Transactions Letters

On Computing the Feedback Capacity of Channels and the Feed-Forward Rate-Distortion Function of Sources

Ramji Venkataramanan, *Member, IEEE,* and S. Sandeep Pradhan, *Member, IEEE*

***Abstract*—The problem of computing the capacity-cost function of channels with feedback and the rate-distortion function of sources with feed-forward is considered. Sufficient conditions are derived on : a) the structure of the cost function for a chosen joint distribution to achieve the optimal feedback capacity-cost function, b) the structure of the distortion function for a chosen joint distribution to achieve the optimal feed-forward rate-distortion function. These structural results are useful since it is infeasible in general to directly compute the optimizations. Examples are provided to show how the results can help compute the performance limits with feedback and feed-forward.**

***Index Terms*—Feed-forward, feedback, computation.**

I. Introduction

FEEDBACK is widely used to improve the reliability of transmission over noisy channels. Though Shannon showed that feedback cannot increase the capacity of a memoryless point-to-point channel [1], feedback can increase the capacity of a channel with memory. Marko was one of the first to develop tools to study feedback in his effort to develop a bidirectional theory of communication[2]. Inspired by Marko's work, Massey [3] introduced the concept of directed information and used it to upper bound the feedback capacity of a point-to-point channel. Tatikonda later established the capacity of a general point-to-point channel with feedback in terms of the directed information flowing from the input to the output [4], [5]. The literature on feedback capacity is vast and we shall not review it here. An incomplete list includes [6]–[9].

Source coding with feed-forward, the dual problem of channel coding with feedback, has been studied recently in [10]–[13]. The notion of feed-forward can be explained in simple terms as follows. Consider a source X to be compressed as $\hat{X}$, within some distortion D. The encoder takes a block of, say, N source samples and maps it to an index in a codebook. The decoder uses this index and reconstructs the N source samples *sequentially*: in order to reconstruct each source sample, the decoder has access to the index *and* some past source samples. More precisely, let $X_n, \hat{X}_n$ denote the source and reconstruction samples at time n, respectively. To produce $\hat{X}_n$, the decoder has knowledge of the index plus the source samples until time $(n-k)$. We call this set-up feed-forward with delay k.

Source coding with feed-forward was first considered in the context of competitive prediction in [10], and later studied in [11]–[13] as a variant of source coding with side information. Later, we shall present an example of feed-forward related to predicting transitions in a Markov chain. The problems of feed-forward and feedback are closely related. In each of these problems, there is a dynamic aspect to either the encoder (feedback) or the decoder (feed-forward). Directed information, which captures the causal flow of information between random sequences, is the information quantity that characterizes the performance limit of both these problems [4], [13].

In this letter, we consider the problem of computing the feedback capacity obtained in [5], and the feed-forward rate-distortion function obtained in [13]. Perhaps the most appealing feature of Shannon's formulas (for channel capacity and source rate-distortion function) is the simplicity of the optimizations involved. The capacity-cost function of a memoryless channel and the rate-distortion function of a memoryless source have 'single-letter' formulas, i.e., we need to optimize over probability distributions of a finite number of random variables. These single-letter optimizations can be computed efficiently using techniques such as the Blahut-Arimoto algorithm [14].

In contrast, to achieve the feedback capacity, we need to optimize over the space of all input policies. In other words, for each time n, we need to pick a function f_n that generates the nth channel input using the message and the past channel outputs. Similarly, to achieve the feed-forward rate-distortion functions, we need to optimize over all reconstruction policies. The space of all input/reconstruction policies is extremely large. The contribution of [5] and [13] was to show that the feedback capacity and feed-forward rate-distortion function can be expressed as optimizations of multi-letter information-theoretic quantities (directed information) over the space of valid distributions. While these formulas are considerably simpler than optimizing over all policies, they are not optimizations over finite-dimensional spaces, and are difficult to compute. There are two reasons we cannot obtain simple, single-letter expressions for these problems:

1) With feedback and feed-forward, the channels and sources of interest are those with memory.

Paper approved by F. Alajaji, the Editor for Source and Source/Channel Coding of the IEEE Communications Society. Manuscript received July 20, 2009; revised December 6, 2009.

R. Venkataramanan was a Ph.D. student at the EECS Dept., University of Michigan, Ann Arbor, and is now at Stanford University (e-mail: vramji@stanford.edu).

S. S. Pradhan is with the EECS Dept., University of Michigan, Ann Arbor (e-mail: pradhanv@eecs.umich.edu).

This work was supported by NSF Grants CCF-0427385 (ITR), and CCF-0448115 (CAREER). It was presented in part at the IEEE International Symp. on Information Theory, Nice, France, June 2007.

Digital Object Identifier 10.1109/TCOMM.2010.07.090115

2) Due to the dynamics introduced by feed-forward and feedback, one cannot guarantee the optimal joint distributions to be stationary and ergodic in general. Thus the optimization is over all joint processes, not just the stationary and ergodic ones.

Consequently, we cannot expect to have a simple algorithm to compute the performance limit of a general problem with feedback/feed-forward. The main contribution of this letter is a pair of structural results which help compute the feedback capacity and the feed-forward rate-distortion function. In particular, we obtain sufficient conditions on:

- The structure of the cost function for a given joint distribution to achieve the feedback capacity-cost function,
- The structure of the distortion function for a given joint distribution to achieve the optimum feed-forward rate-distortion function.

Related Work: Csiszár and Körner [15, p. 147, Problems 2,3] characterized the cost/distortion function in terms of the optimal joint distribution for discrete memoryless channels/sources without feedback/feed-forward. Our structural results may be viewed as extensions of those in [15] to problems with delayed feedback and feed-forward. We also note that the results of [15] are applied directly in [16] to study the optimality of uncoded transmission of sources over channels, and in [17] to study duality between source and channel coding.

