Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source

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

***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 characterize the error exponent for a general source with feed-forward. The results are then extended to feed-forward with an arbitrary delay larger than the block length.**

***Index Terms*—Directed Information, random coding, real-time reconstruction, side information, source coding with feed-forward.**

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						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Fig. 1. Time-line: Instantaneous observations.

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						$\hat{X}_1$	$\hat{X}_2$	$\hat{X}_3$	$\hat{X}_4$	$\hat{X}_5$

Fig. 2. Time-line: delayed observations.

I. Introduction

With the 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 a random process $\{X_n\}_{n=1}^{\infty}$, needs to be encoded in blocks of length N into a message W. W is to be transmitted over a noiseless channel of finite rate to a decoder, which has access to some side information $\{Y_n\}_{n=1}^{\infty}$ that is correlated to the source X. 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 when (X, Y) is a joint independent and identically distributed (i.i.d.) process was obtained by Wyner and Ziv in [2]. The encoder and the decoder are in time-synchrony, i.e., to reconstruct a set of N samples of X, the decoder uses the corresponding set of N samples of Y. This is used to 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 simultaneously observed at the encoder and the decoder, respectively. 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 $\hat{X}_1, \dots \hat{X}_5$ simultaneously as a function of W and $Y_1, \dots Y_5$, though it may display them as shown in Fig. 1.

Often the side-information Y is a noisy version of the source X which is assumed to be available simultaneously at the decoder. The key question that we would like to ask is: what happens if the underlying signal field X is traveling slowly (compared to the speed of electromagnetic wave propagation) from the location of the encoder to that of the decoder, where it is available as Y? In other words, there is a delay between the instant when ith source sample X_i is observed at the encoder and the instant when corresponding noisy version Y_i is observed at the decoder.

We want the reconstruction to be real-time, so we need a new dynamic compression model. Fig. 2 shows such a model when the signal field delay is 6 time units and block length $N = 5$. In Fig. 2, for real-time reconstruction of the ith source sample, all the past $i - 1$ samples of the side information are available. In other words, the decoding 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 noncausal and the decoder is causal.

Manuscript received August 12, 2004; revised July 25, 2005. This work was supported by the NSF under Grants CCR-0219735, CCF-0329715 and CCF-0448115. This work was presented in part at IEEE Information Theory Workshop (ITW), San Antonio, TX, October 2004.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: rvenkata@eecs.umich.edu; pradhanv@eecs.umich.edu).

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2007.896887

0018-9448/$25.00 © 2007 IEEE

In this work, as a first step, we consider an idealized version of this problem where we assume that the traveling source field X is available noiselessly with an arbitrary delay at the decoder, i.e., $Y = X$. We call this problem source coding with noiseless feed-forward. This was first considered by Weissman and Merhav in [3] and [4] in the context of competitive prediction. 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. When the delay is $N + k$, we refer to the problem as source coding with delay k feed-forward. Thus with delay k feed-forward, the decoder has available the source samples until time $i - k$ as side-information to reconstruct $\hat{X}_i$.

The relevance of this problem extends much beyond the sensor networks application outlined above. 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 ith 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.[1]

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 the following.

- 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.

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]. They consider sources with feed-forward delay 1 and a single-letter, difference distortion measure. In [4], the optimal distortion-rate function with feed-forward is derived for sources that can be represented auto-regressively with an innovations process that is either i.i.d. or satisfies the Shannon Lower Bound (SLB) [5] with equality. The distortion-rate function was evaluated in [4] for a symmetric binary Markov source with feed-forward and a stationary Gaussian source with feed-forward as examples of this result. For sources with general innovations processes, [4] provides upper and lower bounds on the distortion-rate function. The block coding error exponent is also derived in [4] for the case where the innovations process is i.i.d. and is shown to be the same as Marton's no-feed-forward error exponent [6]. It was noted in [4] that feed-forward can only decrease the distortion-rate function of a source; however, with single-letter difference distortion measures, feed-forward does not reduce the optimal distortion-rate function for i.i.d. sources and all sources that satisfy SLB with equality.

Later, the model of source coding with general feed-forward was considered in [7], [8] as a variant of the problem of source coding with side information at the decoder, and a quantization scheme with linear processing for i.i.d. 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 [9], an elegant variable-length coding strategy to achieve the optimal Shannon rate-distortion bound for any finite-alphabet i.i.d. source with feed-forward was presented, along with a beautiful illustrative example. In [8], two-channel multiple-description source coding for i.i.d. Gaussian sources with feed-forward was also considered and the optimal rate-distortion function, error exponent were derived. The problem of source coding with feed-forward is also related to source coding with a delay-dependent distortion function [10], causal source coding [11] and real-time source coding [12].

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

1) The optimal rate-distortion function for a general discrete source with a general distortion measure and with noiseless feed-forward, $R_{\text{ff}}(D)$, is given by the minimum of the directed information function [13] flowing from the reconstruction to the source. From the properties of directed information, it will follow that $R_{\text{ff}}(D) \leq R(D)$, where $R(D)$ denotes the optimal Shannon rate-distortion function for the source without feed-forward.
2) We extend the Asymptotic Equipartition Property [5] to define a new kind of typicality that we call 'directed typicality'. This is used to provide a simple, intuitive direct coding theorem for stationary, ergodic sources with feed-forward.
3) 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 characterize the error exponent for a general source with feed-forward.
4) Extension of these results to feed-forward with arbitrary delay. We introduce a generalized form of directed information to analyze the problem of source coding with delayed feed-forward.

We now briefly outline how our results differ from that of [4]. In [4], feed-forward is considered in the context of competitive prediction. The optimal distortion-rate function of a source with feed-forward is completely characterized in [4] only when the source has an autoregressive representation with an innovations process that is either i.i.d. or satisfies the SLB with equality. This characterization of the distortion-rate function is in terms of the innovations process. In our work, we derive the optimal rate-distortion function with feed-forward for any general source

[1] We will use the superscript notation to denote a sequence of random variables. Thus $X^{i-1} = [X_1, \ldots, X_{i-1}]$.

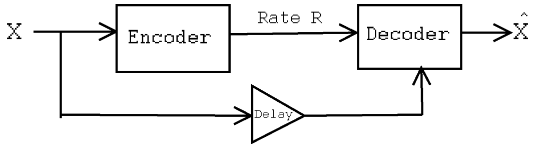

Fig. 3. Source coding system with feed-forward.

with feed-forward. This is expressed in terms of directed information, a quantity involving just the source X and the reconstruction $\hat{X}$. The feed-forward error exponent is derived in [4] for sources with an autoregressive representation with i.i.d. innovations. We characterize the error exponent for a general source. The results in [4] are derived for single-letter, difference distortion measures and feed-forward with delay 1. Our results are derived for arbitrary (not necessarily single-letter) distortion measures and feed-forward with arbitrary delay.

Our paper is organized as follows. In Section II, 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 III which captures the essence of this problem. We must mention here that for stationary, ergodic sources *without* feed-forward with single-letter distortion measures, the optimal rate-distortion function is attained by a jointly stationary and ergodic $(X, \hat{X})$ process [14]. Unfortunately, a similar result may not hold for stationary, ergodic sources with feed-forward even with single-letter distortion measures. This is because the information available at the decoder changes with time. Hence, we can only obtain a direct coding theorem by restricting our attention to stationary, ergodic joint processes.

To obtain a tight rate-distortion theorem, we have to consider general processes. The method of information spectrum introduced by Han and Verdu [15] is a powerful tool to deal with general processes. Using this, we give the direct and converse coding theorems for general sources in Section IV. In that section, we also consider some special cases such as discrete memoryless sources and Gaussian sources. Error exponents are considered in the general setting in Section V. We extend our results to arbitrary delays in Section VI and finally, concluding remarks are given in Section VII.

II. The Source Coding Model

A. Problem Statement

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

Definition 2.1: An $(N, 2^{NR})$ source code with feed-forward of block length N and rate R consists of an encoder mapping e and a sequence of decoder mappings $g_i, i = 1, \ldots, N$, where

$$e : \mathcal{X}^N \rightarrow \{1, \ldots, 2^{NR}\}$$
$$g_i : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{i-1} \rightarrow \hat{\mathcal{X}}, \; i = 1, \ldots, N.$$

The encoder maps each N-length source sequence to an index in $\{1, \ldots, 2^{NR}\}$. The decoder receives the index transmitted by the encoder, and to reconstruct the ith sample, it has access to all the past $(i-1)$ samples of the source. 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 $R_{\mathrm{ff}}(D)$ 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 [16]–[18] of channel coding with feedback.

B. 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 [13]. This notion was motivated by the work of Marko [19] and was also studied in [20]–[22] 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 [23], [24].

Definition 2.2: [13] The directed information flowing from a random vector A^N to another random vector B^N is defined as

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

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. We can write the directed information flowing from the reconstruction $\hat{X}^N$ to the source X^N as

$$I(\hat{X}^N \rightarrow X^N) = I(X^N; \hat{X}^N) - \sum_{n=2}^N I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}). \tag{3}$$

Equation (3) can be derived using the chain rule as follows [25].

$$\begin{aligned}
&I(\hat{X}^N \rightarrow X^N) + \sum_{n=2}^N I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}) \\
&= \sum_{n=1}^N I(\hat{X}^n; X_n | X^{n-1}) + \sum_{n=2}^N I(X^{n-1}; \hat{X}_n | \hat{X}^{n-1}) \\
&= H(X^N) + H(\hat{X}^N) \\
&\quad - \sum_{n=1}^N H(X_n | X^{n-1}, \hat{X}^n) - H(\hat{X}_n | X^{n-1}, \hat{X}^{n-1}) \\
&= H(X^N) + H(\hat{X}^N) - \sum_{n=1}^N H(X_n, \hat{X}_n | X^{n-1}, \hat{X}^{n-1}) \\
&= I(X^N; \hat{X}^N).
\end{aligned} \tag{4}$$

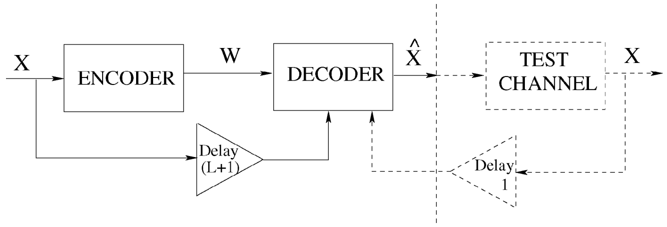

Fig. 4. Backward test channel interpretation.

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 $\hat{X}^N$. At time instant n, since the decoder knows the symbols X^{n-1} to reconstruct $\hat{X}_n$, (3) says we need not spend $I(X^{n-1}; \hat{X}_n|\hat{X}^{n-1})$ 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.

An interesting way to understand any source compression system is to analyze the corresponding backward test channel [5], [26], [27]. 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 N samples of X. The process of reconstruction starts with 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 nonanticipatory 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. In the next clock cycle, it gets X_2. 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 $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N$ as input and $X_1, X_2, \dots, X_N$ as output with a sequence of conditional distributions given by

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

where X^i denotes the vector of $X_1, X_2, \dots, X_i$. 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:

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

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.

III. 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 $\{X_n, \hat{X}_n\}$ isdiscrete, stationary, and ergodic [28]. This assumption is not necessary to prove the rate-distortion theorem for arbitrary sources with feed-forward in Section IV—the purpose is to first give intuition about how feed-forward helps in source coding. This assumption of stationarity and ergodicity leads to a rather simple and intuitive proof of the rate-distortion theorem along the lines of the proof of the rate-distortion theorem for discrete memoryless sources in [5]. 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 [15] is denoted by

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

Similarly, a conditional distribution is denoted by

$$\mathbf{P_{\hat{X}|X}} \triangleq \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}. \tag{6}$$

Finally, for stationary and ergodic joint processes, the directed information rate exists and is defined by [23]

$$I(\hat{X} \to X) = \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N). \tag{7}$$

We use an expected distortion criterion here. For simplicity (only for this section), 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. In fact, since the joint random process $\{X_n, \hat{X}_n\}$ is stationary, $E[d_N(X^N, \hat{X}^N)] = E[d(X, \hat{X})]$.

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

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

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

$$R \ge R^*(D) \triangleq \inf_{\mathbf{P_{\hat{X}|X}}: E[d(X,\hat{X})] \le D} I(\hat{X} \to X)$$

are achievable[2] at expected distortion D.

