CSPL Technical Report No. 361, University of Michigan at Ann Arbor, May 9, 2005
Submitted to IEEE Transactions on Information Theory, August 2004.

Source Coding with Feed-forward: Rate-Distortion Theorems and Error Exponents for a General Source

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

May 9, 2005

Abstract

In this work, we consider a source coding model with feed-forward. We analyze a system with a noiseless, feed-forward link where the decoder has knowledge of all previous source samples while reconstructing the present sample. The rate-distortion function for an arbitrary source with feed-forward is derived in terms of directed information, a variant of mutual information. We further investigate the nature of the rate-distortion function with feed-forward for two common types of sources- discrete memoryless sources and Gaussian sources. We then derive a random coding error exponent which is used to bound the probability of decoding error for a source code (with feed- forward) of finite block length. The results are then extended to feed-forward with an arbitrary delay larger than the block length.

Keywords

Source coding with feed-forward, Real-time reconstruction, Side Information, Directed Information, Random Coding

1 Introduction

With the recent emergence of applications involving sensor networks [1], the problem of source coding with side-information at the decoder [2] has gained special significance. Here the source of information, say modeled as an independent identically distributed (IID) random process $\{X_n\}_{n=1}^{\infty}$, needs to be encoded in blocks of length N into a message (description) W. W is to be transmitted over a noiseless channel of finite rate to a decoder, which has access to some information $\{Y_n\}_{n=1}^{\infty}$ (referred to as side information and also modeled as an IID random process) that is correlated to the source X. The ith sample of X is correlated to the ith sample of Y for all i. The decoder with the help of the side information Y and the bit stream W obtains an optimal estimate of N samples of the source at once, and hence, over time, a reconstruction of the process X. The goal is to minimize the reconstruction distortion for a fixed transmission rate. The optimal rate-distortion performance limit is obtained by Wyner and Ziv in [2]. The encoder and the decoder are in time-synchrony. To reconstruct a set of N samples of X, the decoder uses the corresponding set of N samples of Y. This is used to

*This work was supported by NSF Grants CCR-0219735 and CCF-0329715. This work was presented in part at IEEE Information Theory Workshop (ITW), San Antonio, TX, October 2004.

model the compression problem in general sensor networks where X and Y are the correlated signals captured by the sensor and the destination nodes.

As one can see, the implicit assumption is that the underlying sample pairs (X_i, Y_i) are instantaneously observed respectively at the encoder and the decoder. So after an encoding delay of N samples, when the decoder gets the message W (say being transmitted instantaneously using electromagnetic waves), it has access to the corresponding N samples of Y, so that the decoding can begin immediately. The time-line of the samples of the source, the message and the side information is depicted in Fig. 1 for $N = 5$. Note that in this model, for example, at the 6th time unit, the decoder reconstructs $\hat{X}_1, \ldots \hat{X}_5$ simultaneously as a function of W and $Y_1, \ldots Y_5$, though it may display them as shown in Fig. 1.

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	–	–	–	–	W	–	–	–	–	W
Side info	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Figure 1: Time-line: instantaneous observations

The key question that we would like to ask is *what happens if the underlying signal field is traveling slowly (compared to the speed of electromagnetic wave propagation) from the geographical location of the encoder to that of the decoder*, so that a there is a delay between the instant when ith sample of X is observed at the encoder and the instant when corresponding ith sample of Y is observed at the decoder, with the additional constraint that the reconstruction be real-time. In that case, we need a new dynamic compression model which is depicted in Fig. 2. Here it is assumed that the signal field delay is 6 time units, so that for real-time reconstruction of the ith source sample, all the past samples of the side information are available. In other words, now the decoding

Time	1	2	3	4	5	6	7	8	9	10
Source	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Encoder	–	–	–	–	W	–	–	–	–	W
Side info	–	–	–	–	–	–	Y_1	Y_2	Y_3	Y_4
Decoder						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Figure 2: Time-line: delayed observations

operation consists of a sequence of functions such that the ith reconstruction is a function of W and $(i-1)$ side information samples. The encoding operation, however, remains as in [2], i.e., a mapping from the N-product source alphabet to an index set of size 2^{NR} where R is the rate of transmission. This general compression model takes this important physical signal delay into account in its real-time reconstruction. We refer to this model as source coding with feed-forward. Note that in this problem, the encoder is non-causal and the decoder is causal. In this work, as a first step, we consider an idealized version of this problem called source coding with noiseless feed-forward. In this model, we assume that noiseless source samples are available with a delay at the decoder, i.e. $Y = X$.

From Fig. 2, it is clear that the model with $Y = X$ is meaningful only when the delay is at least $N + 1$, where the block length is N. However, for a general Y, any delay leads to a valid problem.

Related Work: The problem of source coding with noiseless feed-forward was first considered by Weissman and Merhav in the context of competitive prediction in [3, 4], where a complete characterization of attainable performance is provided for any source that can be represented auto-regressively with an IID innovation process, as well as any innovation process satisfying the Shannon lower bound (SLB) with equality. In particular, it was shown that for IID sources, as well as all sources that satisfy SLB with equality and with single-letter difference distortion measures, feed-forward does not reduce the optimal rate-distortion function and does not increase the optimal error exponent with block coding. Later, the model of source coding with general feed-forward was considered in [5] as a variant of the problem of source coding with side information at the decoder, and a quantization scheme with linear processing for IID Gaussian sources with mean squared error distortion function and with noiseless feed-forward was reported. It was also shown that this scheme approaches the optimal rate-distortion function. In [6], an elegant variable-length coding strategy to achieve the optimal Shannon rate-distortion bound for any finite-alphabet IID source with feed-forward was presented, along with a beautiful illustrative example. The problem of source coding with feed-forward is also related to source coding with a delay-dependent distortion function [7] and causal source coding [8].

The main results of this paper can be summarized as follows:

1. The optimal rate-distortion function for general discrete sources with a general distortion measure and with noiseless feed-forward, $R_{ff}(D)$, is given by the minimum of the directed information function [9] flowing from the reconstruction to the source. $R_{ff}(D) \leq R(D)$, where $R(D)$ denotes the optimal Shannon rate-distortion function for the source without feed-forward.

2. The performance of the best possible source code (with feed-forward) of rate R, distortion D and block length N is characterized by an error exponent. We provide a random coding error exponent $E_{N-ff}(R, D)$

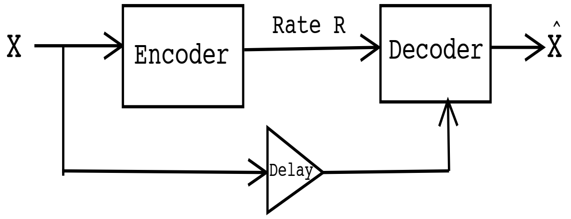

Figure 3: Source coding system with feed-forward.

and show that it is greater than or equal to the random coding error exponent without feed-forward.

3. Feed-forward does not decrease the rate-distortion function of general discrete memoryless sources with memoryless distortion measures.

The paper is organized as follows. In Section 2 we give a fairly formal definition of the above source coding model and the intuition behind the proposed approach. Instead of giving the main result for the most general sources and then considering the special cases, we first consider the special case when the source and the reconstruction processes are jointly stationary and ergodic and give a direct coding theorem in Section 3 which captures the essence of this problem. We then give the direct and the converse coding theorems for general sources in Section 4. In that section we also consider some special cases such as discrete memoryless sources and Gaussian sources. Random coding error exponents are considered in the general setting in Section 5. We extend our results to arbitrary delays in Section 6 and finally, concluding remarks are given in Section 7.

2 The Source Coding Model

2.1 Problem Statement

The model is shown in Figure 3. Consider a general 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 $d_N : \mathcal{X}^N \times \widehat{\mathcal{X}}^N \to \mathbb{R}^+$ on pairs of sequences. 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 $d' : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$.

For an $(N, 2^{NR})$ source code of block length N and rate R, 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 ith sample, it has access to all the past $(i-1)$ samples of the source. In other words, the decoder is a sequence of mappings $g_i : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{i-1} \to \widehat{\mathcal{X}}, \quad i = 1, \ldots, N$. Let $\hat{x}^N$ denote the reconstruction of the source sequence x^N. We want to minimize R for a given distortion constraint. We consider two types of distortion constraints in this work: 1) expected distortion constraint and 2) probability-1 distortion constraint. These constraints are formally defined in the sequel. For any D, let $R_{ff}(D)$ denote the infimum of R over all encoder decoder pairs for any block length N such that the distortion is less than D. It is worthwhile noting that source coding with feed-forward can be considered the dual problem [10, 11] of channel coding with feedback.

The relevance of this problem extends beyond the application outlined in Section 1. As an example, consider a stock market game in which we want to predict the share price of some company over an $N-$day period. Let the share price on day i be X_i. On the morning of the $i-$th day, we have to make our guess $\hat{X}_i$. In the evening, we know X_i- the actual closing price of the share for that day. Let $d(X_i, \hat{X}_i)$ be a measure of our guessing error. Note that to make our guess $\hat{X}_i$, we know X^{i-1}, the actual share prices of the previous days. We want to play this guessing game over an N day period.

Further suppose that at the beginning of this period, we have some a priori information about different possible scenarios over the next N days. For example, the scenarios could be something like

- Scenario 1: Demand high in the third week, low in the fifth week, layoffs in sixth week.
- Scenario 2: Price initially steady; company results expected to be good, declared on day m, steady increase after that.
- ...
- Scenario 2^{NR}.

The a priori information tells us which of the 2^{NR} scenarios is relevant for the $N-$day period. The question we ask is: Over the N-day period, if we want our average prediction error to satisfy

$$\frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_i) \leq D, \tag{1}$$

what is the minimum a priori information needed? Note that it makes sense for the number of possible scenarios to grow as 2^{NR} since we will need more information to maintain the same level of performance D as N gets larger. Clearly, this problem of 'prediction with a priori information' is identical to source coding with feed-forward.

2.2 Intuition behind the proposed approach

To analyze the problem of source coding with feed-forward we need a directional notion of information. This is given by directed information, as defined by Massey [9]. This notion was earlier studied in [12, 13, 14] in the context of dependence and feedback between random processes. More recently, directed information has been used to characterize the capacity of channels with feedback [15, 16].

Definition 2.1. *[9] The directed information flowing from a random vector A^N to another random vector B^N is defined as*

$$I(A^N \to B^N) = \sum_{n=1}^{N} I(A^n; B_n | B^{n-1}). \tag{2}$$

Note that the definition is similar to that of mutual information $I(A^N; B^N)$ except that the mutual information has A^N instead of A^n in the summation on the right. The directed information has a nice interpretation in the context of our problem.

An interesting way to understand any source compression system is to analyze the corresponding backward test channel [17, 18, 19]. This is a fictitious channel which connects the source with the reconstruction, characterized by the conditional distribution of the source given the reconstruction. The decoder first gets the index W (sent by the encoder) containing the information about the first (say) N samples of X. The process of reconstruction starts by first spitting out the reconstruction of the first sample $\hat{X}_1 = g_1(W)$ as a function of W alone. In the next clock cycle, the decoder has W and X_1. This can be interpreted as follows: $\hat{X}_1$ goes through a non-anticipatory fictitious channel to produce X_1 and is fed back to the decoder. Now the decoder reconstructs the second sample $\hat{X}_2 = g_2(W, X_1)$ as a function of W and X_1. As before, we can interpret it as $\hat{X}_2$ going through the test channel to produce X_2 which is fed back to the decoder and so on. So this test channel can be thought of as having $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N$ as input and $X_1, X_2, \ldots, X_N$ as output with a sequence of conditional distributions given by

$$\hat{Q}_1(X_1|\hat{X}_1), \hat{Q}_2(X_2|X_1, \hat{X}_1, \hat{X}_2), \ldots, \hat{Q}_i(X_i|X^{i-1}, \hat{X}^i), \ldots, \hat{Q}_N(X_N|X^{N-1}, \hat{X}^N),$$

where X^i denotes the vector of $X_1, X_2, \ldots, X_i$. This sequence of conditional distributions is related to the source and the encoder transformation in the following way. Note that the source distribution $P_{X^N}(X^N)$ and the quantizer transformation $P_{\hat{X}^N|X^N}(\hat{X}^N|X^N)$ fix the joint distribution $P_{X^N,\hat{X}^N}(X^N, \hat{X}^N)$. This can be factored into two components as follows:

$$P_{X^N,\hat{X}^N}(X^N, \hat{X}^N) = \prod_{i=1}^{N} P_i(X_i, \hat{X}_i|X^{i-1}\hat{X}^{i-1}) = \prod_{i=1}^{N} Q_i(\hat{X}_i|X^{i-1}, \hat{X}^{i-1}) \prod_{i=1}^{N} \hat{Q}_i(X_i|X^{i-1}\hat{X}^i),$$

where Q characterizes the decoder reconstruction function, whereas $\hat{Q}$ denotes the test channel conditional distribution, and both of them are assumed to have memory. This is illustrated in Fig. 4. The rate required to quantize X^N to $\hat{X}^N$ (i.e., the rate of transmission of the message W) can now be interpreted as the rate of flow of information through this test channel. This is given by directed information as follows: $I(\hat{X}^N \to X^N) = \sum_{i=1}^{N} I(\hat{X}^i; X_i|X^{i-1})$. Using simply the chain rule, we get

$$I(\hat{X}^N \to X^N) = I(\hat{X}^N; X^N) - \sum_{i=2}^{N} I(X^{i-1}; \hat{X}_i|\hat{X}^{i-1}). \tag{3}$$

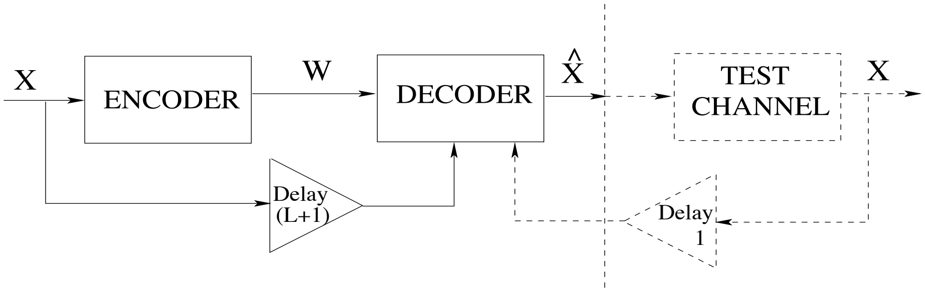

Figure 4: Backward test channel interpretation

We know for the standard source coding problem (without feed-forward) that the mutual information $I(\hat{X}^N; X^N)$ is the number of bits required to represent X^N with $\hat{X}^N$. Since the decoder knows the symbols X^{i-1} to reconstruct $\hat{X}_i$, we need not spend $I(X^{i-1}; \hat{X}_i|\hat{X}^{i-1})$ bits to code this information, hence this rate comes for free. In other words, the performance limit on this problem is given by the minimum of the directed information.

3 Stationary and ergodic joint processes

In this section, we will provide a direct coding theorem for a general source with feed-forward assuming that the joint random process $\{X_n, \hat{X}_n\}$ is stationary and ergodic. This assumption leads to a rather simple and intuitive proof of the rate-distortion theorem on the lines of the proof of the rate distortion theorem for discrete memoryless sources in [18]. The purpose of this section is give intuition about how feed-forward helps in source coding before going into full generality in the following sections. We will use a new kind of typicality, tailored for our problem of source coding with feed-forward. A word about the notation before we state the theorem. All logarithms used in the sequel are assumed to be with base 2, unless otherwise stated. The source distribution, defined by a sequence of finite-dimensional distributions [20] is denoted by

$$\mathbf{P_X} \triangleq \{P_{X^n}\}_{n=1}^{\infty}. \tag{4}$$

Similarly, a conditional distribution is denoted by

$$\mathbf{P_{\hat{X}|X}} \triangleq \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}. \tag{5}$$

Finally, for stationary and ergodic joint processes, the directed information rate exists and is defined by

$$I(\hat{X} \to X) = \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N). \tag{6}$$

We use an expected distortion criterion here. For simplicity, we assume $d_N(x^N, \hat{x}^N) = \frac{1}{N}\sum_{i=1}^N d(x_i, \hat{x}_i)$, where $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. Let d_{max} be the maximum of $d(x, \hat{x}) \quad \forall x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$. Since the distortion measure is bounded, $\lim_{N\to\infty} E[d_N(X^N, \hat{X}^N)]$ exists.

Definition 3.1. *R is an achievable rate at expected distortion D if $\forall \epsilon > 0$, for all sufficiently large N, there exists an $(N, 2^{NR})$ code such that*

$$E_{X^N}\left[d_N(X^N, \hat{X}^N)\right] \leq D + \epsilon,$$

where $\hat{X}^N$ denotes the reconstruction of X^N.

Theorem 1. *For a discrete stationary and ergodic source X characterized by a distribution $\mathbf{P_X}$, all rates R such that*

$$R \geq R^*(D) \triangleq \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: \lim_{N\to\infty} E[d_N(X^N, \hat{X}^N)] \leq D} I(\hat{X} \to X)$$

are achievable at expected distortion D.

