Feed-Forward Rate Distortion Function and Markov Sources

by

Shahab Asoodeh

A project submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
December 2012

Abstract

The problem of channel coding with feedback has been extensively studied over the last 50 years. Using an ideal feedback link, the encoder knows all previously received channel outputs. Recently the duality between channel coding and rate distortion has been established in [15] and [3], leading to the problem of source coding with feed-forward.

In this project we first study the general formula for the feed-forward rate distortion function given by Venkataramanan et al. [8]. They studied the source coding problem when a feed-forward link is available for general sources and general distortion measures. They derived the feed-forward rate distortion function for an arbitrary source in terms of the directed information which was originally introduced by Massey [5]. It is shown that the general formula given for source coding with feed-forward is closely related to the general formula for channel coding with feedback given by Tatikonda [4].

We then study another formula for the feed-forward rate distortion function recently proposed by Naiss et al. which is tractable and computable [17]. They also calculated the exact rate distortion function for first order asymmetric Markov sources.

An original contribution of this project is an alternative achievability proof for the feed-forward rate distortion function of first order asymmetric Markov sources.

Acknowledgments

I am grateful to my supervisors Prof. Fady Alajaji and Prof. Tamás Linder for their sincere and insightful guidance, continuous support and for being so open and helpful for all problems I had had while completing my degree.

I am also grateful to my family. I cannot imagine myself going through with this degree without their unconditional love and support.

Contents

List of Figures

Chapter 1

Introduction

Missing a train is only painful if you run after it!

–N. N. Taleb

The problem of source coding with side information at the decoder has gained special significance since the emergence of wireless sensor networks. The model of this problem is as follows: An information source which is modeled by a random process $X = \{X_i\}_{i=1}^{\infty}$ is to be encoded in blocks of length n into a message W which is then transmitted over a noiseless channel of finite rate to a decoder. The decoder has access to another random process $Y = \{Y_i\}_{i=1}^{\infty}$, which is correlated with the source X. The decoder then estimates n source samples knowing Y and W and produces a reconstruction of the source process over time. The goal is then to minimize the reconstruction distortion subject to a rate constraint. This problem dates back to 1976 when Wyner and Ziv [1] obtained the optimal rate distortion function when (X, Y) is assumed to be a jointly independent identically distributed (i.i.d.) process. In this problem, (X_i, Y_i) are simultaneously observed at the encoder and decoder. Figure 1.1

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

Figure 1.1: Instantaneous side information, $n = 5$ [8].

shows this scenario for blocklength $n = 5$. Note that in this case, the decoder reconstructs[1] X^5 at the 6th time slot, but we display this as shown in Figure 1.1. Often the side information is the noisy version of the source which is assumed to be available at the decoder. The idealized (though interesting) question is what happens if a delayed version of the source process is available at the decoder. In this case, the delay must be greater than n. Figure 1.2 shows this scenario when the delay is 6 and $Y_i = X_i$.

Although the problem of Figure 1.2 is different from the Wyner and Ziv problem, the encoding remains the same, i.e., a mapping from the $n-$fold product of the source alphabet to an index set of size 2^{nR}, where R is the rate of transmission; thus the encoder is noncausal and the decoder is causal. This model is referred to as data compression with feed-forward. Obviously, while reconstructing $\hat{X}_i$, the decoder knows X^{i-1}. Henceforth we assume that the delay in the scenario of Figure 1.2 is 1 despite the fact that the actual delay is $n + 1 = 6$. Note that the delay obviously cannot be less than the blocklength n, so the assumption that delay is 1 shall not be viewed as the actual delay. In other words, in the feed-forward problems delay k refers to an actual delay of $n + k$.

The source coding model in this setting is as follows. Consider a general discrete

[1] As usual, $X^n := (X_1, X_2, \ldots, X_n)$.

source X with alphabet $\mathcal{X}$ and output alphabet $\hat{\mathcal{X}}$. There is a distortion measure $d_n : \hat{\mathcal{X}}^n \times \mathcal{X}^n \to \mathbb{R}^+$ associated to each pair of sequences. We assume here that $d_n(\hat{x}^n, x^n)$ is normalized with respect to n and uniformly bounded. An $(n, 2^{nR})$ source code with feed-forward of delay 1 and block length n and rate R consists of an encoder mapping f and sequence of decoder mappings g_i, $i = 1, 2, \ldots, n$ such that

$$f : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\},$$

and

$$g_i : \{1, 2, \ldots, 2^{nR}\} \times \mathcal{X}^{i-1} \to \hat{\mathcal{X}}.$$

The encoder maps a sequence of length n to an index chosen from set $\{1, 2, \ldots, 2^{nR}\}$ and sends it to the decoder. The decoder then receives the index and to reconstruct the ith sample it is given all previous $(i-1)$ source samples.

This model was first proposed by Weissman and Merhav [6] in the context of competitive prediction. They considered feed-forward of delay 1 and a single letter difference distortion measure. They derived the distortion rate function with feed-forward of delay 1 for sources which can be represented via an auto-regressive model with an innovation process that is either i.i.d. or satisfies the Shannon Lower Bound (SLB) with equality. As examples of such sources, the distortion rate function is evaluated in [6] for a symmetric binary Markov source with feed-forward of delay 1 and a stationary Gaussian source with feed-forward of delay 1. They also showed that for single-letter difference distortion measures, feed-forward does not lower the optimal distortion rate function for i.i.d. sources and all sources that satisfy the SLB with equality.

Later Pradhan [7] considered the model of source coding with general feed-forward as a variant of the problem of source coding with side information at the decoder and

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

Figure 1.2: Delayed side information, $n = 5$ [8].

a quantization scheme with linear processing for i.i.d. Gaussian sources with mean squared error distortion measure. Finally [8] derived the rate distortion function for general sources and general distortion measures in terms of directed information using the information spectrum method [9], which will be discussed in this report.

The main differences between the results of [8] and [6] are as follows:

- The distortion rate function of a source with feed-forward is completely characterized in [6] only when the source has an auto-regressive representation and the characterization of the distortion rate function is in terms of an innovation process. However [8] considered general sources with feed-forward and the resulting distortion rate function is expressed as a directed information which involves only the source distribution and the conditional probability of reconstruction points given source symbols.

- The results of [6] are valid only for single-letter, difference distortion measures and feed-forward with delay one, while, [8] deals with arbitrary distortion measures and feed forward with arbitrary delay.

1.1 Preliminaries

In 1990, Massey noticed that the usual definition of the most basic channel, namely the discrete memoryless channel (DMC), precludes the use of feedback[2] [5]. He pointed out that probabilitic dependence is quite different from causal dependence. For example, statistical dependence, as opposed to causality, has no inherent directivity. In other words, whether X causes Y or Y causes X, the random variables X and Y are statistically dependent. He then came up with the notion of directivity in information theory by considering the *directed information*. The properties of directed information were developed mainly by Kramer in his PhD thesis [12]. Although the directed information was introduced in the problem of channel coding with feedback, the established duality between feedback in channel coding and feed-forward in source coding has caused the directed information to appear in the source coding literature, see e.g. [8], [16], [17], [21]. In this section we briefly define quantities that we need in our problem and list some of their properties.

Definition. *The directed information flowing from a random vector* $X^n := (X_1, X_2, \ldots, X_n)$ *to another random vector* $Y^n := (Y_1, Y_2, \ldots, Y_n)$ *is defined as*

$$I(X^n \to Y^n) := \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}). \tag{1.1}$$

Notice that (1.1) looks like the mutual information between two random vectors, namely $I(X^n; Y^n) = \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1})$, except that the mutual information has X^n in place of X^i. Hence we can say that the directed information is causal but the mutual information is not. It is easy to verify that $I(X^n \to Y^n) \neq I(Y^n \to X^n)$ in general. Directed information appeared in almost all recent results where feedback

[2] As an example refer to [23, page 48] where Ash used the fact $P(y_i|x^n y^{i-1}) = P(y_i|x^i y^{i-1})$ for DMC. In words, he confused probabilistic dependence with causal dependence.

or feedforward are studied (c.f. [3], [4]). Later, Massey provided the relation between mutual information and directed information by [14]

$$I(X^n \to Y^n) = I(X^n; Y^n) - \sum_{i=2}^{n} I(Y^{i-1}; X_i | X^{i-1}). \tag{1.2}$$

which can be justified as follows

$$\begin{aligned}
I(X^n; Y^n) &= H(X^n) + H(Y^n) - \sum_{i=1}^{n} H(X_i, Y_i | X^{i-1}, Y^{i-1}), \\
&= H(X^n) + H(Y^n) - \sum_{i=1}^{n} H(X_i | X^{i-1}, Y^{i-1}) - \sum_{i=1}^{n} H(Y_i | X^i, Y^{i-1}), \\
&= \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}) + \sum_{i=2}^{n} I(X_i; Y^{i-1} | X^{i-1}), \\
&\overset{(a)}{=} I(X^n \to Y^n) + \sum_{i=2}^{n} I(Y^{i-1}; X_i | X^{i-1}),
\end{aligned} \tag{1.3}$$

where (a) holds by definition of directed information 1.1 and we have assumed the convention $I(Y^0; .) = 0$. Equation (1.2) has an interesting implication in source coding. We know that without feedforward, we need $I(\hat{X}^n; X^n)$ bits to represent X^n by $\hat{X}^n$. At time instant i the decoder knows x^{i-1} to reconstruct $\hat{X}_i$, therefore by (1.2) we can save $I(X^{i-1}; \hat{X}_i | \hat{X}^{i-1})$ bits. In other words, we need not spend $I(X^{i-1}; \hat{X}_i | \hat{X}^{i-1})$ bits at time instant i.

Similar to the information rate, we can define the directed information rate as

$$\lim_{n \to \infty} \frac{1}{n} I(X^n \to Y^n),$$

which is guaranteed to exist if the source is stationary and ergodic [12]. Other concepts that we need are the directed counterparts of conditional probabilities defined by

$$\vec{P}_{\hat{X}^n | X^n}(\hat{x}^n | x^n) := \prod_{i=1}^{n} P_{\hat{X}_i | X^{i-1}, \hat{X}^{i-1}}(\hat{x}_i | x^{i-1}, \hat{x}^{i-1}), \tag{1.4}$$

which corresponds to the decoder conditional probability (or input to the test channel)

and which we call "strictly causal conditional distribution with lag 1", and

$$\vec{P}_{X^n|\hat{X}^n}(x^n|\hat{x}^n) := \prod_{i=1}^{n} P_{X_i|X^{i-1},\hat{X}^i}(x_i|x^{i-1},\hat{x}^i). \tag{1.5}$$

which corresponds to the test channel and which we call "causal conditional distribution with lag 1". To unify these two notions, we can use the causal conditional distribution with lag s defined by Kramer [12] as

$$P_{Y^n||X^{n-s}}(y^n||x^{n-s}) := \prod_{i=1}^{n} P_{Y_i|Y^{i-1},X^{i-s}}(y_i|y^{i-1},x^{i-s}),$$

for any $s \geq 0$. Here we assume the convention that $P_{Y_i|Y^{i-1},X^{i-s}}(y_i|y^{i-1},x^{i-s}) = P_{Y_i|Y^{i-1}}(y_i|y^{i-1})$ for $i = 1,2,\ldots,s$. We can then rewrite (1.4) and (1.5) as $P(\hat{X}^n||X^{n-1})$ and $P(X^n||\hat{X}^n)$, respectively. It is easy to show that

$$P_{\hat{X}^N,X^N}(\hat{x}^N,x^N) = P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})P_{X^n||\hat{X}^n}(x^n||\hat{x}^n). \tag{1.6}$$

We can envision the source coding with feedforward via a test channel using the above two notations. The decoder first receives the index W containing the information about a block of n source samples. The reconstruction process starts from reconstructing the first sample as a function of W; $\hat{x}_1 = g(W)$. In the next time instant, $\hat{x}_1$ is fed to the fictitious test channel $P_{X_1|\hat{X}_1}(x_1|\hat{x}_1)$ to produce x_1, which is then fed back to the decoder. Therefore, the decoder knows W and x_1 to reconstruct the second source sample; $\hat{x}_2 = g(W,x_1)$. In the next time instant, as before, $\hat{x}_2$ goes through the test channel $P(x_2|\hat{x}_2)$ to produce x_2 fed back to decoder. Hence, the decoder can use W, x_1 and x_2 to reconstruct the third sample; $\hat{x}_3 = g(W,x^2)$. This procedure goes on till all source samples are reconstructed.

