Source Coding with Feed-Forward

Ranji Venkataramanan
Dept. of EECS
Univ. of Michigan
1301 Beal Avenue
Ann Arbor, Michigan 48109-2122
e-mail: rvenkat@eecs.umich.edu

S. Sandeep Pradhan
Dept. of EECS
Univ. of Michigan
1301 Beal Avenue
Ann Arbor, Michigan 48109-2122
e-mail: pradhanv@eecs.umich.edu

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. The special cases of discrete memoryless sources and Gaussian sources with feed-forward are further examined. We also derive a random coding error exponent which is used to bound the probability of decoding error for a source code (with feed-forward) of finite block length. The results are then extended to feed-forward with an arbitrary delay larger than the block length.

I. INTRODUCTION

With the recent emergence of applications involving sensor networks [1], the problem of source coding with side-information at the decoder [2] has gained special significance. Here, the encoder represents the source with an index based on the knowledge that the decoder has access to some correlated side-information. In a typical setting, at each instant of time, the source produces a symbol X and a sample of the side-information Y appears at the decoder. The time sequence of events at the (encoder, decoder) would look like $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$. We are interested in considering a variant of this problem, where there is a delay in the side-information available at the decoder. For instance, if the delay is 3 time units, the sequence of events at the (encoder, decoder) would be $(X_1, \ldots, X_4, Y_1), (X_5, Y_2), (X_6, Y_3), (X_7, Y_4)$ and so on. We would like to analyze this problem of source coding with delayed side-information.

Frequently the side information Y is a noisy version of X. Thus, we would expect that $Y_1$ be strongly correlated with $X_1, Y_2$ with $X_2$ and so on. Such a model would be relevant in applications involving estimation of an information field (e.g. a seismic/acoustic signal) in a sensor network. A node may have to estimate (compressed) signals received from other nodes and process these signals in real-time. However, the signal to be estimated might be available at the node in a delayed and perhaps, noisy form, i.e., there is a feed-forward path from the source to the decoder. Thus an efficient decoder must take into account all the information available while decoding a particular sample.

In this work, we consider an idealized version of this problem called source coding with feed-forward [3]. In this model, we assume that noiseless source samples are available with a delay at the decoder, i.e. $Y = X$. This is meaningful only when the feed-forward delay is at least $N+1$, where the block length is $N$. However, for a general Y, any delay leads to a valid problem.

Related Work: The problem of source coding with noiseless feed-forward arose in the context of competitive prediction in [4]. In that paper, it was shown that for IID discrete sources, feed-forward does not reduce the optimal rate-distortion function and the optimal error-exponent with block coding. Around the same time, the model of source coding with feed-forward was defined in [8] and a simple, deterministic coding scheme to achieve the rate-distortion bound for an IID Gaussian source with feed-forward was described. At the time of writing this paper, we also became aware of another related work [5], which gives a coding strategy to achieve the rate-distortion bound for any finite-alphabet, IID source with feed-forward. The problem of source coding with feed-forward is also related to source coding with a delay-dependent distortion function [6].

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

1. The optimal rate-distortion function for general discrete sources with a general distortion measure and with noiseless feed-forward, $R_{ff}(D)$, is given by the minimum of the directed information function [7] flowing from the reconstruction to the source. $R_{ff}(D) \leq R(D)$, where $R(D)$ denotes the optimal Shannon rate-distortion function for the source without feed-forward.

2. The performance of the best possible source code (with feed-forward) of rate $R$, distortion $D$ and block length $N$ is characterized by an error exponent. We provide a random coding error exponent $E_{N-ff}(R,D)$ and show that it is greater than or equal to the random coding error exponent without feed-forward.

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

In Section 2 we give a fairly formal definition of the above source coding model and the intuition behind the proposed approach. We then give the rate-distortion theorem for general sources in Section 3. In that section, we also consider two special cases- discrete memoryless sources and Gaussian sources. Random coding error exponents are considered in the general setting in Section 4. We extend our results to feed-forward with an arbitrary delay in Section 5.

