Feedback Capacity of the Compound Channel

Brooke Shrader, *Member, IEEE*, and Haim Permuter, *Member, IEEE*

***Abstract*—In this work, we find the capacity of a compound finite-state channel (FSC) with time-invariant deterministic feedback. We consider the use of fixed length block codes over the compound channel. Our achievability result includes a proof of the existence of a universal decoder for the family of FSCs with feedback. As a consequence of our capacity result, we show that feedback does not increase the capacity of the compound Gilbert–Elliot channel. Additionally, we show that for a stationary and uniformly ergodic Markovian channel, if the compound channel capacity is zero without feedback then it is zero with feedback. Finally, we use our result on the FSC to show that the feedback capacity of the memoryless compound channel is given by $\inf_\theta \max_{Q_X} I(X;Y\,|\,\theta)$.**

***Index Terms*—Causal conditioning probability, code-trees, compound channel, directed information, feedback capacity, finite-state channel (FSC), Gilbert–Elliot channel, Pinsker's inequality, Sanov's theorem, types of code-trees, universal decoder.**

I. Introduction

THE compound channel consists of a set of channels indexed by $\theta \in \Theta$ with the same input and output alphabets but different conditional probabilities. In the setting of the compound channel only one actual channel θ is used in all transmissions. The transmitter and the receiver know the family of channels but they have no prior knowledge of which channel is actually used. There is no distribution law on the family of channels and the communication has to be reliable for all channels in the family.

Blackwell *et al.* [1] and independently Wolfowitz [2] showed that the capacity of a compound channel consisting of memoryless channels only, and without feedback, is given by

$$\max_{Q_X} \inf_{\theta} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) \tag{1}$$

where $Q_X(\cdot)$ denotes the input distribution to the channel, $P_{Y\,|\,X,\theta}(\cdot\,|\,\cdot,\theta)$ denotes the conditional probability of a memoryless channel indexed by θ, and the notation $\mathcal{I}(Q_X; P_{Y\,|\,X,\theta})$ denotes the mutual information of channel $P_{Y\,|\,X,\theta}$ for the input distribution Q_X, i.e.,

$$\mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) \triangleq \sum_{x,y} Q_X(x) P_{Y\,|\,X,\theta}(y\,|\,x,\theta) \times \ln \frac{P_{Y\,|\,X,\theta}(y\,|\,x,\theta)}{\sum_{x'} Q_X(x') P_{Y\,|\,X,\theta}(y\,|\,x',\theta)}. \tag{2}$$

Manuscript received November 05, 2007; revised March 04, 2008. Current version published July 15, 2009. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Nice, France, June 2007. This work was completed while B. Shrader was at the University of Maryland, College Park, MD, and H. Permuter was at Stanford University, Stanford, CA.

B. Shrader is with Lincoln Laboratory, Massachusetts Institute of Technology (MIT), Lexington, MA 02420 USA (e-mail: brooke.shrader@ll.mit.edu).

H. Permuter is with the Electrical and Computer Engineering Department, Ben-Gurion University, Be'er-Sheva 84105, Israel (e-mail: haimp@bgu.ac.il).

Communicated by P. Viswanath, Associate Editor for Communications.

Digital Object Identifier 10.1109/TIT.2009.2023727

The capacity in (1) is in general less than the capacity of every channel in the family. Wolfowitz, who coined the term "compound channel," showed that if the transmitter knows the channel θ in use, then the capacity is given by [3, Ch. 4]

$$\inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) = \inf_{\theta} C_\theta \tag{3}$$

where C_θ is the capacity of the channel indexed by θ. This shows that knowledge at the transmitter of the channel θ in use helps in that the infimum of the capacities of the channels in the family can now be achieved. In the case that Θ is a finite set, then it follows from Wolfowitz's result that $\min_\theta C_\theta$ is the feedback capacity of the memoryless compound channel, since the transmitter can use a training sequence together with the feedback to estimate θ with high probability. In this paper, we show that when Θ is not limited to finite cardinality, the feedback capacity of the memoryless compound channel is given by $\inf_\theta C_\theta$. One might be tempted to think that for a compound channel with memory, feedback provides a means to achieve the infimum of the capacities of the channels in the family. However, this is not necessarily true, as we show in Example 1, which is taken from [4] and applied to the compound Gilbert–Elliot channel with feedback. That example is found in Section V.

A comprehensive review of the compound channel and its role in communication is given by Lapidoth and Narayan [5]. Recent results on the Gaussian compound channel for multiuser and multiple-input multiple-output (MIMO) settings can be found in [6]–[8]. Of specific interest in this paper are compound channels with memory which are often used to model wireless communication in the presence of fading [9]–[11]. Lapidoth and Telatar [4] derived the following formula for the compound channel capacity of the class of finite-state channels (FSCs) when there is no feedback available at the transmitter:

$$\lim_{n\to\infty} \max_{Q_{X^n}} \inf_{s_o,\theta} \frac{1}{n} \mathcal{I}(Q_{X^n}; P_{Y^n\,|\,X^n,s_0,\theta}) \tag{4}$$

where s_0 denotes the initial state of the FSC, and $Q_{X^n}(\cdot)$ and $P_{Y^n\,|\,X^n,s_0,\theta}(\cdot\,|\,\cdot,s_0,\theta)$ denote the input distribution and channel conditional probability for block length n. Lapidoth and Telatar's achievability result makes use of a universal decoder for the family of FSCs. The existence of the universal decoder is proved by Feder and Lapidoth in [12] by merging a finite number of maximum-likelihood (ML) decoders, each tuned to a channel in the family Θ.

Throughout this paper, we use the concepts of causal conditioning and directed information which were introduced by Massey in [13]. Kramer extended those concepts and used them in [14] to characterize the capacity of discrete memoryless networks. Subsequently, three different proofs—Tatikonda and Mitter [15], [16], Permuter, Weissman, and Goldsmith [17], and Kim [18]—have shown that directed information and

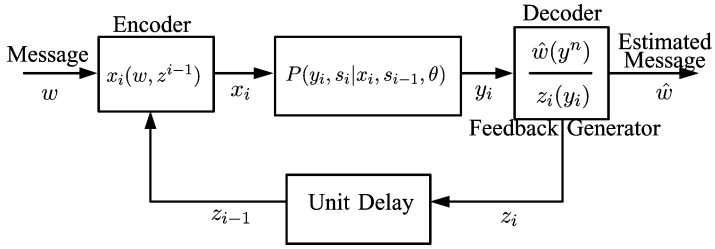

Fig. 1. Compound FSC with feedback that is a time-invariant deterministic function of the channel output.

causal conditioning are useful in characterizing the feedback capacity of a point-to-point channel with memory. In particular, this work uses results from [17] that show that Gallager's [9, Chs. 4, 5] upper and lower bound on capacity of a FSC can be generalized to the case that there is a time-invariant deterministic feedback, $z_{i-1} = f(y_{i-1})$, available at the encoder at time i.

In this paper, we extend Lapidoth and Telatar's work for the case that there is deterministic time-invariant feedback available at the encoder by replacing the regular conditioning with the causal conditioning. Then we use the feedback capacity theorem to study the compound Gilbert–Elliot channel and the memoryless compound channel and to specify a class of compound channels for which the capacity is zero if and only if the feedback capacity is zero. The proof of the feedback capacity of the FSC is found in Section III, which describes the converse result, and Section IV, where we prove achievability. As a consequence of the capacity result, we show in Section V that feedback does not increase the capacity of the compound Gilbert–Elliot channel. We next show in Section VI that for a family of stationary and uniformly ergodic Markovian channels, the capacity of the compound channel is positive if and only if the feedback capacity of the compound channel is positive. Finally, we return to the memoryless compound channel in Section VII and make use of our capacity result to provide a proof of the feedback capacity.[1]

The notation we use throughout is as follows. A capital letter X denotes a random variable and a lower case letter, x, denotes a realization of the random variable. Vectors are denoted using subscripts and superscripts, $x^n = (x_1, \ldots, x_n)$ and $x_i^n = (x_i, \ldots, x_n)$. We deal with discrete random variables where a probability mass function on the channel input is denoted $Q_{X^n}(x^n) = \Pr(X^n = x^n)$ and $P_{Y^n \,|\, X^n, \theta}(y^n \,|\, x^n, \theta) = \Pr(Y^n = y^n \,|\, X^n = x^n, \theta)$ denotes a mass function on the channel output. When no confusion can result, we will omit subscripts from the probability functions, i.e., $Q(x_i \,|\, x^{i-1}, y^{i-1})$ will denote $Q_{X_i \,|\, X^{i-1}, Y^{i-1}}(x_i \,|\, x^{i-1}, y^{i-1})$.

II. Problem Statement and Main Result

The problem we consider is depicted in Fig. 1. A message W from the set $\{1, 2, \ldots, e^{nR}\}$ is to be transmitted over a compound FSC with time-invariant deterministic feedback. The family Θ of FSCs has a common state space $\mathcal{S}$ and common finite input and output alphabets given by $\mathcal{X}$ and $\mathcal{Y}$. For a given channel $\theta \in \Theta$ the channel output at time i is characterized by the conditional probability

$$P(y_i, s_i \,|\, x_i, s_{i-1}, \theta), \quad y_i \in \mathcal{Y}, x_i \in \mathcal{X}, s_i, s_{i-1} \in \mathcal{S} \tag{5}$$

which satisfies the condition $P(y_i, s_i \,|\, x^i, s^{i-1}, y^{i-1}, \theta) = P(y_i, s_i \,|\, x_i, s_{i-1}, \theta)$. The channel θ is in use over the sequence of n channel inputs. The family Θ of channels is known to both the encoder and decoder, however, they do not have knowledge of the channel θ in use before transmission begins.

The message W is encoded such that at time i the codeword symbol X_i is a function of W and the feedback sequence Z^{i-1}. For notational convenience, we will refer to the input sequence $X^i(W, Z^{i-1})$ as simply X^i. The feedback sequence is a time-invariant deterministic function of the output Y_i and is available at the encoder with a single time unit delay. The function performed on the channel output Y_i to form the feedback Z_i is known to both the transmitter and receiver before communication begins. The decoder operates over the sequence of channel outputs Y^n to form the message estimate $\hat{W}$.

For a given initial state $s_0 \in \mathcal{S}$ and channel $\theta \in \Theta$, the channel causal conditioning distribution is given by

$$P(y^n \,\|\, x^n, s_0, \theta) \triangleq \prod_{i=1}^{n} P(y_i \,|\, x^i, y^{i-1}, s_0, \theta). \tag{6}$$

Additionally, we will make use of Massey's directed information [13]. When conditioned on the initial state and channel, the directed information is given by

$$I(X^n \to Y^n \,|\, s_0, \theta) = \sum_{i=1}^{n} I(Y_i; X^i \,|\, Y^{i-1}, s_0, \theta). \tag{7}$$

Our capacity result will involve a maximization of the directed information over the input distribution $Q(x^n \,\|\, z^{n-1})$ which is defined as

$$Q(x^n \,\|\, z^{n-1}) \triangleq \prod_{i=1}^{n} Q(x_i \,|\, x^{i-1}, z^{i-1}). \tag{8}$$

We make use of some of the properties provided in [13], [17] in our work, including the following three which we restate for our problem setting.

1) $P(x^n, y^n \,|\, s_0, \theta) = Q(x^n \,\|\, y^{n-1}) P(y^n \,\|\, x^n, s_0, \theta)$ [13, eq. (3)], [17, Lemma 1].

[1]Although Wolfowitz mentions the feedback problem in discussing the memoryless compound channel [3, Ch. 4], to the best of our knowledge, this result has not been proved in any previous work.

2) $|I(X^n \to Y^n \,|\, \theta) - I(X^n \to Y^n \,|\, S, \theta)| \leq \log|\mathcal{S}|$, where random variable S denotes the state of the FSC [17, Lemma 4].
3) From [17, Lemma 5]

$$\begin{aligned} I(X^n \to Y^n \,|\, s_0, \theta) &= \mathcal{I}(Q_{X^n \,\|\, Y^{n-1}}; P_{Y^n \,\|\, X^n, s_0, \theta}) \\ &= \sum_{x^n, y^n} Q(x^n \,\|\, y^{n-1}) P(y^n \,\|\, x^n, s_0, \theta) \\ &\quad \times \ln \frac{P(y^n \,\|\, x^n, s_0, \theta)}{\sum_{x'^n} Q(x'^n) P(y^n \,\|\, x'^n, s_0, \theta)}. \end{aligned}$$

Note that properties 1) and 3) hold since $Q(x^n \,\|\, y^{n-1}, s_0, \theta) = Q(x^n \,\|\, y^{n-1})$ for our feedback setting, where it is assumed that the state s_0 is not available at the encoder.

For a given initial state s_0 and channel θ, the average probability of error in decoding message w is given by

$$P_{e,w}(s_0, \theta) = \sum_{y^n \in \mathcal{Y}^n : \hat{w} \neq w} P(y^n \,\|\, x^n, s_0, \theta)$$

where x^n is a function of the message w and of the feedback z^{n-1}. The average (over messages) error probability is denoted $P_e(s_0, \theta)$, where $P_e(s_0, \theta) = 1/e^{nR} \sum_w P_{e,w}(s_0, \theta)$. We say that a rate R is achievable for the compound channel with feedback as shown in Fig. 1, if for any $\epsilon > 0$ there exists a code of fixed block length n and rate R, i.e., (n, e^{nR}), such that $P_e(s_0, \theta) < \epsilon$ for all $\theta \in \Theta$ and $s_0 \in \mathcal{S}$. Equivalently, rate R is achievable if there exists a sequence of rate-R codes such that

$$\lim_{n \to \infty} \sup_{s_0, \theta} P_e(s_0, \theta) = 0. \tag{9}$$

This definition of achievable rate is identical to that given in previous work on the compound channel without feedback. A different definition for the compound channel with feedback could also be considered; for instance, in [19], the authors consider codes of variable block length and define achievability accordingly.

