Directed Information for Communication Problems with Common Side Information and Delayed Feedback/Feedforward

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

Abstract

In this work, we consider the problems of channel coding with feedback and source coding with feedforward. We give a new interpretation of directed information [1] which helps us understand why it arises in the capacity and rate-distortion function of these two problems. Comparing the two problems studied, we find that in channel coding feedback boosts capacity due to a larger constraint set of optimization of the same objective function. On the other hand, feedforward reduces the rate-distortion function due to a smaller objective function optimized over an unchanged constraint set. Using our interpretation of directed information, we find the capacity of channels with arbitrarily delayed feedback and causal/non-causal state information. We also solve the corresponding source coding problem.

Keywords

Channels with Feedback, Feedforward, Side Information, Directed Information, Delay

1 Introduction

It is a well-known result in information theory that feedback does not increase the capacity of a discrete memoryless channel [2]. 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 [1, 3, 4]. In these works, the feedback considered was available at the encoder with delay 1. There is a source coding counterpart to channel coding with feedback, viz., source coding with feedforward. The problem of source coding with feedforward was considered in [5, 6, 7, 8]. In [5], a characterization of attainable performance was provided for sources that can be represented auto-regressively with an innovation process satisfying the Shannon lower bound (SLB) with equality. The optimal rate-distortion function for source coding with feedforward for arbitrary sources and distortion measures was characterized using directed information in [7].

In this work, we consider generalized versions of the communication problems described above. In the first part of the paper, we consider channels with arbitrarily delayed feedback and side information (available either causally or non-causally at both the transmitter and receiver). In the second part, we deal with source coding with delayed feedforward and side information. We give a new, intuitive interpretation of directed information. This interpretation is crucial for obtaining performance limits for problems with delayed feedback/feedforward and side information.

*This work was supported by NSF Grant CCF-0329715 and Grant (CAREER) CCF-0448115.

2 Notation and Preliminaries

We first lay down the notation used in the rest of this paper. The symbols X_n and Y_n are used to denote the channel input and output, respectively, at time n. The notation X^n denotes the sequence of random variables $(X_1, X_2 \ldots, X_n)$. A channel is defined as a sequence of probability distributions:

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

We denote the input distribution for a channel without feedback as

$$P_{\mathbf{X}} = \{P_{X_n|X^{n-1}}\}_{n=1}^{\infty}. \tag{2}$$

For a channel with feedback, the input distribution can depend on the fed-back output symbols. We will be dealing with arbitrary delays in feedback. If the feedback has delay k, the channel input at time n can depend on the first $(n-k)$ channel outputs. In this case, the distribution is denoted by

$$\vec{P}_{\mathbf{X}|\mathbf{Y^k}} = \{P_{X_n|X^{n-1},Y^{n-k}}\}_{n=1}^{\infty}. \tag{3}$$

Note that for a channel with feedback delay k, (1) and (3) completely define the joint distribution of the system as follows.

$$P_{\mathbf{XY}} = P^{ch}_{\mathbf{Y}|\mathbf{X}} \cdot \vec{P}_{\mathbf{X}|\mathbf{Y^k}}$$

Throughout this paper, we will assume that the joint process $\{X^n, Y^n\}_{n=1}^{\infty}$ is stationary and ergodic. It is also assumed that the feedback is noiseless.

We will also use the finite dimensional probability distributions corresponding to the random processes in (1) and (3) in order to express directed information and other related quantities. In particular, we will need the following definitions of directed probability distributions [4].

Definition 1.

$$\vec{P}_{X^N|Y^N} \triangleq \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1}},$$

$$\vec{P}^{k}_{X^N|Y^N} \triangleq \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k}},$$

$$\vec{P}^{ch}_{Y^N|X^N} \triangleq \prod_{n=1}^{N} P_{Y_n|X^n,Y^{n-1}}.$$

In the above definition, we note that for $n < k$, $P_{X_n|X^{n-1},Y^{n-k}}$ is taken to be equal to $P_{X_n|X^{n-1}}$. We also observe that, using Bayes' rule, the joint probability distribution of the system at time N can always be written as

$$P_{X^N,Y^N} = \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1}} \cdot P_{Y_n|X^n,Y^{n-1}} = \vec{P}_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}. \tag{4}$$