Proof. Since the AEP holds for stationary and ergodic processes [18], we have

$$\begin{aligned} -\frac{1}{N}\log P(X^N) &\to H(X) \quad \text{w.pr.1}, \\ -\frac{1}{N}\log P(X^N, \hat{X}^N) &\to H(X, \hat{X}) \quad \text{w.pr.1}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} H(X) &= \lim_{N\to\infty} H(X_N | X^{N-1}) = \lim_{N\to\infty} \frac{1}{N} H(X^N), \\ H(X, \hat{X}) &= \lim_{N\to\infty} H(X_N, \hat{X}_N | X^{N-1}, \hat{X}^{N-1}) = \lim_{N\to\infty} \frac{1}{N} H(X^N, \hat{X}^N). \end{aligned}$$

We now define two 'directed' quantities, first introduced in [16] in the context of channels with feedback. These will be frequently used in the rest of this paper. $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$,

$$\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \triangleq \prod_{i=1}^{N} P_{\hat{X}_i|\hat{X}^{i-1}, X^{i-1}}(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}), \tag{8}$$

$$\vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \triangleq \prod_{i=1}^{N} P_{X_i|\hat{X}^i, X^{i-1}}(x_i|\hat{x}^i, x^{i-1}). \tag{9}$$

These can be pictured in terms of the test channel from $\hat{X}$ to X. (8) describes the sequence of input distributions to this test channel and (9) specifies the test channel. Recall that the joint distribution can be split as

$$P_{X^N, \hat{X}^N} = \prod_{i=1}^{N} P_{\hat{X}_i|\hat{X}^{i-1}, X^{i-1}} \cdot P_{X_i|\hat{X}^i, X^{i-1}}. \tag{10}$$

The basic ingredient in our proof is the following Lemma which says that a property analogous to the AEP holds for the directed quantities defined in (8) and (9). Let $H(\hat{X}^N || X^N) = \sum_{i=1}^{N} H(\hat{X}_i | \hat{X}^{i-1}, X^i)$.

Lemma 3.1. *If the process $\{X_i, \hat{X}_i\}_{i=1}^{\infty}$ is stationary and ergodic, we have*

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) \to H(\hat{X}||0X) \quad \textit{w.pr.}1, \tag{11}$$

where

$$\begin{aligned}H(\hat{X}||0X) &\triangleq \lim_{N\to\infty}\frac{1}{N}H(\hat{X}^N||0X^{N-1})\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}H(\hat{X}_i|X^{i-1},\hat{X}^{i-1}) \qquad (12)\\ &= \lim_{N\to\infty}H(\hat{X}_N|X^{N-1},\hat{X}^{N-1}),\end{aligned}$$

where $0X^{N-1}$ *denotes the sequence* $[\_, X_1, X_2, \ldots, X_{N-1}]$.

The proof of the lemma is similar to the Shannon-McMillan-Breiman Theorem in [18] and is given in Appendix A. We now define a new kind of joint distortion typicality. Given the source $\mathbf{P_X}$, fix any conditional distribution $\mathbf{P_{\hat{X}|X}}$ to get a joint distribution $\mathbf{P_{X,\hat{X}}} = \{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$. Also recall that the distortion is given by $d_N(x^N,\hat{x}^N) = \frac{1}{N}\sum_{i=1}^{N} d(x_i,\hat{x}_i)$.

Definition 3.2. *An ordered sequence pair* $(x^N,\hat{x}^N)$ *with* $x^N \in \mathcal{X}^N$ *and* $\hat{x}^N \in \hat{\mathcal{X}}^N$ *is said to be directed distortion* ϵ*-typical if:*

$$\begin{aligned}\left|-\frac{1}{N}\log P_{X^N}(x^N) - H(X)\right| &< \epsilon\\ \left|-\frac{1}{N}\log P_{X^N,\hat{X}^N}(x^N,\hat{x}^N) - H(X,\hat{X})\right| &< \epsilon\\ \left|-\frac{1}{N}\log \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) - H(\hat{X}||0X)\right| &< \epsilon\\ \left|d_N(x^N,\hat{x}^N) - Ed_N(X^N,\hat{X}^N)\right| &< \epsilon\end{aligned}$$

We denote the set of directed distortion ϵ-typical pairs by $\mathcal{A}_\epsilon^N$.

Lemma 3.2. *If an ordered pair* $(X^N,\hat{X}^N)$ *is drawn from* $P_{X^N,\hat{X}^N}$*, then*

$$\Pr((X^N,\hat{X}^N) \in \mathcal{A}_\epsilon^N) \to 1 \quad as \quad N \to \infty. \qquad (13)$$

Proof. From the AEP for stationary and ergodic processes, the first, second and fourth conditions in Definition 3.2 are satisfied with probability 1 as $N \to \infty$. From Lemma 3.1, the third condition is satisfied with probability 1 as $N \to \infty$, proving the lemma. □

Lemma 3.3. *For all* $(x^N,\hat{x}^N) \in \mathcal{A}_\epsilon^N$,

$$\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \geq P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot 2^{-N(I(\hat{X}\to X)+3\epsilon)}. \qquad (14)$$

Proof.

$$\begin{aligned}
P_{\hat{X}^N|X^N}(\hat{x}^N|x^N) &= \frac{P_{X^N,\hat{X}^N}}{P_{X^N}} \\
&= \vec{P}_{\hat{X}^N|X^N} \frac{P_{X^N,\hat{X}^N}}{\vec{P}_{\hat{X}^N|X^N} \cdot P_{X^N}} \\
&\leq \vec{P}_{\hat{X}^N|X^N} \cdot \frac{2^{-N(H(X,\hat{X})-\epsilon)}}{2^{-N(H(\hat{X}||0X)+\epsilon)} \cdot 2^{-N(H(X)+\epsilon)}} \\
&= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \cdot 2^{N(I(\hat{X}\to X)+3\epsilon)},
\end{aligned} \tag{15}$$

from which the lemma follows. For the last inequality in (15), we have used the fact [9] that $I(X^N;\hat{X}^N) = I(\hat{X}^N \to X^N) + I(0X^{N-1} \to \hat{X}^N)$. Hence

$$\begin{aligned}
H(\hat{X}^N||0X^{N-1}) + H(X^N) - H(X^N,\hat{X}^N) &= H(\hat{X}^N||0X^{N-1}) - H(\hat{X}^N|X^N) \\
&= H(\hat{X}^N) - H(\hat{X}^N|X^N) - [H(\hat{X}^N) - H(\hat{X}^N||0X^{N-1})] \\
&= I(X^N;\hat{X}^N) - I(0X^{N-1} \to \hat{X}^N) \\
&= I(\hat{X}^N \to X^N).
\end{aligned} \tag{16}$$

□

We are now ready to prove the achievability of $R^*(D)$. At this point, it is worth comparing the expression in Theorem 1 for $R^*(D)$ with the optimal rate-distortion function for a source without feed-forward. The constraint set for the infimum is the same in both cases, but the objective function in $R^*(D)$ is less than or equal to that in the no-feed-forward rate-distortion function since $I(\hat{X}^N \to X^N) \leq I(\hat{X}^N; X^N)$.

Codebook generation: In source coding with feed-forward, to produce the ith reconstruction symbol $\hat{x}_i$, the decoder knows the first i source samples x^{i-1}. This means that we could have a different reconstruction $\hat{x}_i$ for each x^{i-1}. Thus we can have a codebook of code-trees rather than codewords. A code tree is constructed as follows.

Pick a joint distribution $\mathbf{P}_{\hat{\mathbf{X}},\mathbf{X}} = \{P_{\hat{X}^n,\hat{X}^n}\}_{n=1}^{\infty}$, such that the $X-$marginal has the distribution $\mathbf{P}_{\mathbf{X}}$ and $\lim_{N\to\infty} Ed_N(X^N,\hat{X}^N) \leq D$. This joint distribution is stationary and ergodic by assumption. Fix ϵ and the block length N. Pick the first input symbol $\hat{x}_1$ randomly according to the distribution $P_{\hat{X}_1}$. To choose the next symbol, the encoder knows x_1. Therefore we have $|\mathcal{X}|$ choices for the $\hat{x}_2$ depending on the x_1 observed. Thus, $\hat{x}_2$ is chosen randomly and independently according to the distribution $P_{\hat{X}_2|\hat{x}_1,x_1}$ for each possible x_1. For each of these $\hat{x}_2$, there are $|\mathcal{X}|$ possible $\hat{x}_3$'s (depending on the x_2 observed) picked randomly and independently according to the distribution $P_{\hat{X}_3|\hat{x}^2,x^2}$. We continue picking the input symbols in this manner and finally we pick $\hat{x}_N$ according to $P_{\hat{X}_N|\hat{x}^{N-1},x^{N-1}}$. The process of choosing a code-tree for a binary alphabet in shown in Figure 5. We obtain 2^{NR} such independent and randomly chosen code-trees in the same fashion to form the codebook.

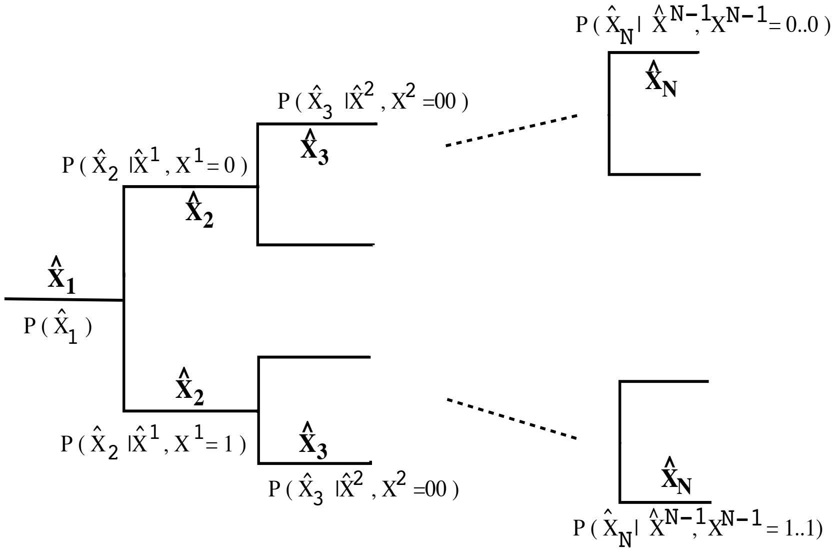

Figure 5: Code function for a binary source.

Encoding: We will use jointly typical encoding. The encoder has the sequence x^N. It traces the path determined by x^{N-1} on each of the 2^{NR} trees of the codebook. Each of these paths corresponds to a reconstruction sequence $\hat{x}^N[i] \quad (i \in \{1, \ldots 2^{NR}\})$. The encoder chooses a $\hat{x}^N[W]$ that is directed distortion ϵ-typical with x^N and sends W to the decoder. If no such typical $\hat{x}^N$ is found, an encoding error is declared.

Decoding: The decoder receives the index W from the encoder ($W \in \{1, \ldots, 2^{NR}\}$). It uses the Wth code-tree and obtains the reconstruction symbols along the path traced by $\{x_k\}_{k=1}^{N-1}$ that are fed-forward. For instance, suppose $\hat{\mathcal{X}}$ and $\mathcal{X}$ are binary alphabets and the code-tree in Figure 5 is used. If the fed-forward sequence, x^{N-1}, is the all zero sequence, the decoder traces the upper-most path on the tree and obtains the reconstruction symbols along that path.

Distortion: There are two types of source sequences x^N- a) Good sequences x^N, that are properly encoded with distortion $\leq D + \epsilon$, b) Bad source sequences x^N, for which the encoder cannot find a distortion-typical path. Let P_e denote the probability of the set of bad source sequences for the code. The expected distortion for the code can be written as

$$E[d_N(X^N, \hat{X}^N)] \leq D + \epsilon + P_e d_{max}. \tag{17}$$

We calculate the expected distortion averaged over all random codebooks. This is given by

$$E_{\mathcal{C}}[E[d_N(X^N, \hat{X}^N)]] \leq D + \epsilon + \overline{P}_e d_{max}, \tag{18}$$

where $\overline{P}_e$ is the expected probability of the set of bad X^N sequences, the expectation being computed over all randomly chosen codes. We will show that when R satisfies the condition given by Theorem 1, P_e goes to 0 as $N \to \infty$. This would prove the existence of at least one rate-R code with expected distortion $\leq D + \epsilon$.

Average Probabilty of Error $\overline{P}_e$: $\overline{P}_e$ is the probability that for a random code $\mathcal{C}$ and a random source sequence X^N, none of the 2^{NR} codewords are jointly typical with X^N. Let $J(\mathcal{C})$ denote the set of good (properly encoded) source sequences for code $\mathcal{C}$. Now,

$$\begin{aligned}\overline{P}_e &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \sum_{x^N : x^N \notin J(\mathcal{C})} P(x^N) && (19)\\ &= \sum_{x^N} P(x^N) \sum_{\mathcal{C} : x^N \notin J(\mathcal{C})} \Pr(\mathcal{C}). && (20)\end{aligned}$$

The inner summation is the probability of choosing a codebook that does not well represent the x^N specified in the outer summation. The probability that a single randomly chosen codeword does not well represent x^N is

$$\Pr\left((x^N, \hat{X}^N) \notin A_\epsilon^N\right) = 1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N). \qquad (21)$$

Thus the probability of choosing a codebook that does not well represent x^N is

$$\left[1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N)\right]^{2^{NR}}. \qquad (22)$$

Substituting this in (20), we get

$$\overline{P}_e = \sum_{x^N} P(x^N) \left[1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N)\right]^{2^{NR}}. \qquad (23)$$

We can now use Lemma 3.3 to obtain

$$\overline{P}_e \leq \sum_{x^N} P(x^N) \left[1 - 2^{-N(I(\hat{X} \to X) + 3\epsilon)} \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} P(\hat{x}^N | x^N)\right]^{2^{NR}}. \qquad (24)$$

As shown in [18], the inequality

$$(1 - xy)^n \leq 1 - y + e^{-xn} \qquad (25)$$

holds for $n > 0$ and $0 \leq x, y \leq 1$. Using this in (24), we get

$$\begin{aligned}\overline{P}_e &\leq \left[\sum_{x^N} P(x^N) \sum_{\hat{x}^N : (x^N, \hat{x}^N) \notin A_\epsilon^N} P(\hat{x}^N | x^N)\right] + e^{-2^{N(R - I(\hat{X} \to X) - 3\epsilon)}} \\ &= \sum_{(x^N, \hat{x}^N) \notin A_\epsilon^N} P(x^N, \hat{x}^N) + e^{-2^{N(R - I(\hat{X} \to X) - 3\epsilon)}}.\end{aligned} \qquad (26)$$

The first term is the probability that a pair $(x^N, \hat{x}^N)$ chosen according to the distribution $P_{X^N, \hat{X}^N}$ is not directed distortion ϵ-typical. From Lemma 3.2, this vanishes as $N \to \infty$. Therefore, $\overline{P}_e \to 0$ as long as $R > I(\hat{X} \to X) + 3\epsilon$. Thus we have shown that there exists a code with rate arbitrarily close to $R^*(D)$ that has expected distortion arbitrarily close to D. □

Remark: We now make some observations connecting the above discussion to channel coding with feedback. Consider a channel with input X_n and output Y_n with perfect feedback, i.e. to determine X_n, the encoder knows Y^{n-1}. The channel, characterized by a sequence of distributions $\vec{P}_{\mathbf{Y}|\mathbf{X}} = \{P_{Y_n|X^n,Y^{n-1}}\}_{n=1}^{\infty}$, is fixed. What the encoder can control is the input distribution $\vec{P}_{\mathbf{X}|\mathbf{Y}} = \{P_{X_n|X^{n-1},Y^{n-1}}\}_{n=1}^{\infty}$. Note that

$$\mathbf{P}_{\mathbf{X},\mathbf{Y}} = \vec{P}_{\mathbf{Y}|\mathbf{X}} \cdot \vec{P}_{\mathbf{X}|\mathbf{Y}}.$$

Under the assumption that the joint process $\{X_n, Y_n\}_{n=1}^{\infty}$ is stationary and ergodic, we can use methods similar to those used in this section to show that all rates less that $sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}}} I(X \to Y)$ are achievable with feedback. Comparing this with the no-feedback capacity of the channel, given by $sup_{\mathbf{P}_\mathbf{X}} I(X;Y)$, we see that although the objective function with feedback is smaller $\left(I(X^N \to Y^N) \leq I(X^N;Y^N)\right)$, the constraint set of optimization is larger when feedback is present since the space of $\mathbf{P}_\mathbf{X}$ is contained in the space of $\vec{P}_{\mathbf{X}|\mathbf{Y}}$.

Compare this with the source coding problem where $\mathbf{P}_\mathbf{X}$ is fixed. With or without feed-forward, the constraint set of optimization remains the same ($\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ subject to distortion constraint). But the objective function with feed-forward- $I(\hat{X} \to X)$- is smaller than in the no-feed-forward case, $I(\hat{X};X)$. In summary, for channels, the boost in capacity due to feedback is due to a larger constraint set of optimization. In contrast, for sources, the decrease in the rate-distortion function due to feed-forward is due to a smaller objective function.