[2] The infimization is over all conditional distributions such that the joint process $(\mathbf{X}, \hat{\mathbf{X}})$ is stationary and ergodic.

Proof: Since the AEP holds for discrete, stationary and ergodic processes [5], we have

$$-\frac{1}{N}\log P(X^N) \to H(X) \quad \text{w.pr.1}$$
$$-\frac{1}{N}\log P(X^N, \hat{X}^N) \to H(X, \hat{X}) \quad \text{w.pr.1} \tag{8}$$

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, introduced in [19] and [29], respectively. These were used in [24] in the context of channels with feedback. These will be frequently used in the rest of this paper. $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$

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

$$\vec{P}_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \triangleq \prod_{i=1}^{N} P_{X_i|\hat{X}^{i},X^{i-1}}(x_i|\hat{x}^{i},x^{i-1}). \tag{10}$$

These can be pictured in terms of the test channel from $\hat{X}$ to X. Equation (9) describes the sequence of input distributions to this test channel and (10) specifies the test channel. Recall that the joint distribution can be split as

$$P_{X^N,\hat{X}^N}(x^N,\hat{x}^N) = \prod_{i=1}^{N} P_{\hat{X}_i|\hat{X}^{i-1},X^{i-1}}(\hat{x}_i|\hat{x}^{i-1},x^{i-1}) \cdot P_{X_i|\hat{X}^{i},X^{i-1}}(x_i|\hat{x}^{i},x^{i-1}). \tag{11}$$

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 (9) and (10). Let

$$H(\hat{X}^N\|X^N) \triangleq \sum_{i=1}^{N} H(\hat{X}_i|\hat{X}^{i-1},X^i).$$

$H(\hat{X}^N\|X^N)$ is known as the entropy of $\hat{X}^N$ causally conditioned on X^N [23], [25]. We will also use $H(\hat{X}^N\|0X^{N-1})$, the entropy of $\hat{X}^N$ causally conditioned on the delayed X sequence $0X^{N-1} \triangleq [\_, X_1, X_2, \ldots, X_{N-1}]$.

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

$$-\frac{1}{N}\log\vec{P}(\hat{X}^N|X^N) \to \vec{H}(\hat{X}\|X) \quad \text{w.p.1} \tag{12}$$

where

$$\begin{aligned} \vec{H}(\hat{X}\|X) &\triangleq \lim_{N\to\infty}\frac{1}{N}H(\hat{X}^N\|0X^{N-1})\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}H(\hat{X}_i|X^{i-1},\hat{X}^{i-1})\\ &= \lim_{N\to\infty}H(\hat{X}_N|X^{N-1},\hat{X}^{N-1}) \end{aligned} \tag{13}$$

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 [5], [30] and is given in Appendix I. 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

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

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

$$P((X^N,\hat{X}^N) \in \mathcal{A}_\epsilon^N) \to 1 \quad \text{as} \quad N \to \infty. \tag{14}$$

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 $(x^N,\hat{x}^N) \in \mathcal{A}_\epsilon^N$

$$\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) \geq P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot 2^{-N(I(\hat{X}\to X)+3\epsilon)}. \tag{15}$$

Proof:

$$\begin{aligned} &P_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\\ &= \frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{P_{X^N}(x^N)}\\ &= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\frac{P_{X^N,\hat{X}^N}(x^N,\hat{x}^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot P_{X^N}(x^N)}\\ &\leq \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot\frac{2^{-N(H(X,\hat{X})-\epsilon)}}{2^{-N(\vec{H}(\hat{X}\|X)+\epsilon)}\cdot 2^{-N(H(X)+\epsilon)}}\\ &= \vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N)\cdot 2^{N(I(\hat{X}\to X)+3\epsilon)} \end{aligned} \tag{16}$$

from which the lemma follows. The last equality in (16) can be proved as follows:

$$\begin{aligned} &H(\hat{X}^N\|0X^{N-1}) + H(X^N) - H(X^N,\hat{X}^N)\\ &= H(\hat{X}^N\|0X^{N-1}) - H(\hat{X}^N|X^N)\\ &= H(\hat{X}^N) - H(\hat{X}^N|X^N) - [H(\hat{X}^N) - H(\hat{X}^N\|0X^{N-1})]\\ &\overset{(a)}{=} I(X^N;\hat{X}^N) - I(0X^{N-1}\to\hat{X}^N)\\ &\overset{(b)}{=} I(\hat{X}^N\to X^N) \end{aligned} \tag{17}$$

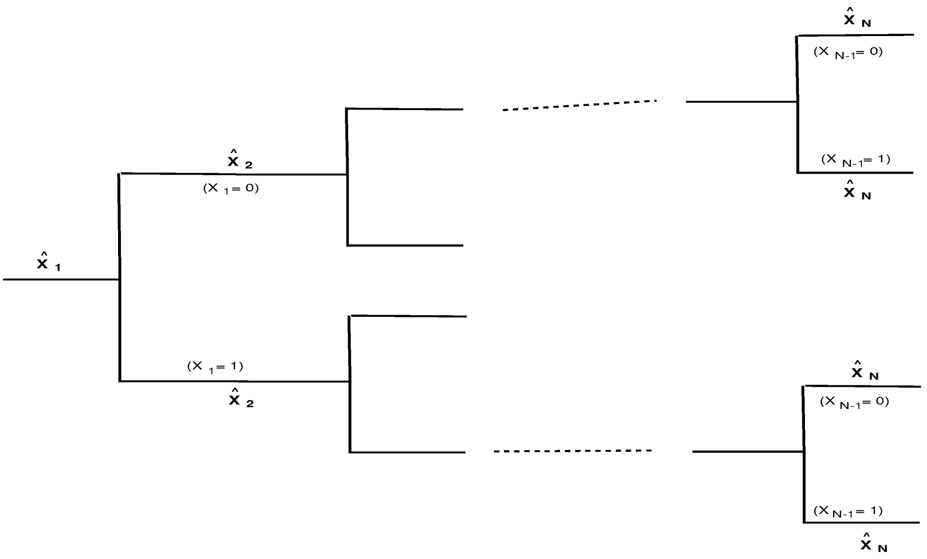

Fig. 5. Code function for a binary source.

where (a) follows by writing the definition of directed information in (2) in terms of entropies and (b) follows from (3). Dividing by N and taking limits we get the result.

We are now ready to prove the achievability of $R^*(D)$.

Codetrees: In source coding with feed-forward, to produce the ith reconstruction symbol $\hat{x}_i$, the decoder knows the first $i-1$ 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. Let the first input symbol be $\hat{x}_1$. 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. For each value of $\hat{x}_2$, we have $|\mathcal{X}|$ choices for $\hat{x}_3$ and so on, thus forming a tree. A code-tree for a system with binary source and reconstruction alphabets is shown in Fig. 5. A rate R source code with feed-forward consists of a codebook of 2^{NR} code-trees. The decoder receives the index of the code-tree chosen by the encoder, traces the path along the code-tree using the fed-forward source symbols and produces the reconstruction. For instance, suppose the code-tree in Fig. 5 is used and the fed-forward sequence, x^{N-1}, is the all zero sequence. The decoder traces the uppermost path on the tree and obtains the reconstruction symbols along that path.

Random Codebook Generation: Pick a joint distribution $\mathbf{P}_{\mathbf{X},\hat{\mathbf{X}}} = \{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$, such that the X–marginal has the distribution $\mathbf{P}_{\mathbf{X}}$ and $E[d(X,\hat{X})] \leq D$. 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, 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 $P_{\hat{X}_2|\hat{X}_1,X_1}(\cdot|\hat{x}_1,x_1)$ 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 $P_{\hat{X}_3|\hat{X}^2,X^2}(\cdot|\hat{x}^2,x^2)$. We continue picking the input symbols in this manner and finally we pick $\hat{x}_N$ according to $P_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\cdot|\hat{x}^{N-1},x^{N-1})$. We obtain 2^{NR} such independent and randomly chosen code-trees in the same fashion to form the codebook.

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.

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

$$E[d_N(X^N,\hat{X}^N)] \leq D+\epsilon+P_e d_{\max}. \tag{18}$$

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

$$E_{\mathcal{C}}[E[d_N(X^N,\hat{X}^N)]] \leq D+\epsilon+\overline{P}_e d_{\max} \tag{19}$$

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, $\overline{P}_e$ 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

$$\overline{P}_e = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \sum_{x^N : x^N \notin J(\mathcal{C})} P(x^N) \tag{20}$$

$$= \sum_{x^N} P(x^N) \sum_{\mathcal{C} : x^N \notin J(\mathcal{C})} \Pr(\mathcal{C}). \tag{21}$$

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

$$\Pr\left((x^N, \hat{X}^N) \notin A_\epsilon^N\right) = 1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N). \tag{22}$$

We emphasize here that we need to use the directed probability $\vec{P}(\hat{x}^N | x^N)$ in (22) because this is the distribution we used to generate the random code. Thus the probability of choosing a codebook that does not well represent x^N is

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

Substituting this in (21), we get

$$\overline{P}_e = \sum_{x^N} P(x^N) \left[1 - \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} \vec{P}(\hat{x}^N | x^N)\right]^{2^{NR}}. \tag{24}$$

We can now use Lemma 3.3 to obtain

$$\overline{P}_e \leq \sum_{x^N} P(x^N) \cdot \left[1 - 2^{-N(I(\hat{X} \to X) + 3\epsilon)} \sum_{\hat{x}^N : (x^N, \hat{x}^N) \in A_\epsilon^N} P(\hat{x}^N | x^N)\right]^{2^{NR}}. \tag{25}$$

As shown in [5], the inequality

$$(1 - xy)^n \leq 1 - y + e^{-xn} \tag{26}$$

holds for $n > 0$ and $0 \leq x, y \leq 1$. Using this in (25), we get

$$\overline{P}_e \leq \sum_{x^N} P(x^N) \sum_{\hat{x}^N : (x^N, \hat{x}^N) \notin A_\epsilon^N} P(\hat{x}^N | x^N) + e^{-2^{N(R - I(\hat{X} \to X) - 3\epsilon)}}$$

$$= \sum_{(x^N, \hat{x}^N) \notin A_\epsilon^N} P(x^N, \hat{x}^N) + e^{-2^{N(R - I(\hat{X} \to X) - 3\epsilon)}}. \tag{27}$$

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. □

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 $I(\hat{X}^N \to X^N) \leq I(\hat{X}^N; X^N)$. 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 than $sup_{\vec{P}_{\mathbf{X}|\mathbf{Y}}} I(X \to Y)$ are achievable with feedback. Compare this with the no-feedback capacity of the channel, given by $sup_{\mathbf{P}_{\mathbf{X}}} I(X; Y)$. It is shown in [13] that when there is no feedback in the channel, $I(X; Y) = I(X \to Y)$. Hence, the no-feedback capacity of the channel can be written as $sup_{\mathbf{P}_{\mathbf{X}}} I(X \to Y)$.

Comparing the expressions for capacity with and without feedback, we see that the objective function $(I(X \to Y))$ is the same; but 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.

IV. General Sources

A. Rate-Distortion Theorem

In this section, we prove the rate-distortion theorem for arbitrary sources with feed-forward. We will use the method of information spectrum introduced by Han and Verdú [15]. This a powerful tool to deal with general processes without making any assumptions. Information spectrum methods have been used to derive formulas for the capacity of general channels with and without feedback [24], [31] and the rate-distortion function of general sources [32]. They have also been used to derive error exponents for both lossless and lossy source coding of general sources [34]–[37].

The apparatus we will use for proving coding theorems for general discrete sources with feed-forward is first described. We define a code-function, which maps the feed-forward information to a source reconstruction symbol $\hat{X}$. These code-functions are the same as the code-trees used in the previous section, but we give a formal definition here. Roughly speaking, a source code with feed-forward is a set of code-functions. The source sequence X^N determines the code-function to be used and the mapping to the reconstruction symbols is done by the decoder using the code-function and fed-forward values.