3 Channel Coding- Feedback with delay

In this section, we consider a channel with feedback with arbitrary delay k, as shown in Figure 1. We want to characterize the capacity of this channel. Toward this end, we give a new interpretation for directed information, a quantity first defined by Massey [1]. We then use this interpretation to deduce the capacity of a channel with arbitrarily delayed feedback. It should

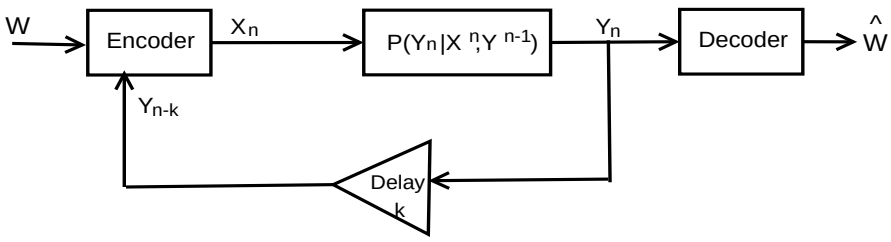

Figure 1: Channel with k-delayed feedback

be mentioned here that finite-state machine channels with arbitrarily delayed feedback were considered in [9].

A technical remark is in order before we proceed. Throughout this work, we will use the word 'capacity' to denote the maximum achievable rate assuming that the joint process $\{X^n, Y^n\}_{n=1}^{\infty}$ is stationary and ergodic. This assumption is made only to keep the expressions intuitive and to give insight into the feedback problem. The stationary and ergodic assumption enables us to use a directed version [10] of the Asymptotic Equipartition Property to give simple proofs of the capacity results. We can also rigorously prove the results for the general non-stationary, non-ergodic case by using information spectrum [11] versions of all the information quantities in the capacity expressions.

The notion of directed information is needed to characterize the feedback capacity. The directed information flowing from random sequence X^N to random sequence Y^N is defined as

$$I(X^N \to Y^N) \triangleq \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}). \tag{5}$$

Using the chain rule, this can be written as

$$I(X^N \to Y^N) = I(X^N; Y^N) - \sum_{n=2}^{N} I(Y^{n-1}; X_n | X^{n-1}) \tag{6}$$

We start with a channel without feedback. The no-feedback capacity is

$$C_{no-FB} = \sup_{P_{\mathbf{X}}} \lim_{N\to\infty} \frac{1}{N} I(X^N; Y^N). \tag{7}$$

When there is no feedback, the interpretation is that $I(X^N; Y^N)$ is the reduction in uncertainty of the input X^N when the decoder observes Y^N. Now consider the same channel with delay 1 feedback, i.e. $k = 1$ in Figure 1. It is known [4] that the capacity of this channel is

$$C_{FB}^{1} = \sup_{P_{\mathbf{X}|\mathbf{Y^1}}} \lim_{N\to\infty} \frac{1}{N} I(X^N \to Y^N). \tag{8}$$

This can be interpreted as follows. When there is feedback with delay 1, to generate the input X_n, the encoder knows all the past outputs Y^{n-1}. Hence the information $I(Y^{n-1}; X_i | X^{n-1})$ is already known at both encoder and decoder due to the feedback and is not 'actually transmitted'. In light of this interpretation, (6) says that when there is feedback, the mutual information $I(X^N; Y^N)$ is still the fundamental quantity that characterizes the capacity, but the information that is known at both ends due to the feedback $\left(\sum_{n=2}^{N} I(Y^n; X_n | X^{n-1})\right)$ should be subtracted out. This might seem surprising because it is obvious that the maximum achievable rate should

not decrease with feedback. Reassuringly, this is indeed true, since we have a larger set of distributions to optimize over when there is feedback. In other words, although the objective function with feedback is smaller $\big(I(X^N \to Y^N) \leq I(X^N;Y^N)\big)$, the constraint set of optimization is larger when feedback is present since the space of $\mathbf{P_X}$ is contained in the space of $\vec{P}_{\mathbf{X|Y}}$.

