step. We also note that our result intersects with [11] and [12] when the noise process in the latter works is restricted to being Markovian, stationary and irreducible.

Here is a brief summary of the paper. We first give the definition of quasi-symmetric FSM channels. This will be followed by a section on their capacity with feedback. Next, we discuss some channels that satisfy the quasi-symmetry condition and hence conclude that their capacity does not increase with feedback. Finally, we end the paper with concluding remarks.

Throughout the paper, we will use the following notations. A random variable will be denoted by an upper case letter $X$ and its particular realization by a lower case letter $x$. The sequence of random variables $X_1, X_2, ..., X_n$ will be denoted by $X^n$ and so its realization will be $x^n$. We will represent a finite-state Markov source by a pair $[S, P]$, where $S$ is the state set and $P$ is the state transition probability matrix. We will also be assuming that the Markov processes in the paper are time-homogeneous, aperiodic and irreducible (hence ergodic).
II. QUASI-SYMMETRIC FINITE STATE MARKOV CHANNEL

A finite-state Markov channel (FSMC) (e.g., [9]) is defined by a pentad \([X, Y, S, P_S, C]\), where \(X\) is the input alphabet, \(Y\) is the output alphabet and the Markov process \(\{S_n\}_{n=1}^{\infty}, S_n \in S\) is represented by the pair \([S, P_S]\) where \(S\) is the state set and \(P_S\) is the state transition probability matrix. We assume that the sets \(X\), \(Y\) and \(S\) are all finite. The set \(C\) is a collection of transition probability distributions, \(p_C(y|x, s)\), on \(Y\) for each \(x \in X, s \in S\). We also assume that the FSM channel satisfies the following properties:

I1 Markov Property: For any integer \(i \geq 1\)
\[
P(s_i|s_{i-1}, y_{i-1}, x_{i-1}) = P(s_i|s_{i-1}).
\]  

I2 For any integer \(i \geq 1\),
\[
P(y_i|s_i, x_i, s_{i-1}, x_{i-1}, y_{i-1}) = p_C(y_i|s_i, x_i),
\]  
where \(p_C(\ldots)\) is defined by \(C\).

In this paper, we are interested in a subclass of FSM channels where the channel transition matrices, \(Q \triangleq [p_C(y|s, x)]_{xy}, s \in S\), carry some notion of symmetry which is similar to the symmetry defined for DMCs in the following sense:

**Definition 1:** A DMC with input alphabet \(X\), output alphabet \(Y\) and channel transition matrix \(Q = [p(y|x)]\) is quasi-symmetric if \(Q\) can be partitioned along its columns into weakly-symmetric sub-arrays, \(Q_1, Q_2, \ldots, Q_m\), with each \(Q_i\) having size \(|X| \times |Y_i|\), where \(Y_1 \cup \cdots \cup Y_m = Y\) and \(Y_i \cap Y_j = \emptyset\), \(\forall i \neq j\) [23]. A weakly-symmetric sub-array is a matrix whose rows are permutations of each other and whose column sums are all identically equal to a constant.

Note that for a quasi-symmetric DMC, the rows of its entire transition matrix, \(Q\), are also permutations of each other. It is also worth pointing out that the above quasi-symmetry\(^1\) notion for DMC’s encompasses Gallager’s symmetry definition [2, p.94]. A simple example of a quasi-symmetric DMC can be given by the following (stochastic, i.e., with row sums equal to 1)

\(^1\)The capacity of a quasi-symmetric DMC is achieved by a uniform input distribution and it can be expressed via a simple closed-form formula [23]: \(C = \sum_{i=1}^{m} \alpha_i C_i\) where \(\alpha_i \triangleq \sum_{y \in Y_i} P(y|x) = \text{sum of any row in } Q_i, \ i = 1, \ldots, m\), and \(C_i = \log_2 |Y_i| - H\left(\text{any row in the matrix } \frac{1}{\alpha_i} Q_i\right), \ i = 1, \ldots, m\).
transition matrix

\[
Q = \begin{pmatrix}
  a & b & c & d & e & f \\
  c & b & a & f & e & d \\
  b & a & c & e & d & f \\
  c & a & b & f & d & e
\end{pmatrix}
\]  \hspace{1cm} (3)

for which \(a + b = 2c\) and \(d + e = 2f\), and it can be partitioned along its columns into two sub-arrays

\[
Q_1 = \begin{pmatrix}
  a & b & c \\
  c & b & a \\
  b & a & c \\
  c & a & b
\end{pmatrix}, \quad \text{and} \quad Q_2 = \begin{pmatrix}
  d & e & f \\
  f & e & d \\
  e & d & f \\
  f & d & e
\end{pmatrix}.
\]

We can now give various definitions of symmetry for FSM channels.

**Definition 2:** (e.g., [9], [10]) An FSM channel is symmetric if for each state \(s \in S\), the rows of \(Q^s\) are permutations of each other such that the row permutation pattern is identical for all states, and similarly, if for each \(s \in S\) the columns of \(Q^s\) are permutations of each other with an identical column permutation pattern across all states.

**Definition 3:** An FSM channel is weakly-symmetric if for each state \(s \in S\), \(Q^s\) is weakly-symmetric and the row permutation pattern is identical for all states.

**Definition 4:** An FSM channel is quasi-symmetric if for each state \(s \in S\), \(Q^s\) is quasi-symmetric and the row permutation pattern is identical for all states.

To make these definitions clear, let us consider the following conditional probability matrices of a two-state quasi-symmetric FSM channel:

\[
Q^1 = \begin{pmatrix}
  a & b & c & d & e & f \\
  c & b & a & f & e & d \\
  b & a & c & e & d & f \\
  c & a & b & f & d & e
\end{pmatrix}, \quad Q^2 = \begin{pmatrix}
  a' & b' & c' & d' & e' & f' \\
  c' & b' & a' & f' & e' & d' \\
  b' & a' & c' & e' & d' & f' \\
  c' & a' & b' & f' & d' & e'
\end{pmatrix}.
\]  \hspace{1cm} (4)

As it can be seen, \(Q^1\) and \(Q^2\) have the same permutation orders. It directly follows that, symmetric and weakly symmetric FSM channels are special cases of quasi-symmetric FSM channels. Therefore, we focus on quasi-symmetric FSM channels for the sake of generality.
Let us define $Z$ (which will serve as a noise alphabet) such that $|\mathcal{Y}| = |Z|$, where $\mathcal{Y}$ is the output alphabet. Then, the symmetry definitions above imply that for each state $s$, we can find functions $f_s(\cdot) : Z \to [0, 1]$ and functions $\Phi_s(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to Z$ that are onto given $x$ (i.e., for each $x \in \mathcal{X}$, $\Phi_s(x, \cdot) : \mathcal{Y} \to Z$ is onto), such that

$$f_s(\Phi_s(x, y)) = p_C(y|x, s).$$

(5)

**Lemma 1:** The function $\Phi_s(\cdot, \cdot)$ is invariant with $s$.

**Proof:** $\Phi_s(\cdot, \cdot)$ is a function which takes the row and column position of a matrix as the input and outputs a noise value. We need to show that $\Phi_{s_i}(x, y) = \Phi_{s_j}(x, y)$, $\forall s_i, s_j \in S$ and $\forall x \in \mathcal{X}, y \in \mathcal{Y}$. Since the rows of a quasi-symmetric channel transition matrix are permutations of each other, we can assume that at state $s_i$ there exists $y_{i_1}, \ldots, y_{i_k} \in \mathcal{Y}$, where $k = |\mathcal{X}|$, which satisfies

$$\Phi_{s_i}(x_1, y_{i_1}) = \Phi_{s_i}(x_2, y_{i_2}) = \cdots = \Phi_{s_i}(x_k, y_{i_k}) = z^*.$$ 

By the unique order of row permutation between states, we have that

$$\Phi_{s_j}(x_1, y_{i_1}) = \Phi_{s_j}(x_2, y_{i_2}) = \cdots = \Phi_{s_j}(x_k, y_{i_k}) = z^*$$

which implies that $\Phi_{s_j} = \Phi_{s_i}$. 