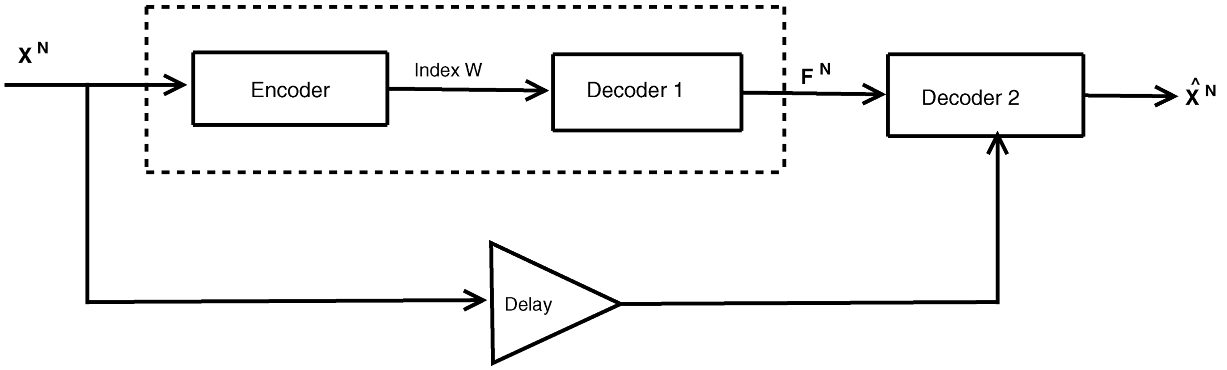

Fig. 6. Representation of a source coding scheme with feed-forward.

Definition 4.1: A source code-function f^N is a set of N functions $\{f_n\}_{n=1}^N$ such that $f_n : \mathcal{X}^{n-1} \rightarrow \widehat{\mathcal{X}}$ 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 $\mathcal{F}^N = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \mathcal{F}_N \triangleq \{f^N : f^N \text{ is a code function}\}$.

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 $f^N[w], \quad w = 1, \ldots, 2^{NR}$.

An encoder is a mapping that maps each source sequence $x^N \in \mathcal{X}^N$ to a code-function in the codebook.

Note that a source code-book and an encoder together automatically define the decoder as follows. 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 Fig. 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 $f^N = \{f_1, \ldots, f_N\}$ where for each $n \in \{1, \ldots, N\}$, the function f_n is a constant mapping.

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 $f^N[i], \quad i \in \{1, \ldots, 2^{NR}\}$ and sends the index of the chosen code function. This is the part inside the dashed box in Fig. 6. If we denote the chosen code-function by f^N, the second component (decoder 2 in Fig. 6) produces the reconstruction given by

$$\hat{X}_i = f_i(X^{i-1}), \qquad i = 1, \ldots, N. \tag{28}$$

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 (28). 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

$$P_{F^N|X^N}(f^N|x^N) = \begin{cases} 1, & \text{if } f^N = \text{the code-function chosen by the encoder.} \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

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

$$P_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N, x^N) = \delta_{\{\hat{x}^N = f^N(x^{N-1})\}}. \tag{30}$$

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

$$\begin{aligned} &Q_{X^N,F^N,\hat{X}^N}(x^N, f^N, \hat{x}^N) \\ &= P_{X^N}(x^N) \cdot P_{F^N|X^N}(f^N|x^N) \cdot P_{\hat{X}^N|F^N,X^N}(\hat{x}^N|f^N, x^N) \\ &= P_{X^N}(x^N) \cdot \delta_{\{f^N = e(x^N)\}} \cdot \delta_{\{\hat{x}^N = f^N(x^{N-1})\}} \end{aligned} \tag{31}$$

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 [32] 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 $d_n(\cdot,\cdot)$ 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 and an associated encoder 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 state the definitions of a few quantities (previously defined in [31], [24]) 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 $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N$

$$i(x^N; \hat{x}^N) \triangleq \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{\hat{X}^N}(\hat{x}^N) P_{X^N}(x^N)} \tag{32}$$

$$\overline{H}(X) \triangleq \limsup_{inprob} \frac{1}{N} \log \frac{1}{P_{X^N}(X^N)} \tag{33}$$

$$\underline{H}(X) \triangleq \liminf_{inprob} \frac{1}{N} \log \frac{1}{P_{X^N}(X^N)} \tag{34}$$

$$\vec{i}(\hat{x}^N; x^N) \triangleq \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{\vec{P}_{\hat{X}^N|X^N}(\hat{x}^N|x^N) P_{X^N}(x^N)} \tag{35}$$

$$\overline{I}(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N} \vec{i}(\hat{X}^N; X^N) \tag{36}$$

$$\underline{I}(\hat{X} \to X) \triangleq \liminf_{inprob} \frac{1}{N} \vec{i}(\hat{X}^N; X^N) \tag{37}$$

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 (9) and (10), respectively.

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

$$I(\hat{X}^N \to X^N) = \sum_{x^N,\hat{x}^N} P_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \vec{i}(\hat{x}^N; x^N). \tag{38}$$

As pointed out in [32], 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 and the 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 [15]).

Lemma 4.1: [24] For any sequence of joint distributions $\{P_{X^n,\hat{X}^n}\}_{n=1}^{\infty}$, we have

$$\begin{aligned} \underline{I}(\hat{X} \to X) &\leq \liminf_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \\ &\leq \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \leq \overline{I}(\hat{X} \to X). \end{aligned} \tag{39}$$

If

$$\underline{I}(\hat{X} \to X) = \overline{I}(\hat{X} \to X) \tag{40}$$

then the limit exists and all the quantities in (39) are equal. The class of processes for which this equality holds includes (but is not limited to) stationary and ergodic joint processes. We are now ready to state and prove the rate distortion theorem for an arbitrary source with feed-forward. In [31], 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 [32] 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 [24] 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

$$R^*_{\text{ff}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D} \overline{I}(\hat{X} \to X) \tag{41}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{n\to\infty} E[d_n(X^n, \hat{X}^n)]. \tag{42}$$

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

$$R_{\text{ff}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \leq D} \overline{I}(\hat{X} \to X) \tag{43}$$

where

$$\begin{aligned} \rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) &\triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) \\ &= \inf\Big\{h : \lim_{n\to\infty} P_{X^n} P_{\hat{X}^n|X^n}\left((x^n, \hat{x}^n)\right. \\ &\qquad \left. : d_n(x^n, \hat{x}^n) > h\right) = 0\Big\}. \end{aligned} \tag{44}$$

Note that if the joint process $\{X_n, \hat{X}_n\}_{n=1}^{\infty}$ satisfies (40), from Lemma 4.1, the rate-distortion function becomes

$$R_{\rm ff}(D) = \inf \lim_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \tag{45}$$

where the infimum is evaluated according to the distortion constraint used. Although the rate-distortion function given by Theorem 2 involves optimizing a multi-letter expression involving X and $\hat{X}$, in a future paper we will show that this can be evaluated in closed form for several classes of sources and distortion measures with memory [33].

The detailed proofs of the direct and converse parts of Theorem 2 are found in Appendices II and III, respectively. The proof of the direct part uses the machinery introduced in [24] for proving the capacity results for channels with feedback. 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 that (45) holds. We want to show the achievability of all rates greater than $R_{\rm ff}(D)$ in (45).

Let $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}} = \{P^*_{\hat{X}^n|X^n}\}$ 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 Fig. 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

$$Q_{X^N,\hat{X}^N} = P_{X^N} P^*_{\hat{X}^N|X^N}. \tag{46}$$

We also impose certain additional constraints on $Q_{F^N,X^N,\hat{X}^N}$ so that[3]

$$I_Q(F^N; X^N) = I_Q(\hat{X}^N \to X^N). \tag{47}$$

Using (46) in the above equation, we get

$$I_Q(F^N; X^N) = I_{P_{X^N} P^*_{\hat{X}^N|X^N}}(\hat{X}^N \to X^N). \tag{48}$$

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 (48), 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.

It should be remarked here that to prove Theorem 2, we do not use the concept of directed distortion typicality introduced in Section III. Notions of typicality are useful only for stationary and ergodic processes. However, when the joint process $\{X_n, \hat{X}_n\}$ is stationary and ergodic, Theorem 2 (a) gives the same rate-distortion function as Theorem 1. The reason for the discussion in Section III was to give intuition about source coding with feed-forward before going into full generality.

B. 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.

Proof: See Appendix IV.

This result was shown in [4] for the sources that were identically distributed, in addition to being memoryless. 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 $d_N(x^N, \hat{x}^N)$.

C. Gaussian Sources With Feed-Forward

In this section, we study the 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 i.i.d. 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 [7].

An explicit characterization of the distortion-rate function for a stationary Gaussian source with feed-forward was given in [4] for an average mean-squared error distortion criterion. Here we consider arbitrary Gaussian sources and prove a result on the structure of the optimum achieving conditional distribution for any quadratic distortion criterion. As in the case of discrete memoryless sources, we use the expected distortion constraint. We now show that for a Gaussian source, $R^*_{\rm ff}(D)$ is achieved by a Gaussian conditional distribution.

Proposition 4.1: Let X be an arbitrary Gaussian source with distribution $\mathbf{P_X}$. Then the optimal rate-distortion function with feed-forward with a quadratic distortion measure is achieved by a Gaussian conditional distribution.

Proof: Suppose the conditional distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} = \{P_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ achieves the optimal rate-distortion function. Let $\mathbf{G}_{\hat{\mathbf{X}}|\mathbf{X}} = \{G_{\hat{X}^n|X^n}\}_{n=1}^{\infty}$ be a Gaussian conditional distribution such that for all N

$$G_{X^N,\hat{X}^N} = P_{X^N} \cdot G_{\hat{X}^N|X^N}$$

is 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 we will show the following:

1) $I_G(\hat{X}^N \to X^N) \le I_P(\hat{X}^N \to X^N)$;

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

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

$$E_P[d_N(X^N,\hat{X}^N)] = E_G[d_N(X^N,\hat{X}^N)]. \tag{49}$$

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

$$\begin{aligned} p_{X^N,\hat{X}^N} &= p_{X^N} p_{\hat{X}^N|X^N} \\ g_{X^N,\hat{X}^N} &= p_{X^N} g_{\hat{X}^N|X^N}. \end{aligned}$$

Using the representation of directed information given in (38), we have the following chain of inequalities:

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

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. Dropping the subscript from $p_{X^N,\hat{X}^N}$ and continuing the chain, we have

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

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 nonnegative.

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 a quadratic distortion measure, the optimizing conditional distribution can be taken to be jointly Gaussian. We also have the following result from [24] for jointly Gaussian distributions. For any jointly Gaussian distribution $\mathbf{P}_{\mathbf{X^N},\hat{\mathbf{X}}^\mathbf{N}} = \{P_{X^N,\hat{X}^N}\}_{n=1}^{\infty}$

$$\overline{I}(\hat{X} \to X) = \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N). \tag{50}$$

This property follows from the asymptotic equipartition property, which is valid for an arbitrary Gaussian random processes (Theorem 5, [38]). Thus the rate-distortion function for an arbitrary Gaussian source with expected mean-squared error distortion criterion can be written as

$$R_{\mathrm{ff}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\le D} \limsup_{N\to\infty} \frac{1}{N} I(\hat{X}^N \to X^N) \tag{51}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) = \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2] \tag{52}$$

and $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ can be taken to be Gaussian.

V. Error Exponents

In this section we study the error exponent for source coding with feed-forward. The error-exponent for lossy, fixed-length source coding of a stationary memoryless source (without feed-forward) with a single-letter distortion measure was derived by Marton [6] and by Blahut [39]. Recently, Iriyama derived the error exponent for lossy, fixed-length coding of a general source without feed-forward with a general distortion measure [37]. For lossless source coding, the reliability function for fixed-length coding of a general source was first studied in [34] and then in [35], [36]. Error exponents for lossy/lossless coding of certain classes of discrete sources were earlier studied in [40]–[44].

The error exponent for fixed-length lossy source coding with feed-forward is derived in [4] for sources that can be auto-regressively represented with an i.i.d. innovations process and is shown to be the same as Marton's no-feed-forward error exponent [6]. In this section, we will use the approach and framework of [37] to obtain a formula for the error exponent for fixed-length lossy coding of any general source with feed-forward.

