A Coding Theorem for a Class of Stationary Channels With Feedback

Young-Han Kim, *Member, IEEE*

***Abstract*—A coding theorem is proved for a class of stationary channels with feedback in which the output $Y_n = f(X_{n-m}^n, Z_{n-m}^n)$ is the function of the current and past m symbols from the channel input X_n and the stationary ergodic channel noise Z_n. In particular, it is shown that the feedback capacity is equal to**

$$\lim_{n\to\infty} \sup_{p(x^n\|y^{n-1})} \frac{1}{n} I(X^n \to Y^n)$$

where $I(X^n \to Y^n) = \sum_{i=1}^n I\left(X^i; Y_i | Y^{i-1}\right)$ denotes the Massey directed information from the channel input to the output, and the supremum is taken over all causally conditioned distributions $p(x^n\|y^{n-1}) = \prod_{i=1}^n p\left(x_i | x^{i-1}, y^{i-1}\right)$. The main ideas of the proof are a classical application of the Shannon strategy for coding with side information and a new elementary coding technique for the given channel model *without feedback*, which is in a sense dual to Gallager's lossy coding of stationary ergodic sources. A similar approach gives a simple alternative proof of coding theorems for finite state channels by Yang–Kavčić–Tatikonda, Chen–Berger, and Permuter–Weissman–Goldsmith.

***Index Terms*—Capacity, channels with memory, coding theorem, directed information, ergodic decomposition, feedback, feedback capacity, Shannon strategy.**

I. Introduction

SHANNON [34] showed that the capacity C of a memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, operationally defined as supremum of all achievable rates [9, Sec. 7.5], is characterized by

$$C = \sup_{p(x)} I(X;Y). \tag{1}$$

When the channel has memory but still maintains certain ergodic properties, then (1) can be extended to the following multiletter expression:

$$C = \lim_{n\to\infty} \sup_{p(x^n)} \frac{1}{n} I(X^n; Y^n). \tag{2}$$

For example, Dobrushin [10] showed that the capacity formula (2) holds if the channel is *information stable*; see also Pinsker [33]. Further extensions and refinements of (2) with more general capacity formulas abound in the literature. For stationary channels, readers are referred to Gray and Ornstein [17], Kieffer [20], and the references therein. A general formula for the capacity is given by Verdú and Han [39] for arbitrary nonstationary channels that can be represented through a sequence of n-dimensional conditional distributions (even without any consistency requirement); see also Han [18].

For memoryless channels *with feedback*, it was again Shannon [35] who showed that feedback does not increase the capacity and hence that the feedback capacity is given by

$$C_{\mathrm{FB}} = C = \sup_{p(x)} I(X;Y). \tag{3}$$

As in the case of nonfeedback capacity (2), the question arises how to extend the feedback capacity formula (3) to channels with memory. The most natural candidate is the following multiletter expression with *directed information* introduced by Massey [26] in place of the usual mutual information in (2):

$$\begin{aligned} C_{\mathrm{FB}} &= \lim_{n\to\infty} \sup_{p(x^n\|y^{n-1})} \frac{1}{n} I(X^n \to Y^n) \\ &= \lim_{n\to\infty} \sup_{p(x^n\|y^{n-1})} \frac{1}{n} \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) \end{aligned} \tag{4}$$

where the supremum is taken over all n-dimensional *causally conditioned* probabilities

$$p(x^n\|y^{n-1}) = \prod_{i=1}^n p\left(x_i | x^{i-1}, y^{i-1}\right).$$

The main goal of this paper is to establish the validity of the feedback capacity formula (4) for a reasonably general class of channels with memory, in the simplest manner.

Massey [26] introduced the mathematical notion of directed information

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

and established its operational meaning by showing that the feedback capacity is upper bounded by the maximum normalized directed information, which can be in general tighter than the usual mutual information. He also showed that (4) reduces to (3) if the channel is memoryless, and to (2) if the channel is used without feedback. Kramer [23], [24] streamlined the notion of directed information further and explored many interesting properties; see also Massey and Massey [27].

For channels with certain structures, the validity of the feedback capacity formula (4) has been established implicitly. For example, Cover and Pombra [8] gives a multi-letter characterization of the Gaussian feedback capacity, and Alajaji [1]

Manuscript received January 7, 2007; revised November 18, 2007. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, June 2007.

The author is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: yhk@ucsd.edu).

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

Digital Object Identifier 10.1109/TIT.2008.917685

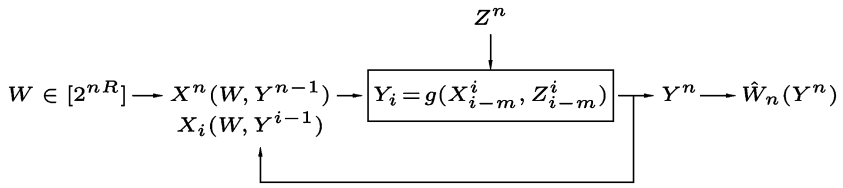

Fig. 1. Feedback communication channel $Y_i = g\left(X_{i-m}^i, Z_{i-m}^i\right)$.

characterizes the feedback capacity of discrete channels with additive noise—feedback does not increase the capacity of discrete additive channels when there is no input cost constraint. Both results can be recast in the form of directed information (see [8, eq. (52)] and [1, eq. (17)]). The notion of directed information in these contexts, however, has a rather limited role as an intermediate step in the proof of converse coding theorems. Indeed, the highlight of Cover–Pombra characterization is the asymptotic equipartition property of arbitrary nonstationary nonergodic Gaussian processes [8, Sec. V]; see also Pinsker [33]. (The case of discrete additive channel is trivial since the optimal input distribution is memoryless and uniform, and hence feedback does not increase the capacity.)

In a heroic effort [37], [38], Tatikonda attacked the general nonanticipatory channel with feedback by combining Verdú–Han formula for nonfeedback capacity, Massey directed information, and Shannon strategy for channel side information [36], as well as dynamic programming for Markov decision processes. As the cost of generality, however, it seems very difficult to establish a simple formula like (4) from the approaches taken therein. See, for example, [38, Ths. 7.3 and 7.5] for nontrivial sufficient conditions that are needed to establish the equivalence between a Verdú–Han-type formula and (4).

More recently, Yang, Kavčić, and Tatikonda [42] and Chen and Berger [6] studied special cases of finite-state channels, based on Tatikonda's framework. A finite-state channel [14, Sec. 4.6] is described by the conditional probability distribution

$$p(y_n, s_n | x_n, s_{n-1}) \tag{5}$$

where s_n denotes the channel state at time n. Using a different approach based on Gallager's proof of the nonfeedback capacity [14, Sec. 5.9], Permuter, Weissman, and Goldsmith [31] proved various coding theorems for finite-state channels with feedback that include *inter alia* the results of [42], [6] and establish the validity of (4) for indecomposable finite-state channels without intersymbol interference (i.e., the channel states evolve as an ergodic Markov chain, independent of the channel input).

As mentioned before, we strive to give a straightforward treatment of the feedback coding theorem. Towards this goal, this paper focuses on stationary nonanticipatory channels of the form

$$Y_n = g\left(X_{n-m}^n, Z_{n-m}^n\right). \tag{6}$$

In words, the channel output Y_n at time n is given as a time-invariant deterministic function of channel inputs $X_{n-m}^n = (X_{n-m}, X_{n-m+1}, \ldots, X_n)$ up to past m symbols and channel noises $Z_{n-m}^n = (Z_{n-m}, Z_{n-m+1}, \ldots, Z_n)$ up to past m symbols. We assume the noise process $\{Z_n\}_{n=1}^{\infty}$ is an arbitrary stationary ergodic process (without any mixing condition) independent of the message sent over the channel.

The channel model (6) is rather simple and physically motivated. Yet this channel model is general enough to include many important feedback communication models such as any additive noise fading channels with intersymbol interference and indecomposable finite-state channels without intersymbol interference.[1]

The channel (6) has finite input memory in the sense of Feinstein [11] and can be viewed as a finite-window sliding-block coder [16, Sec. 9.4] of input and noise processes (cf. primitive channels introduced by Neuhoff and Shields [29] in which the noise process is memoryless). Compared to the general finite-state channel model (5), in which the channel has infinite input memory but the channel noise is memoryless, our channel model (6) has finite input memory but the noise has infinite memory; recall that there is no mixing condition on the noise process $\{Z_n\}_{n=1}^{\infty}$. Thus, the finite-state channel model (5) and the finite sliding-block channel model (6) nicely complement each other.