1.2 Organization of Report

The next chapter considers general sources and gives the feed-forward rate distortion function. In the same chapter, we summarize the results proved in [8]. Chapter 3 gives an alternative formula for feed-forward rate distortion function using the notion of nth order rate distortion function. Chapter 4 provides an achievability proof for the feed-forward rate distortion function of first order asymmetric Markov sources together with its converse thus establishing the feed-forward rate distortion function.

Chapter 2

Feed-Forward Rate Distortion Function

A nice adaptation of conditions will make almost any hypothesis agree with the phenomena. This will please the imagination but does not advance our knowledge.

–J. Black

Consider a sensor network in which a sensor measures a certain physical quantity over time. The main objective of a sensor network is to convey the (processed) measurement, X_i over time $i = 1, 2, \ldots, n$, to the receiver. As an example, each sensor quantizes a measurement and sends it to the receiver. The receiver might have some side information with delay $s \geq 0$, Y_{i-s}. Clearly if the process (X, Y) is i.i.d. and $s > 0$, then side information does not help at all. However this is not the case for $s = 0$ as Wyner and Ziv showed in 1976 [1]. An interesting case for side information is $X_i = Y_{i-s}$, so the whole source field is transferred to the receiver but of course with some non zero delay. Here in this project we always assume that delay is one. Note that in order to have a valid communication model, the delay must be

always greater than the coding blocklength, n, so $s = 1$ is meant to be $s = n + 1$. However in this report we write $s = 1$ for simplicity. In this chapter we first consider the rate distortion function for an stationary ergodic source when a feed-forward of delay one is available and then briefly generalize the result for arbitrary sources. The feed-forward rate distortion function for an arbitrary delay is given in [8].

2.1 Stationary and Ergodic Sources

Although [8] gives the fundamental limit of source coding with feed-forward for general sources, it is instructive to look first at the stationary and ergodic source. Here we assume that $d_n(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i)$, where $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is the distortion measure. We assume that $d(x, \hat{x}), \forall x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$ is bounded and therefore $\lim_{N\to\infty} E\Big[d_n(X^n, \hat{X}^n)\Big]$ exists. Note that when the joint random process $\{X_n, \hat{X}_n\}$ is stationary, then we can write $E\Big[d_n(X^n, \hat{X}^n)\Big] = E[d(X, \hat{X})]$.

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

$$E_{X^n}\Big[d_n(X^n, \hat{X}^n)\Big] \leq D + \epsilon.$$

The distribution, defined by a sequence of finite-dimensional distributions is denoted by

$$\mathbf{P_X} := \{P_{X^i}\}_{i=1}^{\infty},$$

and similarly the conditional distribution is denoted by

$$\mathbf{P_{\hat{X}|X}} := \{P_{\hat{X}^i|X^i}\}_{i=1}^{\infty}.$$

Theorem 2.1.1. *[8] For a discrete stationary and ergodic source X characterized by a distribution $\mathbf{P_X} = \{P_{X^i}\}_{i=1}^{\infty}$, all rates $R \geq R^*(D)$ are achievable at expected distortion D where*[1]

$$R^*(D) := \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: E[d(X,\hat{X})] \leq D} \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n),$$

where the infimum is taken over all conditional distributions $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ for stationary and ergodic joint process $\{\hat{X}_n, X_n\}$.

Note that as we mentioned earlier, the directed information rate exists for stationary and ergodic sources. The proof of achievability is given after two constructive lemmas. The proof is based on a new Asymptotic Equipartition Property (AEP). By the Shannon-McMillan-Breiman theorem, we know that the AEP holds for discrete stationary ergodic sources, which means that with probability one

$$-\frac{1}{n} \log P_{X^n}(X^n) \to H(\mathbf{X}),$$

and

$$-\frac{1}{n} \log P_{X^n,\hat{X}^n}(X^n, \hat{X}^n) \to H(\mathbf{X}, \hat{\mathbf{X}})$$

where

$$H(\mathbf{X}) := \lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} H(X_n|X^{n-1}),$$

and

$$H(\mathbf{X}, \hat{\mathbf{X}}) := \lim_{n\to\infty} \frac{1}{n} H(X^n, \hat{X}^n) = \lim_{n\to\infty} H(X_n, \hat{X}_n|X^{n-1}, \hat{X}^{n-1}).$$

Let $H(\hat{X}^n||X^n)$ denote the entropy of $\hat{X}^n$ causally conditioned on X^n, defined as

$$H(\hat{X}^n||X^n) := \sum_{i=1}^{n} H(\hat{X}_i|\hat{X}^{i-1}, X^i). \tag{2.1}$$

[1]This theorem concerns the achievability of $R^*(D)$, however the converse proof for arbitrary sources is provided in [8].

Similarly we can define[2] $H(\hat{X}^n||X^{n-1})$ with respect to probability distribution (1.4) as follows

$$H(\hat{X}^n||X^{n-1}) := \sum_{i=1}^{n} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}). \tag{2.2}$$

Lemma 2.1.2. *[8] If the process $\{X_i, \hat{X}_i\}_{i=1}^{\infty}$ is stationary and ergodic, then with probability one,*

$$-\frac{1}{n}\log P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) \to H(\hat{\mathbf{X}}||\mathbf{X})$$

as $n \to \infty$ where

$$\begin{aligned} H(\hat{\mathbf{X}}||\mathbf{X}) := \lim_{n\to\infty}\frac{1}{n}H(\hat{X}^n||X^{n-1}) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}) \\ &= \lim_{n\to\infty} H(\hat{X}_n|\hat{X}^{n-1}, X^{n-1}). \end{aligned}$$

The proof follows from the proof of the Shannon-McMillan-Breiman theorem given in [15]. This lemma leads us to the new definition of a distortion typical set. Fix a conditional distribution $\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}$ for a given source distribution $\mathbf{P}_{\mathbf{X}}$, to obtain the joint distribution $\mathbf{P}_{\hat{\mathbf{X}},\mathbf{X}} = \{P_{\hat{x}^n,x^n}\}_{n=1}^{\infty}$. Then for any $x^n \in \mathcal{X}^n$ and $\hat{x}^n \in \hat{\mathcal{X}}^n$, we say that the pair $(x^n, \hat{x}^n)$ belongs to directed distortion typical set, $\mathcal{A}_\epsilon^n$, if

$$\left| -\frac{1}{n}\log P_{X^n}(x^n) - H(\mathbf{X}) \right| < \epsilon, \tag{2.3}$$

$$\left| -\frac{1}{n}\log P_{X^n,\hat{X}^n}(x^n, \hat{x}^n) - H(\mathbf{X}, \hat{\mathbf{X}}) \right| < \epsilon, \tag{2.4}$$

$$\left| -\frac{1}{n}\log P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) - H(\hat{\mathbf{X}}||\mathbf{X}) \right| < \epsilon, \tag{2.5}$$

$$\left| d_n(x^n, \hat{x}^n) - Ed_n(X^n, \hat{X}^n) \right| < \epsilon.$$

Having defined this new typical set, we can show that a result similar to the conventional result in [15] can be obtained. That is, it is easy to show that for any pair

[2]Some references denote this as $H(\hat{X}^n||0X^{n-1})$ where $0X^{n-1} := [-, X_1, X_2, \ldots, X_{n-1}]$.

$(X^n, \hat{X}^n)$ drawn according to $P_{\hat{X}^n, X^n}$;

$$\lim_{n\to\infty} P\big(((X^n, \hat{X}^n) \in \mathcal{A}_\epsilon^n)\big) = 1,$$

or in other words, $\Pr(\mathcal{A}_\epsilon^n) > 1 - \epsilon$ for n sufficiently large, which is proved by an application of Lemma 2.1.2 and the conventional AEP [15]. The following lemma gives a bound on the ratio of conditional distribution and the directed distribution.

Lemma 2.1.3. *[8] For any pair $(x^n, \hat{x}^n) \in \mathcal{A}_\epsilon^n$, we have*

$$\frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})} \leq 2^{(nI(\hat{\mathbf{X}}\to\mathbf{X})+3\epsilon)},$$

where $I(\hat{\mathbf{X}} \to \mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n)$.

Proof.

$$\begin{aligned} P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \frac{P_{X^n,\hat{X}^n}(x^n,\hat{x}^n)}{P_{X^n}(x^n)} &= P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) \frac{P_{X^n,\hat{X}^n}(x^n,\hat{x}^n)}{P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) P_{X^n}(x^n)} \\ &\leq P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) \frac{2^{-n(H(\hat{\mathbf{X}},\mathbf{X})-\epsilon)}}{2^{-n(H(\hat{\mathbf{X}}||\mathbf{X})+\epsilon)} 2^{-n(H(\mathbf{X})-\epsilon)}} \\ &= P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}) 2^{n(I(\hat{\mathbf{X}}\to\mathbf{X})+3\epsilon)}, \end{aligned}$$

where the inequality holds because of (2.3),(2.4) and (2.5) and the last equality holds because $I(\hat{X}^n \to X^n) = H(\hat{X}^n||X^{n-1}) - H(\hat{X}^n|X^n)$, which is proved in (A.6), Appendix A. □

After the preceding lemmas, we can now give a coding scheme which achieves $R^*(D)$. We recall that $d_n(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$. The following proof is adapted from the one given in [8].

Proof of Theorem. Since different x^{i-1} may lead to different reconstruction $\hat{x}_i$, the codebook here consists of codetrees in lieu of codewords. Let the first symbol be $\hat{x}_1$. At the next time instant, the decoder knows x_1, hence to choose $\hat{x}_2$ we have $|\mathcal{X}|$

different choices depending on x_1 observed via the feed-forward link. Similarly, to choose $\hat{x}_3$ we have $\mathcal{X}^2$ different choices given X^2. We can continue this procedure to construct a codebook consisting of 2^{nR} codetrees for a source code with rate R. To generate a codebook, we pick a joint distribution $\mathbf{P}_{\hat{\mathbf{X}},\mathbf{X}}$ whose X-marginal is the source distribution and satisfies $Ed(X,\hat{X}) \leq D$. This joint distribution is stationary and ergodic by assumption. The first symbol $\hat{x}_1$ is chosen randomly according to $P_{\hat{X}_1}$. The second one is chosen independently and randomly according to $P_{\hat{X}_2|\hat{X}_1,X_1}(.|\hat{x}_1,x_1)$ for each possible x_1. For each of $\hat{x}_2$ chosen at last step, there are $|\mathcal{X}|$ possible choices for $\hat{x}_3$ which we pick independently and randomly according to $P_{\hat{X}_3|\hat{X}^2,X^2}(.|\hat{x}^2,x^2)$, We keep on constructing the codebook at each stage till finally we pick $\hat{x}_n$ according to $P_{\hat{X}_n|\hat{X}^{n-1},X^{n-1}}(.|\hat{x}^{n-1},x^{n-1})$. We construct 2^{nR} such codetrees independently and then reveal this codebook to both encoder and decoder.

The encoder can trace the path of each codetree given source sequence x^{n-1}. By this procedure, the encoder gets 2^{nR} different sequences which in fact corresponds to $\hat{x}^n(m)$ for $m \in \{1,2,\ldots,2^{nR}\}$. Then the encoder sends index m for which $(\hat{x}^n,x^n) \in \mathcal{A}_\epsilon^n$.

