Finite State Channels With Time-Invariant Deterministic Feedback

Haim Henry Permuter, *Student Member, IEEE*, Tsachy Weissman, *Senior Member, IEEE*, and Andrea J. Goldsmith, *Fellow, IEEE*

***Abstract*—We consider capacity of discrete-time channels with feedback for the general case where the feedback is a time-invariant deterministic function of the output samples. Under the assumption that the channel states take values in a finite alphabet, we find a sequence of achievable rates and a sequence of upper bounds on the capacity. The achievable rates and the upper bounds are computable for any N, and the limits of the sequences exist. We show that when the probability of the initial state is positive for all the channel states, then the capacity is the limit of the achievable-rate sequence. We further show that when the channel is stationary, indecomposable, and has no intersymbol interference (ISI), its capacity is given by the limit of the maximum of the (normalized) directed information between the input X^N and the output Y^N, i.e.,**

$$C = \lim_{N\to\infty} \frac{1}{N} \max I(X^N \to Y^N)$$

where the maximization is taken over the causal conditioning probability $Q(x^N \| z^{N-1})$ defined in this paper. The main idea for obtaining the results is to add causality into Gallager's results on finite state channels. The capacity results are used to show that the source–channel separation theorem holds for time-invariant determinist feedback, and if the state of the channel is known both at the encoder and the decoder, then feedback does not increase capacity.

***Index Terms*—Causal conditioning, code-tree, directed information, feedback capacity, maximum likelihood, random coding, source–channel coding separation.**

I. Introduction

SHANNON showed in [2] that feedback does not increase the capacity of a memoryless channel, and therefore the capacity of a memoryless channel with feedback is given by maximizing the mutual information between the input X, and the output Y, i.e., $C = \max_{P(X)} I(X;Y)$. In the case where there is no feedback, and the channel is an indecomposable finite-state channel (FSC), the capacity was shown by Gallager [1] and by Blackwell, Breiman and Thomasian [3] to be

$$C_{NF} = \lim_{N\to\infty} \frac{1}{N} \max_{P(x^N)} I(X^N;Y^N). \tag{1}$$

A simple example can show that mutual information is not the right measure for characterizing feedback capacity of FSCs. Consider the binary symmetric channel (BSC) with probability of error $\frac{1}{2}$ and an input to the channel that is the output with one epoch-delay, i.e., $X_i = Y_{i-1}$. It is easy to see that the mutual information between the input and the output to the channel, $\frac{1}{N} I(X^N;Y^N)$ tends to one as $N \to \infty$, despite the fact that the capacity of this memoryless channel is obviously zero.

In 1989, the directed information appeared in an implicit way in a paper by Cover and Pombra [4]. In an intermediate step [4, eq. 52], they showed that the directed information can be used to characterize the capacity of additive Gaussian noise channels with feedback. However, the term directed information was coined only a year later by Massey in a key paper [5].

In [5], Massey introduced directed information, denoted by $I(X^N \to Y^N)$, which he attributes to Marko [6]. Directed information, $I(X^N \to Y^N)$, is defined as

$$I(X^N \to Y^N) \triangleq \sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1}). \tag{2}$$

Massey showed that directed information is the same as mutual information $I(X^N;Y^N)$ in the absence of feedback and that it gives a better upper bound on the information that the channel output Y^N gives about the source sequence in the presence of feedback.

In his Ph.D. dissertation [7] and in [8], Tatikonda generalized the capacity formula of Verdú and Han [9] that deals with arbitrary single-user channels without feedback to the case of arbitrary single-user channels with feedback by using the directed information formula. Tatikonda also introduced the dynamic programming framework for computing the directed information for Markov channels, and derived the directed data processing inequality. Recently, the directed information formula was used by Yang, Kavčić, and Tatikonda [10] and by Chen and Berger [11] to compute the feedback capacity for some special FSCs (In [10], it was assumed that the state channel is a deterministic function of the previous state and input, and in [11] it was assumed that state is a deterministic function of the output).

Directed information also appeared recently in a rate distortion problem. Following the competitive prediction of Weissman and Merhav [12], Pradhan and Venkataramanan [13], [14] formulated a problem of source coding with feed-forward

Manuscript received August 17, 2006; revised March 21, 2008. Current version published February 04, 2009. This work was supported by the National Science Foundation (NSF) under Grant CCR-0311633, by NSF CAREER grant, the U.S. Army under MURI award W911NF-05-1-0246, and by the ONR under award N00014-05-1-0168. The material in this paper was presented in part at IEEE International Symposium on Information (ISIT), Seattle, WA, July 2006.

H. Permuter is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is also with the Department of Electrical and Computer Engineering, Ben Gurion University of the Negev, Beer-Sheva, Israel (e-mail: haim1@stanford.edu).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is also with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa, Israel (e-mail: tsachy@stanford.edu).

A. J. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. (e-mail: andrea@stanford.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2008.2009849

0018-9448/$25.00 © 2009 IEEE

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

and showed that directed information can be used to characterize the rate distortion function for the case of feed-forward. Another source coding context in which directed information has arisen is the recent work by Zamir *et al.* [15], which gives a linear prediction representation for the rate distortion function of a stationary Gaussian source.

The main contribution of this work is extending the achievability proof and the converse proof, given by Gallager in [1] for the case of an FSC without feedback, to the case of na FSC with time-invariant feedback. The extension is done by using the causal conditioning distribution that was introduced by Massey [5] and Kramer [16], rather than the regular conditioning. We first establish properties of causal conditioning that are useful throughout the proofs. We also show that causal conditioning can be used to generate a random code for the feedback setting to represent the channel and to be a metric for the maximum-likelihood (ML) decoder. We then show how Gallager's capacity proofs can be extended to feedback channels by replacing regular conditioning and mutual information with causal conditioning and directed information, respectively. This replacement requires careful justification. Moreover, the extension requires significant work in some cases because of difficulties that arise. For instance, not every property that holds for regular conditioning also holds for causal conditioning, as will be shown in Section IV. Furthermore, feedback introduces dependencies between the input, output, and the state of the channel that do not exist in the absence of feedback, and it cancels the one-to-one mapping between the messages m and the input x^N that exists in the absence of feedback. In most of the theorems and lemmas, the difficulties above are solved by appropriate modifications of Gallager's proofs, except for [1, Theorem 4.6.4], where a modification of the theorem itself is needed.

Time-invariant feedback includes the cases of quantized feedback, delayed feedback, and even noisy feedback where the noise is known to the decoder. In addition, it allows a unified treatment of capacity analysis for two ubiquitous cases: channels without feedback and channels with perfect feedback. These two settings are special cases of time-invariant feedback: in the first case, the time-invariant function of the feedback is the null function and in the second case the time-invariant function of the feedback is the identity function. The capacity of some channels with channel state information at the receiver and transmitter was derived by Viswanthan [17] and by Caire and Shamai in [18]. Note that if the channel state information can be considered part of the channel output and fed beck to the transmitter, then this case is a special case of a channel with time-invariant feedback.

The paper is organized as follows. Section II defines the channel setting and the notation throughout the paper. Section III provides a concise summary of the main results of the paper. Section IV introduces several properties of causal conditioning and directed information that are later used throughout the proofs. Section V provides the proof of achievability of capacity of FSCs with time-invariant feedback. Section VI gives an upper bound on the capacity. Section VII gives the capacity of a stationary indecomposable FSC without intersymbol interference (ISI). Section VIII considers the case of FSCs with feedback and side information and shows that if the state is known both at the encoder and decoder, then feedback does not increase the capacity of the channel. Section IX shows that under some conditions on the source and the channel, the optimality of source–channel separation holds in the presence of time-invariant feedback. We conclude in Section X with a summary of this work and some related future directions.

II. Channel Models and Preliminaries

We use subscripts and superscripts to denote vectors in the following way: $x^i = (x_1 \dots x_i)$ and $x_i^j = (x_i \dots x_j)$ for $i \leq j$. For $i \leq 0$, x^i defines the null string as does x_i^j when $i > j$. Moreover, we use lower case to denote sample values (e.g., x) and upper case to denote random variables (e.g., X) and calligraphic letter to denote alphabets (e.g., $\mathcal{X}$). The cardinality of an alphabet $\mathcal{X}$ is denoted as $|\mathcal{X}|$. Probability mass functions are denoted by P or Q when the arguments specify the distribution, e.g., $P(x|y) = P(X = x|Y = y)$. We usually use the letter Q for describing channel-input distributions and P for all the other distributions. Throughout this paper, we consider only random variables from finite alphabets, and when we write $P_1(x|y) = P_2(x|y)$, we mean that the equality necessarily holds for y that satisfies $P(y) > 0$.

Here, we consider only FSCs. The FSCs are a class of channels rich enough to include channels with memory, e.g., channels with ISI. The input of the channel is denoted by $\{X_1, X_2, \dots\}$, and the output of the channel is denoted by $\{Y_1, Y_2, \dots\}$, both taking values in a finite alphabet $\mathcal{X}, \mathcal{Y}$. In addition, the channel states take values in a finite set of possible states $\mathcal{S}$. The channel is stationary and is characterized by a conditional probability assignment $P(y_i, s_i|x_i, s_{i-1})$ that satisfies

$$P(y_i, s_i|x^i, s^{i-1}, y^{i-1}) = P(y_i, s_i|x_i, s_{i-1}) \tag{3}$$

and by the initial state distribution $P(s_0)$. An FSC is said to be without ISI if the input sequence does not affect the evolution of the state sequence, i.e., $P(s_i|s_{i-1}, x_i) = P(s_i|s_{i-1})$.

We assume a communication setting that includes feedback as shown in Fig. 1. The transmitter (encoder) knows at time i the message m and the feedback samples z^{i-1}. The output of the encoder at time i is denoted by x_i, and it is a function of the message and the feedback. The channel is an FSC, and the output of the channel y_i enters the decoder (receiver). The feedback z_i is a known time-invariant deterministic function of the current output of the channel y_i. For example, z_i could equal y_i or a quantized version of it. The encoder receives the feedback sample with one unit delay. We are using the definition of achievable rate and capacity as given in the book by Cover and Thomas [19].

Definition 1: A rate R is said to be *achievable* if there exists a sequence of block codes $(N, \lceil 2^{NR} \rceil)$ such that the maximal probability of error

$$\max_{m \in [1, \dots, 2^{NR}]} \Pr(\hat{m} \neq m | \text{message } m \text{ was sent})$$

tends to zero as $N \to \infty$[19]. The *capacity* of an FSC is denoted as C and is the supremium of all achievable rates.

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

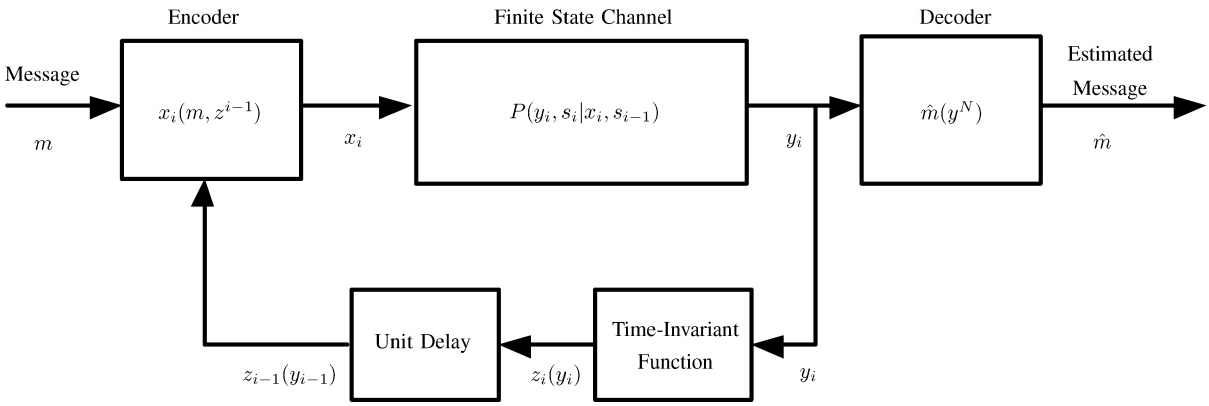

Fig. 1. Channel with feedback that is a time-invariant deterministic function of the output.

Throughout this paper we use the *Causal Conditioning* notation $(\cdot\|\cdot)$, which was introduced and employed by Kramer [16], [20] and by Massey [21]:

$$P(y^N\|x^N) \triangleq \prod_{i=1}^{N} P(y_i|x^i, y^{i-1}). \tag{4}$$

In addition, we introduce the following notation:

$$P(x^N\|y^{N-1}) \triangleq \prod_{i=1}^{N} P(x_i|x^{i-1}, y^{i-1}). \tag{5}$$

The definition given in (5) can be considered to be a particular case of the definition given in (4), where x_0 is set to a dummy zero. Since, we define the causal conditioning $P(y^N\|x^N)$ as a product of $P(y_i|y^{i-1}, x^i)$, then whenever we use $P(y^N\|x^N)$, we implicitly assume that there exists a set of conditional distribution $\{P(y^i|y^{i-1}, x^i)\}_{i=1}^{N}$ that satisfies the equality in (4). We can call $P(y^N\|x^N)$ and $P(x^N\|y^{N-1})$ a causal conditioning *distribution* since they are nonnegative for all x^N, y^N and since they sum to one, i.e., $\sum_{y^N} P(y^N\|x^N) = 1$ and $\sum_{x^N} P(x^N\|y^{N-1}) = 1$. The directed information $I(X^N \to Y^N)$ is defined in (2), and it can be expressed in terms of causal conditioning distribution as

$$I(X^N \to Y^N) = \sum_{i=1}^{N} I(X^i; Y_i|Y^{i-1}) = \mathbf{E}\left[\log \frac{P(Y^N\|X^N)}{P(Y^N)}\right] \tag{6}$$

where $\mathbf{E}$ denotes expectation. The directed information between X^N and Y^N, conditioned on S, is denoted as $I(X^N \to Y^N|S)$ and is defined as

$$I(X^N \to Y^N|S) \triangleq \sum_{i=1}^{N} I(Y_i; X^i|Y^{i-1}, S). \tag{7}$$

III. Main Results

Here we present the main results of the paper. Let $\underline{C}_N$ and $\overline{C}_N$ denote

$$\underline{C}_N \triangleq \frac{1}{N} \max_{Q(x^N\|z^{N-1})} \min_{s_0} I(X^N \to Y^N|s_0) \tag{8}$$

$$\overline{C}_N \triangleq \frac{1}{N} \max_{Q(x^N\|z^{N-1})} \max_{s_0} I(X^N \to Y^N|s_0). \tag{9}$$

- **Achievable rate:** For any FSC with the feedback as in Fig. 1, any rate less then $\lim_{N\to\infty} \underline{C}_N$ is achievable, where the limit of $\underline{C}_N$ exists and equals $\sup_N[\underline{C}_N - \frac{\log|\mathcal{S}|}{N}]$. This implies the following lower bound on the capacity:

$$C \geq \underline{C}_N - \frac{\log|\mathcal{S}|}{N}, \qquad N = 1, 2, 3 \ldots. \tag{10}$$

- **Converse:** For any FSC with the feedback as in Fig. 1, any achievable rate must be less then $\lim_{N\to\infty} \overline{C}_N$, where the limit of $\overline{C}_N$ exists and equals $\inf_N[\underline{C}_N + \frac{\log|\mathcal{S}|}{N}]$. This implies the following upper bound on the capacity:

$$C \leq \overline{C}_N + \frac{\log|\mathcal{S}|}{N}, \qquad N = 1, 2, 3 \ldots. \tag{11}$$

- **Capacity:** For two cases, the following capacity results hold.
  (a) For an FSC where the probability of the initial state is positive for all $s_0 \in \mathcal{S}$, the capacity is shown to be

$$C = \lim_{N\to\infty} \underline{C}_N. \tag{12}$$

  (b) For a stationary indecomposable FSC without ISI, the capacity, is shown to be

$$C = \lim_{N\to\infty} \frac{1}{N} \max_{Q(x^N\|z^{N-1})} I(X^N \to Y^N). \tag{13}$$

Finally, using the achievable rate and the converse, we show that feedback does not increase the capacity of a connected FSC (every state can be reached from every other state with positive probability under some input distribution) when the state of the channel is known both at the encoder and the decoder. And using the directed data processing inequality [7, Lemma 4.8.1], we show in a straightforward manner that the source–channel coding separation is optimal for any stationary and ergodic source and for any channel with time-invariant deterministic feedback, where the capacity is given by (13).