Our main result is to show that the feedback capacity C_{FB} of the channel (6) is characterized by (4). More precisely, we consider a communication problem depicted in Fig. 1. Here one wishes to communicate a message index $W \in [2^{nR}] := \{1, 2, \ldots, 2^{nR}\}$ over the channel

$$Y_i = \begin{cases} \emptyset, & i = 1, 2, \ldots, m \\ g\left(X_{i-m}^i, Z_{i-m}^i\right), & i = m+1, m+2, \ldots \end{cases} \tag{7}$$

where the time-i channel output Y_i on the output alphabet $\mathcal{Y}$ is given by a deterministic map $f : \mathcal{X}^m \times \mathcal{Z}^m \to \mathcal{Y}$ of the current and past m channel inputs X_{i-m}^i on the input alphabet $\mathcal{X}$ and the current and past m channel noises Z_{i-m}^i on the noise alphabet $\mathcal{Z}$. We assume that the channel noise process $\{Z_i\}_{i=1}^{\infty}$ is stationary ergodic and is independent of the message W. The initial values of $Y_1, \ldots, Y_m$, coming from an unspecified initial condition $\left(X_{-m+1}^0, Z_{-m+1}^0\right)$, are set arbitrarily, which we denote by $\emptyset$. Throughout this paper we adopt the convention that the value of $\emptyset$ is taken as a fixed constant for any operation performed on it. For example, we have $f(\emptyset)$ as a constant, $f(\emptyset, x)$ as a function of x only, and $I(\emptyset; X) = 0$. Under this convention, both the encoder and the decoder simply ignore $Y_1, \ldots, Y_m$, which does not affect the long-term performance of communication over the channel.

We specify a $(2^{nR}, n)$ feedback code with the encoding maps

$$X_i : [2^{nR}] \times \mathcal{Y}^{i-1} \to \mathcal{X}, \quad i = 1, 2, \ldots, n$$

[1]A notable exception is a famous finite-state channel called the "trapdoor channel" introduced by Blackwell [3], the feedback capacity of which is established in [30].

that result in codewords

$$X^n(W, Y^{n-1}) = (X_1(W), X_2(W, Y_1), \ldots, X_n(W, Y^{n-1}))$$

and the decoding map

$$\hat{W}_n : \mathcal{Y}^n \to [2^{nR}].$$

The probability of error $P_e^{(n)}$ is defined as

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \Pr\left\{\hat{W}_n(Y^n) \neq w | W = w\right\} = \Pr\left\{\hat{W}_n(Y^n) \neq W\right\}$$

where the message W is uniformly distributed over $[2^{nR}]$ and is independent of $\{Z_i\}_{i=1}^{\infty}$. We say that the rate R is achievable if there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ as $n \to \infty$. The feedback capacity C_{FB} is defined as the supremum of all achievable rates. The nonfeedback capacity C is defined similarly, with codewords $X^n(W) = (X_1(W), \ldots, X_n(W))$ restricted to be a function of the message W only.

We will prove the following result in Section III.

Theorem 1: The feedback capacity C_{FB} of the channel (7) is given by

$$C_{\mathrm{FB}} = \lim_{n\to\infty} \sup_{p(x^n \| y^{n-1})} \frac{1}{n} I(X^n \to Y^n). \tag{8}$$

Our development has two major ingredients. First, we revisit the communication problem over the same channel *without feedback* in Section II and prove that the nonfeedback capacity is given by

$$C = \lim_{n\to\infty} \sup_{p(x^n)} \frac{1}{n} I(X^n; Y^n).$$

Roughly speaking, there are three flavors for the achievability proof of nonfeedback capacity theorems in the literature. The first one is Shannon's original argument [34] based on random codebook generation, asymptotic equipartition property, and joint typicality decoding, which was made rigorous by Forney [13] and Cover [7], and now is used widely in coding theorems for memoryless networks [9, Ch. 15]. This approach, however, does not easily generalize to channels with memory. The second flavor is the method of random coding exponent by Gallager [15], which was later applied to finite-state channels [14, Sec. 5.9]. This approach is perhaps the simplest one for the analysis of general finite-state channels and has been adapted by Lapidoth and Telatar [25] for compound finite-state channels and by Permuter *et al.* [31] for finite-state channels with feedback.

The third and the least intuitive approach is Feinstein's fundamental lemma [12]. This is the most powerful and general method of the three, and has been applied extensively in the literature, say, from Khinchin [19] to Gray [16], Verdú and Han [39], and Tatikonda [37], [38].

Our approach is somewhat different from these three usual approaches. We use the strong typicality (relative frequency) decoding for n-dimensional super letters. The encoding is based on block ergodic decomposition of Nedoma [28], which uses a long codeword on the n-letter super alphabet, constructed as a concatenation of n shorter codewords. While each short codeword and the corresponding output fall into their own ergodic mode, the long codeword as a whole maintains the ergodic behavior. To be fair, codebook construction of this type is far from new in the literature, and our method is intimately related to the one used by Gallager [14, Sec. 9.8] and Berger [2, Sec. 7.2] for lossy compression of stationary ergodic sources. Indeed, when the channel (6) has zero memory ($m = 0$), then the role of the input for our channel coding scheme is equivalent to the role of the covering channel for Gallager's source coding scheme.

Equipped with this coding method for nonfeedback sliding-block coder channels (6), the extension to the feedback case is relatively straightforward. The basic ingredient for this extension is the Shannon strategy for channels with causal side information at the transmitter [36], previously employed by Tatikonda [37], [38] and Permuter *et al.* [31]. As a matter of fact, Shannon himself observed that the major utility of his result is feedback communication. Following is the first sentence of [36]:

> Channels with feedback from the receiving to the transmitting point are a special case of a situation in which there is additional information available at the transmitter which may be used as an aid in the forward transmission system.

As observed by Caire and Shamai [5, Proposition 1], the causality has no cost when the transmitter and the receiver share the same side information—in our case, the past input (if decoded faithfully) and the past output (received from feedback)—and the transmission can fully utilize this side information as if it were known *a priori*. Indeed, the capacity of a memoryless state-dependent channel $p(y|x, s)$ with memoryless state sequence S^n known *causally* at both the transmitter and the receiver is given by

$$C = \max_{p(x|s)} I(X; Y|S) \tag{9}$$

which is equal to the capacity of the same channel when the state sequence S^n is known *noncausally* at the transmitter and the receiver [5, Proposition 1]. (See also Lemma 9 in the Appendix.)

As depicted in Fig. 2, given n and $i = 1, 2, \ldots, n$, a heuristic coding scheme can be constructed that uses the channel input sequence $(X_i, X_{i+n}, X_{i+2n}, \ldots)$ and the output sequence $\left(Y_i^n, Y_{i+n}^{2n}, Y_{i+2n}^{3n}, \ldots\right)$ with the past input and output symbols $\left(X^{i-1}, Y^{i-1}, X_{n+1}^{n+i-1}, Y_{n+1}^{n+i-1}, X_{2n+1}^{2n+i-1}, Y_{2n+1}^{2n+i-1}, \ldots\right)$ as causal side information available at both the transmitter and the receiver. From (9), one could argue that this coding scheme achieves the rate R_i for the ith input symbol in the length-n super symbol as

$$R_i = \max_{p(x_i|x^{i-1}, y^{i-1})} I\left(X_i; Y_i^n | X^{i-1}, Y^{i-1}\right) \tag{10}$$

for each i, given $p(x^{i-1}, y^{i-1})$, and hence the total achievable rate becomes

$$R = \max_{p(x^n \| y^{n-1})} \sum_{i=1}^{n} I(X_i; Y_i^n | X^{i-1}, Y^{i-1})$$

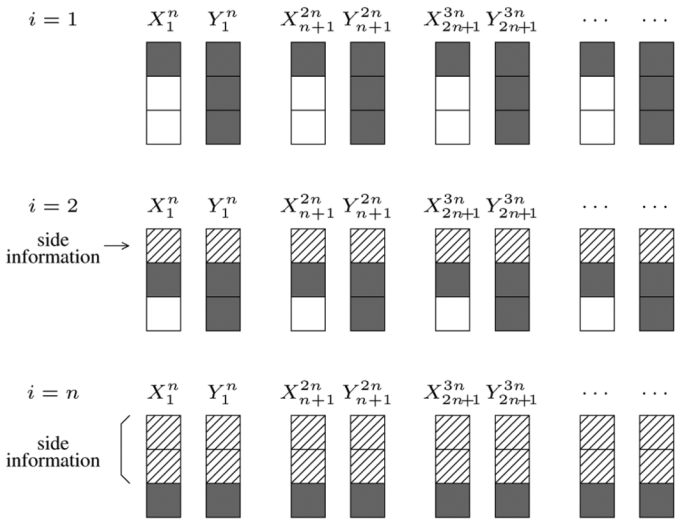

Fig. 2. Heuristic coding scheme to achieve $R_i = I\left(X_i; Y_i^n | X^{i-1}, Y^{i-1}\right)$: $i = 1, 2, 3$, $n = 3$. The solid gray boxes denote active input symbols (codewords) and output symbols for each stage i, and the stripe shaded boxes denote past input and output symbols used as side information. For $i = 1$, one can "achieve" $I(X_1; Y_1, Y_2, Y_3)$ by encoding over $(X_1, X_4, X_7, \ldots)$ and decoding with all three-letter output sequences. For $i = 2$, one can use past input–output pairs as causal side information, encode over copies of X_2, and decode with all two-letter output sequences to "achieve" $I(X_2; Y_2, Y_3 | X_1, Y_1)$. Similarly, for $i = n = 3$ one can "achieve" $I(X_3; Y_3 | X_1, X_2, Y_1, Y_2)$.

per n transmissions. Now a simple algebra shows that this rate is equal to the maximum directed information as

$$\begin{aligned}
&\sum_{i=1}^{n} I\left(X_i; Y_i^n | X^{i-1}, Y^{i-1}\right) \\
&\quad = \sum_{i=1}^{n} \sum_{j=i}^{n} I\left(X_i; Y_j | X^{i-1}, Y^{j-1}\right) \\
&\quad = \sum_{j=1}^{n} \sum_{i=1}^{j} I\left(X_i; Y_j | X^{i-1}, Y^{j-1}\right) \\
&\quad = \sum_{j=1}^{n} I\left(X^j; Y_j | Y^{j-1}\right) \\
&\quad = I(X^n \to Y^n). \qquad (11)
\end{aligned}$$

This argument, although intuitively appealing and apparently applicable to general channel models, is not completely rigorous, however. The channel $p\left(y_i^n | x_i, (x^{i-1}, y^{i-1})\right)$ and its copies sampled every n times are not "ergodic" enough, not to mention memoryless, so one cannot jump to conclude that $I\left(X_i; Y_i^n | X^{i-1}, Y^{i-1}\right)$ is achievable as in the case of memoryless channels with state (9).

