A Coding Theorem for a Class of Stationary Channels with Feedback

Young-Han Kim
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA
yhk@ucsd.edu

***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 $I(X^n \to Y^n) = \sum_{i=1}^n I(X^i; Y_i|Y^{i-1})$ denotes the Massey directed information from the channel input to the output, and the supremum is taken over all causally conditioned distributions $p(x^n||y^{n-1}) = \prod_{i=1}^n p(x_i|x^{i-1}, y^{i-1})$. The main ideas of the proof are 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.

I. Introduction

In [11], Massey introduced the mathematical notion of directed information

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

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. Since then there have been many attempts to show that Massey's directed information is indeed the feedback capacity, namely,

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

where the supremum is taken over all *causally conditioned* probabilities $p(x^n||y^{n-1}) = \prod_{i=1}^n p(x_i|x^{i-1}, y^{i-1})$. For example, in a heroic effort [17], [18], Tatikonda attacked the general nonanticipatory channel with feedback by combining Verdú–Han formula for nonfeedback capacity [19], Massey directed information, and Shannon strategy for channel side information [16], as well as dynamic programming for Markov decision processes. As the cost of generality, however, it is extremely difficult to establish a simple formula like (1). See, for example, [18, Theorem 7.5] for a major hurdle in proving the equivalence between a Verdú–Han-type formula and (1).

More recently, Yang, Kavčić, and Tatikonda [20] and Chen and Berger [3] studied special cases of finite-state channels, based on Tatikonda's framework. A finite-state channel [6, Section 4.6] is described by the conditional probability distribution

$$p(y_n, s_n|x_n s_{n-1}), \qquad (2)$$

where s_n denotes the channel state at time n. Using a different approach based on Gallager's proof of the nonfeedback capacity [6, Section 5.9], Permuter, Weissman, and Goldsmith [15] proved various coding theorems for finite-state channels with feedback that, *inter alia,* subsume many results in [20], [3] and establish the validity of (1) for indecomposable finite-state channels without intersymbol interference (i.e., finite-state channels whose state evolves as an ergodic Markov chain, independent of the channel input).

The main goal of this paper is to establish the validity of the feedback capacity formula (1) for a reasonably general class of channels with memory, in the simplest manner. Towards this goal, we focus on stationary nonanticipatory channels of the form

$$Y_i = \begin{cases} \emptyset, & i = 1, \dots, m,\\ g(X_{i-m}^i, Z_{i-m}^i), & i = m+1, m+2, \dots, \end{cases} \qquad (3)$$

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 noise process $\{Z_n\}_{n=1}^\infty$ is an arbitrary stationary ergodic process (without any mixing condition).

The channel model (3) 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]

[1] A notable exception is a famous finite-state channel called the "trapdoor channel" introduced by Blackwell [1], the feedback capacity of which is established in [14].

1-4244-1429-6/07/$25.00 ©2007 IEEE

The channel (3) has finite input memory in the sense of Feinstein [5] and can be viewed as a finite-window sliding-block coder [7, Section 9.4] of input and noise processes (cf. primitive channels introduced by Neuhoff and Shields [13] in which the noise process is memoryless). Compared to the general finite-state channel model (2), in which the channel has infinite input memory but the channel noise is memoryless, our channel model (3) 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 and the finite sliding-block channel model nicely complement each other.

Our main result is to show that the feedback capacity C_{FB} of the channel (3) is characterized by (1). More precisely, we consider a communication problem depicted in Figure 1. Here one wishes to communicated a message index $W \in$

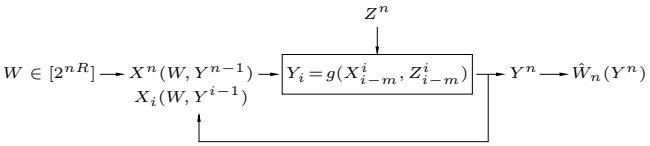

Fig. 1. Feedback communication channel $Y_i = g(X_{i-m}^i, Z_{i-m}^i)$.