IV. Properties of Causal Conditioning and Directed Information

In this section, we present some properties of the causal conditioning distribution and the directed information which are defined in Section II in (4)–(6). The properties hold for any discrete random variables (not necessarily those induced by a FSC) and are used throughout the paper. Some of the properties assumes that $P(x^N\|y^{N-1}) = P(x^N\|z^{N-1})$. This equality is justified

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

in Section V-B, (38), for the setting of deterministic feedback $z_i(y_i)$, but it, actually, holds for any kind of feedback that has a delay of at least one time epoch. The lemmas proven here also help us gain some intuition about the role of causal conditioning and directed information in the proofs.

The first property was given by Massey [5, eq. (3)] and shows that a joint distribution can be decomposed into a multiplication of two causal conditioning distribution.

Lemma 1: (Chain rule for causal conditioning.) For any random variables (X^N, Y^N)

$$P(x^N, y^N) = P(y^N||x^N)P(x^N||y^{N-1}) \tag{14}$$

and, consequently, if Z^N is a random vector that satisfies $P(x^N||y^{N-1}) = P(x^N||z^{N-1})$ then

$$P(x^N, y^N) = P(y^N||x^N)P(x^N||z^{N-1}). \tag{15}$$

Proof:

$$\begin{aligned} P(y^N, x^N) &= \prod_{i=1}^{N} P(y_i, x_i|x^{i-1}, y^{i-1}) \\ &= \prod_{i=1}^{N} P(y_i|x^i, y^{i-1})P(x_i|x^{i-1}, y^{i-1}) \\ &= P(y^N||x^N)P(x^N||y^{N-1}). \end{aligned} \tag{16}$$

□

Let us define

$$P(y^N||x^N, s) \triangleq \prod_{i=1}^{N} P(y_i|x^i, y^{i-1}, s). \tag{17}$$

Lemma 2: For any random variables (X^N, Y^N, Z^{N-1}, S_0) that satisfy $P(x^N||y^{N-1}, s_0) = P(x^N||z^{N-1})$

$$P(x^N, y^N|s_0) = P(y^N||x^N, s_0)P(x^N||z^{N-1}). \tag{18}$$

The proof of Lemma 2 is similar to that of Lemma 1 and therefore is omitted.

The fact that the sequence $\{P(x_i|x^{i-1}, z^{i-1})\}_{i=1}^N$ determines uniquely the term $P(x^N||z^{N-1})$ follows immediately from the definition of the later. The next lemma, shows that the opposite is also true, namely, that $P(x^N||z^{N-1})$ determines uniquely $\{P(x_i|x^{i-1}, z^{i-1})\}_{i=1}^N$. This implies that maximizing the directed information over $P(x^N||z^{N-1})$ is equivalent to maximizing it over the set of sequences $\{P(x_i|x^{i-1}, z^{i-1})\}_{i=1}^N$. This is analogous to the fact that maximization of mutual information over the set $P(x^N)$ is equivalent to the maximization over the set of sequences $\{P(x_i|x^{i-1})\}_{i=1}^N$.

Lemma 3: The causal conditioning distribution $P(x^N||z^{N-1})$ uniquely determines the value of $P(x_i|x^{i-1}, z^{i-1})$ for all $i \leq N$ and all the arguments (x^{i-1}, z^{i-1}), for which $P(x^{i-1}, z^{i-1}) > 0$.

Proof: First we note that if $P(x^{i-1}, z^{i-1}) > 0$, then according to Lemma 1, it also implies that $P(x^{i-1}||z^{i-2}) > 0$. In addition, we always have the equality

$$P(x^{N-1}||z^{N-2}) = \sum_{x_N} P(x^N||z^{N-1}); \tag{19}$$

hence, $P(x^{N-1}||z^{N-2})$ is uniquely determined from $P(x^N||z^{N-1})$. Furthermore, by induction it can be shown that the sequence $\{P(x^i||z^{i-1})\}_{i=1}^N$ is uniquely derived from $P(x^N||z^{N-1})$. Since $P(x^{i-1}||z^{i-2}) > 0$, we can use the equality

$$P(x_i|x^{i-1}, z^{i-1}) = \frac{P(x^i||z^{i-1})}{P(x^{i-1}||z^{i-2})} \tag{20}$$

to derive unique value of $P(x_i|x^{i-1}, z^{i-1})$. □

The next lemma has an important role in the proofs for the capacity of FSCs because it bounds the difference of directed information before and after conditioning on a state by a constant. The proof of the lemma is given in Appendix I.

Lemma 4: (Analogue to $|I(X;Y) - I(X;Y|S)| \leq H(S)$) Let X^N, Y^N be arbitrary random vectors and S a random variable taking values in an alphabet of size $|\mathcal{S}|$. Then

$$\left|I(X^N \to Y^N) - I(X^N \to Y^N|S)\right| \leq H(S) \leq \log|\mathcal{S}|. \tag{21}$$

In the following lemma, we use the notation of mutual information $I(X^N; Y^N)$ as $\mathcal{I}(Q(x^N); P(y^N|x^N))$ where the latter is functional of $Q(x^N)$ and $P(y^N|x^N)$, i.e.,

$$\begin{aligned} &\mathcal{I}(Q(x^N), P(y^N|x^N)) \\ &\triangleq \sum_{y^N}\sum_{x^N} Q(x^N)P(y^N|x^N)\log\frac{P(y^N|x^N)}{\sum_{x'^N} Q(x'^N)P(y^N|x'^N)}. \end{aligned} \tag{22}$$

At the end of the achievability proof we will see that the achievable rate is the same functional, $\mathcal{I}(Q;P)$, as in the case without feedback but with the probability mass function $Q(x^N)$ replaced by $Q(x^N||z^{N-1})$ and $P(y^N|x^N)$ replaced by $P(y^N||x^N)$. Lemma 5 shows that the replacement of regular conditioning with causal conditioning in the functional $\mathcal{I}(Q;P)$, yields the directed information.

Lemma 5: If $P(x^N||y^{N-1}) = Q(x^N||z^{N-1})$ then

$$\mathcal{I}(Q(x^N||z^{N-1}), P(y^N||x^N)) = I(X^N \to Y^N) \tag{23}$$

and, similarly, if $P(x^N||y^{N-1}, s_0) = Q(x^N||z^{N-1})$ then

$$\mathcal{I}(Q(x^N||z^{N-1}), P(y^N||x^N, s_0)) = I(X^N \to Y^N|s_0). \tag{24}$$

Proof:

$$\begin{aligned} &\mathcal{I}(Q(x^N||z^{N-1}), P(y^N||x^N) \\ &\quad \overset{(a)}{=} \sum_{y^N}\sum_{x^N} Q(x^N||z^{N-1}) \\ &\qquad \cdot P(y^N||x^N)\log\frac{P(y^N||x^N)}{\sum_{x^N} Q(x^N||z^{N-1})P(y^N||x^N)} \\ &\quad \overset{(b)}{=} \mathbf{E}\left[\log\frac{P(Y^N||X^N)}{P(Y^N)}\right] \\ &\quad = I(X^N \to Y^N) \end{aligned} \tag{25}$$

where equality (a) is due to the definition of the functional $\mathcal{I}(Q,P)$ which is given in (22), and equality (b) is due to Lemma 1. □

Throughout the proof of the coding theorem of FSC we use the causal conditioning distribution $P(y^N||x^N)$. The next lemma shows how $P(y^N||x^N)$ can be calculated from the

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

FSC definition. Recall, that an FSC is defined by the initial state distribution $P(s_0)$ and the conditional distribution $P(y_i, s_i|x_i, s_{i-1})$.

Lemma 6: (Causal conditioning for an FSC.) For an FSC with time-invariant feedback, as shown in Fig. 1, the causal conditioning distribution can be calculated as follows:

$$P(y^N||x^N, s_0) = \sum_{s^N} \prod_{i=1}^{N} P(y_i, s_i|x_i, s_{i-1}) \tag{26}$$

$$P(y^N||x^N) = \sum_{s_0^N} \left(\prod_{i=1}^{N} P(y_i, s_i|x_i, s_{i-1}) \right) P(s_0) \tag{27}$$

where s_0^N denotes the vector $(s_0, s_1, \ldots, s_N)$.

The lemma is proved in Appendix II. For the case that the channel is memoryless, i.e., $|\mathcal{S}| = 1$, we have that $P(y^N||x^N) = \prod_{i=1}^{N} P(y_i|x_i)$, and it coincides with Massey's definition of a memoryless channel [5]. Two additional properties that hold for FSCs with feedback, and are used in this paper, are given in Appendices VI and VII.

The following lemma is an extension of the conservation law of information given by Massey in [21].

Lemma 7: (Extended conservation law.) For any random variables (X^N, Y^N, Z^{N-1}) that satisfy $P(x^N||y^{N-1}) = P(x^N||z^{N-1})$

$$I(X^N; Y^N) = I(X^N \to Y^N) + I(\{0, Z^{N-1}\} \to X^N) \tag{28}$$

where $\{0, Z^{N-1}\}$ is a concatenation of dummy zero to the beginning of the sequence Z^{N-1}.

Proof:

$$\begin{aligned} I(X^N; Y^N) &\overset{(a)}{=} \mathbf{E}\left[\log \frac{P(Y^N, X^N)}{P(Y^N)P(X^N)}\right] \\ &\overset{(b)}{=} \mathbf{E}\left[\log \frac{P(Y^N||X^N)P(X^N||Z^{N-1})}{P(Y^N)P(X^N)}\right] \\ &= \mathbf{E}\left[\log \frac{P(Y^N||X^N)}{P(Y^N)}\right] + \mathbf{E}\left[\log \frac{P(X^N||Z^{N-1})}{P(X^N)}\right] \\ &\overset{(c)}{=} I(X^N \to Y^N) + I(\{0, Z^{N-1}\} \to X^N). \end{aligned} \tag{29}$$

Equality (a) is due to the definition of mutual information. Equality (b) is due to Lemma 1, and equality (c) is due to the definition of directed information. □

The lemma was proven by induction in [21] for the case where $z_i = y_i$, and here it is shown to hold also for a broader case, in which the feedback is a function of the output. This lemma is not used for the proof of achievability; however, it gives a nice intuition for the relation of directed information and mutual information in the setting of deterministic feedback. In particular, the lemma implies that the mutual information between the input and the output of the channel is equal to the sum of directed information in the forward link and the directed information in the backward link. In addition, it is straightforward to see that in the case of no feedback, i.e., when z_i is null, then $I(X^N; Y^N) = I(X^N \to Y^N)$.

A property that does not holds for causal conditioning [8]: One can see that every property that holds for $Q(x^N||z^{N-1})$, $P(y^N||x^N)$ also holds for $Q(x^M), P(y^M|x^M)$, since we can consider the case $N = 1$ and then replace (x, y) by (x^M, y^M). However, there are properties that hold for regular conditioning but do not hold for causal conditioning. Such a property was shown by Tatikonda in [8, p. 3213]; for any random variables Y^N, U_i, X^N, we have the identity

$$\sum_{u_i} P(\{y_1, y_2, \ldots, y_{i-1}, u_i, y_i, \ldots y_N\}|x^N) = P(y^N|x^N). \tag{30}$$

But, in general, the identity

$$\sum_{u_i} P(\{y_1, y_2, \ldots, y_{i-1}, u_i, y_i, \ldots y_N\}||x^N) = P(y^N||x^N) \tag{31}$$

does *not* hold.

V. Proof of Achievability

The proof of the achievable rate of a channel with feedback given here is based on extending the upper bound on the error of ML decoding derived by Gallager in [1, Ch. 4,5] for FSCs without feedback to the case of FSCs with feedback.

Before presenting the achievability proof, let us first present a short outline.

- *Encoding scheme.* We randomly generate an encoding scheme for blocks of length N by using the causal conditioning distribution $Q(x^N||z^{N-1})$.
- *Decoding.* We assume a ML decoder, and we denote the error probability when message m is sent and the initial state of the channel is s_0 as $P_{e,m}(s_0)$.
- Bounding the error probability. We show that for each $N > N(\epsilon)$ there exists a code for which we can bound the error probability for all messages $1 \le m \le \lfloor 2^{NR} \rfloor$ and all initial states s_0 by the following exponential:

$$P_{e,m}(s_0) \le 2^{-N[E_r(R)-\epsilon]}. \tag{32}$$

In addition, we show that if $R < \underline{C}$, when $\underline{C}$ is defined as

$$\underline{C} \triangleq \lim_{N\to\infty} \underline{C}_N \tag{33}$$

then $E_r(R)$ is strictly positive and, hence, by choosing $\epsilon < E_r(R)$, the probability of error diminishes exponentially for $N > N(\epsilon)$.