In Section II, we review the capacity result for feedback channels and then present the structural result for the feedback capacity-cost function. In Section III, we review source coding with feed-forward and state the structural result for the optimal rate-distortion function. In Section IV, we give two examples to show how our results can be used to compute the performance limit.

Notation: Upper-case notation will be used for random variables, lower-case for their realizations and bold-face to denote a random process. Thus $\mathbf{A}$ shall denote the process $\{A_n\}_{n=1}^{\infty}$, where A_n represents the nth sample. A^n will denote the random vector $(A_1, \ldots, A_n)$. The pmf of a random variable A_n is denoted P_{A_n}, or $P(A_n)$, when there is no possibility of confusion.

II. Channel Coding with Delayed Feedback

Consider a channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Let X_n, Y_n denote the channel input and output at time n, respectively. The channel is defined by the sequence of conditional distributions $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}} \triangleq \{P^{ch}_{Y_n|X^n,Y^{n-1}}\}_{n=1}^{\infty}$. As shown in Figure 1, the channel has noiseless feedback with delay k ($k \geq 1$). At time $n > k$, the encoder has perfect knowledge of the channel outputs until time $(n-k)$ to produce the channel input X_n.

Definition 1: (a) An $(N, 2^{NR})$ channel code (block length N, rate R) for a channel with feedback delay k consists of a sequence of encoder mappings $e_n, n = 1, \ldots, N$ and a decoder g, where

$$e_n : \{1, \ldots, 2^{NR}\} \times \mathcal{Y}^{n-k} \to \mathcal{X}, \quad n = 1, \ldots, N$$
$$g : \mathcal{Y}^N \to \{1, \ldots, 2^{NR}\}.$$

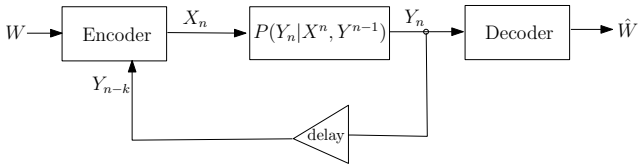

Fig. 1. Channel with delay-k feedback.

In the above and in the sequel, it is understood that $\mathcal{Y}^{n-k} = \phi$, for $n \leq k$. An input distribution for a channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback is a sequence of distributions of the form

$$\mathbf{P}^{k}_{\mathbf{X}|\mathbf{Y}} = \{P_{X_n|X^{n-1},Y^{n-k}}\}_{n=1}^{\infty}. \tag{1}$$

Consider a time-line of how input symbols are produced at the encoder for a message W:

$$X_1(W) \ \ldots \ X_k(W) \ X_{k+1}(W, Y^1) \quad X_{k+2}(W, Y^2) \ldots$$

Thus any channel code corresponds to a unique input distribution. The input distribution of the system until time N specified by $\left\{P_{X_1}, \ldots, P_{X_k|X^{k-1}}, P_{X_(k+1)|X^k,Y^1}, \ldots, P_{X_N|X^{N-1},Y^{N-k}}\right\}$, coupled with the channel distribution, specifies the joint distribution of the input and output at time N as

$$\begin{aligned} P_{X^N,Y^N} &= P_{X_1} \cdot P^{ch}_{Y_1|X_1} \ldots P_{X_{k+1}|X^k,Y^1} \cdot P^{ch}_{Y_{k+1}|X^{k+1},Y^k} \cdots \\ &= \vec{P}^{k}_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}, \end{aligned} \tag{2}$$

where

$$\begin{aligned} \vec{P}^{k}_{X^N|Y^N} &= \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k}}, \\ \vec{P}^{ch}_{Y^N|X^N} &= \prod_{n=1}^{N} P^{ch}_{Y_n|X^n,Y^{n-1}}. \end{aligned} \tag{3}$$

Let $c_N(X^N, Y^N)$ be the cost associated with N uses of the channel. Notice that we allow the cost function at time N to depend on the inputs and the outputs until time N: using feedback, the encoder learns the outputs (with some delay) and can potentially use this information to choose future input symbols so that the cost constraint is satisfied. If W is the message that was transmitted, the probability of error is

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

Definition 2: R is an (ϵ, δ)-achievable rate at cost P with k-delay feedback if for all sufficiently large N, there exists an $(N, 2^{NR})$ channel code such that $P_e < \epsilon$ and $Pr(c_N(X^N, Y^N) > P) < \delta$.

R is an achievable rate at cost P with k-delay feedback if it is (ϵ, δ)-achievable for every $\epsilon, \delta > 0$.

The feedback capacity, the supremum of all achievable rates, was characterized in [4], [5] for a general channel. Recall that the channels we consider have memory and the information available at the encoder changes with time due to feedback. As a result, we cannot assume in general that the optimal joint distribution is stationary and ergodic. A tight capacity result requires the use of information spectrum methods [18]. We briefly state the required definitions below,

along with some intuition on why these are relevant for general channels with feedback.

Definition 3: (a) The *limsup in probability* of a sequence of random variables $\{A_n\}$, denoted $\overline{A}$, is defined as the infimum of all real numbers α such that $\lim_{n\to\infty} Pr[A_n > \alpha] = 0$.

(b) The *liminf in probability* of a sequence of random variables $\{A_n\}$, denoted $\underline{A}$, is defined as the supremum of all real numbers β such that $\lim_{n\to\infty} Pr[A_n < \beta] = 0$.