4 General sources

4.1 Rate-distortion theorem

In this section, we first describe the apparatus we will use for proving coding theorems for general discrete sources with feed-forward. We introduce code-functions, which map the feed-forward information to a source reconstruction symbol $\hat{X}$. The idea of code-functions was introduced by Shannon in 1961 [21]. We first give a formal definition of a code-function and then see how it is useful in analyzing systems with feed-forward.

Definition 4.1. *A source code-function f^N is a set of N functions $\{f_n\}_{n=1}^N$ such that $f_n : \mathcal{X}^{n-1} \to \widehat{\mathcal{X}}$ maps each source sequence $x^{n-1} \in \mathcal{X}^{n-1}$ to a reconstruction symbol $\hat{x}_n \in \widehat{\mathcal{X}}$. Denote the space of all code-functions by $\mathcal{F}^N = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \mathcal{F}_N \triangleq \{f^N : f^N$ is a code function$\}$.*

Definition 4.2. *A $(N, 2^{NR})$ source codebook of rate R and block length N is a set of 2^{NR} code-functions. Denote them by $f^N[w], \quad w = 1, \ldots, 2^{NR}$.*

For each source sequence of length N, the encoder sends an index to the decoder. Using the code-function corresponding to this index, the decoder maps the information fed forward from the source to produce an estimate $\hat{X}$. A code-function can be represented as a tree as in Figure 5. In a system without feed forward, a code-function generates the reconstruction independent of the past source samples. In this case, the code-function reduces to a codeword. In other words, for a system without feed-forward, a source codeword is a source code-function $f^N = \{f_1, \dots, f_N\}$ where for each $n \in \{1, \dots, N\}$, the function f_n is a constant mapping.

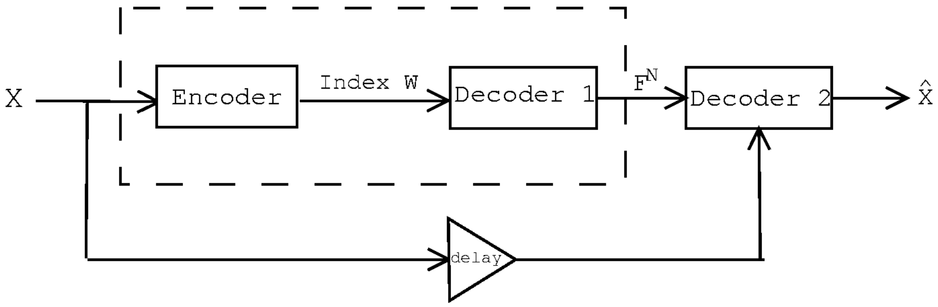

Figure 6: Representation of a source coding scheme with feed-forward.

A source coding system with feed-forward can be thought of as having two components. The first is a usual source coding problem with F^N as the reconstruction for the source sequence X^N. In other words, for each source sequence x^N, the encoder chooses the best code-function among $f^N[i], \quad i \in \{1, \dots, 2^{NR}\}$ and sends the index of the chosen code function. This is the part inside the dashed box in Figure 6. If we denote the chosen code-function by f^N, the second component (decoder 2 in Figure 6) produces the reconstruction given by

$$\hat{X}_i = f_i(X^{i-1}), \qquad i = 1, \dots, N. \tag{27}$$

In the sequel, we will use the notation $\hat{X}^N = f^N(X^{N-1})$ as shorthand to collectively refer to the N equations described by (27). In source coding with feed-forward, the encoder induces a conditional distribution $\forall f^N \in \mathcal{F}^N, x^N \in \mathcal{X}^N$ given by

$$P_{F^N|X^N}(f^N|x^N) = \begin{cases} 1, & \text{if } f^N = \text{the code-function chosen by the encoder.} \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

The reconstruction $\hat{x}^N$ is uniquely determined by f^N and x^N. Thus

$$P_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N, x^N) = \delta_{\{\hat{x}^N = f^N(x^{N-1})\}}. \tag{29}$$

Therefore, given a source distribution P_{X^N} and a source code with feed-forward, a unique joint distribution Q of X^N, F^N and $\hat{X}^N$ is determined: $\forall x^N \in \mathcal{X}^N, \quad f^N \in \{f^N[i] : 1 \leq i \leq 2^{NR}\}, \quad \hat{x}^N \in \hat{\mathcal{X}}^N$,

$$\begin{aligned} Q_{X^N,F^N,\hat{X}^N}(x^N, f^N, \hat{x}^N) &= P_{X^N}(x^N) \cdot P_{F^N|X^N}(f^N|x^N) \cdot P_{\hat{X}^N|F^N,X^N}(\hat{x}^N|f^N, x^N) \\ &= P_{X^N}(x^N) \cdot \delta_{\{f^N = e(x^N)\}} \cdot \delta_{\{\hat{x}^N = f^N(x^{N-1})\}}, \end{aligned} \tag{30}$$

where $e(x^N)$ denotes the code-function chosen by the encoder for a sequence $x^N \in \mathcal{X}^N$.

We now give the general rate-distortion theorem - for arbitrary discrete sources with feed-forward without the assumptions of stationarity or ergodicity. For this, we use the machinery developed in [22] for the standard source coding problem, i.e., without feed-forward. The source distribution is a sequence of distributions denoted by $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$. A conditional distribution is denoted by $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$. We consider a sequence of distortion measures $d_n(x^n, \hat{x}^n)$, and, as before, we assume $d_n(.,.)$ is normalized with respect to n and is uniformly bounded in n.

We give the result for two kinds of distortion criteria. The first is a constraint on the expected distortion. The second criterion is a probability of error criterion- the restriction is on the probability that the distortion is $\geq D$. The probability of error criterion may be more useful for a general source, which may not be ergodic or stationary.

Definition 4.3 (a). (Expected distortion criterion) *R is an ϵ-achievable rate at expected distortion D if for all sufficiently large N, there exists an $(N, 2^{NR})$ source codebook such that*

$$E_{X^N}\left[d_N(x^N, \hat{x}^N)\right] \leq D + \epsilon,$$

where $\hat{x}^N$ denotes the reconstruction of x^N.

R is an achievable rate at expected distortion D if it is ϵ-achievable for every $\epsilon > 0$.

(b) (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 restate the definitions of a few quantities (see [23],[16]) which we will use in our coding theorems. A word about the notation used in the remainder of this paper. We will use the usual notation $P_X(x)$ to indicate the probability mass function of X evaluated at the point x. Often, we will treat the p.m.f of X as a function of the random variable X. In such situations, the function is also random variable and we will use the notation $P(X)$ and $P_X(X)$ interchangeably to refer to this random variable.

Definition 4.4. *The* limsup in probability *of a sequence of random variables $\{X_n\}$ is defined as the smallest extended real number α such that $\forall \epsilon > 0$*

$$\lim_{n \to \infty} Pr[X_n \geq \alpha + \epsilon] = 0.$$

The liminf in probability *of a sequence of random variables* $\{X_n\}$ *is defined as the largest extended real number* β *such that* $\forall\epsilon > 0$

$$\lim_{n\to\infty} Pr[X_n \leq \beta - \epsilon] = 0.$$

Definition 4.5. *For any sequence of joint distributions* $\{P_{X^N,\hat{X}^N}\}_{N=1}^{\infty}$*, define* $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$

$$\begin{aligned}
i(x^N;\hat{x}^N) &\triangleq \log\frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{P_{\hat{X}^N}(\hat{x}^N)P_{X^N}(x^N)} \quad, && (31)\\
\overline{H}(X) &\triangleq \limsup_{inprob} \frac{1}{N}\log\frac{1}{P_{X^N}(X^N)} \quad, && (32)\\
\underline{H}(X) &\triangleq \liminf_{inprob} \frac{1}{N}\log\frac{1}{P_{X^N}(X^N)} \quad, && (33)\\
\vec{i}(\hat{x}^N;x^N) &\triangleq \log\frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)P_{X^N}(x^N)} \quad, && (34)\\
\overline{I}(\hat{X}\to X) &\triangleq \limsup_{inprob} \frac{1}{N}\vec{i}(\hat{X}^N;X^N) \quad, && (35)\\
\underline{I}(\hat{X}\to X) &\triangleq \liminf_{inprob} \frac{1}{N}\vec{i}(\hat{X}^N;X^N) \quad, && (36)
\end{aligned}$$

where $\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)$ and $\vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)$ are given by (8) and (9) respectively.

We also note that the directed information from $\hat{X}^N$ to X^N can be written as

$$I(\hat{X}^N \to X^N) = \sum_{x^N,\hat{x}^N} P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)\vec{i}(\hat{x}^N;x^N). \tag{37}$$

As pointed out in [22], the entropy rate and the mutual information rate, defined by $\lim_{n\to\infty}\frac{1}{n}\log H(X^n)$ and $\lim_{n\to\infty}\frac{1}{n}\log I(X^n;\hat{X}^n)$ respectively, may not exist for an arbitrary random process which may neither be stationary nor ergodic. But the sup-entropy rate, inf-entropy rate ($\overline{H}(X)$ and $\underline{H}(X)$ defined above) always exist, as do the sup-information rate and the inf- information rate ($\overline{I}(X;\hat{X})$ and $\underline{I}(X;\hat{X})$ defined in [20]).

Lemma 4.1. [16] *For any sequence of joint distributions* $\{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$*, we have*

$$\underline{I}(\hat{X}\to X) \leq \liminf_{N\to\infty}\frac{1}{N}I(\hat{X}^N \to X^N) \leq \limsup_{N\to\infty}\frac{1}{N}I(\hat{X}^N \to X^N) \leq \overline{I}(\hat{X}\to X). \tag{38}$$

If

$$\underline{I}(\hat{X}\to X) = \overline{I}(\hat{X}\to X) \tag{39}$$

we say that the process $\{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$ is information stable [24]. We are now ready to state and prove the rate distortion theorem for an arbitrary source with feed-forward. In [23], Verdu and Han showed that the capacity formula for arbitrary channels without feedback is an optimization(sup) of the inf-information rate over all input distributions. Analogously, it was shown in [22] that the rate distortion function (without feed-forward) for an arbitrary source is given by an optimization(inf) of the sup-information rate. Tatikonda and Mitter [16] showed that for arbitrary channels with feedback, the capacity is an optimization of $\underline{I}(X \to Y)$, the inf-directed information rate. Our result is that the rate distortion function for an arbitrary source with feed-forward is an optimization of $\overline{I}(X \to \hat{X})$, the sup-directed information rate.

Theorem 2 (a). (Expected Distortion Constraint) *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 expected distortion D, is given by*

$$R^*_{ff}(D) = \inf_{\mathbf{P_{\hat{X}|X}} : \lambda(\mathbf{P_{\hat{X}|X}}) \leq D} \overline{I}(\hat{X} \to X), \tag{40}$$

where

$$\lambda(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{n \to \infty} E[d_n(X^n, \hat{X}^n)]. \tag{41}$$

(b) (Probability of Error Constraint) *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*

$$R_{ff}(D) = \inf_{\mathbf{P_{\hat{X}|X}} : \rho(\mathbf{P_{\hat{X}|X}}) \leq D} \overline{I}(\hat{X} \to X), \tag{42}$$

where

$$\rho(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) = \inf \left\{ h : \lim_{n \to \infty} P_{X^n} P_{\hat{X}^n|X^n} \left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h \right) = 0 \right\}. \tag{43}$$

Note that if the joint process $\{X_n, \hat{X}_n\}_{n=1}^{\infty}$ is information stable (see (39)), from Lemma 4.1, the rate-distortion function becomes

$$R_{ff}(D) = \inf \lim_{N \to \infty} \frac{1}{N} I(\hat{X}^N \to X^N), \tag{44}$$

where the infimum is evaluated according to the distortion constraint used. The detailed proofs of the direct and converse parts of Theorem 2 are found in Appendix B and C, respectively. The proofs for parts (a) and (b) are very similar. We only give a brief outline here of the direct coding theorem. For the sake of intuition, assume information stability. We want to show the achievability of all rates greater than $R_{ff}(D)$ in (44).

Let $\mathbf{P^*_{\hat{X}|X}} = \{P^*_{\hat{X}^n|X^n}\}$ be the conditional distribution that achieves the infimum (subject to the constraint). Fix the block length N. The source code with source X^N and reconstruction F^N does not contain feed-forward (see Figure 6). Our goal is to construct a joint distribution over $X^N, \hat{X}^N$ and F^N, say $Q_{F^N, X^N, \hat{X}^N}$, such that the marginal over X^N and $\hat{X}^N$ satisfies

$$Q_{X^N, \hat{X}^N} = P_{X^N} P^*_{\hat{X}^N|X^N}. \tag{45}$$

We also impose certain additional constraints on $Q_{F^N, X^N, \hat{X}^N}$ so that [1]

$$I_Q(F^N; X^N) = I_Q(\hat{X}^N \to X^N). \tag{46}$$

[1] For clarity, wherever necessary, we will indicate the distribution used to calculate the information quantity as a subscript of I.

Using (45) in the above equation, we get

$$I_Q(F^N;X^N) = I_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{X}^N \to X^N). \tag{47}$$

Using the usual techniques for source coding without feed-forward, it can be shown that all rates greater than $\frac{1}{N}I_Q(F^N;X^N)$ can be achieved. From (47), it follows that all rates greater than $I_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{X}^N \to X^N)$ are achievable. The bulk of the proof lies in constructing a suitable joint distribution Q.

4.2 Discrete Memoryless Sources

Consider an arbitrary discrete memoryless source (DMS). Such a source is characterized by a sequence of distributions $\{P_{X^n}\}_{n=1}^{\infty}$, where for each n, P_{X^n} is a product distribution.

We prove the following result for a DMS with expected distortion constraint and a memoryless distortion measure $d_N(x^N,\hat{x}^N) = \frac{1}{N}\sum_{i=1}^{N} d_i(x_i,\hat{x}_i)$.

Theorem 3. *Feed-forward does not decrease the rate-distortion function of a discrete memoryless source.*

The proof is found in Appendix D. It should be noted that Theorem 3 may not hold for a general distortion measure $d_N(x^N,\hat{x}^N)$. In other words, even when the source is memoryless, feed-forward could decrease the rate-distortion function when the distortion constraint has memory. The theorem may also not hold when the probability of error distortion constraint (Theorem 2(b)) is used instead of the expected distortion constraint regardless of the nature of the distortion measure $d_N(x^N,\hat{x}^N)$.

4.3 Gaussian sources with feed-forward

In this section, we study the rate-distortion function for the special case of Gaussian sources with feed-forward. A source X is Gaussian if the random process $\{X_n\}_{n=1}^{\infty}$ is jointly Gaussian. A Gaussian source is continuous valued unlike the sources hitherto discussed. However, it is straightforward to extend the results derived earlier for discrete sources to continuous sources. In particular, feed-forward does not decrease the rate-distortion function of a memoryless Gaussian source with expected mean-squared error distortion criterion. Interestingly though, feed-forward in an IID Gaussian source enables us to achieve rates arbitrarily close to the rate-distortion function with a low complexity coding scheme involving just linear processing and uniform scalar quantization (without entropy coding) at all rates [5].

In this section, we consider the commonly used mean-squared error as the distortion measure for Gaussian sources, i.e. $d_N(x^N,\hat{x}^N) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$. As in the case of discrete memoryless sources, we use the expected distortion constraint. We now show that for a Gaussian source, $R^*_{ff}(D)$ is achieved by a jointly Gaussian conditional distribution.

Proposition 4.1. *Let X be a Gaussian source with distribution $\mathbf{P_X}$ and let $\mathbf{P_{\hat{X}|X}}$ be any conditional distribution. Let $G_{X^N,\hat{X}^N} = P_{X^N} \cdot G_{\hat{X}^N|X^N}$ be a jointly Gaussian distribution that has the same second order properties as $P_{X^N,\hat{X}^N} = P_{X^N} \cdot P_{\hat{X}^N|X^N}$. Then:*

1. $I_G(\hat{X}^N \to X^N) \leq I_P(\hat{X}^N \to X^N)$

2. *The average distortion is the same under both distributions, i.e.,*

$$E_P[d_N(X^N, \hat{X}^N)] = E_G[d_N(X^N, \hat{X}^N)]. \tag{48}$$

Proof. 1. We denote the densities corresponding to $P_{X^N,\hat{X}^N}$ and $G_{X^N,\hat{X}^N}$ by

$$p_{X^N,\hat{X}^N} = p_{X^N} p_{\hat{X}^N|X^N}$$

$$g_{X^N,\hat{X}^N} = p_{X^N} g_{\hat{X}^N|X^N}$$

Using the representation of directed information given in (37), we have the following chain of inequalities

$$\begin{aligned}
&I_P(\hat{X}^N \to X^N) - I_G(\hat{X}^N \to X^N) \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&\qquad - \int g_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{g_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{\vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) p_{X^N}(x^N)} dx^N d\hat{x}^N \\
&\qquad - \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{g_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{\vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) p_{X^N}(x^N)} dx^N d\hat{x}^N,
\end{aligned}$$

where the last equality is due to the fact that $p_{X^N,\hat{X}^N}$ and $g_{X^N,\hat{X}^N}$ have the same second order properties. Continuing the chain, we have

$$\begin{aligned}
&I_P(\hat{X}^N \to X^N) - I_G(\hat{X}^N \to X^N) \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \vec{g}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{g_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{\vec{p}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)}{\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{\vec{p}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} dx^N d\hat{x}^N \\
&= \int p_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{p_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{p'_{X^N,\hat{X}^N}(x^N, \hat{x}^N)} dx^N d\hat{x}^N,
\end{aligned}$$

where $p'_{X^N,\hat{X}^N}$ is the joint distribution $\vec{g}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \cdot \vec{p}_{\hat{X}^N|X^N}$. Then last expression is the Kullback-Leibler distance between the distributions p and p' and is thus non-negative.