$[2^{nR}] := \{1, 2, \ldots, 2^{nR}\}$ over the channel (3). 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 $Y_1, \ldots, Y_m$ are set arbitrarily. They depend on the unspecified initial condition (X_{-m+1}^0, Z_{-m+1}^0), the effect of which vanishes from time $m+1$. Thus the long term behavior of the channel is independent of Y_1^m.

We specify a $(2^{nR}, n)$ code with the encoding maps $X^n(W, Y^{n-1}) = (X_1(W), X_2(W, Y_1), \ldots, X_n(W, Y^{n-1}))$, and the decoding map $\hat{W}_n : \mathcal{Y}^n \to [2^{nR}]$. The probability of error $P_e^{(n)}$ is defined as $P_e^{(n)} = \Pr\{\hat{W}_n(Y^n) \neq W\}$, 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 II.

Theorem 1. *The feedback capacity C_{FB} of the channel* (3) *is given by*

$$C_{\text{FB}} = \lim_{n\to\infty} \sup_{p(x^n||y^{n-1})} \frac{1}{n} I(X^n \to Y^n). \tag{4}$$

Our development has two major ingredients. First, we use the coding theorem for the same channel without feedback, fully described in [10]. This approach is somewhat different from the conventional approaches such as Shannon's random codebook generation and typicality decoding, Gallager's random coding exponent method, or Feinstein's maximal coding argument. We use the strong typicality (relative frequency) decoding for n-dimensional super letters. A constructive coding scheme (up to the level of Shannon's random codebook generation) based on block ergodic decomposition of Nedoma [12] is developed, 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 [6, Section 9.8] for lossy compression of stationary ergodic sources. Indeed, when the channel (3) 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 (3), 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 [16]. As a matter of fact, Shannon himself observed that the major utility of his result is feedback communication. Following is the first sentence of [16]:

> 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 [2, 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 the side information as if it were known *a priori*. Thus, intuitively speaking, we can achieve the rate

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

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

$$\sum_{i=1}^{n} I(X_i; Y_i^n | X^{i-1}, Y^{i-1}) = I(X^n \to Y^n). \tag{5}$$

The above argument, while intuitively appealing, is not completely rigorous, however. Therefore, we will take more careful steps, by first proving the achievability of $\frac{1}{n} I(U^n; Y^n)$ for all auxiliary random variables U^n and Shannon strategies $X_i(U_i, X^{i-1}, Y^{i-1})$, $i = 1, \ldots, n$, and then showing that $I(U^n; Y^n)$ reduces to $I(X^n \to Y^n)$ via pure algebra.

II. Proof of Theorem 1

Recall our channel model

$$Y_i = \begin{cases} \emptyset, & i = 1, \ldots, m, \\ g(X_{i-m}^i, Z_{i-m}^i), & i = m+1, m+2, \ldots, \end{cases} \tag{6}$$

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

$$\begin{aligned} C_{\text{FB}} &= \lim_{n\to\infty} C_{\text{FB},n} \\ &= \lim_{n\to\infty} \sup_{p(x^n||y^{n-1})} \frac{1}{n} I(X^n \to Y^n), \end{aligned} \tag{7}$$

where the supremum is over all causally conditioned distributions

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

The following lemma is crucial to the proof of Theorem 1.

Lemma 1. *Suppose a causally conditioned distribution $p(y^n||x^n)$ is given. Then we have*

$$\max_{p(u^n), x_i = f(u_i, x^{i-1}, y^{i-1})} I(U^n; Y^n) = \max_{p(x^n||y^{n-1})} I(X^n \to Y^n), \tag{8}$$

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

$$\begin{aligned} p(u^n, x^n, y^n) &= \prod_{i=1}^{n} \big(p(u_i)p(x_i|u_i, x^{i-1}, y^{i-1})p(y_i|x^i, y^{i-1})\big) \\ &= \Big(\prod_{i=1}^{n} p(u_i)p(x_i|u_i, x^{i-1}, y^{i-1})\Big) p(y^n||x^n) \end{aligned} \tag{9}$$