A. Existence of $\underline{C}$

The following theorem states that the limit of the sequence $\underline{C}_N$ exists.

Theorem 8: (Analogue to [1, Theorem 4.6.1].) For a finite-state channel with $|\mathcal{S}|$ states the limit in (33) exists and

$$\lim_{N\to\infty} \underline{C}_N = \sup_N \left[\underline{C}_N - \frac{\log|\mathcal{S}|}{N} \right]. \tag{34}$$

The basic idea of the proof, similar to the proof of [1, Theorem 4.6.1], is to show that the sequence $N(\underline{C}_N - \frac{\log|\mathcal{S}|}{N})$ is super-additive. A sequence a_N is super-additive, if for any positive integers n, N, where $N > n$, we have $a_N \ge a_n + a_{N-n}$. For such a sequence $\lim_{N\to\infty} \frac{a_N}{N}$ exists, and the limit equals to $\sup_N \frac{a_N}{N}$. The proof differs from the proof of [1, Theorem 4.6.1] since, for an input distribution of the form $Q(x^N) = Q(x^n)Q(x_{n+1}^N)$ we have that X^n is independent of X_{n+1}^N[1,

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

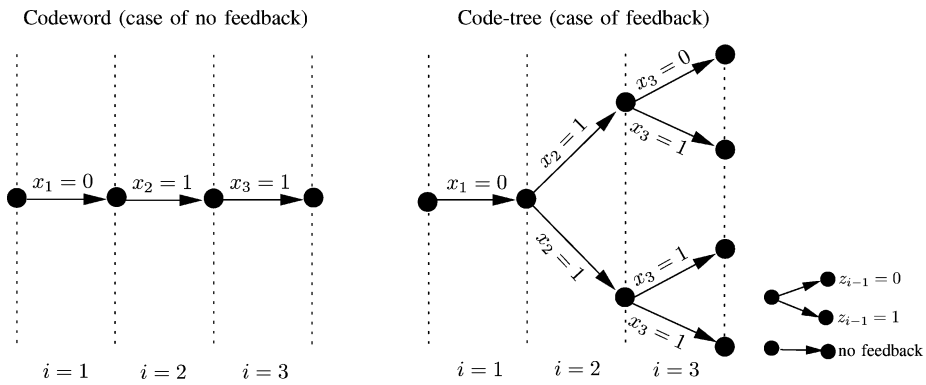

Fig. 2. Illustration of a coding scheme for a setting without feedback and for a setting with feedback. In the case of no feedback, each message is mapped to a codeword, and in the case of feedback each message is mapped to a code tree.

eq. (4A.17)]. In contrast, for an input distribution of the form $Q(x^N||z^{N-1}) = Q(x^n||z^{n-1})Q(x_{n+1}^N||z_{n+1}^{N-1})$ there could be a dependency of X^n and X_{n+1}^N.

Proof: Let n, l be two integers such that $N = n + l$. Let $Q_n(x^n||z^{n-1})$ and $Q_l(x^l||z^{l-1})$ be the probability assignments that achieve $\underline{C}_n$ and $\underline{C}_l$, respectively. Let us consider the probability assignment

$$Q(x^N||z^{N-1}) = Q_n(x^n||z^n)Q_l(x_{n+1}^N||z_{n+1}^{N-1}). \tag{35}$$

Since $Q(x^N||z^{N-1})$ is not necessary the input distribution that achieves $N\underline{C}_N$, we have

$$\begin{aligned}
&N\underline{C}_N \\
&\geq \min_{s_0} I(X^N \to Y^N|s_0) \\
&\overset{(a)}{\geq} \min_{s_0} \sum_{i=1}^{n} I(Y_i; X^i|Y^{i-1}, s_0) + \min_{s_0} \sum_{j=n+1}^{n+l} I(Y_j; X^j|Y^{j-1}, s_0) \\
&\overset{(b)}{\geq} n\underline{C}_n + \min_{s_0} \sum_{j=n+1}^{n+l} I(Y_j; X_{n+1}^j|Y^{j-1}, s_0) \\
&\overset{(c)}{\geq} n\underline{C}_n + \min_{s_0} \sum_{j=n+1}^{n+l} I(Y_j; X_{n+1}^j|Y^{j-1}, S_n, s_0) - \log|\mathcal{S}| \\
&\geq n\underline{C}_n + \min_{s_0} \min_{s_n} \sum_{j=n+1}^{n+l} I(Y_j; X_{n+1}^j|Y^{j-1}, s_n, s_0) - \log|\mathcal{S}| \\
&\overset{(d)}{=} n\underline{C}_n + \min_{s_n} \sum_{j=n+1}^{n+l} I(Y_j; X_{n+1}^j|Y_{n+1}^{j-1}, s_n) - \log|\mathcal{S}| \\
&\overset{(e)}{=} n\underline{C}_n + l\underline{C}_l - \log|\mathcal{S}|.
\end{aligned} \tag{36}$$

Equality (a) is due to the definition of the directed information and the fact that $\min_s[f(s) + g(s)] \geq \min_s f(s) + \min_s g(s)$. Inequality (b) holds because $\underline{C}_n$ is the first term and for the second term we have that $I(X;Y,Z) \geq I(X;Y)$ for any random variables (X, Y, Z). Inequality (c) is due to Lemma 4. Inequality (d) is due to the fact that given the input distribution from (35), we have the Markov chain $(S^{n-1}, X^n, Y^n) - S_n - (Y_{n+1}^j, X_{n+1}^j)$ for any $j \geq n+1$ (see Appendix VII for the proof). Equality (e) is due to the fact that the channel conditional distribution $P(y_i, s_i|s_{i-1}, x_i)$ is fixed over time. Rearranging the inequality we obtain

$$N\left[\underline{C}_N - \frac{\log|\mathcal{S}|}{N}\right] \geq n\left[\underline{C}_n - \frac{\log|\mathcal{S}|}{n}\right] + l\left[\underline{C}_l - \frac{\log|\mathcal{S}|}{l}\right]. \tag{37}$$

Finally, by using the convergence of a super additive sequence, the theorem is proved. □

B. Random Generation of Coding Scheme

In the case of no feedback, a coding block of length N is a mapping of each message m to a codeword of length N and is denoted by $x^N(m)$. In the case of feedback, a coding block is a vector function whose ith component is a function of m and the first $i-1$ components of the received feedback. The mapping of the message m and the feedback z^{i-1} to the input of the channel $x_i(m, z^{i-1})$ is called a *code tree*[22, Ch. 9], *strategy*[23] or *code functions*[7, Ch. 4.3]. Fig. 2 shows an example of a codeword of length $N = 3$ for the case of no feedback and a code tree of depth $N = 3$ for the case of binary feedback.

Randomly chosen coding scheme: We choose the ith channel input symbol $x_i(m, z^{i-1})$ of the codeword m by using a probability mass function (PMF) based on previous symbols of the code $x^{i-1}(m, z^{i-2})$ and previous feedback symbols z^{i-1}. The first channel input symbol of codeword m is chosen by the probability function $Q(x_1)$. The second symbol of codeword m is chosen for all possible feedback observations z_1 by the probability function $Q(x_2|x^1, z^1)$. The ith bit is chosen for all possible z^{i-1} by the probability function $Q(x_i|x^{i-1}, z^{i-1})$. This scheme of communication assumes that the probability assignment of x_i given x^{i-1} and z^{i-1} cannot depend on y^{i-1}, because it is unavailable. Therefore

$$P(x_i|x^{i-1}, z^{i-1}(y^{i-1}), y^{i-1}) = Q(x_i|x^{i-1}, z^{i-1}(y^{i-1})). \tag{38}$$

In this achievability proof, we choose $Q(x^N||z^{N-1})$, or equivalently, the sequence $\{Q(x_i|x^{i-1}, z^{i-1})\}_{i=1}^N$, that attains the maximum of $\max_{Q(x^N||z^{N-1})} \min_{s_0} I(X^N \to Y^N|s_0)$.

Encoding scheme: Each message m has a code tree. Therefore, for any feedback z^{N-1} and message m there is a unique input $x^N(m, z^{N-1})$ that was chosen randomly as described in the previous paragraph. After choosing the coding scheme, the

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

decoder is made aware of the code trees for all possible messages. In our coding scheme the input $x^N(m, z^{N-1})$ is always a function of the message m and the feedback, but in order to make the equations shorter we also use the abbreviated notation x^N for $x^N(m, z^{N-1})$.

Decoding scheme: The decoder is the ML decoder. Since the codewords depend on the feedback, two different messages can have the same codeword for two different outputs, therefore the regular ML $\arg\max_{x^N} P(y^N|x^N)$ cannot be used for decoding the message. Instead, the ML decoder should be $\arg\max_m P(y^N|m)$ where N is the block length. The following equation shows that finding the most likely message m can be done by maximizing the causal conditioning $P(y^N||x^N)$:

$$\arg\max_m P(y^N|m) = \arg\max_m P(y^N||x^N). \tag{39}$$

The equality in (39) is shown as follows:

$$\begin{aligned} P(y^N|m) &= \prod_i P(y_i|y^{i-1}, m) \\ &\overset{(a)}{=} \prod_i P(y_i|y^{i-1}, m, x^i(m, z^{i-1}(y^{i-1}))) \\ &\overset{(b)}{=} \prod_i P(y_i|y^{i-1}, x^i(m, z^{i-1}(y^{i-1}))) \\ &\overset{(c)}{=} P(y^N||x^N). \end{aligned} \tag{40}$$

Equality (a) holds because x^i is uniquely determined by the message m and the feedback z^{i-1}, and the feedback z^{i-1} is a deterministic function of y^{i-1}. Equality (b) holds because according to the channel structure, y_i does not depend on m given x^i. Equality (c) follows from the definition of causal conditioning given in (4).

C. ML Decoding Error Bound

The next theorem bounds the expected ML decoding error probability with respect to the random coding. The theorem was proved in [7, Propsition 4.4.1], for the case of perfect feedback, by introducing the idea of converting the channel with feedback into a new channel without feedback[1] and then applying [1, Theorem 4.6.1]. We present an additional proof in Appendix III, that follows Gallager's proof [1, Theorem 4.6.1], but the coding scheme includes time-invariant feedback and the ML decoder is the one presented in (40).

Let $P_{e,m}$ denote the probability of error using the ML decoder when message m is sent. When the source produces message m, there is a set of outputs denoted by ${Y_m}^c$ that cause an error in decoding the message m, i.e.,

$$P_{e,m} = \sum_{y^N \in {Y_m}^c} P(y^N|m). \tag{41}$$

Theorem 9: (Analogue to [1, Theorem 5.6.1]) Suppose that an arbitrary message $m, 1 \le m \le M$, enters the encoder with feedback and that ML decoding is employed. Then the average probability of decoding error over this ensemble of codes is bounded, for any choice of $\rho, 0 < \rho \le 1$, by

$$\begin{aligned} &\mathbf{E}(P_{e,m}) \\ &\le (M-1)^\rho \sum_{y^N} \left[\sum_{x^N} Q(x^N||z^{N-1}) P(y^N||x^N)^{\frac{1}{(1+\rho)}} \right]^{1+\rho} \end{aligned} \tag{42}$$

where the expectation is with respect to the randomness in the ensemble.

Let us define $P_{e,m}(s_0)$ to be the probability of error given that the initial state of the channel is s_0 and message m was sent. The following theorem, which is proved in Appendix IV, establishes the existence of a code such that $P_{e,m}(s_0)$ is small for all $1 \le m \le M$.

Theorem 10: (Analogue to [1, Theorem 5.9.1]) For an arbitrary finite-state channel with $|\mathcal{S}|$ states, for any positive integer N, and any positive R, there exists an (N, M) code for which for all messages $m, 1 \le m \le M = \lfloor 2^{NR} \rfloor$, all initial states s_0, and all $\rho, 0 \le \rho \le 1$, its probability of error is bounded as

$$P_{e,m}(s_0) \le 4|\mathcal{S}| 2^{\{-N[-\rho R + F_N(\rho)]\}} \tag{43}$$

where

$$\begin{aligned} &F_N(\rho) \\ &= -\frac{\rho \log |\mathcal{S}|}{N} + \max_{Q(x^N||z^{N-1})} \left[\min_{s_0} E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0) \right] \end{aligned} \tag{44}$$

$$\begin{aligned} &E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0) \\ &= -\frac{1}{N} \log \sum_{y^N} \left[\sum_{x^N} Q(x^N||z^{N-1}) P(y^N||x^N, s_0)^{\frac{1}{(1+\rho)}} \right]^{1+\rho}. \end{aligned} \tag{45}$$

The following theorem presents a few properties of the function $E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0)$ which is defined in (45), such as positivity of the function and its derivative, and convexity of the function with respect to ρ.

Theorem 11: (Analogue to [1, Theorem 5.6.3]) The term $E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0)$ has the following properties for $\rho \ge 0$:

$$E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0) \ge 0. \tag{46}$$

$$\begin{aligned} &\frac{1}{N} \mathcal{I}(Q(x^N||y^{N-1}), P(y^N||x^N, s_0)) \\ &\quad \ge \frac{\partial E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0)}{\partial \rho} > 0. \end{aligned} \tag{47}$$

$$\frac{\partial^2 E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0)}{\partial \rho^2} > 0. \tag{48}$$

Furthermore, equality holds in (46) when $\rho = 0$, and equality holds on the left side of (47) when $\rho = 0$.

The proof of the theorem is omitted because it is the same as the proof of Theorem 5.6.3 in [1]. The theorem in [1] states these same properties with $Q(x^N||z^{N-1})$ and $P(y^N||x^N)$ replaced by $Q(x^N)$ and $P(y^N|x^N)$, respectively. The proof of those properties only requires that $\sum_{x^N} Q(x^N||z^{N-1}) = 1$, which follows from (19), and

[1]The idea of converting the channel with feedback into a new channel without feedback as introduced by Tatikonda [7] can be easily extended also for the case of time-invariant feedback. The new channel without feedback is not necessarily an FSC and therefore the method works for theorems when their proofs do not need the assumption of having an FSC. As pointed by one of the reviewers, this method can be used for proving Lemma 4, Theorem 9, and Theorem 10.