For a source-code with feed-forward of block-length N, let $\epsilon_N(D)$ be the probability of the distortion exceeding D.

$$\epsilon_N(D) = \Pr\left\{ d_N(X^N,\hat{X}^N) > D \right\}. \tag{53}$$

We want to determine the infimum of all achievable coding rates such that asymptotically $\epsilon_N(D) \sim e^{-Nr} \quad (N \to \infty)$. This is called the minimum (D,r)-achievable rate with feed-forward for the source $\mathbf{X}$ and is denoted $R_{\mathrm{ff}}(D,r|\mathbf{X})$. We will derive a formula for this in the following.

A. Performance of "Good" Source Codes With Feed-Forward

In this section, we will determine the minimum (D,r)-achievable rate for the source $\mathbf{X}$ with feed-forward. Defining the problem formally, consider a sequence of $\{(N, 2^{NR_N})\}_{N=1}^{\infty}$ source codes with feed-forward. Each code in this sequence is defined according to Definition 2.1. We are interested in a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward such that

$$\limsup_{N\to\infty} R_N \le R$$

and

$$\liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{\epsilon_N(D)} \ge r. \tag{54}$$

Definition 5.1: The minimum (D,r)-achievable rate for the source $\mathbf{X}$ with feed-forward is defined as

$$R_{\text{ff}}(D,r|\mathbf{X}) = \inf\{R : \exists \text{ sequence of codes satisfying } (54)\}. \tag{55}$$

The minimum (D,r)-achievable rate will be expressed in terms of a rate-distortion function with feed-forward. This rate-distortion function is defined according to a distortion constraint that is different from those considered in Theorem 2. This is described as follows. Consider a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward satisfying

$$\limsup_{N\to\infty} R_N \le R$$

and

$$\limsup_{N\to\infty} (1-\epsilon_N(D)) > 0. \tag{56}$$

In other words, we are interested in a sequence of codes with rate R. Further, the probability of correct decoding should be nonzero for infinitely many codes in this sequence. We will need the rate-distortion function with feed-forward defined according to this criterion.

Definition 5.2: The rate-distortion function $R^*_{\text{ff}}(D|\mathbf{X})$ for the source $\mathbf{X}$ with feed-forward is defined as

$$R^*_{\text{ff}}(D|\mathbf{X}) = \inf\{R : \exists \text{ sequence of codes satisfying } (56)\} \tag{57}$$

Finally, we will need a couple of divergence quantities to express the minimum (D,r)-achievable rate. We have $D_u(\mathbf{Y}\|\mathbf{X})$ and $D_l(\mathbf{Y}\|\mathbf{X})$ defined by

$$\begin{aligned} D_u(\mathbf{Y}\|\mathbf{X}) &= \limsup_{n\to\infty} \frac{1}{n} D(Y^n\|X^n) \\ D_l(\mathbf{Y}\|\mathbf{X}) &= \liminf_{n\to\infty} \frac{1}{n} D(Y^n\|X^n). \end{aligned} \tag{58}$$

We can now state our result.

Theorem 4: For any $D, r > 0$

$$\begin{aligned} \sup_{\mathbf{Y}:D_l(\mathbf{Y}\|\mathbf{X})<r} R^*_{\text{ff}}(D|\mathbf{Y}) &\le R_{\text{ff}}(D,r|\mathbf{X}) \\ &\le \sup_{\mathbf{Y}:D_l(\mathbf{Y}\|\mathbf{X})\le r} R^*_{\text{ff}}(D|\mathbf{Y}) \end{aligned} \tag{59}$$

with equalities if $R_{\text{ff}}(D,r|\mathbf{X})$ is continuous at r. Further

$$\begin{aligned} \inf_{\hat{\mathbf{Y}}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D} \underline{I}(\hat{\mathbf{Y}}\to\mathbf{Y}) &\le R^*_{\text{ff}}(D|\mathbf{Y}) \\ &\le \inf_{\hat{\mathbf{Y}}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D_1} \underline{I}(\hat{\mathbf{Y}}\to\mathbf{Y}) \end{aligned} \tag{60}$$

for any $0 < D_1 < D$, with equalities if continuous at D. In the above, $\overline{D}(\mathbf{Y},\hat{\mathbf{Y}}) = \limsup_{\text{inprob}} d_n(y^n, \hat{y}^n)$.

Proof: In Appendix V.

Let us examine the case when $R_{\text{ff}}(D,r|\mathbf{X})$ is continuous. Then the minimum (D,r)-achievable rate can be expressed as

$$R_{\text{ff}}(D,r|\mathbf{X}) = \sup_{\mathbf{Y}:D_l(\mathbf{Y}\|\mathbf{X})\le r} R^*_{\text{ff}}(D|\mathbf{Y}). \tag{61}$$

This can be pictured in a manner analogous to the interpretation of the error exponent for stationary memoryless sources using the type-covering lemma [26], [45]. Loosely speaking, for the error decay with exponent r, we need the code to cover all sequences belonging to source distributions that are at a distance within r from the 'true' distribution $\mathbf{P_X}$. This is possible if we build a code with rate given by (61).

We observe that the minimum (D,r) achievable rate increases with r. As we should expect, we also see that it approaches the feed-forward rate-distortion function of $\mathbf{X}$ as r approaches 0.

From (61), it is also clear that the minimum (D,r)-achievable rate for a source with feed-forward is smaller than for the same source without feed-forward. Without feed-forward, the formula is the supremum of the no-feed-forward rate-distortion function $R^*(D|\mathbf{Y})$ which is clearly greater than the corresponding feed-forward rate-distortion function $R^*_{\text{ff}}(D|\mathbf{Y})$.

B. Performance of "Bad" Source Codes With Feed-Forward

If the coding rate is sufficiently small, then the probability $\epsilon_N(D)$ tends to one. Similar to [37], we can study the performance of bad feed-forward codes. In this section, we will determine the minimum coding rate $R^*_{\text{ff}}(D,r|\mathbf{X})$ for which the probability of distortion being less than or equal to D goes to zero exponentially fast with exponent r. We are interested in a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward such that

$$\limsup_{N\to\infty} R_N \le R$$

and

$$\liminf_{N\to\infty} \frac{1}{N} \log \frac{1}{1-\epsilon_N(D)} \le r. \tag{62}$$

We define a minimum achievable rate with feed-forward $R^*_{\text{ff}}(D,r|\mathbf{X})$

$$R^*_{\text{ff}}(D,r|\mathbf{X}) = \inf\{R : \exists \text{ a sequence of codes satisfying } (62)\}. \tag{63}$$

We will express $R^*_{\text{ff}}(D,r|\mathbf{X})$ in terms of the rate-distortion function defined as follows. Consider a sequence of $\{(N, 2^{NR_N})\}$ codes with feed-forward satisfying

$$\limsup_{N\to\infty} R_N \le R$$

and

$$\limsup_{N\to\infty} \epsilon_N(D) = 0. \tag{64}$$

This condition is similar to (but not the same as) the probability-1 distortion constraint. For a source $\mathbf{Y}$, we define

$$R_{\text{ff}}(D|\mathbf{Y}) = \inf\{R : \exists \text{ a sequence of codes satisfying (64)}\}. \tag{65}$$

We are now ready to state our result in terms of $R_{\text{ff}}(D|\mathbf{Y})$.

Theorem 5: For any $D, r > 0$

$$R^*_{\text{ff}}(D, r|\mathbf{X}) = \inf_{\mathbf{Y}: D_u(\mathbf{Y}\|\mathbf{X}) \le r} R_{\text{ff}}(D|\mathbf{Y}). \tag{66}$$

Further

$$\begin{aligned}\inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}}) \le D} \overline{I}(\hat{\mathbf{Y}} \to \mathbf{Y}) &\le R_{\text{ff}}(D|\mathbf{Y}) \\ &\le \inf_{\hat{\mathbf{Y}}: \overline{D}(\mathbf{Y},\hat{\mathbf{Y}}) \le D_1} \overline{I}(\hat{\mathbf{Y}} \to \mathbf{Y})\end{aligned} \tag{67}$$

for $0 < D_1 < D$, with equalities if continuous at D.

The proof of this theorem is found in Appendix V along with the proof of Theorem 4.

VI. Feed-Forward With Arbitrary Delay

Recall from the discussion in Section I 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. We call this delay k feed-forward. 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 $g_i : \{1, \ldots, 2^{NR}\} \times \mathcal{X}^{i-k} \to \hat{\mathcal{X}}, \quad i = 1, \ldots, N$.

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

$$I(\hat{X}^N \to X^N) = I(\hat{X}^N; X^N) - \sum_{i=2}^{N} I(X^{i-1}; \hat{X}_i|\hat{X}^{i-1}). \tag{68}$$

With delay k feed-forward, 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

$$\begin{aligned}I_k(\hat{X}^N \to X^N) &\triangleq I(\hat{X}^N; X^N) - \sum_{i=k+1}^{N} I(X^{i-k}; \hat{X}_i|\hat{X}^{i-1}) \\ &= I(\hat{X}^N; X^N) - I(0^k X^{N-k} \to \hat{X}^N)\end{aligned} \tag{69}$$

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

Observing (69), 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, we get some information for free and the rate-distortion function is reduced by a quantity equal to the 'free information'. characterize the capacity of channels with feedback delay $k \ge 1$. Similar intuition can be used to understand the rate-distortion function when there is side-information available with some delay at both encoder and decoder, in addition to delayed feed-forward [46].

We now state the rate-distortion theorem for feed-forward with general delay. We omit the proof since it is similar to the ones in the preceding sections.

Definition 6.1:

$$\vec{P}_k(\hat{X}^N|X^N) \triangleq \prod_{i=1}^{N} P(\hat{X}_i|\hat{X}^{i-1}, X^{i-k}) \tag{70}$$

$$\begin{aligned}I_k(\hat{X}^N \to X^N) &\triangleq I(\hat{X}^N; X^N) - \sum_{i=k+1}^{N} I(X^{i-k}; \hat{X}_i|\hat{X}^{i-1}) \\ &= \sum_{x^n, \hat{x}^N} P_{X^N,\hat{X}^N}(x^N, \hat{x}^N) \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{X^N}(x^N)\vec{P}_k(\hat{x}^N|x^N)}\end{aligned} \tag{71}$$

$$\overline{I}_k(\hat{X} \to X) \triangleq \limsup_{inprob} \frac{1}{N} \log \frac{P_{X^N,\hat{X}^N}(x^N, \hat{x}^N)}{P_{x^N}(x^N)\vec{P}_k(\hat{x}^N|x^N)}. \tag{72}$$

Theorem 6 (Rate-Distortion Theorem):

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

$$R^*_{\text{ff}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \le D} \overline{I}_k(\hat{X} \to X) \tag{73}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{n\to\infty} E[d_n(X^n, \hat{X}^n)]. \tag{74}$$

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

$$R_{\text{ff}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}} : \rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \le D} \overline{I}_k(\hat{X} \to X) \tag{75}$$

where

$$\begin{aligned}\rho(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) &\triangleq \limsup_{inprob} d_n(x^n, \hat{x}^n) \\ &= \inf\Big\{h : \lim_{n\to\infty} P_{X^n}P_{\hat{X}^n|X^n}\big((x^n, \hat{x}^n) \\ &\qquad : d_n(x^n, \hat{x}^n) > h\big) = 0\Big\}.\end{aligned} \tag{76}$$

We can also extend the error exponent results of Theorems 4 and 5 to feed-forward with arbitrary delay. As a final remark, we note that when $k \to \infty$, the problem becomes source coding without feed-forward. As we would expect, the delay k feed-forward rate-distortion function given by Theorem 6 then reduces to the no-feed-forward rate distortion function $\inf \overline{I}(\hat{X}; X)$ as $k \to \infty$.

VII. Conclusion

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. This problem was first considered in [4] where the distortion-rate function was characterized for a class of sources. We have derived the optimal rate-distortion function for a general source with feed-forward. We also characterized the error exponent for a general source with feed-forward. Specifically, for a source to be encoded with distortion D, we found the minimum rate at which the probability of error decays with exponent r. 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. In a forthcoming paper, we demonstrate that the rate-distortion function with feed-forward can be evaluated in closed-form for several classes of sources and distortion measures with memory [33]. Extensions to accommodate practical constraints such as a noisy feed-forward path are part of future work.