For any sequence $\{P_{X^N,Y^N}\}_{N=1}^{\infty}$ of joint distributions on the input and output (with P_{X^N,Y^N} as in (2)), define $\forall (x^N, y^N) \in \mathcal{X}^N \times \mathcal{Y}^N$:

$$\vec{i}(x^N; y^N) \triangleq \frac{1}{N} \log \frac{P_{X^N,Y^N}(x^N, y^N)}{\vec{P}^k_{X^N|Y^N}(x^N|y^N) P_{Y^N}(y^N)} \tag{4}$$

$$= \frac{1}{N} \log \frac{\vec{P}^{ch}_{Y^N|X^N}(y^N|x^N)}{P_{Y^N}(y^N)},$$

$$\underline{I}(X \to Y) \triangleq \liminf_{in\ prob} \vec{i}(X^N; Y^N) \tag{5}$$

where $\vec{P}^k_{X^N|Y^N}, \vec{P}^{ch}_{Y^N|X^N}$ are defined by (3), and P_{Y^N} is the marginal from (2).

We may interpret the above definitions as follows. The directed information flowing from X^N to Y^N was defined by Massey [3] as

$$I(X^N \to Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}) \tag{6}$$

who also showed that it upper bounds the feedback capacity of the channel from X to Y for a given valid joint distribution on (X^N, Y^N). It can be verified that the expected value of the *directed information density* $\vec{i}(X^N; Y^N)$ in (4) is exactly $\frac{1}{N} I(X^N \to Y^N)$. The intuition is that for a general joint process characterized by $\mathbf{P_{XY}}$, the asymptotic behavior of $\vec{i}(x^N; y^N)$ determines the maximum rate that can be transmitted over the feedback channel. For an arbitrary joint distribution $\mathbf{P_{XY}}$, the directed information density $\vec{i}(X^N; Y^N)$ may not converge as $N \to \infty$. In general, it is a random quantity asymptotically bounded between its $\liminf_{in prob}$ and the $\limsup_{in\ prob}$. Since the feedback capacity represents the maximum rate of transmission we can guarantee, it is characterized by the $\liminf_{in\ prob}$ of the directed information density, given by (5). In the case where the joint process $\mathbf{P_{XY}}$ is stationary and ergodic, one can show that the quantity $\vec{i}(X^N; Y^N)$ converges to its expected value as $N \to \infty$. In other words, $\frac{\vec{i}(X^N;Y^N)}{\frac{1}{N} I(X^N \to Y^N)} \to 1$ as $N \to \infty$ and Massey's upper bound can be achieved.

The feedback capacity theorem of [4], [5] is given without a cost-function, but it can be easily extended to include a cost constraint as follows.

Fact 1: [4] For an arbitrary channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$, the capacity with k-delay feedback at cost P is

$$C^k_{fb}(P) = \sup_{\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}: \rho(P^k_{\mathbf{X}|\mathbf{Y}}) \le P} \underline{I}(X \to Y), \tag{7}$$

where

$$\rho(\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}) \triangleq \limsup_{in\ prob} c_n(X^n, Y^n)$$

$$= \inf \left\{ h : \lim_{n\to\infty} P_{X^n Y^n}\big((x^n, y^n) : c_n(x^n, y^n) > h\big) = 0 \right\}.$$

A. Computing the Capacity-Cost Function

The above formula for the capacity-cost function involves optimizing the function

$$\underline{I}(X \to Y) \triangleq \liminf_{in\ prob} \frac{1}{N} \log \frac{\vec{P}^{ch}_{Y^N|X^N}}{P_{Y^N}}$$

over an infinite-dimensional space of input distributions $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$. Since computing the optimization directly is difficult, we can pose the following question: given a channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ and an input distribution $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$, for what sequence of cost measures does $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$ achieve the supremum in the capacity formula ?

Our first result is a sufficient condition for a specified joint process to achieve the optimum in the feedback capacity formula. The joint process is assumed to satisfy two conditions. The first is that it is *directed information stable* in the sense defined below. Information stability of random processes is discussed in detail in [19]. In [5], the concept is extended to directed information stability.

Definition 4: A joint process $\mathbf{P_{XY}}$ is directed information stable if

$$\lim_{N\to\infty} P\left(\left| \frac{\vec{i}(X^N \to Y^N)}{I(X^N \to Y^N)} - 1 \right| > \epsilon \right) = 0 \quad \forall \epsilon > 0,$$

where $\vec{i}(X^N \to Y^N)$ and $I(X^N \to Y^N)$ are defined in (4) and (6), respectively.

Recall that $\overline{I}$ and $\underline{I}$ denote the $\limsup_{in\ prob}$ and $\liminf_{in\ prob}$ of $\vec{i}()$, respectively. If a joint process $\mathbf{P_{XY}}$ is directed information stable, it can be shown (cf. [5]) that

$$\begin{aligned} \underline{I}(\hat{X} \to Y) &= \liminf_{N\to\infty} \frac{1}{N} I(X^N \to Y^N) \\ &\le \limsup_{N\to\infty} \frac{1}{N} I(X^N \to Y^N) = \bar{I}(X \to Y). \end{aligned} \tag{8}$$

Jointly stationary and ergodic processes are examples of processes that are directed information stable which further satisfy (8) with equality.

Theorem 1: For a channel $\mathbf{P}^{ch}_{\mathbf{Y}|\mathbf{X}}$ with $k-$delay feedback, let $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$ be an input distribution such that the joint process $\mathbf{P_{XY}} = \{P_{X^n,Y^n}\}_{n=1}^{\infty}$ (given by (2)) is directed information stable and further, equality holds in (8) . Then the input distribution $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$ achieves the k-delay feedback capacity of the channel at cost level P if for all sufficiently large n, the cost measure satisfies

$$c_n(x^n, y^n) = \lambda \cdot \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)} + d_0, \tag{9}$$

where $\vec{P}^{ch}_{Y^n|X^n}$ is defined in (3), λ is any positive number, d_0 is an arbitrary constant and $P = \limsup_{n\to\infty} E[c_n(X^n, Y^n)]$.

