On Evaluating the Rate-Distortion Function of Sources with Feed-Forward and the Capacity of Channels with Feedback.

Ramji Venkataramanan and S. Sandeep Pradhan
Department of EECS, University of Michigan, Ann Arbor, MI 48105
rvenkata@umich.edu, pradhanv@eecs.umich.edu

"THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD" ***Abstract*— We study the problem of computing the rate-distortion function for sources with feed-forward and the capacity for channels with feedback. The formulas (involving directed information) for the optimal rate-distortion function with feed-forward and channel capacity with feedback are multi-letter expressions and cannot be computed easily in general. In this work, we derive conditions under which these can be computed for a large class of sources/channels with memory and distortion/cost measures. Illustrative examples are also provided.**

I. Introduction

Feedback is widely used in communication systems to help combat the effect of noisy channels. It is well-known that feedback does not increase the capacity of a discrete memoryless channel [1]. However, feedback could increase the capacity of a channel with memory. Recently, directed information has been used to elegantly characterize the capacity of channels with feedback [2], [3], [4]. The source coding counterpart to channel coding with feedback- source coding with feed-forward- has recently been studied in [5], [6], [7], [8]. The optimal rate-distortion function with feed-forward was characterized using directed information in [6].

In this work, we study the problem of computing these rate-distortion and capacity expressions. The formulas (involving directed information) for the optimal rate-distortion function with feed-forward [6] and channel capacity with feedback [4] are multi-letter expressions and cannot be computed easily in general. We derive conditions under which these can be computed for a large class of sources (channels) with memory and distortion (cost) measures. We also provide illustrative examples. Throughout, we consider source feed-forward and channel feedback with arbitrary delay. When the delay goes to ∞, we obtain the case of no feed-forward/feedback.

II. Source Coding with Feed-Forward

A. Problem Formulation

In simple terms, source coding with feed-forward is the source coding problem in which the decoder gets to observe some past source samples to help reconstruct the present sample. Consider a general discrete source X with alphabet $\mathcal{X}$ and reconstruction alphabet $\widehat{\mathcal{X}}$. The source is characterized by a sequence of distributions denoted by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$. There is an associated sequence of distortion measures $d_n : \mathcal{X}^n \times \widehat{\mathcal{X}}^n \to \mathbb{R}^+$. It is assumed that $d_n(x^n, \hat{x}^n)$ is normalized with respect to n and is uniformly bounded in n. For example $d_n(x^n, \hat{x}^n)$ may be the average per-letter distortion, i.e., $\frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i)$ for some $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$.

Definition 1: An $(N, 2^{NR})$ source code with delay k feed-forward of block length N and rate R consists of an encoder mapping e and a sequence of decoder mappings $g_i, i = 1, \ldots, N$, where

$$e : \mathcal{X}^N \to \{1, \ldots, 2^{NR}\}$$
$$g_i : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{i-k} \to \widehat{\mathcal{X}}, \quad i = 1, \ldots, N.$$

The encoder maps each N-length source sequence to an index in $\{1, \ldots, 2^{NR}\}$. The decoder receives the index transmitted by the encoder, and to reconstruct the ith sample, it has access to the source samples until time $(i-k)$. We want to minimize R for a given distortion constraint.

Definition 2: (Probability of error criterion) R is an ϵ-achievable rate at probability-1 distortion D if for all sufficiently large N, there exists an $(N, 2^{NR})$ source codebook such that

$$P_{X^N}\left(x^N : d_N(x^N, \hat{x}^N) > D\right) < \epsilon,$$

where $\hat{x}^N$ denotes the reconstruction of x^N.
R is an achievable rate at probability-1 distortion D if it is ϵ-achievable for every $\epsilon > 0$.