II. THE SOURCE CODING MODEL

A. Problem Statement. The model is shown in Figure 1. Consider a discrete source $X$ with $N$th order probability distribution $P_{X^n}$, alphabet $\mathcal{X}$ and reconstruction alphabet $\hat{\mathcal{X}}$. There is an associated distortion measure $d_N$:
\(X^N \times \hat{X}^N \to \mathbb{R}^+\). We assume that \(d_N(,.)\) is normalized with respect to \(N\) and is uniformly bounded in \(N\). For example, \(d_N\) could be a memoryless distortion measure, i.e., \(d_N(x^N, \hat{x}^N) = \frac{1}{N} \sum_{i=1}^{N} d_i(x_i, \hat{x}_i)\) for some \(d_i : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+\), \(\forall i\).

For a source code of block length \(N\) and rate \(R\), the encoder is a mapping to an index set: \(e : X^N \to \{1, \ldots, 2^{NR}\}\). The decoder receives the index transmitted by the encoder, and to reconstruct the \(i\)th sample, it has access to all the past \((i-1)\) samples of the source. In other words, the decoder is a sequence of mappings \(g_i : \{1, \ldots, 2^{NR}\} \times \hat{X}^{i-1} \to \hat{X}\), \(i = 1, \ldots, N\). Let \(\hat{x}^N\) denote the reconstruction of the source sequence \(x^N\). We want to minimize \(E[d_N(X^N, \hat{x}^N)]\) for a given rate \(R\). For any \(D\), let \(R_D(D)\) 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 [8][9] of channel coding with feedback.

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

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

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

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

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.

B. Directed Information. The directed information function was introduced by Massey [7] and has been used to characterize the capacity of channels with feedback [10][11].

Definition 1. The directed information flowing from a sequence \(A^N\) to a sequence \(B^N\) is defined as

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

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. The directed information flowing from \(\hat{X}^N\) to \(X^N\) can be written as

\[
I(\hat{X}^N \to X^N) = \sum_{i=1}^{N} I(\hat{X}_i; X_i|X^{i-1})
\]

\[
= I(\hat{X}^N; X^N) - \sum_{i=2}^{N} I(X^{i-1}; \hat{X}_i|X^{i-1}).
\]

We know that for the usual source coding problem (without feed-forward), the mutual information \(I(X^N; \hat{X}^N)\) represents the minimum number of bits needed to represent \(X^N\) by \(\hat{X}^N\). With feed-forward, the decoder knows the symbols \(X^{i-1}\) to reconstruct \(\hat{X}_i\). This is reflected in the terms subtracted from \(I(X^N; \hat{X}^N)\) in (2). (2) says that since the information \(I(X^{i-1}; \hat{X}_i|X^{i-1})\) is already known through the feed-forward link, we need not spend bits to code this information. Consequently, it is reasonable to expect that the directed information characterizes the rate-distortion function for sources with feed-forward.

We can also interpret the directed information in terms of the backward test-channel \(\hat{X}^N \to X^N\). A source code with feed-forward can be thought of as having feedback in the test-channel and the directed information gives the information flow through the channel with feedback.

III. THE RATE-DISTORTION THEOREM FOR ARBITRARY SOURCES

In this section, we first describe the apparatus we will use for proving coding theorems for sources with feed-forward. We introduce code-functions, which map the feed-forward information to a source reconstruction symbol \(\hat{X}\). The idea of code-functions was introduced by Shannon in 1961 [12]. We
first give a formal definition of a code-function and then see how it is useful in analyzing systems with feed-forward.

**Definition 2.** A source code-function \( f^N \) is a set of \( N \) functions \( \{ f_n \}_{n=1}^N \) such that \( f_n : X^{n-1} \rightarrow \hat{X} \) maps each source sequence \( x^{n-1} \in X^{n-1} \) to a reconstruction symbol \( \hat{x}_n \in \hat{X} \). Denote the space of all code-functions by \( F^N = F_1 \times F_2 \times \ldots \times F_N \). \( \{ f^N : f^N \) is a code function\). \]

**Definition 3.** A \((N, 2^NR)\) source codebook of rate \( R \) and block length \( N \) is a set of \( 2^NR \) code-functions. Denote them by \( f^N[w] \), \( w = 1, \ldots, 2^NR \).