The decoder receives the index m of the codetree the encoder has picked. Given the codetree and the feedforward sequence x_k for $k = 1,2,\ldots,n-1$, the decoder outputs the path on the codetree determined by the sequence x_k. For example, suppose the codetree in Figure 2.1 is used for a binary source and the source sequence is 101. Then tracing the path determined by 101 gives us 010 as reconstruction. There are two types of distortion incurred in the above coding; one corresponding to the sequence x^n properly encoded which is less than $D+\epsilon$ and another one corresponding to the sequences for which the encoder fails to encode (i.e., the encoder could not find a joint

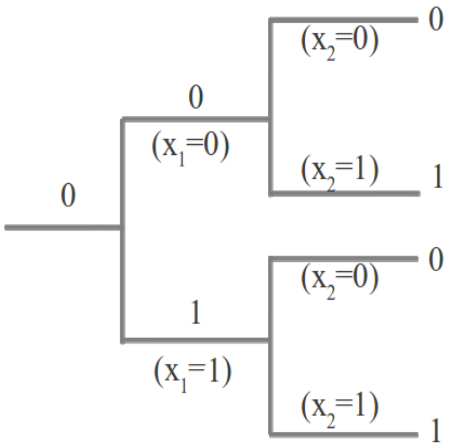

Figure 2.1: Codetree for binary sources

typical path). When the latter happens within a given code for a source sequence, we call the sequence bad. Letting P_e denote the probability of the set of bad source sequence for the code, we can write the expected distortion given the code $\mathcal{C}$ as

$$E[d_n(X^n, \hat{X}^n)|\mathcal{C}] \leq D + \epsilon + P_e d_{max},$$

where d_{max} denotes the maximum of $d(x, \hat{x})$ among all $x \in \mathcal{X}$ and $\hat{x} \in \hat{\mathcal{X}}$. Taking average over all possible random independent codebooks, $\mathcal{C}$, we get

$$E_{\mathcal{C}}\Big[E[d_n(X^n, \hat{X}^n)|\mathcal{C}]\Big] \leq D + \epsilon + \bar{P}_e d_{max},$$

in which $\bar{P}_e$ denotes the probability of the set of bad source sequences averaged over all random codes. It therefore suffices to show that $\bar{P}_e \to 0$ as $n \to \infty$ when $R \geq R^*(D)$. Letting $\mathcal{G}(\mathcal{C})$ be the set of all good sequences (i.e. the ones which can get properly encoded) for code $\mathcal{C}$, we can write

$$\bar{P}_e = \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{x^n : x^n \notin \mathcal{G}(\mathcal{C})} P_{x^n}(x^n). \tag{2.6}$$

Recall that $\bar{P}_e$ is the probability that for a source random sequence X^n and a random

codebook, none of the 2^{nR} paths are directed jointly typical set with X^n. Alternatively, we can calculate $\bar{P}_e$ by first fixing $X^n = x^n$, then finding the probability of bad random codes (the codes all whose codewords corresponding to $X^n = x^n$ are non-typical with x^n) and then summing over all x^n, that is

$$\bar{P}_e = \sum_{x^n} P_{X^n}(x^n) \sum_{\mathcal{C}: x^n \notin \mathcal{G}(\mathcal{C})} P(\mathcal{C}). \tag{2.7}$$

On the other hand, since we generate the codebood using $P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})$, then the probability that a fixed source sequence x^n is not properly represented by a single random codeword $\hat{X}^n$ is

$$\Pr\left((x^n, \hat{X}^n) \notin \mathcal{A}_\epsilon^n\right) = 1 - \sum_{\hat{x}^n:(x^n,\hat{x}^n)\in\mathcal{A}_\epsilon^n} P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1}),$$

which together with the independence of codetrees leads us to calculate the probability of choosing a bad codebook with respect to x^n, that is

$$\sum_{\mathcal{C}: x^n \notin \mathcal{G}(\mathcal{C})} P(\mathcal{C}) = \left(1 - \sum_{\hat{x}^n:(x^n,\hat{x}^n)\in\mathcal{A}_\epsilon^n} P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})\right)^{2^{nR}},$$

and hence together with (2.7), we can write

$$\bar{P}_e = \sum_{x^n} P_{X^n}(x^n) \left(1 - \sum_{\hat{x}^n:(x^n,\hat{x}^n)\in\mathcal{A}_\epsilon^n} P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})\right)^{2^{nR}} \tag{2.8}$$

$$\leq \sum_{x^n} P_{X^n}(x^n) \left(1 - 2^{-n(I(\hat{\mathbf{X}}\to\mathbf{X})+3\epsilon)} \sum_{\hat{x}^n:(x^n,\hat{x}^n)\in\mathcal{A}_\epsilon^n} P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)\right)^{2^{nR}} \tag{2.9}$$

$$\leq \sum_{x^n} P_{X^n}(x^n) \sum_{\hat{x}^n:(x^n,\hat{x}^n)\notin\mathcal{A}_\epsilon^n} P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) + e^{-2^{n(R-I(\hat{\mathbf{X}}\to\mathbf{X})-3\epsilon)}} \tag{2.10}$$

$$= \sum_{(x^n,\hat{x}^n)\notin\mathcal{A}_\epsilon^n} P_{X^n,\hat{X}^n}(x^n, \hat{x}^n) + e^{-2^{n(R-I(\hat{\mathbf{X}}\to\mathbf{X})-3\epsilon)}}, \tag{2.11}$$

where (2.9) holds using Lemma 2.1.3 and (2.10) is due to the inequality $(1-xy)^k \leq 1 - y + e^{-kx}$ for $k > 0$ and $0 \leq x, y \leq 1$. The first term in (2.11) tends to zero as $n \to \infty$ and therefore $\bar{P}_e \to 0$ if $R > I(\hat{\mathbf{X}} \to \mathbf{X}) + 3\epsilon$. □

It is worth comparing the Theorem 2.1.1 with results known for channel coding with perfect feedback. Let X and Y be the channel input and output processes such that $\{X_n, Y_n\}_{n=1}^{\infty}$ is stationary and ergodic. Let $\vec{\mathbf{P}}_{\mathbf{Y}|\mathbf{X}} := \{P_{Y_i|X^i,Y^{i-1}}\}_{i=1}^{\infty}$ and $\vec{\mathbf{P}}_{\mathbf{X}|\mathbf{Y}} := \{P_{X_i|X^{i-1},Y^{i-1}}\}_{i=1}^{\infty}$. The channel is characterized by $\vec{\mathbf{P}}_{\mathbf{Y}|\mathbf{X}}$, so we assume it is fixed. Note that $\mathbf{P}_{\mathbf{X},\mathbf{Y}} = \vec{\mathbf{P}}_{\mathbf{Y}|\mathbf{X}}.\vec{\mathbf{P}}_{\mathbf{X}|\mathbf{Y}}$. Using the direct proof given in this chapter, [16] shows that all rates less than $\sup_{\vec{\mathbf{P}}_{\mathbf{X}|\mathbf{Y}}} I(X \to Y)$ are achievable with feedback. On the other hand, since we know $I(X;Y) = I(X \to Y)$ when there is no feedback, we can conclude that the capacity of channel without feedback can be written as $\sup_{\mathbf{P}_X} I(X \to Y)$. Hence when feedback is available the objective function is the same but the constraint set is larger because the space of $\mathbf{P}_X$ is contained in $\vec{\mathbf{P}}_{\mathbf{X}|\mathbf{Y}}$. This is reversed in source coding. When feedforward is available the objective function is smaller than the no-feedforward case whereas the constraint set is the same.

2.2 General Sources

This section gives the rate distortion function for general sources, which might be non stationary nor ergodic. We can also assume a general distortion measure, single-letter or multi-letter and also with memory. The result is based on the notations introduced by Han in [9], which will be presented briefly in the sequel.

Definition. *The limsup in probability for a sequence of real-valued random variables* $\{X_n\}$ *is defined as the smallest extended real number* α *such that*

$$\lim_{n\to\infty} \Pr(X_n > \alpha) = 0.$$

Definition. *The liminf in probability for a sequence of real-valued random variables*

$\{X_n\}$ *is defined as the largest extended real number* β *such that*

$$\lim_{n\to\infty} \Pr(X_n < \beta) = 0.$$

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

$$i(x^n;\hat{x}^n) := \log \frac{P_{X^n;\hat{X}^n}(x^n;\hat{x}^n)}{P_{X^n}(x^n)P_{\hat{X}^n}(\hat{x}^n)}, \tag{2.12}$$

$$\vec{i}(\hat{x}^n;x^n) := \log \frac{P_{X^n;\hat{X}^n}(x^n;\hat{x}^n)}{P_{\hat{X}^n||X^{n-1}}(\hat{x}^n||x^{n-1})P_{X^n}(x^n)}, \tag{2.13}$$

$$\bar{I}(\hat{\mathbf{X}}\to\mathbf{X}) := \limsup_{inprob} \frac{1}{n}\vec{i}(\hat{X}^n;X^n), \tag{2.14}$$

$$\underline{I}(\hat{\mathbf{X}}\to\mathbf{X}) := \liminf_{inprob} \frac{1}{n}\vec{i}(\hat{X}^n;X^n) \tag{2.15}$$

Verdú and Han [10] showed the interesting result that the capacity without feedback is the sup of inf-information rate (liminf in probability of (2.12)). It was also shown in [11] that the rate distortion function (without feed-forward) for an arbitrary source is given by the inf of the sup-information rate (limsup in probability of (2.12)) and finally [4] proved that for arbitrary channels with feedback, the capacity is an optimization of $\underline{I}(\mathbf{X}\to\mathbf{Y})$, the inf-directed information rate given by (2.15). The following result completes this picture by giving the rate distortion function with feed-forward.

Theorem 2.2.1. *For an arbitrary source* X *characterized by a distribution* $\mathbf{P_X}$*, the rate distortion function with feed-forward at expected distortion* D *is given by*

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}:\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}})\le D} \bar{I}(\hat{\mathbf{X}}\to\mathbf{X})$$

where

$$\lambda(\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}) := E[d_n(X^n;\hat{X}^n)].$$

We have to mention that the same result can be obtained when we impose some constraint on the probability of distortion measure instead of its expectation, that is, we want the achievable rate for which $P_{X^n}\big(x^n : d_n(x^n, \hat{x}^n) \geq D\big) < \epsilon$.

Despite its complicated definition, $R_{ff}(D)$ has been evaluated in closed-form for several classes of sources and distortions measures with memory [13].

Chapter 3

Another Look at Stationary Ergodic Sources

Not everything that can be counted counts, and not everything that counts can be counted.

–A. Einstein

As discussed in the previous chapter, the explicit definition of the feed-forward rate distortion function for an arbitrary normalized distortion function is given by Venkataramana et al. [8]. Borrowing the following measures from the information spectrum method [10],

$$\bar{I}(\hat{\mathbf{X}} \to \mathbf{X}) = \limsup_{inprob} \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}(X^n, \hat{X}^n)}{P_{X^n||\hat{X}^{n-1}}(\hat{X}^n||X^{n-1})P_{X^n}(X^n)},$$

and

$$\underline{I}(\hat{\mathbf{X}} \to \mathbf{X}) = \liminf_{inprob} \frac{1}{n} \log \frac{P_{X^n,\hat{X}^n}(X^n, \hat{X}^n)}{P_{X^n||\hat{X}^{n-1}}(\hat{X}^n||X^{n-1})P_{X^n}(X^n)},$$

they showed that the feed-forward rate distortion function $R_{ff}(D)$, is given by

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: E[d(X,\hat{X})] \leq D} \bar{I}(\hat{\mathbf{X}} \to \mathbf{X}). \tag{3.1}$$

Tatikonda showed that [4]

$$\underline{I}(\hat{\mathbf{X}} \to \mathbf{X}) \leq \liminf_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n) \leq \limsup_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n) \leq \bar{I}(\hat{\mathbf{X}} \to \mathbf{X}),$$

which lets us conclude that for any class of joint processes $(X^n, \hat{X}^n)$ such that $\underline{I}(\hat{\mathbf{X}} \to \mathbf{X}) = \bar{I}(\hat{\mathbf{X}} \to \mathbf{X})$ which includes but is not limited to stationary and ergodic joint processes,

$$R_{ff}(D) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: E[d(X,\hat{X})] \leq D} I(\hat{\mathbf{X}} \to \mathbf{X}) = \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: E[d(X,\hat{X})] \leq D} \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n).$$

Naiss et al.[17] adopted the same approach as [18] to define the nth order feed-forward rate distortion function for stationary and ergodic sources and also proved that it eventually approaches to $R_{ff}(D)$ from above. The new formula for $R_{ff}(D)$ brings a great deal of simplification in terms of calculation as it shows the limit and infimum in the original formula can be interchanged. We briefly mention this result in this chapter and use this to calculate the feed-forward rate distortion function for first order binary asymmetric Markov sources.