The capacity is defined as the supremum over all achievable rates and is given in the following theorem.

Theorem 1: The feedback capacity of the compound FSC is given by

$$C = \lim_{n \to \infty} \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{s_0, \theta} \frac{1}{n} I(X^n \to Y^n \,|\, s_0, \theta). \tag{10}$$

Theorem 1 is proved in Section III, which shows the existence of C and proves the converse, and Section IV, where achievability is established.

III. Existence of C and the Converse

We first state the following proposition, which shows that the capacity C as defined in Theorem 1 exists. The proof is found in Appendix A.

Proposition 1: Let

$$C_n = \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{s_0, \theta} \frac{1}{n} I(X^n \to Y^n \,|\, s_0, \theta). \tag{11}$$

Then C_n is well defined and converges for $n \to \infty$. In addition, let

$$\hat{C}_n = C_n - \frac{\log|\mathcal{S}|}{n}. \tag{12}$$

Then

$$\lim_{n \to \infty} C_n = \sup_n \hat{C}_n. \tag{13}$$

To prove the converse in Theorem 1, we assume a uniform distribution on the message set, for which $H(W) = nR$. Since the message is independent of the channel parameters $H(W) = H(W \,|\, s_0, \theta)$ and we apply Fano's inequality as follows:

$$\begin{aligned} nR &= H(W \,|\, s_0, \theta) \\ &= I(Y^n; W \,|\, s_0, \theta) + H(W \,|\, Y^n, s_0, \theta) \\ &\leq I(Y^n; W \,|\, s_0, \theta) + P_e(s_0, \theta) nR + 1 \\ &= H(Y^n \,|\, s_0, \theta) - H(Y^n \,|\, W, s_0, \theta) + P_e(s_0, \theta) nR + 1 \\ &= \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, s_0, \theta) \\ &\quad - \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, W, s_0, \theta) + P_e(s_0, \theta) nR + 1 \\ &= \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, s_0, \theta) \\ &\quad - \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, W, X^i(W, Z^{i-1}(Y^{i-1})), s_0, \theta) \\ &\quad + P_e(s_0, \theta) nR + 1 \\ &= \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, s_0, \theta) \\ &\quad - \sum_{i=1}^n H(Y_i \,|\, Y^{i-1}, X^i, s_0, \theta) + P_e(s_0, \theta) nR + 1 \\ &= \sum_{i=1}^n I(Y_i; X^i \,|\, Y^{i-1}, s_0, \theta) + P_e(s_0, \theta) nR + 1 \\ &= I(X^n \to Y^n \,|\, s_0, \theta) + P_e(s_0, \theta) nR + 1. \end{aligned}$$

For any code we have

$$I(X^n \to Y^n \,|\, s_0, \theta) \geq nR(1 - P_e(s_0, \theta)) - 1 \tag{14}$$

and therefore

$$\inf_{s_0, \theta} I(X^n \to Y^n \,|\, s_0, \theta) \geq nR \left(1 - \sup_{s_0, \theta} P_e(s_0, \theta)\right) - 1. \tag{15}$$

By combining the above statement with Proposition 1 we have

$$C \geq \hat{C}_n \geq R(1 - \sup_{s_0, \theta} P_e(s_0, \theta)) - \frac{1}{n} - \frac{\log|\mathcal{S}|}{n}. \tag{16}$$

Then for a sequence of codes of rate R with $\lim_{n \to \infty} \sup_{s_0, \theta} P_e(s_0, \theta) = 0$, this implies $R \leq C$.

IV. Achievability

Before proving achievability, we mention a simple case which follows from previous results. If the set Θ has finite cardinality, then achievability follows immediately from the results in [17, Theorem 14], which are true for any FSC with feedback. Hence, we can construct an FSC where the augmented state is (s, θ) and by assuming a positive probability for all initial states (s_0, θ)

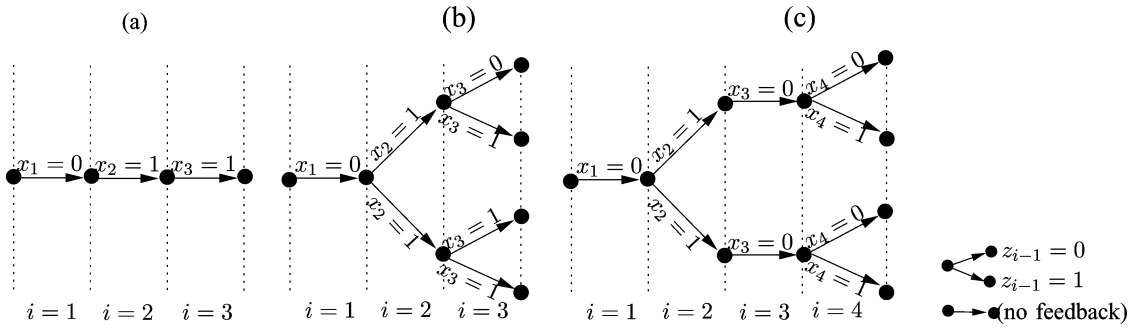

Fig. 2. Illustration of coding scheme for (a) setting without feedback, (b) setting with binary feedback as used in [17], and (c) a code-tree that was created by concatenating smaller code-trees. In the case of no feedback, each message is mapped to a codeword, and in the case of feedback each message is mapped to a code-tree. The third scheme is a code-tree of depth 4 created by concatenating two trees of depth 2.

then we get that for any $\theta \in \Theta, |\Theta| < \infty$ and any $s_0 \in \mathcal{S}$ the rate R is achievable if

$$R < \lim_{n\to\infty} \max_{Q_{X^n \,\|\, Z^{n-1}}} \min_{s_0,\theta} \frac{1}{n} I(X^n \to Y^n \,|\, s_0, \theta). \tag{17}$$

More work is needed in the achievability proof when the set Θ is not restricted to finite cardinality. This is outlined in the following subsections in three steps. In the first step, we assume that the decoder knows the channel θ in use and we show in Theorem 2 that if $R < C$ and if the decoder consists of an ML decoder, then there exist codes for which the error probability decays uniformly over the family Θ and exponentially in the block length. The codes used in showing this result are codes of block length Nm where each subblock of length m is generated independent and identically distributed (i.i.d.) according to some distribution. In the second step, we show in Lemma 3 that if instead the codes are chosen uniformly and independently from a set of possible block-length-Nm codes, then the error probability still decays uniformly over Θ and exponentially in the block length. In the third and final step, we show in Theorem 4 and Lemma 5 that for codes chosen uniformly and independently from a set of block-length-Nm codes, there exists a decoder that for every channel $\theta \in \Theta$ achieves the same error exponent as the ML decoder tuned to θ.

In the sections that follow, $\mathcal{P}(\mathcal{X}^n \,\|\, \mathcal{Z}^{n-1})$ denotes the set of probability distributions on X^n causally conditioned on Z^{n-1}.

A. Achievability for a Decoder Tuned to θ

We begin by proving that if the decoder is tuned to the channel $\theta \in \Theta$ in use, i.e., if the decoder knows the channel θ in use, and if $R < C$, then the average error probability approaches zero. This is proved through the use of random coding and ML decoding.

The encoding scheme consists of randomly generating a *code-tree* for each message w, as shown in Fig. 2(b) for the case of binary feedback. A code-tree has depth n corresponding to the block length and level i designates a set of $|\mathcal{Z}|^{i-1}$ possible codeword symbols. One of the $|\mathcal{Z}|^{i-1}$ symbols is chosen as the input X_i according to the feedback sequence z^{i-1}. The first codeword symbol is generated as $X_1 \sim Q(x_1)$. The second codeword symbol is generated by conditioning on the previous codeword symbol and on the feedback, $X_2 \sim Q(x_2 \,|\, x_1, z_1)$ for all possible values of z_1. For instance, in the binary case, $|\mathcal{Z}| = 2$, two possible values (branches) of X_2 will be generated and the transmitted codeword symbol will be selected from among these two values according to the value of the feedback Z_1. Subsequent codeword symbols are generated similarly, $X_i \sim Q(x_i \,|\, x^{i-1}, z^{i-1})$ for all possible z^{i-1}. For a given feedback sequence z^{n-1}, the input distribution, corresponding to the distribution on a path through the tree of depth n, is

$$Q(x^n \,\|\, z^{n-1}) = \prod_{i=1}^{n} Q(x_i \,|\, x^{i-1}, z^{i-1}). \tag{18}$$

A code-tree of depth n is a vector of $D(n)$ symbols, where

$$D(n) \triangleq \sum_{i=1}^{n} |\mathcal{Z}|^{i-1} = \frac{|\mathcal{Z}|^n - 1}{|\mathcal{Z}| - 1} \tag{19}$$

and each element in the vector takes value from the alphabet $\mathcal{X}$. We denote a random code-tree by $A^{D(n)}$ and a realization of the random code-tree by $a^{D(n)}$. The probability of a tree $a^{D(n)} \in \mathcal{X}^{D(n)}$ is uniquely determined by $Q_{X^n \,\|\, Z^{n-1}}(\cdot \,\|\, \cdot) \in \mathcal{P}(\mathcal{X}^n \,\|\, \mathcal{Z}^{n-1})$. For instance, consider the case of binary feedback, $\mathcal{Z} = \{0, 1\}$, and a tree of depth $n = 2$, for which $D(n) = 3$. A code-tree is a vector $a^3 = (x_1, x_{21}, x_{22})$ where x_1 is the symbol sent at time $i = 1$, x_{21} is the symbol sent at time $i = 2$ for feedback $z_1 = 0$, and x_{22} is the symbol sent at time $i = 2$ for feedback $z_1 = 1$. Then

$$\begin{aligned} &\Pr(A^3 = a^3) \\ &\quad = Q(x_1) Q(x_{21} \,|\, x_1, z_1 = 0) Q(x_{22} \,|\, x_1, z_1 = 1) \end{aligned} \tag{20}$$

which is uniquely determined by $Q_{X^2 \,\|\, Z_1}(\cdot \,\|\, \cdot)$. In general, for a code-tree of depth n, the following holds:

$$\sum_{a^{D(n)} \in \mathcal{X}^{D(n)}} \Pr\left(A^{D(n)} = a^{D(n)}\right) = 1 \tag{21}$$

A code-tree for each message w is randomly generated, and for each message w and feedback sequence z^{n-1} the codeword $x^n(w, z^{n-1})$ is unique. The decoder is made aware of the code-trees for all messages. Assuming that the ML decoder knows the channel θ in use, it estimates the message as follows:

$$\hat{w} = \arg\max_{w} P(y^n \,|\, w, \theta). \tag{22}$$

As shown in [17], since x^i is uniquely determined by w and z^{i-1} and since z^i is a deterministic function of y^i, we have the equivalence

$$P(y^n \,|\, w, \theta) = P(y^n \,\|\, x^n(w, z^{n-1}), \theta) \tag{23}$$

so the ML decoder can be described as

$$\hat{w} = \arg\max_w P(y^n \,\|\, x^n(w, z^{n-1}), \theta). \tag{24}$$

Let $P_e^n(s_0, \theta)$ denote the average (over messages) error probability incurred when a code of block length n is used over channel θ with initial state s_0. The following theorem bounds the error probability uniformly in (s_0, θ) when the decoder knows the channel $\theta \in \Theta$ in use. The theorem is proved in Appendix B.

Theorem 2: For a compound FSC with initial state $s_0 \in \mathcal{S}$, input alphabet $\mathcal{X}$, and output alphabet $\mathcal{Y}$, assuming that the decoder knows the channel θ in use, then there exists a code of rate R and block length Nm, where $N \geq 1$ and m is chosen such that $\hat{C}_m \geq R + \epsilon$, for which the error probability $P_e^{Nm}(s_0, \theta)$ of the ML decoder satisfies

$$P_e^{Nm}(s_0, \theta) \leq |\mathcal{S}| \exp(-Nm\beta(\epsilon, m, |\mathcal{Y}|)) \tag{25}$$

for any $\theta \in \Theta$, where

$$\beta(\epsilon, m, |\mathcal{Y}|) = \begin{cases} m\epsilon^2/(2\log(e|\mathcal{Y}|^m)^2), & \epsilon < \frac{1}{m}(\log(e|\mathcal{Y}|^m))^2 \\ \epsilon - \frac{1}{2m}(\log(e|\mathcal{Y}|^m))^2, & \text{otherwise.} \end{cases} \tag{26}$$

The result in Theorem 2 is shown by the use of a randomly generated code-tree of depth Nm for each message w. For every feedback sequence z^{Nm-1}, the corresponding path in the code-tree is generated by the input distribution $Q_{X^{Nm} \,\|\, Z^{Nm-1}}(\cdot \,\|\, \cdot) \in \mathcal{P}(\mathcal{X}^{Nm} \,\|\, \mathcal{Z}^{Nm-1})$ given by

$$\begin{aligned} Q(x^{Nm} \,\|\, z^{Nm-1}) &= Q_m^*\left(x_1^m \,\|\, z_1^{m-1}\right) \times Q_m^*\left(x_{m+1}^{2m} \,\|\, z_{m+1}^{2m-1}\right) \times \cdots \\ &\quad \times Q_m^*\left(x_{(N-1)m+1}^{Nm} \,\|\, z_{(N-1)m+1}^{Nm-1}\right), \\ &\qquad \forall x^{Nm} \in \mathcal{X}^{Nm}, z^{Nm-1} \in \mathcal{Z}^{Nm-1} \end{aligned} \tag{27}$$

where Q_m^* is the distribution that achieves the supremum in $\hat{C}_m$. The random codebook $\mathcal{C}$ used in proving Theorem 2 consists of e^{NR} code-trees. Each code-tree in the codebook is a concatenated code-tree with depth Nm consisting of N code-trees, each of depth m. For a given feedback sequence z^{Nm-1} (corresponding to a certain path in the concatenated code-tree) the codeword is generated by $Q_{X^{Nm} \,\|\, Z^{Nm-1}}(\cdot \,\|\, \cdot)$. An example of a concatenated code-tree is found in Fig. 2(c).