We now give a brief summary of the rate-distortion results with feed-forward found in [6]. The rate-distortion function with feed-forward (delay 1) is characterized by directed information, a quantity defined in [2]. The directed information flowing from a random sequence $\hat{X}^N$ to a random sequence X^N is defined as

$$I(\hat{X}^N \to X^N) = \sum_{n=1}^{N} I(\hat{X}^n; X_n | X^{n-1}). \tag{1}$$

When the feed-forward delay is k, the rate-distortion function is characterized by the $k-$delay version of the directed

This work was supported by NSF Grant ITR-0427385 and Grant (CAREER) CCF-0448115.

information:

$$I_k(\hat{X}^N \to X^N) = \sum_{n=1}^{N} I(\hat{X}^{n+k-1}; X_n | X^{n-1}). \qquad (2)$$

When we do not make any assumption on the nature of the joint process $\{\mathbf{X}, \hat{\mathbf{X}}\}$, we need to use the information spectrum version of (2). In particular, we will need the quantity[1]

$$\overline{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}}{\vec{P}^k_{\hat{X}^n|X^n} \cdot P_{X^n}}, \qquad (3)$$

where

$$\vec{P}^k_{\hat{X}^n|X^n} = \prod_{i=1}^{n} P_{\hat{X}_i|\hat{X}^{i-1}, X^{i-k}}.$$

It should be noted that (2) and (3) are the same when the joint process $\{\mathbf{X}, \hat{\mathbf{X}}\}$ is stationary and ergodic.

Theorem 1: [6] For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with feed-forward, the infimum of all achievable rates at probability-1 distortion D, is given by

$$R_{ff}(D) = \inf_{\mathbf{P_{\hat{X}|X}}: \rho(\mathbf{P_{\hat{X}|X}}) \leq D} \overline{I}_k(\hat{X} \to X), \qquad (4)$$

where

$$\begin{aligned} &\rho(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) \\ &= \inf \left\{ h : \lim_{n \to \infty} P_{X^n,\hat{X}^n}\left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h\right) = 0 \right\}. \end{aligned} \qquad (5)$$

B. Evaluating the Rate-Distortion Function with Feed-forward

The rate-distortion formula in Theorem 1 is an optimization of a multi-letter expression:

$$\overline{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}}{\vec{P}^k_{\hat{X}^n|X^n} \cdot P_{X^n}},$$

This is an optimization over an infinite dimensional space of conditional distributions $\mathbf{P_{\hat{X}|X}}$. Since this is a potentially difficult optimization, we turn the problem on its head and pose the following question:

Given a source X with distribution $\mathbf{P_X}$ and a conditional distribution $\mathbf{P_{\hat{X}|X}}$, for what sequence of distortion measures does $\mathbf{P_{\hat{X}|X}}$ achieve the infimum in the rate-distortion formula ?

A similar approach is used in [9] (Problem 2 and 3, p. 147) to find optimizing distributions for discrete memoryless channels and sources without feedback/feed-forward. It is also used in [10] to study the optimality of transmitting uncoded source data over channels and in [11] to study the duality between source and channel coding.

Given a source X, suppose we have a hunch about the structure of the optimal conditional distribution. The following theorem (proof omitted) provides the distortion measures for which our hunch is correct.

[1]The $\limsup_{inprob}$ of a random sequence A_n is defined as the smallest number α such that $P(A_n \geq \alpha + \epsilon) = 0$ for all $\epsilon > 0$ and is denoted $\overline{A}$.

Theorem 2: Suppose we are given a stationary, ergodic source X characterized by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$ with feed-forward delay k. Let $\mathbf{P_{\hat{X}|X}} = \{P_{X^n|X^n}\}_{n=1}^{\infty}$ be a conditional distribution such that the joint distribution is stationary and ergodic. Then $\mathbf{P_{\hat{X}|X}}$ achieves the rate-distortion function if for all sufficiently large n, the distortion measure satisfies

$$d_n(x^n, \hat{x}^n) = -c \cdot \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}(x^n, \hat{x}^n)}{\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n)} + d_0(x^n), \qquad (6)$$

where

$$\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \prod_{i=1}^{n} P_{\hat{X}_i|X^{i-k},\hat{X}^{i-1}}(\hat{x}_i|x^{i-k}, \hat{x}^{i-1}),$$

c is any positive number and $d_0(.)$ is an arbitrary function.