Therefore, we will take more careful steps to prove Theorem 1, by first proving the achievability of $\frac{1}{n} I(U^n; Y^n)$ for all auxiliary random variables U^n and Shannon strategies $X_i\left(U_i, X^{i-1}, Y^{i-1}\right)$, $i = 1, \ldots, n$, and then showing that $I(U^n; Y^n)$ reduces to $I(X^n \to Y^n)$ via pure algebra.

The Section II revisits nonfeedback communication over the same channel model and proves the coding theorem, which will be crucial to the proof of the feedback coding theorem in Section III. Technical lemmas are delegated to the Appendix.

II. Nonfeedback Coding Theorem Revisited

This section is devoted to the proof of the following result.

Theorem 2: The nonfeedback capacity C of the stationary channel

$$Y_i = \begin{cases} \emptyset, & i = 1, \ldots, m, \\ g\left(X_{i-m}^i, Z_{i-m}^i\right), & i = m+1, m+2, \ldots \end{cases} \qquad (12)$$

with the input X_i and the stationary ergodic noise process $\{Z_i\}_{i=1}^{\infty}$ depicted in Fig. 1 is given by

$$\begin{aligned}
C &= \lim_{n \to \infty} C_n \\
&= \lim_{n \to \infty} \sup_{p(x^n)} \frac{1}{n} I(X^n; Y^n). \qquad (13)
\end{aligned}$$

Revisiting and proving the nonfeedback coding theorem is rewarding for two reasons. First, our proof is somewhat different from the usual techniques and hence is interesting on its own. (See Section I for the discussion on conventional achievability proofs of nonfeedback capacity theorems.) Second, our exercise here will lead to a straightforward proof of the feedback coding theorem in Section III.

Before we go into the formal proof, observe that we can decompose the achievable rate as

$$\begin{aligned}
R &= \frac{1}{n} \sum_{i=1}^{n} R_i \\
&= \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y^n | X^{i-1}).
\end{aligned}$$

As in the heuristic coding scheme described in Fig. 2 of Section I, the rate

$$R_i = I\left(X_i; Y^n | X^{i-1}\right) \qquad (14)$$

corresponds to the rate achievable for copies of the channel $p\left(y^n | x_i, x^{i-1}\right)$ with the input X_i, the output Y^n, and the noncausal side information X^{i-1}. (Compare this to the feedback case (10), in which there is additional causal side information Y^{i-1}.) Since the channel is not memoryless or even ergodic, we cannot prove the achievability of (14) directly by coding over $(X_i, X_{n+i}, X_{2n+i}, \ldots)$. Instead, we "synchronize" the transmission by concatenating short codewords with pauses between them, so that the overall long codeword and the channel have an ergodic behavior. More details follow.

Proof of Theorem 2: We first note that the capacity expression (13) is well defined because nC_n is superadditive (i.e., $nC_n + kC_k \le (n+k)C_{n+k}$), which implies that the limit exists and

$$\lim_{n \to \infty} C_n = \sup_{n \ge 1} C_n.$$

The converse follows immediately from Fano's inequality [9, Lemma 7.9.1]. For any sequence of $(2^{nR}, n)$ codes $(X^n(W), \hat{W}(Y^n))$ with message W drawn uniformly over $[2^{nR}]$, if

$$P_e^{(n)} = \Pr(W \ne \hat{W}) \to 0$$

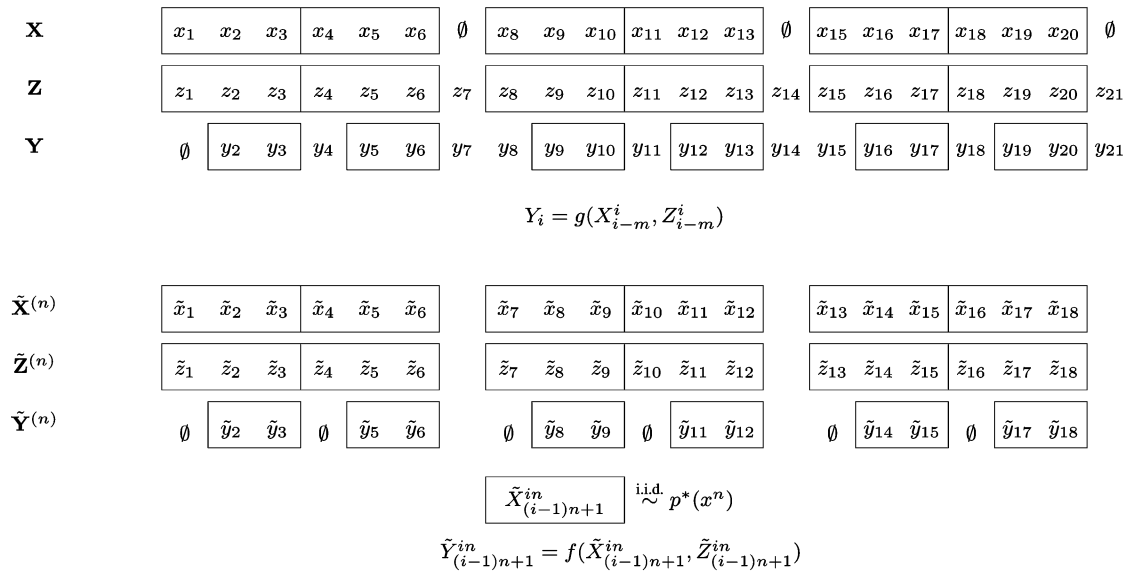

Fig. 3. Input, noise, and output sequences: $n = 3$, $L = 2$, $m = 1$. A comparison can be made to the code construction by Gallager [14, Fig. 9.8.1] for lossy coding of arbitrary stationary ergodic sources, in which test channels (covering codewords) play the role of input codewords here.

then we must have

$$\begin{aligned} nR &\le I(W; Y^n) + n\epsilon_n \\ &\le I(X^n; Y^n) + n\epsilon_n \end{aligned}$$

where $\epsilon_n \to 0$ as $n \to \infty$.

For the achievability, it suffices to show that there exists a sequence of codes that achieves C_n for each $n > m$. (Recall $C_1 = \cdots = C_m = 0$.) Without loss of generality, we assume that the alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are finite. Otherwise, we can partition the space for each n and $\epsilon > 0$ such that

$$\max_{p([x]^n)} \frac{1}{n} I([X]^n; [Y]^n) \ge C_n - \epsilon$$

and prove the achievability on this partitioned space.

Codebook Generation. Fix $n > m$ and let $p^*(x^n)$ denote the input distribution that achieves C_n. For each $L = 1, 2, \ldots$, let $k = k(L, n) = Ln^2 + n$. We generate a sequence[2] of $(2^{kR}, k)$ codes $X^k(w)$ as depicted in Fig. 3.

For each $w \in [2^{kR}]$, generate a codeword $\tilde{\mathbf{X}}^{(n)}(w) = \tilde{X}^{Ln^2}(w)$ of length Ln on the n-letter super alphabet $\mathcal{X}^n$ independently according to

$$p\left(\tilde{x}^{Ln^2}\right) = \prod_{i=1}^{Ln} p^*\left(x_{(n-1)i+1}^{ni}\right).$$

We exhibit the 2^{kR} codewords as the rows of a matrix:

$$\mathcal{C} = \begin{bmatrix} \tilde{X}_1^n(1) & \tilde{X}_{n+1}^{2n}(1) & \cdots & \tilde{X}_{Ln(n-1)+1}^{Ln^2}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_1^n(2^{kR}) & \tilde{X}_{n+1}^{2n}(2^{kR}) & \cdots & \tilde{X}_{Ln(n-1)+1}^{Ln^2}(2^{kR}) \end{bmatrix}.$$

[2]This gives only a subsequence of $(2^{kR}, k)$ codes. But we can easily interpolate to $Ln^2 + n < k < (L+1)n^2 + n$ without any rate loss, since $(Ln^2 + n)/((L+1)n^2 + n) \to 1$ as $L \to \infty$.

Each entry in this matrix is generated independent and identically distributed (i.i.d.) according to $p^*(x^n)$.

Using the construction as depicted in Fig. 3 (see also Lemma 6 in the Appendix[3]), the *actual* codewords $\mathbf{X}(w) = X^k(w)$, $w = 1, 2, \ldots, 2^{nR}$, which will be transmitted over the channel, are generated from $\tilde{\mathbf{X}}^{(n)}(w) = \tilde{X}^{Ln^2}(w)$ as follows:

$$X_{(i-1)Ln+i}^{iLn+i} = \left(\tilde{X}_{(i-1)Ln+1}^{iLn},\ \emptyset\right), \quad i = 1, 2, \ldots, n.$$

In other words, X^k is a verbatim copy of $\tilde{X}^{Ln^2}$ with fixed symbol $\emptyset$ separating the subsequences of length Ln.