Therefore, for a quasi-symmetric FSM channel, there exist a function $\Phi(\cdot, \cdot)$ such that the random variable $Z = \Phi(X, Y)$ has the conditional distribution

$$P(z|x, s) = \frac{P(y = \nu(x, z), z, x, s)}{P(x, s)} = \frac{P(z|x, y = \nu(x, z))p_C(y|x, s)P(x, s)}{P(x, s)} \overset{(a)}{=} p_C(y|x, s) = f_s(z).$$

(6)

where $\nu(\cdot, \cdot) : \mathcal{X} \times Z \to \mathcal{Y}$ and $(a)$ is due to the fact that $p(z|x, y = \nu(x, z)) = 1$. This important observation first given in [9], reduces the set of conditional probability distributions which identifies the quasi-symmetric FSM channel to an $|S| \times |Z|$ matrix $T$ defined by

$$T[s, z] = f_s(z).$$

(7)

Therefore, for quasi-symmetric FSM channels besides properties [11] to [12], we have an additional property defined as follows:

2Note that since each function $\Phi_s(x, \cdot) : \mathcal{Y} \to Z$ is onto given $x$ and since $|\mathcal{Y}| = |Z|$, then it is also one-to-one given $x$; i.e., $\Phi_s(x, y) = \Phi_s(x, y') \Rightarrow y = y'$.
For a quasi-symmetric Markov channel, for any \( n \), \( P(z_n|x_n, s_n) = P(z_n|s_n) = T[s_n, z_n] \).

To make this statement explicit, let us consider the FSM channel given above with \( \mathcal{X} = \{1, 2, 3, 4\} \), \( \mathcal{Y} = \mathcal{Z} = \{1, 2, 3, 4, 5, 6\} \) and \( \mathcal{S} = \{1, 2\} \). For this channel, we can define the function \( z = \Phi(x, y) \) and \( f_s(z) \) such that for each \((x, y)\) pair, which has the same conditional probability within that state, \( \Phi(x, y) \) returns the same value and the value of \( f_s(z) \) at that value is \( p_C(y_i|x_i, s_i) \), e.g., \( \Phi(1, 1) = \Phi(2, 3) = \Phi(3, 2) = \Phi(4, 2) = 1 \) and \( f_1(1) = a \) and \( f_2(1) = a' \). Therefore, the channel conditional probabilities for each state can now be defined by \( \Phi \) and the matrix \( T \), where

\[
T = \begin{pmatrix}
a & b & c & d & e & f \\
a' & b' & c' & d' & e' & f'
\end{pmatrix}.
\]

Hence, the fundamental property for quasi-symmetric FSM channels is the existence of a noise process given by \( Z_n = \Phi(X_n, Y_n) \) such that \( Z^n \) is independent of \( X^n \) given \( S^n \). The class of FSM channels having this property, when there is no feedback, are termed variable noise channels [10].

The properties that we have developed so far are valid for any quasi-symmetric FSM channel. However, while discussing the feedback capacity of these channels we assume that the channels also satisfy the following assumption:

**Assumption 1:** We assume that for a fixed \( y \in \mathcal{Y} \), \( \sum_x f_s(\Phi_s(x, y)) \) is invariant with \( s \in \mathcal{S} \).

In other words, the channel transition matrices satisfy the condition that the sums of the columns of channel transition matrices are identical across all states.

This requirement will be needed in our dynamic programming approach which we use to determine the optimal feedback policy (as will be seen in the next section).

**III. Feedback Capacity of Quasi-Symmetric FSM Channels**

In this section, we will show that feedback does not increase the capacity of quasi-symmetric FSM channels defined in the previous section. By feedback, we mean that there exists a channel from the receiver to the transmitter which is noiseless, delayless and has large capacity. Thus at any given time, all previously received outputs are unambiguously known by the transmitter and can be used for encoding the message into the next code symbol.

A feedback code with blocklength \( n \) and rate \( R \) consists of a sequence of mappings

\[
\psi_i : \{1, 2, ..., 2^{nR}\} \times \mathcal{Y}_i \rightarrow \mathcal{X}
\]
for $i = 1, 2, \ldots n$ and an associated decoding function

$$
\phi : Y^n \to \{1, 2, \ldots, 2^{nR}\}.
$$

Thus, when the transmitter wants to send a message, say $W \in \mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$, it sends the codeword $X^n$, where $X_1 = \psi_1(W)$ and $X_i = \psi_i(W, Y_1, \ldots, Y_{i-1})$, for $i = 2, \ldots, n$. For a received $Y^n$ at the channel output, the receiver uses the decoding function to estimate the transmitted message as $\hat{W} = \phi(Y^n)$. A decoding error is made when $\hat{W} \neq W$. We assume that the message $W$ is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. Therefore, the probability of error is given by

$$
P_e^{(n)} = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} P\{\phi(Y^n) \neq W | W = k\}.
$$

The capacity with feedback, $C_{FB}$, is the supremum of all admissible rates; i.e., rates for which there exists sequences of feedback codes with asymptotically vanishing probability of error. From Fano’s inequality, we have

$$
H(W|Y_n) \leq h_b(P_e^{(n)}) + P_e^{(n)} \log_2(2^{nR} - 1) \leq 1 + P_e^{(n)} nR
$$

where the first inequality holds since $h_b(P_e^{(n)}) \leq 1$, where $h_b(\cdot)$ is the binary entropy function. Since $W$ is uniformly distributed,

$$
nR = H(W) = H(W|Y^n) + I(W; Y^n) \leq 1 + P_e^{(n)} nR + I(W; Y^n)
$$

where $R$ is any admissible rate. Dividing both sides by $n$ and taking the $\lim \inf$ yields

$$
C_{FB} \leq \lim \inf_{n \to \infty} \sup_{\{\psi_i\}_{i=1}^n} \frac{1}{n} I(W; Y^n).
$$

It should be noted that

$$
P(x_i|x^{i-1}, y^{i-1}) = \sum_{w \in \mathcal{W}} P(w) 1_{\{\psi_i(w, y^{i-1}) = x_i\}} = \frac{1}{2^{nR}} \sum_{w \in \mathcal{W}} 1_{\{\psi_i(w, y^{i-1}) = x_i\}}
$$

where $1_{\{\cdot\}}$ is the indicator function. Furthermore,
\[ \lim \inf \sup_{n \to \infty} \frac{1}{n} I(W; Y^n) = \lim \inf \sup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|W, Y^{i-1}) \]
\[ = \lim \inf \sup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|W, Y^{i-1}, X^i) \]
\[ \leq \lim \inf \sup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i) \]
\[ = \lim \inf \sup_{n \to \infty} \{P(x_i|x^{i-1}, y^{i-1})\}_{i=1}^{n} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i) \]

where \( (a) \) holds since \( H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i) \) is a function of \( \{P(x_i|x^{i-1}, y^{i-1})\} \).