2. Since $P_{X^N,\hat{X}^N}$ and $G_{X^N,\hat{X}^N}$ have the same second order properties, it follows that the expected distortion is the same under both distributions. □

Thus for Gaussian sources with expected mean-squared error distortion criterion, the optimizing conditional distribution can be taken to be jointly Gaussian.

We also have the following result from [16] for jointly Gaussian distributions. For any jointly Gaussian distribution $\mathbf{P}_{\mathbf{X^N},\hat{\mathbf{X}}^\mathbf{N}} = \{P_{X^N,\hat{X}^N}\}_{n=1}^{\infty}$,

$$\overline{I}(\hat{X} \to X) = \limsup_{N\to\infty} \frac{1}{N} I(\hat{X} \to X). \tag{49}$$

This property follows from the asymptotic equipartition property, which is valid for an arbitrary Gaussian random processes (Theorem 5, [25]). Thus the rate-distortion function for an arbitrary Gaussian source with expected mean-squared error distortion criterion can be written as

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N), \tag{50}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) = \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2] \tag{51}$$

and $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ can be taken to be jointly Gaussian without loss of generality.

5 Error Exponents

In this section we obtain the random coding error exponent for the problem of source coding with feed-forward. The error-exponent for a source code of block-length N for a discrete memoryless source is derived by Blahut[26] and by Marton in [27]. Error exponents for discrete sources have also been studied in [28, 29, 30, 31, 32]. A procedure identical to the proof of Theorem 6.5.1 in [26] yields the error exponent for an arbitrary discrete source (without feed-forward). Therefore, we have the following fact for discrete sources without feed-forward.

Given a source with Nth order distribution P_{X^N}, there exists a $(N, 2^{NR})$ source code (without feed-forward) such that the probability that a source sequence of length N cannot be encoded with distortion $\leq D$ satisfies

$$P_e \leq e^{-NE_N(R,D)+o(N)}, \tag{52}$$

where $E_N(R,D)$ is the random coding error exponent for the source (without feed-forward) and is given by

$$E_N(R,D) = \max_{s\geq 0}\min_{t\leq 0}\max_{q_{\hat{X}^N}} \left[sR - stD - \frac{1}{N}\log_2 \sum_{x^N} P_{X^N}(x^N) \left(\sum_{\hat{x}^N} q_{\hat{X}^N}(\hat{x}^N) e^{tNd_N(x^N,\hat{x}^N)} \right)^{-s} \right], \tag{53}$$

and for large enough N, $o(N) = 0$. We state the result in the following theorem, whose proof is found in Appendix E.

Theorem 4. *Given a source with Nth order distribution P_{X^N}, there exists a $(N, 2^{NR})$ source code with feed-forward so that the probability that a source sequence of length N cannot be encoded with distortion $\leq D$ satisfies*

$$P_e \leq e^{-NE_{ff-N}(R,D)+o(N)}, \tag{54}$$

where $E_{ff-N}(R,D)$ is the random coding error exponent for the source with feed-forward and is given by

$$E_{ff-N}(R,D) = \max_{s\geq 0} \min_{t\leq 0} \max_{\vec{q}_{\hat{X}^N|X^N}} \left[sR - stD - \frac{1}{N} \log_2 \sum_{x^N} P_{X^N}(x^N) \left(\sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{tNd_N(x^N,\hat{x}^N)} \right)^{-s} \right], \quad (55)$$

where $\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) = \prod_{i=1}^N q_{\hat{X}_i|X^{i-1},\hat{X}^{i-1}}(\hat{x}_i|x^{i-1},\hat{x}^{i-1})$ and for large enough N, $o(N) = 0$.

We now compare the error exponents for a source with and without feed-forward given by (55) and (53), respectively. Denote the space of all distributions of the form $q_{\hat{X}^N}$ by $\mathcal{S}_q$ and the space of all distributions of the form $\vec{q}_{\hat{X}^N|X^N}$ by $\mathcal{S}_{\vec{q}}$. The only difference between the expressions for the error exponents with and without feed-forward is that the former involves a maximization over distributions in $\mathcal{S}_q$, while the latter involves a maximization over $\mathcal{S}_{\vec{q}}$.

Now, every distribution $q_{\hat{X}^N} = \prod_{i=1}^N q_{\hat{X}_i|\hat{X}^{i-1}}$ belongs to the space of distributions of the form $\vec{q}_{\hat{X}^N|X^N} = \prod_{i=1}^N q_{\hat{X}_i|\hat{X}^{i-1},X^{i-1}}$. Therefore,

$$\mathcal{S}_q \subset \mathcal{S}_{\vec{q}}.$$

Thus in the no feed-forward case, we are maximizing over a subset of the distributions available to us in the feed-forward case. Equivalently, we have proved the following theorem.

Theorem 5. *For any source X, the error exponent with feed-forward is at least as large as the error exponent without feed-forward.*

Equation (54) guarantees an exponentially small probability of error only when $E_{ff-N}(R,D)$ is positive. An alternate definition of the error exponent is better suited to determine the values of R for which $E_{ff-N}(R,D)$ is positive. We first have the following definition.

Definition 5.1.

$$B_N(\hat{p}_{X^N}, D) \triangleq \min_{q_{\hat{X}^N|X^N}: E_{\hat{p}q}[d_N(x^N,\hat{x}^N)\leq D]} \frac{1}{N} I_{\hat{p}_{X^N} q_{\hat{X}^N|X^N}}(\hat{X}^N \rightarrow X^N), \quad (56)$$

where the subscripts denote the joint distribution used to calculate the directed information.

Theorem 6. *An equivalent representation of $E_{ff-N}(R,D)$ is*

$$E_{ff-N}(R,D) = \min_{\hat{p}_{X^N} \in \mathcal{P}(R,D)} \frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)}, \quad (57)$$

where

$$\mathcal{P}(R,D) = \{\hat{p}_{X^N} : B_N(\hat{p}_{X^N}, D) \geq R\}. \quad (58)$$

The proof of the above theorem is found in Appendix F. The quantity on the right hand side of (57) is a discrimination. It is 0 iff the source distribution $P_{X^N} \in \mathcal{P}$ and positive otherwise. From the definition of $\mathcal{P}$, it follows that $P_{X^N} \in \mathcal{P}$ if $R \leq B_N(P_{X^N}, D)$. Therefore, $E_{ff-N}(R, D)$ is strictly positive for rates R such that

$$R > B_N(P_{X^N}, D). \tag{59}$$

Using the representation of the error exponent in Theorem 6, it is easy to see that the error exponent for a discrete memoryless source does not change with feed-forward. This result was obtained in [4].

6 Feed-Forward with Arbitrary Delay

Recall from the discussion in Section 1 that our problem of source coding with noiseless feed-forward is meaningful for any delay larger than the block length N. Our results in the preceding sections assumed that the delay was $N+1$, i.e., to reconstruct the ith sample the decoder had perfect knowledge of first $i-1$ samples.

We now extend our results for a general delay $N+k$, where N is the block length. 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 ith sample, it has access to all the past $(i-k)$ samples of the source. In other words, the decoder is a sequence of mappings $g_i : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{i-k} \to \widehat{\mathcal{X}}, \quad i = 1, \ldots, N$.

The key to understanding feed-forward with arbitrary delay is the interpretation of directed information in Section 2.2. Recall from (3) that the directed information can be expressed as

$$I(\hat{X}^N \to X^N) = I(\hat{X}^N; X^N) - \sum_{i=2}^{N} I(X^{i-1}; \hat{X}_i | \hat{X}^{i-1}). \tag{60}$$

When the feed-forward delay is $N+k$, the decoder knows X^{i-k} to reconstruct $\hat{X}_i$. Here, we need not spend $I(X^{i-k}; \hat{X}_i | \hat{X}^{i-1})$ bits to code this information, hence this rate comes for free. In other words, the performance limit on this problem is given by the minimum of

$$\begin{aligned} I_k(\hat{X}^N \to X^N) &\triangleq I(\hat{X}^N; X^N) - \sum_{i=k+1}^{N} I(X^{i-k}; \hat{X}_i | \hat{X}^{i-1}) && (61) \\ &= I(\hat{X}^N; X^N) - I(0^k X^{N-k} \to \hat{X}^N), && (62) \end{aligned}$$

where $0^k X^{N-k}$ is the $N-$length sequence $[\_, \_, \ldots \_, X_1, X_2, \ldots, X_{N-k}]$.

Observing (61), we make the following comment. In any source coding problem, the mutual information $I(\hat{X}^N, X^N)$ is the fundamental quantity to characterize the rate-distortion function. With feed-forward, the rate-distortion function is reduced by a quantity equal to the information we get for free because of the feed-forward. One can use very similar arguments to characterize the capacity of channels with feedback delay $k \geq 1$.

We now state our two main theorems- the rate-distortion theorem and the random coding error exponent- for feed-forward with general delay. We omit the proofs since they are similar to the ones in the preceding sections.

Definition 6.1.

$$\begin{aligned}
\vec{P}_k(\hat{X}^N|X^N) &\triangleq \prod_{i=1}^{N} P(\hat{X}_i|\hat{X}^{i-1}, X^{i-k}), && (63)\\
I_k(\hat{X}^N \rightarrow X^N) &\triangleq I(\hat{X}^N; X^N) - \sum_{i=k+1}^{N} I(X^{i-k}; \hat{X}_i|\hat{X}^{i-1}) && (64)\\
&= \sum_{x^n, \hat{x}^N} P_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N)\vec{P}_k(\hat{X}^N|X^N)}, &&\\
\overline{I}_k(\hat{X} \rightarrow X) &\triangleq \limsup_{inprob} \frac{1}{N} \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N)\vec{P}_k(\hat{X}^N|X^N)}. && (65)
\end{aligned}$$

Theorem 7 (Rate-Distortion Theorem).

(a) (Expected Distortion Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with $N+k$ delayed feed-forward, the infimum of all achievable rates at expected distortion D, is given by*

$$R^{*}_{ff}(D) = \inf_{\mathbf{P_{\hat{X}|X}}:\lambda(\mathbf{P_{\hat{X}|X}}) \leq D} \overline{I}_k(\hat{X} \rightarrow X), \qquad (66)$$

where

$$\lambda(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{n\rightarrow\infty} E[d_n(X^n, \hat{X}^n)]. \qquad (67)$$

(b) (Probability of Error Constraint) *For an arbitrary source X characterized by a distribution $\mathbf{P_X}$, the rate-distortion function with $N+k$ delayed feed-forward, the infimum of all achievable rates at probability-*1 *distortion D, is given by*

$$R_{ff}(D) = \inf_{\mathbf{P_{\hat{X}|X}}:\rho(\mathbf{P_{\hat{X}|X}}) \leq D} \overline{I}_k(\hat{X} \rightarrow X), \qquad (68)$$

where

$$\rho(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) = \inf\left\{h : \lim_{n\rightarrow\infty} P_{X^n}P_{\hat{X}^n|X^n}\left((x^n, \hat{x}^n) : d_n(x^n, \hat{x}^n) > h\right) = 0\right\}. \qquad (69)$$

Theorem 8 (Error Exponent). *Given a source with N-th order distribution P_{X^N}, there exists a $(N, 2^{NR})$ source code with $N+k$ delayed feed-forward so that the probability that a source sequence of length N cannot be encoded with distortion $\leq D$ satisfies*

$$P_e \leq e^{-NE_{ff-N}(R,D)+o(N)}, \qquad (70)$$

where $E_{ff-N}(R, D)$ is the error exponent for the source with feed-forward and is given by

$$E_{ff-N}(R,D) = \max_{s\geq 0} \min_{t\leq 0} \max_{\vec{q}_{k\hat{X}^N|X^N}} \left[sR - stD - \frac{1}{N}\log_2 \sum_{x^N} P_{X^N}(x^N)\left(\sum_{\hat{x}^N} \vec{q}_{k\hat{X}^N|X^N}(\hat{x}^N|x^N)e^{tNd_N(x^N,\hat{x}^N)}\right)^{-s}\right], \qquad (71)$$

where $\vec{q}_{k\hat{X}^N|X^N}(\hat{x}^N|x^N) = \prod_{i=1}^{N} q_{\hat{X}_i|X^{i-k},\hat{X}^{i-1}}(\hat{x}_i|x^{i-k}, \hat{x}^{i-1})$ and for large enough N, $o(N) = 0$.

7 Conclusions

In this work, we have defined and analyzed a source coding model with feed-forward. This is a source coding system in which the decoder has knowledge of all previous source samples while reconstructing the present sample. We have shown that feed-forward does not decrease the rate-distortion function of a discrete memoryless source. For an arbitrary source with feed-forward, we have derived the rate-distortion function and a bound on the worst-case performance of a source code with finite block length. We proved that the random coding error exponent for a source with feed-forward is at least as large as the exponent for the same source without feed-forward and found the range of rates for which the error exponent is strictly positive. We then extended our results to the feed-forward model with an arbitrary delay larger than the block length. The problem of source coding with feed-forward can be considered the dual of channel coding with feedback. Extensions to accommodate practical constraints such as a noisy feed-forward path are part of future work.

Acknowledgments: The authors wish to thank Sekhar Tatikonda of Yale University for inspiring discussions and comments.

APPENDIX

A Proof of Lemma 3.1(AEP)

The proof is similar to that of the Shannon-McMillan Breiman theorem in [18, 33]. We first state the definitions and three lemmas required for the proof. Recall that

$$\begin{aligned}
\hat{x}^N &= (\hat{x}_1, \hat{x}_2 \ldots, \hat{x}_N), \\
\vec{P}(\hat{x}^N|x^N) &= \prod_{i=1}^{N} P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}), \\
\vec{P}(x^N|\hat{x}^N) &= \prod_{i=1}^{N} P(x_i|\hat{x}^{i}, x^{i-1}).
\end{aligned}$$

We want to show that

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) \to H(\hat{X}\|0X) \triangleq \lim_{N\to\infty} H(\hat{X}_N|X^{N-1}, \hat{X}^{N-1}) \tag{72}$$

Definition A.1. *Let*

$$\begin{aligned}
H^{\infty}(\hat{X}\|0X) &= E\left[-\log P(\hat{X}_0|\hat{X}_{-1}, \hat{X}_{-2}, \ldots, X_{-1}, X_{-2}, \ldots)\right], \\
H^k &= E\left[-\log P(\hat{X}_0|\hat{X}^{-1}_{-k}, X^{-1}_{-k})\right], \\
\vec{P}^k(\hat{X}^N|X^N) &= \vec{P}(\hat{X}^k|X^k)\prod_{i=k+1}^{N} P(\hat{X}_i|\hat{X}^{i-1}_{i-k}, X^{i-1}_{i-k}), \\
\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) &= \prod_{i=1}^{N} P(\hat{X}_i|\hat{X}^{i-1}_{-\infty}, X^{i-1}_{-\infty}).
\end{aligned}$$

Lemma A.1.

$$\begin{aligned}
-\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N) &\to H^k, \\
-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) &\to H^{\infty}(\hat{X}\|0X).
\end{aligned}$$

Proof.

$$\begin{aligned}
-\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N) &= -\frac{1}{N}\vec{P}(\hat{X}^k|X^k) - \frac{1}{N}\sum_{i=k+1}^{N}\log P(\hat{X}_i|\hat{X}^{i-1}_{i-k}, X^{i-1}_{i-k}) \\
&\to 0 + H^k \quad \text{by the ergodic theorem.}
\end{aligned} \tag{73}$$

$$\begin{aligned}
-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty}) &= -\frac{1}{N}\sum_{i=1}^{N}\log P(\hat{X}_i|, \hat{X}^{i-1}_{-\infty}, X^{i-1}_{-\infty}) \\
&\to H^{\infty}(\hat{X}\|0X) \quad \text{by the ergodic theorem.}
\end{aligned} \tag{74}$$

□

Lemma A.2.

$$H^k \to H^{\infty}(\hat{X}\|0X), \quad H(\hat{X}\|0X) = H^{\infty}(\hat{X}\|0X).$$

Proof. We know that $H^k \to H(\hat{X}||0X)$, since the joint process is stationary and $\{H_k\}$ is a non-increasing sequence of non-negative numbers. So we only need to show that $H^k \to H^\infty(\hat{X}||0X)$. The Martingale convergence theorem says that

$$P(\hat{x}_0|\hat{X}_{-k}^{-1}, X_{-k}^{-1}) \to P(\hat{x}_0|\hat{X}_{-\infty}^{-1}, X_{-\infty}^{-1}).$$

Since $\hat{\mathcal{X}}$ is a finite alphabet and $p \log p$ is bounded, by the dominated convergence theorem,

$$\begin{aligned} \lim_{k\to\infty} H^k &= \lim_{k\to\infty} E\left[-\sum_{\hat{x}_0\in\hat{\mathcal{X}}} P(\hat{x}_0|\hat{X}_{-k}^{-1}, X_{-k}^{-1}) \log P(\hat{x}_0|\hat{X}_{-k}^{-1}, X_{-k}^{-1})\right] \\ &= E\left[-\sum_{\hat{x}_0\in\hat{\mathcal{X}}} P(\hat{x}_0|\hat{X}_{-\infty}^{-1}, X_{-\infty}^{-1}) \log P(\hat{x}_0|\hat{X}_{-\infty}^{-1}, X_{-\infty}^{-1})\right] \\ &= H^\infty(\hat{X}||0X). \end{aligned}$$

Thus $H^k \to H^\infty(\hat{X}||0X)$. □

Lemma A.3.