The proof of the theorem is given in Appendix A. Though the theorem requires the chosen input distribution to be such that it makes the induced joint distribution $\mathbf{P_{XY}}$ information stable, we emphasize that for the cost function given by the theorem, the optimality of this distribution is among all valid input distributions for the given channel, not just among the ones that make the joint distribution information stable.

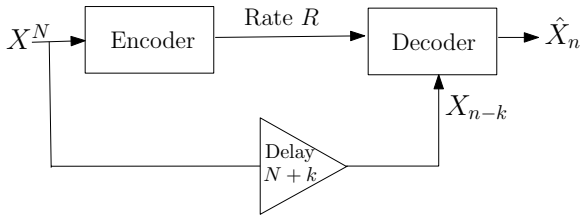

Fig. 2. Source coding with delay-k feed-forward.

III. Source Coding with Delayed Feed-Forward

Consider a general discrete source X with alphabet $\mathcal{X}$, distribution $\mathbf{P_X} = \{P_{X^N}\}_{N=1}^{\infty}$ and reconstruction alphabet $\widehat{\mathcal{X}}$ [1]. The feed-forward model with delay k is shown in Figure 2. The reconstruction at the decoder is sequential: for $n > k$ the decoder has access to source samples X^{n-k} to produce reconstruction symbol $\hat{X}_n$ (for $n \leq k$, $\hat{X}_n$ is produced using the index alone).

Definition 5: An $(N, 2^{NR})$ source code with k-delay feed-forward with block length N, and rate R consists of an encoder mapping e and a sequence of decoder mappings $g_n, n = 1, \ldots, N$, where

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

There is a distortion measure $d_N : \mathcal{X}^N \times \widehat{\mathcal{X}}^N \to \mathbb{R}^+$ on pairs of sequences of length N. We assume that $d_N(x^N, \hat{x}^N)$ is normalized with respect to N and is uniformly bounded in N.

Definition 6: R is an ϵ-achievable rate at distortion D with k-delay feed-forward if for all sufficiently large N, there exists an $(N, 2^{NR})$ source code 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 distortion D with k-delay feed-forward if it is ϵ-achievable for every $\epsilon > 0$.

The rate-distortion function, $R_{ff}^k(D)$, is the infimum of all achievable rates at distortion D with k-delay feed-forward. This was characterized for general sources and distortion measures in [13]. With feed-forward, the information at the decoder changes with time. Hence one needs to consider general joint processes $(\mathbf{X}, \hat{\mathbf{X}})$, and information spectrum methods are needed to obtain a tight rate-distortion theorem. For any sequence of joint distributions $\{P_{X^N,\hat{X}^N}\}_{N=1}^{\infty}$, define $\forall (x^N, \hat{x}^N) \in \mathcal{X}^N \times \widehat{\mathcal{X}}^N$:

$$\vec{i}_k(\hat{x}^N; x^N) \triangleq$$
$$\frac{1}{N} \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N) \prod_{n=1}^{N} P_{\hat{X}_n|\hat{X}^{n-1},X^{n-k}}(\hat{x}_n|\hat{x}^{n-1}, x^{n-k})}, \tag{10}$$

$$\overline{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \vec{i}_k(\hat{X}^N; X^N). \tag{11}$$

In (10), $P_{\hat{X}_n|\hat{X}^{n-1},X^{n-k}} = P_{\hat{X}_n|\hat{X}^{n-1}}$ for $n < k$. It is interesting to compare the k-directed information density in (10) with the directed information density defined in (4). In channel coding, recall that the feedback delay k restricts the input distribution to be of the form in (1) and hence the joint distribution is restricted to be as in (2). As a consequence of this, the directed information density in (4) has the same formula for every k. It is important to note that though the formula in (4) does not change with feedback delay k, the directed information density *does* depend on k since P_{Y^N} is derived from the joint distribution, which varies with k. In source coding, there is no restriction on the joint distributions $\{P_{X^N,\hat{X}^N}\}$ that can be chosen regardless of the feed-forward delay k. Hence (10) represents a different formula for each k, which we call the k-directed information density. The asymptotic behavior of $\vec{i}_k(\hat{x}^N; x^N)$ determines the minimum achievable rate with feed-forward delay k.

Fact 2 ([13]): For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with k-delay feed-forward is given by

$$R_{ff}^k(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D} \overline{I}_k(\hat{X} \to X), \tag{12}$$

where $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ and

$$\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{in\ prob} d_n(x^n, \hat{x}^n)$$
$$= \inf\{h : \lim_{n\to\infty} P_{X^n,\hat{X}^n}((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h) = 0\} \tag{13}$$

A. Computing the feed-forward rate-distortion function

Since the optimization in Theorem 2 is intractable in general, we can pose the following question: given a source distribution $\mathbf{P_X}$ and a conditional distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$, for what sequence of distortion measures does $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ achieve the infimum in the rate-distortion formula ?

The following theorem gives a sufficient condition for a given joint process to achieve the optimum in the rate-distortion formula, provided it satisfies two conditions. We first need a generalization of Definition 4 to accomodate for the fact that in source coding with feed-forward, the objective function in Theorem 2 changes with k.