3.1 nth Order Feed-Forward RDF

Suppose the source is stationary and ergodic. As stated in the last chapter, a feed-forward rate-distortion pair (R, D) is achievable if there exists an $(n, 2^{nR})$ code such that $E[d(\hat{X}^n, X^n)] \leq D+\epsilon$. We recall that the operational feed-forward rate distortion function is the infimum of R for which (R, D) is achievable. Let $R_{n,ff}(D)$ be the nth order feed-forward rate distortion function for the source defined between two blocks

X^n and $\hat{X}^n$ and normalized distortion measure, i.e.[1]

$$R_{n,ff}(D) := \inf_{\mathbf{P}_{\hat{\mathbf{X}}|\mathbf{X}}: E[d(X^n,\hat{X}^n)]\leq D} \frac{1}{n} I(\hat{X}^n \to X^n), \tag{3.2}$$

in similar way as (9.8.2) in [18]. Let $R^I_{ff}(D)$ be defined by

$$R^I_{ff}(D) := \lim_{n\to\infty} R_{n,ff}(D). \tag{3.3}$$

whenever the limit exists. The following theorem, adapted from [17], shows that the operational definition of feed-forward rate distortion function is equal to the expression given in (3.3).

Theorem 3.1.1. *For any stationary and ergodic source and any distortion D, $R_{ff}(D) = R^I_{ff}(D)$.*

To prove this theorem we first need to show that the limit in (3.3) exists, then show that it is achievable ($R_{ff}(D) \leq R^I_{ff}(D)$), and finally show the converse ($R_{ff}(D) \geq R^I_{ff}(D)$).

Proof. **[Achievability]**

We first show the achievability and assume, for the moment, that the stationary source is block ergodic in blocks of length n. This means that considering each block of length n as a super letter from the super alphabet $\mathcal{X}^n$,we will obtain an ergodic super source. In this setting, we want to prove that for any sufficiently large L, there exists a codebook of trees $\mathcal{T}_\mathcal{C}$ of length L whose cardinality satisfies $|\mathcal{T}_\mathcal{C}| \leq 2^{L(R_{n,ff}(D)+\delta)}$ for which $E[d(X^L, \hat{X}^L)] \leq D + \delta$. This is a generalization of [18, Theorem 9.8.2] for an ensemble of codetrees generated by $P(\hat{x}^n||x^{n-1})$ instead of codewords generated by $P(\hat{x}^n)$.

[1] In fact, we can use min in (3.2) in lieu of inf, as we know that the directed information, like mutual information, is a convex function of the conditional probability distribution.

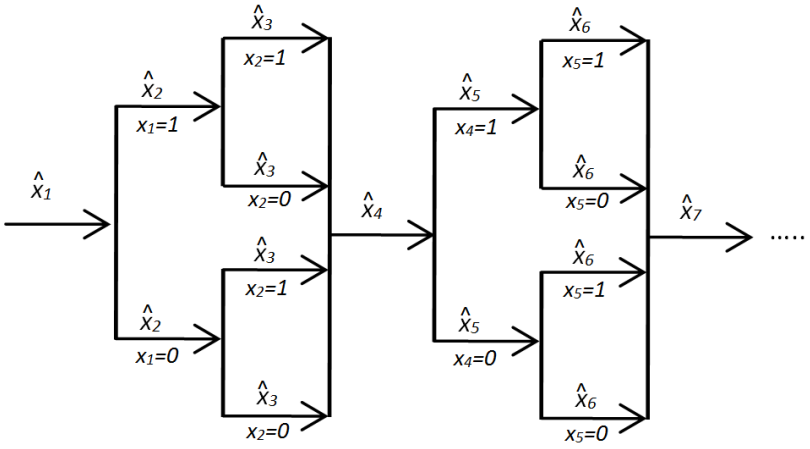

Figure 3.1: Concatenation of two sub-codetrees each whose length is $n = 3$.

Fix $P_{X^n}(x^n)$. Let $P_{X^n|\hat{X}^n}(\hat{x}^n|x^n)$ be the conditional distribution that achieves the $R_{n,ff}(D)$ from which we obtain $P_{X^n||\hat{X}^{n-1}}(\hat{x}^n||x^{n-1})$. For every L, consider the codebook $\mathcal{T}_\mathcal{C}$ ensemble of M codetrees, each of which $\tau^L \in \mathcal{T}_\mathcal{C}$ is a concatenation of $\frac{L}{n}$ sub-codetrees of length n. Each sub-codetree is generated independently according to $P_{X^n||\hat{X}^{n-1}}(\hat{x}^n||x^{n-1})$ as explained in the previous chapter and illustrated in Figure 3.1.

The encoder maps each source sequence x^L to the codetree τ^L whose path determined by x^L has the minimum distortion with x^L, that is, $d(x^L, \hat{x}^L(\tau^L, x^{L-1}))$ is minimum where $\hat{x}^L(\tau^L, x^{L-1})$ denotes the path over τ^L determined by x^L. The encoder then sends the index of that codetree. In other words, the encoder sends the index of codetree τ^{L*} defined as follows

$$\tau^{L*} := \arg\min_{\tau^L \in \mathcal{T}_\mathcal{C}} d(x^L, \hat{x}^L(\tau^L, x^{L-1})). \tag{3.4}$$

The decoder simply picks the tree whose index is received and then follows the path determined by X^L sequentially, that is, at time k decoder returns $\hat{x}_k(\tau^{L*}, x^{k-1})$.

In this setting the test channel and the causal conditional probability can be

written as

$$P_{\hat{X}^L|X^L}(\hat{x}^L|x^L) = \prod_{i=0}^{L/n-1} P_{\hat{X}_{ni+1}^{ni+n}|X_{ni+1}^{ni+n}}(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n}), \tag{3.5}$$

$$P_{\hat{X}^L||X^L}(\hat{x}^L||x^L) = \prod_{i=0}^{L/n-1} P_{\hat{X}_{ni+1}^{ni+n}||X_{ni+1}^{ni+n-1}}(\hat{x}_{ni+1}^{ni+n}||x_{ni+1}^{ni+n-1}), \tag{3.6}$$

each of whose terms is simply the same by stationarity of the source, that is

$$P(\hat{X}_{ni+1}^{ni+n} = \hat{x}^n|X_{ni+1}^{ni+n} = x^n) = P(\hat{X}^n = \hat{x}^n|X^n = x^n), \tag{3.7}$$

$$P(\hat{X}_{ni+1}^{ni+n} = \hat{x}^n||X_{ni+1}^{ni+n-1} = x^{n-1}) = P(\hat{X}^n = \hat{x}^n||X^{n-1} = x^{n-1}). \tag{3.8}$$

To follow the proof of Gallager we need to establish a result similar to [18, Lemma 9.3.1] for our setting. To do so we modify the definition of (9.8.8) in [18]. For every tree τ^L we define the measure

$$I_n(\tau^L \to x^L) = \log \frac{P_{\hat{X}^L|X^L}(\hat{x}^L(\tau^L, x^{L-1})|x^L)}{P_{\hat{X}^L||X^{L-1}}(\hat{x}^L(\tau^L, x^{L-1})||x^{L-1})}. \tag{3.9}$$

Notice that (3.9) does not specify the directed information between $\hat{x}^L$ and x^L, but its average does. Following (9.8.9) in [18] we define the following set

$$A = \{\tau^L \in \mathcal{T}^L, x^L \in \mathcal{X}^L : \text{ either } I_n(\tau^L \to x^L) > L\tilde{R} \text{ or } d(x^L, \hat{x}^L(\tau^L, x^{L-1})) > L\tilde{D}\}, \tag{3.10}$$

where $\tilde{R} = (R_{n,ff}(D) + \delta/2)$, $\tilde{D} = (D + \delta/2)$ and $\mathcal{T}^L$ denotes the set of all trees with length L.

The following lemma gives an upper bound to the probability (over the ensemble of $\mathcal{T}_\mathcal{C}$ and ensemble of source outputs) that the distortion between source sequence and the codeword into which it is mapped exceeds $L\tilde{D}$. This is similar to [18, Lemma 9.3.1].

Lemma 3.1.2. *For a given source, distortion measure and test channel, we have the inequality*

$$P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big) \leq P(A) + e^{-(M2^{-L\tilde{R}})},$$

where the set A is defined in (3.10)*, $P(A)$ is the probability of A on the test channel ensemble, M is the number of codetrees and L is the size of codewords.*

Proof. The proof is given in Appendix B. □

Notice that the normalized average distortion over the ensemble of codes satisfies

$$E\big[d(X^L, \hat{X}^L)\big] \leq D + \delta/2 + P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big) \sup_{x^L, \hat{x}^L} d(x^L, \hat{x}^L), \quad (3.11)$$

obtained using upper-bounding the distortion by $D + \delta/2$ when $d(x^L, \hat{x}^L) \leq D + \delta/2$ and by $\sup_{x^L, \hat{x}^L} d(x^L, \hat{x}^L)$ otherwise. We know by the Lemma 3.1.2 that the term $P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big)$ in (3.11) is upper-bounded by $P(A) + e^{-(M2^{-L\tilde{R}})}$. $P(A)$ can be further upper-bounded using union bound as follows

$$P(A) \leq P\Big(\{x^L \in \mathcal{X}^L, \tau^L \in \mathcal{T}^L : I_n(\tau^L \to x^L) > L\tilde{R}\}\Big) + P\Big(\{x^L \in \mathcal{X}^L, \tau^L \in \mathcal{T}^L : d(x^L, \hat{x}^L(\tau^L, x^{L-1})) > L\tilde{D}\Big). \quad (3.12)$$

The first term in (3.12) tends to zero because with probability 1,

$$\frac{1}{n} \lim_{L\to\infty} \frac{1}{L/n} \sum_{i=1}^{L/n-1} \log \frac{P_{\hat{X}_{ni+1}^{ni+n}|X_{ni+1}^{ni+n}}(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n})}{P_{\hat{X}_{ni+1}^{ni+n}||X_{ni+1}^{ni+n-1}}(\hat{x}_{ni+1}^{ni+n}||x_{ni+1}^{ni+n-1})} = \frac{1}{n} E\Big[\frac{P(\hat{X}^n|X^n)}{P(\hat{X}^n||X^{n-1})}\Big] = R_{n,ff}(D), \quad (3.13)$$

where the first equality is due to ergodicity of joint process $(X^n, \hat{X}^n)$ and the second equality follows from the definition of directed information; $E\Big[\frac{P(\hat{X}^n|X^n)}{P(\hat{X}^n||X^{n-1})}\Big] = I(\hat{X}^n \to X^n)$ (c.f. Appendix A). The joint process $(X^n, \hat{X}^n)$ is ergodic because we assume that the source is ergodic in blocks of length n and the test channel is defined to be memoryless for the blocks of length n and therefore the joint process is ergodic [18,

Lemma 9.8.1]. Hence,

$$P\Big(I_n(\tau^L \to x^L) \le L(R_{n,ff}(D) + \delta/2)\Big) \to 1, \qquad L \to \infty. \tag{3.14}$$

The same argument applies to the second term in (3.12) indicating it also tends to zero, in other words with probability 1,

$$\lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} d(x_i, \hat{x}_i(\tau^L, x^{i-1})) = E[d(X, \hat{X})] \le D + \delta/2,$$

which showes that $P(A)$ tends to zero as $L \to \infty$. The term $e^{-(M2^{-L\tilde{R}})}$ also vanishes with sufficiently large L if $M = \lfloor 2^{L(R_{n,ff}(D)+\delta)} \rfloor$. We can thus write

$$P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big) \to 0, \qquad L \to \infty, \tag{3.15}$$

which together with the assumption $\sup_{x^L,\hat{x}^L} d(x^L, \hat{x}^L) < \infty$ shows that the second term in (3.11) tends to zero as $L \to \infty$. Therefore we can claim that if the source is ergodic in blocks of length n then there exists a codebook of size $\lfloor 2^{L(R_{n,ff}(D)+\delta)} \rfloor$ which satisfies $E\big[d(X^L, \hat{X}^L)\big] \le D + \delta$.