Now, let us consider the following equation
\[ \sup_{\{P(x_i|x^{i-1}, y^{i-1})\}_{i=1}^{n}} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i). \]  \hspace{1cm} (10)

We next establish three Lemmas in order to prove the main contribution of the paper. In the first Lemma, we show that the term \( H(Y_i|X^i, Y^{i-1}) \) is equal to \( H(Z_i|Z^{i-1}) \), and in the other two Lemmas we show that \( \sum_{i=1}^{n} H(Y_i|Y^{i-1}) \) is maximized by uniform feedback policies \( \{P(x_i|x^{i-1}, y^{i-1})\}_{i=1}^{n} \).

**Lemma 2:** A quasi-symmetric FSM satisfies
\[ H(Y_i|X^i, Y^{i-1}) = H(Z_i|Z^{i-1}), \quad \forall i = 1, \ldots, n. \]

The proof of the above lemma is given in Appendix I. As the next step, we show that all of the conditional output entropies \( H(Y^i|Y^{i-1}) \) in (10) are maximized by uniform feedback policies. We solve this problem using dynamic programming.

We now recast the optimization problem, maximization of the sum of conditional output entropies over all feedback policies, using dynamic programming [24]. Let us denote \( P(x_i|x^{i-1}, y^{i-1}) \) by \( \varphi_i \), for \( i = 1, \ldots, n \) and let \( \pi = \{ \varphi_i, 1 \leq i \leq n \} \). Let us recall our optimization problem:
\[ \max_{\{\varphi_1, \ldots, \varphi_n\}} \{H(Y_n|Y^{n-1}) + H(Y_{n-1}|Y^{n-2}) + \cdots + H(Y_1)\}. \]  \hspace{1cm} (11)

From (11), we observe that the optimization problem is nested; that is the policy at time \( n \), i.e., \( \varphi_n \), can depend on the previous policies and should maximize \( H(Y_n|Y^{n-1}) \) and on the other hand, \( \varphi_i \) can depend on previous policies but it should maximize both the current cost \( H(Y_i|Y^{i-1}) \) and the reward-to-go (see (12)).

March 28, 2010

DRAFT
Let $V_i(Y^{i-1}) = \max_{\varphi_i} [H(Y_i|Y^{i-1}) + V_{i+1}(Y^i)]$, where $V_{i+1}(Y^n) = 0$ and the $V_i(Y^{i-1})$ terms are explicitly given in (12) for $i = 1, \ldots, n$. $V_{i+1}(Y^i)$ denotes the reward-to-go at time $i$ which is the future reward generated by the control policy at the current time.

\[
\begin{align*}
V_n(Y^{n-1}) &= \max_{\varphi_n} H(Y_n|Y^{n-1}) \\
V_{n-1}(Y^{n-2}) &= \max_{\varphi_{n-1}} \left\{ H(Y_{n-1}|Y^{n-2}) + \max_{\varphi_n} \left\{ H(Y_n|Y^{n-1}) \right\} \right\} \\
V_{n-2}(Y^{n-3}) &= \max_{\varphi_{n-2}} \left\{ H(Y_{n-2}|Y^{n-3}) + \max_{\varphi_{n-1}} \left\{ H(Y_{n-1}|Y^{n-2}) + \max_{\varphi_n} \left\{ H(Y_n|Y^{n-1}) \right\} \right\} \right\} \\
&\vdots \\
V_1 &= \max_{\varphi_1} \left\{ H(Y_1) + \cdots + \max_{\varphi_{n-1}} \left\{ H(Y_{n-1}|Y^{n-2}) + \max_{\varphi_n} \left\{ H(Y_n|Y^{n-1}) \right\} \right\} \right\} 
\end{align*}
\]

(12)

Therefore, the optimization problem turns out to be finding the policy, $\pi$, which achieves $V_1$. We next show that the policy achieving $V_1$ is composed of the uniform feedback policies for $i = 1, \ldots, n$. Toward this goal, we find a condition that the policies at times $(i-1), \ldots, 1$ do not affect the value attained by the conditional output entropy at time $i$ when the policy at time $i$ is uniform. A sufficient condition to manage this problem is requiring $\sum_x f_s(\Phi_s(x, y))$ to be invariant with $s \in S$. This will be explicitly shown in Lemma 4. We first have the following:

**Lemma 3:** For a quasi-symmetric FSM channel, each conditional output entropy $H(Y_i|Y^{i-1})$, $i = 1, \ldots, n$ in (10), given the past policies $\{\varphi_1, \varphi_2, \cdots, \varphi_{i-1}\}$, is maximized by a uniform feedback policy:

\[
\arg\max_{\varphi_i} H(Y_i|Y^{i-1}) = \varphi^*_i(x) = \frac{1}{|X|}, \quad \forall x \in X
\]

\[\forall i = 1, \ldots, n.\]  

(13)

The proof of the above can be found in Appendix II. With this Lemma, we have shown that for each $i$, $H(Y_i|Y^{i-1})$ is maximized by the uniform input policy. However, this is not sufficient to conclude that the optimal policy attaining $V_1(Y^n)$, i.e., the optimal policy maximizing $\sum_{i=1}^n H(Y_i|Y^{i-1})$, consists of a sequence uniform input policies for $i = 1, \cdots, n$. This is because Lemma 3 only maximizes the current conditional entropy via a uniform input (that is it is optimal in a myopic sense); however, it is still possible that a non-uniform policy might result in a higher value function through the rewards-to-go. Let us now look at $P(y_i|y^{i-1})$ when we apply a uniform
policy at time $i$ (current time). We obtain using (23) that

$$P(y_i|y^{i-1}) = \sum_{x_i,x^{i-1},s_i,s^{i-1}} p_C(y_i|x_i,s_i)P(x_i|x^{i-1},y^{i-1})P(s_i|s^{i-1})P(x^{i-1},s^{i-1}|y^{i-1})$$

$$= \frac{1}{|X|} \sum_{x_i,x^{i-1},s_i,s^{i-1}} p_C(y_i|x_i,s_i)P(s_i|s^{i-1})P(x^{i-1},s^{i-1}|y^{i-1})$$

$$= \frac{1}{|X|} \sum_{x_i,s_i} p_C(y_i|x_i,s_i)P(s_i|s^{i-1}) \sum_{x^{i-1}} P(x^{i-1},s^{i-1}|y^{i-1})$$

$$= \frac{1}{|X|} \sum_{x_i,s_i} p_C(y_i|x_i,s_i)P(s_i|y^{i-1})$$

where $(i)$ is valid since the input policy $\varphi_i$ is uniform. Note that the dependency on past input policies comes through $P(s_i|y^{i-1})$ which includes transition probabilities between states, on which we have no control.

**Lemma 4:** Assume that the feedback policy $\varphi_i$, at (current) time $i$, is uniform. Then the value of the conditional entropy $H(Y_i|Y^{i-1})$, at time $i$, is independent of past feedback policies at times $(i-1), \ldots, 1$ if $\sum_x f_s(\Phi_s(x,y))$ is invariant with $s \in S$.