Appendix I
Proof of Lemma 3.1 (AEP)

The proof is similar to that of the Shannon–McMillan Breiman Theorem in [5], [30]. We first state the definitions and three lemmas required for the proof. Recall that

$$\vec{P}(\hat{x}^N|x^N) = \prod_{i=1}^{N} P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}),$$
$$\vec{P}(x^N|\hat{x}^N) = \prod_{i=1}^{N} P(x_i|\hat{x}^{i}, x^{i-1}).$$

We want to show that

$$-\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) \to \vec{H}(\hat{X}\|X), \text{ where}$$
$$\vec{H}(\hat{X}\|X) \triangleq \lim_{N\to\infty} H(\hat{X}_N|X^{N-1}, \hat{X}^{N-1}). \quad \text{(A1)}$$

Definition 1.1: Let

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

Lemma 1.1:

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

Proof: We have

$$\begin{aligned}-\frac{1}{N}&\log \vec{P}^k(\hat{X}^N|X^N)\\ &= -\frac{1}{N}\vec{P}(\hat{X}^k|X^k) - \frac{1}{N}\sum_{i=k+1}^{N}\log P(\hat{X}_i|\hat{X}_{i-k}^{i-1}, X_{i-k}^{i-1})\\ &\to 0 + H^k \quad \text{by the ergodic theorem.} \qquad \text{(A2)}\end{aligned}$$

We also have

$$\begin{aligned}-\frac{1}{N}&\log \vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0)\\ &= -\frac{1}{N}\sum_{i=1}^{N}\log P(\hat{X}_i|, \hat{X}_{-\infty}^{i-1}, X_{-\infty}^{i-1})\\ &\to \vec{H}^{\infty}(\hat{X}\|X) \text{ by the ergodic theorem.} \quad \text{(A3)}\end{aligned}$$

□

Lemma 1.2:

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

Proof: We know that $H^k \to \vec{H}(\hat{X}\|X)$, since the joint process is stationary and $\{H_k\}$ is a nonincreasing sequence of nonnegative numbers. So we only need to show that $H^k \to \vec{H}^{\infty}(\hat{X}\|X)$. 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, we have the first equation shown at the bottom of the next page. Thus $H^k \to \vec{H}^{\infty}(\hat{X}\|X)$. □

Lemma 1.3:

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

where

$$\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0) \triangleq \prod_{i=1}^N P(\hat{X}_i|X_{-\infty}^{i-1}, \hat{X}_{-\infty}^{i-1}). \tag{A4}$$

Proof:

$$\begin{aligned} E&\left[\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)}\right] \\ &= \sum_{\hat{x}^N, x^N} P(\hat{x}^N, x^N) \\ &\quad \cdot \frac{\prod_{i=1}^k P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) \prod_{i=k+1}^N P(\hat{x}_i|\hat{x}_{i-k}^{i-1}, x_{i-k}^{i-1})}{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1})} \\ &= \sum_{\hat{x}^N, x^N} P(\hat{x}^k, x^k) \\ &\quad \cdot \prod_{i=k+1}^N P(x_i|x^{i-1}, \hat{x}^i) P(\hat{x}_i|\hat{x}_{i-k}^{i-1}, x_{i-k}^{i-1}) = 1 \end{aligned} \tag{A5}$$

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

$$\Pr\left\{\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq N^2\right\} \leq \frac{1}{N^2} \tag{A6}$$

or

$$\Pr\left\{\frac{1}{N}\log\frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \geq \frac{1}{N}\log N^2\right\} \leq \frac{1}{N^2}. \tag{A7}$$

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 (A8), shown at the bottom of the page. The inner expectation can be written as (A9) shown at the bottom of the page. In (A9), 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

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

We are now ready to prove the AEP.

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

$$E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0)}\right] = E_{\hat{X}_{-\infty}^0, X_{-\infty}^0}\left[E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0)}\middle|\hat{X}_{-\infty}^0, X_{-\infty}^0\right]\right]. \tag{A8}$$

$$\begin{aligned} E\left[\frac{\vec{P}(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X_{-\infty}^N, \hat{X}_{-\infty}^0)}\middle|\hat{X}_{-\infty}^0, X_{-\infty}^0\right] &= \sum_{\hat{x}^N, x^N} P(\hat{x}^N, x^N|\hat{X}_{-\infty}^0, X_{-\infty}^0) \frac{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1})}{\prod_{i=1}^N P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}, \hat{X}_{-\infty}^0, X_{-\infty}^0)} \\ &= \sum_{\hat{x}^N, x^N} \prod_{i=1}^N P(x_i|\hat{x}^i, x^{i-1}, \hat{X}_{-\infty}^0, X_{-\infty}^0) P(\hat{x}_i|\hat{x}^{i-1}, x^{i-1}) = 1. \end{aligned} \tag{A9}$$

Proof: [**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 $\vec{H}^{\infty}(\hat{X}\|X)$ for all $k \geq 0$. From Lemma 1.3, we have

$$\limsup_{N\to\infty} \frac{1}{N}\log \frac{\vec{P}^k(\hat{X}^N|X^N)}{\vec{P}(\hat{X}^N|X^N)} \leq 0. \tag{A10}$$

Since the limit $\frac{1}{N}\log \vec{P}^k(\hat{X}^N|X^N)$ exists (Lemma 1.1), we can write (A10) as

$$\limsup_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \leq \lim_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}^k(\hat{X}^N|X^N)} = H^k. \tag{A11}$$

The second part of Lemma 1.3 can be written as

$$\liminf_{N\to\infty} \frac{1}{N}\log \frac{\vec{P}(\hat{X}^N|X_{-\infty}^{N},\hat{X}_{-\infty}^{0})}{\vec{P}(\hat{X}^N|X^N)} \geq 0. \tag{A12}$$

Since the limit $\frac{1}{N}\log \vec{P}(\hat{X}^N|X_{-\infty}^{N},\hat{X}_{-\infty}^{0})$ exists (Lemma 1.1), we can rewrite (A12) as

$$\liminf_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \geq \lim_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}(\hat{X}^N|X_{-\infty}^{N},\hat{X}_{-\infty}^{0})} = \vec{H}^{\infty}(\hat{X}\|X). \tag{A13}$$

Combining (A11) and (A13), we have

$$\begin{aligned}\vec{H}^{\infty}(\hat{X}\|X) &\leq \liminf_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}(\hat{X}^N|X^N)}\\ &\leq \limsup_{N\to\infty} \frac{1}{N}\log \frac{1}{\vec{P}(\hat{X}^N|X^N)} \leq H^k \text{ for all } k.\end{aligned} \tag{A14}$$

By Lemma 1.2, $H^k \to \vec{H}(\hat{X}\|X) = \vec{H}^{\infty}(\hat{X}\|X)$. Thus

$$\lim_{N\to\infty} -\frac{1}{N}\log \vec{P}(\hat{X}^N|X^N) = \vec{H}(\hat{X}\|X). \tag{A15}$$

□

Appendix II
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 Fig. 6 (Section IV), 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 $R_{\text{ff}}(D)$.

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. In the second part, we will construct random codes for the system without feed-forward and show the achievability of $R_{\text{ff}}(D)$ 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. In this part, we wish 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

$$Q^*_{X^N,\hat{X}^N} = P_{X^N}P^*_{\hat{X}^N|X^N}. \tag{B1}$$

To do this, as will be shown in the sequel, the only distribution we can choose is the code-function distribution P_{F^N}. We pick P_{F^N} such that the induced distribution $Q^*_{F^N,X^N,\hat{X}^N}$ has certain desired properties and (B1) is also satisfied.

For any N, the joint distribution $P_{X^N}P^*_{\hat{X}^N|X^N}$ can be split, as in (11), as

$$P_{X^N}P^*_{\hat{X}^N|X^N} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|\hat{X}^n,X^{n-1}} \tag{B2}$$

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 test-channel, respectively.

Let P_{F^N} be any distribution on the space of code-functions. Given P_{F^N} and the test channel P_{ch} in (B2), 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.

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

$$Q_{\hat{X}_n|F_n,X^{n-1}}(\hat{x}_n|f_n,x^{n-1}) = \begin{cases} 1, & \text{if } \hat{x}_n = f_n(x^{n-1}) \\ 0, & \text{otherwise.}\end{cases} \tag{B3}$$

2)

$$Q_{F_n|F^{n-1},X^{n-1}}(f_n|f^{n-1},x^{n-1}) = P_{F_n|F^{n-1}}(f_n|f^{n-1}) \quad n = 1, \ldots, N. \tag{B4}$$

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

$$Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n,\hat{x}^n,x^{n-1}) = P^{ch}_{X_n|\hat{X}^n,X^{n-1}}(x_n|\hat{x}^n,x^{n-1}),\ n = 1, \ldots, N. \tag{B5}$$

A joint distribution 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 three constraints above hold. It is important to note that in general, 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 encoder-decoder pair does not satisfy these conditions. We just want to construct a 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 $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^{N}$, 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:

$$\begin{aligned} &Q_{F^N,X^N,\hat{X}^N} \\ &= \left\{ \prod_{n=1}^{N} Q_{X_n|F^n,X^{n-1}} \cdot Q_{F_n|F^{n-1},X^{n-1}} \right\} \cdot Q_{\hat{X}^N|F^N,X^N} \\ &= \left\{ \prod_{n=1}^{N} Q_{X_n|F^n,X^{n-1}} \cdot P_{F_n|F^{n-1}} \right\} \cdot \delta_{\hat{X}^N=F^N(X^{N-1})} \end{aligned} \quad \text{(B6)}$$

where we have used (B3) and (B4) to obtain the second equality. Now we can use the fact that $\hat{x}_n = f_n(x^{n-1})$ to write

$$\begin{aligned} &Q_{X_n|F^n,X^{n-1}}(x_n|f^n, x^{n-1}) \\ &= Q_{X_n|F^n,\hat{X}^n,X^{n-1}}(x_n|f^n, \hat{x}^n, x^{n-1}) \\ &= P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\left(x_n|x^{n-1}, f^n(x^{n-1})\right) \end{aligned} \quad \text{(B7)}$$

where we have used (B5) for the second equality. Thus, the unique nice joint distribution is given by (B8) shown at the bottom of the page. Keeping P^{ch} fixed, (B8) says that choosing P_{F^N} automatically 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

$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \quad \text{(B9)}$$

so that (B1) is satisfied.

Definition 2.1: 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 $Q_{F^N,X^N,\hat{X}^N}$:

$$\begin{aligned} &\prod_{n=1}^{N} Q_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1}, x^{n-1}) \\ &= \prod_{n=1}^{N} \qquad P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1}, x^{n-1}) \\ &\qquad x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \hat{\mathcal{X}}^N. \end{aligned} \quad \text{(B10)}$$

This definition of 'good' is equivalent to, but slightly different from that in [24]. 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 [24] in a different flavor.

Lemma 2.1: For a test-channel $\{P^{ch}_{X_n|X^{n-1},\hat{X}^n}\}_{n=1}^{N}$, there exists a code-function distribution P_{F^N} good with respect to a decoder distribution $\{P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}\}_{n=1}^{N}$.