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

$$\underline{C}_N = \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{\{s_0, y_{2-D}, \ldots, y_0\}} I(X^N \to Y^N | s_0, y_{2-D}, \ldots, y_0) \tag{59}$$

$\sum_{x^N, y^N} Q(x^N||z^{N-1}) P(y^N||x^N, s_0) = 1$ which follows Lemma 2. By using Lemma 5 we can substitute $\mathcal{I}(Q(x^N||y^{N-1}), P(y^N||x^N, s_0))$ in (47) by the directed mutual information $I(X^N \to Y^N|s_0)$.

In this paper, we use Theorem 11 to claim (in the proof of Theorem 14) that if $R < \mathcal{I}(Q(x^N||y^{N-1}), P(y^N||x^N, s_0))$, then there is a range of $\rho > 0$ for which

$$E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0) - \rho R > 0. \tag{49}$$

An alternative to the use of Theorem 11 is to use [24, Lemma 2], given by Lapidoth and Telatar. It is possible to extend [24, Lemma 2], in a straightforward manner, and to obtain that

$$E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0) \geq \rho \mathcal{I}(Q(x^N||y^{N-1}), P(y^N||x^N, s_0)) - \frac{1}{2}\rho^2 [\ln(e|\mathcal{Y}|)]^2. \tag{50}$$

Obviously, (50) also implies that (49) holds for some range of $\rho > 0$.

Lemma 12: (Super additivity of $NF_N(\rho)$, analogue to [1, Lemma 5.9.1].) For any given finite-state channel, $F_N(\rho)$, as given by (44), satisfies

$$F_N(\rho) \geq \frac{n}{N} F_n(\rho) + \frac{l}{N} F_l(\rho) \tag{51}$$

for all positive integers n and l with $N = n + l$.

The proof of the following lemma is given in Appendix V.

Lemma 13: (Convergence of $F_N(\rho)$, analogue to [1, Lemma 5.9.2].) Let

$$F_\infty(\rho) = \sup_N F_N(\rho) \tag{52}$$

then

$$\lim_{N \to \infty} F_N(\rho) = F_\infty(\rho) \tag{53}$$

for $0 \leq \rho \leq 1$. Furthermore, the convergence is uniform in ρ and $F_\infty(\rho)$ is uniformly continuous for $\rho \in [0, 1]$.

The proof of the lemma is identical to the proof of [1, Lemma 5.9.2] and therefore omitted. In the proof, Gallager uses the fact that $\frac{1}{N}\mathcal{I}(Q(x^N), P(y^N|x^N, s_0)) \leq \log|\mathcal{Y}|$ to bound the derivative of $F_N(\rho)$ by $\log|\mathcal{Y}|$ in the case of no feedback. The same bound applies in the case of feedback, i.e.,

$$\frac{1}{N}\mathcal{I}(Q(x^N||y^{N-1}), P(y^N||x^N, s_0)) \leq \log|\mathcal{Y}|. \tag{54}$$

The following theorem states that any rate R that satisfies $R < \underline{C}$ is achievable.

Theorem 14: (Analogue to [1, Theorem 5.9.2].) For any given finite-state channel, let

$$E_r(R) = \max_{0 \leq \rho \leq 1} [F_\infty(\rho) - \rho R]. \tag{55}$$

Then, for any $\epsilon > 0$, there exists $N(\epsilon)$ such that for $N \geq N(\epsilon)$ there exists an (N, M) code such that for all $m, 1 \leq m \leq M = \lceil 2^{NR} \rceil$, and all initial states

$$P_{e,m}(s_0) \leq 2^{-N[E_r(R) - \epsilon]}. \tag{56}$$

Furthermore, for $0 \leq R < \underline{C}$, $E_r(R)$ is strictly positive, and therefore the error can be arbitrarily small for N large enough.

The proof is identical to the proof of [1, Theorem 5.9.2] and therefore omitted. It uses Theorems 5.9.1, 5.6.1 and Lemma 5.9.2 that correspond to Theorem 10, 8, and Lemma 13 in this paper, to prove that for any s_0 there exists a ρ^* such that $F_\infty(\rho^*) - \rho^* R > 0$, for all $R < \underline{C}$.

D. Feedback That Is a Deterministic Function of a Finite Tuple of the Output

The proof of Theorem 14 holds for the case that the feedback z_i is a deterministic function of the output at time i, i.e., $z_i = z(y_i)$. We now extend the theorem to the case where the feedback is a deterministic function of a finite tuple of the output, i.e., $z_i = z(y_{i-D-1}, \ldots, y_i)$.

Consider the case $D = 2$. Let us construct a new finite state channel, with input x_i, and output $\tilde{y}_i$ that is the tuple $\{y_{i-1}, y_i\}$. The state of the new channel $\tilde{s}_i$ is the tuple $\{s_i, y_i\}$.

Let us verify that the definition of an FSC holds for the new channel:

$$\begin{aligned} P(\tilde{y}_i, \tilde{s}_i | \tilde{y}^{i-1}, \tilde{s}^{i-1}, x^i) &= P(y_i, y_{i-1}, s_i, y_i | y^{i-1}, s^{i-1}, y^{i-1}, x^i) \\ &= P(y_i, y_{i-1}, s_i | y_{i-1}, s_{i-1}, x_i) \\ &= P(\tilde{y}_i, \tilde{s}_i | \tilde{s}_{i-1}, x_i). \end{aligned} \tag{57}$$

Both channels are equivalent, and because the feedback z_i is a deterministic function of the output of the new channel $\tilde{y}_i$, we can apply Theorem 14 and obtain that any R that satisfies

$$\begin{aligned} R \leq \underline{C}_N &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{\tilde{s}_0} I(X^N \to \tilde{Y}^N | \tilde{s}_0) \\ &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0, y_0} I(X^N \to \{Y^N, Y_0^{N-1}\} | s_0, y_0) \\ &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0, y_0} \sum_{i=1}^{N} I(X^i; Y_i, Y_{i-1} | Y^{i-1}, Y^{i-2}, s_0, y_0) \\ &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0, y_0} \sum_{i=1}^{N} H(Y_i, Y_{i-1} | Y^{i-1}, Y^{i-2}, s_0, y_0) \\ &\quad - H(Y_i, Y_{i-1} | Y^{i-1}, Y^{i-2}, X^i, s_0, y_0) \\ &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0, y_0} \sum_{i=1}^{N} H(Y_i | Y^{i-1}, s_0, y_0) \\ &\quad - H(Y_i | Y^{i-1}, X^i, s_0, y_0) \\ &= \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0, y_0} I(X^N \to Y^N | s_0, y_0) \end{aligned} \tag{58}$$

is achievable for any initial state (s_0, y_0). This result can be extended by induction to the general case where the feedback z_i depends on a tuple of $D \geq 2$ outputs. It leads to the result that

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

any rate smaller than $\lim_{N\to\infty}\underline{C}_N$, given in (59) at the top of the page, is achievable for any initial state $(s_0, y_{2-D}, \ldots, y_0)$.

VI. Upper Bound on the Feedback Capacity

Recall the definitions

$$\overline{C}_N \triangleq \frac{1}{N}\max_{Q(x^N\|z^{N-1})}\max_{s_0} I(X^N \to Y^N|s_0) \tag{60}$$

and

$$\overline{C} \triangleq \lim_{N\to\infty}\overline{C}_N; \tag{61}$$

a limit that will be shown to exist in Theorem 16.

Theorem 15: The capacity of an FSC with a time-invariant deterministic feedback, as presented in Fig. 1, is upper-bounded by

$$C_{FB} \le \overline{C}; \tag{62}$$

in addition, for any positive integer N, we have the following computable upper bound:

$$C_{FB} \le \overline{C}_N + \frac{\log|\mathcal{S}|}{N}. \tag{63}$$

The first upper bound, given in (62), follows from Fano's inequality in a manner similar to the derivation of Massey's upper bound in [5]. The second upper bound, given in (62), is a direct consequence of the following lemma.

Theorem 16: (Analogue to [1, Theorem 4.6.1]). For a finite-state channel with $|\mathcal{S}|$ states, the limit in (61) exists, and

$$\lim_{N\to\infty}\overline{C}_N = \inf_N\left[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}\right]. \tag{64}$$

Similar to Gallager's proof, we show that the sequence $N[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}]$ is subadditive, i.e., for any N, n, l, such that $N = n + l$

$$N\left[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}\right] \le n\left[\overline{C}_n + \frac{\log|\mathcal{S}|}{n}\right] + l\left[\overline{C}_l + \frac{\log|\mathcal{S}|}{l}\right] \tag{65}$$

and this implies that

$$\lim_{N\to\infty}\left[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}\right] = \inf_N\left[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}\right]. \tag{66}$$

The proof differs from [1, Theorem 4.6.1]), since Gallager used the fact that $I(X_{n+1}^N; Y^n|X^n) = 0$[1, eq. (4A.24)], which obviously does not hold if feedback is used.

Proof of Theorem 16: Let $Q(x^N\|z^{N-1})$ and s_0 be the input distribution and the initial state that achieves $\overline{C}_N$. The distribution of the variables X^N, Y^N in the following sequence of equations is determined by the input distribution $Q_N(x^N\|z^{N-1})$ and the channel:

$$\begin{aligned}
N\overline{C}_N &= I(X^N \to Y^N|s_0)\\
&= \sum_{i=1}^{n} I(Y_i; X^i|Y^{i-1}, s_0) + \sum_{j=n+1}^{n+l} I(Y_j; X^j|Y^{j-1}, s_0)\\
&\le n\overline{C}_n + \sum_{j=n+1}^{n+l} I(Y_j; X^j|Y^{j-1}, s_0)\\
&\overset{(a)}{\le} n\overline{C}_n + \sum_{j=n+1}^{n+l} I(Y_j; X^j|Y^{j-1}, S_n, s_0) + \log|\mathcal{S}|\\
&= n\overline{C}_n + \sum_{j=n+1}^{n+l} H(Y_j|Y^{j-1}, S_n, s_0)\\
&\quad - H(Y_j|X^j, Y^{j-1}, S_n, s_0) + \log|\mathcal{S}|\\
&\overset{(b)}{\le} n\overline{C}_n + \sum_{j=n+1}^{n+l} H(Y_j|Y_{n+1}^{j-1}, S_n, s_0)\\
&\quad - H(Y_j|X_{n+1}^{j}, Y_{n+1}^{j-1}, S_n, s_0) + \log|\mathcal{S}|\\
&= n\overline{C}_n + \sum_{j=n+1}^{n+l} I(Y_j; X_{n+1}^{j}|Y_{n+1}^{j-1}, S_n, s_0) + \log|\mathcal{S}|\\
&= n\overline{C}_n + I(X_{n+1}^N \to Y_{n+1}^N|S_n, s_0) + \log|\mathcal{S}|\\
&\le n\overline{C}_n + \max_{s_n} I(X_{n+1}^N \to Y_{n+1}^N|s_n, s_0) + \log|\mathcal{S}|\\
&\overset{(c)}{\le} n\overline{C}_n + l\overline{C}_l + \log|\mathcal{S}|.
\end{aligned} \tag{67}$$

Inequality (a) is due to Lemma 4 where we treat the vector X^n to be the first element in the sequence, i.e., $(X^n, X_{n+1}, X_{n+2}, \ldots, X_{n+l})$. Inequality (b) results because conditioning reduces entropy and because

$$p(y_j|x^j, y^{j-1}, s_n, s_0) = p(y_j|x_{n+1}^j, y_{n+1}^{j-1}, s_n) \tag{68}$$

for $j > n$, and any FSC with and without feedback (see Appendix VI for the proof of (68)). Inequality (c) follows the following argument. Let

$$\overline{s}_n = \arg\max_{s_n} I(X_{n+1}^N \to Y_{n+1}^N|s_n, s_0)$$

and let Q_l denote the causally conditioned distribution $Q_l(x_{n+1}^N\|y_{n+1}^{N-1}, \overline{s}_n, s_0)$ induced by the input $Q_N(x^N\|z^{N-1})$ and the channel. Such a distribution exists, since, by Lemma 2, any joint distribution can be decomposed into casually conditioned distributions as

$$\begin{aligned}
&P(x_{n+1}^N, y_{n+1}^N|\overline{s}_n, s_0)\\
&\qquad = Q_l(x_{n+1}^N\|y_{n+1}^{N-1}, \overline{s}_n, s_0)P(y_{n+1}^N\|x_{n+1}^N, \overline{s}_n).
\end{aligned} \tag{69}$$

Since $\overline{s}_n$ and s_0 are fixed, Q_l is a legitimate input distribution, and since the dependency of the joint distribution $P(x_{n+1}^N, y_{n+1}^N|\overline{s}_n, s_0)$ on s_0 is only through the input distribution distribution Q_l, we have

$$I(x_{n+1}^N \to y_{n+1}^N|\overline{s}_n, s_0) = I_{Q_l}(x_{n+1}^N \to y_{n+1}^N|\overline{s}_n) \le \overline{C}_l. \tag{70}$$

Rearranging the last inequality of (67), we deduce that (65) holds, and since

$$\lim_{N\to\infty}\left[\overline{C}_N + \frac{\log|\mathcal{S}|}{N}\right] = \lim_{N\to\infty}\overline{C}_N$$

the lemma holds. □

Proof of Theorem 15: Let W be the message, chosen according to a uniform distribution $\Pr(W = w) = 2^{-NR}$. The input to the channel x_i is a function of the message W and the arbitrary deterministic feedback output $z^{i-1}(y^{i-1})$. For a code $(2^{NR}, N)$ with average probability $P_e^{(N)}$, we have

$$\begin{aligned}
NR &= H(W)\\
&= I(W; Y^N) + H(W|Y^N)
\end{aligned}$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

$$
\begin{aligned}
&\overset{(a)}{\leq} I(Y^N;W)+1+P_e^{(N)}NR \\
&= H(Y^N)-H(Y^N|W)+1+P_e^{(N)}NR \\
&= \sum_{i=1}^{N} H(Y_i|Y^{i-1}) - \sum_{i=1}^{N} H(Y_i|W,Y^{i-1})+1+P_e^{(N)}NR \\
&\overset{(b)}{=} \sum_{i=1}^{N} H(Y_i|Y^{i-1}) \\
&\quad - \sum_{i=1}^{N} H(Y_i|W,Y^{i-1},X^i(W,z^{i-1}(Y^{i-1}))) \\
&\quad +1+P_e^{(N)}NR \\
&\overset{(c)}{=} \sum_{i=1}^{N} H(Y_i|Y^{i-1}) - \sum_{i=1}^{N} H(Y_i|Y^{i-1},X^i) \\
&\quad +1+P_e^{(N)}NR \\
&= \sum_{i=1}^{N} I(Y_i;X^i|Y^{i-1})+1+P_e^{(N)}NR \\
&\overset{(d)}{\leq} I(X^N \to Y^N|S_0)+\log|\mathcal{S}|+1+P_e^{(N)}NR \\
&= \sum_{s_0\in\mathcal{S}} p(s_0) I(X^N \to Y^N|s_0)+\log|\mathcal{S}|+1+P_e^{(N)}NR \\
&\leq \max_{s_0} I(X^N \to Y^N|s_0)+\log|\mathcal{S}|+1+P_e^{(N)}NR. \quad (71)
\end{aligned}
$$

Inequality (a) follows from Fano's inequality. Equality (b) follows from the fact that x_i is a deterministic function given the message W and the feedback z^{i-1}. Equality (c) follows from the fact that the random variables W, X_i, Y^{i-1}, Y_i form the Markov chain $W-(X_i,Y^{i-1})-Y_i$, and inequality (d) follows Lemma 4. By dividing both sides of the equation by N, maximizing over all possible input distributions, and letting $N\to\infty$, we find that in order to have an error probability arbitrarily small, the rate R must satisfy

$$R \leq \overline{C} \overset{(a)}{\leq} \overline{C}_L + \frac{\log|\mathcal{S}|}{L}, \quad L=1,2,3\ldots \qquad (72)$$

where inequality (a) in (72) is due to Theorem 16. □

The upper bound in the previous theorem holds for any FSC, and does not depend on the initial state distribution $P(s_0)$. For a case in which the initial state s_0 has a positive probability for all $s_0\in\mathcal{S}$, an upper bound that coincides with the achievable rate can be derived.