Definition 7: A joint process $\mathbf{P}_{\mathbf{X}\hat{\mathbf{X}}}$ is k-directed information stable if

$$\lim_{N\to\infty} P\left(\left|\frac{\vec{i}_k(\hat{X}^N \to X^N)}{E[\vec{i}_k(\hat{X}^N \to X^N)]} - 1\right| > \epsilon\right) = 0 \quad \forall \epsilon > 0.$$

If a joint process $\mathbf{P}_{\hat{\mathbf{X}}\mathbf{X}}$ is k-directed information stable, it can be shown (cf. [5]) that

$$\underline{I}_k(\hat{X} \to X) = \liminf_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N)$$
$$\leq \limsup_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N) = \bar{I}_k(\hat{X} \to X). \tag{14}$$

Theorem 2: Let X be a source characterized by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$ and feed-forward delay k. Let $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be a conditional distribution such that the joint process is k-directed information stable and equality holds in

[1] In Section II, X denoted the channel input. With some abuse of notation, in this section we shall use X to represent the source, and $\hat{X}$ the reconstruction.

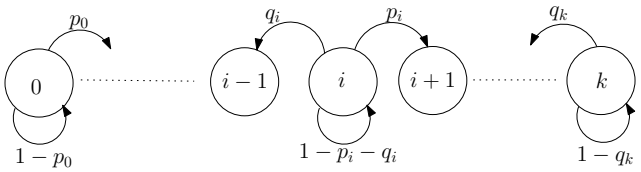

Fig. 3. Markov chain representing the source.

(14). Then $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ achieves the rate-distortion function with k-delay feed-forward at distortion level D if for all sufficiently large n, the distortion measure satisfies

$$d_n(x^n,\hat{x}^n) = -c\cdot\frac{1}{n}\log\frac{P_{X^n,\hat{X}^n}(x^n,\hat{x}^n)}{\vec{P}^k_{\hat{X}^n|X^n}(\hat{x}^n|x^n)} + d_0(x^n), \quad (15)$$

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, $d_0(.)$ is an arbitrary function, and $D = \limsup_{n\to\infty} Ed_n(X^n,\hat{X}^n)$.

The proof is along the lines as that of Theorem 1, with a few differences. It is omitted due to space constraints and can be found in [20].

IV. Examples

In this section, we present two examples where Theorems 1 and 2 are used to compute the capacity-cost function and the feed-forward rate-distortion function, respectively.

A. Source Coding Example

Consider a source $\mathbf{X} = \{X_n\}$, where X_n evolves according to the Markov chain shown in Figure 3. The source can take values in the set $\{0,1,\ldots,k\}$. If $X_n = i$ (state i), X_{n+1} is equal to $i+1$ with probability p_i, or $i-1$ with probability q_i, or i with probability $1-p_i-q_i$. If X_n denotes the price of a stock at the end of day n, this is a reasonable model for how the value of the stock varies over time. Suppose we are interested in predicting drops in the stock price over a period of N days. Our reconstruction $\hat{X}_n$ is binary: we predict $\hat{X}_n = 1$ if we expect the price to drop from day $n-1$ to n, otherwise $\hat{X}_n = 0$. The distortion is modeled using a Hamming criterion as follows.

$$d_n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(\hat{x}_i,x_{i-1},x_i), \quad (16)$$

where $d(.,.,.)$ is the *per-letter* distortion specified as follows. $d(\hat{x}_i,x_{i-1},x_i) = 1$ if either we fail to predict a drop: $\hat{x}_i = 0$ and $x_i < x_{i-1}$, or we falsely predict a drop: $\hat{x}_i = 1$ and $x_i \geq x_{i-1}$. In all other cases, $d(\hat{x}_i,x_{i-1},x_i) = 0$.

Now assume there is an insider who has a priori information about the behavior of the stock over the N days. If he is willing to share this information, what is the minimum information we need from him in order to predict drops with distortion D? Note that before making the decision $\hat{X}_n$, we know the values of the stock on all previous days, i.e., X^{n-1}. Thus feed-forward is built into this problem, and the minimum rate of information (in bits/sample) we need to predict drops in value with distortion D is $R^1_{ff}(D)$.

Proposition 1: For the problem described above,

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

where $h(\cdot)$ 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: The source is characterized by the Markov chain transition probabilities $P_{X_i|X_{i-1}}$, $\forall i$. We shall show that the optimal rate-distortion function is achieved by a joint distribution of the form $P_{X^n,\hat{X}^n} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i|X_{i-1}}$ $\forall n$. For this to hold, the structure of the cost function from Theorem 2 is

$$d_n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n}\Big(-c\log_2 P(x_i|\hat{x}_i,x_{i-1}) + d_0(x_{i-1},x_i)\Big). \quad (17)$$

Equivalently, using (16), the above condition holds when

$$d(\hat{x}_i,x_{i-1},x_i) = -c\log_2 P(x_i|\hat{x}_i,x_{i-1}) + d_0(x_{i-1},x_i). \quad (18)$$

Guided by the structure of the distortion function $d(\cdot)$, we choose 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 $P(\hat{X}_i = 0|x_{i-1} = 0, x_i = 0) = 1$, which using Bayes' rule yields $P(X_i = 0|x_{i-1} = 0,\hat{x}_i = 0) = 1-p$. The event $(X_{i-1} = 0, \hat{X}_i = 1)$ has zero probability. Thus we obtain the first two columns of Table I. When $(X_{i-1} = j, \hat{X}_i = 0)$, $1 \leq j \leq k$, 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 I.

Substituting the values from Table I in the relation $P(x_2|x_1,\hat{x}_2) = \frac{P(x_2|x_1)P(\hat{x}_2|x_2,x_1)}{\sum_{x_2}P(x_2|x_1)P(\hat{x}_2|x_2,x_1)}$, we obtain the conditional distribution $P(\hat{X}_2|x_1,x_2)$ shown in Table II. To show that the conditional distribution in Table II is optimal, we need to check that the resulting joint distribution can be made to satisfy (18). This can be done by using the values from Table I in (18) to determine c and $d_0(.,.)$.