**Proof:**

$$P(y_i|y^{i-1}) = \frac{1}{|X|} \sum_{x_i,s_i} p_C(y_i|x_i,s_i)P(s_i|y^{i-1})$$

$$= \frac{1}{|X|} \sum_{s_i} P(s_i|y^{i-1}) \sum_{x_i} p_C(y_i|x_i,s_i)$$

$$= \frac{1}{|X|} \sum_{s_i} P(s_i|y^{i-1}) \sum_{x_i} f_s(\Phi(x_i,y_i))$$

and since the underbraced term is invariant with $s$, the proof is complete as the final sum will be $\frac{1}{|X|} \sum_x f_s(\Phi(x_i,y_i))$. □

We have so far shown that $H(Y_i|X^i,Y^{i-1}) = H(Z_i|Z^{i-1})$ and that $\sum_{i=1}^n H(Y_i|Y^{i-1})$ is maximized by uniform input polices. With these results in hand, we have thus shown the following upperbound for the feedback capacity

$$C_{FB} \leq \lim_{n \to \infty} \inf \frac{1}{n} [H(\tilde{Y}^n) - H(Z^n)]$$

(14)

where $H(\tilde{Y}^n)$ is the output entropy when the input is uniform.
Let us now define a Hidden Markov Process (HMP) [25] which we will use while discussing the ergodicity of the noise and output processes. An HMP is denoted by a quadruple \([S, P, Z, T]\) in which \([S, P]\) is a Markov process and \(T\) is the observation matrix defined by (7). The non-Markov process \(\{Z_i\}_{i=1}^{\infty}\) with alphabet \(Z\) is called HMP and it is the noisy version of the state process observed through a DMC determined by \(T\).

**Lemma 5:** For a quasi-symmetric FSM channel with feedback, the noise process is an HMP with parameters \([S, P, Z, T]\) if the feedback policy is uniform.

**Proof:** We slightly modify the proof in [9, Lemma 1]. It should be noted that this proof is valid if the input process is i.i.d. Therefore, this proof is inapplicable when feedback exists. However, for quasi-symmetric FSM channels, the capacity achieving distribution is uniform; therefore the statement holds with feedback as well. To show this result, it suffices to show that

\[
P(z_i|s_i, z^{i-1}) = P(z_i|s_i).
\]

Since \(\{S_i\}_{i=1}^{\infty}\) is Markovian, it directly implies that

\[
P(s_i|s_{i-1}, z^{i-1}) = P(s_i|s_{i-1}).
\]

Note that

\[
P(z_i|s_i, z^{i-1}) = \sum_{x^{i-1}} \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} P(y_i, x_i, x^{i-1}|s_i, z^{i-1})
\]

\[
= \sum_{x^{i-1}} \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} p_C(y_i|x_i, s_i) P(x_i|x^{i-1}, s_i, z^{i-1}) P(x^{i-1}|s_i, z^{i-1})
\]

\[
\overset{(i)}{=} \sum_{x^{i-1}} \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} f_s(\Phi(x_i, y_i)) P(x_i|x^{i-1}, s_i, z^{i-1}) P(x^{i-1}|s_i, z^{i-1})
\]

\[
\overset{(ii)}{=} \frac{1}{|\mathcal{X}|} \sum_{x^{i-1}} P(x^{i-1}|s_i, z^{i-1}) \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} f_s(\Phi(x_i, y_i))
\]

\[
= \frac{1}{|\mathcal{X}|} \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} f_s(\Phi(x_i, y_i)) = \frac{1}{|\mathcal{X}|} \sum \{ (x_i, y_i): z_i = \Phi(x_i, y_i) \} p_C(y_i|x_i, s_i)
\]

\[
\overset{(iii)}{=} \frac{1}{|\mathcal{X}|} |\mathcal{X}| P(z_i|s_i)
\]

where \((i)\) is valid by (5), \((ii)\) is valid since the input policy is uniform and \((iii)\) is valid since each \(z_i\) is satisfied by \(|\mathcal{X}|\) number of \((x_i, y_i)\) pairs.

It should also be noted that, the output process, \(\{\tilde{Y}_i\}_{i=1}^{\infty}\), for uniform and independent \(\{X_i\}_{i=1}^{\infty}\) is also a hidden Markov process since

\[
P(\tilde{y}_i|s_i, \tilde{y}^{i-1}) = \sum_{x_i} P(\tilde{y}_i, x_i|s_i, \tilde{y}^{i-1}) = \sum_{x_i} p_C(\tilde{y}_i|x_i, s_i) P(x_i|s_i, \tilde{y}^{i-1})
\]

\[
= \sum_{x_i} p_C(\tilde{y}_i|x_i, s_i) P(x_i|s_i) = P(\tilde{y}_i|s_i)
\]
where (a) is due to (2) and (b) is due to the fact that $X_i$ is uniformly distributed. The channel associated with the HMP is memoryless and as such it is stationary. Therefore, since the state process is stationary and ergodic both the output and noise processes are stationary and ergodic.

Theorem 1: The feedback capacity of the quasi-symmetric FSM channel $[X,Y,S,P_S,Z,T,\Phi]$ satisfying the condition that $\sum_x f_s(\Phi_s(x,y))$ is invariant with $s \in S$ is given by

$$C_{FB} = H(\tilde{Y}) - H(Z)$$

where $H(\tilde{Y})$ is the entropy rate of the output process for uniform i.i.d. $X^n$ and $H(Z)$ is the entropy rate of the HMP $\{Z_i\}_{i=1}^\infty$.

Proof: With (14) we already have a converse for the feedback capacity. We need to show that this bound is achievable. We first note that the noise and output processes are stationary which imply that

$$C_{FB} \leq \liminf_{n \to \infty} \sup_{P(x_i|x^{i-1},y^{i-1})} \frac{1}{n} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},X^i)$$

$$= \liminf_{n \to \infty} \frac{1}{n} [H(\tilde{Y}^n) - H(Z^n)]$$

$$= \lim_{n \to \infty} \frac{1}{n} [H(\tilde{Y}^n) - H(Z^n)] = H(\tilde{Y}) - H(Z).$$

(15)

It is sufficient to show that the bound in (15) is achievable. We now remark that there exists a coding policy which achieves this bound. Note that since the noise process is stationary and ergodic, it can be shown that $H(\tilde{Y}) - H(Z)$ is an admissible rate (e.g. see [15, Theorem 5.3] and [26, Theorem 2]). Thus,

$$C_{FB} \geq \lim_{n \to \infty} \frac{1}{n} [H(\tilde{Y}^n) - H(Z^n)] = H(\tilde{Y}) - H(Z)$$

and this completes the proof.

Corollary 1: Feedback does not increase capacity of quasi-symmetric FSM channels for which $\sum_x f_s(\Phi_s(x,y))$ is invariant with $s \in S$.

Proof: The result follows by noting that a non-feedback code is a special case of a feedback code and that the non-feedback capacity is also achieved by uniform input policies. This can be
shown more explicitly as follows
\[ C_{FB} = \lim_{n \to \infty} \frac{1}{n} H(\tilde{Y}^n) - \lim_{n \to \infty} \frac{1}{n} H(Z^n) \]
\[ \overset{(i)}{=} \lim_{n \to \infty} \frac{1}{n} H(Y^n) \bigg|_{P(X^n) = \frac{1}{2}} - \lim_{n \to \infty} \frac{1}{n} H(Z^n) \]
\[ \leq \lim_{n \to \infty} \frac{1}{n} \sup_{P(x^n)} I(X^n, Y^n) = C_{NFB} \]
where \( C_{NFB} \) denotes the non-feedback capacity and (i) is valid since the input process is uniform and i.i.d. Finally, since \( C_{FB} \geq C_{NFB} \), we obtain that \( C_{FB} = C_{NFB} \).