Encoding. If $W = w$, the transmitter sends the codeword $\mathbf{X}(w) = X^k(w)$ over the channel.

Decoding. Upon receiving the sequence $\mathbf{Y} = Y^k$, the receiver forms the sequence $\tilde{\mathbf{Y}}^{(n)} = \tilde{Y}^{Ln^2}$ of length Ln in the n-letter super alphabet $\mathcal{Y}^n$, as depicted in Fig. 3:

$$\tilde{Y}_{(i-1)n+1}^{in} = \begin{cases} \left(\emptyset^m, Y_{(i-1)n+m+1}^{in}\right), & i = 1, \ldots, L \\ \left(\emptyset^m, Y_{(i-1)n+m+2}^{in+1}\right), & i = L+1, \ldots, 2L \\ \quad\vdots & \\ \left(\emptyset^m, Y_{(i-1)n+m+n-1}^{in+n}\right), & i = L(n-1)+1, \ldots, Ln. \end{cases}$$

Now we consider $\tilde{\mathbf{X}}^{(n)} = \tilde{X}^{Ln^2}$ and $\tilde{\mathbf{Y}}^{(n)} = \tilde{Y}^{Ln^2}$ as sequences of length Ln on the super alphabet $\mathcal{X}^n \times \mathcal{Y}^n$. The receiver declares that the message $\hat{W}$ was sent if there is a unique $\hat{W}$ such that

$$(\tilde{\mathbf{X}}^{(n)}(\hat{W}), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}(X^n, Y^n)$$

that is, $(\tilde{\mathbf{X}}^{(n)}(\hat{W}), \tilde{\mathbf{Y}}^{(n)})$ is jointly typical with respect to the joint distribution $p(x^n, y^n)$ specified by $p^*(x^n)p(z^n)$ and the definition of the channel (12). Otherwise, an error is declared.

[3]All technical lemmas referenced in Sections II and III are located in the Appendix.

Analysis of the probability of error. Without loss of generality, we assume $W = 1$ was sent. We define the following events:

$$E_w = \left\{(\tilde{\mathbf{X}}^{(n)}(w), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}(X^n, Y^n)\right\}$$

for $w \in [2^{kR}]$, where E_w is the event that the codeword $\tilde{\mathbf{X}}^{(n)}(w)$ and $\tilde{\mathbf{Y}}^{(n)}$ are jointly typical. By Bonferonni's inequality, we have

$$\begin{aligned}\Pr(\hat{W} \neq W) &= \Pr(\hat{W} \neq W | W = 1) \\ &= \Pr\left(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{kR}}\right) \\ &\leq \Pr\left(E_1^c\right) + \sum_{w=2}^{2^{kR}} \Pr(E_w).\end{aligned}$$

In order to bound $\Pr\left(E_1^c\right)$, we define $\tilde{\mathbf{Z}}^{(n)}$ as the nth-order super process of length Ln on the super alphabet $\mathcal{Z}^n$ constructed from the noise process $\{Z_i\}_{i=1}^\infty$ as in Fig. 3 (see also Lemma 6.) Since $\tilde{\mathbf{X}}^{(n)}(1)$ is blockwise i.i.d. $\sim p^*(x^n)$ and independent of $\mathbf{Z} = \{Z_1\}_{i=1}^\infty$, we have from Lemma 7 that

$$\Pr\left((\tilde{\mathbf{X}}^{(n)}(1), \tilde{\mathbf{Z}}^{(n)}) \in A_\epsilon^{*(Ln)}(X^n, Z^n)\right) \to 1 \text{ as } \quad L \to \infty.$$

Furthermore, $\tilde{\mathbf{Y}}^{(n)}$ is the n-letter blockwise function of $(\tilde{\mathbf{X}}^{(n)}(1), \tilde{\mathbf{Z}}^{(n)})$, that is,

$$\tilde{Y}_{(i-1)n+1}^{in} = f\left(\tilde{X}_{(i-1)n+1}^{in}(1), \tilde{Z}_{(i-1)n+1}^{in}\right)$$

with the time-invariant function f induced by the channel function g in (12). Thus by Lemma 4

$$\Pr\left((\tilde{\mathbf{X}}^{(n)}(1), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}(X^n, Y^n)\right) \to 1 \text{ as } L \to \infty$$

and

$$\Pr(E_1^c) \leq \epsilon \text{ for } L \text{ sufficiently large.}$$

On the other hand, the joint typicality of $(\tilde{\mathbf{X}}^{(n)}(w), \tilde{\mathbf{Y}}^{(n)})$ implies the typicality of $\tilde{\mathbf{Y}}^{(n)}$; see Lemma 4. Hence, by Lemma 8 we have for each $w \neq 1$

$$\begin{aligned}\Pr(E_w) &= \Pr\left((\tilde{\mathbf{X}}^{(n)}(w), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}\right) \\ &= \sum_{\tilde{\mathbf{y}}^{(n)} \in A_\epsilon^{*(Ln)}} \Pr\left((\tilde{\mathbf{X}}^{(n)}(w), \tilde{\mathbf{y}}^{(n)}) \in A_\epsilon^{*(Ln)}\right) \\ &\leq 2^{-Ln\left(I\left(X^n; Y_{m+1}^n\right) - \delta\right)}\end{aligned}$$

where $\delta \to 0$ as $\epsilon \to 0$. Consequently,

$$\begin{aligned}\Pr(\hat{W} \neq W) &\leq \Pr\left(E_1^c\right) + \sum_{w=2}^{2^{kR}} \Pr(E_w) \\ &\leq \epsilon + 2^{kR} 2^{-Ln\left(I\left(X^n; Y_{m+1}^n\right) - \delta\right)} \\ &\leq 2\epsilon\end{aligned}$$

if L is sufficiently large and

$$kR < Ln\left(I\left(X^n; Y_{m+1}^n\right) - \delta\right)$$

or equivalently,

$$R < \frac{Ln}{Ln^2 + n}\left(I\left(X^n; Y_{m+1}^n\right) - \delta\right).$$

Since ϵ and δ can be made arbitrarily small and $(Ln)/(Ln^2 + n) \to 1/n$ as $L \to \infty$, we have a sequence of $(2^{kR}, k)$ codes with $P_e^{(k)} \to 0$ if

$$R < \frac{1}{n} I\left(X^n; Y_{m+1}^n\right) = \frac{1}{n} I(X^n; Y^n) = C_n.$$

This completes the proof of achievability. □

III. Proof of Theorem 1

Recall the channel model:

$$Y_i = \begin{cases} \emptyset, & i = 1, \ldots, m, \\ g\left(X_{i-m}^i, Z_{i-m}^i\right), & i = m+1, m+2, \ldots \end{cases} \tag{15}$$

with the input X_i and the stationary ergodic noise process $\{Z_i\}_{i=1}^\infty$ depicted in Fig. 1. We prove that the feedback capacity is given by

$$\begin{aligned}C_{\mathrm{FB}} &= \lim_{n\to\infty} C_{\mathrm{FB},n} \\ &= \lim_{n\to\infty} \sup_{p(x^n \| y^{n-1})} \frac{1}{n} I(X^n \to Y^n)\end{aligned} \tag{16}$$

where the supremum is over all causally conditioned distributions

$$p(x^n \| y^{n-1}) = \prod_{i=1}^n p\left(x_i | x^{i-1}, y^{i-1}\right).$$

We will combine the coding technique developed in the previous section with the Shannon strategy for channels with side information, in particular, Lemma 10.

That the limit in (16) is well defined follows from the superadditivity of $nC_{\mathrm{FB},n}$, which is an easy consequence of the stationarity of the process $\{Z_i\}_{i=1}^\infty$ and the definition of the channel model (15), in particular, $Y_i = \emptyset,\ i = 1, \ldots, m$. Thus

$$C_{\mathrm{FB}} = \lim_{n\to\infty} C_{\mathrm{FB},n} = \sup_{n \geq 1} C_{\mathrm{FB},n}.$$

The converse was proved by Massey [26, Th. 3]. We repeat the proof here for completeness. For any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$, we have from Fano's inequality

$$\begin{aligned}nR &\leq I(W; Y^n) + n\epsilon_n \\ &= \sum_{i=1}^n I(W; Y_i | Y^{i-1}) + n\epsilon_n \\ &= \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) + n\epsilon_n \\ &= I(X^n \to Y^n) + n\epsilon_n\end{aligned} \tag{17}$$

where $\epsilon_n \to 0$ as $n \to \infty$. Here (17) follows from the codebook structure $X_i(W, Y^{i-1})$ and the Markovity $W \to (X^i, Y^{i-1}) \to Y_i$.