B. Achievability for Codewords Chosen Uniformly Over a Set

In this subsection, we show that the result in Theorem 2 implies that the error probability can be similarly bounded when codewords are chosen uniformly over a set. In other words, we convert the random coding exponent given in Theorem 2, where it is assumed that the codebook consists of concatenated code-trees of depth Nm in which each subtree of depth m is generated i.i.d. according to Q_m^*, to a new random coding exponent for which the concatenated code-trees in the codebook are chosen uniformly from a set of concatenated code-trees. This alternate type of random coding, where the concatenated code-trees are chosen uniformly from a set, is the coding approach subsequently used to prove the existence of a universal decoder.

We first introduce the notion of types on code-trees. Let $a^{ND(m)}$ denote the concatenation of N depth-m code-trees $a^{D(m)}$, where $D(m)$ is defined in (19) and $a^{ND(m)} \in \mathcal{X}^{ND(m)}$. The type (or empirical probability distribution) of a concatenated code-tree $a^{ND(m)}$ is the relative proportion of occurrences of each code-tree $a^{D(m)} \in \mathcal{X}^{D(m)}$. Equivalently, N multiplied by the type of $a^{ND(m)}$ indicates the number of times each depth-m code-tree from the set $\mathcal{X}^{D(m)}$ occurs in the concatenated code-tree $a^{ND(m)}$. Let $\mathcal{P}_N(\mathcal{X}^{D(m)})$ denote the set of types of concatenated code-trees of depth Nm.

Let $P_e(n, R, Q, P)$ denote the average probability of error incurred when a code-tree of depth n and rate R drawn according to a distribution $Q \in \mathcal{P}(\mathcal{X}^n \,\|\, \mathcal{Z}^{n-1})$ is used over the channel P. We now prove the following result.

Lemma 3: Given $Q_m \in \mathcal{P}(\mathcal{X}^m \,\|\, \mathcal{Z}^{m-1})$, let $Q_{Nm} \in \mathcal{P}(\mathcal{X}^{Nm} \,\|\, \mathcal{Z}^{Nm-1})$ denote the distribution given by the N-fold product of Q_m, i.e.,

$$\begin{aligned} Q_{Nm}(x^{Nm} \,\|\, z^{Nm-1}) &= \prod_{i=1}^{N} Q_m\left(x_{(i-1)m+1}^{im} \,\|\, z_{(i-1)m+1}^{im-1}\right), \\ &\qquad \forall x^{Nm} \in \mathcal{X}^{Nm}, z^{Nm-1} \in \mathcal{Z}^{Nm-1}. \end{aligned} \tag{28}$$

For a given type $\hat{Q}_{Nm} \in \mathcal{P}_N(\mathcal{X}^{D(m)})$, let $\bar{Q}_{Nm} \in \mathcal{P}(\mathcal{X}^{Nm} \,\|\, \mathcal{Z}^{Nm-1})$ denote the distribution that is uniform over the set of concatenated code-trees of type $\hat{Q}_{Nm}$. For every distribution $Q_m \in \mathcal{P}(\mathcal{X}^m \,\|\, \mathcal{Z}^{m-1})$ there exists a type $\hat{Q}_{Nm} \in \mathcal{P}_N(\mathcal{X}^{D(m)})$ whose choice depends on Q_m and N but not on P such that

$$\begin{aligned} P_e(Nm, R, \overline{Q}_{Nm}, P) &\leq \exp(2Nm\delta(N, m, |\mathcal{Z}|)) \\ &\quad \times P_e(Nm, R + m\delta(N, m, |\mathcal{Z}|), Q_{Nm}, P) \end{aligned} \tag{29}$$

for all P, where $\delta(N, m, |\mathcal{Z}|) = |\mathcal{X}|^{D(m)} \log(N+1)/Nm$ tends to 0 as $N \to \infty$.

Proof: The proof follows the approach of [4, Lemma 3] except that our codebook consists of code-trees rather than codewords; we include this proof for completeness in describing the notion of types on code-trees. Given a codebook $\mathcal{C}$ of rate $R + m\delta(N, m, |\mathcal{Z}|)$ chosen according to Q_{Nm}, we can construct a subcode $\mathcal{C}'$ of rate R in the following way. Let Q' denote the type with the highest occurrence in $\mathcal{C}$. The number of types in $\mathcal{C}$ is upper-bounded by $(N+1)^{|\mathcal{X}|^{D(m)}} = \exp(Nm\delta(N, m, |\mathcal{Z}|))$, so the number of concatenated code-trees of type Q' is lower-bounded by $\exp(N(R + m\delta(N, m, |\mathcal{Z}|)))/\exp(Nm\delta(N, m, |\mathcal{Z}|)) = \exp(NR)$. We construct the code $\mathcal{C}'$ by picking the first e^{NR} concatenated code-trees of type Q'. Since $\mathcal{C}'$ is a subcode of $\mathcal{C}$, its average probability of error is upper-bounded by the average probability of error of $\mathcal{C}$ times $|\mathcal{C}|/|\mathcal{C}'| = \exp(Nm\delta(N, m, |\mathcal{Z}|))$.

Conditioned on Q', the codewords in $\mathcal{C}'$ are mutually independent and uniformly distributed over a set of concatenated

code-trees of type Q'. Since $\mathcal{C}$ is a random code, the type Q' is also random, and let π denote the distribution of Q'. Pick a realization of the type Q', denoted $\hat{Q}_{Nm}$, that satisfies $\pi(\hat{Q}_{Nm}) \geq \exp(-Nm\delta(N,m,|\mathcal{Z}|))$. (This is possible since the number of types is upper-bounded by $\exp(Nm\delta(N,m,|\mathcal{Z}|))$.) Then

$$\begin{aligned}
&\pi(\hat{Q}_{Nm})P_e(Nm,R,\bar{Q}_{Nm},P)\\
&\quad\leq \sum_{Q'}\pi(Q')P_e(Nm,R,Q',P) &(30)\\
&\quad\leq \exp(Nm\delta(N,m,|\mathcal{Z}|))\\
&\qquad\times P_e(Nm,R+m\delta(N,m,|\mathcal{Z}|),Q_{Nm},P) &(31)
\end{aligned}$$

and

$$\begin{aligned}
&P_e(Nm,R,\bar{Q}_{Nm},P)\\
&\quad\leq \frac{\exp(Nm\delta(N,m,|\mathcal{Z}|))}{\pi(\hat{Q}_{Nm})}\\
&\qquad\times P_e(Nm,R+m\delta(N,m,|\mathcal{Z}|),Q_{Nm},P) &(32)\\
&\quad\leq \exp(2Nm\delta(N,m,|\mathcal{Z}|))\\
&\qquad\times P_e(Nm,R+m\delta(N,m,|\mathcal{Z}|),Q_{Nm},P). &(33)
\end{aligned}$$

□

Combining this result with Theorem 2, we have that there exists a type $\hat{Q}_{Nm}\in\mathcal{P}_N(\mathcal{X}^{D(m)})$ such that when the codewords are chosen uniformly from the type class of $\hat{Q}_{Nm}$, given by the distribution $\bar{Q}_{Nm}$, the average probability of error is bounded as

$$\begin{aligned}
&P_e(Nm,R,\bar{Q}_{Nm},P)\\
&\quad\leq \exp(2Nm\delta(N,m,|\mathcal{Z}|))|\mathcal{S}|\\
&\qquad\times\exp(-Nm\beta(\epsilon-m\delta(N,m,|\mathcal{Z}|)/2,m,|\mathcal{Y}|)) &(34)\\
&\quad= |\mathcal{S}|\exp\left\{-Nm\left[\beta\left(\epsilon-\frac{1}{2}m\delta(N,m,|\mathcal{Z}|),m,|\mathcal{Y}|\right)\right.\right.\\
&\qquad\left.\left.-2\delta(N,m,|\mathcal{Z}|)\right]\right\}. &(35)
\end{aligned}$$

It is then possible to choose N_0 such that for all $N>N_0$

$$\frac{1}{2}|\mathcal{X}|^{D(m)}\frac{\log(N+1)}{N}<\frac{\epsilon}{2} \tag{36}$$

and

$$2|\mathcal{X}|^{D(m)}\frac{\log(N+1)}{Nm}<\frac{1}{2}\beta\left(\frac{\epsilon}{2},m,|\mathcal{Y}|\right) \tag{37}$$

which implies that the probability of error is bounded as

$$P_e(Nm,R,\bar{Q}_{Nm},P)\leq|\mathcal{S}|\exp\left(-Nm\frac{1}{2}\beta\left(\frac{\epsilon}{2},m,|\mathcal{Y}|\right)\right). \tag{38}$$

C. Existence of a Universal Decoder

We next show that when a codebook is constructed by choosing code-trees uniformly from a set, there exists a universal decoder for the family of finite-state channels with feedback. This result is shown in the following four steps.

- We define the notion of a strongly separable family Θ of channels given by the causal conditioning distribution. The notion of strong separability means that the family is well approximated by a finite subset of the channels in Θ.
- We prove that for strongly separable Θ and code-trees chosen uniformly from a set, there exists a universal decoder.
- We describe the universal decoder which "merges" the ML decoders tuned to a finite subset of the channels in Θ.
- We show that the family of FSCs given by the causal conditioning distribution is a strongly separable family.

Our approach follows precisely the approach of Feder and Lapidoth [12] except that our codebook consists of concatenated code-trees (rather than codewords) and our channel is given by the causal conditioning distribution.

Let $a^{ND(m)}$ denote a concatenated code-tree of depth $Nm, a^{ND(m)}\in\mathcal{X}^{ND(m)}$ where $D(m)=(|\mathcal{Z}|^m-1)/(|\mathcal{Z}|-1)$, and let B_{Nm} denote a set of such code-trees, $B_{Nm}\subseteq\mathcal{X}^{ND(m)}$. As described in Lemma 3, B_{Nm} will be the set of code-trees of type $\hat{Q}_{Nm}\in\mathcal{P}_N(\mathcal{X}^{D(m)})$ and the code-tree for each message will be chosen uniformly from this set, i.e., $\bar{Q}_{Nm}(a^{ND(m)})=1/|B_{Nm}|$ for any $a^{ND(m)}\in B_{Nm}$. As described below, for a given output sequence y^{Nm}, ML decoding will correspond to comparing the functions $P_\theta(y^{Nm}\,|\,a^{ND(m)})$, $a^{ND(m)}\in B_{Nm}$. Note that comparing the functions $P_\theta(y^{Nm}\,|\,a^{ND(m)})$ is equivalent to comparing the channel causal conditioning distributions since $P_\theta(y^{Nm}\,|\,a^{ND(m)})=P_\theta(y^{Nm}\,\|\,x^{Nm})$ as shown below.

$$\begin{aligned}
P_\theta\left(y^{Nm}\,|\,a^{ND(m)}\right)&=\prod_{i=1}^{Nm}P_\theta\left(y_i\,|\,y^{i-1},a^{ND(m)}\right) &(39)\\
&\overset{(a)}{=}\prod_{i=1}^{Nm}P_\theta\left(y_i\,|\,y^{i-1},a^{ND(m)},z^{i-1}\right) &(40)\\
&\overset{(b)}{=}\prod_{i=1}^{Nm}P_\theta(y_i\,|\,y^{i-1},x^i) &(41)\\
&=P_\theta(y^{Nm}\,\|\,x^{Nm}). &(42)
\end{aligned}$$

In the above, (a) holds since z^{i-1} is a known, deterministic function of y^{i-1} and (b) holds since the code-tree $a^{ND(m)}$ together with the feedback sequence z^{i-1} uniquely determines the channel input x^i.

For notational convenience, the results below on the universal decoder are stated for block length n, where $A^{D(n)}$ denotes a code-tree of depth n and B_n denotes a set of such code-trees. These results extend to the set of concatenated code-trees B_{Nm} and any exceptions are described in the text. Furthermore, we introduce the following notation: ϕ_θ denotes the ML decoder tuned to channel θ; $P_e(\theta,\phi)$ denotes the average (over messages and codebooks chosen uniformly from a set) error probability when decoder ϕ is used over channel θ; and $P_e(\theta,\phi\,|\,\mathcal{C})$ denotes the average (over messages) error probability when codebook $\mathcal{C}$ and decoder ϕ is used over channel θ.