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

Proof. Let $q(u^n, x^n, y^n)$ be any joint distribution of the form (9) such that $q(x_i|u_i, x^{i-1}, y^{i-1})$, $i = 1, \ldots, n$ 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 $U^{i-1} \to (X^{i-1}, Y^{i-1}) \to Y_i^n$ forms 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 forms a Markov chain. Similarly, we have the Markovity for $U^i \to (Y^{i-1}, X^i) \to Y_i^n$ and $X^i \to (U^i, Y^{i-1}) \to Y_i^n$. Therefore, we have

$$\begin{aligned} I(U_i; Y^n|U^{i-1}) &= I(U_i; Y_i^n|Y^{i-1}, U^{i-1}) && (10) \\ &= H(Y_i^n|Y^{i-1}, U^{i-1}) - H(Y_i^n|Y^{i-1}, U^i) \\ &= H(Y_i^n|Y^{i-1}, X^{i-1}) - H(Y_i^n|Y^{i-1}, X^i) && (11) \\ &= I(X_i; Y_i^n|X^{i-1}, Y^{i-1}), \end{aligned}$$

where (10) follows from the independence of U_i and (U^{i-1}, Y^{i-1}), and (11) follows from Markov relationships observed above. Now from the alternative expansion of the directed information shown in (5), we have

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

Finally, by using distributions of the form

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

with appropriately chosen $p(u_i)$ and deterministic $p(x_i|u_i, x^{i-1}, y^{i-1})$, we can represent any causally conditioned distribution

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

which implies the desired result. □

That the limit in (7) is well-defined follows from the superadditivity[2] of $nC_{\text{FB},n}$. Thus,

$$C_{\text{FB}} = \lim_{n\to\infty} C_{\text{FB},n} = \sup_{n\ge 1} C_{\text{FB},n}.$$

The converse was proved by Massey [11, Theorem 3]. For any sequence of $(2^{nR}, n)$ codes with $P_e^{(n)}$, we have from Fano's inequality

$$\begin{aligned} nR &\le I(W; Y^n) + n\epsilon_n \\ &= \sum_{i=1}^{n} I(W; Y_i|Y^{i-1}) + n\epsilon_n \\ &= \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}) + n\epsilon_n && (12) \\ &= I(X^n \to Y^n) + n\epsilon_n, \end{aligned}$$

where $\epsilon_n \to 0$ as $n \to \infty$. Here (12) 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 achieves $C_{\text{FB},n}$ for each n. For simplicity, we assume that the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. A complete argument is given in [10].

In the light of Lemma 1, it suffices to show that

$$C'_{\text{FB},n} = \max_{p(u^n), x_i = f(u_i, x^{i-1}, y^{i-1})} I(U^n; Y^n) \tag{13}$$

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

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

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

Codebook generation and encoding. Fix n and let $p_i^*(u_i)$, $i = 1, \ldots, n$, and $f_i^* : (u_i, x^{i-1}, y^{i-1}) \mapsto x_i$, $i = 1, \ldots, n$, achieve the maximum of (13). We will also use the notation $p^*(u^n) = \prod_{i=1}^{n} p_i^*(u_i)$ and $f^*(u^n, x^{n-1}, y^{n-1}) = (f_1^*(u_1), \ldots, f_n^*(u_n, x^{n-1}, y^{n-1}))$.

For each $k = k(L, n) = Ln^2 + n$, $L = 1, 2, \ldots$, we generate a $(2^{kR}, k)$ code $\{X_i(W, Y^{i-1})\}_{i=1}^{k}$ as summarized in

[2] that is, $n_1 C_{\text{FB},n_1} + n_2 C_{\text{FB},n_2} \le (n_1+n_2) C_{\text{FB},n_1+n_2}$, which is an easy consequence of the stationarity of the process $\{Z_i\}_{i=1}^{\infty}$ and the definition of the channel model (6), in particular, $Y_i = \emptyset, i = 1, \ldots, m$.

Figure 2. Here $\tilde{\mathbf{X}}^{(n)}$, $\tilde{\mathbf{Y}}^{(n)}$, and $\tilde{\mathbf{Z}}^{(n)}$ are copies the underlying sequences $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ with every $(Ln+1)st$ symbol omitted.