For the achievability, we show that there exists a sequence of codes that achieve $C_{\mathrm{FB},n}$ for each n. As before, we assume that

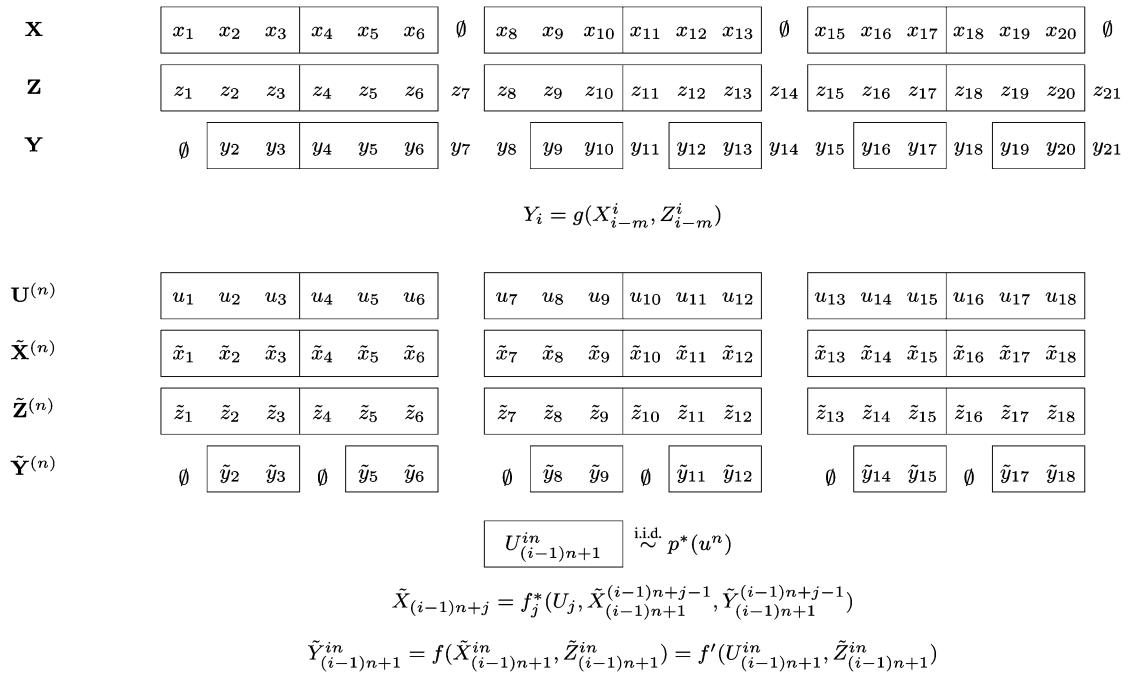

Fig. 4. Code, input, noise, and output sequences: $n = 3$, $L = 2$, $m = 1$.

the alphabets are finite. In the light of Lemma 10, it suffices to show that

$$C'_{FB,n} = \max_{p(u^n), x_i = f(u_i, x^{i-1}, y^{i-1})} I(U^n; Y^n) \qquad (18)$$

is achievable, where the auxiliary random variables U_i has the cardinality bounded by $|\mathcal{U}_i| \le |\mathcal{X}|^i |\mathcal{Y}|^{i-1}$, and the maximization is over all joint distributions of the form

$$p(u^n, x^n, y^n) = \left(\prod_{i=1}^n p(u_i) p\left(x_i | u_i, x^{i-1}, y^{i-1}\right) \right) p(y^n \| x^n)$$

with deterministic[4] $p\left(x_i | u_i, x^{i-1}, y^{i-1}\right)$, $i = 1, \ldots, n$.

Codebook generation and encoding. Fix n and let $p_i^*(u_i)$, $i = 1, \ldots, n$, and $f_i^* : \left(u_i, x^{i-1}, y^{i-1}\right) \mapsto x_i$, $i = 1, \ldots, n$, achieve the maximum of (18). We will also use the notation $p^*(u^n) = \prod_{i=1}^n p_i^*(u_i)$ and $f^*(u^n, x^{n-1}, y^{n-1}) = \left(f_1^*(u_1), \ldots, f_n^*\left(u_n, x^{n-1}, y^{n-1}\right)\right)$.

For each $k = k(L, n) = Ln^2 + n$, $L = 1, 2, \ldots$, we generate a $(2^{kR}, k)$ code $\left\{X_i(W, Y^{i-1})\right\}_{i=1}^k$ as summarized in Fig. 4. As before, $\tilde{\mathbf{X}}^{(n)}$, $\tilde{\mathbf{Y}}^{(n)}$, and $\tilde{\mathbf{Z}}^{(n)}$ are copies of the underlying sequences $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, respectively, with every $(Ln+1)$st symbol omitted.

For each $w \in [2^{kR}]$, we generate a codeword $\mathbf{U}^{(n)}(w) = U^{Ln^2}(w)$ of length Ln on the n-letter alphabet $\mathcal{U}_1 \times \cdots \times \mathcal{U}_n$ independently according to

$$p(u^{Ln^2}) = \prod_{i=1}^{Ln} p^*\left(u_{(n-1)i+1}^{ni}\right).$$

This gives a $2^{kR} \times Ln$ codebook matrix with each entry drawn i.i.d. according to $p^*(u^n)$.

[4]That is, $p\left(x_i | u_i, x^{i-1}, y^{i-1}\right) = 0$ or 1.

To communicate the message $W = w$, the transmitter chooses the codeword $\mathbf{U}^{(n)}(w) = U^{Ln^2}(w)$ and sends

$$\tilde{X}_{(i-1)n+j} = f_j^*\left(U_j(w), \tilde{X}_{(i-1)n+1}^{(i-1)n+j-1}, \tilde{Y}_{(i-1)n+1}^{(i-1)n+j-1}\right), \quad i = 1, \ldots, Ln, \ j = 1, \ldots, n.$$

Thus, the code function $X^n(w, Y^{n-1})$ utilizes the codeword $\mathbf{U}^{(n)}$ and the channel feedback $\tilde{\mathbf{Y}}^{(n)}$ only within the frame of n transmissions (each box in Fig. 4).

Decoding. Upon receiving Y^k, the receiver declares that the message $\hat{W}$ was sent if there is a unique $\hat{W}$ such that

$$(\mathbf{U}^{(n)}(\hat{W}), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}(U^n, Y^n)$$

that is, $(\mathbf{U}^{(n)}(\hat{W}), \tilde{\mathbf{Y}}^{(n)})$ is jointly typical with respect to the joint distribution $p(u^n, y^n)$ specified by $p^*(u^n)p(z^n)$, $x_i = f_i^*\left(u_i, x^{i-1}, y^{i-1}\right)$, and the definition of the channel (15). Otherwise, an error is declared.

Analysis of the probability of error. We define the following events:

$$E_w = \left\{ \left(\mathbf{U}^{(n)}(w), \tilde{\mathbf{Y}}^{(n)} \right) \in A_\epsilon^{*(Ln)}(U^n, Y^n) \right\}, \quad w \in [2^{kR}].$$

Without loss of generality, we assume $W = 1$ was sent.

From Lemma 7, $\mathbf{U}^{(n)}(1)$ and $\mathbf{Z}^{(n)}$ are jointly typical with high probability for L sufficiently large. Furthermore, $\tilde{\mathbf{Y}}^{(n)}$ is an n-letter blockwise function of $(\tilde{\mathbf{X}}^{(n)}(1), \tilde{\mathbf{Z}}^{(n)})$, and thus of $(\mathbf{U}^{(n)}(1), \tilde{\mathbf{Z}}^{(n)})$. Therefore, the probability of the event E_1^c that the intended codeword $\mathbf{U}^{(n)}(1)$ is not jointly typical with $\tilde{\mathbf{Y}}^{(n)}$ vanishes as $L \to \infty$.

On the other hand, $\mathbf{U}^{(n)}(w)$, $w \ne 1$, is generated blockwise i.i.d. $\sim p^*(u^n)$ independent of $\mathbf{Y}^{(n)}$. Hence, from Lemma 8, the probability of the event E_w that $\mathbf{U}^{(n)}(w)$ is jointly typical with $\mathbf{Y}^{(n)}$ is bounded by

$$\Pr(E_w) \le 2^{-Ln(I(U^n; Y^n) - \delta)}, \ \text{for all } w \ne 1$$

where $\delta \to 0$ as $\epsilon \to 0$. Consequently, we have

$$\begin{aligned}\Pr(\hat{W} \neq W) &\leq \Pr(E_1^c) + \sum_{w=2}^{2^{kR}} \Pr(E_w) \\ &\leq \epsilon + 2^{kR} 2^{-Ln(I(U^n;Y^n)-\delta)} \\ &\leq 2\epsilon\end{aligned}$$

if L is sufficiently large and

$$kR = (Ln^2 + n)R < Ln(I(U^n; Y^n) - \delta).$$

Thus by letting $L \to \infty$ and then $\epsilon \to 0$, we can achieve any rate $R < C'_{\mathrm{FB},n}$.

Finally by Lemma 10, this implies that we can achieve

$$C'_{FB,n} = C_{FB,n} = \max_{p(x^n \| y^{n-1})} \frac{1}{n} I(X^n \to Y^n)$$

which completes the proof of Theorem 1.

IV. Concluding Remarks

Trading off generality for transparency, we have focused on the stationary channels of the form

$$Y_n = f\left(X_{n-m}^n, Z_{n-m}^n\right)$$

and presented a straightforward proof of the feedback coding theorem. As is manifested in previous approaches by Tatikonda [37], [38] and Permuter *et al.* [31], the Shannon strategy (also called code-function or code-tree) has a fundamental role in transforming the feedback coding problem into a nonfeedback one. For the given channel model, the equivalent nonfeedback problem can be then solved by a *scalable* coding scheme that constructs a long typical (ergodic) input–output sequence pair by concatenating shorter nonergodic ones with appropriate phase shifts.