Proof: For all f^N and $n = 1, \ldots, N$, we define two quantities given by (B11) and (B12) shown at the bottom of the page. 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 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 of the proof uses this expression to show that (B12) defines a good code-function distribution. Actually, we first need to show that for all n, $P_{F_n|F^{n-1}}$ defined above is a valid probability distribution. This part is omitted since it can be shown using arguments similar to those in Part B

Part A: Define $\forall n \in \{1, \ldots, N\}$

$$\Gamma_n(x^{n-1}, \hat{x}_n) \triangleq \{f_n : f_n(x^{n-1}) = \hat{x}_n\} \quad \text{(B13)}$$

$$\Gamma^n(x^{n-1}, \hat{x}^n) \triangleq \{f^n : f_i(x^{i-1}) = \hat{x}_i,\ i = 1, \ldots, n\}. \quad \text{(B14)}$$

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

$$\begin{aligned} &Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) \\ &= P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1})\right), \\ &\quad n = 1, \ldots, N. \end{aligned} \quad \text{(B15)}$$

This is Lemma 5.2 in [24], 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

$$\begin{aligned} &Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1}, \hat{x}^{n-1}) \\ &\qquad = Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}). \end{aligned} \quad \text{(B16)}$$

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

$$\begin{aligned} &Q_{\hat{X}_n|F^{n-1},X^{n-1}}(\hat{x}_n|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}) \\ &= Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1}, \hat{x}_n)|\Gamma^{n-1}(x^{n-2}, \hat{x}^{n-1}), x^{n-1}). \end{aligned} \quad \text{(B17)}$$

$$Q_{F^N,X^N,\hat{X}^N}(f^N, x^N, \hat{x}^N) = \prod_{n=1}^{N} P_{F_n|F^{n-1}}(f_n|f^{n-1}) \cdot \prod_{n=1}^{N} P^{ch}_{X_n|X^{n-1},\hat{X}^{n-1}}(x_n|f^n(x^{n-1}), x^{n-1}) \cdot \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}. \quad \text{(B8)}$$

$$\text{graph}(f_n) \triangleq \{(x^{n-1}, \hat{x}_n) : f_n(x^{n-1}) = \hat{x}_n\} \subset \mathcal{X}^{n-1} \times \hat{\mathcal{X}}, \quad \text{(B11)}$$

$$P_{F_n|F^{n-1}}(f_n|f^{n-1}) \triangleq \prod_{(b^{n-1},a_n)\in\text{graph}(f_n)} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(a_n|f_1, \ldots, f_{n-1}(b^{n-2}), b^{n-1}). \quad \text{(B12)}$$

Since Q is nice, it satisfies (B4). Hence

$$\begin{aligned}&Q_{F_n|F^{n-1},X^{n-1}}(\Gamma_n(x^{n-1},x_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1}),x^{n-1})\\ &\quad = P_{F_N|F^{N-1}}(\Gamma_n(x^{n-1},\hat{x}_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1})).\end{aligned} \tag{B18}$$

Combining (B16), (B17) and (B18), we obtain the expression in (B15).

Part B: We now show that P_{F^N} defined by (B12) is good with respect to P^{dec}. For a pair $x^{N-1} \in \mathcal{X}^{N-1}, \hat{x}^N \in \mathcal{X}^N$, consider

$$\begin{aligned}&\textstyle\sum_{f^N:f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N)\\ &\quad = P_{F^N}(\Gamma^N(x^{N-1},\hat{x}^N))\\ &\quad = P_{F^N}\left(\Gamma_1(\hat{x}_1),\dots,\Gamma_n(x^{n-1},\hat{x}_n),\dots,\Gamma_N(x^{N-1},\hat{x}_N)\right)\\ &\quad = \prod_{n=1}^{N} P_{F_n|F^{n-1}}\left(\Gamma_n(x^{n-1},\hat{x}_n)|\Gamma^{n-1}(x^{n-2},\hat{x}^{n-1})\right).\end{aligned} \tag{B19}$$

Substituting (B15) in the above equation, we get

$$\begin{aligned}&\sum_{f^N:f^N(x^{N-1})=\hat{x}^N} P_{F^N}(f^N)\\ &\qquad = \prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1}).\end{aligned} \tag{B20}$$

We can also write the LHS of (B19) as in (B21) shown at the bottom of the page.

We evaluate the Nth inner summation in (B21) as shown in (B22) at the bottom of the page. This equation is explained as follows. In (B22), $f_1,\dots,f_{N-1}$ are specified by the $N-1$ outer summations and "gr" has been used as shorthand for graph. (a) in (B22) follows from (B12) 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 (B21) 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 (B21) sequentially from $n=N$ down to $n=1$. Substituting in (B21), we get

$$\begin{aligned}&\sum_{\substack{f^N:\\ f^N(x^{N-1})=\hat{x}^N}} P_{F^N}(f^N)\\ &\qquad = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1}).\end{aligned} \tag{B23}$$

From (B20) and (B23), we have

$$\begin{aligned}&\textstyle\prod_{n=1}^{N} Q_{\hat{X}_n|X^{n-1},\hat{X}^{n-1}}(\hat{x}_n|x^{n-1},\hat{x}^{n-1})\\ &\quad = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(\hat{x}_n|\hat{x}^{n-1},x^{n-1})\\ &\quad n=1,\dots,N.\end{aligned} \tag{B24}$$

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 (B8).
- For a test-channel $\{P^{ch}_{X_n|\hat{X}^n,X^{n-1}}\}_{n=1}^{N}$, we can find a code function distribution $P^*_{F^N}$ to be good with respect to P^{dec},

$$\begin{aligned}\sum_{\substack{f^N:\\ f^N(x^{N-1})=\hat{x}^N}} P_{F^N}(f^N) &= \sum_{f_1:f_1=\hat{x}_1} \cdots \sum_{\substack{f_n:\\ f_n(x^{n-1})=\hat{x}_n}} \cdots \sum_{\substack{f_N:\\ f_N(x^{N-1})=\hat{x}_N}} \prod_{n=1}^{N} P_{F_n|F^{n-1}}(f_n|f^{n-1})\\ &= \sum_{f_1:f_1=\hat{x}_1} P_{F_1}(f_1)\dots \sum_{\substack{f_n:\\ f_n(x^{n-1})=\hat{x}_n}} P_{F_n|F^{n-1}}(f_n|f^{n-1})\dots \sum_{\substack{f_N:\\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N|f^{N-1}).\end{aligned} \tag{B21}$$

$$\begin{aligned}&\textstyle\sum_{\substack{f_N:\\ f_N(x^{N-1})=\hat{x}_N}} P_{F_N|F^{N-1}}(f_N|f^{N-1}) \overset{(a)}{=} \sum_{\substack{f_N:\\ f_N(x^{N-1})=\hat{x}_N}} \prod_{(b^{N-1},a_N)\in\mathrm{gr}(f_N)} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\dots,f_{N-1}(b^{N-2}),b^{N-1})\\ &\quad = P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}) \cdot \sum_{\substack{f_N:\\ f_N(x^{N-1})=\hat{x}_N}} \prod_{\substack{(b^{N-1},a_N)\in\mathrm{gr}(f_N)\\ b^{N-1}\neq x^{N-1}}} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\dots,f_{N-1}(b^{N-2}),b^{N-1})\\ &\quad \overset{(b)}{=} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}) \prod_{b^{N-1}\neq x^{N-1}} \sum_{a_N} P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(a_N|f_1,\dots,f_{N-1}(b^{N-2}),b^{N-1})\\ &\quad = P^{dec}_{\hat{X}_N|\hat{X}^{N-1},X^{N-1}}(\hat{x}_N|\hat{x}^{N-1},x^{N-1}).\end{aligned} \tag{B22}$$

i.e., the set of induced "decoder" distributions of Q^* satisfying the relation

$$\prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} = \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \tag{B25}$$

for $n = 1, \ldots, N$. Hence, we have

$$\begin{aligned} Q^*_{X^N,\hat{X}^N} &= \prod_{n=1}^{N} Q^*_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot Q^*_{X_n|X^{n-1},\hat{X}^n} \\ &= \prod_{n=1}^{N} P^{dec}_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}} \cdot P^{ch}_{X_n|X^{n-1},\hat{X}^n} \\ &= P_{X^N} \cdot P^*_{\hat{X}^N|X^N}. \end{aligned} \tag{B26}$$

Equation (B26) 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

$$\begin{aligned} &A(\mathcal{C}_N) \\ &\quad = \left\{x^N \in \mathcal{X}^N : \exists f^N \in \mathcal{C}_N \; with \; d'_N(x^N, f^N) \le D + \delta\right\}. \end{aligned} \tag{B27}$$

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 $R > R_{\text{ff}}(D)$. Indeed

$$\begin{aligned} &E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] \\ &\quad = \sum_{\mathcal{C}_N} P^*_{F^N}(\mathcal{C}_N) \sum_{x^N \notin A(\mathcal{C}_N)} P_{X^N}(x^N) \\ &\quad = \sum_{x^N} P_{X^N}(x^N) \sum_{\mathcal{C}_N : x^N \notin A(\mathcal{C}_N)} P^*_{F^N}(\mathcal{C}_N). \end{aligned} \tag{B28}$$

The last sum on the right-hand side of (B28) is the probability of choosing a codebook that does not represent the particular x^N with a distortion $D + \delta$. Define the set $B_{N,\delta}$ given by (B29) shown at the bottom of the page. In (B29), $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is as in Theorem 2, and $\overline{I}(\hat{X} \to X)$ is computed with the distribution $\mathbf{P}_{\mathbf{X}}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ and is therefore equal to $R_{\text{ff}}(D)$. Also define an indicator function

$$K(x^N, f^N) = \begin{cases} 1, & \text{if } (x^N, f^N) \in B_{N,\delta} \\ 0, & \text{otherwise.} \end{cases} \tag{B30}$$

We will also need the following Lemma, whose proof is given on the following page.

Lemma 2.2:

a)

$$\begin{aligned} Q^*_{F^N|X^N}(f^N|x^N) \le P^*_{F^N}(f^N) \exp_2[N(R_{\text{ff}}(D) + \delta)] \\ \forall (x^N, f^N) \in B_{N,\delta}. \end{aligned}$$

b)

$$Q^*_{X^N,F^N}(B_{N,\delta}) \to 1 \text{ as } N \to \infty.$$

Since $\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}$ achieves $R_{\text{ff}}(D)$ 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

$$\begin{aligned} &P^*_{F^N}\left(d'_N(x^N, F^N) \ge D + \delta\right) \\ &\quad \le P^*_{F^N}\left(K(x^N, F^N) = 0\right) \\ &\quad = 1 - \sum_{f^N} P^*_{F^N}(f^n) K(x^N, f^N). \end{aligned} \tag{B31}$$

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

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

Using this in (B28), we obtain (B32) shown at the top of the next page. In (B28), the last inequality follows from part a) of Lemma 1.5. Using the inequality

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

in (B32), we get (B33) shown at the top of the next page.When $R > R_{\text{ff}}(D) + \delta$, using part b) of Lemma 1.5, we have

$$\lim_{N \to \infty} E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] = 0. \tag{B34}$$

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

$$\begin{aligned} \lim_{N \to \infty} \Pr\left(x^N \in \mathcal{X}^N : d_N(x^N, f^N(x^{N-1})) > D + \delta, \; \forall f^N \in \mathcal{C}_N\right) \\ = 0. \end{aligned} \tag{B35}$$

The theorem follows.

$$B_{N,\delta} = \left\{(x^N, f^N) : \quad d'_N(x^N, f^N) < \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta, \quad \frac{1}{N} i_{Q^*}(x^N; f^N) < \overline{I}_{\mathbf{P}_{\mathbf{X}}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X} \to X) + \delta\right\}. \tag{B29}$$

$$
\begin{aligned}
E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &\leq \sum_{x^N} P_{X^N}(x^N)\left(1 - \sum_{f^N} P^*_{F^N}(f^n)K(x^N, f^N)\right)^{2^{NR}} \\
&\leq \sum_{x^N} P_{X^N}(x^N)\left(1 - \exp_2\left\{-N(R_{\text{ff}}(D)+\delta)\right\}\sum_{f^N} Q^*_{F^N|X^N}(f^N|x^N)K(x^N, f^N)\right)^{2^{NR}}. \qquad \text{(B32)}
\end{aligned}
$$

$$
\begin{aligned}
E\left[P_{X^N}(A^c(\mathcal{C}_N))\right] &\leq 1 + \exp_2\left[-\exp_2[N(R - R_{\text{ff}}(D) - \delta)]\right] - \sum_{x^N, f^N} P_{X^N}(x^N)Q^*_{F^N|X^N}(f^N|x^N)K(x^N, f^N) \\
&= 1 - Q^*_{F^N,X^N}(B_{N,\delta}) + \exp_2\Big(-\exp_2[N(R - R_{\text{ff}}(D) - \delta)]\Big). \qquad \text{(B33)}
\end{aligned}
$$

Proof of Lemma 2.2:

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

$$
\begin{aligned}
&Q^*_{F^N|X^N}(f^N|x^N) \\
&= Q^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N; f^N)] \\
&= P^*_{F^N}(f^N)\exp_2[i_{Q^*_{X^N,F^N}}(x^N; f^N)] \qquad \text{(B36)}
\end{aligned}
$$

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

$$
\frac{1}{N} i_{Q^*}(x^N; f^N) < \overline{I}_{\mathbf{P_X P^*_{\hat{X}|X}}}(\hat{X} \to X) + \delta, \forall (x^n, f^N) \in B_{N,\delta}. \qquad \text{(B37)}
$$