However, an ergodic source need not be ergodic in blocks. As an example, consider a binary source with alphabet $\{0, 1\}$ for which the output consists of pairs of identical digits. With probability 1/2, each digit with even index is chosen independently and equipropably from the alphabet and each digit with odd index is a repetition of its preceding even-numbered digit. Similarly, with probability 1/2 each digit with odd index is chosen independently and equipropably from the alphabet and each digit with even index is the same as the preceding odd-numbered digit. Obviously the second-order super source (the source whose outputs are assumed to be pairs) is not ergodic regardless whether the source is ergodic. Indeed, the second-order super source has two modes; in one mode the super source is memoryless giving out 00 and 11 with probability 1/2 and in the other mode all four possible pairs are equally likely and in

this mode the last digit of one pair and the first digit of the next pair are the same. Note that each of these two modes are ergodic.

To extend the above result for ergodic sources which are not necessarily ergodic in blocks of length n, we need to follow Gallager's notion of *ergodic modes*. We know that all stationary processes that are not ergodic can be modeled as a mixture of ergodic sources. If we consider the blocks of length n as a single symbol, then we have a *super source*. As Gallager showed in[18, Lemma 9.8.2] , the set of sequences from the super source can be decomposed into n' ergodic modes, each of which has equal probability $1/n'$, where n' divides n. In fact, if we consider an invariant set S_0, $P(S_0) > 0$, with respect to the n-shift operator, T^n, then we can decompose the source S into n' invariant subsets $\{S_i = T^i(S_0)\}_{i=0}^{n'-1}$, $P(S_i) = 1/n'$ which are called ergodic modes. Moreover, the modes are disjoint, except for an intersection of sets of zero probability. This ensures that conditional on an ergodic mode, S_i, all its invariant subsets under one-shift operator, T, are of probability either 0 or 1. The readers are refereed to [18, Section 9.8] for a detailed study of ergodic modes.

It will be more convenient, in what follows, to assume that there are n ergodic modes where only n' of them are different. We also need to define the ith-phase source, $0 \leq i \leq n-1$, as the source that produces the sequences in S_i according to the original probability distribution on sequences of letters conditional on the occurrence of S_i.

Now let $I(\hat{X}^n \rightarrow X^n|i)$ denote the directed information between a super letter of the ith-phase source $0 \leq i \leq n-1$, and a letter of the super destination alphabet using the conditional probability $P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)$ that achieves $R_{n,ff}(D)$. We can now conclude that $\frac{1}{n}I(\hat{X}^n \rightarrow X^n|i)$ is an upper bound to the n-th order feed-forward

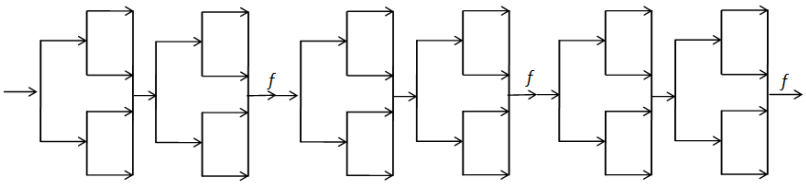

Figure 3.2: A codetree structure from the ith codebook, $n = 3$ and $L = 6$. Letters indicated by f are fixed letters.

rate distortion function of the ith-phase source. Since we can apply the achievability method given above for each ergodic mode, there exists a codebook $\mathcal{T}_{\mathcal{C}_i}$ with $\lfloor 2^{L(\frac{1}{n}I(\hat{X}^n \to X^n|i)+\delta)} \rfloor$ many codetrees of length L such that the distortion constraint for i-th phase is satisfied. We further know that the output of one mode is statistically identical to that of the next mode shifted by one digit. Therefore, if we encode the $(i-1)$th-phase source with $\mathcal{T}_{\mathcal{C}_{i-1}}$, then we can encode ith-phase source with $\mathcal{T}_{\mathcal{C}_i}$. In the following we mention the codebook construction, encoding and decoding processes.

For any L and any ergodic mode S_i, $0 \leq i \leq n$, we use a codebook $\mathcal{T}_{\mathcal{C}_i}$ constructed as explained before with $M_i = \lfloor 2^{L\left(\frac{1}{n}I(\hat{X}^n \to X^n|i)+\delta\right)} \rfloor$ many codetrees of length L according to $P_{\hat{X}^L||X^{L-1}}(\hat{x}^L||x^{L-1})$. Let L be large enough so that such a code can be selected for each of the n phases and consider such a set of n codes. We use these codetrees as the constituent elements of bigger codetrees. For every $0 \leq i \leq n-1$ the ith codebook is an ensemble of 'big' codetrees which consists of n 'little' constituent codetrees starting from one in $\mathcal{T}_{\mathcal{C}_i}$ and followed by one from $\mathcal{T}_{\mathcal{C}_{i+1}}$ to one from $\mathcal{T}_{\mathcal{C}_{i+n-1}}$, where all indices are modulo n. We place some fixed additional letters from $\hat{\mathcal{X}}$ at the end of each little tree like the structure given by Gallager for codewords [18, Figure 9.8.1] which are to shift the sequence and encode it with a codetree from the sequential codebook. In the Figure 3.2, the letters indicated by f are the fixed letters. In this setting, each codetree is therefore of length $L' = nL + n$.

For every i, the encoder assigns for each source sequence $x^{L'} \in S_i$ a codetree, $\tau^{L'}$, from the ith codebook such that $d(x^{L'}, \hat{x}^{L'}(\tau^{L'}, x^{L'-1}))$ is minimal. The decoder simply picks the corresponding codetree (upon receiving the index) and then follows the path determined by $x^{L'}$ and returns $\hat{x}^{L'}$.

Note that since the distortion constraint for ith-phase source is satisfied, the total distortion over all ergodic modes also satisfies the constraint. We also need to check the distortions of fixed letters added between little codetrees. However those distortions are upper-bounded by $n \sup d(x, \hat{x})$ and therefore is negligible in the total normalized distortion for large L.

The total number of codetrees in the ith big code is thus $\prod_{i=0}^{n-1} M_i$ where M_i is the number of codetrees in the ith little code and since for every ergodic mode, the codebook is of the same size, the number of overall codetrees is therefore:

$$\begin{aligned} M = n\prod_{i=0}^{n-1} M_i &\leq n\prod_{i=0}^{n-1} 2^{L(\frac{1}{n}I(\hat{X}^n \to X^n|i)+\delta)} \\ &= 2^{L\left(\frac{1}{n}\sum_{i=0}^{n-1} I(\hat{X}^n \to X^n|i)+n\delta+\frac{\log n}{L}\right)} \\ &\leq 2^{L\left(I(\hat{X}^n \to X^n)+n\delta+\frac{\log n}{L}\right)} = 2^{nL\left(\frac{1}{n}I(\hat{X}^n \to X^n)+\delta+\frac{\log n}{nL}\right)} \\ &\leq 2^{L'\left(\frac{1}{n}I(\hat{X}^n \to X^n)+\delta'\right)} = 2^{L'(R_{n,ff}(D)+\delta')}, \end{aligned} \tag{3.16}$$

where $\delta' = \delta + \frac{\log n}{nL}$ and the first inequality is due to the concavity of directed information over the input probability $P(x^n)$, i.e.,

$$I(\hat{X}^n \to X^n) \geq \frac{1}{n}\sum_{i=0}^{n-1} I(\hat{X}^n \to X^n|i),$$

which completes the proof. □

So far we proved that the $R_{n,ff}(D)$ defined in (3.2) is achievable. Hence to complete the proof of achievability of $R_{ff}^I(D)$, we need to show that $\lim_{n\to\infty} R_{n,ff}(D)$

exists and is also achievable. The following theorem shows this.

Theorem 3.1.3. *The sequence $R_{n,ff}(D)$ is sub-additive and thus*

$$\inf_n R_{n,ff}(D) = \lim_{n\to\infty} R_{n,ff}(D).$$

Proof. To show that a sequence, a_n is sub-additive, we need to prove that for all m and l, $(m+l)a_{m+l} \leq ma_m + la_l$. Let $P_m(\hat{x}^m|x^m)$ and $P_l(\hat{x}^l|x^l)$ be two conditional probabilities that achieve $R_{m,ff}(D)$ and $R_{l,ff}(D)$, respectively. Consider two source sequences x^m and x^l generated independently according to $P_{X^m}(x^m)$ and $P_{X^l}(x^l)$ and then append them to obtain the sequence x^{m+l}. Hence by the construction,

$$H(\hat{X}^{m+l}|X^{m+l}) = H(\hat{X}^m|X^m) + H(\hat{X}_{m+1}^{m+l}|X_{m+1}^{m+l}). \tag{3.17}$$

According to the formula $I(\hat{X}^n \to X^n) = H(\hat{X}^n||X^{n-1}) - H(\hat{X}^n|X^n)$ proved in (A.6), we need to calculate $H(\hat{X}^{n+l}||X^{n+l-1})$. We can write

$$\begin{aligned} H(\hat{X}^{m+l}||X^{m+l-1}) &= \sum_{i=1}^{m+l} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}) = H(\hat{X}^m||X^{m-1}) + \sum_{i=m+1}^{m+l} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}) \\ &\leq H(\hat{X}^m||X^{m-1}) + \sum_{i=m+1}^{m+l} H(\hat{X}_i|\hat{X}_{m+1}^{i-1}, X_{m+1}^{i-1}) \\ &= H(\hat{X}^m||X^{m-1}) + H(\hat{X}_{m+1}^{m+l}||X_{m+1}^{m+l-1}). \end{aligned} \tag{3.18}$$

Combining (3.17) and (3.18), we can write

$$\begin{aligned} (m+l)R_{m+l,ff}(D) \leq I(\hat{X}^{m+l} \to X^{m+l}) &\leq I(\hat{X}^m \to X^m) + I(\hat{X}_{m+1}^{m+l} \to X_{m+1}^{n+l}) \\ &= nR_{n,ff}(D) + lR_{l,ff}(D), \end{aligned} \tag{3.19}$$

where the equality holds because $R_{n,ff}(D) = \frac{1}{n} I(\hat{X}^n \to X^n)$ and due to stationarity $R_{l,ff}(D) = \frac{1}{l} I(\hat{X}_{m+1}^{m+l} \to X_{m+1}^{n+l})$. It just remains to invoke [18, Lemma 4A.2] to conclude that $\inf_n R_{n,ff}(D) = \lim_{n\to\infty} R_{n,ff}(D)$. □

This theorem shows that the limit in the definition of $R_{ff}^I(D)$ exists and is equal

to the infimum of sequence $R_{n,ff}(D)$. We showed that $R_{n,ff}(D)$ is achievable for any stationary and ergodic sources so is its infimum. Therefore $R^I_{ff}(D)$ is achievable, i.e., $R_{ff}(D) \geq R^I_{ff}(D)$. The following shows the converse.