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} &\leq 0, \\ \limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0)} &\leq 0, \end{aligned}$$

where

$$\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0) \triangleq \prod_{i=1}^N P(\hat{X}_i|X_{-\infty}^{i-1}, \hat{X}_{-\infty}^{i-1}). \tag{75}$$

Proof.

$$\begin{aligned} E\left[\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)}\right] &= \sum_{\hat{x}^N, x^N} P(\hat{x}^N, x^N) \frac{\prod_{i=1}^k P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) \cdot \prod_{i=k+1}^N P(\hat{x}_i|\hat{x}_{i-k}^{i-1}, x_{i-k}^{i-1})}{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1})} \\ &= \sum_{\hat{x}^N, x^N} P(\hat{x}^k, x^k) \cdot \prod_{i=k+1}^N P(x_i|x^{i-1}, \hat{x}^i) P(\hat{x}_i|\hat{x}_{i-k}^{i-1}, x_{i-k}^{i-1}) \\ &= 1, \end{aligned} \tag{76}$$

where the last equality follows by evaluating the sum first over x_N, then over $\hat{x}_N$, then over x_{N-1} and so on. Using the above in Markov's inequality, we have

$$\Pr\left\{\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq N^2\right\} \leq \frac{1}{N^2} \tag{77}$$

or

$$\Pr\left\{\frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq \frac{1}{N} \log N^2\right\} \leq \frac{1}{N^2}. \tag{78}$$

Since $\sum_{N=1}^\infty \frac{1}{N^2} < \infty$, the Borel-Cantelli lemma says that, with probability 1, the event

$$\left\{\frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq \frac{1}{N} \log N^2\right\}$$

occurs only for finitely many N. Thus

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \leq 0 \quad \text{with probability 1.}$$

The second part of the lemma is proved in a similar manner. Using conditional expectations, we can write

$$E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})}\right] = E_{\hat{X}^0_{-\infty}, X^0_{-\infty}}\left[E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})}\middle|\hat{X}^0_{-\infty}, X^0_{-\infty}\right]\right]. \tag{79}$$

The inner expectation can be written as

$$\begin{aligned} E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})}\middle|\hat{X}^0_{-\infty}, X^0_{-\infty}\right] &= \sum_{\hat{x}^N, x^N} P(\hat{x}^N, x^N|\hat{X}^0_{-\infty}, X^0_{-\infty}) \frac{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1})}{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}, \hat{X}^0_{-\infty}, X^0_{-\infty})} \\ &= \sum_{\hat{x}^N, x^N} \prod_{i=1}^N P(x_i|\hat{x}^i, x^{i-1}, \hat{X}^0_{-\infty}, X^0_{-\infty}) P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) \\ &= 1, \end{aligned} \tag{80}$$

where the last equality is obtained by evaluating the sum first over x_N, then over $\hat{x}_N$, then over x_{N-1} and so on. Using the Borel-Cantelli lemma as in the previous part, we obtain

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})} \leq 0.$$

□

Proof of Lemma 3.1- AEP. We will show that the sequence of random variables $-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N)$ is sandwiched between the upper bound H^k and the lower bound $H^\infty(\hat{X}||0X)$ for all $k \geq 0$. From Lemma A.3, we have

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \leq 0. \tag{81}$$

Since the limit $\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N)$ exists (Lemma A.1), we can write (81) as

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \leq \lim_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}^k(\hat{X}^N|X^N)} = H^k. \tag{82}$$

The second part of Lemma A.3 can be written as

$$\liminf_{N\to\infty} \frac{1}{N} \log \frac{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})}{\vec{P}(\hat{X}^N|X^N)} \geq 0. \tag{83}$$

Since the limit $\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})$ exists (Lemma A.1), we can rewrite (83) as

$$\liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \geq \lim_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}(\hat{X}^N|X^N_{-\infty}, \hat{X}^0_{-\infty})} = H^\infty(\hat{X}||0X) \tag{84}$$

Combining (82) and (84), we have

$$H^\infty(\hat{X}||0X) \leq \liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \leq \limsup_{N\to\infty} \frac{1}{N} \log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \leq H^k \quad \text{for all } k. \tag{85}$$

By Lemma A.2, $H^k \to H(\hat{X}||0X) = H^\infty(\hat{X}||0X)$. Thus

$$\lim_{N\to\infty} -\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) = H(\hat{X}||0X). \tag{86}$$

□

B Proof of Direct Part of Theorem 2

The approach we will take is as follows. We build a source code for the $X - F$ block in Figure 6 (Section 4), a system without feed-forward. Here, the code-functions themselves are considered 'reconstructions' of the source sequences. We will then connect the $X - F$ and the $X - \hat{X}$ systems to prove the achievability of $R_{ff}(D)$.

For the sake of clarity, we present the proof in two parts. The first part establishes the background for making the connection between the $X - F$ and $X - \hat{X}$ systems. This part is common to parts (a) and (b) of Theorem 2. In the second part, we will construct random codes for the system without feed-forward and show the achievability of $R_{ff}(D)$ using the results of the first part. We describe the second part in detail for the probability of error criterion. The proof for the expected distortion case is omitted since it is similar.

Part I

Let $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^*_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be the sequence of distributions that achieves the infimum in Theorem 2. We want to construct a joint distribution over $X^N, \hat{X}^N$ and F^N, say $Q^*_{F^N,X^N,\hat{X}^N}$, such that the marginal over X^N and $\hat{X}^N$ satisfies

$$Q^*_{X^N,\hat{X}^N} = P_{X^N}P^*_{\hat{X}^N|X^N}. \tag{87}$$

For any N, the joint distribution $P_{X^N}P^*_{\hat{X}^N|X^N}$ can be split, as in (10), as

$$P_{X^N}P^*_{\hat{X}^N|X^N} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|\hat{X}^n,X^{n-1}}, \tag{88}$$

where the marginals, given by P^{ch} and P^{dec}, can be considered the fictitious test-channel from $\hat{X}$ to X and the set of 'decoder' distributions to this channel, respectively.

Let P_{F^N} be any distribution on the space of code-functions. We now define a joint distribution over $Q_{X^N,F^N,\hat{X}^N}$ over $(\mathcal{X}^N, \mathcal{F}^N, \hat{\mathcal{X}}^N)$, imposing the following constraints. Q is said to be nice with respect to P_{F^N} and $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^N$ if $\forall x^N \in \mathcal{X}^N, f^N \in \mathcal{F}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$, the following hold:

1. For $n = 1, \ldots, N$,

$$Q_{\hat{X}_n|F_n,X^{n-1}}(\hat{x}_n|f_n, x^{n-1}) = \begin{cases} 1, & \text{if } \hat{x}_n = f_n(x^{n-1}) \\ 0, & \text{otherwise.} \end{cases} \tag{89}$$

2.

$$Q_{F_n|F^{n-1},X^{n-1}}(f_n|f^{n-1}, x^{n-1}) = P_{F_n|F^{n-1}}(f_n|f^{n-1}) \qquad n = 1, \ldots, N. \tag{90}$$

3. For $\hat{x}^n = f^n(x^{n-1})$,

$$Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n, \hat{x}^n, x^{n-1}) = P^{ch}_{X_n|\hat{X}^n,X^{n-1}}(x_n|\hat{x}^n, x^{n-1}), \quad n = 1, \ldots, N. \tag{91}$$

It is important to note that for a given problem of source coding with feed-forward, the joint distribution on $X^N, F^N, \hat{X}^N$ induced from an arbitrary pair of encoder and decoder does not satisfy these conditions in general. We just want to construct an arbitrary joint distribution Q over the variables of interest satisfying the above conditions for the direct coding theorem. Given a code-function distribution P_{F^N} and the test channel $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^{\infty}$, there exists a unique joint distribution $Q_{F^N,X^N,\hat{X}^N}$ that is nice with respect to them. This follows from the following arguments.

$$\begin{aligned} Q_{F^N,X^N,\hat{X}^N} &= \left\{ \prod_{n=1}^{N} Q_{X_n|F^n,X^{n-1}} \cdot Q_{F_n|F^{n-1},X^{n-1}} \right\} \cdot Q_{\hat{X}^N|F^N,X^N} \\ &= \left\{ \prod_{n=1}^{N} Q_{X_n|F^n,X^{n-1}} \cdot P_{F_n|F^{n-1}} \right\} \cdot \delta_{\hat{X}^N = F^N(X^{N-1})} \quad , \end{aligned} \tag{92}$$

where we have used (89) and (90) to obtain the second equality. Now we can use the fact that $\hat{x}_n = f_n(x^{n-1})$ to write

$$\begin{aligned} Q_{X_n|F^n,X^{n-1}}(x_n|f^n, x^{n-1}) &= Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n, \hat{x}^n, x^{n-1}) \\ &= P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\left(x_n|x^{n-1}, f^n(x^{n-1})\right), \end{aligned} \tag{93}$$

where we have used (91) for the second equality. Thus the unique nice joint distribution is given by

$$Q_{F^N,X^N,\hat{X}^N}(f^N, x^N, \hat{x}^N) = \prod_{n=1}^{N} P_{F_n|F^{n-1}}(f_n|f^{n-1}) \cdot \prod_{n=1}^{N} P^{ch}_{X_n|X^{n-1},\hat{X}^{n-1}}(x_n|f^n(x^{n-1}), x^{n-1}) \cdot \delta_{\{\hat{x}^N = f^N(x^{N-1})\}}. \tag{94}$$

Keeping P^{ch} fixed, (94) says that choosing a P_{F^N} determines a unique nice distribution. We want to choose P_{F^N} such that the resulting nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$ satisfies

$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}, \tag{95}$$

so that (87) is satisfied.

Definition B.1. *For a test-channel* $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^N$*, we call a code-function distribution* P_{F^N} *'good' with respect to a decoder distribution* $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^N$ *if the following holds for the nice induced distribution* $Q_{F^N,X^N,\hat{X}^N}$*:*

$$\prod_{n=1}^N Q_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}) = \prod_{n=1}^N P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}), \qquad \forall x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \hat{\mathcal{X}}^N. \tag{96}$$

This definition of 'good' is equivalent to, but slightly different from that in [16]. The next Lemma says that it is possible to find such a good P_{F^N}. For the sake of clarity, we give the proof although it is found in [16] in a different flavor.

Lemma B.1. *There exists a code-function distribution* P_{F^N} *good with respect to a decoder distribution* $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^{\infty}$*.*

Proof. Define $\forall n \in \{1,\cdots,N\}$

$$\begin{aligned}
\text{graph}(f_n) &\triangleq \{(x^{n-1},\hat{x}_n) : f_n(x^{n-1}) = \hat{x}_n\} \subset \mathcal{X}^{n-1} \times \hat{\mathcal{X}} && (97)\\
\Gamma_n(x^{n-1},\hat{x}_n) &\triangleq \left\{f_n : f_n(x^{n-1}) = \hat{x}_n\right\}, && (98)\\
\Gamma^n(x^{n-1},\hat{x}^n) &\triangleq \left\{f^n : f_i(x^{i-1}) = \hat{x}_i, \quad i = 1,\ldots,n\right\}. && (99)
\end{aligned}$$

Now for all f^N, define for $n = 1,\ldots,N$

$$P_{F_n|F^{n-1}}(f_n|f^{n-1}) \triangleq \prod_{(b^{n-1},a_n)\in\text{graph}(f_n)} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(a_n|f_1,\ldots,f_{n-1}(b^{n-2}),b^{n-1}). \tag{100}$$

We will show that $P_{F^N} = \prod_{n=1}^N P_{F_n|F^{n-1}}$ is good with respect to $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^{\infty}$. We first need to show that for all n, $P_{F_n|F^{n-1}}$ defined above is a valid probability distribution. This can be shown using arguments similar to those in Part B of this proof. We give the proof in two parts. In part A, we obtain an expression for the induced decoder distribution given P_{F^N} and P^{ch}. Part B uses this expression to show that (100) defines a good code-function distribution.

Part A

Given the test-channel $\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\}_{n=1}^N$ and a code function distribution P_{F^N}, a unique nice distribution $Q_{F^N,X^N,\hat{X}^N}$ is determined. We now show that the induced decoder distribution is given by

$$Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}) = P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1},\hat{x}_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1})\right), \qquad n = 1,\ldots,N. \tag{101}$$

This is Lemma 5.2 in [16], but we repeat the proof here for the sake of completeness.

Note that $(\hat{x}^{n-1},x^{n-1})$ uniquely determines $(\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1}),x^{n-1})$ and vice versa. Therefore,

$$Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}) = Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1}),x^{n-1}). \tag{102}$$

Now $(x^{n-1}, \hat{x}_n)$ uniquely determines $(\Gamma_n(x^{n-1}, \hat{x}_n), x^{n-1})$ and vice versa. Thus we must have

$$Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}) = Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}). \quad (103)$$

Since Q is nice, it satisfies (90). Hence

$$Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1}, x_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}) = P_{F_N|F^{N-1}}(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})). \quad (104)$$

Combining (102), (103) and (104), we obtain the expression in (101).

Part B

We now show that P_{F^N} defined by (100) is good with respect to P^{ch}. For a pair $x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \mathcal{X}^N$, consider

$$\begin{aligned}\sum_{f^N: f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N) &= P_{F^N}(\Gamma^N(x^{N-1}, \hat{x}^N)) \\ &= P_{F^N}\left(\Gamma_1(\hat{x}_1), \ldots, \Gamma_n(x^{n-1}, \hat{x}_n), \ldots, \Gamma_N(x^{N-1}, \hat{x}_N)\right) \\ &= \prod_{n=1}^{N} P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})\right).\end{aligned} \quad (105)$$

Substituting (101) in the above equation, we get

$$\sum_{f^N: f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N) = \prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}). \quad (106)$$

We can also write

$$\begin{aligned}\sum_{\substack{f^N: \\ f^N(x^{N-1})=\hat{x}^N}} P_{F^N}(f^N) &= \sum_{f_1: f_1=\hat{x}_1} \cdots \sum_{\substack{f_n: \\ f_n(x^{n-1})=\hat{x}_n}} \cdots \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{n=1}^{N} P_{F_n|F^{n-1}}(f_n|f^{n-1}) \\ &= \sum_{f_1: f_1=\hat{x}_1} P_{F_1}(f_1) \ldots \sum_{\substack{f_n: \\ f_n(x^{n-1})=\hat{x}_n}} P_{F_n|F^{n-1}}(f_n|f^{n-1}) \ldots \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N|f^{N-1}).\end{aligned} \quad (107)$$

We evaluate the Nth inner summation in the above equation as

$$\begin{aligned}\sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N|f^{N-1}) &\overset{(a)}{=} \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{(b^{N-1},a_N)\in \mathrm{gr}(f_N)} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1, \ldots, f_{N-1}(b^{N-2}), b^{N-1}) \\ &= P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1}, x^{N-1}) \\ &\quad \cdot \sum_{\substack{f_N: \\ f_N(x^{N-1})=\hat{x}_N}} \prod_{\substack{(b^{N-1},a_N)\in \mathrm{gr}(f_N) \\ b^{N-1}\neq x^{N-1}}} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1, \ldots, f_{N-1}(b^{N-2}), b^{N-1}) \\ &\overset{(b)}{=} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1}, x^{N-1}) \\ &\quad \cdot \prod_{b^{N-1}\neq x^{N-1}} \sum_{a_N} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1, \ldots, f_{N-1}(b^{N-2}), b^{N-1}) \\ &= P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1}, x^{N-1}),\end{aligned} \quad (108)$$

where $f_1, \ldots, f_{N-1}$ are specified by the $N-1$ outer summations and 'gr' has been used as shorthand for graph. (a) follows from (100) and (b) follows from an observation similar to

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}} xyz = \sum_{x \in \mathcal{X}} x \cdot \sum_{y \in \mathcal{Y}} y \cdot \sum_{z \in \mathcal{Z}} z.$$

Now, the $(N-1)$th inner sum in (107) can be shown to be equal to $P^{dec}_{\hat{X}_{N-1}|\hat{X}^{N-2},X^{N-2}}(\hat{x}_{N-1}|\hat{x}^{N-2}, x^{N-2})$ in a similar fashion. Thus we can compute the summations in (107) sequentially from $n = N$ down to $n = 1$. Substituting in (107), we get

$$\sum_{\substack{f^N: \\ f^N(x^{N-1}) = \hat{x}^N}} P_{F^N}(f^N) = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1}, x^{n-1}). \tag{109}$$

From (106) and (109), we have

$$\prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1}, x^{n-1}) \qquad n = 1, \ldots, N. \tag{110}$$

completing the proof of the lemma. □

To summarize, we have the following:

- The code-function distribution $P^*_{F^N}$ and the test-channel $\left\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\right\}_{n=1}^{N}$ determine a unique nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$ given by (94).
- We can find a code function distribution $P^*_{F^N}$ to be good with respect to P^{dec}, i.e., the set of induced 'decoder' distributions of Q^* satisfying the relation

$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}, \qquad n = 1, \ldots, N. \tag{111}$$

Hence we have

$$\begin{aligned} Q^*_{X^N,\hat{X}^N} &= \prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot Q^*_{X_n|X^{n-1},\hat{X}^n} \\ &= \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|X^{n-1},\hat{X}^n} \\ &= P_{X^N} \cdot P^*_{\hat{X}^N|X^N}. \end{aligned} \tag{112}$$

Equation (112) is the key to connect the $X - F$ source code without feed-forward to the $X - \hat{X}$ code with feed-forward. We are now ready to prove Theorem 2.