This interpretation of directed information lends itself to deducing the capacity when the feedback has arbitrary delay k (Figure 1). Here, the encoder knows the outputs Y^{n-k} to generate input X_n. Hence the information $I(Y^{n-k};X_n|X^{n-1})$ is already known at both encoder and decoder due to the feedback and is not 'actually transmitted'. One can guess that the capacity should be

$$C_{FB}^k = \sup_{\vec{P}_{\mathbf{X|Y^k}}} \lim_{N\to\infty} \frac{1}{N}\left[I(X^N;Y^N) - \sum_{n=k+1}^{N} I(Y^{n-k};X_n|X^{n-1})\right]. \tag{9}$$

Although the proof is omitted here, we can show that this is actually the capacity with k-delayed feedback. For brevity, we define

$$I_k(X^N \to Y^N) = I(X^N;Y^N) - \sum_{n=k+1}^{N} I(Y^{n-k};X_n|X^{n-1}). \tag{10}$$

In the next section, we will examine and compare in detail the three capacity expressions given by (8), (7) and (9). Before proceeding, we need to express the directed information and it's more general counterpart $I_k(X^N \to Y^N)$ using the directed quantities in Definition 1.

Proposition 1.

$$I_k(X^N \to Y^N) = \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{\vec{P}^k(x^N|y^N)\cdot P(y^N)}.$$

Proof.

$$\begin{aligned}
I_k(X^N \to Y^N) &= I(X^N;Y^N) - \sum_{n=k+1}^{N} I(Y^{n-k};X_n|X^{n-1}) \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)P(y^N)} - \sum_{n=k+1}^{N}\sum_{x^n,y^{n-k}} P(x^n,y^{n-k})\log\frac{P(x_n,y^{n-k}|x^{n-1})}{P(x_n|x^{n-1})P(y^{n-k}|x^{n-1})} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)\cdot P(y^N)} - \sum_{n=k+1}^{N}\sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x_n|x^{n-1},y^{n-k})}{P(x_n|x^{n-1})} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)\cdot P(y^N)} - \sum_{x^N,y^N} P(x^N,y^N)\log\prod_{n=k+1}^{N}\frac{P(x_n|x^{n-1},y^{n-k})}{P(x_n|x^{n-1})}.
\end{aligned} \tag{11}$$

Using Definition 1, we can write this as

$$\begin{aligned}
I_k(X^N \to Y^N) &= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{P(x^N)\cdot P(y^N)} - \sum_{x^N,y^N} P(x^N,y^N)\log\frac{\vec{P}^k(x^N|y^N)}{P(x^N)} \\
&= \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{\vec{P}^k(x^N|y^N)\cdot P(y^N)}.
\end{aligned} \tag{12}$$

□

In particular, we have

$$I(X^N \to Y^N) = \sum_{x^N,y^N} P(x^N,y^N)\log\frac{P(x^N,y^N)}{\vec{P}(x^N|y^N)\cdot P(y^N)}. \tag{13}$$

3.1 They are all the same - Directed Information!

The capacity expressions for a channel with no feedback, unit delay feedback and k-delayed feedback are given by (7),(8) and (9) respectively. We will now compare the three expressions and show that the objective function in all three cases can be written as directed information flowing from X^N to Y^N.

First consider the case of k-delayed feedback. Consider a time-line of how input symbols are produced at the encoder until time N.

$$X_1 \quad X_2 \quad \dots \quad X_k \quad X_{k+1}(Y^1) \quad X_{k+1}(Y^2) \quad \dots \quad X_N(Y^{N-k}).$$

The input distribution of the system until time N is given by

$$\left\{P_{X_1}, P_{X_2|X_1}, \dots, P_{X_k|X^{k-1}}, P_{X_k|X^{k-1},Y^1}, \dots, P_{X_N|X^{N-1},Y^{N-k}}\right\}.$$

This coupled with the channel distribution P^{ch} specifies the joint distribution of the system at time N as

$$\begin{aligned} P_{X^N,Y^N} &= P_{X_1} \cdot P^{ch}_{Y_1|X_1} \dots P_{X_{k+1}|X^k,Y^1} \cdot P^{ch}_{Y_{k+1}|X^{k+1},Y^k} \dots P_{X_N|X^{N-1},Y^{N-k}} \cdot P^{ch}_{Y_N|X^N,Y^{N-k}} \\ &= \vec{P}^k_{X^N|Y^N} \cdot \vec{P}^{ch}_{Y^N|X^N}. \end{aligned} \tag{14}$$