Proof. **[Converse]**

To prove the converse, we assume that there is a feed-forward code $(n, 2^{nR})$ defined in Chapter 2 satisfying the distortion constraint $E[d(x^n, \hat{x}^n)] \leq D + \epsilon$ for sufficiently large n. Suppose the encoder function is f and the index transmitted is $T = f(X^n)$. Then we can write:

$$\begin{aligned} nR \geq H(T) \geq I(X^n;T) &= \sum_{i=1}^{n} I(X_i;T|X^{i-1}) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}) - H(X_i|T,X^{i-1}) \overset{(a)}{=} \sum_{i=1}^{n} H(X_i|X^{i-1}) - H(X_i|T,X^{i-1},\hat{X}^i) \\ &\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_i|X^{i-1}) - H(X_i|X^{i-1},\hat{X}^i) \\ &\overset{(c)}{=} I(\hat{X}^n \to X^n), \end{aligned} \tag{3.20}$$

where (a) holds because given T and X^{i-1} the decoder knows X^i, (b) is due to the fact that conditioning reduces the entropy and (c) is the definition of directed information. Taking n to infinity we can conclude that

$$R \geq \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n) \geq \lim_{n\to\infty} R_{n,ff}(D) = R^I_{ff}(D),$$

which proves the converse. □

Chapter 4

Markov Sources

The supreme task of the physicist is to arrive at those universal elementary laws from which the cosmos can be built up by pure deduction. There is no logical path to these laws; only intuition, resting on sympathetic understanding of experience, can reach them.

–A. Einstein

As discussed in the preceding chapters, two formulae have been proposed for calculating $R_{ff}(D)$. The latter discussed in Chapter 3, gives a better means to calculate the feed-forward rate distortion function. To calculate $R_{ff}(D)$ using this formula given in (3.3), we need to first compute the nth order feed-forward rate distortion function, $R_{n,ff}(D)$, take its limit as n tends to infinity and then apply Theorem 3.1.1. Although this is much easier than the original formula given in (3.1), the computational complexity required for solving the convex optimization in (3.3) grows exponentially with n. However it is shown in this chapter that (3.3) is helpful for computing $R_{ff}(D)$ for Markov sources. We apply this approach to obtain the feed-forward rate distortion function for the first order asymmetric Markov source (FOAMS). Throughout this chapter we assume that $D \leq 1/2$.

The attempt to calculate the rate distortion function for Markov source dates back to 1970 when Gray could explicitly compute the rate distortion function for a binary symmetric Markov source with transition probability q, BSMS(q), only in a small distortion region [19]. His result is given by

$$R(D) = H_b(q) - H_b(D), \qquad 0 \leq D \leq D_c, \tag{4.1}$$

where $H_b(.)$ is the binary entropy and for $q \leq 1/2$

$$D_c = \frac{1}{2}\Big(1 - \sqrt{1 - \big(\frac{q}{1-q}\big)^2}\Big).$$

Beyond D_c currently only lower and upper bounds on $R(D)$ are known. In 1977, Berger found explicit lower and upper bounds for $R(D)$, $R_\ell(D)$ and $R_u(D)$ respectively, which do not depend on n and hence can easily be computed [20]. The lower bound is given by

$$R_\ell(D) = \begin{cases} H_b(q) - H_b(D), & 0 \leq D \leq D_2, \\ \max_{q/2 \leq \alpha \leq 1} \big[D \log \alpha - \log(1+\alpha) - (1-q)\log p_\theta - q \log q_\theta\big], & D_2 \leq D \leq \frac{1}{2}, \end{cases}$$

where

$$D_2 = \frac{1}{2}\big(1 - \sqrt{1-2q}\big),$$

$$r = \frac{q}{1-q},$$

and

$$p_\theta = 1 - q_\theta = (1 + r^\theta)^{-1}$$

and θ and α are related via the following expression

$$\frac{r^\theta}{p(1+r^\theta)^2(1+r^{1-\theta})} = \frac{\alpha}{(1+\alpha)^2}.$$

We can observe that for the small distortion region, $R_\ell(D)$ coincides with Gray's result; however, as distortion increases it deviates and can be shown to be strictly

better than the one proposed by Gray. The upper bound, $R_u(D)$, is given by

$$R_u(D_\alpha) = D_\alpha \log \alpha - \log(1+\alpha) - (1 - q \log p_\alpha) - q \log q_\alpha, \tag{4.2}$$

where

$$q_\alpha = 1 - p_\alpha = 2\sqrt{\alpha}(1+\sqrt{\alpha})^{-2},$$

$$r_\alpha = \frac{q_\alpha}{p_\alpha} = \frac{2\sqrt{\alpha}}{1+\alpha},$$

and

$$D_\alpha = \Big(\frac{\alpha}{1-\alpha^2}\Big)[(pr_\alpha + qr_\alpha^{-1})^2 - \alpha].$$

The advantage of these two bounds over the one given by Gray is that they can be easily computed with little computational effort as they do not depend on n.

Feed-forward was introduced by Weissman et al.[6]. They actually called the problem of feed-forward competitive prediction in which they defined a set of functions, F_i, that predicts X_i given X^{i-1}. They also defined the innovation process, $W_i = X_i - F_i(X^{i-1})$, and showed that if W_i is i.i.d. then the feed-forward distortion rate function is equal to the standard distortion rate function of W_i without feed-forward. They also showed that the same result can be obtained if the innovation process meets the Shannon lower bound with equality. As an immediate consequence of the former result, we can conclude that if X_i is a memoryless source and thus an i.i.d. process, then $W_i = X_i$ and therefore the presence of the feed-forward link does not improve the rate-distortion function. Since for Hamming loss the innovation process for BSMS(q), $q \leq 1/2$ satisfies the Shannon lower bound with equality, the result of [6] can be used to show that for BSMS(q), $R_{ff}(D) = H_b(q) - H_b(D)$ which is equal to the lower bound obtained by Berger (4.2) and Gray (4.1) and thus the feed-forward helps us achieve the lower bound of the rate distortion function for the

binary symmetric Markov source. In this chapter we calculate the $R_{ff}(D)$ for an asymmetric Markov sources, FOAMS, that can serve as the lower bound for $R(D)$. In the sequel, we first present the converse and then present an achievability proof whose spirit is borrowed from [21].

4.1 Binary Asymmetric Markov Source

Let $\mathcal{B}(p)$ denote a Bernoulli distribution with transition probability p, that is, if $W \sim \mathcal{B}(p)$ then $W = 1$ with probability p and $W = 0$ with probability $1 - p$. Any FOAMS, X, can be represented by two Bernoulli sources as follows:

$$X_i = X_{i-1}W_i^1 + (1 - X_{i-1})W_i^2, \tag{4.3}$$

where W^1 and W^2 are two independent processes and $W_i^1 \sim \mathcal{B}(1-q)$ and $W_i^2 \sim \mathcal{B}(p)$ and X_{i-1}, W_i^1 and W_i^2 are independent for every i. To show that the process X_i with the above representation is indeed a Markov process, we need to show that $P(X_i = x_i|X^{i-1} = x^{i-1}) = \Pr(X_i = x_i|X_{i-1} = x_{i-1})$ for $x_i, x_{i-1} \in \{0, 1\}$ and $x^{i-1} \in \{0, 1\}^{i-1}$.

$$\begin{aligned}\mathrm{P}(X_i = 0|X^{i-1} = x^{i-1}) &= \mathrm{P}\Big(x_{i-1}W_i^1 + (1 - x_{i-1})W_i^2 = 0|X^{i-1} = x^{i-1}\Big),\\ &= 1_{\{x_{i-1}=0\}}\mathrm{P}(W_i^2 = 0) + 1_{\{x_{i-1}=1\}}\mathrm{P}(W_i^1 = 0),\end{aligned}$$

and similarly,

$$\mathrm{P}(X_i = 1|X^{i-1} = x^{i-1}) = 1_{\{x_{i-1}=0\}}\mathrm{P}(W_i^2 = 1) + 1_{\{x_{i-1}=1\}}\mathrm{P}(W_i^1 = 1),$$

which show that $\mathrm{P}(X_i = x_i|X^{i-1} = x^{i-1})$ is a function of only x_{i-1}.

Having represented FOAMS like this, its easy to show that

$$H(X_i|X_{i-1}) = \pi_1 H(p) + \pi_2 H(q), \qquad 2 \le i \le n, \tag{4.4}$$

where $\pi = (\pi_1, \pi_2)$ is the invariant distribution of Markov source. For the FOAM represented by (4.3), we can write

$$\pi_1 = \frac{q}{p+q}, \qquad \pi_2 = \frac{p}{p+q}.$$

4.2 Converse

The general formula for the feed-forward rate distortion function is given in (3.1) in terms of directed information. The formula for the stationary and ergodic sources is

$$R_{ff}(D) = \inf_{P(\hat{\mathbf{X}}|\mathbf{X}), E[d(X,\hat{X})\leq D]} \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n). \tag{4.5}$$

Note that if we use the single letter distortion function, i.e., $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$, then for jointly stationary process $\{X_n, \hat{X}_n\}$, $E[d(X^n, \hat{X}^n) = E[d(X, \hat{X})]$. The formula (4.5) is hard to compute even for the easiest sources; however as shown in Chapter 3, (3.3) can instead be used. That is, we can simply exchange the inf and lim to compute the $R_{ff}(D)$ which makes the computation much easier. For FOAMS we can therefore write

$$\begin{aligned}
\frac{1}{n} I(\hat{X}^n \to X^n) &= \frac{1}{n}\sum_{i=1}^n I(\hat{X}^i; X_i | X^{i-1}), \\
&= \frac{1}{n}\sum_{i=1}^n H(X_i|X^{i-1}) - H(X_i|X^{i-1}, \hat{X}^i), \\
&\overset{(a)}{=} \frac{1}{n}\Big[H(X_1) + (n-1)H(X_n|X_{n-1}) - \sum_{i=1}^n H(X_i|X^{i-1}, \hat{X}^i)\Big], \\
&\overset{(b)}{\geq} \frac{1}{n}\Big[H(X_1) + (n-1)H(X_n|X_{n-1}) - \sum_{i=1}^n H(X_i|\hat{X}_i)\Big], \\
&\overset{(c)}{\geq} \frac{1}{n} H(\pi) + \frac{n-1}{n}[\pi_1 H(p) + \pi_2 H(q)] - H(D).
\end{aligned}$$

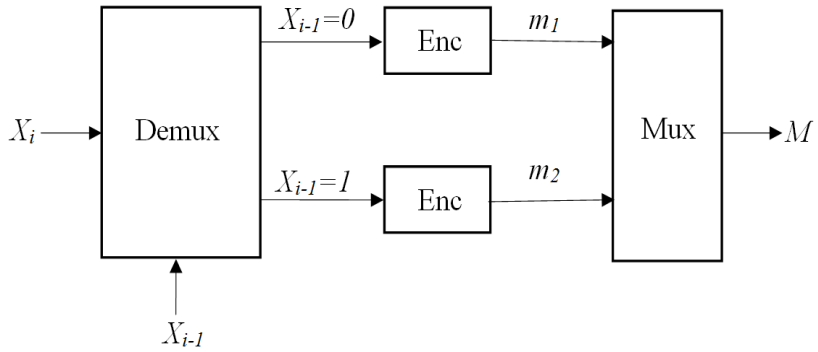

Figure 4.1: The block diagram of encoder

where (a) holds due to Markovity and stationarity of source, (b) follows from the fact that the conditioning reduces the entropy and (c) follows from the fact that $P(X_i \neq \hat{X}_i) \leq D$ and $H(D)$ increases with D for $D \leq \frac{1}{2}$. We can hence conclude that

$$R_{ff}(D) \geq \pi_1 H(p) + \pi_2 H(q) - H(D). \tag{4.6}$$

4.3 Achievability

This section describes the encoding and decoding scheme that achieves the lower bound given in (4.6). The encoder in this setting is a mapping $f\colon \{0,1\}^n \to \{1,2,\ldots,2^{nR}\}$ and the decoder is a series of functions $g_i\colon \{1,2,\ldots,2^{nR}\} \times \{0,1\}^{i-1} \to \{0,1\}$ and the distortion function is single letter. We first partition the source sequence into two sub-sequences, the x_i's following a 0 and the x_i's following a 1 and then encode separately these two sub-sequences. We describe in detail the encoding process for one sub-sequence as the other one is similar. Figure 4.1 shows the schematic structure for the proposed encoder.