Lemma 10 provides a transparent (and in a sense, mechanical) method for this feedback–nonfeedback, directed information–mutual information transformation, and can be applied to other channel models, provided that the equivalent nonfeedback problem has a straightforward coding theorem. For example, we can show that the finite-state channel

$$p(y_n, s_n | s_{n-1}, x_n) = p(y_n | s_{n-1}, x_n) p(s_n | s_{n-1}, x_n, y_n)$$

with $s_n = f(s_{n-1}, x_n, y_n)$ for some deterministic function f (but without assumption of indecomposability) has the feedback capacity lower bounded by

$$C_{\mathrm{FB}} \geq \sup_{n \geq 1} \max_{p(x^n \| y^{n-1})} \min_{s_0} \frac{1}{n} I(X^n \to Y^n | s_0).$$

This result was previously shown by Permuter *et al.* [31, Sec. V] via a generalization of Gallager's random coding exponent method for finite state channels without feedback [14, Sec. 5.9]. Here we sketch a simple alternative proof.

From a trivial modification of Lemma 10, the problem reduces to showing that

$$\max_{p(u^n), x_i = f(u_i, x^{i-1}, y^{i-1})} \min_{s_0} \frac{1}{n} I\left(U^n; Y^n | s_0\right) \tag{19}$$

is achievable for each n. But the given Shannon strategy $\left(p^*(u^n), x^n = f^*\left(u^n, x^{n-1}, y^{n-1}\right)\right)$ induces a new time-invariant finite-state channel on the n-letter super alphabet as $p(\mathbf{y}_k, \mathbf{s}_k | \mathbf{s}_{k-1}, \mathbf{u}_k)$. Hence we can use Gallager's random coding exponent method directly to achieve

$$\lim_{k \to \infty} \max_{p(\mathbf{u}^k)} \min_{\mathbf{s}_0} \frac{1}{k} I\left(\mathbf{U}^k; \mathbf{Y}^k | \mathbf{s}_0\right)$$

which can be shown to be no less than our target

$$\frac{1}{n} I(\mathbf{U}_1; \mathbf{Y}_1 | \mathbf{s}_0)$$

because of the deterministic evolution of the state $S_n = f(S_{n-1}, X_n, Y_n)$.

Finally, we comment on an important question that is not addressed in this paper. Our characterization of the feedback capacity

$$C_{\mathrm{FB}} = \lim_{n \to \infty} \max_{p(x^n \| y^{n-1})} \frac{1}{n} I(X^n \to Y^n) \tag{20}$$

or any similar multiletter expressions are in general not computable and do not provide much insight on the structure of the capacity achieving coding scheme. One may ask whether a stationary or even Markov distribution is asymptotically optimal for the sequence of maximizations in (20). This problem has been solved for a few specific channel models such as certain classes of finite-state channels [6], [42], [31], [30] and stationary additive Gaussian noise channels [21], [22], sometimes with analytic expressions for the feedback capacity. In this context, the current development is just the first step toward the complete characterization of the feedback capacity.

Appendix

Here we collect relevant materials from ergodic theory and information theory. Some of the lemmas are classical and are presented in order to make the paper self-contained. Throughout the Appendix, $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$ denotes a generic stochastic process on a finite alphabet $\mathcal{Z}$ with associated probability measure P defined on Borel sets under the usual topology on $\mathcal{Z}^{\infty}$.

A. Ergodicity

Given a *stationary* process $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$, let $T : \mathcal{Z}^{\infty} \to \mathcal{Z}^{\infty}$ be the associated measure preserving shift transformation. Intuitively, T maps the infinite sequence $(z_1, z_2, z_3, \ldots)$ to $(z_2, z_3, z_4, \ldots)$. We say the transformation T (or the process $\mathbf{Z}$ itself) is *ergodic* if every measurable set A with $TA = A$ satisfies either $P(A) = 0$ or $P(A) = 1$.

The following characterization of ergodicity is well known; see, for example, Petersen [32, Exercise 2.4.4] or Wolfowitz [41, Lemma 10.3.1].

Lemma 1: Suppose $\{Z_i\}_{i=1}^{\infty}$ be a stationary process and let T denote the associated measure preserving shift transformation. Then, $\{Z_i\}$ is ergodic if and only if

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} P(T^{-i} A \cap B) = P(A) \cdot P(B)$$

for all measurable A and B.

When $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$ are independent stationary ergodic processes, they are not necessarily jointly ergodic. For example, if we take

$$\mathbf{X} = \begin{cases} 01010101\ldots, & \text{with probability } 1/2 \\ 10101010\ldots, & \text{with probability } 1/2 \end{cases}$$

and $\mathbf{Z}$ is independent and identically distributed as $\mathbf{X}$, then it is easy to verify that $\{Y_i = X_i + Z_i \bmod 2\}_{i=1}^{\infty}$ is *not* ergodic. However, if one of the processes is mixing reasonably fast, then they are jointly ergodic. The following result states a sufficient condition for joint ergodicity.

Lemma 2: If $\mathbf{X}$ is i.i.d., and $\mathbf{Z}$ is stationary ergodic, independent of $\mathbf{X}$, then the pair $(\mathbf{X}, \mathbf{Z}) = \{(X_i, Z_i)\}_{i=1}^{\infty}$ is jointly stationary ergodic.

A stronger result is true, which assumes $\mathbf{X}$ to be weakly mixing only. The proof is an easy consequence of Lemma 1; for details refer to Brown [4, Proposition 1.6] or Wolfowitz [41, Th. 10.3.1].

B. Strong Typicality

The next lemma due to Gallager [14, Lemma 9.8.2] deals with the ergodic decomposition of the n-letter super process that is built from a single-letter stationary ergodic one; see also Nedoma [28] and Berger [2, Sec. 7.2].

Lemma 3: Suppose $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$ be stationary ergodic on $\mathcal{Z}$, and let T be the associated shift transformation. Define the nth-order super process $\mathbf{Z}^{(n)} = \left\{Z_i^{(n)}\right\}_{i=1}^{\infty}$ on $\mathcal{Z}^n$ as

$$Z_i^{(n)} = (Z_{n(i-1)+1}, Z_{n(i-1)+2}, \ldots, Z_{ni}), \quad i = 1, 2, \ldots.$$

Then, the super process $\mathbf{Z}^{(n)}$ has n' ergodic modes, each with probability $1/n'$ and disjoint up to measure zero, where n' divides n. Furthermore, in the space $\mathcal{Z}^{\infty}$ of the original process $\mathbf{Z}$, the sets $S_1, S_2, \ldots, S_{n'}$ corresponding to these ergodic modes can be related by $T(S_i) = S_{i+1}$, $1 \le i \le n-1$, and $T(S_n) = S_1$.

We will use the notation $P(\cdot|S_k)$, $k = 1, \ldots, n'$, for the probability measure under each ergodic mode.

We use the strong typicality [9, Sec. 10.6] as the basic method of decoding. Here we review a few basic properties of strongly typical sequences.

First definitions. Let $N(a|x^n)$ denote the number of occurrences of the symbol a in the sequence x^n. We say a sequence $x^n \in \mathcal{X}^n$ is *ϵ-strongly typical* (or typical in short) with respect to a distribution $P(x)$ on $\mathcal{X}$ if

$$\left|\frac{1}{n}N(a|x^n) - P(a)\right| < \frac{\epsilon}{|\mathcal{X}|}$$

for all $a \in \mathcal{X}$ with $P(a) > 0$, and $N(a|x^n) = 0$ for all $a \in \mathcal{X}$ with $P(a) = 0$. Consistent with this definition, we say a pair of sequences (x^n, y^n) are *jointly ϵ-strongly typical* (or jointly typical in short) with respect to a distribution $P(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ if

$$\left|\frac{1}{n}N(a, b|x^n, y^n) - P(a, b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}$$

for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $P(a, b) > 0$, and $N(a, b|x^n) = 0$ for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $P(a, b) = 0$.

The set of strongly typical sequences $x^n \in \mathcal{X}^n$ with respect to $X \sim P(x)$ is denoted $A_{\epsilon}^{*(n)}(X)$. We similarly define a joint typical set $A_{\epsilon}^{*(n)}(X, Y)$ for $(X, Y) \sim P(x, y)$.

The following statement is a trivial consequence of the definition of typical sequences.

Lemma 4: Suppose $X \sim P(x)$. If $x^n \in A_{\epsilon}^{*(n)}(X)$ and $y^n = f(x^n) := (f(x_1), f(x_2), \ldots, f(x_n))$, then $y^n \in A_{\delta}^{*(n)}(f(X))$ with $\delta = \epsilon \cdot (|\mathcal{X}| - 1)$.

As a special case, if (x^n, y^n) is ϵ-strongly typical with respect to a joint distribution $P(x, y)$, then x^n is ϵ-strongly typical with respect to the marginal $P(x) = \sum_y P(x, y)$.

Our discussion on the typical sequences so far has not given a specific context on how they are generated. Now we connect the notion of strong typicality with ergodic processes. First, from Birkhoff's ergodic theorem [32, Th. 2.2.3] and the definition of ergodicity, the following lemma is immediate.