Part II (The probability of error constraint)

For any N, pick $M = 2^{NR}$ N-length code-functions independently according to $P^*_{F^N}$. Denote this $(N, 2^{NR})$ codebook by $\mathcal{C}_N$. Define the "distortion" $d'_N(x^N, f^N) = d\left(x^N, f^N(x^{N-1})\right)$. Let

$$A(\mathcal{C}_N) = \{x^N \in \mathcal{X}^N : \exists f^N \in \mathcal{C}_N \quad with \quad d'_N(x^N, f^N) \le D + \delta\} \tag{113}$$

The set $A^c(\mathcal{C}_N)$ represents the set of x^N's that are not well represented by the chosen codebook. We will show that $P_{X^N}(A^c(\mathcal{C}_N))$, averaged over all realizations of $\mathcal{C}_N$, goes to 0 as $N \to \infty$ as long as $R > R_{ff}(D)$. Indeed,

$$\begin{aligned} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &= \sum_{\mathcal{C}_N} P^*_{F^N}(\mathcal{C}_N) \sum_{x^N \notin A(\mathcal{C}_N)} P_{X^N}(x^N) \\ &= \sum_{x^N} P_{X^N}(x^N) \sum_{\mathcal{C}_N : x^N \notin A(\mathcal{C}_N)} P^*_{F^N}(\mathcal{C}_N). \end{aligned} \tag{114}$$

The last sum on the right-hand side of (114) is the probability of choosing a codebook that does not represent the particular x^N with a distortion $D + \delta$. Define the set

$$B_{N,\delta} = \left\{(x^N, f^N) : \quad d'_N(x^N, f^N) < \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta, \quad \frac{1}{N} i_{Q^*}(x^N; f^N) < \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X} \to X) + \delta\right\}, \tag{115}$$

where $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is as in Theorem 2, and $\overline{I}(\hat{X} \to X)$ is computed with the distribution $\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ and is therefore equal to $R_{ff}(D)$. Define an indicator function

$$K(x^N, f^N) = \begin{cases} 1, & \text{if } (x^N, f^N) \in B_{N,\delta} \\ 0, & \text{otherwise.} \end{cases} \tag{116}$$

We will also need the following Lemma, whose proof is given in Appendix B.1.

Lemma B.2 (a). $Q^*_{F^N|X^N}(f^N|x^N) \le P^*_{F^N}(f^N) \exp_2[N(R_{ff}(D) + \delta)], \qquad \forall (x^N, f^N) \in B_{N,\delta}$.
(b)$Q^*_{X^N,F^N}(B_{N,\delta}) \to 1$ *as* $N \to \infty$.

Since $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ achieves $R_{ff}(D)$ we have $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) \le D$. Hence, for any f^N that does not represent a given x^N with distortion $\le D + \delta$, the pair (x^N, f^N) does not belong to $B_{N,\delta}$. The probability that a code function chosen randomly according to $P^*_{F^N}$ does not represent a given x^N with distortion within $D + \delta$ is

$$\begin{aligned} P^*_{F^N}\left(d'_N(x^N, F^N) \ge D + \delta\right) &\le P^*_{F^N}\left(K(x^N, F^N) = 0\right) \\ &= 1 - \sum_{f^N} P^*_{F^N}(f^n) K(x^N, f^N). \end{aligned} \tag{117}$$

Thus, the probability that none of 2^{NR} code functions, each independently chosen according to $P^*_{F^N}$, represent a given x^N with distortion $D + \delta$ is upper bounded by

$$\left(1 - \sum_{f^N} P^*_{F^N}(f^N) K(x^N, f^N)\right)^{2^{NR}} .$$

This implies

$$\begin{aligned} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &\leq \sum_{x^N} P_{X^N}(x^N)\left(1-\sum_{f^N} P^*_{F^N}(f^n)K(x^N,f^N)\right)^{2^{NR}} \\ &\leq \sum_{x^N} P_{X^N}(x^N)\left(1-\exp_2\{-N(R_{ff}(D)+\delta)\}\sum_{f^N} Q^*_{F^N|X^N}(f^N|x^N)K(x^N,f^N)\right)^{2^{NR}} \quad (118) \end{aligned}$$

where the last inequality follows from Lemma B.2(a). Using the inequality

$$(1-xy)^N \leq 1-x+2^{-yN} \qquad \text{for } 0 \leq x, y \leq 1,$$

in (118), we get

$$\begin{aligned} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &\leq 1+\exp_2\left[-\exp_2[N(R-R_{ff}(D)-\delta)]\right] \\ &\quad -\sum_{x^N,f^N} P_{X^N}(x^N)Q^*_{F^N|X^N}(f^N|x^N)K(x^N,f^N) \\ &= 1-Q^*_{F^N,X^N}(B_{N,\delta})+\exp_2\Big(-\exp_2[N(R-R_{ff}(D)-\delta)]\Big). \quad (119) \end{aligned}$$

When $R > R_{ff}(D)+\delta$, using part(b) of Lemma B.2, we have

$$\lim_{N\to\infty} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] = 0. \quad (120)$$

Therefore, there exists at least one sequence of codes $\{\mathcal{C}_N\}$ such that

$$\limsup_{N\to\infty} P_{X^N}\left(A^c(\mathcal{C}_N)\right) = 0.$$

In other words, there exists a sequence of codebooks $\{\mathcal{C}_N\}$ of code-functions for which

$$\lim_{N\to\infty} \Pr\left\{x^N \in \mathcal{X}^N : \quad d_N(x^N, f^N(x^{N-1})) > D+\delta, \qquad \forall f^N \in \mathcal{C}_N\right\} = 0. \quad (121)$$

The theorem follows. □

B.1 Proof of Lemma B.2

Proof: (a) From the definition, we have

$$i_{Q^*_{X^N,F^N}}(x^N;f^N) = \log\frac{Q^*_{F^N|X^N}(f^N|x^N)}{Q^*_{F^N}(f^N)}$$

Therefore,

$$Q^*_{F^N|X^N}(f^N|x^N) = Q^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N;f^N)] = P^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N;f^N)], \quad (122)$$

where the second equality follows because $P^*_{F^N}$ is used to construct Q^*. Moreover,

$$\frac{1}{N}i_{Q^*}(x^N;f^N) < \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X)+\delta, \qquad \forall (x^n,f^N) \in B_{N,\delta}. \quad (123)$$

Substituting the above in (122), we get part (a) of the lemma.

(b) The code function distribution $P^*_{F^N}$, the test-channel $\left\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\right\}_{n=1}^{N}$ determines a nice joint distribution $Q^*_{F^N,X^N,\hat{X}^N}$, given by (94). Under these conditions Q^* satisfies

$$\begin{aligned}
\frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} &= \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^N}}{Q^*_{X^N}} \\
&\overset{a}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^n}}{Q^*_{X^N}} \\
&\overset{b}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},\hat{X}^n}}{Q^*_{X^N}} \\
&= \frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}},
\end{aligned} \tag{124}$$

where, as before, $\vec{Q}^*_{\hat{X}^N|X^N} = \prod_{n=1}^N Q^*_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}$. (a) holds because it can be shown [34] that the condition $Q_{X_n|X^{n-1},F^N} = Q_{X_n|X^{n-1},F^n}$ is equivalent to (90) , while (b) follows from (90) and (91). (124) is essentially Lemma 5.1 in [16]. Thus we have

$$i_{Q^*_{X^N,F^N}}(f^N;x^N) = \frac{1}{N}\log\frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} = \frac{1}{N}\log\frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}} = \vec{i}_{Q^*_{\hat{X}^N,X^N}}(\hat{x}^N;x^N). \tag{125}$$

Define

$$\begin{aligned}
\vec{P}^{dec}_{\hat{X}^N|X^N} &= \prod_{n=1}^N P^{dec}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}, \\
\vec{P}^{ch}_{X^N|\hat{X}^N} &= \prod_{n=1}^N P^{ch}_{X_n|X^{n-1},\hat{X}^n}.
\end{aligned}$$

Since $P^*_{F^N}$ is chosen to be good with respect to $\vec{P}^{dec}_{\hat{X}^N|X^N}$ for the test channel P^{ch}, we have from (112)

$$Q^*_{X^N,\hat{X}^N} = \vec{Q^*}_{\hat{X}^N|X^N}\vec{Q^*}_{X^N|\hat{X}^N} = \vec{P}^{dec}_{\hat{X}^N|X^N}\vec{P}^{ch}_{X^N|\hat{X}^N} = P_{X^N}P^*_{\hat{X}^N|X^N}. \tag{126}$$

Using (126) in (125), we get

$$i_{Q^*_{X^N,F^N}}(f^N;x^N) = \vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N;x^N). \tag{127}$$

Now,

$$\begin{aligned}
Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta}) \quad &= \quad Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad d'_N(x^N,f^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta \\
&\qquad\qquad \text{or} \quad \frac{1}{N}i_{Q^*}(x^N;f^N) \geq \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X) + \delta\Big) \\
&\leq \quad Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad d'_N(x^N,f^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) \\
&\quad + Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \quad \frac{1}{N}i_{Q^*}(x^N;f^N) \geq \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X}\to X) + \delta\Big).
\end{aligned} \tag{128}$$

Since

$$d'_N(x^N,f^N) = d_N(x^N,f^N(x^{N-1})) = d_N(x^N,\hat{x}^N), \tag{129}$$

the first term in (128) equals

$$\begin{aligned} &Q^*_{X^N,\hat{X}^N}\left((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta\right) \\ &\qquad = P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta\right), \end{aligned} \tag{130}$$

where we have used (126). Since $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is the lim sup in probability of $d_N(x^N,\hat{x}^N)$,

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta\right) = 0. \tag{131}$$

The second term in (128) equals

$$\begin{aligned} &Q^*_{F^N,X^N,\hat{X}^N}\left((f^N,x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\right) \\ &\qquad = P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\right), \end{aligned} \tag{132}$$

where we have used (127) for the first equality and (126) for the next. Since $\overline{I}_{\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)$ is the lim sup in probability of $\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N;x^N)$,

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\left((x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\right) = 0. \tag{133}$$

Equations (131) and (133) imply

$$\lim_{N\to\infty} Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta}) = 0, \tag{134}$$

proving part (b) of the lemma. □

C Proof of Converse Part of Theorem 2

Let $\{\mathcal{C}_N\}_{N=1}^{\infty}$ be any sequence of codes, with rate R, that achieve distortion D (either expected distortion D or probability-1 distortion D depending on the criterion used). For any given block length N, there is an induced $P_{F^N|X^N}$. (equal to 1 for the code function f^N chosen to represent X^N and 0 for the other $2^{NR}-1$ code functions). This, along with the source distribution $P(X^N)$, determines P_{F^N}, a 2^{NR}-point discrete distribution. Thus, given the source distribution and the encoding and decoding rules, a joint distribution is induced. $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N, f^N \in \{f^N[i], i=1,\cdots,2^{NR}\}$, the induced distribution is given by

$$\hat{Q}_{X^N,F^N,\hat{X}^N}(x^N,f^N,\hat{x}^N) = P_{X^N}(x^N)\cdot P_{F^N|X^N}(f^N|x^N)\cdot \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}. \tag{135}$$

All probability distributions in the remainder of this section are marginals drawn from the induced joint distribution in (135). We first show that for any such induced distribution, we have

$$\overline{H}(F) = \limsup_{inprob} \frac{1}{N}\log\frac{1}{P(F^N)} \leq R. \tag{136}$$

Equivalently, we show that for any $\delta > 0$,

$$\lim_{N\to\infty} \Pr\left(\frac{1}{N}\log\frac{1}{P(F^N)} > R+\delta\right) = 0. \tag{137}$$

We have

$$\begin{aligned}\Pr\left(\frac{1}{N}\log\frac{1}{P(F^N)} > R+\delta\right) &= \Pr\left(P(F^N) < 2^{-N(R+\delta)}\right)\\ &= \sum_{f^N: 0<P_{F^N}(f^N)<2^{-N(R+\delta)}} P_{F^N}(f^N)\\ &\leq \sum_{f^N: P_{F^N}(f^N)>0} 2^{-N(R+\delta)} \qquad (138)\\ &= 2^{NR}\cdot 2^{-N(R+\delta)}\\ &= 2^{-N\delta} \to 0 \quad \text{as} \quad N\to\infty,\end{aligned}$$

thereby proving (136). Thus we have

$$R \geq \overline{H}(F) \geq \overline{H}(F) - \underline{H}(F|X) \geq \overline{I}(F;X), \tag{139}$$

where the last inequality follows from Lemma 2 in [22]. We need the following lemma, whose proof is found in Section C.1.

Lemma C.1. *For any sequence of codes as defined above, we have*

$$\overline{I}(F;X) \geq \overline{I}(\hat{X}\to X), \tag{140}$$

where the above quantities are computed with joint distribution induced by the code.

Using this lemma in (139), we obtain

$$R \geq \overline{I}(\hat{X}\to X). \tag{141}$$

By assumption, the sequence of codes with rate R achieves distortion D. This means that the induced output distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ satisfies the distortion constraint in Theorem 2. Therefore, we have

$$R \geq \overline{I}(\hat{X}\to X) \geq R_{ff}(D). \tag{142}$$

□

C.1 Proof of Lemma C.1

Let $\hat{Q}_{X^N,F^N,\hat{X}^N}$ be the joint distribution induced by the source code as in (135). From Definition 4.5, we have

$$i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N) = \log\frac{P(X^N|F^N)}{\vec{P}(X^N|\hat{X}^N)} = \log\frac{P(X^N|F^N)}{\prod_{n=1}^{N} P(X_n|X^{n-1},\hat{X}^n)}, \tag{143}$$

where the distributions are those induced from the source code. The upper-case notation we have used indicates that we want to consider the probabilities and the information quantities as random variables. We will first show that

$$\liminf_{inprob} \frac{1}{N}\left(i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N)\right) \geq 0. \tag{144}$$

This is equivalent to proving that for any $\delta > 0$,

$$\lim_{N\to\infty} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) = 0. \tag{145}$$

Since $F^n(X^{n-1}) = \hat{X}^n$, we have

$$P(X^N|F^N) = \prod_{n=1}^N P(X_n|X^{n-1},F^N) = \prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n). \tag{146}$$

Therefore,

$$\begin{aligned} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) &= P\left(\prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n) < 2^{-N\delta}\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)\right) \\ &= \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N), \end{aligned} \tag{147}$$

where

$$\mathcal{G} = \left\{(f^N,x^N,\hat{x}^N) : \prod_{n=1}^N P_{X_n|X^{n-1},F^N,\hat{X}^n}(x_n|x^{n-1},f^N,\hat{x}^n) < 2^{-N\delta}\prod_{n=1}^N P_{X_n|X^{n-1},\hat{X}^n}(x_n|x^{n-1},\hat{x}^n)\right\}.$$

In the remainder of this section, we drop the subscripts of the probabilities since the arguments make it clear what P refers to in each case.