But from (4), we know that the joint distribution for the system can always be written as

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

Therefore, for a channel with k-delayed feedback, we must have

$$\vec{P}^k_{X^N|Y^N} = \vec{P}_{X^N|Y^N}. \tag{15}$$

Using this in Proposition 1 and (13), we get

$$I_k(X^N \to Y^N) = I(X^N \to Y^N)$$

for a channel with k-delayed feedback. It follows that the capacity can be written as

$$C^k_{FB} = \sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}^\mathbf{k}}} \lim_{N\to\infty} \frac{1}{N} I(X^N \to Y^N). \tag{16}$$

The no feedback case is just a special case of the above ($k = \infty$). When there is no feedback in the channel, the joint distribution is given by

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

Comparing with (4) again, we obtain

$$P_{X^N} = \vec{P}_{X^N|Y^N}. \tag{17}$$

which implies that when there is no feedback,

$$I(X^N; Y^N) = I(X^N \to Y^N).$$

Therefore,

$$C_{no-FB} = \sup_{P_\mathbf{X}} \lim_{N\to\infty} \frac{1}{N} I(X^N \to Y^N). \tag{18}$$

Thus for all three cases- unit-delayed feedback, k-delayed feedback and no feedback the objective function in the capacity expression is always directed information. But the space of optimization gets progressively smaller from unit-delay feedback ($\sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}^\mathbf{1}}}$) to k-delay feedback ($\sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}^\mathbf{k}}}$) to no feedback ($\sup_{P_\mathbf{X}}$).

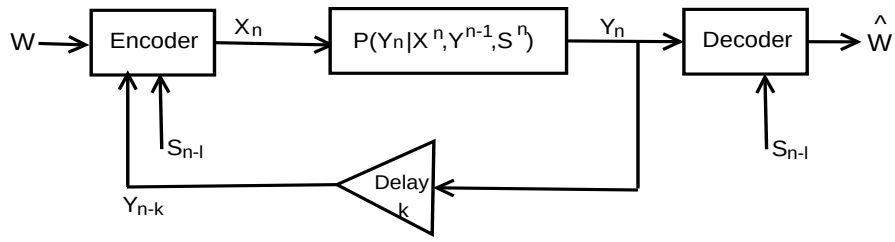

Figure 2: Channel with k-delayed feedback and l-delayed state-information

4 Channels with feedback and State information

In this section, we analyze the general model shown in Figure 2. There is a channel with input X, channel state S and output Y. It is defined by the sequence of distributions

$$P^{ch}_{\mathbf{Y}|\mathbf{X},\mathbf{S}} = \{P_{Y_n|X^n,Y^{n-1},S^n}\}_{n=1}^{\infty}. \tag{19}$$

There is k-delayed feedback, as in the previous section. In addition, the state information is known with delay l at *both* the encoder and decoder. The communication is performed over N channel uses. Thus at time n, the encoder has Y^{n-k} and S^{n-l} available to generate channel input X_n. We allow the possibility that l could be negative, i.e., the channel state information is available non-causally. For instance, $l = -3$ means that at time n, state symbols S^{n+3} are available to both encoder and decoder. It is understood that for negative l, $S^{n-l} = S^N$ when $n - l \geq N$.

4.1 Causal state information

Consider the case with causal state-information, i.e., $l \geq 0$. Since the channel input can depend on the fed-back symbols and the available state information, the input distribution is of the form

$$\vec{P}_{\mathbf{X}|\mathbf{Y^k},\mathbf{S^l}} = \{P_{X_n|X^{n-1},Y^{n-k},S^{n-l}}\}_{n=1}^{\infty}. \tag{20}$$

Consider N uses of the channel. First, we assume that the channel states S^N are produced *a priori* randomly according to $\{P_{S_n|S^{n-1}}\}_{n=1}^{\infty}$. Then, the joint distribution at time N is given by

$$\begin{aligned} P_{X^N,Y^N,S^N} &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^N} \cdot P_{Y_n|Y^{n-1},X^n,S^N} \\ &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{21}$$

where we have made two practical assumptions to obtain the second equality. The first one is the physical constraint on the channel input distribution that at time n, it can produce input X_n based only on what is available to it, viz., $(X^{n-1}, Y^{n-k}, S^{n-l})$. The second assumption stems from the definition of the channel.