Substituting the above in (B36), 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 (B8). Under these conditions Q^* satisfies

$$
\begin{aligned}
\frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} &= \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^N}}{Q^*_{X^N}} \\
&\overset{a}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},F^n}}{Q^*_{X^N}} \\
&\overset{b}{=} \frac{\prod_{n=1}^N Q^*_{X_n|X^{n-1},\hat{X}^n}}{Q^*_{X^N}} \\
&= \frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}} \qquad \text{(B38)}
\end{aligned}
$$

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 the condition $Q_{X_n|X^{n-1},F^N} = Q_{X_n|X^{n-1},F^n}$ is equivalent to (B4). This is shown in [25] as a condition for a channel not to have have feedback. (b) follows from (B4) and (B5). (B38) is essentially Lemma 5.1 in [24]. Thus, we have

$$
\begin{aligned}
i_{Q^*_{X^N,F^N}}(f^N; x^N) &= \frac{1}{N}\log \frac{Q^*_{F^N,X^N}}{Q^*_{F^N}Q^*_{X^N}} \\
&= \frac{1}{N}\log \frac{Q^*_{X^N,\hat{X}^N}}{\vec{Q}^*_{\hat{X}^N|X^N}Q^*_{X^N}} \\
&= \vec{i}_{Q^*_{\hat{X}^N,X^N}}(\hat{x}^N; x^N). \qquad \text{(B39)}
\end{aligned}
$$

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 (B26)

$$
\begin{aligned}
Q^*_{X^N,\hat{X}^N} &= \vec{Q^*}_{\hat{X}^N|X^N}\vec{Q^*}_{X^N|\hat{X}^N} = \vec{P}^{dec}_{\hat{X}^N|X^N}\vec{P}^{ch}_{X^N|\hat{X}^N} \\
&= P_{X^N}P^*_{\hat{X}^N|X^N}. \qquad \text{(B40)}
\end{aligned}
$$

Using (B40) in (B39), we get

$$
i_{Q^*_{X^N,F^N}}(f^N; x^N) = \vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N; x^N). \qquad \text{(B41)}
$$

Next, we express the probability of the set $B^c_{N,\delta}$ as given in (B42) shown at the top of the next page. Since

$$
d'_N(x^N, f^N) = d_N(x^N, f^N(x^{N-1})) = d_N(x^N, \hat{x}^N) \qquad \text{(B43)}
$$

the first term in (B42) equals

$$
\begin{aligned}
&Q^*_{X^N,\hat{X}^N}\Big((x^N, \hat{x}^N) : d_N(x^N, \hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) \\
&= P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N, \hat{x}^N) : d_N(x^N, \hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}) + \delta\Big) \qquad \text{(B44)}
\end{aligned}
$$

$$\begin{aligned} Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta}) &= Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N):\ d'_N(x^N,f^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta \\ &\qquad \text{or } \frac{1}{N} i_{Q^*}(x^N;f^N) \geq \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\Big) \\ &\leq Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N):\quad d'_N(x^N,f^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta\Big) \\ &\quad + Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N):\quad \frac{1}{N} i_{Q^*}(x^N;f^N) \geq \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\Big). \end{aligned} \tag{B42}$$

where we have used (B40). Since $\rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})$ is the $\lim\sup$ in probability of $d_N(x^N,\hat{x}^N)$

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N,\hat{x}^N): d_N(x^N,\hat{x}^N) \geq \rho(\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}})+\delta\Big) = 0. \tag{B45}$$

Using (B41) and (B40), the second term in (B42) can be written as (B46) shown at the bottom of the page. Since $\overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X} \to X)$ is the $\lim\sup$ in probability of $\frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(\hat{x}^N;x^N)$, we obtain the limit as in (B47), shown at the bottom of the page. Equations (B45) and (B47) imply

$$\lim_{N\to\infty} Q^*_{F^N,X^N,\hat{X}^N}(B^c_{N,\delta}) = 0 \tag{B48}$$

proving part (b) of the lemma. □

Appendix III
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. $\forall x^N \in \mathcal{X}^N, \hat{x}^N \in \hat{\mathcal{X}}^N, f^N \in \{f^N[i], i=1,\ldots,2^{NR}\}$, the induced distribution is given by

$$\begin{aligned} &\hat{Q}_{X^N,F^N,\hat{X}^N}(x^N,f^N,\hat{x}^N) \\ &\quad = P_{X^N}(x^N)\cdot P_{F^N|X^N}(f^N|x^N)\cdot \delta_{\{\hat{x}^N=f^N(x^{N-1})\}}. \end{aligned} \tag{C1}$$

All probability distributions in the remainder of this section are marginals drawn from the induced joint distribution in (C1). We first show that for any such induced distribution, we have

$$\overline{H}(F) = \limsup_{inprob} \frac{1}{N}\log\frac{1}{P(F^N)} \leq R. \tag{C2}$$

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

$$\lim_{N\to\infty} \Pr\left(\frac{1}{N}\log\frac{1}{P(F^N)} > R+\delta\right) = 0. \tag{C3}$$

We have

$$\begin{aligned} &\Pr\left(\frac{1}{N}\log\frac{1}{P(F^N)} > R+\delta\right) \\ &= \Pr\left(P(F^N) < 2^{-N(R+\delta)}\right) \\ &= \sum_{f^N:0<P_{F^N}(f^N)<2^{-N(R+\delta)}} P_{F^N}(f^N) \\ &\leq \sum_{f^N:P_{F^N}(f^N)>0} 2^{-N(R+\delta)} \\ &= 2^{NR}\cdot 2^{-N(R+\delta)} \\ &= 2^{-N\delta} \to 0 \quad \text{as} \quad N\to\infty \end{aligned} \tag{C4}$$

thereby proving (C2). Thus, we have

$$R \geq \overline{H}(F) \geq \overline{H}(F) - \underline{H}(F|X) \geq \overline{I}(F;X) \tag{C5}$$

where the last inequality follows from Lemma 2 in [32]. We need the following lemma, whose proof is given subseequently.

Lemma 3.1: For any sequence of codes as defined above, we have

$$\overline{I}(F;X) \geq \overline{I}(\hat{X}\to X) \tag{C6}$$

$$\begin{aligned} &Q^*_{F^N,X^N,\hat{X}^N}\Big((f^N,x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\Big) \\ &\qquad = P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\Big). \end{aligned} \tag{B46}$$

$$\lim_{N\to\infty} P_{X^N}P^*_{\hat{X}^N|X^N}\Big((x^N,\hat{x}^N): \frac{1}{N}\vec{i}_{P_{X^N}P^*_{\hat{X}^N|X^N}}(x^N;\hat{x}^N) \geq \overline{I}_{\mathbf{P_X}\mathbf{P}^*_{\hat{\mathbf{X}}|\mathbf{X}}}(\hat{X}\to X)+\delta\Big) = 0. \tag{B47}$$

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

Using this lemma in (C5), we obtain

$$R \geq \overline{I}(\hat{X} \to X). \tag{C7}$$

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

$$R \geq \overline{I}(\hat{X} \to X) \geq R_{\text{ff}}(D). \tag{C8}$$

Proof of Lemma 3.1: Let $\hat{Q}_{X^N,F^N,\hat{X}^N}$ be the joint distribution induced by the source code as in (C1). From Definition 4.5, we have

$$\begin{aligned} i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N) &= \log \frac{P(X^N|F^N)}{\vec{P}(X^N|\hat{X}^N)} \\ &= \log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} \end{aligned} \tag{C9}$$

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

$$\liminf_{inprob} \frac{1}{N} \left(i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N) \right) \geq 0. \tag{C10}$$

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

$$\lim_{N\to\infty} P\left(\frac{1}{N} \log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta \right) = 0. \tag{C11}$$

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

$$\begin{aligned} P(X^N|F^N) &= \prod_{n=1}^N P(X_n|X^{n-1},F^N) \\ &= \prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n). \end{aligned} \tag{C12}$$

Hence the quantity in (C11) can be expressed as in (C13) shown at the bottom of the page. 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.

Next, we write the series of equations (C14) shown at the bottom of the page. In (C14), (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 (C13) becomes

$$\begin{aligned} P\left(\frac{1}{N} \log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta \right) \\ = \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N) < 2^{-N\delta}. \end{aligned} \tag{C15}$$

$$\begin{aligned} P\left(\frac{1}{N} \log \frac{P(X^N|F^N)}{\prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n)} < -\delta \right) &= P\left(\prod_{n=1}^N P(X_n|X^{n-1},F^N,\hat{X}^n) < 2^{-N\delta} \prod_{n=1}^N P(X_n|X^{n-1},\hat{X}^n) \right) \\ &= \sum_{(f^N,x^N,\hat{x}^N)\in\mathcal{G}} \hat{Q}(f^N,x^N,\hat{x}^N). \end{aligned} \tag{C13}$$

where

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

$$\begin{aligned} \sum_{\mathcal{G}} \hat{Q}_{F^N,X^N,\hat{X}^N}(f^N,x^N,\hat{x}^N) &= \sum_{\mathcal{G}} P(f^N)P(x^N,\hat{x}^N|f^N) = \sum_{\mathcal{G}} P(f^N) \prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n,f^N)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &< 2^{-N\delta} \sum_{\mathcal{G}} P(f^N) \prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\leq 2^{-N\delta} \sum_{x^N,f^N,\hat{x}^N} P(f^N) \prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\overset{(a)}{=} 2^{-N\delta} \sum_{f^N} P(f^N) \sum_{(x^N,\hat{x}^N):f^N(x^{N-1})=\hat{x}^N} \prod_{n=1}^N P(x_n|x^{n-1},\hat{x}^n)P(\hat{x}_n|x^{n-1},\hat{x}^{n-1},f^N) \\ &\overset{(b)}{=} 2^{-N\delta} \sum_{f^N} P(f^N) \sum_{x^N} \prod_{n=1}^N P(x_n|x^{n-1},f^n(x^{n-1})) \overset{(c)}{=} 2^{-N\delta} \cdot 1. \end{aligned} \tag{C14}$$

Hence

$$\lim_{N\to\infty} P\left(\frac{1}{N}\log\frac{P(X^N|F^N)}{\prod_{n=1}^{N} P(X_n|X^{n-1},\hat{X}^n)} < -\delta\right) = 0. \tag{C16}$$

Thus we have proved (C10). Now, using the inequality

$$\liminf_{inprob}(a_n+b_n) \leq \limsup_{inprob} a_n + \liminf_{inprob} b_n \tag{C17}$$

in (C10), we get

$$\begin{aligned} 0 &\leq \liminf_{inprob}\frac{1}{N}\left(i(F^N;X^N) - \vec{i}(\hat{X}^N;X^N)\right) \\ &\leq \limsup_{inprob}\frac{1}{N}i(F^N;X^N) + \liminf_{inprob} -\frac{1}{N}\vec{i}(\hat{X}^N;X^N) \\ &= \limsup_{inprob}\frac{1}{N}i(F^N;X^N) - \limsup_{inprob}\frac{1}{N}\vec{i}(\hat{X}^N;X^N). \end{aligned} \tag{C18}$$

Or

$$\overline{I}(F;X) \geq \overline{I}(\hat{X}\to X) \tag{C19}$$

completing the proof of the lemma.