For each $w \in \{1, 2, \ldots, 2^{kR}\}$, 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

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

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

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

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

$i = 1, \ldots, Ln$, $j = 1, \ldots, n$. 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 Figure 2).

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 $p^*(u^n)p(z^n)$, $x_i = f_i^*(u_i, x^{i-1}, y^{i-1})$, and the definition of the channel (6). Otherwise, an error is declared.

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

$$E_i = \{(\mathbf{U}^{(n)}(i), \tilde{\mathbf{Y}}^{(n)}) \in A_\epsilon^{*(Ln)}(U^n, Y^n)\}, \quad i \in [2^{kR}]$$

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

From [10, 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, $\mathbf{U}^{(n)}(i)$, $i \neq 1$, is generated blockwise i.i.d. $\sim p^*(u^n)$ independent of $\mathbf{Y}^{(n)}$. Hence, from [4, Lemma 10.6.2], the probability of the event E_i that $\mathbf{U}^{(n)}(i)$ is jointly typical with $\mathbf{Y}^{(n)}$ is bounded by

$$\Pr(E_i) \leq 2^{-Ln(I(U^n;Y^n)-\delta)}, \qquad \text{for all } i \neq 1,$$

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

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

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 $R < C'_{\text{FB},n}$.

Finally by Lemma 1, this implies that we can achieve

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

which completes the proof of Theorem 1.

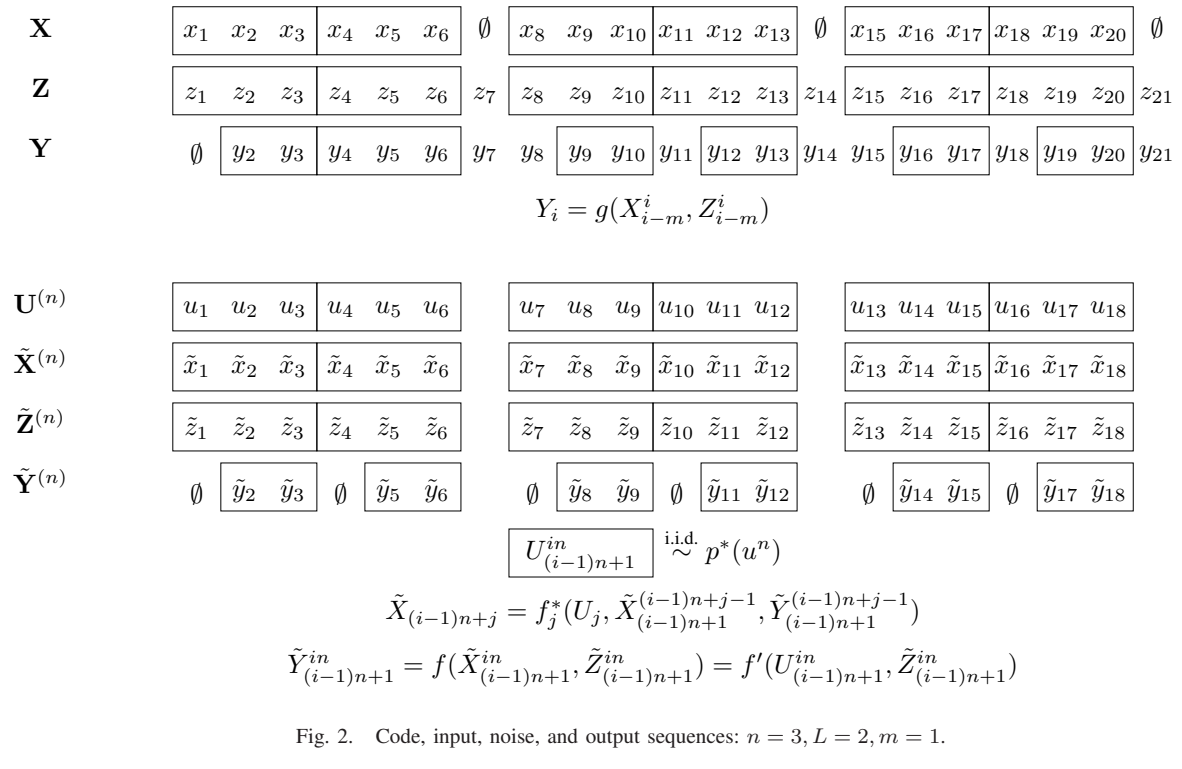

Fig. 2. Code, input, noise, and output sequences: $n = 3, L = 2, m = 1$.