On the other hand, we can also assume that the channel states are produced in real time, i.e., at time n, S_n, X_n are produced and the channel acts on S^n and X^n to produce Y_n. Then,

the joint distribution at time N is determined as

$$\begin{aligned} P_{X^N,Y^N,S^N} &= \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-1},S^n} \cdot P_{Y_n|Y^{n-1},X^n,S^n} \\ &= \prod_{n=1}^{N} P_{S_n|S^{n-1}} \cdot P_{X_n|X^{n-1},Y^{n-k},S^{n-l}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{22}$$

where we have again made the assumption relating to the physical constraints on the channel input. (21) and (22) arise from two different physical models that result in the same joint distribution because of the practical assumptions inherent in each case. Without loss of generality, we will consider the first model where S^N is produced *a priori*, since this model can also be used when the state information is known non-causally.

It is easy to see that the state symbols $\{S_n\}$ can be considered as additional outputs of the channel which are fed back to the encoder with delay l. This is justified since the state information is available with delay l at the encoder and the decoder acts on the channel outputs and state information only at the end of all reception (at time $N+l$). Hence the encoder knows Y^{n-k} and S^{n-l} to produce the input X_n at time n. Hence the information $I(Y^{n-k}S^{n-l}; X_n|X^{n-1})$ is not 'really transmitted' and comes for free. Using the same line of reasoning as before, we state the following theorem, omitting the proof.

Theorem 1. *The capacity of a channel whose output Y is fed back with delay k and state information S is available with delay l at both the transmitter and the receiver is*

$$C^{k,l}_{FB} = \sup_{\vec{P}_{\mathbf{X}|\mathbf{Y^k},\mathbf{S^l}}} \lim_{N\to\infty} \frac{1}{N} \left[I(X^N; Y^N S^N) - \sum_{n=\min(k,l)+1}^{N} I(Y^{n-k}S^{n-l}; X_n|X^{n-1}) \right]. \tag{23}$$

However, unlike the previous case, we cannot reduce this expression to a directed information quantity.

4.2 Non-Causal state information

In this case, the channel state information is produced *a priori* and the state information available at the encoder and decoder at time n is S^{n+m}, with $m > 0$. Hence, to generate input X_n, the encoder can use m channel states from the future. The input distribution is given by

$$\vec{P}_{\mathbf{X}|\mathbf{Y^k},\mathbf{S^{-m}}} = \{P_{X_n|X^{n-1},Y^{n-k},S^{n+m}}\}_{n=1}^{\infty}. \tag{24}$$

The joint distribution at time N is given by

$$\begin{aligned} P_{X^N,Y^N,S^N} &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^N} \cdot P_{Y_n|Y^{n-1},X^n,S^N} \\ &= P_{S^N} \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},Y^{n-1},S^{n+m}} \cdot P^{ch}_{Y_n|Y^{n-1},X^n,S^n}, \end{aligned} \tag{25}$$

where we have used the physical constraint on the channel input distribution and the definition of the channel to obtain the second equality. The arguments in the previous section hold here too, and we can show that the capacity is given by

$$C^{k,-m}_{FB} = \sup_{\vec{P}_{\mathbf{X}|\mathbf{Y^k},\mathbf{S^{-m}}}} \lim_{N\to\infty} \frac{1}{N} \left[I(X^N; Y^N S^N) - \sum_{n=1}^{N} I(Y^{n-k}S^{n+m}; X_n|X^{n-1}) \right]. \tag{26}$$

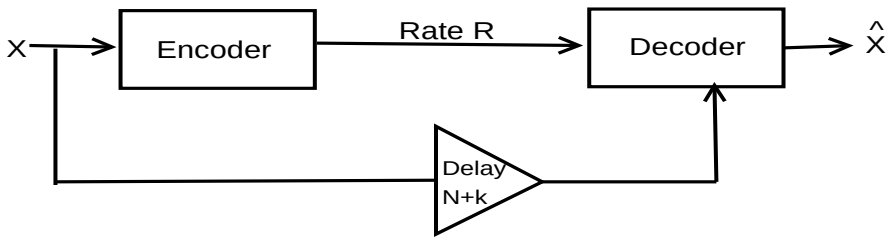

Figure 3: Source coding system with k-delayed feedforward.