C. Markov Sources with Feed-forward

A stationary, ergodic mth order Markov source X is characterized by a distribution $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$ where

$$P_{X^n} = \prod_{i=1}^{n} P_{X_i|X_{i-m}^{i-1}}, \quad \forall n. \qquad (7)$$

Let the source have feed-forward with delay k. We first ask: *When is the optimal joint distribution also mth order Markov in the following sense*:

$$P_{X^n,\hat{X}^n} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i|X_{i-m}^{i-1}}, \quad \forall n. \qquad (8)$$

In other words, when does the optimizing conditional distribution have the form

$$P_{\hat{X}^n|X^n} = \prod_{i=1}^{n} P_{\hat{X}_i|X_{i-m}^{i}}, \quad \forall n. \qquad (9)$$

The answer, provided by Theorem 2, is stated below. In the sequel, we drop the subscripts on the probabilities to keep the notation clean.

Corollary 1: For an mth order Markov source (described in (7)) with feed-forward delay k, an mth order conditional distribution (described in (9)) achieves the optimum in the rate-distortion function for a sequence of distortion measures $\{d_n\}$ given by

$$d_n(x^n, \hat{x}^n) = -c \cdot \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(x_i, \hat{x}_i | x_{i-m}^{i-1})}{P(\hat{x}_i | \hat{x}_{i-k+1}^{i-1}, x_{i-k+1-m}^{i-k})} + d_0(x^n), \qquad (10)$$

where c is any positive number and $d_0(.)$ is an arbitrary function.

Proof: The proof involves substituting (7) and (9) in (6) and performing a few manipulations.

In the following section, we provide two examples to illustrate how Theorem 2 can be used to determine the rate-distortion function of sources with feed-forward.

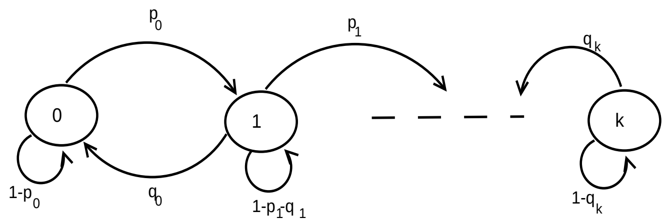

Fig. 1. Markov chain representing the stock value

TABLE I
DISTORTION $e(\hat{x}_i, x_{i-1} = j, x_i)$

	(x_{i-1}, x_i) $j, j+1$	j, j	$j, j-1$
$\hat{x}_i = 0$	0	0	1
$\hat{x}_i = 1$	1	1	0

III. Examples

A. Stock-market example

Suppose that we wish to observe the behavior of a particular stock in the stock market over an $N-$day period. Assume that the value of the stock can take $k+1$ different values and is modeled as a $k+1$-state Markov chain, as shown in Fig. 1. If on a particular day, the stock is in state i, $1 \leq i < k$, then on the next day, one of the following can happen.

- The value increases to state $i+1$ with probability p_i.
- The value drops to state $i-1$ with probability q_i.
- The value remains the same with probability $1 - p_i - q_i$.

When the stock-value is in state 0, the value cannot decrease. Similarly, when in state k, the value cannot increase. Suppose an investor invests in this stock over an $N-$day period and desires to be forewarned whenever the value drops. Assume that there is an insider (with a priori information about the behavior of the stock) who can send information to the investor at a finite rate.

The value of the stock is modeled as a Markov source $\mathbf{X} = \{X_n\}$. The decision $\hat{X}_n$ of the investor is binary: $\hat{X}_n = 1$ indicates that the price is going to drop from day $n-1$ to n, $\hat{X}_n = 0$ means otherwise. Before day n, the investor knows all the previous values of the stock X^{n-1} and has to make the decision $\hat{X}_n$. Thus feed-forward is automatically built into the problem.

The investor makes an error either when he fails to predict a drop or when he falsely predicts a drop. The distortion is modeled using a Hamming distortion criterion as follows.

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} e(\hat{x}_i, x_{i-1}, x_i), \qquad (11)$$

where $e(.,.,.)$ is the *per-letter* distortion given Table I. The minimum amount of information (in bits/sample) the insider needs to convey to the investor so that he can predict drops in value with distortion D is denoted $R_{ff}(D)$.