Lemma 5: Let $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$ be stationary ergodic with $Z_1 \sim P(z)$. Then

$$\Pr\left(Z^n \in A_{\epsilon}^{*(n)}(Z_1)\right) \to 1 \text{ as } n \to \infty.$$

As we mentioned above, the nth order super process $\mathbf{Z}^{(n)} = \left\{Z_i^{(n)}\right\}_{i=1}^{\infty}$ defined as

$$Z_i^{(n)} = Z_{(n-1)i+1}^{ni}, \quad i = 1, 2, \ldots$$

is not necessarily ergodic, but is a mixture of disjoint ergodic modes. Thus, a "typical" sequence from the super process $\mathbf{Z}^{(n)}$ is not necessarily typical with respect to $P(z^n)$ on the n-letter alphabet $\mathcal{Z}^n$. The following construction by Gallager [14, pp. 498–499], however, gives a typical sequence in the n-letter super alphabet by shifting through each ergodic phase.

Lemma 6: Given positive integers n, L and a stationary ergodic process $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$, construct $\tilde{\mathbf{Z}} = \{\tilde{Z}_i\}_{i=1}^{Ln^2}$ as follows (see Fig. 5):

$$\tilde{Z}_i = \begin{cases} Z_i, & i = 1, \ldots, Ln \\ Z_{i+1}, & i = Ln+1, \ldots, 2Ln \\ \quad\vdots & \\ Z_{i+n-1}, & i = Ln(n-1)+1, \ldots, Ln^2. \end{cases}$$

In other words, $\{\tilde{Z}_i\}_{i=1}^{Ln^2}$ is a verbatim copy of $\{Z_i\}_{i=1}^{Ln^2+n}$ with every $(Ln+1)$st position skipped. Now let $\tilde{\mathbf{Z}}^{(n)} = \left(\tilde{Z}_1^n, \tilde{Z}_{n+1}^{2n}, \ldots, \tilde{Z}_{Ln^2-n+1}^{Ln^2}\right)$ be the associated nth order super process of length Ln. Then,

$$\Pr\left(\tilde{\mathbf{Z}}^{(n)} \in A_{\epsilon}^{*(Ln)}(Z^n)\right) \to 1 \text{ as } L \to \infty.$$

Proof: From Lemma 3 and the given construction of skipping one position after every Ln symbols, each of n sequences

$$\begin{aligned} &\left(\tilde{Z}_1^n, \ldots, \tilde{Z}_{Ln-n+1}^{Ln}\right) \\ &\quad = \left(Z_1^n, \ldots, Z_{Ln-n+1}^{Ln}\right) \\ &\left(\tilde{Z}_{Ln+1}^{Ln+n}, \ldots, \tilde{Z}_{2Ln-n+1}^{2Ln}\right) \\ &\quad = \left(Z_{Ln+2}^{L(n+1)+1}, \ldots, Z_{2Ln-n+2}^{2Ln+1}\right) \\ &\quad\vdots \end{aligned}$$

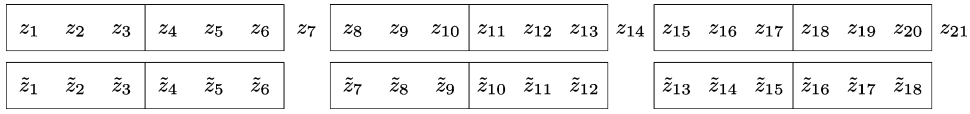

Fig. 5. Construction of $\bar{Z}^{Ln^2}$ from Z^{Ln^2+n}: $n = 3$, $L = 2$.

$$\left(\tilde{Z}_{Ln(n-1)+1}^{Ln(n-1)+n}, \ldots, \tilde{Z}_{Ln^2-n+1}^{Ln^2}\right) = \left(Z_{Ln(n-1)+n}^{Ln(n-1)+2n-1}, \ldots, Z_{Ln^2}^{Ln^2+n-1}\right)$$

falls in one of ergodic modes $(S_1, \ldots, S_{n'})$ with n/n' sequences in each mode. Now for each sequence with corresponding ergodic mode S_k, the relative frequencies of all super symbols $a^n \in \mathcal{Z}^n$ converge to the corresponding distribution $P(a^n|S_k)$ as $L \to \infty$. But each ergodic mode is visited evenly, each by n/n' sequences. Therefore, the relative frequencies of all $a^n \in \mathcal{Z}^n$ in the entire sequence $\tilde{Z}_1^{Ln^2}$ converge to

$$\frac{1}{n'}\sum_{k=1}^{n'} P(a^n|S_k) = P(a^n)$$

as $L \to \infty$. □

Combining Lemma 2 with the proof of Lemma 6, we have the following result.

Lemma 7: Under the condition of Lemma 6, let further $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ be blockwise i.i.d. $\sim P(x^n)$, that is, $X_i^{(n)} = X_{(n-1)i+1}^{ni}$, $i = 1, 2, \ldots$, i.i.d. $\sim P(x^n)$, independent of $\mathbf{Z}$. Then

$$\Pr\left((\mathbf{X}^{(n)}, \tilde{\mathbf{Z}}^{(n)}) \in A_\epsilon^{*(Ln)}(X^n, Z^n)\right) \to 1 \text{ as } L \to \infty.$$

Finally we recall a key result that links the typicality with mutual information [9, Lemma 10.6.2].

Lemma 8: Suppose $(X, Y) \sim P(x, y)$ and let $X_1, X_2, \ldots$ be i.i.d. $\sim P(x)$. For $y^n \in A_\epsilon^{*(n)}(Y)$, the probability that $(X^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$ is upper bounded by

$$\Pr\left((X^n, y^n) \in A_\epsilon^{*(n)}(X, Y)\right) \le 2^{-n(I(X;Y)-\delta)}$$

where $\delta \to 0$ as $\epsilon \to 0$.

C. Channels With Side Information

We prove the following identity, originally proved by Caire and Shamai [5, Proposition 1], in a purely algebraic manner, then review its meaning in information theory. Here we assume every alphabet is finite.

Lemma 9: Suppose $S \sim p(s)$. For a given conditional distribution $p(y|x, s)$ on the product space $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, we have

$$\max_{p(u), x=f(u,s)} I(U; Y, S) = \max_{p(x|s)} I(X; Y|S) \tag{21}$$

where the maximum on the left-hand side is taken over all conditional distributions of the form $p(u, x|s) = p(u)p(x|u, s)$ with deterministic $p(x|u, s)$ (that is, $p(x|u, s) = 0$ or 1), and the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \le (|\mathcal{X}| - 1)|\mathcal{S}|$.

Proof: For any joint distribution $p(u, x, s, y)$ of the form $p(u, x, s, y) = p(u)p(s)p(x|u, s)p(y|x, s)$ with deterministic $p(x|u, s)$, we have the following Markov chains: $U \to (X, S) \to Y$ and $X \to (U, S) \to Y$. Combined with the independence of U and S, these Markov relationships imply that

$$\begin{aligned}\max_{p(u), x=f(u,s)} I(U; Y, S) &= \max_{p(u), x=f(u,s)} I(U; Y|S) \\ &= \max_{p(u), x=f(u,s)} I(X; Y|S).\end{aligned} \tag{22}$$

But it can be easily verified [40, eq. (44)] that any conditional distribution $p(x|s)$ can be represented as

$$p(x|s) = \sum_u p(u)p(x|u, s)$$

for appropriately chosen $p(u)$ and *deterministic* $p(x|u, s)$ with cardinality of U upper bounded by $|\mathcal{U}| \le (|\mathcal{X}| - 1)|\mathcal{S}| + 1$. Therefore, we have

$$\max_{p(u), x=f(u,s)} I(X; Y|S) = \max_{p(x|s)} I(X; Y|S)$$

which proves the desired result. □

It is well known that the capacity of a memoryless state-dependent channel $p(y|x, s)$ is given as

$$C = \max_{p(x|s)} I(X; Y|S)$$

if the state information is known at both the encoder and the decoder prior to the actual communication. What will happen if the transmitter learns the state information on the fly, so that only the past and present state realization can be utilized for communication?

Shannon [36] considered the communication over a memoryless state-dependent channel $p(y|x, s)$ with state information available *only* at the transmitter on the fly, and showed that the capacity is given by

$$C = \max_{p(u), x=f(u,s)} I(U; Y) \tag{23}$$

where the cardinality of U is bounded as $|\mathcal{U}| \le |\mathcal{X}|^{|\mathcal{S}|}$, counting all functions $f_u : \mathcal{S} \to \mathcal{X}$. This capacity is achieved by attaching a physical device $X = f(U, S)$ in front of the actual channel as depicted in Fig. 6, which maps the channel state S to the channel input X according to the function (index) U. Now treating U as the input to the newly generated channel

$$p(y|u) = \sum_{x,s} p(s)p(x|u, s)p(y|x, s)$$

and coding as in the case of usual memoryless channels, we can easily achieve $I(U; Y)$. This method, surprisingly simple yet optimal, is sometimes called the Shannon strategy.

Now when the decoder also knows the channel state S, it is equivalent for the decoder to receive the augmented channel output $Y' = (Y, S)$. Thus, the capacity of the same channel

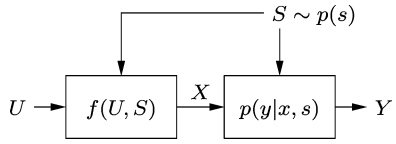

Fig. 6. Shannon strategy for coding with side information.