Given the source sequence $\{X_n\}_{n=1}^{\infty}$, let N_i be the time index of ith zero in the sequence and $Y_i := X_{N_i+1}$. We can show that $\{Y_n\}$ is an i.i.d. process generated by

$\mathcal{B}(p)$. To achieve this end, we write

$$\begin{aligned}
P(Y^i = y^i) &= \prod_{j=1}^{i} P(Y_j = y_j | Y^{j-1} = y^{j-1}) \\
&= \prod_{j=1}^{i} \sum_{n=1}^{\infty} P(Y_j = y_j | Y^{j-1} = y^{j-1}, N_j = n) P(N_j = n | Y^{j-1} = y^{j-1}) \\
&\overset{(a)}{=} \prod_{j=1}^{i} \sum_{n=1}^{\infty} P(X_{n+1} = y_j | X_n = 0) P(N_j = n | Y^{j-1} = y^{j-1}) \\
&\overset{(b)}{=} \prod_{j=1}^{i} p^{y_j} (1-p)^{1-y_j},
\end{aligned}$$

where (a) is due to the fact that $1_{\{Y^{j-1}=y^{j-1}, N_j=n\}}$ is a measurable function of W_1^1, $W_2^1, \ldots, W_{j-1}^1, W_1^2, W_2^2, \ldots, W_{j-1}^2$ and hence due to the Markovity of source $P(Y_j = y_j | Y^{j-1} = y^{j-1}, N_j = n, X_n = 0) = P(X_{n+1} = y_j | X_n = 0)$ and (b) holds because from (4.3), $P(X_{n+1} = y_j | X_n = 0) = p^{y_j}(1-p)^{1-y_j}$.

The key idea of the encoding scheme is to apply the rate distortion code of a Bernoulli $\mathcal{B}(p)$ source for the sequence Y^i. By the strong law of large numbers for Markov chains, we can conclude that the number of zeros in a sufficiently large source sequence X^n is approximately $n\pi_1$, in other words, as $n \to \infty$ with probability one[1]

$$\frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i = 0\}} \to \pi_1.$$

Let $k_n^p = \lceil n(\pi_1 - \delta) \rceil$ and $\mathcal{E}_n$ be a binary random variable defined as follows

$$\mathcal{E}_n = \begin{cases} 0 & \text{if } N_{k_n^p} \leq n, \\ 1 & \text{if } N_{k_n^p} > n. \end{cases} \tag{4.7}$$

When $\mathcal{E}_n = 0$ we encode $(Y_1, Y_2, \ldots, Y_{k_n^p})$ using an optimal rate distortion code for the source $\mathcal{B}(p)$ at rate R and if $\mathcal{E}_n = 1$ we do not encode and simply send a particular vector. Note that $Y^{k_n^p}$ is no longer an i.i.d. sequence when conditioned on the event

[1]For further details on this result and similar ones, refer to [22, Section 5.5].

$\mathcal{E}_n = 0$. Let $(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_{k_n^p})$ be the reproduction sequence and assume that the distortion between the two sequences is normalized, that is,

$$d(y^{k_n^p}, \hat{y}^{k_n^p}) = \frac{1}{k_n^p} \sum_{i=1}^{k_n^p} d(y_i, \hat{y}_i),$$

and the per-letter distortion is assumed to be Hamming. The total distortion in encoding $Y^{k_n^p}$ using an optimal rate distortion code is

$$\begin{aligned} D_n &:= E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})\big] \\ &= E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})|\mathcal{E}_n = 0\big]P(\mathcal{E}_n = 0) + E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})|\mathcal{E}_n = 1\big]P(\mathcal{E}_n = 1). \quad (4.8) \end{aligned}$$

Since the sequence $Y_1, Y_2, \ldots$ is an i.i.d. sequence with distribution $\mathcal{B}(p)$, then obviously

$$\lim_{n\to\infty} D_n = D_p(R), \quad (4.9)$$

where $D_p(R)$ is the distortion rate function of a Bernoulli source $\mathcal{B}(p)$ operating at rate R. Note that since all terms in (4.8) are nonnegative, we have

$$D_p(R) = \lim_{n\to\infty} D_n \geq \limsup_{n\to\infty} E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})|\mathcal{E}_n = 0\big]P(\mathcal{E}_n = 0). \quad (4.10)$$

On the other hand, since $d(x, y) \leq 1$ for $x, y \in \{0, 1\}$, the distortion of our scheme is deterministically upper bounded by 1 when $\mathcal{E}_n = 1$. Thus, if D_n^p denotes the expected distortion of our scheme, we have

$$D_n^p \leq E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})|\mathcal{E}_n = 0\big]P(\mathcal{E}_n = 0) + P(\mathcal{E}_n = 1). \quad (4.11)$$

Hence together with the fact that $P(\mathcal{E}_n = 1) \to 0$ as $n \to \infty$, we obtain

$$\limsup_{n\to\infty} D_n^p \leq \limsup_{n\to\infty} E\big[d(Y^{k_n^p}, \hat{Y}^{k_n^p})|\mathcal{E}_n = 0\big]P(\mathcal{E}_n = 0) \leq D_p(R). \quad (4.12)$$

The encoding scheme for the other sub-sequence is similar to the above. Let M_i be the time index of ith one in the source sequence x^n and $Z_i := X_{M_i+1}$. We can again show that sequence $\{Z_i\}$ is i.i.d. with distribution $\mathcal{B}(q)$. Letting k_n^q be equal

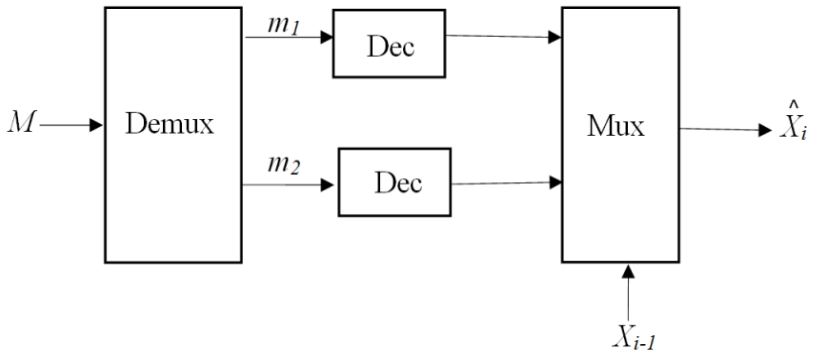

Figure 4.2: The block diagram of decoder

to $\lceil n(\pi_2 - \delta) \rceil$, we can imitate the same coding scheme as before for sequence $Z^{k_n^q}$. Similarly, let D_n^q define the distortion of the encoding scheme in this case.

In the receiver side, we receive two indices regarding two encoders for $Y^{k_n^p}$ and $Z^{k_n^q}$ and hence are able to reconstruct $\hat{Y}^{k_n^p}$ and $\hat{Z}^{k_n^q}$. We then need the causal information, i.e., X^{i-1} at time i to reconstruct the source sequence. In other words, at time i, causal information X^{i-1} helps the decoder pick the appropriate letter between $\hat{Y}_i$ and $\hat{Z}_i$ depending on whether $X_{i-1} = 0$ or $X_{i-1} = 1$. The decoder diagram is schemed in Figure 4.2

The total distortion for encoding the source sequence X^n using our parallel encoding scheme is the sum of the distortion of each sub-sequence and therefore can be obtained in terms of k_n^p, D_n^p, k_n^q and D_n^q. Note that thanks to the way we define k_n^p and k_n^q, there are at most $2n\delta$ many source letters which are not encoded and hence contributes to the total normalized distortion at most 2δ. For the total normalized distortion we can write

$$D_{tot} \leq \frac{1}{n}\big(k_n^p D_n^p + k_n^q D_n^q + 2n\delta\big), \tag{4.13}$$

where $2n\delta$ is the contribution of uncoded bits. Letting $n \to \infty$, we can write:

$$\begin{aligned} D_{tot} &\leq (\pi_1 - \delta)D_P(R) + (\pi_2 - \delta)D_q(R) + 2\delta \\ &\leq \pi_1 D_P(R) + \pi_2 D_q(R) + \delta(\underbrace{1 - D_p(R)}_{>0}) + \delta(\underbrace{1 - D_q(R)}_{>0}) \\ &= \pi_1 D_P(R) + \pi_2 D_q(R) + \epsilon. \end{aligned} \tag{4.14}$$

The entire encoding function can be described as the following mapping

$$\{0,1\}^{k_n^p + k_n^q} \to \{1, 2, \ldots, 2^{k_n^p R}, 2^{k_n^p R} + 1\} \times \{1, 2, \ldots, 2^{k_n^q R}, 2^{k_n^q R} + 1\},$$

which emphasizes that for the sequence $Y^{k_n^p}$ we need an index chosen from $\{1, 2, \ldots, 2^{k_n^p R}\}$ and also one extra index for the case of $\mathcal{E}_n = 1$ and similarly for $Z^{k_n^q}$. Clearly the rate of this encoding is

$$\begin{aligned} R_{tot} &= \frac{1}{n} \log \left[(2^{k_n^p R} + 1)(2^{k_n^q R} + 1) \right] \leq \frac{1}{n}(k_n^p R + k_n^q R + 2) \\ &\leq R + \epsilon, \end{aligned} \tag{4.15}$$

where we use the obvious inequality $\log(1 + x) \leq 1 + \log x$ for $x \geq 1$.

Chapter 5

Summary and Conclusions

In this project, we considered the problem of lossy source coding in the presence of a feed-forward link which conveys the source symbols to the receiver with a non-zero delay. The fundamental limit of the lossy source coding, namely the rate distortion function, is characterized for general sources and any delay in [8] in terms of a multie-letter expression. Naiss et al.[17] uses the Gallager's idea of nth order rate distortion function to prove another formula for feed-forward rate distortion function for station-ary and ergodic soruces. They showed the existence of a sequence of rate-distortion function which converge from above to the formula given by [8].

We used the latter formula to calculate the rate distortion function of first order asymmetric Markov source and proposed an optimal coding scheme to achieve it.

We are currently working on obtaining upper and lower bounds for the rate distortion function of Markov sources when the feed-forward link has delay 2.

Appendix A

Directed Information

Since the directed information (DI) is used in expression for the feed-forward rate distortion function, we provide some basic formulae for DI in this appendix. The directed information from random sequence X^n to Y^n was introduced by Massey [5] as follows:

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}). \tag{A.1}$$

Getting back to the mutual information between two sequences;

$$I(X^n; Y^n) = \sum_{i=1}^{n} I(X^n; Y_i | Y^{i-1}),$$

we can conclude that DI is the causal version of mutual information. Based on this formula, we can easily find the following more illuminating one:

$$I(X^n \to Y^n) = I(X^n; Y^n) - \sum_{i=2}^{n} I(Y^{i-1}; X_i | X^{i-1}), \tag{A.2}$$

which is justified in (1.3).

This equation shows how the feed-forward can reduce the rate distortion function.

In the source coding setting, we can rewrite (A.2) as

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

In fact the second term in (A.3) is the rate which comes for free when the feed-forward link is available. This makes clear as to why DI characterizes the performance limit.