Theorem 17: An achievable rate R of an FSC, where the initial state S_0 has a positive probability for all $s_0\in\mathcal{S}$, must satisfy

$$R \leq \underline{C} \qquad (73)$$

where $\underline{C}$ is defined as in the previous section

$$\underline{C} = \lim_{n\to\infty} \max_{Q(x^n||z^{n-1})} \min_{s_0} I(X^n \to Y^n|s_0). \qquad (74)$$

Since $\underline{C}$ is an achievable rate for any FSC (Theorem 14), we conclude that for an FSC, where $P(s_0)>0, \forall s_0\in\mathcal{S}$, the capacity is $\underline{C}$.

Proof: Recall that a rate R is achievable if there exists a sequence of block codes $(N,\lceil 2^{NR}\rceil)$ such that the probability of error $P_e^{(N)}$ goes to zero as $N\to\infty$. We denote $P_e^{(N)}(s_0)$ the error probability of a code given that the initial state is s_0. Since every initial state $s_0\in\mathcal{S}$ has a positive probability, we can infer that rate R is achievable, if there exists a sequence of block codes $(N,\lceil 2^{NR}\rceil)$ such that the probability of error $P_e(s_0)$ goes to zero for all $s_0\in\mathcal{S}$.

Since the message W is independent of s_0, it is possible to repeat the sequence of inequalities (a)–(c) given in (71), but with conditioning the random variables on s_0, i.e., $NR=H(W|s_0)$. Hence, we have

$$NR \leq I(X^n \to Y^n|s_0)+1+P_e^{(N)}(s_0)NR, \quad \forall s_0\in\mathcal{S}. \qquad (75)$$

And, in particular

$$NR \leq \min_{s_0}\left[I(X^n \to Y^n|s_0)+1+P_e^{(N)}(s_0)NR\right]. \qquad (76)$$

Finally, since an achievable rate R requires $P_e^{(N)}(s_0)\to 0$ as $N\to\infty$ for all s_0, we have

$$R \leq \lim_{n\to\infty} \max_{Q(x^n||z^{n-1})} \min_{s_0} I(X^n \to Y^n|s_0). \qquad (77)$$

□

Remark: The converse proofs are with respect to the average error over all messages. This, of course, implies that it is also true with respect to the maximum error over all messages. In the achievability part, we proved that the maximum error over all messages goes to zero when $R\leq\underline{C}$ which, of course, also implies that the average error goes to zero. Hence, both the achievability and the converse are true with respect to average error probability and maximum error probability over all messages.

VII. Stationary Indecomposable FSC Without ISI

In this section, we assume that the channel states evolve according to a Markov chain that does not depend on the input, namely, $P(y_i,s_i|s_{i-1},x_i)=P(s_i|s_{i-1})P(y_i|s_i,s_{i-1},x_i)$. In addition, we assume that the Markov chain is indecomposable according to Gallager's definition. Such a channel is called a finite state Markovian indecomposable channel (FSMIC) in [25]; however, another suitable name, which we adopt henceforth, is an indecomposable FSC without ISI.

The definition of an indecomposable FSC for the case of no ISI [1, p. 106] implies that there is one ergodic class.

Definition 2: An FSC without ISI is *indecomposable* if, for every $\epsilon>0$, there exists an N_0 such that for $N\geq N_0$

$$|P(s_N|s_0)-P(s_N|s_0')|\leq\epsilon \qquad (78)$$

for all s_N, s_0, s_0'.

Gallager also provides in [1, Theorem 4.6.3] a necessary and sufficient condition for verifying that the channel is indecomposable. The condition is that for some fixed $n\leq 2^{|\mathcal{S}|^2}$, there exists a state s_n such that

$$P(s_n|s_0)>0, \quad \forall s_0\in\mathcal{S}. \qquad (79)$$

This condition can be verified in a finite time, and it also implies [26, Theorem 6.3.2] that there exists a unique steady-state distribution (stationary distribution), i.e.,

$$\lim_{N\to\infty} \Pr(S_N=s|s_0)=\pi(s) \qquad (80)$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

where $\pi(s)$ is the stationary distribution. If $P(s_0)=\pi(s_0)$ we say that the channel is stationary.

Theorem 18: For a stationary and indecomposable FSC without ISI, the capacity of the channel is given by

$$C_{FB}=\lim_{N\to\infty}\frac{1}{N}\max_{Q(x^N||z^{N-1})} I(X^N\to Y^N). \quad (81)$$

Proof: Since $\underline{C}$ is achievable (Theorem 14), and since the right-hand side of(81) is an upper bound on the capacity (step (c) in (71)), it is enough to show that

$$\lim_{N\to\infty}\left(\frac{1}{N}\max_{Q(x^N||z^{N-1})} I(X^N\to Y^N)-\underline{C}_N\right)=0. \quad (82)$$

We will prove (82) by considering an input distribution for $\underline{C}_N$ that is arbitrary for the first n epochs time and then equal to the causally conditioned distribution obtained by maximizing $I(X^N\to Y^N)$. Since the channel is indecomposable and without ISI, the distribution of S_n can be made arbitrarily close to $\pi(s)$ by choosing n large enough, and this will allow us to bound the difference between $\underline{C}_N$ and $\frac{1}{N}I(X^N\to Y^N)$.

Let $n+l=N$, where n,l are positive integers. Then

$$\begin{aligned}
&I(X^N\to Y^N|s_0)\\
&\geq I(X^N\to Y^N|S_n,s_0)-\log|\mathcal{S}|\\
&=H(Y^N|S_n,s_0)-\sum_{i=1}^{N}H(Y_i|Y^{i-1},X^i,S_n,s_0)-\log|\mathcal{S}|\\
&\geq H(Y_{n+1}^N|S_n,s_0)-\sum_{i=n+1}^{N}H(Y_i|Y^{i-1},X^i,S_n,s_0)\\
&\quad -n\log|\mathcal{Y}|-\log|\mathcal{S}|\\
&\geq H(Y_{n+1}^N|S_n,s_0)-\sum_{i=n+1}^{N}H(Y_i|Y_{n+1}^{i-1},X_{n+1}^i,S_n,s_0)\\
&\quad -n\log|\mathcal{Y}|-\log|\mathcal{S}|\\
&=I(X_{n+1}^N\to Y_{n+1}^N|S_n,s_0)-n\log|\mathcal{Y}|-\log|\mathcal{S}|. \quad (83)
\end{aligned}$$

Let $Q_N^*(x^N||z^{N-1})=\arg\max_Q I(X^N\to Y^N)$. Q_N^* induces an input distribution

$$Q_l(x^l||z^{l-1})=\sum_{x_{l+1}^N}Q^*(x^N||z^{N-1}).$$

Now consider an input distribution

$$Q_N(x^N||z^{N-1})=Q_n(x^n)Q_l(x_{n+1}^N||z_{n+1}^{N-1})$$

where $Q_n(x^n)$ is an arbitrary distribution that does not depend on the feedback. Then, the joint distribution induced by Q_N satisfies

$$\begin{aligned}
&P(x_{n+1}^N,y_{n+1}^N|s_n,s_0)\\
&\quad=Q(x_{n+1}^N||y_{n+1}^{N-1},s_n,s_0)P(y_{n+1}^N||x_{n+1}^N,s_n,s_0)\\
&\quad=Q(x_{n+1}^N||y_{n+1}^{N-1})P(y_{n+1}^N||x_{n+1}^N,s_n)\\
&\quad=P(x_{n+1}^N,y_{n+1}^N|s_n). \quad (84)
\end{aligned}$$

Now, consider the following difference:

$$\begin{aligned}
&I_{Q_N^*}(X^N\to Y^N)-I_{Q_N}(X_{n+1}^N\to Y_{n+1}^N|S_n,s_0)\\
&\quad\leq I_{Q_N^*}(X^N\to Y^N|S_0)-I_{Q_N}(X_{n+1}^N\to Y_{n+1}^N|S_n,s_0)\\
&\qquad+\log|S|\\
&\quad\overset{(a)}{=} I_{Q_N^*}(X^N\to Y^N|S_0)\\
&\qquad-\sum_{s_n}P(s_n|s_0)I_{Q_N}(X_{n+1}^N\to Y_{n+1}^N|s_n)+\log|S|\\
&\quad\overset{(b)}{\leq} I_{Q_N^*}(X^N\to Y^N|S_0)\\
&\qquad-\sum_{s_n}(\pi(s_n)-\epsilon(n))I_{Q_N}(X_{n+1}^N\to Y_{n+1}^N|s_n)+\log|S|\\
&\quad\overset{(c)}{=} I_{Q_N^*}(X^N\to Y^N|S_0)\\
&\qquad-\sum_{s_0}(\pi(s_0)-\epsilon(n))I_{Q_l}(X^l\to Y^l|s_0)+\log|S|\\
&\quad\leq n\log|\mathcal{Y}|+I_{Q_l}(X^l\to Y^l|S_0)\\
&\qquad-\sum_{s_0}(\pi(s_0)-\epsilon(n))I_{Q_l}(X^l\to Y^l|s_0)+\log|S|\\
&\quad\overset{(d)}{\leq} n\log|\mathcal{Y}|+|\mathcal{S}|\epsilon(n)l|\log|\mathcal{Y}|+\log|S| \quad (85)
\end{aligned}$$

where (a) follows from (84), (b) follows from the assumption that the channel is indecomposable (see (78)) and has no ISI and its state distribution, hence, converges to a stationary distribution, (c) follows from a time shift, and (d) follows because stationarity of the channel implies that $\sum_{s_0}\pi(s_0)I_{Q_l}(X^l\to Y^l|s_0)=I_{Q_l}(X^l\to Y^l|S_0)$.

Combining (83) and (85), we obtain

$$\begin{aligned}
\frac{1}{N}\left(\max_{Q(x^N||y^N)}I(X^N\to Y^N)-\max_{Q(x^N||y^N)}\min_{s_0}(X^N\to Y^N|s_0)\right)\\
\leq\frac{1}{N}\left(2n+|\mathcal{S}|\epsilon(n)l\right)\log|\mathcal{Y}|+2\log|S|). \quad (86)
\end{aligned}$$

Finally, for any $\delta>0$, we can choose n such that $|\mathcal{S}|\epsilon(n)\log|\mathcal{Y}|\leq\frac{\delta}{2}$, and we can choose N such that $\frac{2n\log|\mathcal{Y}|+2\log|S|}{N}\leq\frac{\delta}{2}$, and therefore we conclude that (82) holds. □

From the definition of channel capacity, it follows that the probability of error $P_e^{(n)}(s_0)$ has to go to zero for all $\{s_0: P(s_0)>0\}$. Hence, the stationarity condition in Theorem 18 can be relaxed to the condition that the support of $P(s_0)$ contains the support of $\pi(s_0)>0$, i.e., $P(s_0)>0$ if $\pi(s_0)>0$; however, the calculation of $\lim_{N\to\infty}\frac{1}{N}\max_{Q(x^N||z^{N-1})}I(X^N\to Y^N)$ in Theorem 18 should be still done as with an initial distribution $\pi(s_0)$.

VIII. Feedback and Side Information

The results of the previous sections can be extended to the case where side information is available at the decoder that might also be fed back to the encoder. Let l_i be the side information available at the decoder and the setting of communication that us shown Fig. 3. If the side information l_i satisfies

$$P(l_i,y_i,s_i|s^{i-1},x^i,y^{i-1},l^{i-1})=P(l_i,y_i,s_i|s_{i-1},x_i) \quad (87)$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

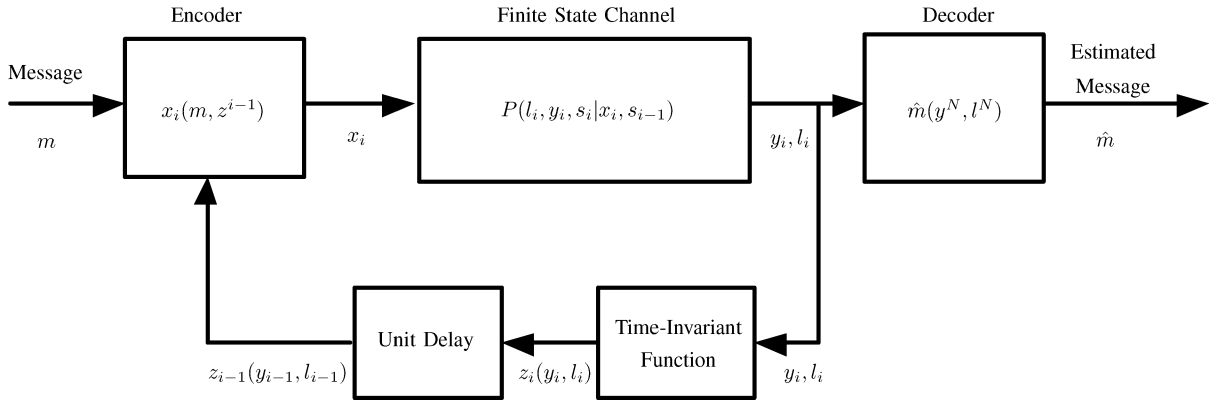

Fig. 3. Channel with feedback and side information l_i.

then it follows that

$$P(\bar{y}_i, s_i|s^{i-1}, x^i, \bar{y}^{i-1}) = P(\bar{y}_i, s_i|s_{i-1}, x_i) \tag{88}$$

where $\bar{y}_i = (l_i, y_i)$. We can now apply Theorem 14 and get

$$\underline{C}_N = \frac{1}{N} \max_{Q(x^N||z^{N-1})} \min_{s_0} I(X^N \to \{Y^N, L^N\}|s_0) \tag{89}$$

where z_{i-1} denotes the feedback available at the receiver at time i, which is a time-invariant function of l_{i-1} and y_{i-1}.