Definition 1: A family of channels $\{P_{Y^n\,\|\,X^n,\theta}(\cdot\,\|\,\cdot,\theta),\theta\in\Theta\}$ defined over common input and output alphabets $\mathcal{X},\mathcal{Y}$ is said to be *strongly separable* for the input code-tree sets $\{B_n\}$, $B_n\subseteq\mathcal{X}^{(|\mathcal{Z}|^n-1)/(|\mathcal{Z}|-1)}$, if there exists some $\mu>0$ that upper-bounds the error exponents in the family, i.e., that satisfies

$$\limsup_{n\to\infty}\sup_\theta-\frac{1}{n}\log P_e(\theta,\phi_\theta)<\mu \tag{43}$$

such that for every $\epsilon>0$ and block length n, there exists a subexponential number $K(n)$ (that may depend on μ and on ϵ) of channels $\{\theta_k^{(n)}\}_{k=1}^{K(n)}\subseteq\Theta$

$$\lim_{n\to\infty}\frac{1}{n}\log K(n)=0 \tag{44}$$

that well approximate any $\theta \in \Theta$ in the following sense: For any $\theta \in \Theta$ there exists $\theta_{k^*}^{(n)} \in \Theta, 1 \leq k^* \leq K(n)$, so that

$$P(y^n \,\|\, x^n, \theta) \leq 2^{n\epsilon} P\left(y^n \,\|\, x^n, \theta_{k^*}^{(n)}\right), \quad \forall (x^n, y^n) : P(y^n \,\|\, x^n, \theta) > 2^{-n(\mu + \log|\mathcal{Y}|)} \tag{45}$$

and

$$P(y^n \,\|\, x^n, \theta) \geq 2^{-n\epsilon} P\left(y^n \,\|\, x^n, \theta_{k^*}^{(n)}\right), \quad \forall (x^n, y^n) : P\left(y^n \,\|\, x^n, \theta_{k^*}^{(n)}\right) > 2^{-n(\mu + \log|\mathcal{Y}|)}. \tag{46}$$

The notion of strong separability means that the family Θ is well-approximated by a finite subset $\{\theta_k^{(n)}\}_{k=1}^{K(n)} \subseteq \Theta$ of the channels in the family. In order to prove that the family of finite-state channels with feedback is separable, we will need a value μ that satisfies (43). The error probability $P_e(\theta, \phi_\theta)$ is lower-bounded by the probability that the output sequence Y^{Nm} corresponding to two different messages is the same for a given realization of the channel and code-tree. For a random code-tree this is lower-bounded by a uniform memoryless distribution on the channel output. Then $P_e(\theta, \phi_\theta) \geq |\mathcal{Y}|^{-Nm}$ and a suitable candidate for μ is $1 + \log|\mathcal{Y}|$. The following theorem shows the existence of a universal decoder for a strongly separable family and input code-tree sets B_n. The proof follows from the proof of Theorem 2 in [12] except that we replace the channel conditional distribution $P(y^n \,|\, x^n, \theta)$ with the causal conditioning distribution $P(y^n \,\|\, x^n, \theta)$.

Theorem 4: If a family of channels defined over common finite-input and -output alphabets $\mathcal{X}, \mathcal{Y}$ is strongly separable for the input code-tree sets $\{B_n\}$, then there exists a sequence of rate-R block-length-n codes $\mathcal{C}_n$ and a sequence of decoders $\{u_n\}$ such that

$$\lim_{n\to\infty} \sup_{\theta} \frac{1}{n} \log\left(\frac{P_e(\theta, u_n \,|\, \mathcal{C}_n)}{P_e(\theta, \phi_\theta)}\right) = 0. \tag{47}$$

The universal decoder u_n in Theorem 4 is given by "merging" the ML decoders tuned to channels θ_k, $1 \leq k \leq K(n)$, that are used to approximate the family Θ. In order to describe the merging of the ML decoders, we first present the ranking function M_θ. An ML decoder tuned to the channel θ can be described by a ranking function M_θ defined as the mapping

$$M_\theta : B_{Nm} \times \mathcal{Y}^{Nm} \to \{1, 2, \ldots, |B_{Nm}|\} \tag{48}$$

where a rank of 1 denotes the code-tree $a^{ND(m)}$ that is most likely given output y^{Nm}, rank 2 denotes the second most likely code-tree, and so on. For a given received sequence y^{Nm}, every code-tree in the set B_{Nm} is assigned a rank. For code-trees $a_i^{ND(m)}, a_j^{ND(m)} \in B_{Nm}$

$$P_\theta\left(y^{Nm} \,|\, a_i^{ND(m)}\right) > P_\theta\left(y^{Nm} \,|\, a_j^{ND(m)}\right) \implies M_\theta\left(a_i^{ND(m)}, y^{Nm}\right) < M_\theta\left(a_j^{ND(m)}, y^{Nm}\right). \tag{49}$$

By (42), comparing the function $P_\theta(y^{Nm} \,|\, a^{ND(m)})$ is equivalent to comparing the channel causal conditioning distribution $P_\theta(y^{Nm} \,\|\, x^{Nm})$. Letting ϕ_θ denote the ML decoder tuned to θ, we can describe the decoder as

$$\phi_\theta(y^{Nm}) = w, \quad \text{iff } M_\theta(a^{ND(m)}(w), y^{Nm}) < M_\theta(a^{ND(m)}(w'), y^{Nm}), \ \forall w' \neq w \tag{50}$$

where $a^{ND(m)}(w)$ represents the code-tree chosen for message $w, 1 \leq w \leq e^{NR}$. In the case that multiple code-trees maximize the likelihood $P_\theta(y^{Nm} \,|\, a^{ND(m)})$ for a given y^{Nm}, the ranking function M_θ determines which code-tree (and, correspondingly, message) is chosen by the decoder. In the case that the same code-tree from B_{Nm} is chosen for more than one message, the ranks will be identical and a decoding error will occur. Note that for a given output sequence y^{Nm}, the decoder $\phi_\theta(y^{Nm})$ will not always return the code-tree $a^{ND(m)} \in B_{Nm}$ for which $M_\theta(a^{ND(m)}, y^{Nm}) = 1$, since the code-tree $a^{ND(m)}$ may or may not be in the codebook.

Now consider a set of K channels from the family Θ, given by $\theta_k \in \Theta, 1 \leq k \leq K$. The codebooks for these K channels will be drawn randomly from the set B_{Nm}. (Note that the same set B_{Nm} is used for all channels θ_k since, as shown in Lemma 3, the type $\hat{Q}_{Nm} \in \mathcal{P}_N(\mathcal{X}^{D(m)})$ is chosen independent of the channel P.) The K ML decoders matched to these channels, denoted $\phi_{\theta_1}, \phi_{\theta_2}, \ldots, \phi_{\theta_K}$, can be merged as shown in [12]. The merged decoder u_K is described by its ranking function M_{u_K} which is a mapping

$$M_{u_K} : B_{Nm} \times \mathcal{Y}^{Nm} \to \{1, 2, \ldots, |B_{Nm}|\} \tag{51}$$

that ranks all of the code-trees in B_{Nm} for each output sequence y^{Nm}. The ranking M_{u_K} is established for a given y^{Nm} by assigning rank 1 to the code-tree for which $M_{\theta_1} = 1$, rank 2 to the code-tree for which $M_{\theta_2} = 1$, rank 3 to the code-tree for which $M_{\theta_3} = 1$, and so on. After considering the code-trees with rank 1 for all M_{θ_k}, the code-trees with rank 2 in M_{θ_k}, $1 \leq k \leq K$ are considered in order and added into the ranking M_{u_K}. The process continues until the code-trees with rank $|B_{Nm}|$ for all M_{θ_k} have been assigned a rank in M_{u_K}. Throughout this process, if a code-tree has already been ranked, it is simply skipped over, and its original (higher) ranking is maintained. The rank of a code-tree in M_{u_K} can be upper-bounded according to its rank in M_{θ_k} as shown in [12] and stated as follows:

$$M_{\theta_k}\left(a^{ND(m)}, y^{Nm}\right) = j \implies M_{u_K}\left(a^{ND(m)}, y^{Nm}\right) \leq (j-1)K + k, \quad \forall a^{ND(m)} \in B_{Nm}, \forall k, 1 \leq k \leq K. \tag{52}$$

This bound on the rank in M_{u_K} implies another (looser) upper bound.

$$M_{u_K}\left(a^{ND(m)}, y^{Nm}\right) \leq K M_{\theta_k}\left(a^{ND(m)}, y^{Nm}\right), \quad \forall \left(a^{ND(m)}, y^{Nm}\right) \in B_{Nm} \times \mathcal{Y}^{Nm}, \forall k, 1 \leq k \leq K. \tag{53}$$

Equation (53) can be used to upper-bound the error probability when sequences output from the channel $\theta \in \Theta$ are decoded

by the merged decoder u_K. This is a key element of the proof of Theorem 4. Finally, we state the following lemma, which shows that the family of FSCs defined by the causal conditioning distribution is strongly separable. Together with Theorem 4, this establishes existence of a universal decoder for the problem we consider, and completes our proof of achievability.

Lemma 5: The family of all causal-conditioning FSCs defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ is strongly separable in the sense of Definition 1 for any input code-tree sets $\{B_n\}$.

Proof: See Appendix C. □

V. Compound Gilbert–Elliot Channel

The Gilbert–Elliot channel is a widely used example of an FSC. It has a state space consisting of "good" and "bad" states, $\mathcal{S} = \{G, B\}$ and in either of these two states, the channel is a binary-symmetric channel (BSC). The Gilbert–Elliot channel is a stationary and ergodic Markovian channel, i.e., $P(y_i, s_i \,|\, x_i, s_{i-1}, \theta) = P(s_i \,|\, s_{i-1}, \theta) P(y_i \,|\, x_i, s_{i-1}, \theta)$ is satisfied and the Markov process described by $P(s_i \,|\, s_{i-1}, \theta)$ is a stationary and ergodic process. For a given channel θ, the BSC crossover probability is given by $P_B(\theta)$ for $s_i = B$ and $P_G(\theta)$ for $s_i = G$. The channel state S_i forms a stationary Markov process with transition probabilities

$$g(\theta) = P(S_i = G \,|\, S_{i-1} = B) = 1 - P(S_i = B \,|\, S_{i-1} = B) \tag{54}$$

$$b(\theta) = P(S_i = B \,|\, S_{i-1} = G) = 1 - P(S_i = G \,|\, S_{i-1} = G). \tag{55}$$

For a given θ, the Gilbert–Elliot channel is equivalent to the following additive noise channel:

$$Y_i = X_i \oplus V_i \tag{56}$$

where $\oplus$ denotes modulo-2 addition and $V_i \in \{0, 1\}$. Conditioned on the state process $\{S_i\}_{-\infty}^{+\infty}$, the noise V_i forms a Bernoulli process given by

$$\mathrm{P}(V_i = 1 \,|\, \{S_i\}_{-\infty}^{+\infty}, \theta) = \begin{cases} P_B(\theta), & S_i = B \\ P_G(\theta), & S_i = G. \end{cases} \tag{57}$$

For a given channel θ, the capacity of the Gilbert–Elliot channel is found in [11] and is achieved by a uniform Bernoulli input distribution.

The following example illustrates that the feedback capacity of a channel with memory is in general *not* given by

$$C_{\mathrm{FB}} = \inf_{\theta} C_\theta, \tag{58}$$

as in the memoryless case.

Example 1: [4] Consider the example of a Gilbert–Elliot channel where $P_G(\theta) = 0, P_B(\theta) = 0.5, b(\theta) = g(\theta) = 2^{-\theta}$ for $\theta = 1, 2, 3, \ldots$ with feedback. The compound feedback capacity of this channel is zero because assuming that we start in the bad state, for any block length n, the channel that corresponds to $\theta = n$ will remain in the bad state for the duration of the transmission with probability $(1 - 2^{-n})^n > 1 - n2^{-n} \geq \frac{1}{2}$. While the channel is in the bad state the probability of error for decoding the message is positive with or without feedback, hence no reliable communication is possible.

However, if we fix θ, then the capacity C_θ is at least $1 - h_b(\frac{1}{4})$, because we can use a deep enough interleaver to make the channel look like memoryless BSC with crossover probability $\frac{1}{4}$.

A Gilbert–Elliot channel is described by the four parameters $g(\theta), b(\theta), P_G(\theta)$, and $P_B(\theta)$ that lie between 0 and 1 and for any fixed n, $P(y^n \,\|\, x^n, s_0)$ is continuous in those parameters. The continuity of $P(y^n \,\|\, x^n, s_0)$ follows from the fact that $P(y_i, s_i \,|\, x_i, s_{i-1})$ is continuous in the four parameters for any $i \geq 1$, and also because (as shown in Appendix C in (111) and (113)) we can express $P(y^n \,\|\, x^n, s_0)$ as

$$\begin{aligned} P(y^n \,\|\, x^n, s_0) &= \sum_{s^n} P(y^n, s^n \,\|\, x^n, s_0) \\ &= \sum_{s^n} \prod_{i=1}^{n} P(y_i, s_i \,|\, x_i, s_{i-1}). \end{aligned} \tag{59}$$

Let us denote by $\bar{\Theta}$ the closure of the family of channels. Hence, instead of $\inf_{\theta \in \Theta}$ we can write $\min_{\theta \in \bar{\Theta}}$ since $\bar{\Theta}$ is compact and since $\mathcal{I}(Q; P)$ is continuous in P. Now, let $Q_u(x^n)$ denote the uniform distribution over $\mathcal{X}^n$. We have

$$\begin{aligned} \max_{Q} \min_{s_0, \theta} \mathcal{I}(Q; P) &\overset{(a)}{\leq} \min_{s_0, \bar{\theta}} \max_{Q} \mathcal{I}(Q; P) \\ &\overset{(b)}{=} \min_{s_0, \bar{\theta}} I(Q_u; P) \end{aligned} \tag{60}$$

where (a) follows from the fact that $\max \min \leq \min \max$ and (b) follows from the fact that for any channel a uniform distribution maximizes its capacity. Therefore, we can restrict the maximization to the uniform distribution Q_u instead of $Q(x^n \,\|\, y^{n-1})$. Hence, feedback does not increase the capacity of the compound Gilbert–Elliot channel. This result holds for any family of FSCs for which the uniform input distribution achieves the capacity of each channel in the family and is closely related to Alajaji's result [20] that feedback does not increase the capacity of discrete additive noise channels.