$p(y|x,s)$ with the state information *causally* known at both the encoder and the decoder[5] follows from (23) as

$$C = \max_{p(u),x=f(u,s)} I(U;Y,S).$$

Therefore, Lemma 9 states that when the same side information is available at the receiver, the causal encoder with the best Shannon strategy performs no worse than the noncausal encoder who can preselect the entire codeword compatible with the whole state sequence. For an operational (rather than algebraic) proof of the same result, refer to [5, Proposition 1].

For the last lemma needed for main results, we recall the notation of causally conditioned distributions

$$p(x^n||y^{n-1}) = \prod_{i=1}^{n} p\left(x_i|x^{i-1},y^{i-1}\right) \tag{24}$$

and

$$p(y^n||x^n) = \prod_{i=1}^{n} p\left(y_i|x^{i},y^{i-1}\right). \tag{25}$$

(The notation (24) and (25) can be unified if we define

$$p(a^n||b^m) = \begin{cases} \prod_{i=1}^{n} p(a_i|b^i,a^{i-1}), & n=m \\ p(a^n||\emptyset^{n-m}b^m), & n>m \\ p(\emptyset^{m-n}a^n||b^m), & n<m. \end{cases}$$

By chain rule, we have

$$p(x^n||y^{n-1})p(y^n||x^n) = p(x^n,y^n) = p(x^n)p(y^n|x^n)$$

for any joint distribution $p(x^n,y^n)$. Thus, given a causally conditioned distribution (or a channel) $p(y^n||x^n)$, the causally conditioned distribution (or the input) $p(x^n||y^{n-1})$ completely specifies the joint distribution $p(x^n,y^n)$.

As a corollary of Lemma 9, we have the following result (cf. [38, Lemma 5.2]):

Lemma 10: Suppose a causally conditioned distribution $p(y^n||x^n)$ is given. Then we have

$$\max_{p(u^n),x_i=f(u_i,x^{i-1},y^{i-1})} I(U^n;Y^n) = \max_{p(x^n||y^{n-1})} I(X^n \to Y^n) \tag{26}$$

where the maximum on the left-hand side is taken over all joint distributions of the form

$$\begin{aligned} p(u^n,x^n,y^n) &= \prod_{i=1}^{n} p(u_i)p\left(x_i|u_i,x^{i-1},y^{i-1}\right)p\left(y_i|x^i,y^{i-1}\right) \\ &= \left(\prod_{i=1}^{n} p(u_i)p\left(x_i|u_i,x^{i-1},y^{i-1}\right)\right)p(y^n||x^n) \end{aligned} \tag{27}$$

[5]For the usual block coding, the decoder causality is irrelevant. The message is decoded only after the entire block is received.

with deterministic $p(x_i|u_i,x^{i-1},y^{i-1})$, $i=1,\ldots,n$, and the auxiliary random variables U_i have the cardinalities bounded by $|\mathcal{U}_i| \le |\mathcal{X}|^i|\mathcal{Y}|^{i-1}$, respectively.

Proof: Let $q(u^n,x^n,y^n)$ be any joint distribution of the form (27) such that $q\left(x_i|u_i,x^{i-1},y^{i-1}\right)$, $i=1,\ldots,n$, are deterministic and that $q(y^n||x^n) = p(y^n||x^n)$, i.e., the joint distribution $q(u^n,x^n,y^n)$ is consistent with the given causally conditioned distribution $p(y^n||x^n)$. For $(U^n,X^n,Y^n) \sim q(u^n,x^n,y^n)$, it is easy to verify that U_i^n is independent of $(U^{i-1},X^{i-1},Y^{i-1})$, which implies that $U^{i-1} \to (X^{i-1},Y^{i-1}) \to Y_i^n$ form a Markov chain. On the other hand, X^{i-1} is a deterministic function of (U^{i-1},Y^{i-1}) and thus $X^{i-1} \to (U^{i-1},Y^{i-1}) \to Y_i^n$ also form a Markov chain. Similarly, we have the Markovity for $U^i \to (X^i,Y^{i-1}) \to Y_i^n$ and $X^i \to (U^i,Y^{i-1}) \to Y_i^n$. Therefore, we have

$$\begin{aligned} I\left(U_i;Y^n|U^{i-1}\right) &= I\left(U_i;Y_i^n|Y^{i-1},U^{i-1}\right) && (28) \\ &= H\left(Y_i^n|Y^{i-1},U^{i-1}\right) - H\left(Y_i^n|Y^{i-1},U^{i}\right) \\ &= H\left(Y_i^n|Y^{i-1},X^{i-1}\right) - H\left(Y_i^n|Y^{i-1},X^{i}\right) \\ &= I\left(X_i;Y_i^n|X^{i-1},Y^{i-1}\right) && (29) \end{aligned}$$

where the equality (28) follows from the independence of U_i and (U^{i-1},Y^{i-1}), and (29) follows from Markov relationships observed above. Now from the alternative expansion of the directed information shown in (11), we have

$$\max_q I(U^n;Y^n) = \max_q I(X^n \to Y^n).$$

Finally, by using distributions of the form

$$p\left(x_i|x^{i-1},y^{i-1}\right) = \sum_{u_i} p(u_i)p\left(x_i|u_i,x^{i-1},y^{i-1}\right)$$

for each $i=1,\ldots,n$, with appropriately chosen $p(u_i)$ and deterministic $p\left(x_i|u_i,x^{i-1},y^{i-1}\right)$, we can represent any causally conditioned distribution

$$\begin{aligned} p(x^n||y^{n-1}) &= \prod_{i=1}^{n} p(x_i|x^{i-1},y^{i-1}) \\ &= \sum_{u^n} \prod_{i=1}^{n} \left(p(u_i)p(x_i|u_i,x^{i-1},y^{i-1})\right) \end{aligned}$$

which implies that

$$\max_q I(X^n \to Y^n) = \max_{p(x^n||y^{n-1})} I(X^n \to Y^n)$$

and completes the proof. □

ACKNOWLEDGMENT

The author wishes to thank Tom Cover, Bob Gray, and Haim Permuter for helpful discussions. He also thanks the anonymous reviewer and the Associate Editor for their valuable comments, which helped to improve the quality of the paper.

REFERENCES

[1] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 546–549, Mar. 1995.

[2] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[3] D. Blackwell, "Information theory," in *Modern Mathematics for the Engineer: Second Series.*, New York, 1961, pp. 182–193.
[4] J. R. Brown, *Ergodic Theory and Topological Dynamics*. New York: Academic, 1976.
[5] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2007–2019, Sep. 1999.
[6] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.
[7] T. M. Cover, "An achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 4, pp. 399–404, Jul. 1975.
[8] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
[10] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, 1959.
[11] A. Feinstein, "On the coding theorem and its converse for finite-memory channels," *Inf. Contr.*, vol. 2, pp. 25–44, 1959.
[12] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory*, vol. IT-4, pp. 2–22, 1954.
[13] G. D. Forney, Jr, *Information Theory*. Stanford, CA: Stanford Univ., 1972, Unpublished course notes.
[14] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[15] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 1, pp. 3–18, Jan. 1965.
[16] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
[17] R. M. Gray and D. S. Ornstein, "Block coding for discrete stationary $\bar{d}$ -continuous noisy channels," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 3, pp. 292–306, May 1979.
[18] T. S. Han, *Information-Spectrum Methods in Information Theory*. New York: Springer, 2003.
[19] A. I. Khinchin, *Mathematical Foundations of Information Theory*. New York, Dover: , 1957.
[20] J. C. Kieffer, "Block coding for weakly continuous channels," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 6, pp. 721–727, Nov. 1981.
[21] J. C. Kieffer, "Feedback capacity of the first-order moving average Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3063–3079, Jul. 2006.
[22] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," Feb. 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0602091/
[23] G. Kramer, "Directed information for channels with feedback," Dr. sc. thchn, Hartung-Gorre Verlag, Swiss Federal Inst. Technol. (ETH) Zurich, Konstanz, 1998.
[24] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[25] A. Lapidoth and İ. E. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 973–983, May 1998.
[26] J. L. Massey, "Causality, feedback, and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Honolulu, HI, Nov. 1990, pp. 303–305.
[27] J. L. Massey and P. C. Massey, "Conservation of mutual and directed information," in *Proc. Int. Symp. Inf. Theory*, Adelaide, Australia, Sep. 2005, pp. 157–158.
[28] J. Nedoma, "Über die Ergodizität und r-Ergodizität stationärer Wahr-scheinlichkeitsmasse," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 2, pp. 90–97, 1963.
[29] D. L. Neuhoff and P. C. Shields, "Channels with almost finite memory," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 4, pp. 440–447, Jul. 1979.
[30] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0610047/
[31] H. Permuter, T. Weissman, and A. Goldsmith, "Finite-state channels with time-invariant deterministic feedback," 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0608070
[32] K. Petersen, *Ergodic Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
[33] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.
[34] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[35] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 8–19, Sep. 1956.
[36] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289–293, 1958.
[37] S. Tatikonda, "Control under communication constraints," Ph.D., Mass. Inst. Technol., Cambridge, 2000.
[38] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0609139/
[39] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
[40] F. M. J. Willems and E. C. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 313–327, 1985.
[41] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1978.
[42] 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.