For each source sequence of length \( N \), the encoder sends an index to the decoder. Using the code-function corresponding to this index, the decoder maps the information fed forward from the source to produce an estimate \( \hat{X} \). A code-function can be represented as a tree. Figure 3 shows a code-function for a binary source with a binary reconstruction alphabet. Using the code-function shown in the figure, a source sequence (001) would be reconstructed as (000) and (101) would be reconstructed as (010). 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 feedforward, a source codeword is a source code-function \( f^N \) where for each \( n \in \{1, \ldots, N\} \), the function \( f_n \) is a constant mapping.

**Fig. 4:** Representation of a source coding scheme with feed-forward.

A source code with feed-forward can be thought of as having two components. The first is a usual source coding problem with \( P^N \) 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 \( f^N[i] \), \( i \in \{1, \ldots, 2^NR\} \) and sends the index of the chosen code function. This is the part inside the dashed box in Figure 4. If we denote the chosen code-function by \( f^N \), the second component (decoder 2 in Figure 4) produces the reconstruction given by

\[
\hat{X}_i = f_i(X_i^{-1}), \quad i = 1, \ldots, N,
\]

We now give the rate-distortion function for arbitrary sources with feed-forward. Before stating the general result, we need the following definitions (see [13], [11]).

**Definition 4.** The limsup in probability of a sequence of random variables \( \{ X_n \} \) is defined as the smallest extended real number \( \alpha \) such that \( \forall \epsilon > 0 \)

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

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

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

**Definition 5.** For any sequence of joint distributions \( \{ P_{X^N, \hat{X}^N} \}_{n=1}^N \), define \( \forall \epsilon > 0 \) \( \hat{X}^N \in \hat{X}^N \)

\[
P_{X^N, \hat{X}^N} \left( z^N | \hat{z}^N \right) \triangleq \prod_{i=1}^{N} P_{X_i, \hat{X}_i-1}(z_i, \hat{z}_i^{-1}, z_i^{-1}),
\]

\[
P_{X^N, \hat{X}^N} \left( x^N | \hat{z}^N \right) \triangleq \prod_{i=1}^{N} P_{X_i, \hat{X}_i}(x_i, \hat{z}_i^{-1}, z_i^{-1}).
\]

\[
\bar{T}(X \rightarrow \hat{X}) \triangleq \limsup_{N \to \infty} \frac{1}{N} \log \frac{P_{X^N, \hat{X}^N}(x^N, \hat{z}^N)}{P_{X^N, \hat{X}^N}(x^N \mid \hat{z}^N)P_{X^N}(x^N)}
\]

\[
\bar{L}(X \rightarrow \hat{X}) \triangleq \liminf_{N \to \infty} \frac{1}{N} \log \frac{P_{X^N, \hat{X}^N}(x^N, \hat{z}^N)}{P_{X^N, \hat{X}^N}(x^N \mid \hat{z}^N)P_{X^N}(x^N)}
\]

As pointed out in [11], the directed information rate, defined by \( \lim_{n \to \infty} \frac{1}{n} \log I(X^n \rightarrow \hat{X}^n) \) may not exist for an arbitrary random process which may not be stationary. But the sup-directed information rate \( \bar{T}(X \rightarrow \hat{X}) \) and the inf-directed information rate \( \bar{L}(X \rightarrow \hat{X}) \) always exist. Tatikonda and Mitra [11] showed that for arbitrary channels with feedback, the capacity is an optimization of \( \bar{L}(X \rightarrow \hat{X}) \), the inf-directed information rate. Our result is that the rate distortion function for an arbitrary source with feed-forward is an optimization of \( \bar{T}(X \rightarrow \hat{X}) \), the sup-directed information rate.

The source distribution, defined by a sequence of finite-dimensional distributions [14], is denoted by

\[
P_X \triangleq \{ P_{X^N} \}_{n=1}^N.
\]

Similarly, a conditional distribution is denoted by

\[
P_{X \mid x} \triangleq \{ P_{X^N \mid x} \}_{n=1}^N.
\]

We give all our results here for the expected distortion criterion.