VI. Feedback Capacity Is Positive if and Only if Capacity Without Feedback Is Positive

In this section, we show that the capacity of a compound channel that consists of stationary and uniformly ergodic Markovian channels is positive if and only if it is positive for the case that feedback is allowed. The intuition of this result comes mainly from Lemma 9 that states that

$$\max_{Q_{X^n \,\|\, Y^{n-1}}} I(X^n \to Y^n) = 0 \iff \max_{Q_{X^n}} I(X^n \to Y^n) = 0. \tag{61}$$

The reason our proof is restricted to the family of channels that are stationary and uniformly ergodic Markovian is because for this family of channels we can show that the capacity is zero only if for every finite n,

$$\max_{Q_{X^n \,\|\, Y^{n-1}}} \inf_{\theta} I(X^n \to Y^n \,|\, \theta) = 0. \tag{62}$$

A stationary and ergodic Markovian channel is an FSC where the state of the channel is a stationary and ergodic Markov process that is not influenced by the channel input and output. In other words, the conditional probability of the channel output and state given the input and previous state is given by

$$P(y_i, s_i \,|\, x_i, s_{i-1}, \theta) = P(s_i \,|\, s_{i-1}, \theta) P(y_i \,|\, x_i, s_{i-1}, \theta) \quad (63)$$

where the Markov process, described by the transition probability $P(s_i \,|\, s_{i-1}, \theta)$, is stationary and ergodic. We say that the family of channels is *uniformly ergodic* if all channels in the family are ergodic and for all $\epsilon > 0$ there exists an $M(\epsilon)$ such that for all $n > M$

$$\left|\Pr(S_n = s \,|\, s_0, \theta) - P(s \,|\, \theta)\right| \leq \epsilon, \ \ \forall s_0 \in \mathcal{S}, s \in \mathcal{S}, \theta \in \Theta \quad (64)$$

where $P(s \,|\, \theta)$ is the stationary (equilibrium) distribution of the state for channel θ. We define the sequence $C_n^{\text{Markovian}}$ as

$$C_n^{\text{Markovian}} = \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta} \frac{1}{n} I(X^n \to Y^n \,|\, \theta). \quad (65)$$

Theorem 6: The channel capacity of a family of stationary and uniformly ergodic Markovian channels is positive if and only if the feedback capacity of the same family is positive.

Since a memoryless channel is an FSC with only one state, the theorem implies that the feedback capacity of a memoryless compound channel is positive if and only if it is positive without feedback. The theorem also implies that for a stationary and ergodic point-to-point channel (not compound), feedback does not increase the capacity for cases that the capacity without feedback is zero. The stationarity of the channels in Theorem 6 is not necessary since according to our achievability definition, if a rate is less than the capacity, it is achievable regardless of the initial state. We assume stationarity here in order to simplify the proofs. The uniform ergodicity is essential to the proof that is provided here but there are also other family of channels that have this property. For instance, for the regular point-to-point Gaussian channel this result can be concluded from factor two result that claims that feedback at most doubles capacity (cf., [21]–[23]).

The proof of Theorem 6 is based on the following lemmas. We refer the reader to Appendix D for the proofs of these lemmas.

Lemma 7: For any channel with feedback, if the input to the channel is distributed according to

$$Q(x^n \,\|\, z^{n-1}) = Q\left(x_1^k \,\|\, z_1^{k-1}\right) Q\left(x_{k+1}^n \,\|\, z_{k+1}^{n-1}\right)$$

then

$$I(X^n \to Y^n) \geq I(X^k \to Y^k) + I\left(X_{k+1}^n \to Y_{k+1}^n\right). \quad (66)$$

Lemma 8: The feedback capacity of a family of stationary and uniformly ergodic Markovian channels is

$$\lim_{n \to \infty} C_n^{\text{Markovian}}. \quad (67)$$

The limit of $C_n^{\text{Markovian}}$ exists and is equal to $\sup_n C_n^{\text{Markovian}}$.

Lemma 9: Let the input distribution to an arbitrary channel be uniform over the input $\mathcal{X}^n$, i.e., $Q(x^n) = \frac{1}{|\mathcal{X}|^n}$. If under this input distribution $I(X^n \to Y^n)$=0, then the channel has the property that $P(y^n \,||\, x^n) = P(y^n)$ for all $x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$ and this implies that

$$\max_{Q_{X^n \,\|\, Y^{n-1}}} I(X^n \to Y^n) = 0. \quad (68)$$

Proof of Theorem 6: Let C_{NFB} denote the capacity without feedback and C_{FB} denote the capacity with feedback. $C_{\text{NFB}} = 0 \Leftarrow C_{\text{FB}} = 0$ is trivial. To show that $C_{\text{NFB}} = 0 \Longrightarrow C_{\text{FB}} = 0$, we use Lemma 8 to conclude that since $C_{\text{NFB}} = 0$ then $\sup_n C_n^{\text{Markovian}} = 0$ and therefore for any $n \geq 1$

$$\max_{Q_{X^n}} \inf_{\theta} I(X^n \to Y^n \,|\, \theta) = 0. \quad (69)$$

In order to conclude the proof, we show that if (69) holds, then it also holds when we replace Q_{X^n} by $Q_{X^n \,\|\, Y^{n-1}}$. Since $I(X^n \to Y^n)$ is continuous in $P(y^n \,||\, x^n)$ and since the set Θ is a subset of the unit simplex which is bounded, then the infimum over the set Θ can be replaced by the minimum over the closure of the set Θ. Since (69) holds also for the case that Q_{X^n} is restricted to be the uniform distribution, then Lemma 9 implies that the channel that satisfies $P(y^n \,||\, x^n) = P(y^n)$ for all $x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$ is in the closure of Θ and therefore

$$\max_{Q_{X^n \,\|\, Y^{n-1}}} \inf_{\theta} I(X^n \to Y^n \,|\, \theta) = 0. \quad (70)$$

□

VII. Feedback Capacity of the Memoryless Compound Channel

Recall that the capacity of the memoryless compound channel (without feedback) is [1], [2]

$$\max_{Q_X} \inf_{\theta} \mathcal{I}(Q_X; P_{Y \,|\, X, \theta}). \quad (71)$$

Wolfowitz also showed [3] that when θ is known to the encoder, the capacity of the memoryless compound channel is given by switching the inf and the max, i.e.,

$$\inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y \,|\, X, \theta}). \quad (72)$$

In this section, we make use of Theorem 1 to show that (72) is equal to the feedback capacity of the memoryless compound channel.

A. Finite Family of Memoryless Channels

Based on Wolfowitz's result it is straightforward to show that if the family of memoryless channels is finite, $|\Theta| < \infty$, then the feedback capacity of the compound channel is given by switching the max and the min

$$\min_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y \,|\, X, \theta}). \quad (73)$$

This result can be achieved in two steps. Given a probability of error $P_e > 0$, first, the encoder will use M uses of the channels in order to estimate the channel with probability of error less than $\frac{P_e}{2}$. Since the number of channels is finite such an M exists. In the second step, the encoder will use a coding scheme with block length N adapted for the estimated channel to obtain an error probability that is smaller than $\frac{P_e}{2}$. Hence, we get that the total error of the code of length $M + N$ is smaller than P_e.

B. Arbitrary Family of Memoryless Channels

For the case that the number of channels is infinite, the argument above does not hold, since there is no guarantee that for any $P_e > 0$ there exists a block length $n(P_e)$ such that an (e^{nR}, n) code achieves an error less than P_e for all channels in the family.[2] However, we are able to establish the feedback capacity using our capacity theorem for the compound FSC, and the result is stated in the following theorem.

Theorem 10: The feedback capacity of the memoryless compound channel is

$$\inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}). \tag{74}$$

Theorem 10 is a direct result of Theorem 1 and the following lemma.

Lemma 11: For a family Θ of memoryless channels we have

$$\lim_{n\to\infty} \frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) = \inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}). \tag{75}$$

The proof of Lemma 11 requires two lemmas, which we state below. The proofs of Lemmas 12 and 13 are found in Appendix E.

Lemma 12: Let $Q_X^1 = \arg\max_{Q_X} \mathcal{I}(Q_X, P_{Y\,|\,X,\theta_1})$ and $Q_X^2 = \arg\max_{Q_X} \mathcal{I}(Q_X, P_{Y\,|\,X,\theta_2})$. For two conditional distributions $P_{Y\,|\,X,\theta_1}$ and $P_{Y\,|\,X,\theta_2}$ with

$$\Delta = \|P_{Y\,|\,X,\theta_1} - P_{Y\,|\,X,\theta_2}\|_1 = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} |P_{Y\,|\,X,\theta_1}(y\,|\,x,\theta_1) - P_{Y\,|\,X,\theta_2}(y\,|\,x,\theta_2)| \tag{76}$$

there exists an upper bound

$$|\mathcal{I}\left(Q_X^2, P_{Y\,|\,X,\theta_1}\right) - \mathcal{I}\left(Q_X^1, P_{Y\,|\,X,\theta_1}\right)| \le \eta(\Delta) \tag{77}$$

where $\eta(\Delta) \to 0$ as $\Delta \to 0$.

Lemma 13: For any $\delta > 0$, any $\epsilon > 0$, and any channel $P_{Y\,|\,X}$, there exists an M such that we can choose a channel $P_{Y\,|\,X,\hat{\theta}}$ as a function of M inputs and outputs such that

$$\Pr\{\Delta > \epsilon\} \le \delta \tag{78}$$

where Δ denotes the L_1 distance between the estimated channel $P_{Y\,|\,X,\hat{\theta}}$ and the actual channel $P_{Y\,|\,X}$, i.e.,

$$\Delta = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \left| P_{Y\,|\,X,\hat{\theta}}(y\,|\,x,\hat{\theta}) - P_{Y\,|\,X}(y\,|\,x) \right|. \tag{79}$$

[2]In a private communication with A. Tchamkerten [24], it was suggested that the feedback capacity of the memoryless compound channel with an infinite family can also be established using the results in [12] (which show that the family of all discrete memoryless channels is strongly separable). The family is finitely quantized, a training scheme is used to estimate the appropriate quantization cell, the coding is performed according to the representative channel of that cell, and the decoding is done universally as in [12].

Proof of Lemma 11: We prove the equality by showing the following two inequalities hold:

$$\frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) \le \inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) \tag{80}$$

$$\frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) \ge \inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) - \epsilon_n \tag{81}$$

where $\epsilon_n \to 0$ as $n \to \infty$. Inequality (80) is proved by the fact that $\max\inf$ is less than or equal to $\inf\max$ and by the fact that for a memoryless channel an i.i.d input maximizes the directed information

$$\begin{aligned} &\frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) \\ &\quad \le \frac{1}{n} \inf_{\theta} \max_{Q_{X^n\,\|\,Y^{n-1}}} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) \\ &\quad = \inf_{\theta} \max_{Q_X} \mathcal{I}(Q_X; P_{Y\,|\,X,\theta}). \end{aligned} \tag{82}$$

In order to prove inequality (81) we consider the following input distribution. The first M inputs are used to estimate the channel and we denote the estimated channel as $\hat{\theta}$. After the first M inputs, the input distribution is the i.i.d. distribution that maximizes the mutual information between the input and the output for the channel $\hat{\theta}$. According to Lemma 13, we can estimate the channel to within an L_1 distance smaller than $\epsilon > 0$ with probability greater than $1 - \delta$, where $\delta > 0$. According to Lemma 12, by adjusting the input distribution to a channel that is at L_1 distance less than ϵ from the actual channel in use, we lose an amount that goes to zero as $\epsilon \to 0$. Under the input distribution described above we have the following sequence of inequalities:

$$\begin{aligned} &\frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \mathcal{I}(Q_{X^n\,\|\,Y^{n-1}}; P_{Y^n\,\|\,X^n,\theta}) \\ &\overset{(a)}{=} \frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} I(X^n \to Y^n\,|\,\theta) \\ &\overset{(b)}{\ge} \frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \sum_{i=M(\delta,\epsilon)+1}^{n} I(X^i; Y_i\,|\,Y^{i-1}) \\ &\overset{(c)}{\ge} \frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \sum_{i=M+1}^{n} I\left(X_{M+1}^i; Y_i\,|\,Y^{i-1}, X^M\right) \\ &\overset{(d)}{=} \frac{1}{n} \max_{Q_{X^n\,\|\,Y^{n-1}}} \inf_{\theta} \\ &\qquad \sum_{i=M+1}^{n} I\left(X_{M+1}^i; Y_i\,|\,Y_{M+1}^{i-1}, X^M, Y^M, \hat{\Theta}(X^M, Y^M)\right) \\ &\overset{(e)}{\ge} \frac{1}{n} \max_{Q_{X\,|\,\hat{\theta}}} \inf_{\theta} (n-M) I(X; Y\,|\,\theta, \hat{\Theta}) \\ &\overset{(f)}{=} \frac{1}{n} \max_{Q_{X\,|\,\hat{\theta}}} \inf_{\theta} (n-M) \sum_{\hat{\theta}} P(\hat{\theta}) \mathcal{I}(Q_{X\,|\,\hat{\theta}}; P_{Y\,|\,X,\theta}) \\ &\overset{(g)}{\ge} \frac{1}{n} \max_{Q_{X\,|\,\theta}} \inf_{\theta} (n-M)(1-\delta)(\mathcal{I}(Q_{X\,|\,\theta}; P_{Y\,|\,X,\theta}) - \eta(\epsilon)) \\ &\overset{(h)}{=} \frac{1}{n} \inf_{\theta} \max_{Q_X} (n-M)(1-\delta)(\mathcal{I}(Q_X; P_{Y\,|\,X,\theta}) - \eta(\epsilon)) \end{aligned} \tag{83}$$

where

(a) and (f) follow from a change of notation.