Proposition 1: For the stock-market problem described above,

$$R_{ff}(D) = \sum_{i=1}^{k-1} \pi_i \left(h(p_i, q_i, 1 - p_i - q_i) - h(\epsilon, 1 - \epsilon) \right) + \pi_k \left(h(q_k, 1 - q_k) - h(\epsilon, 1 - \epsilon) \right),$$

where $h()$ is the entropy function, $[\pi_0, \pi_1, \cdots, \pi_k]$ is the stationary distribution of the Markov chain and $\epsilon = \frac{D}{1-\pi_0}$.

Proof: We will use Corollary 1 to verify that a first-order Markov conditional distribution of the form

$$P_{\hat{X}_n | \hat{X}^{n-1}, X^n} = P_{\hat{X}_n | X_n, X_{n-1}}, \quad \forall n \qquad (12)$$

achieves the optimum.

Due to the structure of the distortion function in Table I, we can guess the structure of $P(x_i | \hat{x}_i, x_{i-1})$ as follows. When $X_{i-1} = 0$, the decoder can always declare $\hat{X}_i = 0$ - there is no error irrespective of the value of X_i. So we assign $P(\hat{X}_i = 0 | x_{i-1} = 0, x_i = 0) = P(\hat{X}_i = 0 | x_{i-1} = 0, x_i = 1) = 1$, which gives

$$P(X_i = 0 | x_{i-1} = 0, \hat{x}_i = 0) = 1 - p.$$

The event $(X_{i-1} = 0, \hat{X}_i = 1)$ has zero probability. Thus we obtain the first two columns of Table II.

When $(X_{i-1} = j, \hat{X}_i = 0)$, $0 < j < k$, an error occurs when $X_i = j - 1$. This is assigned a probability ϵ. The remaining probability $1 - \epsilon$ is split between $P(X_i = j | x_{i-1} = j, \hat{x}_i = 0)$ and $P(X_i = j+1 | x_{i-1} = j, \hat{x}_i = 0)$ according to their transition probabilities. In a similar fashion, we obtain all the columns in Table II.

We now show that the distortion criterion (11) can be cast in the form

$$d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} \Big(-c \log_2 P(x_i | \hat{x}_i, x_{i-1}) + d_0(x_{i-1}, x_i) \Big), \qquad (13)$$

or equivalently

$$e(\hat{x}_i, x_{i-1}, x_i) = -c \log_2 P(x_i | \hat{x}_i, x_{i-1}) + d_0(x_{i-1}, x_i), \qquad (14)$$

thereby proving that the distribution in Table II is optimal. This is done by substituting values from Tables I and II into (14) to determine c and $d_0(.,.)$.

Since the process $\{\mathbf{X}, \hat{\mathbf{X}}\}$ is jointly stationary and ergodic, the distortion constraint is equivalent to

$$E[e(\hat{x}_2, x_1, x_2)] \leq D.$$

To calculate the expected distortion

$$E[e(\hat{x}_2, x_1, x_2)] = \sum_{x_1, x_2, \hat{x}_2} P(x_1, x_2) P(\hat{x}_2 | x_1, x_2) \cdot e(\hat{x}_2, x_1, x_2), \qquad (15)$$

we need the (optimum achieving) conditional distribution $P(\hat{X}_2 | x_1, x_2)$. This is found by substituting the values from Tables I and II in the relation

$$P(x_2 | x_1, \hat{x}_2) = \frac{P(x_2 | x_1) P(\hat{x}_2 | x_2, x_1)}{\sum_{x_2} P(x_2 | x_1) P(\hat{x}_2 | x_2, x_1)}. \qquad (16)$$