5 Source Coding with Feedforward

We will now look at the source coding versions of the problems hitherto discussed. The dual of channel coding with feedback is source coding with feedforward [7, 8, 5, 6]. Feedforward means that in addition to the index, the decoder has access to some past source samples too. More precisely, for a system with feedforward with delay k, the decoder has access to source samples X^{n-k} to produce reconstruction symbol $\hat{X}_n$.

The model is shown in Figure 3. Consider a discrete source X with Nth order probability distribution P_{X^N}, alphabet $\mathcal{X}$ and reconstruction alphabet $\widehat{\mathcal{X}}$. There is an associated distortion measure $d : \mathcal{X}^N \times \widehat{\mathcal{X}}^N \to \mathbb{R}^+$.

A source code with feedforward of block length N and rate R is defined as follows. The encoder is a mapping to an index set: $e : \mathcal{X}^N \to \{1, \ldots, 2^{NR}\}$. The decoder receives the index transmitted by the encoder, and to reconstruct the nth sample, it has access to all the past $(n-k)$ samples of the source. In other words, the decoder is a sequence of mappings $g_n : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{n-k} \to \widehat{\mathcal{X}}, \quad n = 1, \ldots, N$. Let $\hat{x}^N$ denote the reconstruction of the source sequence x^N. The goal is to find the smallest rate R such that $Ed(x^N, \hat{x}^N) \leq D$.

We remark that in the following, for the sake of simplicity and intuition, we only consider the case where the joint process $\{X_n, \hat{X}_n\}_{n=1}^{\infty}$ is stationary and ergodic. So the term 'rate distortion function' will mean the minimum achievable rate assuming that the joint process is stationary and ergodic. As in the case of channels, the results can be extended to the non-stationary, non-ergodic case using information spectrum versions of the information quantities used here.

5.1 The Rate distortion function for Feedforward

We know that the rate-distortion function of a source without feedforward is [12]

$$R_{no-ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D} \lim_{N \to \infty} \frac{1}{N} I(\hat{X}^N; X^N), \tag{27}$$

where

$$\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty} \qquad \text{and} \qquad \lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) = \limsup_{N \to \infty} Ed(X^N, \hat{X}^N).$$

$I(X^N; \hat{X}^N)$ is the minimum number of bits required to represent the sequence X^N by the sequence $\hat{X}^N$. When there is feedforward with delay k, the decoder already knows X^{n-k} to reconstruct $\hat{X}_n$. Hence, we need not spend $I(X^{n-k}; \hat{X}_n | \hat{X}^{n-1})$ bits to code this information-this rate comes for free. In other words, the performance limit on this problem is characterized by

$$I_k(\hat{X}^N \to X^N) = I(\hat{X}^N; X^N) - \sum_{i=k+1}^{N} I(X^{i-k}; \hat{X}_i | \hat{X}^{i-1}) \tag{28}$$

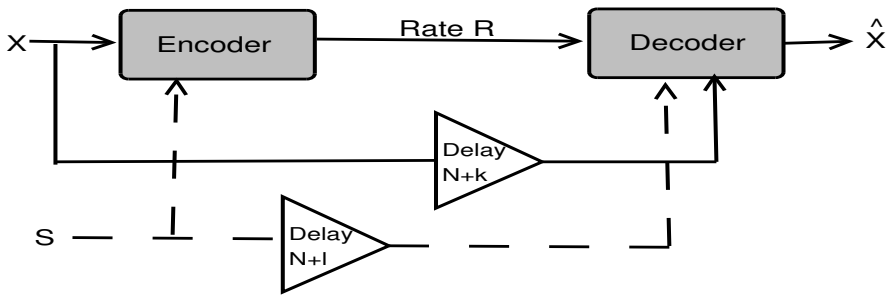

Figure 4: Source coding system with k-delayed feedforward and l-delayed side information.

The rate-distortion function for source coding with k-delay feedforward is [7]

$$R_{ff}^{k}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \lim_{N\to\infty} \frac{1}{N} I_k(\hat{X}^N \to X^N). \tag{29}$$

It can be seen from the above that for delay 1 feedforward,

$$R_{ff}^{1}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N), \tag{30}$$

which is analogous to the case of channel coding with delay 1 feedback.