IV. EXAMPLES OF QUASI-SYMMETRIC FINITE STATE MARKOV CHANNELS

In this section, we present some well-known channels that satisfy the quasi-symmetry condition presented in the paper.

A. Gilbert-Elliot Channel (e.q., [3]) One of the widely used FSM channels is the Gilbert-Elliot channel denoted by \([X, Y, S, P, C]\), where \( X = Y = S = \{0, 1\} \). The two states are called "bad" state and "good" state, respectively, and the state transition matrix is given by:
\[ P = \begin{pmatrix} 1 - g & g \\ b & 1 - b \end{pmatrix}, \]
where \( 0 < g < 1, 0 < b < 1 \) and in either of these two states, the channel is a binary symmetric channel (BSC) with the following transition matrixes for states \( s = 0 \) and \( s = 1 \), respectively:
\[ \begin{pmatrix} 1 - p_G & p_G \\ p_G & 1 - p_G \end{pmatrix}, \begin{pmatrix} 1 - p_B & p_B \\ p_B & 1 - p_B \end{pmatrix}. \]

From the above channel transition matrixes, it can be observed that the Gilbert-Elliot channel is a symmetric FSM channel by Definition 2. Then, there exists a random variable \( Z = \Phi(X, Y) \) with alphabet \( Z = \{0, 1\} \) and a function \( f_s(z) \) such that, \( f_0(0) = 1 - p_G \) and \( f_0(1) = p_G \), \( f_1(0) = 1 - p_B \) and \( f_0(1) = p_B \). Therefore, one can define the \( T(s, z) \) matrix for this channel as
\[ T = \begin{pmatrix} 1 - p_G & p_G \\ 1 - p_B & p_B \end{pmatrix}, \]
and we obtain that \( \Phi(X, Y) = X \oplus Y \), where \( \oplus \) represents modulo addition, and \( T(s, z) \) defined above. By Corollary 1, feedback does not increase the capacity of the Gilbert-Elliot channel. It should be noted that this result is a special case of [11] and [13].
B. Discrete Modulo Additive Channel with Markovian Noise

Consider the discrete channel with a common alphabet $\mathcal{A} = \{0, 1, \ldots, q-1\}$ for the input, output and noise processes. The channel is described by the equation $Y_n = X_n \oplus Z_n$, for $n = 1, 2, 3, \ldots$, and $Y_n, X_n$ and $Z_n$ denotes the output, input and noise processes respectively. The noise process, $\{Z_n\}_{n=1}^{\infty}$, is Markovian and it is independent of the input process. It is straightforward to see that the channel transition matrix at state $s_i$, $Q^{s_i}$, will be as follows:

$$Q^{s_i} = \begin{pmatrix}
    P(Z_i = 0|Z_{i-1} = s_i) & P(Z_i = 1|Z_{i-1} = s_i) & P(Z_i = 2|Z_{i-1} = s_i) \\
    P(Z_i = 2|Z_{i-1} = s_i) & P(Z_i = 0|Z_{i-1} = s_i) & P(Z_i = 1|Z_{i-1} = s_i) \\
    P(Z_i = 1|Z_{i-1} = s_i) & P(Z_i = 2|Z_{i-1} = s_i) & P(Z_i = 0|Z_{i-1} = s_i)
\end{pmatrix}.$$

For each state, the channel transition matrix will still be symmetric with the same row permutation order. Furthermore, it also satisfies Assumption 1 since column sums are always one. Therefore, the discrete modulo additive channel is a symmetric FSM channel with $\mathcal{A} = \{0, 1, 2\}$ and $\Phi(X, Y) = X \oplus Y$. Hence, by Corollary 1, feedback does not increase the capacity of the discrete modulo additive channel with Markovian noise.

This result is a special case of [11]. It has been recently extended to finite-state multiple access channels in [27].

C. A Symmetric Discrete Channel with Markovian Noise

Consider a discrete, not necessarily additive, channel with Markovian noise [12]. More precisely, consider the channel given by $Y_i = f(X_i, Z_i)$ for $i = 1, 2, \ldots$ where $X_i, Z_i$ and $Y_i$ are the input, noise and output of the channel, respectively, and $f: \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$ is a given function. Assume also that $\{X_i\}$ and $\{Z_i\}$ are independent from each other and the channel satisfies the following properties\(^4\):

1. $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = q.$
2. Given the input $x$, $f(x, \cdot)$ is one-to-one; i.e., $\forall x \in \mathcal{X}$, $f(x, z) = f(x, \bar{z}) \Rightarrow z = \bar{z}$.
3. $f^{-1}$ exists such that $z = f^{-1}(x, y)$ and given $y$, $f^{-1}(\cdot, y)$ is one-to-one; i.e., $\forall y \in \mathcal{Y}$, $f^{-1}(x, y) = f^{-1}(\bar{x}, y) \Rightarrow x = \bar{x}$.

We note that a channel satisfying these conditions has a symmetric channel transition matrix for each state, where the state is given by the previous noise variable: $S_i = Z_{i-1}$. Therefore,

\(^4\) In [12] it is stated that $|\mathcal{X}| = |\mathcal{Z}| = q$. However, following the proof, it can be evidently seen that $|\mathcal{Y}| = q$ is also assumed.
This channel is a symmetric FSM channel with the same permutation order determined by the function $f$. It also satisfies Assumption 1 as the column sums are one for each state. Therefore, by Corollary 1, feedback does not increase the capacity of these channels. This result is first shown in [12], where the noise process may be non-Markovian and non-ergodic in general.

**D. Binary Channel with Erasures, Errors and Markovian State** We now present an example which illustrates the result of the paper when the column sums are different than one. Consider the two-state channel given by $\mathcal{X} = \{0, 1\}$, $\mathcal{S} = \{s_1, s_2\}$, where $\{S_i\}$ is Markovian, $\mathcal{Y} = \{0, E, 1\}$ with the following channel transition matrices

\[
Q^{s_1} = \begin{pmatrix}
1 - \varepsilon - \xi & \varepsilon \\
\varepsilon & 1 - \varepsilon - \xi
\end{pmatrix}, \quad Q^{s_2} = \begin{pmatrix}
1 - \varepsilon' - \xi' & \xi' & \varepsilon' \\
\varepsilon' & \xi' & 1 - \varepsilon' - \xi'
\end{pmatrix}
\]

where $0 < \varepsilon, \xi, \varepsilon' < 1$ are fixed. We first note that this channel is a two-state quasi-symmetric FSM channel, since we can partition $Q^{s_1}$ and $Q^{s_2}$ in two symmetric sub-arrays given by

\[
Q_{Y_1}^{s_1} = \begin{pmatrix}
1 - \varepsilon - \xi & \varepsilon \\
\varepsilon & 1 - \varepsilon - \xi
\end{pmatrix}, \quad Q_{Y_2}^{s_1} = \begin{pmatrix}
\xi \\
\xi
\end{pmatrix}
\]

and

\[
Q_{Y_1}^{s_2} = \begin{pmatrix}
1 - \varepsilon' - \xi' & \xi' \\
\varepsilon' & 1 - \varepsilon' - \xi'
\end{pmatrix}, \quad Q_{Y_2}^{s_2} = \begin{pmatrix}
\xi' \\
\xi'
\end{pmatrix}
\]

respectively, where $\mathcal{Y}_1 = \{0, 1\}$ and $\mathcal{Y}_2 = \{E\}$ with identical permutation order between states. For this channel, if we set $\xi = \xi'$, then we automatically satisfy Assumption 1 since the column sums will be $1 - \xi$, $2\xi$ and $1 - \xi$ respectively. In other words, although the error probabilities are different across the states ($\epsilon \neq \epsilon'$ in general), we still have identical column sums. Therefore, by Corollary 1, feedback does not increase the capacity of this channel.