TABLE II

THE DISTRIBUTION $P\left(X_i|x_{i-1},\hat{x}_i\right)$

	$(x_{i-1},\hat{x}_i)$							
	$0,0$	$0,1$	$\cdots$	$j,0$	$j,1$	$\cdots$	$k,0$	$k,1$
$x_i=0$	$1-p$	$-$	$\cdots$	$-$	$-$	$-$	$-$	$-$
$x_i=1$	p	$-$	$\cdots$	$-$	$-$	$-$	$-$	$-$
$x_i=\vdots$	$-$	$-$	$\ddots$	$-$	$-$	$-$	$-$	$-$
$x_i=j-1$	$-$	$-$	$-$	ϵ	$1-\epsilon$	$-$	$-$	$-$
$x_i=j$	$-$	$-$	$-$	$\frac{(1-\epsilon)(1-p_j-q_j)}{1-q_j}$	$\frac{\epsilon(1-p_j-q_j)}{1-q_j}$	$-$	$-$	$-$
$x_i=j+1$	$-$	$-$	$-$	$\frac{(1-\epsilon)p_j}{1-q_j}$	$\frac{\epsilon p_j}{1-q_j}$	$-$	$-$	$-$
$x_i=\vdots$	$-$	$-$	$-$	$-$	$-$	$\ddots$	$-$	$-$
$x_i=k-1$	$-$	$-$	$\cdots$	$-$	$-$	$-$	ϵ	$1-\epsilon$
$x_i=k$	$-$	$-$	$\cdots$	$-$	$-$	$-$	$1-\epsilon$	ϵ

TABLE III

THE CONDITIONAL DISTRIBUTION $P(\hat{X}_i|x_{i-1},x_i)$

	(x_{i-1},x_i)								
	$0,0$	$0,1$	$\cdots$	$j,j-1$	j,j	$j,j+1$	$\cdots$	$k,k-1$	k,k
$\hat{x}_i=0$	1	1	$\cdots$	$\frac{\epsilon(1-q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\cdots$	$\frac{\epsilon(1-q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{(1-\epsilon)(1-q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$
$\hat{x}_i=1$	0	0	$\cdots$	$\frac{(1-\epsilon)(q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$	$\cdots$	$\frac{(1-\epsilon)(q_j-\epsilon)}{q_j(1-2\epsilon)}$	$\frac{\epsilon(q_j-\epsilon)}{(1-q_j)(1-2\epsilon)}$

Thus we obtain the conditional distribution $P(\hat{X}_2|x_1,x_2)$ shown in Table III. Using this in (15), we get

$$E[e(\hat{x}_2,x_1,x_2)]=(1-\pi_0)\epsilon\le D \tag{17}$$

We can now calculate the rate distortion function as

$$\begin{aligned} R_{ff}(D) &= \sum_{x_1,x_2,\hat{x}_2} P(x_1,x_2,\hat{x}_2)\log_2\frac{P(x_2|x_1,\hat{x}_2)}{P(x_2|x_1)} \\ &= H(X_2|X_1)-H(X_2|\hat{X}_2,X_1) \end{aligned} \tag{18}$$

to obtain the expression in Proposition 1. ∎

B. Gauss-Markov Source

Consider a stationary, ergodic, first-order Gauss-Markov source X with mean 0, correlation ρ and variance σ^2:

$$X_n=\rho X_{n-1}+N_n,\quad \forall n, \tag{19}$$

where $\{N_n\}$ are independent, identically distributed Gaussian random variables with mean 0 and variance $(1-\rho^2)\sigma^2$. Assume the source has feed-forward with delay 1 and we want to reconstruct at every time instant n the linear combination aX_n+bX_{n-1}, for any constants a,b. We use the mean-squared error distortion criterion:

$$d_n(x^n,\hat{x}^n)=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i-(ax_i+bx_{i-1})\right)^2. \tag{20}$$

The feed-forward distortion-rate function for this source with average mean-squared error distortion was given in [5]. The feed-forward rate-distortion function can also be obtained using Theorem 2 (proof omitted):

$$R_{ff}(D)=\frac{1}{2}\log\frac{\sigma^2(1-\rho^2)}{D/a^2}. \tag{21}$$

We must mention here that the rate-distortion function in the first example cannot be computed using the techniques in [5].