$$\begin{aligned} \sum_{\mathcal{G}} \hat{Q}_{F^N,X^N,\hat{X}^N}(f^N,x^N,\hat{x}^N) &= \sum_{\mathcal{G}} P(f^N)P(x^N,\hat{x}^N|f^N) \\ &= \sum_{\mathcal{G}} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n,f^N)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &< 2^{-N\delta}\sum_{\mathcal{G}} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\leq 2^{-N\delta}\sum_{x^N,f^N,\hat{x}^N} P(f^N)\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\overset{(a)}{=} 2^{-N\delta}\sum_{f^N} P(f^N)\sum_{(x^N,\hat{x}^N):f^N(x^{N-1})=\hat{x}^N}\prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\overset{(b)}{=} 2^{-N\delta}\sum_{f^N} P(f^N)\sum_{x^N}\prod_{n=1}^N P(x_n|x^{n-1},f^n(x^{n-1})) \\ &\overset{(c)}{=} 2^{-N\delta}\cdot 1, \end{aligned} \tag{148}$$

where (a) follows from the fact that $\hat{x}^N = f^N(x^{N-1})$ and (b) since the term $P(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}, f^N)$ is equal to 1 when $\hat{x}_n = f_n(x^{n-1})$ and zero otherwise. (c) is obtained by evaluating the inner summation first over x_N, then over x_{N-1} and observing that all the f_n's are constant in the inner summation. Therefore (147) becomes

$$\begin{aligned} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) &= \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N) \\ &< 2^{-N\delta}. \end{aligned} \tag{149}$$

Hence

$$\lim_{N\to\infty} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) = 0. \tag{150}$$

Thus we have proved (144). Now, using the inequality

$$\liminf_{inprob}(a_n + b_n) \leq \limsup_{inprob} a_n + \liminf_{inprob} b_n \tag{151}$$

in (144), we get

$$\begin{aligned} 0 \leq \liminf_{inprob} \frac{1}{N}\left(i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N)\right) &\leq \limsup_{inprob} \frac{1}{N} i(F^N;X^N) + \liminf_{inprob} -\frac{1}{N}\vec{i}(\hat{X}^N;X^N) \\ &= \limsup_{inprob} \frac{1}{N} i(F^N;X^N) - \limsup_{inprob} \frac{1}{N}\vec{i}(\hat{X}^N;X^N). \end{aligned} \tag{152}$$

Or,

$$\overline{I}(F;X) \geq \overline{I}(\hat{X} \to X), \tag{153}$$

completing the proof of Lemma C.1. □

D Proof of Theorem 3

The source distribution is a sequence of distributions $\mathbf{P_X} = \{P_{X^n}\}_{n=1}^{\infty}$, where for each n, P_{X^n} is a product distribution. The rate-distortion function for an arbitrary memoryless source without feed-forward is

$$R_{DMS}(D) = \inf_{\mathbf{P_{\hat{X}|X}}:\lambda(\mathbf{P_{\hat{X}|X}})\leq D} \overline{I}(\hat{X};X), \tag{154}$$

where

$$\lambda(\mathbf{P_{\hat{X}|X}}) \triangleq \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^N d_i(X_i,\hat{X}_i)]. \tag{155}$$

Part 1: We first show that for a memoryless distortion measure with an expected distortion constraint, a memoryless conditional distribution achieves the infimum. Let $\mathbf{P_{\hat{X}|X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be any conditional distribution, for which the sup-directed information is $\overline{I}(\hat{X};X)$ and expected distortion is D. We will show that there exists a memoryless conditional distribution $\mathbf{P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}}$ such that $\overline{I}_{ML}(\hat{X};X) \le \overline{I}(\hat{X};X)$ and the expected distortion with $\mathbf{P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}}$ is the same, i.e., D. From the corresponding joint distribution $\mathbf{P_X P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{X^n,\hat{X}^n}\}$, form a memoryless joint distribution $\mathbf{P_X P}^{ML}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^{ML}_{X^n,\hat{X}^n}\}$ as follows. Set

$$P^{ML}_{X^n,\hat{X}^n} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i}, \tag{156}$$

where $P_{X_i,\hat{X}_i}, i \in \{1,\cdots,n\}$ are the marginals of $P_{X^n,\hat{X}^n}$. Clearly, for any N, the expected distortion with $P_{X^N,\hat{X}^N}$

$$E_{P_{X^N,\hat{X}^N}}[\frac{1}{N}\sum_{i=1}^{N} d_i(X_i,\hat{X}_i)] = \frac{1}{N}\sum_{i=1}^{N} E_{P_{X_i,\hat{X}_i}} d_i(X_i,\hat{X}_i) \tag{157}$$

is the same for $P^{ML}_{X^N,\hat{X}^N}$. We need to show

$$\overline{I}_{ML}(\hat{X};X) \le \overline{I}(\hat{X};X) \qquad \text{or}$$

$$\limsup_{inprob} \frac{1}{N} i_{ML}(\hat{X}^N;X^N) \le \limsup_{inprob} \frac{1}{N} i(\hat{X}^N;X^N).$$

To prove that

$$\limsup_{inprob} a_n \ge \limsup_{inprob} b_n, \tag{158}$$

it is enough to show that $\liminf_{inprob} a_n - b_n \ge 0$. This would imply

$$\begin{aligned} 0 \le \liminf_{inprob} a_n - b_n &\le \limsup_{inprob} a_n + \liminf_{inprob} -b_n \\ &= \limsup_{inprob} a_n - \limsup_{inprob} b_n. \end{aligned} \tag{159}$$

We have

$$\begin{aligned} \frac{1}{N}\left(i(\hat{X}^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) &= \frac{1}{N}\log \frac{P(\hat{X}^N,X^N)}{P(\hat{X}^N)\prod_{i=1}^{N} P(X_i)} \cdot \prod_{i=1}^{N} \frac{P(X_i)P(\hat{X}_i)}{P(X_i,\hat{X}_i)} \\ &= \frac{1}{N}\log \frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^{N} P(X_i|\hat{X}_i)}. \end{aligned} \tag{160}$$

We want to show that the $\liminf_{inprob}$ of the expression in (160) is ≥ 0. This is equivalent to showing that for any $\delta > 0$,

$$\lim_{N\to\infty} \Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) < -\delta\right] = 0. \tag{161}$$

Let

$$\mathcal{G} = \left\{(x^N,\hat{x}^N) : P_{X^N|\hat{X}^N}(x^N|\hat{x}^N) < 2^{-N\delta}\prod_{i=1}^{N} P_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\right\}.$$

Then,

$$\begin{aligned}\Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)<-\delta\right] &= \Pr\left[\frac{1}{N}\log\frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^N P(X_i|\hat{X}_i)}<-\delta\right]\\ &= \Pr\left[P(X^N|\hat{X}^N)<2^{-N\delta}\prod_{i=1}^N P(X_i|\hat{X}_i)\right]\\ &= \sum_{(x^N,\hat{x}^N)\in\mathcal{G}} P_{\hat{X}^N}(\hat{x}^N)P_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \qquad (162)\\ &\overset{(a)}{\leq} 2^{-N\delta}\sum_{(x^N,\hat{x}^N)\in\mathcal{G}}\prod_{i=1}^N P_{\hat{X}_i|X^{i-1}}(\hat{x}_i|x^{i-1})P_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\\ &\overset{(b)}{=} 2^{-N\delta}\cdot 1,\end{aligned}$$

where (a) follows from the definition of $\mathcal{G}$ and (b) is obtained by evaluating the sum first over x_N, then over $\hat{x}_N$ and so on. The arguments in (158) and (159) complete the proof that the infimum achieving distribution can be assumed to be memoryless in source coding without feed-forward. We now show that feed-forward does not change the rate-distortion function of the memoryless source.

Part 2: Let $\{\mathcal{C}_N\}_{N=1}^{\infty}$ be any sequence of codes with feed-forward, with rate R, that is achievable at distortion D. For any given block length N, a joint distribution described by (135) is induced:

$$\hat{Q}_{X^N,F^N,\hat{X}^N}=P_{X^N}\cdot P_{F^N|X^N}\cdot\delta_{\{\hat{X}^N=F^N(X^{N-1})\}}. \qquad (163)$$

All probability distributions in the remainder of this section are marginals drawn from this induced joint distribution. As in Part 1, define a memoryless distribution $\hat{Q}^{ML}_{X^N,\hat{X}^N}\triangleq\prod_{n=1}^N\hat{Q}_{X_n,\hat{X}_n}$. The subscript ML on an information quantity will imply that $\hat{Q}^{ML}_{X^N,\hat{X}^N}$ is the distribution being used to compute it. As shown in Appendix C ((136) to (139)), for this joint distribution we have

$$R\geq\overline{H}(F)\geq\overline{H}(F)-\underline{H}(F|X)\geq\overline{I}(F;X). \qquad (164)$$

It remains to show that when the source is memoryless,

$$\begin{gathered}\overline{I}(F;X)\geq\overline{I}_{ML}(\hat{X};X)\quad\text{or}\\ \limsup_{inprob}\frac{1}{N}i(F^N;X^N)\geq\limsup_{inprob}\frac{1}{N}i_{ML}(\hat{X}^N;X^N).\end{gathered}\qquad (165)$$

As in Part 1 of this proof, it suffices to show that $\liminf_{inprob}\frac{1}{N}\left(i(F^N;X^N)-i(\hat{X}^N;X^N)\right)\geq 0$ or equivalently that for all $\delta>0$,

$$\lim_{N\to\infty}\Pr\left[\frac{1}{N}\left(i(F^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right)<-\delta\right]=0. \qquad (166)$$

Noting that $\hat{Q}^{ML}_{X^N,\hat{X}^N}$ is memoryless, we have

$$\begin{aligned}\frac{1}{N}\left(i(F^N;X^N)-i_{ML}(\hat{X}^N;X^N)\right) &= \frac{1}{N}\log\frac{\hat{Q}(F^N,X^N)}{\hat{Q}(F^N)\prod_{n=1}^N P(X_n)}\cdot\prod_{n=1}^N\frac{P(X_n)\hat{Q}(\hat{X}_i)}{\hat{Q}(X_n,\hat{X}_n)}\\ &= \frac{1}{N}\log\frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^N\hat{Q}(X_n|\hat{X}_n)}.\end{aligned}\qquad (167)$$

Hence, we have

$$\begin{aligned}
\Pr\left[\frac{1}{N}\left(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) < -\delta\right] &= \Pr\left[\frac{1}{N}\log\frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^N \hat{Q}(X_n|\hat{X}_n)} < -\delta\right] \\
&= \Pr\left[\hat{Q}(X^N|F^N) < 2^{-N\delta}\prod_{n=1}^{N}\hat{Q}(X_n|\hat{X}_n)\right] \\
&= \Pr\left[(x^N, f^N, \hat{x}^N) : \hat{Q}_{X^N|F^N}(x^N|f^N) < 2^{-N\delta}\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right] \qquad (168)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\hat{Q}_{X^N|F^N}(x^N|f^N)\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) \\
&\leq 2^{-N\delta}\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N),
\end{aligned}$$

where

$$\nu(f^N) \triangleq \left\{(x^N,\hat{x}^N) : \hat{Q}_{X^N|F^N}(x^N|f^N) < 2^{-N\delta}\prod_{i=1}^{N}\hat{Q}_{X_i|\hat{X}_i}(x_i|\hat{x}_i)\right\}.$$

Since f^N and x^N determine the reconstruction $\hat{x}^N$, $\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) = 1$ if $\hat{x}^N = f^N(x^{N-1})$ and 0 otherwise. Thus we have

$$\begin{aligned}
&\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) \\
&\qquad = \sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{x^N}\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|f_n(x^{n-1})) = 1,
\end{aligned} \qquad (169)$$

where the inner summation is computed first over x_N, then x_{N-1} and so on up to x_1. Thus

$$\begin{aligned}
\Pr\left[\frac{1}{N}\left(i(F^N;X^N) - i(\hat{X}^N;X^N) < -\delta\right)\right] &\leq 2^{-N\delta} \\
&\to 0 \quad \text{as} \quad N\to\infty,
\end{aligned} \qquad (170)$$

proving (166). We have shown that any achievable rate R (with feed-forward) satisfies

$$R \geq \overline{I}_{ML}(\hat{X};X).$$

This implies that the rate-distortion function with feed-forward is the same as that without feed-forward. □

E Proof of Theorem 4

As shown in Figure 6, $X - F$ is a system without feed forward with F^N considered to be the 'reconstruction' of X^N. The distortion between F^N and X^N is defined as $d_N(X^N, F^N(X^{N-1}))$. As in (53) in Section 5, by randomly choosing code-functions with distribution r_{F^N}, we know there exists a code of 2^{NR} N-length code-functions F^N with error exponent

$$E_N(R,D,r_{F^N}) = \max_{s\geq 0}\min_{t\leq 0}\left[sR - stD - \frac{1}{N}\log_2 \sum_{x^N} P_{X^N}(x^N)\left(\sum_{f^N} r_{f^N}(f^N)e^{tNd_N\left(x^N, f^N(X^{N-1})\right)}\right)^{-s}\right]. \tag{171}$$

In the summation over f^N on the right hand side, x^N is constant. Noting that the reconstruction is given by $\hat{x}^N = f^N(x^{N-1})$, we can write the previous equation as

$$E_N(R,D,r_{F^N}) = \max_{s\geq 0}\min_{t\leq 0}\left[sR - stD - \frac{1}{N}\log_2 \sum_{x^N} P_{X^N}(x^N)\left(\sum_{\hat{x}^N}\sum_{f^N : f^N(x^{N-1}) = \hat{x}^N} r_{F^N}(f^N)e^{tNd_N\left(x^N, \hat{x}^N\right)}\right)^{-s}\right] \tag{172}$$

Let $\vec{q}^{*}_{\hat{X}^N|X^N} = \prod_{n=1}^{N} q^{*}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}$ be the distribution that achieves the maximum in the definition of $E_{ff-N}(R,D)$ in (55). Fix any test-channel distribution $\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\}_{n=1}^{N}$. For this channel, choose r_{f^N} to be good with respect to the distribution $\{q^{*}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}\}_{n=1}^{N}$. Recall from Definition B.1 in Appendix B that r_{F^N} is good with respect to $\{q^{*}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}\}_{n=1}^{N}$ for a channel P^{ch} if the unique, nice joint distribution $Q_{F^N,X^N\hat{X}^N}$ determined by r_{F^N} and the channel P^{ch} has induced decoder distribution that satisfies

$$\prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}} = \prod_{n=1}^{N} q^{*}_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}. \tag{173}$$

Define $\forall n \in \{1,\cdots,N\}$

$$\begin{aligned}
\Gamma_n(x^{n-1},\hat{x}_n) &\triangleq \left\{f_n : f_n(x^{n-1}) = \hat{x}_n\right\}, &(174)\\
\Gamma^n(x^{n-1},\hat{x}^n) &\triangleq \left\{f^n : f_i(x^{i-1}) = \hat{x}_i, \quad i = 1,\ldots,n\right\}. &(175)
\end{aligned}$$

Now, for any pair $x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \hat{\mathcal{X}}^N$, we have

$$\begin{aligned}
\sum_{f^N : f^N(x^{N-1}) = \hat{x}^N} r_{F^N}(f^N) &= r_{F^N}(\Gamma^N(x^{N-1},\hat{x}^N))\\
&= r_{F^N}\left(\Gamma_1(\hat{x}_1),\ldots,\Gamma_n(x^{n-1},\hat{x}_n),\ldots,\Gamma_N(x^{N-1},\hat{x}_N)\right)\\
&= \prod_{n=1}^{N} r_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1},\hat{x}_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1})\right).
\end{aligned} \tag{176}$$

It was shown in Part A of the proof of Lemma B.1 that the input distribution induced from $Q_{F^N,X^N,\hat{X}^N}$ is given by

$$Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}) = r_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1},\hat{x}_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1})\right), \quad n = 1,\ldots,N. \tag{177}$$

Using this in (176), we have

$$\sum_{f^N : f^N(x^{N-1}) = \hat{x}^N} r_{F^N}(f^N) = \prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}). \tag{178}$$

Since r_{F^N} is good with respect to $\vec{q}^*_{\hat{X}^N|X^N}$, combining (178) and (173), we get

$$\sum_{f^N : f^N(x^{N-1}) = \hat{x}^N} r_{F^N}(f^N) = \prod_{n=1}^{N} q^*_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) = \vec{q}^*_{\hat{X}^N|X^N}(\hat{x}^N|x^N). \tag{179}$$

Substituting the above in (172), we get

$$E_N(R, D, r_{F^N}) = E_{ff-N}(R, D). \tag{180}$$

Since there exists a rate R source code with error exponent $E_N(R, D, r_{F^N})$, the theorem follows from the above equation. □

F Proof of Theorem 6

We start with the definition of the error exponent $E_{ff-N}(R, D)$ given by Theorem 6 and show that it is the same as that given by Theorem 4. This is a generalization of Theorem 6.6.5 in [26]. The proof requires three lemmas, first two of which are needed to prove the third.

Lemma F.1. *Given the distributions P_{X^N} and $P_{\hat{X}^N|X^N}$, the directed information can be written as*

$$I(\hat{X}^N \to X^N) = \min_{\vec{q}_{\hat{X}^N|X^N}} \sum_{x^N} \sum_{\hat{x}^N} P_{X^N}(x^N) P_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \log \frac{P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}, \tag{181}$$

where $\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) = \prod_{n=1}^{N} q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1})$.