(b) follows the fact that we sum fewer elements. The parameter M is a function of $\epsilon > 0$ and $\delta > 0$ and is determined according to Lemma 13. For brevity of notation we denote $M(\epsilon, \delta)$ simply as M.

(c) follows from the fact that

$$H(Y_i|Y^{i-1}) \geq H(Y_i|Y^{i-1}, X^M).$$

(d) follows from the fact that the estimated channel is a random variable denoted as $\hat{\Theta}$ and it is a deterministic function of X^M, Y^M as described in Lemma 13.

(e) follows by restricting the input distribution $Q_{X^n \,\|\, Y^{n-1}}$ to one that uses first M uses of the channel to estimate as described in Lemma 13, and then uses an i.i.d. distribution, i.e., for $i > M$,

$$\begin{aligned} Q(x_i \,|\, x^{i-1}, y^{i-1}) &= Q(x_i \,|\, x^{i-1}, y^{i-1}, \hat{\theta}(x^M, y^M))) \\ &= Q(x_i \,|\, \hat{\theta}). \end{aligned}$$

(g) follows from the fact that with probability $1-\delta$ we have that the L_1 distance $\|P_{Y \,|\, X,\theta} - P_{Y \,|\, X,\hat{\theta}}\|_1 \leq \epsilon$ and by applying Lemma 12, which states that for this case we lose $\eta(\epsilon)$ where $\eta(\epsilon) \to 0$ as $\epsilon \to 0$.

(h) follows from the fact that $\inf_\theta \max_{Q_X}$ is identical to $\max_{Q_{X \,|\, \theta}} \inf_\theta$.

Finally, since M is fixed for any $\epsilon > 0, \delta > 0$ then we can achieve any value below $\inf_\theta \max_{Q_X} \mathcal{I}(Q_X; P_{Y \,|\, X,\theta})$ for large n. Therefore inequality (81) holds. □

VIII. Conclusion

The compound channel is a simple model for communication under channel uncertainty. The original work on the memoryless compound channel without feedback characterizes the capacity [1], [2], which is less than the capacity of each channel in the family, but the reliability function remains unknown. An adaptive approach to using feedback on an unknown memoryless channel is proposed in [19], where coding schemes that universally achieve the reliability function (the Burnashev error exponent) for certain families of channels (e.g., for a family of BSCs) are provided. By using the variable-length coding approach in [19], the capacity of the channel in use can be achieved. In our work, we consider the use of fixed-length block codes and aim to ensure reliability for every channel in the family; as a result, our capacity is limited by the infimum of the capacities of the channels in the family. For the compound channel with memory that we consider, we have characterized an achievable random coding exponent, but the reliability function remains unknown.

The encoding and decoding schemes used in proving our results have a number of practical limitations, including the memory requirements for storing codebooks consisting of concatenated code-trees at both the transmitter and receiver as well as the complexity involved in merging the ML decoders tuned to a number of channels that is polynomial in the block length. As such, our work motivates a search for more practical schemes for feedback communication over the compound channel with memory.

Appendix A
Proof of Proposition 1

The proposition is nearly identical to [4, Proposition 1], except that we replace $I(X^n; Y^n \,|\, s_0, \theta)$ by $I(X^n \to Y^n \,|\, s_0, \theta)$ and $Q(x^n)$ by $Q(x^n \,\|\, z^{n-1})$ using results from [17] on directed mutual information and causal conditioning. We first prove the following lemma, which is needed in the proof of Proposition 1. The lemma shows that directed information is uniformly continuous in $Q_{X^n \,\|\, Y^{n-1}}$. For our time-invariant deterministic feedback model, $Q(x^n \,\|\, y^{n-1}) = Q(x^n \,\|\, z^{n-1})$, and the lemma holds for any such feedback.

Lemma 14: (*Uniform continuity of directed information*) If $Q^1_{X^n \,\|\, Y^{n-1}}$ and $Q^2_{X^n \,\|\, Y^{n-1}}$ are two causal conditioning distributions such that

$$\sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} |Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1})| \leq \Delta \leq \frac{1}{2} \tag{84}$$

then for a fixed $P_{Y^n \,\|\, X^n}$

$$\begin{aligned} \left| \mathcal{I}\left(Q^1_{X^n \,\|\, Y^{n-1}}; P_{Y^n \,\|\, X^n}\right) - \mathcal{I}\left(Q^2_{X^n \,\|\, Y^{n-1}}; P_{Y^n \,\|\, X^n}\right) \right| \\ \leq -\Delta \log \frac{\Delta}{|\mathcal{Y}^n|^2}. \end{aligned} \tag{85}$$

Proof: Directed information can be expressed as a difference between two terms $I(X^n \to Y^n) = H(Y^n) - H(Y^n \,\|\, X^n)$. Let us consider the total variation of $P^1_{Y^n}(\cdot) - P^2_{Y^n}(\cdot)$.

$$\begin{aligned} &\sum_{y^n} |P^1(y^n) - P^2(y^n)| \\ &\quad = \sum_{y^n} \left| \sum_{x^n} P^1(x^n, y^n) - P^2(x^n, y^n) \right| \\ &\quad = \sum_{y^n} \Bigg| \sum_{x^n} Q^1(x^n \,\|\, y^{n-1}) P(y^n \,\|\, x^n) \\ &\qquad\qquad - Q^2(x^n \,\|\, y^{n-1}) P(y^n \,\|\, x^n) \Bigg| \\ &\quad \leq \sum_{y^n} \sum_{x^n} P(y^n \,\|\, x^n) \left| Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1}) \right| \\ &\quad \leq \sum_{y^n} \sum_{x^n} \left| Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1}) \right| \\ &\quad \leq \Delta \end{aligned} \tag{86}$$

By invoking the continuity lemma of entropy [25, p, 33, Theorem 2.7], we get

$$\left| H^1(Y^n) - H^2(Y^n) \right| \leq -\Delta \log \frac{\Delta}{|\mathcal{Y}^n|} \tag{87}$$

where $H^1(Y^n)$ and $H^2(Y^n)$ are the entropies induced by $P^1_{Y^n}(\cdot)$ and $P^2_{Y^n}(\cdot)$, respectively. Now let us consider the difference $H^1(Y^n \,\|\, X^n) - H^2(Y^n \,\|\, X^n)$:

$$\begin{aligned}
&\left|H^1(Y^n \,\|\, X^n) - H^2(Y^n \,\|\, X^n)\right| \\
&\quad = \left| \sum_{x^n, y^n} -P^1(x^n, y^n) \log P(y^n \,\|\, x^n) \right. \\
&\qquad \left. + P^2(x^n, y^n) \log P(y^n \,\|\, x^n) \right| \\
&\quad = \left| \sum_{x^n, y^n} -P(y^n \,\|\, x^n) Q^1(x^n \,\|\, y^{n-1}) \log P(y^n \,\|\, x^n) \right. \\
&\qquad \left. + P(y^n \,\|\, x^n) Q^2(x^n \,\|\, y^{n-1}) \log P(y^n \,\|\, x^n) \right| \\
&\quad = \left| \sum_{x^n, y^n} -P(y^n \,\|\, x^n) \log P(y^n \,\|\, x^n) \right. \\
&\qquad \left. \times \left(Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1})\right) \right| \\
&\quad \le \left| \sum_{x^n, y^n} -P(y^n \,\|\, x^n) \log P(y^n \,\|\, x^n) \right| \\
&\qquad \left. Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1}) \right| \\
&\quad \le \left(\sum_{x^n, y^n} -P(y^n \,\|\, x^n) \log P(y^n \,\|\, x^n) \right) \\
&\qquad \times \left(\sum_{x^n, y^n} \left| Q^1(x^n \,\|\, y^{n-1}) - Q^2(x^n \,\|\, y^{n-1}) \right| \right) \\
&\quad \le \log |\mathcal{Y}^n| \Delta
\end{aligned} \tag{88}$$

By combining inequalities (87) and (88) we conclude the proof of the lemma. □

By Lemma 14, $I(X^n \to Y^n \,|\, s_0, \theta)$ is uniformly continuous in $Q_{X^n \,\|\, Z^{n-1}}$. Since $Q_{X^n \,\|\, Z^{n-1}}$ is a member of a compact set, the maximum over $Q_{X^n \,\|\, Z^{n-1}}$ is attained and C_n is well-defined.

Next, we invoke a result similar to [4, Lemma 5]. Given integers k and m such that $k + m = n$, input sequences $x_1^k = (x_1, \ldots, x_k)$, and $x_{k+1}^n = (x_{k+1}, \ldots, x_n)$ with corresponding output sequences y_1^k and y_{k+1}^n, let $Q_{X^n \,\|\, Z^{n-1}}$ be defined as

$$Q(x^n \,\|\, z^{n-1}) = Q\left(x_1^k \,\|\, z_1^{k-1}\right) Q\left(x_{k+1}^n \,\|\, z_{k+1}^{n-1}\right).$$

Then

$$\begin{aligned}
&\inf_{s_0, \theta} I(X^n \to Y^n \,|\, s_0, \theta) \\
&\quad \ge \inf_{s_0, \theta} I\left(X_1^k \to Y_1^k \,|\, s_0, \theta\right) \\
&\qquad + \inf_{s_0, \theta} I\left(X_{k+1}^n \to Y_{k+1}^n \,|\, s_k, \theta\right) - \log |\mathcal{S}|.
\end{aligned}$$

This result follows from [4, Lemma 5] and [17, Lemma 4].

Finally, if we let $Q(x_1^k \,\|\, z_1^{k-1})$ and $Q(x_{k+1}^n \,\|\, z_{k+1}^{n-1})$ achieve the maximizations in C_k and C_m, respectively, then we have

$$\begin{aligned}
nC_n &\ge \inf_{s_0, \theta} I(X^n \to Y^n \,|\, s_0, \theta) \\
&\ge \inf_{s_0, \theta} I\left(X_1^k \to Y_1^k \,|\, s_0, \theta\right) \\
&\quad + \inf_{s_0, \theta} I\left(X_{k+1}^n \to Y_{k+1}^n \,|\, s_k, \theta\right) - \log |\mathcal{S}| \\
&= kC_k + mC_m - \log |\mathcal{S}|,
\end{aligned}$$

or equivalently

$$n\hat{C}_n \ge k\hat{C}_k + m\hat{C}_m.$$

Clearly, $\lim_{n \to \infty} C_n = \lim_{n \to \infty} \hat{C}_n$, and by the convergence of a super-additive sequence, $\lim_{n \to \infty} \hat{C}_n = \sup_n \hat{C}_n$.

Appendix B
Proof of Theorem 2

The theorem is proved through a collection of results in [4] and [17]. Let $P^n_{e,w}(\theta)$ denote the error probability of the ML decoder when a random code-tree of block length n is used at the encoder

$$P^n_{e,w}(\theta) = \sum_{y^n \in \mathcal{Y}^n : \hat{w} \ne w} P(y^n \,\|\, x^n(w, z^{n-1}), \theta). \tag{89}$$

The following corollary to [17, Theorem 9] bounds the expected value $E[P^n_{e,w}(\theta)]$, where the expectation is with respect to the randomness in the code. The result holds for any initial state s_0.

Corollary 15: Suppose that an arbitrary message $w, 1 \le w \le e^{nR}$ enters the encoder with feedback and that ML decoding tuned to θ is employed. Then the average probability of decoding error over the ensemble of codes is bounded, for any choice of $\rho, 0 < \rho \le 1$, by

$$\begin{aligned}
&E\left[P^n_{e,w}(\theta)\right] \\
&\le (e^{nR} - 1)^\rho \sum_{y^n} \left[\sum_{x^n} Q(x^n \,\|\, z^{n-1}) P(y^n \,\|\, x^n, \theta)^{\frac{1}{1+\rho}} \right]^{1+\rho}.
\end{aligned} \tag{90}$$

Proof: Identical to [17, Proof of Theorem 9] except that $P(y^n \,\|\, x^n)$ is replaced by $P(y^n \,\|\, x^n, \theta)$. □

Next, we let $P^n_e(s_0, \theta)$ denote the average (over messages) error probability incurred when a code-tree of block length n is used over channel θ with initial state s_0. Using Corollary 15, we can bound $P^n_e(s_0, \theta)$ as in the following Corollary to [17, Theorem 10].

Corollary 16: For a compound FSC with $|\mathcal{S}|$ states, where the codewords are drawn independently according to a given distribution $Q_n \in \mathcal{P}(\mathcal{X}^n \,\|\, \mathcal{Z}^{n-1})$ and ML decoding tuned to θ is employed, the average probability of error $P^n_e(s_0, \theta)$ for any initial state $s_0 \in \mathcal{S}$, channel $\theta \in \Theta$, and $\rho, 0 \le \rho \le 1$ is bounded as

$$P^n_e(s_0, \theta) \le |\mathcal{S}| \exp(-n(F^n(\rho, Q_n, \theta) - \rho R)) \tag{91}$$

where

$$F^n(\rho, Q_n, \theta) = \frac{-\rho \log |\mathcal{S}|}{n} + \min_{s_0} E_0(\rho, Q_n, s_0, \theta)$$

and

$$E_0(\rho, Q_n, s_0, \theta) = -\frac{1}{n} \log \sum_{y^n} \left[\sum_{x^n} Q_n P(y^n \,\|\, x^n, s_0, \theta)^{\frac{1}{1+\rho}} \right]^{1+\rho}. \tag{92}$$

Proof: Identical to [17, Proof of Theorem 10] except for: (i) we replace $P(y^n \,\|\, x^n, s_0)$ by $P(y^n \,\|\, x^n, s_0, \theta)$, (ii) we consider the error averaged over all messages (rather than the error for an arbitrary message w), and (iii) we assume a fixed input distribution $Q_{X^n \,\|\, Z^{n-1}}$ rather than minimizing the error probability over all $Q_{X^n \,\|\, Z^{n-1}}$. □

The two results stated above provide us with a bound on the error probability, however, the bound depends on the channel θ in use. Instead, we would like to bound the error probability uniformly over the class Θ. To do so, we cite the following two lemmas from previous work.