IV. Channel Coding with Feedback

In this section, we consider channels with feedback and the problem of evaluating their capacity. A channel is defined as a sequence of probability distributions:

$$P^{ch}_{\mathbf{Y}|\mathbf{X}}=\{P^{ch}_{Y_n|X^n,Y^{n-1}}\}_{n=1}^{\infty}. \tag{22}$$

In the above, X_n and Y_n are the channel input and output symbols at time n, respectively. The channel is assumed to have $k-$delay feedback ($1\le k<\infty$). This means at time instant n, the encoder has perfect knowledge of the channel outputs until time $n-k$ to produce the input x_n. The input distribution to the channel is denoted by $P^k_{\mathbf{X}|\mathbf{Y}}=\{P_{X_n|X^{n-1},Y^{n-k}}\}_{n=1}^{\infty}$. In the sequel, we will need the following product quantities corresponding to the channel and the input.

$$\begin{aligned} \vec{P}^{ch}_{Y^n|X^n} &\triangleq \prod_{i=1}^{n} P_{Y_i|X^i,Y^{i-1}}, \\ \vec{P}^{k}_{X^n|Y^n} &\triangleq \prod_{i=1}^{n} P_{X_i|X^{i-1},Y^{i-k}}. \end{aligned} \tag{23}$$

The joint distribution of the system is given by $P_{\mathbf{X},\mathbf{Y}}=\{P_{X^n,Y^n}\}_{n=1}^{\infty}$, where

$$P_{X^n,Y^n}=\vec{P}^k_{X^n|Y^n}\cdot\vec{P}^{ch}_{Y^n|X^n}. \tag{24}$$

Definition 3: An $(N,2^{NR})$ channel code with delay k feed-forward of block length N and rate R consists of a sequence of encoder mappings $e_i, _1=1,\ldots,N$ and a decoder g, where

$$\begin{aligned} e_i:\{1,\ldots,2^{NR}\}\times\mathcal{Y}^{i-k}\to\mathcal{X},\quad i=1,\ldots,N \\ g:\mathcal{Y}^N\to\{1,\ldots,2^{NR}\} \end{aligned}$$

Thus it is desired to transmit one of 2^{NR} messages over the channel in N units of time. There is an associated cost function for using the channel given by $c_N(X^N, Y^N)$. For example, this could be the average power of the input symbols.

If W is the message that was transmitted, then the probability of error is

$$P_e = Pr(g(Y^N) \neq W).$$

Definition 4: R is an (ϵ, δ)-achievable rate at probability-1 cost C if for all sufficiently large N, there exists an $(N, 2^{NR})$ channel code such that

$$\begin{aligned} P_e &< \epsilon, \\ Pr(c_N(X^N, Y^N) > C) &< \delta \end{aligned}$$

R is an achievable rate at probability-1 cost C if it is (ϵ, δ)-achievable for every $\epsilon, \delta > 0$.

Theorem 3: [6], [4] For an arbitrary channel $P^{ch}_{\mathbf{Y}|\mathbf{X}}$, the capacity with $k-$delay feedback, the infimum of all achievable rates at probability-1 cost C, is given by

$$C_{fb}(C) = \sup_{P^k_{\mathbf{X}|\mathbf{Y}}: \rho(P^k_{\mathbf{X}|\mathbf{Y}}) \leq C} \underline{I}(X \rightarrow Y), \tag{25}$$

where[2]

$$\underline{I}(X \rightarrow Y) \triangleq \liminf_{inprob} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}$$

and

$$\begin{aligned} \rho(P^k_{\mathbf{X}|\mathbf{Y}}) &\triangleq \limsup_{inprob} c_n(X^n, Y^n) \\ &= \inf\{h : \lim_{n \to \infty} P_{X^n Y^n}\left((x^n, y^n) : c_n(x^n, y^n) > h\right)\} = 0. \end{aligned}$$

In the above, we note that

$$P_{Y^n} = \sum_{X^n} P_{X^n, Y^n} = \sum_{X^n} \vec{P}^k_{X^n|Y^n} \cdot \vec{P}^{ch}_{Y^n|X^n}.$$

A. Evaluating the Channel Capacity with Feedback

The capacity formula in Theorem 3 is a multi-letter expression involving optimizing the function $\underline{I}(X \rightarrow Y)$ over an infinite dimensional space of input distributions $P^k_{\mathbf{X}|\mathbf{Y}}$. Just like we did with sources, we can pose the following question: *Given a channel $P^{ch}_{\mathbf{Y}|\mathbf{X}}$ and an input distribution $P^k_{\mathbf{X}|\mathbf{Y}}$, for what sequence of cost measures does $P^k_{\mathbf{X}|\mathbf{Y}}$ achieve the supremum in the capacity formula ?*

The following theorem (proof omitted) provides an answer.