Appendix IV
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

$$R_{\mathrm{DMS}}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\leq D} \overline{I}(\hat{X};X), \tag{D1}$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) \triangleq \limsup_{N\to\infty} E[\frac{1}{N}\sum_{i=1}^{N} d_i(X_i,\hat{X}_i)]. \tag{D2}$$

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

$$P_{X^n,\hat{X}^n}^{ML} = \prod_{i=1}^{n} P_{X_i,\hat{X}_i}, \tag{D3}$$

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

$$E_{P_{X^N,\hat{X}^N}}[\frac{1}{N}\sum_{i=1}^{N} d_i(X_i,\hat{X}_i)] = \frac{1}{N}\sum_{i=1}^{N} E_{P_{X_i,\hat{X}_i}} d_i(X_i,\hat{X}_i) \tag{D4}$$

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

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

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

To prove that

$$\limsup_{\text{inprob}} a_n \geq \limsup_{\text{inprob}} b_n, \tag{D5}$$

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

$$\begin{aligned} 0 \leq \liminf_{\text{inprob}} a_n - b_n &\leq \limsup_{\text{inprob}} a_n + \liminf_{\text{inprob}} -b_n \\ &= \limsup_{\text{inprob}} a_n - \limsup_{\text{inprob}} b_n. \end{aligned} \tag{D6}$$

We have

$$\begin{aligned} &\frac{1}{N}\left(i(\hat{X}^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) \\ &= \frac{1}{N}\log\frac{P(\hat{X}^N,X^N)}{P(\hat{X}^N)\prod_{i=1}^{N}P(X_i)}\cdot\prod_{i=1}^{N}\frac{P(X_i)P(\hat{X}_i)}{P(X_i,\hat{X}_i)} \\ &= \frac{1}{N}\log\frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^{N}P(X_i|\hat{X}_i)}. \end{aligned} \tag{D7}$$

We want to show that the $\liminf_{inprob}$ of the expression in (D7) is ≥ 0. This is equivalent to showing that for any $\delta > 0$

$$\lim_{N\to\infty}\Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) < -\delta\right] = 0. \tag{D8}$$

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,

$$\begin{aligned} &\Pr\left[\frac{1}{N}\left(i(\hat{X}^N;X^N) - i_{ML}(\hat{X}^N;X^N)\right) < -\delta\right] \\ &= \Pr\left[\frac{1}{N}\log\frac{P(X^N|\hat{X}^N)}{\prod_{i=1}^{N}P(X_i|\hat{X}_i)} < -\delta\right] \\ &= \Pr\left[P(X^N|\hat{X}^N) < 2^{-N\delta}\prod_{i=1}^{N}P(X_i|\hat{X}_i)\right] \\ &= \sum_{(x^N,\hat{x}^N)\in\mathcal{G}} P_{\hat{X}^N}(\hat{x}^N)P_{X^N|\hat{X}^N}(x^N|\hat{x}^N) \\ &\overset{(a)}{\leq} 2^{-N\delta}\sum_{(x^N,\hat{x}^N)\in\mathcal{G}}\prod_{i=1}^{N}P_{\hat{X}_i|X^{i-1}}(\hat{x}_i|x^{i-1})P_{X_i|\hat{X}_i}(x_i|\hat{x}_i) \\ &\overset{(b)}{=} 2^{-N\delta}\cdot 1 \end{aligned} \tag{D9}$$

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 (D5) and (D6) 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 (C1) is induced

$$\hat{Q}_{X^N,F^N,\hat{X}^N} = P_{X^N} \cdot P_{F^N|X^N} \cdot \delta_{\{\hat{X}^N = F^N(X^{N-1})\}}. \quad \text{(D10)}$$

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 III ((C2) to (C5)), for this joint distribution we have

$$R \geq \overline{H}(F) \geq \overline{H}(F) - \underline{H}(F|X) \geq \overline{I}(F;X). \quad \text{(D11)}$$

It remains to show that when the source is memoryless

$$\overline{I}(F;X) \geq \overline{I}_{ML}(\hat{X};X) \quad \text{or}$$
$$\limsup_{inprob} \frac{1}{N} i(F^N;X^N) \geq \limsup_{inprob} \frac{1}{N} i_{ML}(\hat{X}^N;X^N). \quad \text{(D12)}$$

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

$$\lim_{N\to\infty} \Pr\left[\frac{1}{N}\Big(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N)\Big) < -\delta\right] = 0. \quad \text{(D13)}$$

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

$$\begin{aligned} &\frac{1}{N}\Big(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N)\Big) \\ &= \frac{1}{N} \log \frac{\hat{Q}(F^N,X^N)}{\hat{Q}(F^N)\prod_{n=1}^{N} P(X_n)} \cdot \prod_{n=1}^{N} \frac{P(X_n)\hat{Q}(\hat{X}_i)}{\hat{Q}(X_n,\hat{X}_n)} \\ &= \frac{1}{N} \log \frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^{N} \hat{Q}(X_n|\hat{X}_n)}. \end{aligned} \quad \text{(D14)}$$

Hence, we can write the series of equations (D15) at the bottom of the page, 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 (D16) shown at the bottom of the page, where the inner summation is computed first over x_N, then x_{N-1} and so on up to x_1. Thus

$$\Pr\left[\frac{1}{N}\Big(i(F^N;X^N) - i(\hat{X}^N;X^N) < -\delta\Big)\right] \leq 2^{-N\delta} \to 0, \quad \text{as } N \to \infty \quad \text{(D17)}$$

proving (D13). 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.

$$\begin{aligned} &\Pr\left[\frac{1}{N}\Big(i(F^N;X^N) - i_{ML}(\hat{X}^N;X^N)\Big) < -\delta\right] = \Pr\left[\frac{1}{N}\log\frac{\hat{Q}(X^N|F^N)}{\prod_{n=1}^{N}\hat{Q}(X_n|\hat{X}_n)} < -\delta\right] = \Pr\left[\hat{Q}(X^N|F^N) < 2^{-N\delta}\prod_{n=1}^{N}\hat{Q}(X_n|\hat{X}_n)\right] \\ &= \Pr\left[(x^N,f^N,\hat{x}^N) : \hat{Q}_{X^N|F^N}(x^N|f^N) < 2^{-N\delta}\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right] \\ &= \sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\hat{Q}_{X^N|F^N}(x^N|f^N)\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) \\ &\leq 2^{-N\delta}\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) \end{aligned} \quad \text{(D15)}$$

$$\begin{aligned} &\sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{(x^N,\hat{x}^N)\in\nu(f^N)}\left[\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|\hat{x}_n)\right]\hat{Q}_{\hat{X}^N|X^N,F^N}(\hat{x}^N|f^N,x^N) \\ &\qquad = \sum_{f^N}\hat{Q}_{F^N}(f^N)\sum_{x^N}\prod_{n=1}^{N}\hat{Q}_{X_n|\hat{X}_n}(x_n|f_n(x^{n-1})) = 1 \end{aligned} \quad \text{(D16)}$$

Appendix V
Proof of Theorems 4 and 5

We first give the proof of Theorem 4. We will use the error-exponent result proved by Iriyama for a general source without feed-forward. For source coding without feed-forward in [37, Th. 1] gives the formula for the minimum achievable rate with error exponent r

$$\sup_{\mathbf{Y}:D_l(\mathbf{Y}\|\mathbf{X})<r} R^*(D|\mathbf{Y}) \le R(D,r|\mathbf{X}) \le \sup_{\mathbf{Y}:D_l(\mathbf{Y}\|\mathbf{X})\le r} R^*(D|\mathbf{Y}) \tag{E1}$$

with equalities if $R(D,r|\mathbf{X})$ is continuous at r. In (E1), the quantities have the same definitions as those in Section V, except that there is no feed-forward.

Recall from Section IV-A that every source coding system with feed-forward is equivalent to a source coding system without feed-forward defined in terms of code-functions. For the no-feed-forward version, the reconstruction is a code-function F^N and the distortion is given by

$$d_n(X^n, F^n) = d_n(X^n, F^n(X^{n-1})), \quad \forall n.$$

Hence, (E1) holds for the source coding problem with source X and reconstruction F. Clearly, any rate-distortion function for the no-feed-forward $X - F$ system is the same as the rate-distortion function for the system $X - \hat{X}$ with feed-forward. Thus we obtain (59).

To prove the second part of Theorem 4, we use Theorem 5 from [37]. Applying this theorem to the $X - F$ source coding system (no feed-forward), we obtain

$$\inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D} \underline{I}(F;\mathbf{Y}) \le R^*_{\mathrm{ff}}(D|\mathbf{Y}) \le \inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D_1} \underline{I}(F;\mathbf{Y}) \qquad 0 < D_1 < D. \tag{E2}$$

We can use the same procedure used in Appendix II to prove the direct part of Theorem 2 to show that

$$\inf_{\mathbf{F}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D} \underline{I}(\mathbf{F};\mathbf{Y}) = \inf_{\hat{\mathbf{Y}}:\overline{D}(\mathbf{Y},\hat{\mathbf{Y}})\le D} \underline{I}(\hat{\mathbf{Y}} \to \mathbf{Y})$$

completing the proof.

Theorem 5 can be proved in a similar fashion, using code-functions and appealing to Theorems 2 and 4 in [37].

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers for their constructive comments and suggestions which led to an improved manuscript. The authors would also like to thank Sekhar Tatikonda of Yale University, New Haven, CT, for inspiring discussions and his comments.

References

[1] J. M. Kahn, R. H. Katz, and K. S. J. Pister, "Mobile networking for smart dust," in *Proc. ACM/IEEE Int. Conf. Mobile Comput.*, Seattle, WA, Aug. 1999.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 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. Inf. Theory*, Jun. 2003, p. 81.

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

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

[6] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-20, pp. 197–199, Mar. 1974.

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

[8] S. S. Pradhan, "On the role of feedforward in Gaussian sources: Point-to-point source coding and multiple description source coding," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 331–349, 2007.

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

[10] S. I. Krich, "Coding for a delay-dependent fidelity criterion," *IEEE Trans. Inf. Theory*, pp. 77–85, Jan. 1974.

[11] D. Neuhoff and R. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, pp. 701–713, Sep. 1982.

[12] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, pp. 1437–1451, 1979.

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

[14] R. Gray, D. Neuhoff, and J. Omura, "Process definitions of distortion-rate functions and source coding theorems," *IEEE Trans. Inf. Theory*, vol. 21, pp. 524–532, Sep. 1975.

[15] T. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, pp. 752–772, May 1993.

[16] T. Cover and M. Chiang, "Duality between channel capacity and rate-distortion with two-sided state information," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1629–1638, Jun. 2002.

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

[18] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1159–1180, May 2003.

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

[20] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. Amer. Stat. Assoc.*, vol. 77, pp. 304–313, Jun. 1982.

[21] P. E. Caines and C. W. Chan, "Feedback between stationary processes," *IEEE Trans. Autom. Contr.*, vol. AC-20, no. 378, pp. 498–508, 1975.

[22] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series," *IEEE Trans. Inf. Theory*, vol. IT-33, pp. 598–601, Jul. 1987.

[23] G. Kramer, "Directed Information for Channels with Feedback," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, Switzerland, 1998.

[24] S. Tatikonda, "Control Under Communications Constraints," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.

[25] J. L. Massey, "Network information theory- some tentative definitions," in *Proc. DIMACS Workshop Network Inf. Theory*, Apr. 2003.

[26] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[27] S. S. Pradhan, "Approximation of test channels in source coding," in *Proc. Conf. Inform. Syst. Sci. (CISS)*, Mar. 2004.

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

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

[30] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *The Ann. Probab.*, vol. 16, pp. 899–909, Apr. 1988.

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

[32] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 43, pp. 63–86, Jan. 1996.

[33] R. Venkataramanan and S. S. Pradhan, "On evaluating the rate-distortion function of sources with feed-forward and channels with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, 2007.

[34] T. S. Han, "The reliability functions of the general source with fixed-length coding," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2117–2132, 2000.

[35] K. Iriyama, "Probability of error for the fixed-length source coding of general sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 2466–2473, 2001.

[36] K. Iriyama and S. Ihara, "The error exponent and minimum achievable rates for the fixed-length coding of general sources," *IEICE Trans. Fund. Electron., Commun. Comput. Sci.*, vol. E84-A, no. 10, pp. 1537–1543, 2001.

[37] K. Iriyama, "Probability of error for the fixed-length lossy coding of general sources," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1498–1507, 2005.

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

[39] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison Wesley, 1988.

[40] J. K. Omura, "A coding theorem for discrete time sources," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 490–498, Jul. 1973.

[41] G. Longo, "On the error exponent for Markov sources," in *Proc. 2nd IEEE Int. Symp. Inf. Theory (ISIT)*, Budapest, Hungary, 1971.

[42] K. Vasek, "On the error exponent for ergodic Markov sources," *Kybernetica*, vol. 16, no. 3, pp. 318–329, 1980.

[43] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Trans. Inf. Theory*, vol. IT-31, pp. 360–365, May 1985.

[44] V. Ananthram, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. IT-4, pp. 938–943, Jul. 1990.

[45] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[46] R. Venkataramanan and S. S. Pradhan, "Directed information for communication problems with side-information and feedback/feed-forward," in *Proceedings of the 43rd Annual Allerton Conference*, Monticello, IL, 2005.