**Definition 6.** (Expected distortion criterion) \( R \) is an achievable rate at expected distortion \( D \) if for all sufficiently large \( N \), there exists an \((N, 2^NR)\) source codebook such that

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

where \( \hat{x}^N \) denotes the reconstruction of \( x^N \). \( R \) is an achievable rate at expected distortion \( D \) if it is achievable for every \( \epsilon > 0 \).
Theorem 1. For an arbitrary source X characterized by a distribution \( P_X \), the rate-distortion function with feed-forward, the infimum of all achievable rates at expected distortion D, is given by

\[
R_{ff}(D) = \inf_{\mathbf{P}_{X|X} : \lambda(\mathbf{P}_{X|X}) \leq D} \mathcal{T}(\hat{X} \to X),
\]

where

\[
\lambda(\mathbf{P}_{X|X}) = \lim_{n \to \infty} \sup_n E[dn(X^n, \hat{X}^n)].
\]

From [11], we have the following result. For any sequence of joint distributions \( \{P_{X_n, X_n}^{\infty}\}_{n=1}^{\infty} \), we have

\[
\mathcal{L}(\hat{X} \to X) \leq \liminf_{N \to \infty} \frac{1}{N} \mathcal{I}(\hat{X}^N \to X^N)
\]

\[
\leq \limsup_{N \to \infty} \frac{1}{N} \mathcal{I}(\hat{X}^N \to X^N) \leq \mathcal{T}(\hat{X} \to X).
\]

If

\[
\mathcal{L}(\hat{X} \to X) = \mathcal{T}(\hat{X} \to X),
\]

we say that the process \( \{P_{X_n, X_n}^{\infty}\}_{n=1}^{\infty} \) is information stable [15], and all four quantities in (7) are equal. Note that if the joint process \( \{X_n, \hat{X}_n\}^{\infty}_{n=1} \) is information stable, the rate-distortion function becomes

\[
R_{ff}(D) = \inf_{\mathbf{P}_{X|X} : \lambda(\mathbf{P}_{X|X}) \leq D} \lim_{N \to \infty} \frac{1}{N} \mathcal{I}(\hat{X}^N \to X^N).
\]

We do not give the detailed proofs of the direct and converse parts of Theorem 1. Instead, we give a brief idea of the direct part here. For the sake of intuition, assume information stability. We want to show the achievability of all rates greater than the \( R_{ff}(D) \) in (8).

Let \( P_{X_n, X_n}^{\infty} \) be the distribution that maximizes \( \mathcal{I}(\hat{X}^N \to X^N) \), subject to the constraint. 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

\[
Q_{X_n, \hat{X}_n} = P_{X_n} P_{X_n|X_n}^{\infty}.
\]

We also impose certain additional constraints on \( Q_{F_n, X_n, \hat{X}_n} \) so that \( I_Q(F^N; X^N) = I_Q(X^N \to X^N) \).

Using (9) in the above equation, we get

\[
I_Q(F^N; X^N) = I_{P_{X_n} P_{X_n|X_n}^{\infty}}(X^N \to X^N).
\]

Using the techniques for source coding without feed-forward [16], it can be shown that all rates greater than \( \frac{1}{N} I_Q(F^N; X^N) \) can be achieved. From (11), it follows that all rates greater than \( I_{P_{X_n} P_{X_n|X_n}^{\infty}} \) are achievable. The bulk of the proof lies in constructing a suitable joint distribution \( Q \).

We now study the rate-distortion function for two important kinds of sources.

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

We state without proof the following theorem for a DMS with expected distortion constraint and a memoryless distortion measure \( d_N(x^n, \hat{x}^N) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, \hat{x}_i) \).

\footnote{For clarity, wherever necessary, we will indicate the distribution used to calculate the information quantity as a subscript.}

\section*{IV. ERRORS EXPOSURES}

We now consider error exponents for source coding with feed-forward.