Theorem 4: Suppose we are given a channel $P^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback and an input distribution $P^k_{\mathbf{X}|\mathbf{Y}}$ such that the joint process $P_{\mathbf{X},\mathbf{Y}}$ given by (24) is stationary, ergodic. Then the input distribution $P^k_{\mathbf{X}|\mathbf{Y}}$ achieves the $k-$delay feedback capacity of the channel if for all sufficiently large n, the cost measure satisfies

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)} + d_0, \tag{26}$$

where λ is any positive number and d_0 is an arbitrary constant.

[2]The $\liminf_{inprob}$ of a random sequence A_n is defined as the largest number α such that $P(A_n \leq \alpha - \epsilon) = 0$ for all $\epsilon > 0$ and is denoted $\underline{A}$.

B. Markov channels with feedback

The problem of evaluating the capacity of finite state machine channels was studied recently in [12] and [13]. In [12], it was shown that the capacity of such a channel is achieved by a feedback dependent Markov source, i.e. , the optimal input distribution is of the form $\{P_{X_n|X_{n-1},Y^{n-1}}\}$, i.e., Markov in X but depends on all the past Y symbols.

We consider a simple Markov channel with feedback delay 1 and the problem of evaluating its capacity. The channel we study is characterized by

$$P^{ch}_{Y_n|X^n,Y^{n-1}} = P^{ch}_{Y_n|X_n,Y_{n-1}}. \tag{27}$$

Let the channel have feedback with delay 1. We are interested in finding cost measures for which the capacity of the channel in (27) is easily evaluated. We first ask: *When is the optimal joint distribution first order Markov in the following sense*:

$$P_{X^n,Y^n} = \prod_{i=1}^{n} P_{X_i,Y_i|Y_{i-1}}, \quad \forall n. \tag{28}$$

In other words, when does the optimizing input distribution to have the form

$$\vec{P}_{X^n|Y^n} = \prod_{i=1}^{n} P_{X_i|Y_{i-1}}, \quad \forall n. \tag{29}$$

From Theorem 4, it is seen that this happens when the cost-function has the form:

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \sum_{i=1}^{n} \log \frac{\vec{P}^{ch}_{Y_i|X_i,Y_{i-1}}(y_i|x_i, y_{i-1})}{P_{Y_i|Y_{i-1}}(y_i|y_{i-1})} + d_0. \tag{30}$$

References

[1] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. IT-2, pp. 8–19, 1956.

[2] J. Massey, "Causality, Feedback and Directed Information," *Proc. 1990 Symp. on Inf. Theory and Applications (ISITA-90)*, pp. 303–305, 1990.

[3] G. Kramer, *Directed Information for channels with Feedback*. PhD thesis, Swiss Federal Institute of Technology, Zurich, 1998.

[4] S. Tatikonda, *Control Under Communications Constraints*. PhD thesis, Massachusetts Inst. of Technology, Cambridge, MA, September 2000.

[5] T. Weissman and N. Merhav, "On competitive prediction and its relation to rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. IT-49, pp. 3185–3194, December 2003.

[6] R. Venkataramanan and S. S. Pradhan, "Directed Information for Communication problems with Side-information and Feedback/Feed-forward," *Proc. 43rd Annual Allerton Conf. (Monticello, IL)*, 2005.

[7] E. Martinian and G. W. Wornell, "Source Coding with Fixed Lag Side Information," *Proc. 42nd Annual Allerton Conf. (Monticello, IL)*, 2004.

[8] S. S. Pradhan, "Source coding with feedforward: Gaussian sources," in *Proc. IEEE Int. Symp. on Inf. Theory*, p. 212, June 2004.

[9] I. Csisz'ar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[10] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited.," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, 2003.

[11] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side-information case," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1181–1203, May 2003.

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

[13] 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, 2005.