Since the process $\{\mathbf{X},\hat{\mathbf{X}}\}$ is jointly stationary and ergodic, when $(x^n,\hat{x}^n) \sim P_{X^n,\hat{X}^n}$, the distortion $d_n(x^n,\hat{x}^n) \to E[d(\hat{x}_2,x_1,x_1)]$ as $n\to\infty$ $w.p.1$. Hence the distortion constraint is equivalent to $E[d(\hat{x}_2,x_1,x_2)] \leq D$. Using Table II and the stationary distribution of the source $[\pi_0,\ldots,\pi_k]$, the expected distortion can be calculated as

$$\begin{aligned} E[d(\hat{x}_2,x_1,x_2)] &= \sum_{x_1,x_2,\hat{x}2} P(x_1,x_2)P(\hat{x}_2|x_1,x_2)\cdot d(\hat{x}_2,x_1,x_2) \\ &= (1-\pi_0)\epsilon \leq D. \end{aligned} \quad (19)$$

We can now calculate the rate distortion function as

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

	$(x_{i-1},\hat{x}_i)$							
	00	01	$\cdots$	$j0$	$j1$	$\cdots$	$k0$	$k1$
$x_i=0$	$1-p$	–	$\cdots$	–	–	–	–	–
$x_i=1$	p	–	$\cdots$	–	–	–	–	–
$x_i=j-1$	–	–	–	ϵ	$1-\epsilon$	–	–	–
$x_i=j$	–	–	–	$\frac{(1-\epsilon)(1-p_j-q_j)}{1-q_j}$	$\frac{\epsilon(1-p_j-q_j)}{1-q_j}$	–	–	–
$x_i=j+1$	–	–	–	$\frac{(1-\epsilon)p_j}{1-q_j}$	$\frac{\epsilon p_j}{1-q_j}$	–	–	–
$x_i=k-1$	–	–	$\cdots$	–	–	–	ϵ	$1-\epsilon$
$x_i=k$	–	–	$\cdots$	–	–	–	$1-\epsilon$	ϵ

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

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

$$\begin{aligned} R_{ff}(D) &= \lim_{N\to\infty}\frac{1}{N}\sum_{x^N,\hat{x}^N} P(x^N,\hat{x}^N)\log_2\prod_{n=1}^{N}\frac{P(x_n|\hat{x}^n,x^{n-1})}{P(x_n|x^{n-1})} \\ &= \sum_{x_1,x_2,\hat{x}_2} P(x_1,x_2,\hat{x}_2)\log_2\frac{P(x_2|x_1,\hat{x}_2)}{P(x_2|x_1)} \\ &= H(X_2|X_1) - H(X_2|\hat{X}_2,X_1) \end{aligned} \tag{20}$$

to obtain the expression in Proposition 1. ∎

In [10], the distortion-rate function for a symmetric binary Markov source with feed-forward and a stationary Gaussian source with feed-forward were evaluated using a result that applies to sources that can be described by an i.i.d innovations process. We note that it is also possible to compute these using Theorem 2. We also remark that the example presented above cannot be computed using the result in [10] since the innovations process of the Markov chain is not i.i.d.

B. Channel Coding Example

Consider a binary Markov channel with feedback delay 1 defined as follows for all time instants i: $P^{ch}(Y_i|X^i,Y^{i-1}) = P^{ch}(Y_i|X_i,Y_{i-1})$ with $X_i, Y_i \in \{0,1\}$

$$\begin{aligned} P^{ch}(Y_i|X_i,Y_{i-1}=1) &= \delta_{(Y_i=X_i)}, \\ P^{ch}(Y_i|X_i,Y_{i-1}=0) &= 0.5. \end{aligned} \tag{21}$$

In other words, if the channel output at time $i-1$ is 1, we have a noiseless channel at time i. If the channel output at time $i-1$ is 0, at time i we have a binary symmetric channel with crossover probability 0.5. Suppose we wish to impose a cost function defined as follows.

1) At time i, if the channel is in the 'good' state ($Y_{i-1}=1$),

$$c(X_i,Y_{i-1}=1) = \begin{cases} \alpha_0 & \text{if } X_i=0 \\ \alpha_1 & \text{if } X_i=1 \end{cases}, \alpha_0,\alpha_1 \in \mathbb{R}. \tag{22}$$

2) If the channel is in the bad state, we impose a constant cost: $c(X_i,Y_{i-1}=0)=\alpha, \quad X_i\in\{0,1\}$.

The cost for n uses of the channel is $c_n(x^n,y^n) = \frac{1}{n}\sum_{i=1}^{n} c(x_i,y_{i-1})$. We ask: given a cost constraint P, does a stationary, first-order Markov input distribution of the form $\{P_{X_i|Y_{i-1}}\}_{i=1}^{\infty}$ achieve the feedback capacity $C_{fb}^1(P)$? Let the input distribution be specified as

$$P(X_i=0|Y_{i-1}=0)=q_0,\ P(X_i=0|Y_{i-1}=1)=q_1\ \forall i, \tag{23}$$

where q_0, q_1 have to be determined. The joint distribution is then

$$\begin{aligned} P(X^n,Y^n) &= \prod_{i=1}^{n} P(X_i,Y_i|Y_{i-1}) \\ &= \prod_{i=1}^{n} P(X_i|Y_{i-1})\cdot P^{ch}(Y_i|X_i,Y_{i-1}), \end{aligned} \tag{24}$$

with $P(X_i|Y_{i-1})$ and $P^{ch}(Y_i|X_i,Y_{i-1})$ given by and (23) and (21), respectively. Using Theorem 1 it follows that this joint distribution is optimal if the cost function satisfies

$$c_n(x^n,y^n) = \lambda\cdot\frac{1}{n}\sum_{i=1}^{n}\log\frac{P^{ch}_{Y_i|X_i,Y_{i-1}}(y_i|x_i,y_{i-1})}{P_{Y_i|Y_{i-1}}(y_i|y_{i-1})} + d_0. \tag{25}$$