5.2 Are they all the same?

For channels with feedback, we saw in Section 3.1 that the capacity expressions for both the no-feedback case and the k-delayed feedback case have the same objective function- directed information flowing from input to output. It is the constraint space of optimization that gets progressively smaller with increasing k and is the smallest for no-feedback. So does this happen in source coding with feedforward too?

We have a reversal of roles in source coding. Observe that the rate-distortion function with k-delayed feedforward given by (29) and the rate-distortion function for no feedforward given by (27). The set of optimization is always the same- $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$, subject to the distortion constraint. It is clear from (28) that the objective function increases with increasing k and is the largest when there is no feedforward ($I(\hat{X}^N; X^N)$).

In channel coding, for any k, the constraint on the input distribution ensures that

$$\vec{P}^{k}_{X^N|Y^N} = \vec{P}_{X^N|Y^N}.$$

holds. It is this constraint that makes the objective function equal to the directed information in each case. In source coding, both with and without feedforward, there is no such constraint on the conditional distribution that can be picked. Hence the objective functions are all different. In summary, for channels, the boost in capacity of channels due to feedback is because of a larger constraint set available to optimize the same objective function. In contrast, for sources, the decrease in the rate-distortion function due to feedforward is because we optimize a smaller objective function over the same constraint set.

5.3 Feed Forward with Delay and Side Information

In this section, we add side information to the feedforward problem (Figure 4). More precisely, there is a source X with feedforward delay k and side-information S available along with the

source at the encoder and fed forward with delay l to the decoder. We also allow the possibility that l could be negative.

The encoder is a mapping to an index set: $e : \mathcal{X}^N \times \mathcal{S}^N \to \{1, \ldots, 2^{NR}\}$. The decoder is a sequence of mappings $g_n : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{n-k} \times \mathcal{S}^{n-l} \to \widehat{\mathcal{X}}, \quad n = 1, \ldots, N$. Clearly, the side-information can be considered as 'another source' fed-forward with delay l. To reconstruct the nth sample, the decoder has access to all the past $(n-k)$ samples of the source and all the $(n-l)$ samples of side-information. Hence, we need not spend bits to code the information $I(\hat{X}_n; X^{n-k}, S^{n-l}|X^{n-1}), \quad \forall n$. We state the following theorem for source coding with feedforward and side-information.

Theorem 2. *The rate-distortion function for a source X with side-information S where the source has delay k feedforward and the side-information has delay l feedforward is*

$$R_{ff}^{k,l}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{XS}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{XS}})\leq D} \lim_{N\to\infty} \frac{1}{N}\left[I(\hat{X}^N; X^N S^N) - \sum_{n=1}^{N} I(\hat{X}_n; X^{n-k} S^{n-l}|\hat{X}^{n-1})\right],$$

where

$$\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{XS}} = \{P_{\hat{X}^n|X^n,S^n}\}_{n=1}^{\infty} \qquad \text{and} \qquad \lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{XS}}) = \limsup_{N\to\infty} Ed(X^N, \hat{X}^N).$$

References

[1] J. Massey, "Causality, Feedback and Directed Information," *Proceedings of the 1990 Symposium on Information Theory and its Applications (ISITA-90)*, pp. 303–305, 1990.

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

[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 Institute of Technology, Cambridge, MA, September 2000.

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

[6] S. S. Pradhan, "Source coding with feedforward: Gaussian sources," in *Proc. of IEEE International Symposium on Information Theory*, p. 212, June 2004.

[7] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward," in *Proc. Information Theory Workshop, San Antonio*, October 2004.

[8] E. Martinian and G. W. Wornell, "Source Coding with Fixed Lag Side Information," *Proceedings of the 42nd Annual Allerton Conference (Monticello, IL)*, 2004.

[9] S. Yang, A. Kavcic, and S. Tatikonda, "Feedback Capacity of Finite-State Machine Channels," *IEEE Trans. on Information Theory*, vol. 51, pp. 799–810, March 2005.

[10] R. Venkataramanan and S. S. Pradhan, "Source Coding with Feedforward: Rate Distortion Theorems and Error Exponents for a General Source," CSPL Technical Report No. 361, University of Michigan, May 2005.

[11] T. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, pp. 752–772, May 1993.

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