A. Upper bound on the probability of error. The random coding error-exponent for a source code of block-length \( N \) for a discrete memoryless source was derived by Blahut [17] and Marton in [18]. A procedure identical to the proof of Theorem 6.5.1 in [17] yields the random coding error exponent for an arbitrary source (without feed-forward). Therefore, we have the following fact for discrete sources without feed-forward.

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

\[
P_e \leq e^{-N E_{x}}(sD) + o(N),
\]

where \( E_{x}(sD) \) is the error exponent for the source (without feed-forward) and is given by

\[
E_{x}(sD) = \max_{s \geq 0} \max_{s \in \mathbb{R}} \left[ \sum_{s} s R - \sum_{s} \log_{2} \left( \frac{\sum_{x} P_{x}(x) e^{s D_{x}(x, x')} - 1}{\sum_{x} P_{x}(x) e^{s D_{x}(x, x')}} \right) \right] ^{s},
\]

for large enough \( N \), \( o(N) = 0 \).

The proof of this involves choosing random codewords with distribution \( \tilde{q}_{x}^{N} \). For a source code with feed-forward, the decoder knows \( x_{i-1} \) to decode \( x_{i} \). So we can choose codewords with distribution

\[
\tilde{q}_{x}^{N} = \prod_{i} q_{x_{i}, x_{i-1}, x_{i-1}},
\]

By randomly picking codewords with the above distribution, we can derive the error exponent for a source with feed-forward.

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

\[
P_e \leq e^{-N E_{x}(sD) + o(N)},
\]

where \( E_{x}(sD) \) is the error exponent for the source (with feed-forward) and is given by

\[
E_{x}(sD) = \max_{s \geq 0} \max_{s \in \mathbb{R}} \left[ \sum_{s} s R - \sum_{s} \log_{2} \left( \frac{\sum_{x} P_{x}(x) e^{s D_{x}(x, x')} - 1}{\sum_{x} P_{x}(x) e^{s D_{x}(x, x')}} \right) \right] ^{s},
\]

where

\[
\tilde{q}_{x}^{N}(x, x') = \prod_{i} q_{x_{i}, x_{i-1}, x_{i-1}}(\hat{x}_{i}, x_{i-1}, x_{i-1}).
\]

We now compare the error exponents for a source with and without feed-forward given by Theorem 4 and (14), respectively. Denote the space of all distributions of the form \( \tilde{q}_{x}^{N} \) by \( S_{q} \) and the space of all distributions of the form \( \tilde{q}_{x}^{N} \) by \( S_{\tilde{q}} \). The only difference between the expressions for the error exponents with and without feed-forward is that the former involves a maximization over distributions in \( S_{q} \), while the latter involves a maximization over \( S_{\tilde{q}} \).

Now, every distribution \( \tilde{q}_{x}^{N} = \prod_{i} q_{x_{i}, x_{i-1}} \) belongs to the space of distributions of the form \( \tilde{q}_{x}^{N} \) is \( \prod_{i} q_{x_{i}, x_{i-1}, x_{i-1}} \). Therefore,

\[
S_{q} \subset S_{\tilde{q}}.
\]

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

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

V. 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 \( N \), i.e., to reconstruct the \( i \)th sample the decoder had perfect knowledge of the first \( i - 1 \) samples.

We now extend our results for a general delay \( N + k \), where \( N \) is the block length. The encoder is a mapper to an index set: \( e: X^{N} \rightarrow \{1, \ldots, 2^{NR}\} \). The decoder receives the index transmitted by the encoder, and to reconstruct the \( i \)th sample, it has access to the first \((i - k)\) samples of the source. In other words, the decoder is a sequence of mappings \( g_{1}, \ldots, 2^{NR} \times X^{N-k} \rightarrow X_{i} \).