Substituting all the possible values for (Y_{i-1},X_i,Y_i) in (25), we see that the conditions are:

$$d_0=\alpha, \quad \lambda\log_2\frac{1}{q_1}+d_0=\alpha_0, \quad \lambda\log_2\frac{1}{1-q_1}+d_0=\alpha_1 \tag{26}$$

Further, the parameter q_1 has to be chosen to satisfy the cost constraint. Since the joint process $\{P_{X^n,Y^n}\}_{n=1}^{\infty}$ (with P_{X^n,Y^n} as in (24)) is jointly stationary and ergodic, the cost constraint reduces to $E[c(X_i,Y_{i-1})] \le P$. In terms of our joint distribution, this can be written as

$$\begin{aligned} &\sum_{x_i} P(Y_{i-1}=0)P(x_i|Y_{i-1}=0)c(x_i,0) \\ &\qquad + P(Y_{i-1}=1)P(x_i|Y_{i-1}=1)c(x_i,1) \\ &\overset{(a)}{=} P(Y_{i-1}=0)\alpha + P(Y_{i-1}=1)[q_1\alpha_0+(1-q_1)\alpha_1] = P, \end{aligned} \tag{27}$$

where we have used (21) and (23) and the cost function to obtain (a). $[P(Y_{i-1}=0),\ P(Y_{i-1}=1)]$ is just the stationary distribution of the Markov chain $\{Y_i\}$, whose transition probabilities are given by $[P(Y_i=0|Y_{i-1}=0)=0.5\ P(Y_i=0|Y_{i-1}=1)=q_1]$. This can be computed to be $[P(Y_{i-1}=0)=\frac{2q_1}{1+2q_1},\ P(Y_{i-1}=1)=\frac{1}{1+2q_1}]$. Using this in (27) and rearranging terms, we obtain

$$q_1\alpha_0+(1-q_1)\alpha_1+2q_1\alpha=P(1+2q_1). \tag{28}$$

If we fix the cost parameters α_0,α_1,α and the cost constraint P, there are four conditions (given by (26) and (28)) to be satisfied. Since there are three variables (λ,d_0,q_1), one can assert that a first-order Markov input distribution of the form of (23) achieves the optimum for this problem only when the system of four equations in three unknowns ((26) and (28)) has a solution. Of course, if we specify only three among the four parameters $\alpha_0,\alpha_1,\alpha,P$, a solution always exists and the fourth parameter gets automatically determined. When there exist λ,d_0,q_1 such that (26) and (28) are satisfied, the feedback capacity-cost function can be evaluated as

$$\begin{aligned}
C_{fb}^1(P)&=\underline{I}(X\to Y)=\lim_{N\to\infty}\frac{1}{N}I(X^N\to Y^N)\\
&=\lim_{N\to\infty}\frac{1}{N}\sum_{x^N,y^N}P(x^N,y^N)\log_2\frac{\prod_{n=1}^N P(y_n|x^n,y^{n-1})}{\prod_{n=1}^N P(y_n|y^{n-1})}\\
&=\sum_{y_1,x_2,y_2}P(y_1,x_2,y_2)\log_2\frac{P(y_2|x_2,y_1)}{P(y_2|y_1)}\\
&=\frac{1}{1+2q_1}\left[q_1\log_2\frac{1}{q_1}+(1-q_1)\log_2\frac{1}{1-q_1}\right]+\frac{2q_1}{1+2q_1}\cdot 0\\
&=\frac{1}{1+2q_1}h(q_1)
\end{aligned}$$

where $h()$ is the binary entropy function.

Example: If we set $\alpha=0$, $\alpha_1=2\alpha_0$ for any $\alpha_0>0$ - when the channel is 'bad', the associated cost is 0; when the channel is 'good', there is a positive cost and the cost is higher to keep the channel in the good state. For these parameters, we obtain from (26) $d_0=0$ and $q_1=0.6180$. If we set the cost constraint P to satisfy (28), we obtain $P=1.382\alpha_0$ and from (IV-B), $C_{fb}(1.382\alpha_0)=0.4291$ bits/channel use.

V. Conclusion

The problem of computing the feedback capacity-cost function of channels and the feed-forward rate-distortion function of sources was studied. Due to the memory and dynamics inherent in these problems, computing these functions involves optimizing a multi-letter expression over all valid processes (not necessarily stationary or ergodic). Since it is infeasible to compute the optimizations directly, a structural approach to the problem can be useful. We derived sufficient conditions on the structure of the distortion (cost) function in order for a chosen joint distribution to achieve the optimal rate-distortion (capacity-cost) function. While the structural results do not yield the feedback capacity for every channel and cost function, the examples show that one can compute performance limits for interesting problems which may otherwise be intractable. Whether the structural conditions on the distortion/cost function are also necessary for a joint distribution to achieve the optimum is an open question.

Appendix A
Proof of Theorem 1

Let $\mathbf{P}'^k_{\mathbf{X}|\mathbf{Y}}$ denote any other input distribution that achieves lower cost than $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$ over the channel. The symbol $'$ will denote that the joint distribution $P'_{XY}\triangleq\vec{P}'^k_{X|Y}\cdot\vec{P}^{ch}_{Y|X}$ is being used. We will also use the notation $P_{XY}\triangleq\vec{P}^k_{X|Y}\cdot\vec{P}^{ch}_{Y|X}$. We note that while the joint distribution P_{XY} is information stable (by the assumptions of the theorem), P'_{XY} need not be since $\mathbf{P}'^k_{\mathbf{X}|\mathbf{Y}}$ is an arbitrary input distribution.