III. Concluding Remarks

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

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

and presented a simple and constructive proof of the feedback coding theorem. The Shannon strategy (Lemma 1) has a fundamental role in transforming the feedback coding problem into a nonfeedback one, which is then solved by a scalable coding scheme of constructing a long typical input-output sequence pair by concatenating shorter nonergodic ones with appropriate phase shifts.

This two-stage approach can be applied to other channel models and can give a straightforward coding theorem. For example, we can show that the semi-deterministic 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 the assumption of indecomposability) has the feedback capacity lower bounded by

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

This result was previously shown by Permuter *et al.* [15, Section V] via a generalization of Gallager's random coding exponent method for finite state channels without feedback [6, Section 5.9]. Here we sketch a simple alternative proof.

From a trivial modification of Lemma 1, the problem reduces to showing that

$$\max_{p(u^n), x_i = f(u_i, x^{i-1}, y^{i-1})} \min_{s_0} \frac{1}{n} I(U^n; Y^n|s_0) \qquad (14)$$

is achievable for each n. But the given Shannon strategy $(p^*(u^n), x^n = f^*(u^n, x^{n-1}, y^{n-1}))$ 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

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

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 $S_n = f(S_{n-1}, X_n, Y_n)$.

We finally mention an important question that is not dealt with in this paper. Our characterization of the feedback capacity

$$C_{\text{FB}} = \lim_{n\to\infty} \max_{p(x^n||y^{n-1})} \frac{1}{n} I(X^n \to Y^n) \qquad (15)$$

or any similar multi-letter 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 (15). This problem has been solved for a few specific channel models such as certain classes of finite-state channels [3], [20], [15], [14] and stationary additive Gaussian noise channels [8], [9], 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.

Acknowledgment

The author wishes to thank Tom Cover, Bob Gray, and Haim Permuter for helpful discussions.

References

[1] D. Blackwell, "Information theory," in *Modern Mathematics for the Engineer: Second Series*. New York: McGraw-Hill, 1961, pp. 182–193.

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

[3] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. IT-51, no. 3, pp. 780–798, Mar. 2005.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[5] A. Feinstein, "On the coding theorem and its converse for finite-memory channels," *Information and Control*, vol. 2, pp. 25–44, 1959.

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

[7] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[8] ——, "Feedback capacity of the first-order moving average Gaussian channel," *IEEE Trans. Inf. Theory*, vol. IT-52, no. 7, pp. 3063–3079, 2006.

[9] ——, "Feedback capacity of stationary Gaussian channels," submitted to *IEEE Trans. Inf. Theory,* February 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0602091/

[10] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," submitted to *IEEE Trans. Inf. Theory,* January 2007. [Online]. Available: http://arxiv.org/abs/cs.IT/0701041/

[11] J. L. Massey, "Causality, feedback, and directed information," in *Proc. International Symposium on Information Theory and its Applications*, Honolulu, Hawaii, Nov. 1990, pp. 303–305.

[12] J. Nedoma, "Über die Ergodizität und r-Ergodizität stationärer Wahrscheinlichkeitsmasse," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 2, pp. 90–97, 1963.

[13] D. L. Neuhoff and P. C. Shields, "Channels with almost finite memory," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 4, pp. 440–447, 1979.

[14] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," submitted to *IEEE Trans. Inform. Theory*, 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0610047/

[15] H. Permuter, T. Weissman, and A. Goldsmith, "Finite-state channels with time-invariant deterministic feedback," submitted to *IEEE Trans. Inform. Theory*, 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0608070/

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

[17] S. Tatikonda, "Control under communication constraints," Ph.D. Thesis, Massachusetts Institute of Technology, Sept. 2000.

[18] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," Sept. 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0609139/

[19] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. IT-40, no. 4, pp. 1147–1157, July 1994.

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