Proof. The joint distribution $P_{X^N} P_{\hat{X}^N|X^N}$ can be split as

$$\begin{aligned} P_{X^N} P_{\hat{X}^N|X^N}(x^n, \hat{x}^N) &= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \\ &= \prod_{n=1}^{N} P_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) \cdot \prod_{n=1}^{N} P_{X_n|X^{n-1},\hat{X}^n}(x_n|x^{n-1}, \hat{x}^n). \end{aligned} \tag{182}$$

As in (37), we can write the directed information as

$$I(\hat{X}^N \to X^N) = \sum_{x^N} \sum_{\hat{x}^N} P_{X^N} P_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}. \tag{183}$$

Therefore,

$$\begin{aligned}
&-I(\hat{X}^N \to X^N) + \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} P_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\
&\quad = \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} P_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\
&\quad \overset{(a)}{\geq} \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} P_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \left(1 - \frac{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{p}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}\right) \\
&\quad = \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} P_{\hat{X}^N|X^N}(x^N, \hat{x}^N) - \sum_{x^N}\sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \\
&\quad \overset{(b)}{=} 1 - 1 = 0,
\end{aligned} \tag{184}$$

where we have used the inequality $\log x \geq 1 - \frac{1}{x}$ to obtain (a), and (b) follows because $\vec{P}_{X^N|\hat{X}^N} \vec{q}_{\hat{X}^N|X^N}$ is a valid joint distribution. The lemma follows. □

Our next two lemmas express $B_N(P_{X^N}, D)$, described in Definition 5.1, in terms of a parameter s. This parameterization is then used to show that the expressions for the error exponent in Theorems 4 and 6 are the same.

Lemma F.2. *For a given source distribution P_{X^N}, the function $B_N(P_{X^N}, D)$, described in Definition 5.1, can be represented in terms of a parameter s as*

$$B_N(P_{X^N}, D_s) = sD_s + \min_{\vec{q}_{\hat{X}^N|X^N}} \left[-\frac{1}{N} \sum_{x^N} P_{X^N}(x^N) \log \sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N, \hat{x}^N)} \right], \tag{185}$$

where

$$D_s = \sum_{x^N}\sum_{\hat{x}^N} P_{X^N}(x^N) \cdot \frac{\vec{q}^{*}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N, \hat{x}^N)}}{\sum_{\hat{x}^N} \vec{q}^{*}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N, \hat{x}^N)}} \cdot d_N(x^N, \hat{x}^N), \tag{186}$$

where $\vec{q}^{*}_{\hat{X}^N|X^N}$ achieves $B_N(P_{X^N}, D_s)$.

Proof. It can be shown that $B_N(P_{X^N}, D)$ is convex in D for a fixed P_{X^N}. This follows from the fact that $I(\hat{X}^N \to X^N)$ is a convex function of the distribution $\vec{P}(\hat{X}^N|X^N)$. From Definition 5.1, it also follows that $B_N(P_{X^N}, D)$ is a nonnegative, non-increasing function which is zero at $D = D_{max}$, the maximum distortion. This, along with convexity implies that $B_N(P_{X^N}, D)$ is strictly decreasing in D, which means the constraint on D in the definition must be satisfied with equality. Thus we can write $B_N(P_{X^N}, D)$ in terms of a Lagrange multiplier as

$$\begin{aligned}
B_N(P_{X^N}, D_s) = \min_{\hat{q}_{\hat{X}^N|X^N}} \Bigg[&\frac{1}{N} \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} \hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{\hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\
&-s \left(\sum_{x^N}\sum_{\hat{x}^N} P_{X^N} \hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) - D_s \right) \Bigg],
\end{aligned} \tag{187}$$

where $\vec{P}_{\hat{X}^N|X^N}$ is induced from $P_{X^N}\hat{q}_{\hat{X}^N|X^N}$ and

$$D_s = \sum_{x^N}\sum_{\hat{x}^N} P_{X^N} q^*_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N), \tag{188}$$

where $q^*_{\hat{X}^N|X^N}$ achieves the minimum. Using the previous lemma, we can write (187) as a double minimum

$$\begin{aligned} B_N(P_{X^N}, D_s) = sD_s + \min_{\vec{q}_{\hat{X}^N|X^N}} \min_{\hat{q}_{\hat{X}^N|X^N}} \Bigg[&\frac{1}{N}\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}\hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{\hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\ &- s\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}\hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) \Bigg], \end{aligned} \tag{189}$$

where now

$$D_s = \frac{1}{N}\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}\hat{q}^*_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N). \tag{190}$$

For a fixed $\vec{q}_{\hat{X}^N|X^N}$, we find the optimal $\hat{q}_{\hat{X}^N|X^N}$. Introducing Lagrange multipliers λ_{x^N} to constrain $\sum_{\hat{x}^N} \hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) = 1$, and differentiating to obtain the minimum, we get

$$\begin{aligned} \frac{\partial}{\partial \hat{q}_{\hat{X}^N|X^N}} \Bigg(&\frac{1}{N}\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}\hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{\hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\ &- s\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}\hat{q}_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) + \sum_{x^N}\lambda_{x^N}\sum_{\hat{x}^N}\hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \Bigg) = 0. \end{aligned} \tag{191}$$

Or,

$$\begin{aligned} &\frac{1}{N}P_{X^N}(x^N)\log \hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) - \frac{1}{N}P_{X^N}(x^N)\log \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \\ &\qquad + \frac{1}{N}P_{X^N}(x^N) - sP_{X^N}(x^N) d_N(x^N, \hat{x}^N) + \lambda_{x^N} = 0. \end{aligned} \tag{192}$$

Solving for $\hat{q}_{\hat{X}^N|X^N}$ and choosing λ_{x^N} so that

$$\sum_{\hat{x}^N} \hat{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) = 1$$

gives

$$\hat{q}^*_{\hat{X}^N|X^N}(\hat{x}^N|x^N) = \frac{\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)}}{\sum_{\hat{x}^N}\vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)}}. \tag{193}$$

Substituting the above in (189) and (190), we get the lemma. □

Lemma F.3. *An alternate representation of $B_N(P_{X^N}, D)$ is given by*

$$B_N(P_{X^N}, D) = \max_{s\le 0} \min_{\vec{q}_{\hat{X}^N|X^N}} \left[sD - \frac{1}{N}\sum_{x^N} P_{X^N}(x^N) \log \sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)} \right] \tag{194}$$

Proof. From the definition of $B_N(P_{X^N}, D)$ (see Definition 5.1), for any $s \le 0$, we can write

$$\begin{aligned} B_N(P_{X^N}, D) \ge \min_{Q_{\hat{X}^N|X^N} \in \mathcal{Q}_D} \Bigg[&\frac{1}{N}\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}Q_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{Q_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \\ &- s\left(\sum_{x^N}\sum_{\hat{x}^N} P_{X^N}Q_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) - D \right) \Bigg], \end{aligned} \tag{195}$$

where

$$\mathcal{Q}_D = \left\{ Q_{\hat{X}^N|X^N} : \sum_{x^N} \sum_{\hat{x}^N} P_{X^N} Q_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) \leq D \right\}.$$

The inequality holds because the term multiplying s is always negative for $s \leq 0$. If the constraint set is enlarged, the minimum cannot increase and hence we have,

$$B_N(P_{X^N}, D) \geq \min_{Q_{\hat{X}^N|X^N}} \left[\frac{1}{N} \sum_{x^N} \sum_{\hat{x}^N} P_{X^N} Q_{\hat{X}^N|X^N}(x^N, \hat{x}^N) \log \frac{Q_{\hat{X}^N|X^N}(\hat{x}^N|x^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)} \right. \\ \left. -s \left(\sum_{x^N} \sum_{\hat{x}^N} P_{X^N} Q_{\hat{X}^N|X^N}(x^N, \hat{x}^N) d_N(x^N, \hat{x}^N) - D \right) \right]. \tag{196}$$

Now, using Lemma F.1, we can write the directed mutual information (the first term in the above equation) in terms of a double minimum and explicitly evaluate the minimum over $Q_{\hat{X}^N|X^N}$ as in the previous lemma. This gives

$$B_N(P_{X^N}, D_s) \geq \min_{\vec{q}_{\hat{X}^N|X^N}} \left[sD_s - \frac{1}{N} \sum_{x^N} P_{X^N}(x^N) \log \sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N, \hat{x}^N)} \right], \tag{197}$$

where D_s is as in (186). The above holds for all for all $s \leq 0$. Therefore,

$$B_N(P_{X^N}, D_s) \geq \max_{s \leq 0} \min_{\vec{q}_{\hat{X}^N|X^N}} \left[sD_s - \frac{1}{N} \sum_{x^N} P_{X^N}(x^N) \log \sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N, \hat{x}^N)} \right], \tag{198}$$

with D_s as in (186). But Lemma F.2 says that for some value of the Lagrange multiplier $s \leq 0$, the inequality is achieved with equality, thereby completing the proof. □

We are now ready to prove Theorem 6.

Proof of Theorem 6

We start with the representation given by Theorem 6. We can show that $E_{ff-N}(R, D)$ is convex and increasing in R. (For this, we use an alternate representation similar to Definition 6.6.1 and Theorem 6.6.4 in [26]) Hence, in the range of R for which $E_{ff-N}(R, D)$ is strictly increasing, we can express $E_{ff-N}(R, D)$ using a Lagrange multiplier as

$$E_{ff-N}(R, D) = \min_{\hat{p}_{X^N}} \frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)} + s\left[R - B_N(\hat{p}_{X^N}, D)\right]. \tag{199}$$

Equivalently for each $s \geq 0$, we can write

$$E_{ff-N}(R, D) \geq \min_{\hat{p}_{X^N} \in \mathcal{P}} \left[\frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)} + s\left[R - B_N(\hat{p}_{X^N}, D)\right] \right], \tag{200}$$

since for $s \geq 0$, the term multiplying s is non-positive. Here, $\mathcal{P}$ is as in (58) given by

$$\mathcal{P} = \{\hat{p}_{X^N} : B_N(\hat{p}_{X^N}, D) \geq R\}. \tag{201}$$

The inequality is still valid if the constraint set is enlarged to include all $\hat{p}_{X^N}$. Further, since the inequality holds for all $s \geq 0$, we can choose the maximizing s to give

$$E_{ff-N}(R, D) \geq \max_{s \geq 0} \min_{\hat{p}_{X^N}} \left[\frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)} + s\left[R - B_N(\hat{p}_{X^N}, D)\right] \right]. \tag{202}$$

But, from the Lagrange multiplier formulation in (199), we see that $E_{ff-N}(R,D)$ has the form on the right side for some value of $s \geq 0$. Therefore, the inequality can be replaced by equality

$$E_{ff-N}(R,D) = \max_{s\geq 0} \min_{\hat{p}_{X^N}} \left[\frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)} + s\left[R - B_N(\hat{p}_{X^N}, D)\right] \right]. \tag{203}$$

We now use Lemma F.3 to express $B_N(\hat{p}_{X^N}, D)$. Using t in place of the s used in Lemma F.3 and noting that $B_N(\hat{p}_{X^N}, D)$ appears with a negative sign, we can write

$$\begin{aligned} E_{ff-N}(R,D) = \max_{s\geq 0} \min_{\hat{p}_{X^N}} \min_{t \leq 0} \max_{\vec{q}_{\hat{X}^N|X^N}} & \left[\frac{1}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \frac{\hat{p}_{X^N}(x^N)}{P_{X^N}(x^N)} + sR \right. \\ & \left. -stD + \frac{s}{N} \sum_{x^N} \hat{p}_{X^N}(x^N) \log \sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)} \right]. \end{aligned} \tag{204}$$

Now, the argument is convex in $\hat{p}_X^N$ and concave in $\vec{q}_{\hat{X}^N|X^N}$. This implies that the solution occurs at a saddle point, i.e., the minimax equals the maximin. Therefore we can interchange $\min_{\hat{p}_{X^N}}$ and $\max_{\vec{q}_{\hat{X}^N|X^N}}$ and evaluate the minimum over $\hat{p}_{X^N}$. By differentiation it can be verified that the minimum occurs at

$$\hat{p}_{X^N}(x^N) = \frac{P_{X^N}(x^N) \left[\sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)} \right]^{-s}}{\sum_{x^N} P_{X^N}(x^N) \left[\sum_{\hat{x}^N} \vec{q}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) e^{sNd_N(x^N,\hat{x}^N)} \right]^{-s}} . \tag{205}$$

Substituting this into the previous equation, we get the expression for $E_{ff-N}(R,D)$ given by Theorem 4, completing the proof of the theorem. □

References

[1] J. M. Kahn, R. H. Katz, and K. S. J. Pister, “Mobile networking for smart dust,” *ACM/IEEE International Conference on Mobile Computing*, Seattle, WA, August 1999.

[2] A. Wyner and J. Ziv, “The Rate-Distortion Function for Source Coding with Side Information at the Decoder,” *IEEE Transactions on Information Theory*, vol. 22, pp. 1–10, January 1976.

[3] T. Weissman and N. Merhav, “On competitive prediction and its relation to rate-distortion theory and channel capacity theory,” in *Proc. IEEE Int. Symp. on Inform. Theory (ISIT), Yokohama, Japan*, June/July 2003.

[4] 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.

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

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

[7] S. I. Krich, “Coding for a Delay-Dependent Fidelity Criterion,” *IEEE Transactions on Information Theory*, pp. 77–85, January 1974.

[8] D. Neuhoff and R. Gilbert, “Causal source codes,” *IEEE Trans. Inform. Theory*, pp. 701–713, Sept. 1982.

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

[10] S. S. Pradhan, J. Chou, and K. Ramchandran, “Duality between source coding and channel coding and its extension to the side-information case,” *IEEE Transactions on Information Theory*, vol. 49, pp. 1181–1203, May 2003.

[11] R. J. Barron, B. Chen, and G. W. Wornell, “The Duality between Information Embedding and Source Coding with Side Information and Some Applications,” *IEEE Transactions on Information Theory*, vol. 49, pp. 1159–1180, May 2003.

[12] J. Geweke, “Measurement of linear dependence and feedback between multiple time series,” *Journal of the American Statistical Association*, vol. 77, pp. 304–313, June 1982.

[13] P. E. Caines and C. W. Chan, “Feedback between stationary processes,” *IEEE Transactions on Automatic Control*, vol. AC-20, no. 378, pp. 498–508, 1975.

[14] J. Rissanen and M. Wax, “Measures of mutual and causal dependence between two time series,” *IEEE Transactions on Information Theory*, vol. IT-33, pp. 598–601, July 1987.

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

[16] S. Tatikonda and S. Mitter, “The Capacity of Channels with Feedback- Part 1:the General Case,” submitted to *IEEE Transactions on Information Theory*, October 2001.

[17] T. Berger, *Rate-distortion theory: A mathematical basis for data compression.* Massachusetts: Prentice-Hall, 1971.

[18] T. Cover and J. Thomas, *Elements of Information Theory.* New York: John Wiley and Sons, 1991.

[19] S. S. Pradhan, “Approximation of test channels in source coding,” in *Proc. of Conf. on Inform. Systems and Sciences (CISS)*, March 2004.

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

[21] C. E. Shannon, “Two-Way Communication Channels,” *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA*, vol. 1, pp. 611–644, 1961.

[22] Y. Steinberg and S. Verdú, “Simulation of random processes and rate-distortion theory,” *IEEE Transactions on Information Theory*, vol. 43, pp. 63–86, January 1996.

[23] S. Verdú and T.Han, “A General formula for Channel Capacity,” *IEEE Transactions on Information Theory*, vol. 40, pp. 1147–1157, July 1994.

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

[25] T. M. Cover and S. Pombra, “Gaussian feedback capacity,” *IEEE Transactions on Information Theory*, vol. IT-35, pp. 37–43, January 1989.

[26] R. E. Blahut, *Principles and Practice of Information Theory.* Massachusetts: Addison Wesley, 1988.

[27] K. Marton, “Error exponent for source coding with a fidelity criterion,” *IEEE Transactions on Information Theory*, vol. IT-20, pp. 197–199, March 1974.

[28] J. K. Omura, “A coding theorem for discrete time sources,” *IEEE Transactions on Information Theory*, vol. IT-19, pp. 490–498, July 1973.

[29] G. Longo, “On the error exponent for markov sources,” in *Proc. of the 2nd IEEE International symposium on Informtaion Theory (ISIT), Tsakhadsor, Soviet Union*, Budapest: Akademiai Kiado, 1971.

[30] K. Vasek, “On the error exponent for ergodic markov sources,” *Kybernetica*, vol. 16, no. 3, pp. 318–329, 1980.

[31] S. Natarajan, “Large deviations, hypotheses testing, and source coding for finite markov chains,” *IEEE Transactions on Information Theory*, vol. IT-31, pp. 360–365, May 1985.

[32] V. Ananthram, “A large deviations approach to error exponents in source coding and hypothesis testing,” *IEEE Transactions on Information Theory*, vol. IT-4, pp. 938–943, July 1990.

[33] P. H. Algoet and T. M. Cover, “A Sandwich Proof of the Shannon-McMillan-Breiman Theorem,” *The Annals of Probability*, vol. 16, pp. 899–909, April 1988.

[34] J. L. Massey, “Network Information Theory- Some Tentative Definitions.” DIMACS Workshop on Network Information Theory, April 2003.