We have

$$\limsup_{in\ prob\ P'_{XY}} c_n(X^n,Y^n)<\limsup_{in\ prob\ P_{XY}} c_n(X^n,Y^n). \tag{29}$$

We will show that if (29) is satisfied, then

$$\underline{I}_{P_{XY}}(X\to Y)>\underline{I}_{P'_{XY}}(X\to Y) \tag{30}$$

under the conditions of the theorem, thus proving the optimality of $\mathbf{P}^k_{\mathbf{X}|\mathbf{Y}}$.

Step 1: We will first show that

$$\begin{aligned}
\underline{I}_{P'_{XY}}(X\to Y)&\triangleq\liminf_{in\ prob\ P'_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P'_{Y^n}}\\
&\le\liminf_{in\ prob\ P'_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}.
\end{aligned} \tag{31}$$

Due to the inequality [18]

$$\liminf_{in\ prob}a_n-\liminf_{in\ prob}b_n\ge\liminf_{in\ prob}(a_n-b_n),$$

to prove (31), it is enough to show that $\liminf_{in\ prob\ P'_{XY}}\frac{1}{n}\log\frac{P'_{Y^n}}{P_{Y^n}}\ge 0$, which is true from Theorem $8(a)$ in [21].

Step 2: Here we use Step 1 to prove (30). We have

$$\begin{aligned}
&\underline{I}_{P_{XY}}(X\to Y)-\underline{I}_{P'_{XY}}(X\to Y)\\
&\overset{(a)}{\ge}\liminf_{in\ prob\ P_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}-\liminf_{in\ prob\ P'_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}\\
&\overset{(b)}{>}\liminf_{in\ prob\ P_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}-\liminf_{in\ prob\ P'_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}\\
&\quad+\limsup_{in\ prob\ P'_{XY}}\ [\beta\cdot c_n(X^n,Y^n)+b_0]\\
&\quad-\limsup_{in\ prob\ P_{XY}}\ [\beta\cdot c_n(X^n,Y^n)+b_0]
\end{aligned} \tag{32}$$

where $\beta>0$ and (a) follows from (31) in Step 1, (b) is from (29). Now set

$$\beta c_n(X^n,Y^n)+b_0=\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)}. \tag{33}$$

Since $\mathbf{P}_{\mathbf{XY}}$ is directed information stable, we have from (8)

$$\begin{aligned}
\underline{I}_{P_{XY}}(X\to Y)&\triangleq\liminf_{in\ prob\ P_{XY}}\frac{1}{n}\log\frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}\\
&=\liminf_{N\to\infty}I_{P_{XY}}(X^N\to Y^N).
\end{aligned} \tag{34}$$

Further, as a consequence of equality in (8),

$$\begin{aligned}
\bar{I}_{P_{XY}}(X\to Y)&=\limsup_{N\to\infty}I_{P_{XY}}(X^N\to Y^N)\\
&=\liminf_{N\to\infty}I_{P_{XY}}(X^N\to Y^N)=\underline{I}_{P_{XY}}(X\to Y)
\end{aligned} \tag{35}$$

and so

$$\limsup_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} = \liminf_{in\ prob\, P_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}}$$

Hence (32) becomes

$$\underline{I}_{P_{XY}}(X \to Y) - \underline{I}_{P'_{XY}}(X \to Y)$$
$$> \limsup_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} - \liminf_{in\ prob\, P'_{XY}} \frac{1}{n} \log \frac{\vec{P}^{ch}_{Y^n|X^n}}{P_{Y^n}} \geq 0 \tag{36}$$

because the $\limsup_{in\ prob}$ is always greater than or equal to the $\liminf_{in\ prob}$. Rearranging (33), we get the result.

ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and the anonymous reviewers for their comments and suggestions, which led to a much improved manuscript.

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] H. Marko, "The bidirectional communication theory—a generalization of information theory," *IEEE Trans. Commun.*, vol. COM-21, pp. 1345-1351, Dec. 1973.

[3] J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Symp. Inf. Theory Appl. (ISITA-90)*, pp. 303-305, 1990.

[4] S. Tatikonda, "Control under communications constraints," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, Sep. 2000.

[5] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, pp. 323-349, Jan. 2009.

[6] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, pp. 4-20, Jan. 2003.

[7] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. IT-35, pp. 37-43, Jan. 1989.

[8] Y. H. Kim, "The feedback capacity of the first-order moving average Gaussian channel," *IEEE Trans. Inf. Theory*, vol. IT-52, pp. 3063-3079, July 2006.

[9] Y. H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. IT-25, pp. 1488-1499, Apr. 2008.

[10] 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, Dec. 2003.

[11] S. S. Pradhan, "On the role of feedforward in Gaussian sources: point-to-point source coding and multiple description source coding," *IEEE Trans. Inf. Theory*, vol. 53, pp. 331-349, Jan. 2007.

[12] E. Martinian and G. W. Wornell, "Source coding with fixed lag side inf," in *Proc. 42nd Annual Allerton Conf.*, Monticello, IL, 2004.

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

[14] R. E. Blahut, "Computation of channel capacity and rate-distortion function," *IEEE Trans. Inf. Theory*, vol. 18, pp. 460-473, July 1972.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[16] 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.

[17] 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.

[18] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, 2002.

[19] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964 (translated by A. Feinstein).

[20] R. Venkataramanan, "Information-theoretic results on communication problems with feed-forward and feedback," Ph.D. thesis, University of Michigan, 2008.

[21] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, pp. 1147-1157, July 1994.