Another important formula for DI is due to the following simple algebraic manipulations:

$$\begin{aligned} I(X^n \to Y^n) &= \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}), \\ &= \sum_{i=1}^{n} \sum_{j=1}^{i} I(X_j; Y_i | X^{j-1}, Y^{i-1}), \\ &= \sum_{j=1}^{n} \sum_{i=j}^{n} I(X_j; Y_i | X^{j-1}, Y^{i-1}), \\ &= \sum_{j=1}^{n} I(X_j; Y_j^n | X^{j-1}, Y^{j-1}), \end{aligned}$$

leading us to the following

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X_i; Y_i^n | X^{i-1}, Y^{i-1}), \tag{A.4}$$

each term of which corresponds to the achievable rate at time i given side information (X^{i-1}, Y^{i-1}). There are two other formulae for directed information which have to do with some other directed quantities. Let the entropy of X^n causally conditioned on Y^n be denoted by $H(X^n||Y^n)$, that is

$$H(X^n||Y^n) = \sum_{i=1}^{n} H(X_i | X^{i-1} Y^i),$$

and

$$H(X^n||Y^{n-1}) = \sum_{i=1}^{n} H(X_i | X^{i-1} Y^{i-1}).$$

Having defined these two quantities, we are able to derive two other formulae for DI

as follows:

$$\begin{aligned} I(X^n \to Y^n) &= \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}), \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i), \\ &= H(Y^n) - H(Y^n||X^n), \end{aligned}$$

which shows

$$I(X^n \to Y^n) = H(Y^n) - H(Y^n||X^n). \tag{A.5}$$

Similarly we can obtain another formula using (A.4) as follows

$$\begin{aligned} I(X^n \to Y^n) &= \sum_{i=1}^{n} I(X_i; Y_i^n | Y^{i-1}, X^{i-1}), \\ &= \sum_{i=1}^{n} H(X_i|Y^{i-1}, X^{i-1}) - H(X_i|Y^n, X^i), \\ &= H(X^n||Y^{n-1}) - H(X^n|Y^n). \end{aligned}$$

and therefore

$$I(X^n \to Y^n) = H(X^n||Y^{n-1}) - H(X^n|Y^n). \tag{A.6}$$

Appendix B

Proof of Lemma 3.1.2

In this appendix, we prove Lemma 3.1.2. We state the lemma again in the following for the sake of completeness. Recall that the set A is defined in (3.10) as

$$A = \{\tau^L \in \mathcal{T}^L, x^L \in \mathcal{X}^L : \text{ either } I_n(\tau^L \to x^L) > L\tilde{R} \text{ or } d(x^L, \hat{x}^L(\tau^L, x^{L-1})) > L\tilde{D}\}, \tag{B.1}$$

Where $I_n(\tau^L \to x^L)$ is defined for every tree τ^L in (3.9). Lemma 3.1.2 gives an upper bound for $P(A)$; namely

Lemma B.0.1. *For a given source, distortion measure and test channel, we have the following inequality*

$$P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big) \leq P(A) + e^{-(M2^{-L\tilde{R}})},$$

where set A is defined in (3.10)*, $P(A)$ is the probability of the set A on the test channel ensemble, M is the number of codetrees and L is the size of codewords.*

Proof. The proof here follows closely the proof of [18, Lemma 9.3.1]. Recall that $P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big)$ is defined over the ensemble of codes. We can

expand this in the following way

$$P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big) = \sum_{x^L \in \mathcal{X}^L} P(x^L) P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D} | X^L = x^L\Big). \tag{B.2}$$

We can also partition set A corresponding to each $x^L \in \mathcal{X}^L$, that is, for every x^L, A_p is the set of all codetrees $\tau^L \in \mathcal{T}^L$ for which $(\tau^L, x^L) \in A$ and hence

$$A_p = \{\tau^L \in \mathcal{T}^L : \text{ either } I_n(\tau^L \to x^L) > L\tilde{R} \text{ or } d(x^L, \hat{x}^L(\tau^L, x^{L-1})) > L\tilde{D}\}. \tag{B.3}$$

Notice that the distortion between given x^L and its corresponding path on the best tree, τ^{L*}, exceeds D if and only if it exceeds D for every trees, in other words, we have $d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}$ for a given x^L if $d\big(X^L, \hat{X}^L(\tau^L, x^{L-1})\big) > L\tilde{D}$ for every $\tau^L \in \mathcal{T}^L$. Thus, loosely speaking, $d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}$ only if $\tau^L \in A_p$ for a given x^L. Since τ^L is independently chosen,

$$P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D} | X^L = x^L\Big) \leq \big(1 - P(A_p^c)\big)^M, \tag{B.4}$$

where A_p^c is the complement of A_p. Since we are dealing with codetrees (as opposed to codewords), we have to consider all codetrees whose paths determined by the given x^l are similar. This is because the probability that $\tau^L \in A_p^c$ depends only on the $\hat{x}^L$ associated with x^L. Hence we need to partition $\mathcal{T}^L$ into disjoint sub-sets:

$$B_{x^L, \hat{x}^L} = \big\{\tau^L \in \mathcal{T}^L \colon \tau^L(x^{L-1}) = \hat{x}^L\big\},$$

where $\tau^L(x^{L-1})$ denotes the path on τ^L determined by x^L. Note that since the trees are constructed according to distribution $P_{\hat{X}^L||X^{L-1}}(\hat{x}^L||x^{L-1})$, then we can conclude that $P(B_{x^L, \hat{x}^L}) = P_{\hat{X}^L||X^{L-1}}(\hat{x}^L||x^{L-1})$.

For every $\tau^L \in B_{x^L, \hat{x}^L} \subset A_p^c$, we have

$$I_n(\tau^L \to x^L) \leq L\tilde{R},$$

and therefore

$$P_{\hat{X}^L||X^{L-1}}(\hat{x}^L||x^{L-1}) \geq P_{\hat{X}^L|X^L}(\hat{x}^L|x^L)2^{-L\tilde{R}}. \tag{B.5}$$

Getting back to (B.4) and letting

$$P_{\tilde{D}}(x^L) := P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}|X^L = x^L\Big),$$

then we can write

$$\begin{aligned}
P_{\tilde{D}}(x^L) &\leq \big(1 - P(A_p^c)\big)^M \\
&= \Big(1 - \sum_{B_{x^L,\hat{x}^L} \subset A_p^c} P(B_{x^L,\hat{x}^L})\Big)^M \\
&= \Big(1 - \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{\hat{X}^L||X^{L-1}}(\hat{x}^L||x^{L-1})\Big)^M \\
&\leq \Big(1 - 2^{-L\tilde{R}} \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{\hat{X}^L|X^L}(\hat{x}^L|x^L)\Big)^M,
\end{aligned} \tag{B.6}$$

where the last inequality follows (B.5). Applying the inequality $(1-ab)^k \leq 1 - a + \exp\{-bk\}$ when $a = \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{\hat{X}^L|X^L}(\hat{x}^L|x^L)$ and $2^{-L\tilde{R}}$, we have

$$P_{\tilde{D}}(x^L) \leq 1 - \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{\hat{X}^L|X^L}(\hat{x}^L|x^L) + \exp\{-M2^{-L\tilde{R}}\}, \tag{B.7}$$

and therefore if we let $P_{\tilde{D}}$ denote $P\Big(d\big(X^L, \hat{X}^L(\tau^{L*}, x^{L-1})\big) > L\tilde{D}\Big)$, we can write

$$\begin{aligned}
P_{\tilde{D}} &\leq \sum_{x^L \in \mathcal{X}^L} P_{X^L}(x^L)\Big[1 - \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{\hat{X}^L|X^L}(\hat{x}^L|x^L) + \exp\{-M2^{-L\tilde{R}}\}\Big] \\
&= 1 - \sum_{x^L \in \mathcal{X}^L} \sum_{\hat{x}^L : B_{x^L,\hat{x}^L} \subset A_p^c} P_{X^L,\hat{X}^L}(x^L, \hat{x}^L) + \exp\{-M2^{-L\tilde{R}}\}.
\end{aligned} \tag{B.8}$$

Note that we can rewrite the double sum in (B.8) as follows

$$\sum_{x^L\in\mathcal{X}^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset A_p^c} P_{X^L,\hat{X}^L}(x^L,\hat{x}^L) = \sum_{x^L\in\mathcal{X}^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset A_p^c}\sum_{\tau^L\in\mathcal{T}^L} P(x^L,\hat{x}^L,\tau^L)$$
$$\geq \sum_{x^L\in\mathcal{X}^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset A_p^c}\sum_{\tau^L\in B_{x^L,\hat{x}^L}} P(x^L,\hat{x}^L,\tau^L). \quad \text{(B.9)}$$

Notice that if $\tau^L \in B_{x^L,\hat{x}^L}$, then $\hat{x}^L$ is deterministically determined by x^L, hence we can continue from (B.9) as follows

$$\sum_{x^L\in\mathcal{X}^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset A_p^c} P_{X^L,\hat{X}^L}(x^L,\hat{x}^L) = \sum_{x^L\in\mathcal{X}^L}\sum_{B_{x^L,\hat{x}^L}\subset A_p^c}\sum_{\tau^L\in B_{x^L,\hat{x}^L}} P(x^L,\tau^L)$$
$$= \sum_{x^L\in\mathcal{X}^L}\sum_{\tau^L\in A_p^c} P(x^L,\tau^L) = P(A^c). \quad \text{(B.10)}$$

Now if we plug (B.10) into (B.8) we get

$$P\Big(d\big(X^L,\hat{X}^L(\tau^{L*},x^{L-1})\big) > L\tilde{D}\Big) \leq 1 - P(A^c) + \exp\{-M2^{-L\tilde{R}}$$
$$= P(A) + \exp\{-M2^{-L\tilde{R}}\}, \quad \text{(B.11)}$$

which completes the proof. □ □

Bibliography

[1] 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.

[2] T. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided state information", *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp.1629-1638, June 2002.

[3] 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, No. 5, pp.1181-1203, May 2003.

[4] S. Tatikonda, *Control Under Communication Constraints*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 2000.

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

[6] T. Weissman and N. Merhav. "On competitive prediction and its relation to rate distortion theory", *IEEE Trans. Inf. Theory*, vol. 49, No. 12, pp.3185-3194, Dec. 2003.

[7] 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, Jan. 2007.

[8] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: rate distortion theorems and error exponents for a general source", *IEEE Trans. Inf. Theory*, vol. 53, No. 6, pp.331-349, June 2007.

[9] T. Han and S. Verdú, "Approximation theory of output statistics", *IEEE Trans. Inf. Theory*, vol. 39, No. 3, pp.752-772, May 1993.

[10] S. Verdú and T. Han, "A general formula for channel capacity", *IEEE Trans. Inf. Theory*, vol. 40, No. 3, pp.1137-1147, July 1994.

[11] Y. Steinberg and S. Verdú, "Simulation of random processes and rate distortion theory", *IEEE Trans. Inf. Theory*, vol. 43, No. 1, pp.63-86, July 1996.

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

[13] R. Venkataramanan and S. S. Pradhan, "On evaluating the rate distortion function of sources with feed-forward and the capacity of channels with feedback", in *Proc. IEEE Inter. Symp. Inf. Theory (ISIT)*, Nice, France, pp. 41-46, June 2007.

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

[15] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, second edition, 2006.

[16] R. Venkataramanan and S. S. Pradhan, "On computing the feedback capacity of channels and the feed-forward rate distortion function of sources", *IEEE Trans. Commun.*, vol. 58, pp. 1889-1896, July 2010.

[17] I. Naiss and H. Permuter, "Computable bounds for rate distortion with feedforward for stationary and ergodic sources", *submitted to IEEE Trans. Info. Theory.*, June 2012. available at http://arxiv.org/pdf/1106.0895v1.pdf.

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

[19] R. M. Gray, "Information rates of autoregressive processes", *IEEE Trans. Inf. Theory*, vol. 16, pp. 412-421, July 1970.

[20] T. Berger, "Explicit bounds to R(D) for a binary symmetric Markov source", *IEEE Trans. Inf. Theory*, vol. 23, pp. 52-59, Jan. 1977.

[21] O. Simeone and H. Permuter, "Source coding when the side information may be delayed", *submitted to IEEE Trans. Inf. Theory, July, 2012*, (available at http://arxiv.org/abs/1109.1293.pdf).

[22] R. Durret, *Probability: Theory and Examples, 3rd edition*, Duxbury Press, 2005.

[23] R. Ash, *Information Theory*, New York, Wiley Press, 1965.

[24] S. Jalali and T. Weissman, "New bounds on the rate distortion function of a binary Markov source", in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nice, France, pp. 571-575, June 2007.