Lemma 17: Given $Q_k \in \mathcal{P}(\mathcal{X}^k \,\|\, \mathcal{Z}^{k-1})$ and $Q_m \in \mathcal{P}(\mathcal{X}^m \,\|\, \mathcal{Z}^{m-1})$, let $m = n - k$ and define

$$Q_n \left(x_1^n \,\|\, z_1^{n-1}\right) = Q_k \left(x_1^k \,\|\, z_1^{k-1}\right) Q_m \left(x_{k+1}^n \,\|\, z_{k+1}^{n-1}\right). \tag{93}$$

Then $F^n(\rho, Q_n, \theta)$ as defined in Corollary 16 satisfies

$$F^n(\rho, Q_n, \theta) \geq \frac{k}{n} F^k(\rho, Q_k, \theta) + \frac{m}{n} F^m(\rho, Q_m, \theta). \tag{94}$$

Proof: Identical to[17, Proof of Lemma 12] except that we replace $P(y^n \,\|\, x^n, s_0)$ by $P(y^n \,\|\, x^n, s_0, \theta)$. □

Lemma 18:

$$E_0(\rho, Q_n, s_0, \theta) \geq \frac{1}{n} \rho \mathcal{I}(Q_n; P_{Y^n \,\|\, X^n, s_0, \theta}) - \frac{1}{2n} \rho^2 (\log(e|\mathcal{Y}|))^2. \tag{95}$$

Proof: The lemma follows from [4, Lemma 2]), which holds for a channel P and input distribution Q satisfying $\sum_{x^n} Q(x^n \,\|\, z^{n-1}) = 1$ and

$$\sum_{x^n, y^n} Q(x^n \,\|\, z^{n-1}) P(y^n \,\|\, x^n) = 1. \qquad \square$$

We now follow the technique in [4] by using Lemmas 17 and 18 to bound the error probability independent of both s_0 and θ. For a given rate $R < C$, let $\epsilon = (C - R)/2$ and pick m in such a way that $\hat{C}_m \geq R + \epsilon$. Then

$$\max_{Q_{X^m \,\|\, Z^{m-1}}} \inf_{s_0, \theta} \frac{1}{m} \mathcal{I}(Q_{X^m \,\|\, Z^{m-1}}; P_{Y^m \,\|\, X^m, s_0, \theta}) - \frac{\log |\mathcal{S}|}{m} \geq R + \epsilon. \tag{96}$$

Let $Q_m^* \in \mathcal{P}(\mathcal{X}^m \,\|\, \mathcal{Z}^{m-1})$ be the input distribution that achieves the supremum in $\hat{C}_m$, i.e.,

$$\inf_{s_0, \theta} \frac{1}{m} \mathcal{I}(Q_m^*; P_{Y^m \,\|\, X^m, s_0, \theta}) - \frac{\log |\mathcal{S}|}{m} \geq R + \epsilon. \tag{97}$$

Next, we use Q_m^* to define a distribution $Q_{Nm} \in \mathcal{P}(\mathcal{X}^{Nm} \,\|\, \mathcal{Z}^{Nm-1})$ for a sequence of length Nm, $N \geq 1$, as follows:

$$\begin{aligned} Q(x^{Nm} \,\|\, z^{Nm-1}) &\triangleq Q_m^* \left(x_1^m \,\|\, z_1^{m-1}\right) \times Q_m^* \left(x_{m+1}^{2m} \,\|\, z_{m+1}^{2m-1}\right) \times \cdots \\ &\quad \times Q_m^* \left(x_{(N-1)m+1}^{Nm} \,\|\, z_{(N-1)m+1}^{Nm-1}\right) \qquad (98) \\ &= \prod_{i=1}^{N} Q_m^* \left(x_{(i-1)m+1}^{im} \,\|\, z_{(i-1)m+1}^{im-1}\right). \qquad (99) \end{aligned}$$

For this new input distribution and sequence of length Nm, we can bound the error exponent

$$F^{Nm}(\rho, Q_{Nm}, \theta) - \rho R \tag{100}$$

as follows:

$$\begin{aligned} &\overset{(a)}{\geq} F^m(\rho, Q_m^*, \theta) - \rho R \qquad (101) \\ &= \min_{s_0} E_0(\rho, Q_m^*, s_0, \theta) - \rho \left(R + \frac{\log |\mathcal{S}|}{m} \right) \qquad (102) \\ &\overset{(b)}{\geq} \min_{s_0} \frac{1}{m} \rho I(Q_m^*; P_{Y^m \,\|\, X^m, s_0, \theta}) - \frac{1}{2m} \rho^2 (\log(e|\mathcal{Y}^m|))^2 - \rho \left(R + \frac{\log |\mathcal{S}|}{m} \right) \qquad (103) \\ &\geq \rho \left(\inf_{s_0, \theta} \frac{1}{m} I(Q_m^*; P_{Y^m \,\|\, X^m, s_0, \theta}) - R - \frac{\log |\mathcal{S}|}{m} \right) - \frac{1}{2m} \rho^2 (\log(e|\mathcal{Y}^m|))^2 \qquad (104) \\ &\overset{(c)}{\geq} \rho \epsilon - \frac{1}{2m} \rho^2 (\log(e|\mathcal{Y}^m|))^2 \qquad (105) \end{aligned}$$

where (a) is due to Lemma 17, (b) follows from Lemma 18, and (c) follows from (97). As in [4], we can maximize the lower bound on the error exponent by setting $\rho = \min(1, m\epsilon/(\log(e|\mathcal{Y}^m|))^2)$. With this choice of ρ we have

$$F^{Nm}(\rho, Q_{Nm}, \theta) - \rho R \geq \begin{cases} m\epsilon^2/(2\log(e|\mathcal{Y}|^m)^2), & \epsilon < \frac{1}{m}(\log(e|\mathcal{Y}|^m))^2 \\ \epsilon - \frac{1}{2m}(\log(e|\mathcal{Y}|^m))^2, & \text{otherwise.} \end{cases} \tag{106}$$

Theorem 2 follows by combining (106) with the result in Corollary 16 (for block length Nm).

Appendix C
Proof of Lemma 5

To prove the lemma, we must first establish two equalities relating the channel causal conditioning distribution $P(y^n \,\|\, x^n, s_0, \theta)$ to the channel probability law $P(y_i, s_i \,|\, x_i, s_{i-1}, \theta)$. The following set of equalities hold.

$$\begin{aligned} P(y^n, x^n \,|\, s_0, \theta) &= \sum_{s^n \in \mathcal{S}^n} P(y^n, x^n, s^n \,|\, s_0, \theta) \qquad (107) \\ &\overset{(a)}{=} \sum_{s^n \in \mathcal{S}^n} P(x^n \,\|\, y^{n-1}, s^{n-1}, s_0, \theta) P(y^n, s^n \,\|\, x^n, s_0, \theta) \qquad (108) \end{aligned}$$

$$\stackrel{(b)}{=} \sum_{s^n \in \mathcal{S}^n} P(x^n \,\|\, y^{n-1}, s_0, \theta) P(y^n, s^n \,\|\, x^n, s_0, \theta) \quad (109)$$

$$= P(x^n \,\|\, y^{n-1}, s_0, \theta) \sum_{s^n \in \mathcal{S}^n} P(y^n, s^n \,\|\, x^n, s_0, \theta) \quad (110)$$

where (a) is due to [17, Lemma 2] and (b) follows from our assumption that the input distribution x^n does not depend on the state sequence s^{n-1}. By the chain rule for causal conditioning [17, Lemma 1], (110) implies that

$$P(y^n \,\|\, x^n, s_0, \theta) = \sum_{s^n \in \mathcal{S}^n} P(y^n, s^n \,\|\, x^n, s_0, \theta). \quad (111)$$

Also

$$P(y^n, s^n \,\|\, x^n, s_0, \theta) = \prod_{i=1}^{n} P(y_i, s_i \,|\, x^{i-1}, y^{i-1}, s^{i-1}, \theta) \quad (112)$$

$$\stackrel{(c)}{=} \prod_{i=1}^{n} P(y_i, s_i \,|\, x_i, s_{i-1}, \theta) \quad (113)$$

where (c) follows from the definition of the compound finite-state channel. Having established (111) and (113), Lemma 5 follows immediately from [12, Lemma 12], where the conditional probability $P(y_i, s_i \,|\, x_i, s_{i-1}, \theta)$ is quantized and the quantization cells are represented by channels $\{\theta_1^{(n)}, \ldots, \theta_{K(n)}^{(n)}\}$. The proof of our result differs only in that the upper bound on the error exponents in the family is given by $\mu = 1 + \log |\mathcal{Y}|$.

Appendix D
Proof of Lemmas 7, 8, and 9

The proof of Lemma 7 is based on an identity that is given by Kim in [18, eq. (9)]

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right). \quad (114)$$

Proof of Lemma 7: Using Kim's identity we have

$$\begin{aligned} I(X^n \to Y^n) &= \sum_{i=1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &= \sum_{i=1}^{k} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &\quad + \sum_{i=k+1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &\geq \sum_{i=1}^{k} I\left(X_i; Y_i^k \,|\, X^{i-1}, Y^{i-1}\right) \\ &\quad + \sum_{i=k+1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &= I(X^k \to Y^k) \\ &\quad + \sum_{i=k+1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right). \end{aligned} \quad (115)$$

Now we bound the sum in the last equality

$$\begin{aligned} &\sum_{i=k+1}^{n} I\left(X_i; Y_i^n \,|\, X^{i-1}, Y^{i-1}\right) \\ &\quad = \sum_{i=k+1}^{n} H\left(X_i \,|\, X^{i-1}, Y^{i-1}\right) \\ &\qquad - H\left(X_i \,|\, X^{i-1}, Y^{i-1}, Y_i^n\right) \\ &\quad \stackrel{(a)}{=} \sum_{i=k+1}^{n} H\left(X_i \,|\, X_{k+1}^{i-1}, Y_{k+1}^{i-1}\right) \\ &\qquad - H\left(X_i \,|\, X^{i-1}, Y^{i-1}, Y_i^n\right) \\ &\quad \geq \sum_{i=k+1}^{n} H\left(X_i \,|\, X_{k+1}^{i-1}, Y_{k+1}^{i-1}\right) \\ &\qquad - H\left(X_i \,|\, X_{k+1}^{i-1}, Y_{k+1}^{i-1}, Y_i^n\right) \\ &\quad = I\left(X_{k+1}^n \to Y_{k+1}^n\right) \end{aligned} \quad (116)$$

where (a) follows from the assumption that

$$Q(x^n \,\|\, z^{n-1}) = Q(x_1^k \,\|\, z_1^{k-1}) Q(x_{k+1}^n \,\|\, z_{k+1}^{n-1}). \qquad \square$$

Proof of Lemma 8: The proof consists of two parts. In the first part, we show that $nC_n^{\text{Markovian}}$ is sup-additive and therefore $\lim_{n\to\infty} C_n^{\text{Markovian}} = \sup_n C_n^{\text{Markovian}}$. In the second part, we prove the capacity of the family of stationary and uniformly ergodic Markovian channels by showing that

$$\lim_{n\to\infty} C_n = \lim_{n\to\infty} C_n^{\text{Markovian}} \quad (117)$$

where C_n is defined in (11).

First part: We show that the sequence $C_n^{\text{Markovian}}$ is sup-additive and therefore the limit exists. Let integers k and m be such that $k + m = n$ and denote input distributions $Q(x^n \,\|\, z^{n-1}), Q(x_1^k \,\|\, z_1^{k-1})$, and $Q(x_{k+1}^n \,\|\, z_{k+1}^{n-1})$ in shortened forms as Q_n, Q_k, and Q_m. We have

$$\begin{aligned} nC_n^{\text{Markovian}} &= \max_{Q_n} \inf_{\theta} I(X^n \to Y^n \,|\, \theta) \\ &\stackrel{(a)}{\geq} \max_{Q_k Q_m} \inf_{\theta} I(X^n \to Y^n \,|\, \theta) \\ &\stackrel{(b)}{\geq} \max_{Q_k Q_m} \inf_{\theta} \left[I(X^k \to Y^k \,|\, \theta) \right. \\ &\qquad \left. + I\left(X_{k+1}^n \to Y_{k+1}^n \,|\, \theta\right) \right] \\ &\geq \max_{Q_k Q_m} \left[\inf_{\theta} I(X^k \to Y^k \,|\, \theta) \right. \\ &\qquad \left. + \inf_{\theta} I\left(X_{k+1}^n \to Y_{k+1}^n \,|\, \theta\right) \right] \\ &= \max_{Q_k} \inf_{\theta} I(X^k \to Y^k \,|\, \theta) \\ &\qquad + \max_{Q_m} \inf_{\theta} I\left(X_{k+1}^n \to Y_{k+1}^n \,|\, \theta\right) \\ &\stackrel{(c)}{=} \max_{Q_k} \inf_{\theta} I(X^k \to Y^k \,|\, \theta) \\ &\qquad + \max_{Q(x^m \,\|\, z^{m-1})} \inf_{\theta} I\left(X^m \to Y^m \,|\, \theta\right) \\ &= kC_k^{\text{Markovian}} + mC_m^{\text{Markovian}} \end{aligned} \quad (118)$$

where (a) follows by restricting the maximization to causal conditioning probabilities of the product form

$$Q(x^n \,\|\, z^{n-1}) = Q(x_1^k \,\|\, z_1^{k-1}) Q(x_{k+1}^n \,\|\, z_{k+1}^{n-1}),$$

(b) follows from Lemma 7, and (c) follows from stationarity of the channel.