Here we consider only the case in which the side information is the state of the channel, i.e., $l_i = s_i$, and we show in the next theorem that if the state is known both to the encoder and decoder, then output feedback does not increase the capacity of a connected FSC. In this section, we no longer assume that there is no ISI; rather we assume that the FSC is connected, which we define as follows.

Definition 3: We say that a finite-state channel is *connected* if there exists an input distribution $\{Q(x_t|s_{t-1})\}_{t\geq 1}$ and integer T such that

$$\Pr\{S_t = s \text{ for some } 1 \leq t \leq T | S_0 = s'\} > 0, \quad \forall s' \in \mathcal{S}, s \in \mathcal{S}. \tag{90}$$

Theorem 19: Feedback does not increase the capacity of a connected FSC when the state of the channel is known both at the encoder and the decoder.

The theorem is proved in Appendix VIII by using the lower and upper bound of capacity of FSC with time-invariant feedback. For several particular cases, this result has been already shown. For instance, Shannon showed in [2] that feedback does not increase the capacity of a discrete memoryless channel (DMC). A DMC can be considered as an FSC with only one state, and therefore the state of the channel is known to the encoder and the decoder. For the case that the channel has no ISI, namely, $P(s_i|s^{i-1}, y^i, x^i) = P(s_i|s_{i-1})$, the result was shown by Viswanathan in [17]. And, if the input distribution is restricted to the form $\{Q(x_i|y^{i-1}, s^{i-1})\}_{i=1}^N$ (as opposed to $\{Q(x_i|x^{i-1}, y^{i-1}, s^{i-1})\}_{i=1}^N$), then results from Tatikonda's thesis [7, Lemmas 4.5.3–4.5.5] can directly prove the theorem.

IX. Source–Channel Separation

For channels that their feedback capacity is given by

$$C = \lim_{N\to\infty} \frac{1}{N} \max_{Q(x^N||z^{N-1})} I(X^N \to Y^N), \tag{91}$$

and for ergodic sources, a simple derivation can show that the optimality of the source–channel separation holds. This means that the distortion that ca be achieved with the communication scheme in Fig. 4 can also be achieved with the communication scheme presented in Fig. 5. Conditions on the source and the channel for having a separation are needed, since even for the case of no feedback, the separation does not always hold [27, Sec. III]. Sufficient and necessary conditions for the separation in the absence of feedback are given in [27]. Here, sufficient conditions are given for the case that deterministic feedback is allowed.

Theorem 20: Consider a channel with its capacity given in (91) (e.g., stationary indecomposable FSC without ISI). Let $\epsilon > 0$ and $D \geq 0$ be given. Let $R(\cdot)$ be the rate distortion function of a discrete, stationary, ergodic source with respect to a single-letter criterion generated by a bounded distortion measure ρ. Then the source output can be reproduced with fidelity $D + \epsilon$ at the receiving end if $C > R(D)$. Conversely, fidelity D is unattainable at the receiving end if $C < R(D)$.

Remark: For simplicity of presentation, we assumed one channel use per source symbol. The derivation below extends to the general case where the average number of channel uses per letter is $\frac{\tau_s}{\tau_c}$, analogously as in [1, Ch. 9].

Proof: The direct proof, namely, that if $C > R(D)$ it is possible to reproduce the source with fidelity D, is straightforward by using the source–channel separation scheme from Fig. 5. The encoder first encodes the source into an index using a rate distortion scheme at a rate $R(D)$ and then sends this index as a message through the channel with feedback. Since the maximum probability of error is arbitrary small at a rate less then C, the fidelity $D + \epsilon$ is achieved, where ϵ is arbitrarily small.

For the converse, namely that $R(D)$ has to be less or equal to C, we use the directed data processing inequality which was derived by Tatikonda [7, Lemma 4.8.1], primarily, for this purpose (e.g., see [7, Theorem 5.3.2] for its use in a converse for

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

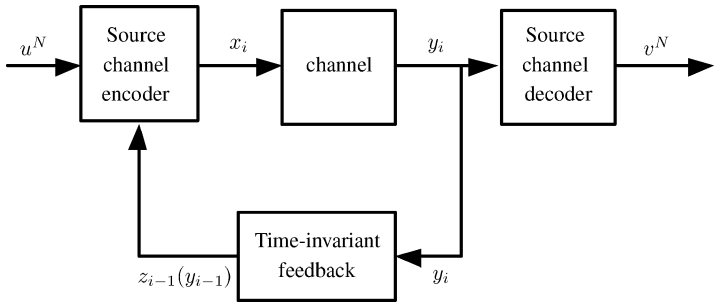

Fig. 4. Source and channel coding, where the channel has time-invariant feedback.

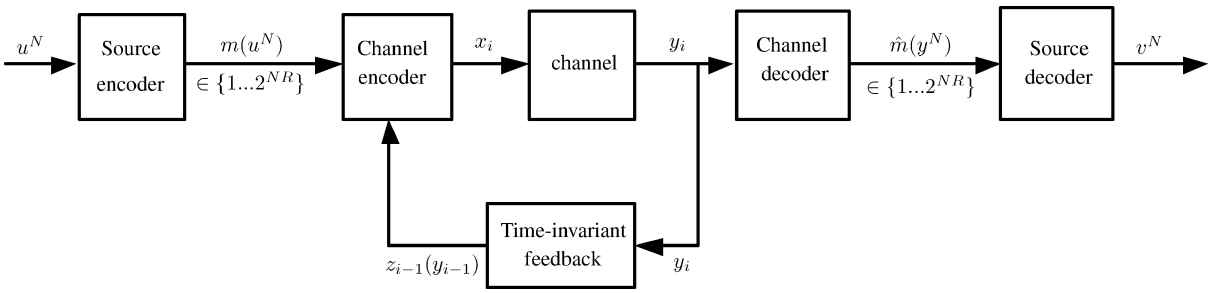

Fig. 5. Source and channel separation.

sequential rate distortion through a channel).[2] The directed data processing inequality says that in the setting given in Fig. 4, for any channel, we have the inequality

$$I(U^N;V^N) \le I(X^N \to Y^N). \tag{92}$$

Based on this inequality, we have

$$\begin{aligned} NR(D) &\overset{(a)}{\le} I(U^N;V^N) \\ &\overset{(b)}{\le} I(X^N \to Y^N) \\ &\overset{(c)}{\le} N(C+\epsilon_N). \end{aligned} \tag{93}$$

Inequality (a) follows the converse for rate distortion [28, Theroem 7.2.5]. Inequality (b) follows the directed data processing inequality, and (c) follows the converse of channel with feedback, where $\epsilon_N \to 0$ as $N \to \infty$. □

X. Conclusion and Future Work

We determined a sequence of achievable rates and a sequence of upper bounds on the capacity of FSCs with feedback that is a deterministic function of the channel output. All bounds are computable for any N. The achievable rates are obtained via a random generated coding scheme that utilizes feedback, along with an ML decoder; the upper bounds are obtained via Fano's inequality. The techniques in this paper extend Gallager's technique by using causal conditioning rather than regular conditioning and by including deterministic feedback. If the initial state S_0 has positive probability for all states $s_0 \in \mathcal{S}$, or if the channel is a stationary indecomposable FSC without ISI, then the capacity is established.

In addition to the coding theorem that is presented here, there are two additional coding theorems that appeared in parallel. The first, by Tatikonda and Mitter [29] is based on Tatikonda's thesis, and the second is by Kim [30]. The assumptions, the techniques, and the final results complement each other. Tatikonda and Mitter[3] assume an arbitrary channel and use Feinstein's lemma and the notion of $\limsup$ in probability. Kim assumes a stationary channel (channels in which their state process evolves independently of the input and is a stationary and ergodic process) and uses ergodic theory to analyze the error probability of a decoder based on the method of types. Here, we assume a finite-state channel and use Gallager's methods to bound the error probability of an ML decoding.

By using the directed information formula for the capacity of FSCs with feedback developed in this work, it was shown in [31] that the feedback capacity of a channel introduced by Blackwell in 1961 [32], also known as the trapdoor channel [26], is the logarithm of the golden ratio. The capacity of Blackwell's channel without feedback is still unknown. One future work is to find the capacity of additional channels with time-invariant feedback.

Appendix I
Proof of Lemma 4

$$\begin{aligned} &|I(X^N \to Y^N) - I(X^N \to Y^N, S)| \\ &\quad \overset{(a)}{=} \left| \sum_{i=1}^{N} I(Y_i;X^i|Y^{i-1}) - I(Y_i;X^i|Y^{i-1},S) \right| \\ &\quad = \Bigg| \sum_{i=1}^{N} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},X^i) \\ &\qquad - H(Y_i|Y^{i-1},S) + H(Y_i|Y^{i-1},X^i,S) \Bigg| \end{aligned}$$

[2]As mentioned by one of the reviewers, the data processing inequality can be also derived from Massey's result in [5, Theorem 3], $I(U^N;Y^N) \le I(X^N \to Y^N)$, and the fact that we have the Markov form $U^N - Y^N - V^N$.

[3]In [29], in addition to the general coding theorem, the authors also treat Markov channels, provide mixing conditions that insure information stability, formulate the capacity problem as a Markov decision process, and formulate an ML decoder.

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

$$
\begin{aligned}
&= \left| \sum_{i=1}^{N} I(Y_i; S|Y^{i-1}) - I(Y_i; S|Y^{i-1}, X^i) \right| \\
&\overset{(b)}{\leq} \max\left(\sum_{i=1}^{N} I(Y_i; S|Y^{i-1}), \sum_{i=1}^{N} I(Y_i; S|Y^{i-1}, X^i) \right) \\
&\overset{(c)}{=} \max\left(I(Y^N; S), I(Y^N, X_2^N; S|X_1) \right) \\
&\overset{(d)}{\leq} \max\left(H(S), H(S) \right) \\
&\leq \log|\mathcal{S}|. \qquad (94)
\end{aligned}
$$

Equality (a) is due to the definition of the directed information. Inequality (b) holds because the magnitude of the difference between two positive numbers is smaller than the maximum of the numbers. Inequality (c) results from the fact that $I(Y_i; S|Y^{i-1}, X^i) \leq I(Y_i, X_{i+1}; S|Y^{i-1}, X^i)$ and then uses the chain rule of mutual information. Inequality (d) is because the mutual information of two variables is smaller than the entropy of each variable, and the last inequality holds because the cardinality of the alphabet of S is $|\mathcal{S}|$. □

APPENDIX II
PROOF OF LEMMA 6

The joint distribution $P(y^N, s_0^N, x^N)$ can be calculated recursively by the following recursive, shown in (95) at the bottom of the page, where $P(y_1, s_0^1, x_1) = P(s_0)Q(x_1)P(y_1, s_1|x_1, s_0)$. It follows from (95) that

$$P(y^N, s_0^N, x^N) = P(s_0)Q(x^N||y^{N-1}) \prod_{i=1}^{N} P(y_i, s_i|x_i, s_{i-1}). \qquad (96)$$

By summing over s_0^N and dividing by $Q(x^N||y^{N-1})$, we get (27), and, similarly, by only summing over s^N and dividing by $Q(x^N||y^{N-1})P(s_0)$, we get (26). □

APPENDIX III
PROOF OF THEOREM 9

The proof follows [1, pp. 136–137], but we take into account that we have codewords generated by $Q(x^N||z^{N-1})$ rather than codewords generated by $Q(x^N)$ and that the ML is $P(y^N||x^N)$ rather than $P(y^N|x^N)$. The proof hinges on the fact that given an output y^N, there is a mapping from m to a unique x^N. For the case of noisy feedback, this property does hold and because of that the theorem is not valid for noisy feedback.