There is one more quasi-symmetric FSM channel that needs further attention. We now investigate how its channel properties directly satisfy the condition that the previous feedback policies do not affect the current value of the conditional output entropy. In other words, the example below satisfies Lemma 4 without having the condition that the column sums should be identical among different states, (i.e., it does not satisfy Assumption 1).

**E. Simplified Binary Erasure Channel with Markovian State** Consider the following binary erasure channel [21], which is a simplified (special) case of the erasure channel of Example D and has been used to model packet losses in a packet communication network, such as the
Internet. The channel has binary input and ternary output; \( \mathcal{X} = \{0, 1\}, \mathcal{Y} = \{0, E, 1\} \). Let \( S_i \) denote the state of the erasure channel when the packet \( i \) arrives such that when \( S_i = 1 \), the packet is erased, and when \( S_i = 0 \), the packet gets through. For a given input, the channel output is identical to the input if there is no erasure, and it is equal to the erasure symbol (\( E \)) if an erasure occurs. Therefore, the channel transition matrices at states 0, 1 will be as follows

\[
Q^0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad Q^1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.
\]

This channel can be considered as a special case of deletion channel in which the erased packet is assumed to be known by the decoder. Therefore, in an erasure channel, the receiver has also the side information about the state. In [21], this channel is considered as a finite buffer queue, which can be viewed as an FSM channel, and the state of the finite buffer channel is determined by the state of the buffer and it is shown that feedback does not increase the capacity of this channel. We herein note that the approach presented in the paper gives the same result.

**Proposition 1:** Feedback does not increase capacity of simplified binary erasure channel with Markovian state.

**Proof:** We first note that since the channel is quasi-symmetric for each state, the conditional output entropy is maximized by uniform input policies. What we further need to show is the independence of the value attained by \( H(Y_i|Y^{i-1} = y^{i-1}) \) from previous input policies. In particular, we need to show that \( P(s_i|y^{i-1}) \) is independent of past input policies (see Lemma 4). It should be noted that

\[
P(s_i|y^{i-1}) = \sum_{s_{i-1}} P(s_i|s_{i-1})P(s_{i-1}|y^{i-1}).
\]

Thus, given \( y^{i-1}, s_{i-1} \) is deterministic and independent of \( x^{i-1} \). Integrating this fact in our approach proves the desired result.

This particular example has the benefit of learning the state deterministically by only observing the output. We should remark that availability of both the state information and output feedback has also been considered within different setups in some other works and the situations for which feedback does not help to increase the capacity is determined (see [17, Theorem 19] and [20]).
V. Conclusion

In this work, we presented a class of symmetric channels which encapsulates a variety of discrete channels with memory. Motivated by several results in the literature, we established a class of symmetric finite-state Markovian channels for which feedback does not increase their capacity. We showed this result by first reformulating the optimization problem in terms of dynamic programming and then proving that, under feedback, the capacity achieving distribution is uniform. An important observation should be highlighted again: when feedback exists, one can learn the channel via the past policies and as such may apply a nonuniform policy which will result in a higher output entropy and capacity. We present a sufficient condition, Assumption 1, under which it is still possible to learn the channel via these past policies however, this learning does not affect the optimal policy. Although this result covers a large class of discrete channels with memory, we believe that by adopting the approach of this work, it is possible to show a similar result for a further general class of symmetric channels; both in the single user and multiple user settings.

Appendix I

Proof of Lemma 2

Proof: Let us define $\eta(x) = x \log(x)$. Then,

$$H(Y_i|X_i, X^{i-1}, Y^{i-1}) = -E_{X_i, Y^{i-1}} \left[ \sum_{y_i} \eta \left( P(y_i|x_i, x^{i-1}, y^{i-1}) \right) \right]$$

$$\overset{(a)}{=} -E_{X_i, Y^{i-1}} \left[ \sum_{y_i} \eta \left( \sum_{s_i} P_C(y_i|x_i, s_i) P(s_i|x_i, x^{i-1}, y^{i-1}) \right) \right]$$

$$\overset{(b)}{=} -E_{X_i, Y^{i-1}} \left[ \sum_{y_i} \eta \left( \sum_{s_i} P_C(y_i|x_i, s_i) \frac{P(s_i, x_i, x^{i-1}, y^{i-1})}{P(x_i, x^{i-1}, y^{i-1})} \right) \right]$$

$$\overset{(c)}{=} -E_{X_i, Y^{i-1}} \left[ \sum_{y_i} \eta \left( \sum_{s_i} P(z_i|s_i) \frac{P(s_i, x^{i-1}, y^{i-1})}{P(x^{i-1}, y^{i-1})} \right) \right]$$

$$\overset{(d)}{=} -E_{X_i, Y^{i-1}} \left[ \sum_{y_i} \eta \left( \sum_{s_i} P(z_i|s_i) P(s_i|x^{i-1}, y^{i-1}, z^{i-1}) \right) \right] \quad (16)$$
where (a) is valid since, given $s_i$ and $x_i$, $y_i$ is conditionally independent of $(x^{i-1}, y^{i-1})$, (b) is valid since the feedback input depends only on $(x^{i-1}, y^{i-1})$, (c) follows from (6), (7) and I3 and finally (d) is valid since $\Phi(x^{i-1}, y^{i-1}) = z^{i-1}$.

We next show that $P(s_i|x^{i-1}, y^{i-1}, z^{i-1}) = P(s_i|z^{i-1})$. It should be noted that

$$P(s_i|x^{i-1}, y^{i-1}, z^{i-1}) = \sum_{s_i^{i-1}} P(s_i|x^{i-1}, y^{i-1}, z^{i-1}) = \sum_{s_i^{i-1}} P(s_i|s_i^{i-1})P(s_i^{i-1}|x^{i-1}, y^{i-1}, z^{i-1}).$$