Second Part: We show that

$$\lim_{n\to\infty} C_n = \lim_{n\to\infty} C_n^{\text{Markovian}}.$$

Due to Lemma 4 in [17]

$$|I(X^n \to Y^n \,|\, \theta) - I(X^n \to Y^n \,|\, S_0, \theta)| \le \log |\mathcal{S}|$$

therefore, it is enough to prove that

$$\lim_{n\to\infty} \frac{1}{n} \left[\max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta} I(X^n \to Y^n \,|\, S_0, \theta) - \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta, s_0} I(X^n \to Y^n, \,|\, s_0, \theta) \right] = 0. \quad (119)$$

The difference in (119) is always positive, hence, it is enough to upper-bound it by an expression that goes to zero as $n \to \infty$. Again by Lemma 4 in [17], we can bound the second term in (119)

$$\begin{aligned} &\max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta, s_0} I(X^n \to Y^n \,|\, s_0, \theta) \\ &\quad \ge \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta, s_0} I(X^n \to Y^n \,|\, S_k, s_0, \theta) - \log |\mathcal{S}| \\ &\quad \overset{(a)}{\ge} \max_{Q_{X_k^n \,\|\, Z_k^{n-1}}} \inf_{\theta, s_0} I\left(X_k^n \to Y_k^n \,|\, S_k, s_0, \theta\right) - \log |\mathcal{S}| \\ &\quad \overset{(b)}{=} \max_{Q_{X^{n-k} \,\|\, Z^{n-k-1}}} \inf_{\theta, s_{-k}} I\left(X^{n-k} \to Y^{n-k} \,|\, S_0, s_{-k}, \theta\right) \\ &\qquad - \log |\mathcal{S}| \end{aligned} \quad (120)$$

where (a) holds for every $k > 1$ and is due to Lemma 7 and (b) holds by the stationarity of the channel. Hence, (120) implies that we can bound the difference

$$\begin{aligned} &\max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta} I(X^n \to Y^n \,|\, S_0, \theta) \\ &\qquad - \max_{Q_{X^n \,\|\, Z^{n-1}}} \inf_{\theta, s_0} I(X^n \to Y^n, \,|\, s_0, \theta) \\ &\quad \overset{(a)}{\le} \left(k \log |\mathcal{Y}| + \max_{Q_{X^{n-k} \,\|\, Z^{n-k-1}}} \inf_{\theta} I(X^{n-k} \to Y^{n-k} \,|\, S_0, \theta) \right) \\ &\qquad - \left(\max_{Q_{X^{n-k} \,\|\, Z^{n-k-1}}} \inf_{\theta, s_{-k}} I\left(X_1^{n-k} \to Y^{n-k}, \,|\, S_0, s_{-k}, \theta\right) - \log |\mathcal{S}| \right) \\ &\quad \overset{(b)}{\le} k \log |\mathcal{Y}| + \epsilon(n-k) \log |\mathcal{Y}| + \log |\mathcal{S}|. \end{aligned} \quad (121)$$

Inequality (a) is due to the fact that $I(X^n \to Y^n) \le k \log |\mathcal{Y}| + I(X^{n-k} \to Y^{n-k})$ and due to (120). Inequality (b) holds since for a uniformly ergodic family of channels, $|P(s_0 \,|\, s_{-k}, \theta) - P(s_0 \,|\, \theta)| \le \epsilon$ for all $s_0 \in \mathcal{S}$ implies that for any input distribution $Q_{X^{n-k} \,\|\, Z^{n-k-1}}$ and any channel θ

$$\left| I(X^{n-k} \to Y^{n-k} \,|\, \theta, S_0) - I\left(X_1^{n-k} \to Y^{n-k}, \,|\, S_0, s_{-k}, \theta\right) \right| \le \epsilon(n-k) \log |\mathcal{Y}|.$$

After dividing (121) by n, and since ϵ can be arbitrarily small and k is fixed for a given ϵ, then (119) holds. □

Proof of Lemma 9: From the assumption of the lemma we have

$$\sum_{x^n, y^n} Q(x^n) P(y^n \,\|\, x^n) \log \frac{Q(x^n) P(y^n \,\|\, x^n)}{P(y^n) Q(x^n)} = 0. \quad (122)$$

By assuming a uniform input distribution, $Q(x^n) = \frac{1}{|\mathcal{X}|^n}$ and by using the fact that if the Kullback–Leibler divergence $D(p \,\|\, q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ is zero, then $p(x) = q(x)$ for all $x \in \mathcal{X}$, we get that (122) implies that $P(y^n \,\|\, x^n) = P(y^n)$ for all $x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n$. It follows that

$$\max_{Q_{X^n \,\|\, Y^{n-1}}} I(X^n \to Y^n) = \max_{Q_{X^n \,\|\, Y^{n-1}}} E\left[\log \frac{P(Y^n \,\|\, X^n)}{P(Y^n)} \right] \quad (123)$$

$$= \max_{Q_{X^n \,\|\, Y^{n-1}}} E[0] = 0. \quad (124)$$

□

Appendix E
Proof of Lemmas 12 and 13

Proof of Lemma 12: The proof is based on the fact that $\mathcal{I}(Q_X, P_{Y \,|\, X})$ is uniformly continuous in $P_{Y \,|\, X}$, namely, for any Q_X

$$|\mathcal{I}(Q_X, P_{Y \,|\, X, \theta_1}) - \mathcal{I}(Q_X, P_{Y \,|\, X, \theta_2})| \le \tau(\Delta) \quad (125)$$

where $\tau(\Delta) \to 0$ as $\Delta \to 0$. (The uniform continuity of mutual information is a straightforward result of the uniform continuity of entropy [25, Theorem 2.7].) We have

$$\begin{aligned} &\left| \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_1}\right) - \mathcal{I}\left(Q_X^1, P_{Y \,|\, X, \theta_1}\right) \right| \\ &= \left| \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_1}\right) - \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) \right. \\ &\quad \left. + \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) - \mathcal{I}\left(Q_X^1, P_{Y \,|\, X, \theta_1}\right) \right| \\ &\le \tau(\Delta) + \left| \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) - \mathcal{I}\left(Q_X^1, P_{Y \,|\, X, \theta_1}\right) \right| \end{aligned} \quad (126)$$

where the last inequality is due to (125). We conclude the proof by bounding the last term in (126) by $\tau(\Delta)$, which implies that if we let $\eta(\Delta) = 2\tau(\Delta)$ then (77) holds.

$$\begin{aligned} &\mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) - \mathcal{I}\left(Q_X^1, P_{Y \,|\, X, \theta_1}\right) \\ &\quad \le \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) - \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_1}\right) \\ &\quad \le \tau(\Delta). \end{aligned} \quad (127)$$

Similarly, we have

$$\mathcal{I}(Q_X^1, P_{Y \,|\, X, \theta_1}) - \mathcal{I}(Q_X^2, P_{Y \,|\, X, \theta_2}) \le \tau(\Delta)$$

and therefore

$$\left| \mathcal{I}\left(Q_X^2, P_{Y \,|\, X, \theta_2}\right) - \mathcal{I}\left(Q_X^1, P_{Y \,|\, X, \theta_1}\right) \right| \le \tau(\Delta). \quad (128)$$

□

APPENDIX F
PROOF OF LEMMA 13

The channel $P_{Y\,|\,X,\hat{\theta}}$ is chosen by finding the conditional empirical distribution induced by an input sequence consisting of $\frac{M}{|\mathcal{X}|}$ copies of each symbol of the alphabet $\mathcal{X}$. We estimate the conditional distribution $P_{Y\,|\,a}$ separately for each $a \in \mathcal{X}$. We insert $x = a$ for $m = \frac{M}{|\mathcal{X}|}$ uses of the channel and we estimate the channel distribution when the input is $x = a$ as the type of the output which is denoted as $P_{Y^m\,|\,a}$. From Sanov's theorem (cf. [26, Theorem 12.4.1]) we have that the probability that type $P_{Y^m\,|\,a}$ will be at L_1-distance larger than $\epsilon_1 = \frac{\epsilon}{|\mathcal{X}|}$ from $P_{Y\,|\,a}$ is upper-bounded by

$$\Pr\{\,\|\,P_{Y^m\,|\,a} - P_{Y\,|\,a}\,\|_1 \geq \epsilon_1\} \leq (m+1)^{|\mathcal{Y}|} \times \exp\left(-m \min_{P_Y:\,\|\,P_Y - P_{Y\,|\,a})\,\|_1 \geq \epsilon_1} D(P_Y\,\|\,P_{Y\,|\,a})\right) \quad (129)$$

where $D(P_Y\,\|\,P_{Y\,|\,a}) = \sum_{y\in\mathcal{Y}} P_Y(y) \log \frac{P_Y(y)}{P_{Y\,|\,a}(y\,|\,a)}$ denotes the divergence between the two distributions. Using Pinsker's inequality [26, Lemma 12.6.1] we have that

$$\min_{P_Y:\,\|\,P_Y - P_{Y\,|\,a})\,\|_1 \geq \epsilon_1} D(P_Y\,\|\,P_{Y\,|\,a}) \geq \frac{\epsilon_1^2}{2} \quad (130)$$

and therefore

$$\Pr\{\,\|\,P_{Y^m} - P_{Y\,|\,a}\,\|_1 \geq \epsilon_1\} \leq (m+1)^{|\mathcal{Y}|} \exp\left(-m\frac{\epsilon_1^2}{2}\right). \quad (131)$$

The term $(m+1)^{|\mathcal{Y}|}\exp(-m\frac{\epsilon_1^2}{2})$ goes to zero as m goes to infinity for $\epsilon_1 > 0$ and, therefore, for any $\frac{\delta}{|\mathcal{X}|} > 0$ we can find an m such that $(m+1)^{|\mathcal{Y}|}\exp(-m\frac{\epsilon_1^2}{2}) \leq \frac{\delta}{|\mathcal{X}|}$. Finally, we have

$$\Pr\{\Delta > \epsilon\} \leq \Pr\left\{\bigcup_{a\in\mathcal{X}} \|\,P_{Y\,|\,a,\hat{\theta}} - P_{Y\,|\,a}\,\|_1 > \frac{\epsilon}{|\mathcal{X}|}\right\} \leq |\mathcal{X}|\frac{\delta}{|\mathcal{X}|} \quad (132)$$

where the inequality on the right is due to the union bound. □

ACKNOWLEDGMENT

The authors would like to thank Anthony Ephremides, Tsachy Weissman, Prakash Narayan, and Andrea Goldsmith for their support of this work.

REFERENCES

[1] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, no. 4, pp. 1229–1241, 1959.
[2] J. Wolfowitz, "Simultaneous channels," *Arch. Rational Mech. and Anal.*, vol. 4, no. 1, pp. 371–386, 1959.
[3] J. Wolfowitz, *Coding Theorems of Information Theory*, 2nd ed. New York: Springer-Verlag, 1964.
[4] A. Lapidoth and İ. E. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 973–983, May 1998.
[5] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.
[6] H. Weingarten, G. Kramer, and S. Shamai (Shitz), "On the compound MIMO broadcast channel," in *Proc. Information Theory and Applications (ITA 2007)*, San Diego, CA, Jan. /Feb. 2007.
[7] A. Raja, V. M. Prabhakaran, and P. Viswanath, "The two user Gaussian compound interference channel," *IEEE Trans. Inf. Theory*, submitted for publication.
[8] H. Weingarten, T. Liu, S. Shamai (Shitz), Y. Steinberg, and P. Viswanath, "Capacity region of the degraded MIMO compound broadcast channel," *IEEE Trans. Inf. Theory*, to be published.
[9] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[10] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.
[11] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert Elliot channel," *IEEE Trans. Inf. Theory*, vol. 35, no. 5, pp. 1277–1290, Nov. 1989.
[12] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1726–1745, Sep. 1998.
[13] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory and Its Application (ISITA-90)*, Hawaii, Nov. 1990, pp. 303–305.
[14] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[15] S. Tatikonda, "Control Under Communication Constraints," Ph.D. disertation, MIT, , Cambridge, MA, 2000.
[16] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
[17] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.
[18] Y. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1488–1499, Apr. 2008.
[19] A. Tchamkerten and İ. E. Telatar, "Variable length coding over an unknown channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.
[20] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 546–549, Mar. 1995.
[21] M. Pinsker, Talk Delivered at the Soviet Information Theory Meeting (no abstract published), 1969.
[22] P. Ebert, "The capacity of the Gaussian channel with feedback," *Bell Syst. Tech. J.*, pp. 1705–1712, 1970.
[23] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
[24] A. Tchamkerten, private communication 2007.
[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[26] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

Brooke Shrader (S'01–M'08) received the B.S. degree from Rice University, Houston, TX, in 2000, the M.S. degree from the Swedish Royal Institute of Technology (KTH), Stockholm, in 2002, and the Ph.D. degree from the University of Maryland, College Park, in 2008, all in electrical engineering.

She is currently a Member of Technical Staff at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory, Lexington.

Haim Permuter (S'08–M'09) received the B.Sc. (*summa cum laude*) and M.Sc. (*summa cum laude*) degrees in electrical and computer engineering from the Ben-Gurion University, Be'er-Sheva, Israel, in 1997 and 2003, respectively, and the Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2008.

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

Dr. Permuter is a recipient of the Fulbright Fellowship and the Stanford Graduate Fellowship (SGF).