Proof:

$$
\begin{aligned}
&\mathbf{E}(P_{e,m}) \\
&= \sum_{y^N} \sum_{x^N} P(x^N, y^N) P[error|m, x^N, y^N] \\
&= \sum_{y^N} \sum_{x^N} Q(x^N||z^{N-1}) P(y^N||x^N) P[error|m, x^N, y^N] \qquad (97)
\end{aligned}
$$

where $P[\text{error}|m, x^N, y^N]$ is the probability of decoding error conditioned on the message m, the output y^N, and the input x^N. The second equality is due to Lemma 1. Throughout the reminder of the proof we fix the message m. For a given tuple (m, x^N, y^N), define the event $A_{m'}$, for each $m' \neq m$, as the event, in which the message m' is selected in such a way that $P(y^N|m') > P(y^N|m)$, which according to (40) is the same as $P(y^N||x'^N) > P(y^N||x^N)$, where x'^N is a shorthand notation for $x^N(m', z^{N-1}(y^{N-1}))$, and x^N is a shorthand notation for $x^N(m, z^{N-1}(y^{N-1}))$. From the definition of $A_{m'}$ we have

$$
\begin{aligned}
&P(A_{m'}|m, x^N, y^N) \\
&= \sum_{x'^N} Q(x'^N||z^{N-1}) \cdot \mathbf{1}[P(y^N||x'^N) > P(y^N||x^N)] \\
&\leq \sum_{x'^N} Q(x'^N||z^{N-1}) \left[\frac{P(y^N||x'^N)}{P(y^N||x^N)} \right]^s, \quad \text{any } s > 0 \qquad (98)
\end{aligned}
$$

where $\mathbf{1}(x)$ denotes the indicator function

$$
\begin{aligned}
&P[\text{error}|m, x^N, y^N] \\
&= P\Big(\bigcup_{m' \neq m} A_{m'} | m, x^N, y^N \Big) \\
&\leq \min\left\{ \sum_{m' \neq m} P(A_{m'}|m, x^N, y^N), 1 \right\} \\
&\leq \left[\sum_{m' \neq m} P(A_{m'}|m, x^N, y^N) \right]^\rho, \qquad \text{any } 0 \leq \rho \leq 1 \\
&\leq \left[(M-1) \sum_{x'^N} Q(x'^N||z^{N-1}) \left[\frac{P(y^N||x'^N)}{P(y^N||x^N)} \right]^s \right]^\rho, \\
&\qquad\qquad 0 \leq \rho \leq 1, s > 0 \qquad (99)
\end{aligned}
$$

where the last inequality is due to inequality (98). By substituting inequality (99) in (97), we obtain

$$
\begin{aligned}
\mathbf{E}[P_{e,m}] \leq (M-1)^\rho \sum_{y^N} &\left[\sum_{x^N} Q(x^N||z^{N-1}) P(y^N||x^N)^{1-s\rho} \right] \\
&\times \left[\sum_{x'^N} Q(x'^N||z^{N-1}) P(y^N||x'^N)^s \right]^\rho. \qquad (100)
\end{aligned}
$$

By substituting $s = 1/(1+\rho)$, and recognizing that x' is a dummy variable of summation, we obtain (42) and complete the proof. □

APPENDIX IV
PROOF OF THEOREM 10

This proof follows Gallager's proof in [1, Theorem 5.9.1] with a modification that is presented here. First, Gallager argues that under the assumption that the initial state is uniformly distributed over the alphabet $\mathcal{S}$, if $\mathbf{E}(P_{e,m}) \leq \epsilon$, then we can infer

$$
\begin{aligned}
P(y^N, s_0^N, x^N) &= P(x_N, y_N, s_N|y^{N-1}, x^{N-1}, s_0^{N-1}) P(y^{N-1}, x^{N-1}, s_0^{N-1}) \\
&= Q(x_N|y^{N-1}, x^{N-1}) P(y_N, s_N|x_N, s_{N-1}) P(y^{N-1}, x^{N-1}, s^{N-1}) \qquad (95)
\end{aligned}
$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

that for any state $s_0 \in \mathcal{S}$, $\mathbf{E}(P_{e,m}(s_0)) \leq |\mathcal{S}|\epsilon$. Hence, we start the proof by assuming uniform distribution of the initial state, i.e., $P(s_0) = \frac{1}{|\mathcal{S}|}$.

In the case that there is no feedback and the initial state has a uniform distribution, then we trivially have

$$P(y^N|x^N) = \sum_{s_0} \frac{1}{|\mathcal{S}|} P(y^N|x^N, s_0). \tag{101}$$

For the case of a channel with feedback, we need several steps to derive a similar equality for the causal conditioning distributions.

$$\begin{aligned} P(y^N||x^N) &\overset{(a)}{=} P(y^N|m) \\ &= \sum_{s_0} P(y^N, s_0|m) \\ &\overset{(b)}{=} \sum_{s_0} P(s_0) P(y^N|m, s_0) \\ &= \sum_{s_0} \frac{1}{|\mathcal{S}|} \prod_{i=1}^{N} P(y_i|y^{i-1}, m, s_0) \\ &= \sum_{s_0} \frac{1}{|\mathcal{S}|} \prod_{i=1}^{N} P(y_i|y^{i-1}, m, x^i, s_0) \\ &= \sum_{s_0} \frac{1}{|\mathcal{S}|} \prod_{i=1}^{N} P(y_i|y^{i-1}, x^i, s_0) \\ &= \sum_{s_0} \frac{1}{|\mathcal{S}|} P(y^N||x^N, s_0). \end{aligned} \tag{102}$$

Equality (a) is shown in (40), and equality (b) holds due to the assumption that the initial state S_0 and the message m are independent. Thus, assuming that s_0 is uniformly distributed, the bound on error probability under ML decoding given in Theorem 9 becomes (103), shown at the bottom of the page. Expression (103) is almost identical to [1, eq. (5.9.1)], only here the regular conditioning is replaced by causal conditioning. Now we can apply steps identical to Gallager's in [1, eqs. (5.9.2)-(5.9.5)], and we get

$$\begin{aligned} P_{e,m}(s_0) \leq 4|\mathcal{S}|(M-1)^\rho |\mathcal{S}|^\rho \min_{Q(x^N||z^{N-1})} \max_{s_0} \sum_{y^N} \\ \times \left\{ \sum_{x^N} Q(x^N||z^{N-1}) \left[P(y^N||x^N, s_0) \right]^{\frac{1}{(1+\rho)}} \right\}^{1+\rho}. \end{aligned} \tag{104}$$

These steps are: deriving a bound on the maximum error probability (over the messages) from the average error probability, and using the inequalities, $(\sum_i a_i)^r \leq \sum_i (a_i)^r$, for $0 \leq r \leq 1$ and Jensen's inequality $(\sum_i P_i a_i)^r \leq \sum_i P_i (a_i)^r$, for $0 \leq r \leq 1$.

Finally, by substituting $M = 2^{NR}$ and (44) and (45) into (104), the theorem is proved. □

Appendix V
Proof of Lemma 12

The proof follows closely the proof of [1, Lemma 5.9.1]. The main difference is in the step of obtaining (108) and is due to the fact that not every property that holds for regular conditioning also holds for causal conditioning.

Let us divide the input x^N into two sets $\boldsymbol{x}_1 = x_1^n$ and $\boldsymbol{x}_2 = x_{n+1}^N$. Similarly, let us divide the output y^N into two sets $\boldsymbol{y}_1 = y_1^n$ and $\boldsymbol{y}_2 = y_{n+1}^N$ and the feedback z^N into $\boldsymbol{z}_1 = z_1^{n-1}$ and $\boldsymbol{z}_2 = z_{n+1}^{N-1}$. Let

$$Q_n(\boldsymbol{x}_1||\boldsymbol{z}_1) = \prod_{i=1}^{n} Q(x_i|x^{i-1}, z^{i-1})$$

and

$$Q_l(\boldsymbol{x}_2||\boldsymbol{z}_2) = \prod_{i=1}^{l} Q(x_{n+i}|x_{n+1}^{n+i}, z_{n+1}^{n+i-1})$$

be the probability assignments that achieve the maxima $F_n(\rho)$ and $F_{\boldsymbol{l}}(\rho)$, respectively. Let us consider the probability assignment $Q(x^N||z^{N-1}) = Q_n(\boldsymbol{x}_1||\boldsymbol{z}_1) Q_l(\boldsymbol{x}_2||\boldsymbol{z}_2)$. Then

$$F_N \geq -\frac{\rho \log |\mathcal{S}|}{N} + E_{o,N}(\rho, Q(x^N||z^{N-1}, s_0')) \tag{105}$$

where s_0' is the state that minimizes $E_{o,N}(\rho, Q(x^N||z^{N-1}), s_0')$.

Now

$$\begin{aligned} P(y^N||x^N, s_0') &\overset{(a)}{=} P(y^N|m, s_0') \\ &= \sum_{s_n} P(y^N, s_n|m, s_0') \\ &= \sum_{s_n} P(\boldsymbol{y}_1, s_n|m, s_0') P(\boldsymbol{y}_2|m, s_n, \boldsymbol{y}_1, s_0') \\ &= \sum_{s_n} P(\boldsymbol{y}_1, s_n|m, s_0') P(\boldsymbol{y}_2||\boldsymbol{x}_2, s_n). \end{aligned} \tag{106}$$

Equality (a) can be proved in the same way as (40) was proved. The term $P(\boldsymbol{y}_1, s_n|m, s_0')$ can be also expressed in terms of $\boldsymbol{y_1}, \boldsymbol{x_1}$ in the following way:

$$\begin{aligned} P(\boldsymbol{y}_1, s_n|m, s_0') &= P(s_n|m, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1|m, s_0') \\ &= P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0). \end{aligned} \tag{107}$$

Hence, we obtain

$$\begin{aligned} &P(y^N||x^N, s_0') \\ &\quad = \sum_{s_n} P(\boldsymbol{y}_2||\boldsymbol{x}_2, s_n) P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0). \end{aligned} \tag{108}$$

$$\mathbf{E}(P_{e,m}) \leq (M-1)^\rho \sum_{y^N} \left\{ \sum_{x^N} Q(x^N||z^{N-1}) \left[\sum_{s_0} \frac{1}{|\mathcal{S}|} P(y^N||x^N, s_0) \right]^{\frac{1}{(1+\rho)}} \right\}^{1+\rho}, \qquad 0 \leq \rho \leq 1. \tag{103}$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

$$
\begin{aligned}
2^{[-NF_N(\rho)]} &\overset{(a)}{\leq} |\mathcal{S}|^{\rho} \sum_{y^N} \left[\sum_{x^N} Q(x^N||z^{N-1}) P(y^N||x^N, s_0')^{1/(1+\rho)} \right]^{1+\rho} \\
&\overset{(b)}{=} |\mathcal{S}|^{\rho} \sum_{\boldsymbol{y}_1 \boldsymbol{y}_2} \left\{ \sum_{\boldsymbol{x}_1 \boldsymbol{x}_2} Q(\boldsymbol{x}_1||\boldsymbol{z}_1) Q(\boldsymbol{x}_2||\boldsymbol{z}_2) \left[\sum_{s_n} P(\boldsymbol{y}_2||\boldsymbol{x}_2, s_n) P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0) \right]^{1/(1+\rho)} \right\}^{1+\rho} \\
&\overset{(c)}{\leq} |\mathcal{S}|^{2\rho} \sum_{s_n} \sum_{\boldsymbol{y}_1 \boldsymbol{y}_2} \left\{ \sum_{\boldsymbol{x}_1 \boldsymbol{x}_2} Q(\boldsymbol{x}_1||\boldsymbol{z}_1) Q(\boldsymbol{x}_2||\boldsymbol{z}_2) \left[P(\boldsymbol{y}_2||\boldsymbol{x}_2, s_n) P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0) \right]^{1/(1+\rho)} \right\}^{1+\rho} \\
&\leq |\mathcal{S}|^{2\rho} \sum_{s_n} \sum_{\boldsymbol{y}_1 y_2} \left[\sum_{\boldsymbol{x}_1} Q(\boldsymbol{x}_1||\boldsymbol{z}_1) [P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0)]^{1/(1+\rho)} \right]^{1+\rho} \\
&\qquad \times \left[\sum_{\boldsymbol{x}_2} Q(\boldsymbol{x}_2||\boldsymbol{z}_2) P(\boldsymbol{y}_2||\boldsymbol{x}_2, s_n)^{1/(1+\rho)} \right]^{1+\rho} \\
&\leq |\mathcal{S}|^{2\rho} \sum_{s_n} \sum_{\boldsymbol{y}_1} \left[\sum_{\boldsymbol{x}_1} Q(\boldsymbol{x}_1||\boldsymbol{z}_1) [P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0)]^{1/(1+\rho)} \right]^{1+\rho} 2^{[-lF_l(\rho)]} \\
&\overset{(d)}{\leq} |\mathcal{S}|^{\rho} \sum_{\boldsymbol{y}_1} \left\{ \sum_{\boldsymbol{x}_1} Q(\boldsymbol{x}_1||\boldsymbol{z}_1) \left[\sum_{s_n} P(s_n|\boldsymbol{x}_1, \boldsymbol{y}_1, s_0') P(\boldsymbol{y}_1||\boldsymbol{x}_1, s_0) \right]^{1/(1+\rho)} \right\}^{1+\rho} 2^{[-lF_l(\rho)]} \\
&\leq 2^{[-nF_n(\rho) - lF_l(\rho)]}. \qquad (109)
\end{aligned}
$$

As in [1, eqs. (5.9.11)–(5.9.15)], we get (109), shown at the top of the page. Inequality (a) is due to inequality (105). Equality (b) is due to (108). Inequality (c) results from $(\sum_i a_i)^r \leq \sum_i (a_i)^r$. Inequality (d) is due to Minkowski's inequality

$$\left[\sum_j P_j \left(\sum_k a_{jk} \right)^{1/r} \right]^r \geq \sum_k \left(\sum_j P_j a_{jk}^{1/r} \right)^r$$

for $r > 1$. □

Appendix VI
Proof of the Markov Chain $(X^n, Y^n, S^{n-1}) - (S_n, X_{n+1}^N, Y_{n+1}^{N-1}) - Y_N$

Lemma 21: For an FSC , with and without feedback, we have

$$p(y_N|x^N, y^{N-1}, s^n) = p(y_N|x_{n+1}^N, y_{n+1}^{N-1}, s_n) \qquad (110)$$

for any $N > n$.