Then,

$$P(s_1|x_1, y_1) = \frac{P(s_1, x_1, y_1)}{\sum_{s_1} P(s_1, x_1, y_1)} = \frac{pc(y_1|x_1, s_1)p(x_1)p(s_1)}{\sum_{s_1} pc(y_1|x_1, s_1)p(x_1)p(s_1)} \overset{(e)}{=} \frac{f_{s_1}(z_1)p(x_1)p(s_1)}{\sum_{s_1} f_{s_1}(z_1)p(x_1)p(s_1)} \overset{(f)}{=} \frac{p(z_1|s_1)p(x_1)p(s_1)}{\sum_{s_1} p(z_1|s_1)p(x_1)p(s_1)} = p(s_1|z_1) \quad (17)$$

where (e) is due to (5)-(7) and (f) is due to I3. Similarly,

$$P(s_2|x_1, y_1) = \sum_{s_1} P(s_2|s_1)p(s_1|x_1, y_1) \overset{(g)}{=} \sum_{s_1} P(s_2|s_1)p(s_1|z_1) = P(s_2|z_1) \quad (18)$$

where (g) is due to (17). Now, let us consider $P(s_3|x^2, y^2) = \sum_{s_2} P(s_3|s_2)p(s_2|x^2, y^2)$. Then,

$$P(s_2|x^2, y^2) = P(s_2|x^2, z^2) = \frac{P(s_2, x_1, x_2, z_1, z_2)}{\sum_{s_2} P(s_2, x_1, x_2, z_1, z_2)} \overset{(h)}{=} \frac{P(z_2|s_2, x_2, x_1, z_1)p(x_2|x_1, z_1)p(s_2, x_1, z_1)}{\sum_{s_2} P(z_2|s_2, x_2, x_1, z_1)p(x_2|x_1, z_1)p(s_2, x_1, z_1)} \overset{(k)}{=} \frac{P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|x_1, z_1)p(x_1)p(z_1)}{\sum_{s_2} P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|x_1, z_1)p(x_1)p(z_1)} \overset{(l)}{=} \frac{P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|x_1, z_1, y_1)p(x_1)p(z_1)}{\sum_{s_2} P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|x_1, z_1, y_1)p(x_1)p(z_1)} \overset{(m)}{=} \frac{P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|z_1)}{\sum_{s_2} P(z_2|s_2)p(x_2|x_1, z_1)p(s_2|z_1)} = \frac{P(x_2|x_1, z_1) \sum_{s_2} P(z_2|s_2)p(s_2|z_1)}{P(x_2|x_1, z_1) \sum_{s_2} P(z_2|s_2)p(s_2|z_1)} \overset{(n)}{=} \frac{pc(y_2|x_2, s_2)p(s_2|z_1)}{\sum_{s_2} pc(y_2|x_2, s_2)p(s_2|z_1)} \overset{(o)}{=} \frac{pc(y_2|x_2, s_2, z_1)p(s_2|z_1)}{\sum_{s_2} pc(y_2|x_2, s_2, z_1)p(s_2|z_1)} \overset{(p)}{=} \frac{P(z_2|s_2, z_1)p(s_2|z_1)}{\sum_{s_2} P(z_2|s_2, z_1)p(s_2|z_1)} = P(s_2|z_2) \quad (19)$$
where \((h)\) holds since \(x_2 \rightarrow (x_1, z_1) \rightarrow s_2\), \((k)\) is valid by \((6)\), \((l)\) is valid since given \(x\), \(\Phi(., .)\) is one-to-one, \((m)\) is valid by \((18)\), \((n)\) is valid due to \((5)-(7)\), \((p)\) is valid due to \(I_2\) and \((r)\) is due to \((5)\). Using \((17)\), \((18)\) and \((19)\) recursively for \(i = 2, \ldots, n\) we obtain

\[
P(s_i|x^{i-1}, y^{i-1}, z^{i-1}) = P(s_i|x^{i-1}, y^{i-1}) = P(s_i|z^{i-1}).
\]

Substituting \((20)\) into \((16)\) yields

\[
-E_{X^i, Y^{i-1}} \left[ \sum_{z_i} \eta \left( \sum_{s_i} p(z_i|s_i)P(s_i|z^{i-1}) \right) \right] = -E_{X^i, Y^{i-1}} \left[ \sum_{z_i} \eta \left( p(z_i|z^{i-1}) \right) \right] = H(Z_i|Z^{i-1}).
\]

\[\square\]

**APPENDIX II**

**PROOF OF LEMMA 3**

**Proof:** Let us first write the conditional output entropy \(H(Y_i|Y^{i-1})\) as

\[
H(Y_i|Y^{i-1}) = \sum_{y^{i-1}} P(y^{i-1})H(Y_i|Y^{i-1} = y^{i-1})
\]

where

\[
H(Y_i|Y^{i-1} = y^{i-1}) = - \sum_{y_i} P(y_i|y^{i-1}) \log P(y_i|y^{i-1}).
\]

To show that \(H(Y_i|Y^{i-1})\) in \((21)\) is maximized by a uniform input policy, it is enough to show that such a uniform policy maximizes each of the \(H(Y_i|Y^{i-1} = y^{i-1})\) terms.

We now expand \(P(y_i|y^{i-1})\) as follows

\[
P(y_i|y^{i-1}) = \sum_{x_i} \sum_{x^{i-1}} \sum_{s_i} \sum_{s^{i-1}} P(y_i, x_i, s_i, x^{i-1}, s^{i-1}|y^{i-1})
\]

\[
= \sum_{x_i, x^{i-1}, s_i, s^{i-1}} P(y_i|x_i, s_i, x^{i-1}, s^{i-1}, y^{i-1})P(x_i, s_i, x^{i-1}, s^{i-1}|y^{i-1})
\]

\[
= \sum_{x_i, x^{i-1}, s_i, s^{i-1}} P(y_i|x_i, s_i)P(x_i, x^{i-1}, s^{i-1}|y^{i-1})P(s_i|x_i, x^{i-1}, s^{i-1}, y^{i-1})
\]

\[
= \sum_{x_i, x^{i-1}, s_i, s^{i-1}} P(y_i|x_i, s_i)P(s_i|x^{i-1}, s^{i-1}, y^{i-1})P(x_i, x^{i-1}, s^{i-1}, y^{i-1})
\]

\[
= \sum_{x_i, x^{i-1}, s_i, s^{i-1}} P(y_i|x_i, s_i)P(s_i|x^{i-1}, y^{i-1})P(s_i|x^{i-1}, s^{i-1})P(x_i, x^{i-1}, s^{i-1}, y^{i-1}).
\]
where \((i)\) follows by (2), \((ii)\) is valid due to the property I2 and finally \((iii)\) is due to the fact that the feedback input depends only on \((x^{i-1}, y^{i-1})\). Thus

\[
P(y_i | y^{i-1}) = \sum_{x_i,x^{i-1}} \sum_{s_i,s^{i-1}} P(y_i | x_i, s_i) P(s_i | s^{i-1}) P(x_i | x^{i-1}, y^{i-1}) P(x^{i-1}, s^{i-1} | y^{i-1}). \tag{23}
\]

The key observation in equation (23) is the existence of an equivalent channel. More specifically, \(\sum_{s_i} P(y_i | x_i, s_i) P(s_i | s^{i-1})\) actually represents a quasi-symmetric channel transition matrix such that its entries are determined by the entries of the channel transition matrices of each state and the transition distribution of state probabilities. To continue, by (5),

\[
P(y_i | y^{i-1}) = \sum_{x_i,x^{i-1}} \sum_{s_i,s^{i-1}} f_{s_i}(\Phi_{s_i}(x_i, y_i)) P(s_i | s^{i-1}) P(x_i | x^{i-1}, y^{i-1}) P(x^{i-1}, s^{i-1} | y^{i-1}). \tag{24}
\]