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

\[
I(X^{N} \rightarrow X^{N}) = I(X_{1}; X^{N}) - \sum_{i=2}^{N} I(X_{i-1}; X_{i}|X^{i-1}).
\]

When the feed-forward delay is \( N + k \), the decoder knows \( X^{N-k} \) to reconstruct \( X_{i} \). Here, we need not spend \( I(X^{N-k}; X_{i}|X^{i-1}) \) 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

\[
I_{b}(X^{N} \rightarrow X^{N}) = I(X_{1}; X^{N}) - \sum_{i=k+1}^{N} I(X_{i-k}; X_{i}|X^{i-1})
\]

\[
= I(X_{1}; X^{N}) - I(0X^{N-k} \rightarrow X^{N})
\]

where \( 0X^{N-k} \) is the \( N \)-length sequence \( [x_{1}, x_{2}, \ldots, x_{k}, \ldots, x_{N-k}] \).

Observing (16), we make the following comment. In any source coding problem, the mutual information \( I(X^{N}, X^{N}) \) is the fundamental quantity to characterize the rate-distortion function. With feed-forward, the rate-distortion function is reduced by a quantity equal to the information we get for free because of the feed-forward. One can use very similar arguments to characterize the capacity of channels with feedback delay \( k \geq 1 \).

We now state the two main theorems: the rate-distortion theorem and the random coding error exponent for feed-forward with delay.

**Definition 7.**

\[
\tilde{I}_{b}(X^{N} \rightarrow X^{N}) \triangleq \frac{1}{N} \sum_{i=1}^{N} P_{X_{i}}(X_{i}^{i-1}, X_{i-k}^{i-k}),
\]

\[
I_{b}(X^{N} \rightarrow X^{N}) \triangleq I(X_{1}; X^{N}) - \sum_{i=k+1}^{N} I(X_{i-k}; X_{i}|X^{i-1})
\]

\[
= \sum_{x^{N-k}} P_{X_{i}}(x_{N}, x_{i}^{N}) \log \frac{P_{X_{i}}(x_{N}) P_{X_{i}}(x_{i}^{N})}{P_{X_{i}}(x_{N}) P_{X_{i}}(x_{i}^{N})},
\]

\[
\log \frac{P_{X_{i}}(x_{N}) P_{X_{i}}(x_{i}^{N})}{P_{X_{i}}(x_{N}) P_{X_{i}}(x_{i}^{N})}.
\]
\[ \mathcal{T}_k(\hat{X} \to X) \triangleq \limsup_{N \to \infty} \frac{1}{N} \log \frac{P_{X^N, \hat{X}^N}(x^n, \hat{x}^n)}{P_{X^N}(x^n)P_{\hat{X}^N|X^N}(\hat{x}^n|x^n)} \]

**Theorem 6 (Rate-Distortion Theorem).** (Expected Distortion Constraint) For an arbitrary source \( X \) characterized by a distribution \( P_X \), the rate-distortion function with \( N + k \) delayed feed-forward, the infimum of all achievable rates at expected distortion \( D \), is given by

\[ R_{f-f}(D) = \inf_{\mathcal{P}_{\hat{X}|X} : \lambda(\mathcal{P}_{\hat{X}|X}) \leq D} \mathcal{T}_k(\hat{X} \to X), \]

where

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

**Theorem 7 (Error Exponent).** Given a source with \( N \)-th order distribution \( P_{X^N} \), there exists a \( (N, 2^NR) \) source code with \( N + k \) delayed feed-forward so that the probability that a source sequence of length \( N \) cannot be encoded with distortion \( \leq D \) satisfies

\[ P \leq e^{-N E_{f-f}(R,D) + o(N)}, \]

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

\[ E_{f-f}(N, R, D) = \min_{s \geq 0} \max_{t \leq 0} \left( \frac{sR - stD -}{\frac{1}{N} \log_2 \sum_{x^N} P_{X^N}(x^n) \left( \sum_{x^N} q_{\hat{X}^N|X^N}(\hat{x}^n|x^n)e^{DN}(x^n, \hat{x}^n) \right)^{-s}} \right) \]

where

\[ q_{\hat{X}^N|X^N}(\hat{x}^n|x^n) = \prod_{i=1}^{N} q_{\hat{X}_{i}^{N-i}X_{i}^{1-i}}(\hat{x}_i|x_i^{i-1}, \hat{x}_{i-1}^{i-1}) \]

and for large enough \( N \), \( o(N) = 0 \).

**REFERENCES**