Proof:

$$
\begin{aligned}
P(x^N, y^N, s^n) &= \sum_{s_{n+1}, \ldots, s_N} P(x^N, y^N, s^N) \\
&= \sum_{s_{n+1}, \ldots, s_N} P(x^n, y^n, s^n) \\
&\quad \times \prod_{j=n+1}^{N} P(x_j|x^{j-1}, y^{j-1}) P(y_j, s_j|x_j, s_{j-1}) \\
&= P(x^n, y^n, s^n) \prod_{j=n+1}^{N} P(x_j|x^{j-1}, y^{j-1}) \\
&\quad \times \sum_{s_{n+1}, \ldots, s_N} \prod_{k=n+1}^{N} P(y_k, s_k|x_k, s_{k-1}). \qquad (111)
\end{aligned}
$$

Similarly, $P(x^N, y^{N-1}, s^n)$ has the expansion

$$
\begin{aligned}
&P(x^N, y^{N-1}, s^n) \\
&\quad = P(x^n, y^n, s^n) \prod_{j=n+1}^{N} P(x_j|x^{j-1}, y^{j-1}) \sum_{s_{n+1}, \ldots, s_N} \\
&\quad \times \prod_{k=n+1}^{N-1} P(y_k, s_j|x_k, s_{k-1}) P(s_N|x_N, s_{N-1}). \qquad (112)
\end{aligned}
$$

Therefore, we get that

$$
\begin{aligned}
P(y_N|x^N, y^{N-1}, s^n) &= \frac{P(x^N, y^N, s^n)}{P(x^N, y^{N-1}, s^n)} \\
&= \frac{\sum_{s_{n+1}, \ldots, s_N} \prod_{j=n+1}^{N} P(y_j, s_j|x_j, s_{j-1})}{\sum_{s_{n+1}, \ldots, s_N} \left(\prod_{j=n+1}^{N-1} P(y_j, s_j|x_j, s_{j-1}) \right) P(s_N|x_N, s_{N-1})}. \qquad (113)
\end{aligned}
$$

Since x^n, y^n, s^{n-1} does not appear in the last expression, we can conclude that

$$P(y_N|x^N, y^{N-1}, s^n) = P(y_N|x_{n+1}^N, y_{n+1}^{N-1}, s_n)$$

for $N > n$. □

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

APPENDIX VII
PROOF OF THE MARKOV CHAIN $(X^n, Y^n, S^{n-1}) - S_n - (X_{n+1}^N, Y_{n+1}^N)$

Lemma 22: For an FSC , with an input distribution of the form

$$Q(x^N \| z^{N-1}) = Q(x^n \| z^{n-1}) Q(x_{n+1}^N \| z_{n+1}^{N-1}), \tag{114}$$

the following holds:

$$P(x_{n+1}^j, y_{n+1}^j | x^n, y^n, s^n) = P(x_{n+1}^j, y_{n+1}^j | s_n), \quad j = n+1, \ldots, N. \tag{115}$$

Proof:

$$\begin{aligned} & P(x_{n+1}^j, y_{n+1}^j | x^n, y^n, s^n) \\ &= \prod_{i=n+1}^{j} P(x_i, y_i | x^{i-1}, y^{i-1}, s^n) \\ &= \prod_{i=n+1}^{j} P(x_i | x^{i-1}, y^{i-1}, s^n) P(y_i | x^i, y^{i-1}, s^n) \\ &\overset{(a)}{=} \prod_{i=n+1}^{j} Q(x_i | x_{n+1}^{i-1}, y_{n+1}^{i-1}) P(y_i | x^i, y^{i-1}, s^n) \\ &\overset{(b)}{=} \prod_{i=n+1}^{j} Q(x_i | x_{n+1}^{i-1}, y_{n+1}^{i-1}) P(y_i | x_{n+1}^i, y_{n+1}^{i-1}, s_n) \end{aligned} \tag{116}$$

where equality (a) is due to the assumption in (114), and equality (b) is due to Lemma 21, which is given in the preceding appendix. □

APPENDIX VIII
PROOF OF THEOREM 19

Proof: First, we note that because the state of the channel is known both to the encoder and the decoder, and because the FSC is connected, we can assume that with probability $1-\epsilon$, where ϵ is arbitrarily small, the FSC channel can be driven, in a finite time, to the state that maximizes the achievable rate. Hence, the achievable rate $\underline{C}$ (Theorem 14) equals the upper bound $\overline{C}$ (Theorem 15), and therefore the capacity of the channel in the present of feedback, which we denote as $C^{(f)}$, is given by $\lim_{N\to\infty} C_N^{(f)}$, where $C_N^{(f)}$ satisfies

$$\begin{aligned} C_N^{(f)} &= \frac{1}{N} \max_{Q(x^N \| z^{N-1})} I(X^N \to \{Y^N, L^N\}) \\ &\overset{(a)}{=} \frac{1}{N} \max_{Q(x^N \| z^{N-1})} \sum_{i=1}^{N} I(X^i; Y_i, S_i | Y^{i-1}, S^{i-1}) \\ &= \frac{1}{N} \max_{Q(x^N \| z^{N-1})} \sum_{i=1}^{N} H(Y_i, S_i | Y^{i-1}, S^{i-1}) \\ &\quad - H(Y_i, S_i | Y^{i-1}, S^{i-1}, X^i) \\ &\overset{(b)}{=} \frac{1}{N} \max_{Q(x^N \| z^{N-1})} \sum_{i=1}^{N} H(Y_i, S_i | Y^{i-1}, S^{i-1}) \\ &\quad - H(Y_i, S_i | S_{i-1}, X_i) \\ &\overset{(c)}{\le} \frac{1}{N} \max_{Q(x^N \| z^{N-1})} \sum_{i=1}^{N} H(Y_i, S_i | S_{i-1}) \\ &\quad - H(Y_i, S_i | S_{i-1}, X_i) \\ &\overset{(d)}{=} \frac{1}{N} \max_{\{Q(x_i | s_{i-1})\}} \sum_{i=1}^{N} I(Y_i, S_i; X_i | S_{i-1}). \end{aligned} \tag{117}$$

Equality (a) follows by replacing L_i with S_i according to the communication setting. Equality (b) follows from the FSC property. Inequality (c) holds because conditioning reduces entropy. Equality (d) holds because maximizing over the set of causal conditioning probability $Q(x^N \| z^{N-1})$ is the same as maximizing over the set of probabilities $\{Q(x_i | s_{i-1})\}_{i=1}^N$, as shown in the following argument. The sum $\sum_{i=1}^N I(Y_i, S_i; X_i | S_{i-1})$ is determined uniquely by the sequence of probabilities $\{P(y_i, s_i, x_i, s_{i-1}\}_{i=1}^N$. Let us prove by induction that this sequence of probabilities is determined by $\{Q(x_i | x^{i-1}, y^{i-1}, s^{i-1})\}_{i=1}^N$ only through $\{Q(x_i | s_{i-1})\}_{i=1}^N$. For $i = 1$, we have

$$P(y_1, s_1, x_1, s_0) = P(s_0) Q(x | s_0) p(y_1, s_1 | x_1, s_0). \tag{118}$$

Since $P(s_0)$ and $P(y_1, s_1 | x_1, s_0)$ are determined by the channel properties, the input distribution to the channel can influence only the term $Q(x | s_0)$. Now, let us assume that the argument is true for $i-1$ and let us prove it for i.

$$P(y_i, s_i, x_i, s_{i-1}) = P(s_{i-1}) Q(x_i | s_{i-1}) P(y_i, s_i | x_i, s_{i-1}). \tag{119}$$

The term $P(s_{i-1})$ is the same under both sequences of probabilities because of the assumption that the argument holds for $i-1$. The term $P(y_i, s_i | x_i, s_{i-1})$ is determined by the channel, so the only term influenced by the input distribution is $Q(x_i | s_{i-1})$. This proves the validity of the argument for all i and, consequently, the equality (d).

Inequality (117) proves that the achievable rate, when there is feedback and state information, cannot exceed

$$\lim_{N\to\infty} \frac{1}{N} \max_{\{Q(x_i | s_{i-1})\}} \sum_{i=1}^{N} I(Y_i, S_i; X_i | S_{i-1}).$$

Now let us prove that if the state of the channel is known at the encoder and the decoder and there is no feedback, we can achieve this rate. For this setting we denote the capacity as $C^{(nf)}$ and as in the case of feedback, the capacity does not depend on the initial state and is given as $\lim_{N\to\infty} C_N^{(nf)}$, where $C_N^{(nf)}$ satisfies

$$\begin{aligned} C_N^{(nf)} &\overset{(b)}{=} \frac{1}{N} \max_{Q(x^N \| s^{N-1})} \sum_{i=1}^{N} H(Y_i, S_i | Y^{i-1}, S^{i-1}) \\ &\quad - H(Y_i, S_i | S_{i-1}, X_i) \\ &\overset{(c)}{\ge} \frac{1}{N} \max_{\{Q(x_i | s_{i-1})\}} \sum_{i=1}^{N} H(Y_i, S_i | Y^{i-1}, S^{i-1}) \\ &\quad - H(Y_i, S_i | S_{i-1}, X_i) \\ &\overset{(d)}{=} \frac{1}{N} \max_{\{Q(x_i | s_{i-1})\}} \sum_{i=1}^{N} H(Y_i, S_i | S_{i-1}) \\ &\quad - H(Y_i, S_i | S_{i-1}, X_i) \\ &= \frac{1}{N} \max_{\{Q(x_i | s_{i-1})\}} \sum_{i=1}^{N} I(Y_i, S_i; X_i | S_{i-1}) \\ &\overset{(e)}{\ge} C_N^{(f)}. \end{aligned} \tag{120}$$

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

Equality (b) in (120) follows by the same sequence of equalities that lead to step (b) in (117), and replacing Z^i with S^i. Equality (b) follows from the FSC property. Inequality (c) holds because we restrict the range of probabilities over which the maximization is performed. Equality (d) holds because under an input distribution $Q(x_i|s_{i-1})$, we have the following Markov chain: $(Y_i, S_i) - S_{i-1} - (Y^{i-1}, S^{i-2})$. Inequality (e) holds due to (117).

Taking the limit $N \to \infty$ on both sides of (120) shows that $C^{(nf)} \geq C^{(f)}$. Since trivially also $C^{(nf)} \leq C^{(f)}$ we have $C^{(nf)} = C^{(f)}$. □

ACKNOWLEDGMENT

The authors would like to thank Thomas Cover, Paul Cuff, Gerhard Kramer, Amos Lapidoth, Taesup Moon, and Sekhar Tatikonda for helpful discussions, and are indebted to Young-Han Kim for suggesting a simple proof of Lemma 4. The authors are also grateful to the reviewer who pointed out a few errors in a previous version of the paper, and for his many helpful comments.

REFERENCES

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

[2] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 8–19, Sep. 1956.

[3] L. Breiman, D. Blackwell, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 29, pp. 1209–1220, 1958.

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

[5] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303–305.

[6] H. Marko, "The bidirectional communication theory- A generalization of information theory," *IEEE Trans. Commun.*, vol. COM-21, no. 12, pp. 1335–1351, Dec. 1973.

[7] S. C. Tatikonda, "Control Under Communication Constraints," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[8] S. C. Tatikonda, "A Markov decision approach to feedback channel capacity," in *Proc. 44th IEEE Conf. Decision and Control, 2005 and 2005 European Control Conf., CDC-ECC*, Spain, Dec. 2005, pp. 3213–3218.

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

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

[11] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–789, Mar. 2005.

[12] T. Weissman and N. Merhav, "On competitive prediction and its relation to rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3185–3194, Dec. 2003.

[13] S. S. Pradhan, "Source coding with feedforward: Gaussian sources," in *Proc. 2004 Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 212.

[14] R. Venkataramanan and S. S. Pradhan, "Source coding with feedforward: Rate-distortion theorems and error exponents for a general source," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2154–2179, Jun. 2007.

[15] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," *IEEE Trans. Inf. Theory*, submitted for publication.

[16] G. Kramer, "Directed Information for Channels With Feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 1998.

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

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

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

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

[21] J. Massey, "Conservation of mutual and directed information," in *Proc. Int. Symp. Information Theory (ISIT-05)*, Adelaide, Australia, Sep. 2005, pp. 157–158.

[22] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[23] C. E. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, Berkeley, CA, 1961, pp. 611–614.

[24] A. Lapidoth and İ. E. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 44, pp. 973–983, 1998.

[25] F. Jelinek, "Indecomposable channels with side information at the transmitter," *Inf. Contr.*, vol. 8, pp. 36–55, 1965.

[26] R. Ash, *Information Theory*. New York: Wiley, 1965.

[27] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.

[28] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[29] S. C. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, submitted for publication.

[30] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. 54, pp. 1488–1499, Apr. 2008.

[31] H. H. Permuter, P. W. Cuff, B. Van-Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," in *Proc. Allerton Conf. Communications, Control, and Computing*, Monticello, IL, Sep. 2006, pp. 3150–3165.

[32] D. Blackwell, "Information theory," in *Modern Mathematics for the Engineer: Second Series*. New York: McGraw-Hill, 1961, pp. 183–193.

Haim Henry Permuter (S'08) 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.

Between 1997 and 2004, he was an officer at a research and development unit of the Israeli Defense Forces. He is currently working toward the Ph.D. degree in electrical engineering at Stanford University, Stanford, CA.

Mr. Permuter is a recipient of the Fullbright Fellowship and the Stanford Graduate Fellowship (SGF).

Tsachy Weissman (S'99–M'02–SM'07) received the B.Sc. and Ph.D. degrees in electrical engineering from the Technion–Israel Institute of Technology, Haifa, Israel.

He has held postdoctoral appointments with the Statistics Department at Stanford University and at Hewlett-Packard Laboratories. Currently, he is with the Departments of Electrical Engineering at Stanford University and at the Technion. His research interests span information theory and its applications, and statistical signal processing. His papers thus far have focused mostly on data compression, communications, prediction, denoising, and learning. He is also inventor or coinventor of several patents in these areas and has been involved in several high-tech companies as a researcher or advisor.

Prof. Weissman's recent prizes include the NSF CAREER award and a Horev fellowship for Leaders in Science and Technology. He has been a Robert N. Noyce Faculty Scholar of the School of Engineering at Stanford, and is a recipient of the 2006 IEEE joint IT/COM Societies best paper award.

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.

Andrea J. Goldsmith (S'90–M'93–SM'99–F'05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Berkeley.

She is a Professor of Electrical Engineering at Stanford University, Stanford, CA, and was previously an Assistant Professor of Electrical Engineering at Caltech, Pasadena, CA. She has also held industry positions at Maxim Technologies and at AT&T Bell Laboratories, and is cofounder and CTO of Quantenna Communications, Inc. Her research includes work on capacity of wireless channels and networks, wireless communication and information theory, energy-constrained wireless communications, wireless communications for distributed control, and cross-layer design of wireless networks. She is author of the book *Wireless Communications* and coauthor of the book *MIMO Wireless Communications*, both published by Cambridge University Press.

Dr. Goldsmith is a Fellow of Stanford. She has received several awards for her research, including the National Academy of Engineering Gilbreth Lectureship, the Alfred P. Sloan Fellowship, the Stanford Terman Fellowship, the National Science Foundation CAREER Development Award, and the Office of Naval Research Young Investigator Award. In addition, she was a corecipient of the 2005 IEEE Communications Society and Information Theory Society joint paper award. She currently serves as Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY and as Editor for the *Journal on Foundations and Trends in Communications and Information Theory and in Networks*. She previously served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and for the IEEE WIRELESS COMMUNICATIONS MAGAZINE, as well as Guest Editor for several IEEE journal and magazine special issues. She participates actively in committees and conference organization for the IEEE Information Theory and Communications Societies and is an elected member of the Board of Governors for both societies. She is the President of the IEEE Information Theory Society, a distinguished lecturer for the IEEE Communications Society, and was the Technical Program Co-Chair for the 2007 IEEE International Symposium on Information Theory. She also founded the student committee of the IEEE Information Theory Society and is an inaugural recipient of Stanford's postdoc mentoring award.

Authorized licensed use limited to: BEN GURION UNIVERSITY. Downloaded on May 6, 2009 at 04:48 from IEEE Xplore. Restrictions apply.