By definition of quasi-symmetry, there exists \(m\) weakly symmetric sub-arrays in the channel transition matrix at each state \(s_i\). Among these sub-arrays, let us pick \(Q_j^s\) of size \(|\mathcal{X}| \times |\mathcal{Y}|\). (We assume that the partition of \(\mathcal{Y}\) is identical across all states.) Let \(Y_{jt}\), for \(t = 1, \ldots, |\mathcal{Y}|\), denote the output values in sub-array \(j\). Therefore, we obtain

\[
P(y_j | y^{i-1}) = \sum_{x_i,x^{i-1}} \sum_{s_i,s^{i-1}} f_{s_i}(\Phi_{s_i}(x_i, y_{jkt})) P(s_i | s^{i-1}) P(x_i | x^{i-1}, y^{i-1}) P(x^{i-1}, s^{i-1} | y^{i-1}). \tag{25}
\]

We desire to maximize (21) over the feedback policies \(P(X_i | X^{i-1}, Y^{i-1})\). Let \(\mathcal{X} = \{x_1, x_2, \ldots, x_k\}\), where \(k = |\mathcal{X}|\), \(\kappa(i-1) = P(s_i | s^{i-1})\), \(\chi(i-1) = P(x^{i-1}, s^{i-1} | y^{i-1})\) and denote the feedback policies by

\[
P(X_i = x_l | x^{i-1}, y^{i-1}) = \varphi_l(x_i) \quad \text{for} \quad l = 1, \ldots, k. \tag{26}
\]

Then we can write

\[
P(y_{j1} | y^{i-1}) = \sum_{s_i,s^{i-1}} \kappa(i-1) \sum_{s_i} \chi(i-1) \{ \varphi_1(x_1) f_{s_i}(\Phi_{s_i}(x_1, y_{j1})) + \cdots + \varphi_k(x_k) f_{s_i}(\Phi_{s_i}(x_k, y_{j1})) \}
\]

\[
P(y_{j2} | y^{i-1}) = \sum_{s_i,s^{i-1}} \kappa(i-1) \sum_{s_i} \chi(i-1) \{ \varphi_1(x_1) f_{s_i}(\Phi_{s_i}(x_1, y_{j2})) + \cdots + \varphi_k(x_k) f_{s_i}(\Phi_{s_i}(x_k, y_{j2})) \}
\]

\[
\vdots
\]

\[
P(y_{j|\mathcal{Y}|} | y^{i-1}) = \sum_{s_i,s^{i-1}} \chi(i-1) \sum_{s_i} \kappa(i-1) \left\{ \varphi_1(x_1) f_{s_i}(\Phi_{s_i}(x_1, y_{j|\mathcal{Y}|})) + \cdots + \varphi_k(x_k) f_{s_i}(\Phi_{s_i}(x_k, y_{j|\mathcal{Y}|})) \right\}.
\]
It should be noted that, each \( f_{s_i}(\Phi_{s_i}(x_i, y_{jt})) \) corresponds to an entry in the channel transition matrix \( Q^{s_i} \) at state \( s_i \). We also know that, the rows of the sub-array \( Q^{s_i}_j \) are permutations of each other. In other words, each \( f_{s_i}(\Phi_{s_i}(x_i, y_{jt})) \) value appears exactly \( k \) times (once in each row) in the sub-array \( Q^{s_i}_j \). Thus, the feedback policy \( \varphi_i(x_i) \) is multiplied by a different \( f_{s_i}(\Phi_{s_i}(x_i, y_{jt})) \) value in each of the equations of \( P(y_{jt}|y^{i-1}) \) given above. Therefore, \( \sum_{t=1}^{\mid Y_j \mid} P(Y_t = y_{jt}|y^{i-1}) \) is equal to

\[
\sum_{s_i, s_{i-1}, x_{i-1}} P(s_i|s^{i-1})P(x^{i-1}, s^{i-1}|y^{i-1}) \sum_{t=1}^{\mid Y_j \mid} f_{s_i}(\Phi_{s_i}(x_i, y_{jt})) \tag{27}
\]

since \( \sum_{i=1}^{k} \varphi_i(x_i) = 1 \), where \( \sum_{t=1}^{\mid Y_j \mid} f_{s_i}(\Phi_{s_i}(x_i, y_{jt})) \) is the sum of any row in the sub-array \( Q^{s_i}_j \) and is equal to a constant (by the weak-symmetry property of sub-array \( Q^{s_i}_j \)). The critical observation is that the value attained by (27) is independent of the feedback policies. Similarly, for all the other \( m-1 \) sub-arrays, their conditional output sums will be independent of the feedback policies. Let us denote these sums by \( \Omega_1, \ldots, \Omega_m \). More specifically for sub-array \( j \), let \( \Omega_j = \sum_{t=1}^{\mid Y_j \mid} P(Y_t = y_{jt}|y^{i-1}) \). Then the optimization problem in (22) now becomes,

\[
\arg\max_{\Omega_{j,t}} - \sum_{j=1}^{m} \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} \log \Omega_{j,t} \tag{28}
\]

where \( \sum_{j=1}^{m} \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} = 1 \) and \( \Omega_{j,t}, \ t = 1, \ldots, \mid Y_j \mid \) denotes conditional output probabilities in sub-array \( j \). For each sub-array \( j \), we need to find the \( \Omega_{j,t} \) values that maximize \( \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} \log \Omega_{j,t} \).

By the log-sum inequality, we have that

\[
- \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} \log \Omega_{j,t} \leq - \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} \log \sum_{t=1}^{\mid Y_j \mid} \Omega_{j,t} \mid Y_j \mid \tag{29}
\]

with equality if and only if

\[
\Omega_{j,t} = \Omega_{s,w} \ \forall s, w \in \{1, \ldots, \mid Y_j \mid \}. \tag{30}
\]

In other words, for the sub-array \( j \), the conditional entropy is maximized if and only if the conditional output probabilities in this sub-array are identical. Since this fact is valid for the other sub-arrays, to maximize the conditional entropy we need to (30) to be valid for all sub-arrays.

At this point, we have shown that the conditional output entropy is maximized if the conditional output probabilities are identical for each sub-array. In order to complete this step, we have to show that this is achieved by uniform feedback policies.
Now, let us consider two conditional output probabilities, \( P(Y_i = y_{js} | y^{i-1}) \) and \( P(Y_i = y_{jt} | y^{i-1}) \) in sub-array \( j \). Then \( P(Y_i = y_{js} | y^{i-1}) = P(Y_i = y_{jt} | y^{i-1}) \) which implies that
\[
\sum_{l=1}^{k} \varphi_i(x_l) f_{s_i}(\Phi(x_l, y_{js})) = \sum_{l=1}^{k} \varphi_i(x_l) f_{s_i}(\Phi(x_l, y_{jt})).
\] (31)

However, for a fixed output \( \sum_{l=1}^{k} f_{s_i}(\Phi(x_l, y_{js})) \) returns the sum of the column corresponding to output \( y_{js} \) (similarly for \( y_{jt} \)) and since sub-array \( j \) is weakly symmetric, the column sums are identical. Therefore, (31) can be achieved if \( \varphi_i(x_l) = \varphi_i(x_m) = \frac{1}{k} \) for all \( l, m = 1, \ldots, k \), by which we get \( P(Y_i = y_{js} | y^{i-1}) = P(Y_i = y_{jt} | y^{i-1}) = \frac{1}{|X|} \sum_{l=1}^{k} f_{s_i}(\Phi(x_l, y_{js})) \). Thus for other sub-arrays since they are also weakly-symmetric, the uniform feedback policy will also satisfy the equivalence of conditional output probabilities.

REFERENCES


