IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Structure Theorems for Real-Time Variable Rate Coding With and Without Side Information

Yonatan Kaspi and Neri Merhav

CCIT Report #794
August 2011

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #794 August 2011

Structure Theorems for Real–Time Variable Rate Coding With and Without Side Information

Yonatan Kaspi and Neri Merhav
Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
Email: {kaspi@tx, merhav@ee}.technion.ac.il

Abstract

The output of a discrete Markov source is to be encoded instantaneously by a variable–rate encoder and decoded by a finite–state decoder. Our performance measure is a linear combination of the distortion and the instantaneous rate. Structure theorems, pertaining to the encoder and next–state functions are derived for every given finite–state decoder, which can have access to side information.

I. Introduction

We consider the following source coding problem. Symbols produced by a discrete Markov source are to be encoded, transmitted noiselessly and reproduced by a decoder which can have causal access to side information (SI) correlated to the source. Operation is in real time, that is, the encoding of each symbol and its reproduction by the decoder must be performed without any delay and the distortion measure does not tolerate delays.

The decoder is assumed to be a finite–state machine with a fixed number of states. With no SI, the scenario where the encoder is of fixed rate was investigated by Witsenhausen [1]. It was shown that for a given decoder, in order to minimize the distortion at each stage for a Markov source of order k, an optimal encoder can be found among those for which the encoding function depends on the k last source symbols and the decoder's state (in contrast to the general case where its a function of all past source symbols). Walrand and Varaiya [2] extended this finding to a joint source–channel setup with noiseless feedback. Teneketzis [3] used ideas from both [1] and [2] and considered the joint source–channel setup for a given finite state decoder without feedback. A causal variant of the Wyner Ziv problem [4] was also considered by Teneketzis [3]. It is shown in [3] that the optimal (fixed rate) encoder for this case is a function of the current source symbol and the probability mass function of the decoder's state for the symbols sent so far. Borkar, Mitter and Tatikonda [5] derived structure theorems of a similar spirit when the cost function is a linear combination (Lagrangian) of the conditional entropy of the reproduction sequence and the mean square error of the reproduction. The scenario where the encoder is also a finite state machine was considered by Gaarder and Slepian in [6]. In some cases, the minimization of the distortion (or cost) can be cast as a stochastic control problem. In this case, tools developed for Markov decision processes

are employed to either solve the optimization problem or to get insights on the structure of the optimal solution. Examples of this technique include [2],[3],[5],[7],[8].

When the time horizon and alphabets are finite, there is a finite number of possible deterministic encoding, decoding and memory update rules. In principle, a brute force search would yield the optimal choice. However, since the number of possibilities increases doubly exponentially in the duration of the communication and exponentially in the alphabet size, it is not trackable even for very short time horizons. Recently, using the results of [3], Mahajan and Teneketzis [9] proposed a search frame that is linear in the communication duration and doubly exponential in the alphabet size.

Real time codes are a subclass of causal codes, as defined by Neuhoff and Gilbert [10]. In [10], entropy coding is used on the whole sequence of reproduction symbols, introducing arbitrarily long delays. In the real time case, entropy coding has to be instantaneous, symbol–by–symbol (possibly taking into account past transmitted symbols). It was shown in [10], that for a discrete memoryless source (DMS), the optimal causal encoder consists of time–sharing between no more than two memoryless encoders. Weissman and Merhav [11] extended [10] to the case where SI is also available at the decoder, encoder or both. Error exponents for real time coding with finite memory for a DMS where derived in [12].

This work extends [1] in several directions: The first is extending the result of [1] from fixed–rate coding to variable–rate coding, where accordingly, the cost function is redefined so as to incorporate both the expected distortion and the expected coding rate. Secondly, we allow the decoder access to causal side information. Unlike [1] and [3], we do not a-priori restrict the encoders to be deterministic and thus the encoders can be any stochastic function of all causally available data. While in [1] and [3], it is quite clear that deterministic encoders are a–priori optimal, it is not immediately clear in our case, as we discuss in the sequel. We show that structure theorems, in the same spirit as those of Witsenhausen [1] and Teneketzis [3], continue to hold in this setting as well. Moreover, the structure can be simplified when the decoder has infinite memory. Finally, we upper bound the loss incurred by using a suboptimal next–state function which uses a "sliding–window" over the past decoder inputs. We refer to such memory update functions as *Markov memory update functions*. The upper bound is given in terms of the original state alphabet and the window length. The suboptimal system that uses Markov memory update functions is analytically more tractable and its optimization is easier since in order to find the best sub–optimal system, effectively, as discussed in the sequel, only the encoders need to be optimized.

In contrast to [1] and [3], where fixed–rate coding was considered, and hence the performance measure was just the expected distortion, here, since we allow variable–rate coding, our cost function incorporates both rate and distortion. This is done by defining our cost function in terms of the Lagrangian

$$(\text{distortion}) + \lambda \cdot (\text{code length}).$$

where $\lambda > 0$ is a fixed parameter that controls the tradeoff between rate and distortion. In [1], the proof of the structure theorem relied on two lemmas. The proofs of the extensions of those lemmas to our case are more involved than the proofs of their respective original versions in [1]. To intuitively see why, remember that the proof of the lemmas in [1], relied on the fact that for every decoder state, source symbol and a given decoder, since there is a finite number of possible encoder outputs (governed by the fixed rate), we could choose the one minimizing the distortion. However, in our case, such a choice might entail a

large expected coding rate, and although minimizes the distortion, it will not minimize the overall cost function (especially for large λ). Furthermore, unlike the case of [1], in our setting, the cost in future stages depends non–linearly on the choices of earlier encoders and in contrast to [1] and [3], there is no reason, as we discuss in the sequel, to a–priori assume that deterministic encoders are optimal.

The remainder of the paper is organized as follows: In Section II, we give the formal setting and notation used throughout the paper. In Section III, we start with the simpler setting without SI. Structure theorems regarding the encoder are derived for both the finite and infinite memory models. In Section IV, we upper bound the loss incurred when Markov memory functions are used instead of the optimal next–state functions. In Section V, we exend the setting of Section III by allowing the decoder access to SI. We begin each section by stating and discussing its main result. Finally, we conclude this work in Section VI.

II. Preliminaries

We begin with notation conventions. Capital letters represent scalar random variables (RV's), specific realizations of them are denoted by the corresponding lower case letters, and their alphabet – by calligraphic letters. For $i < j$ (i, j - positive integers), x_i^j will denote the vector $(x_i, \ldots, x_j)$, where for $i = 1$ the subscript will be omitted. $P_X(\cdot)$ will denote a probability measure over $\mathcal{X}$. When there is no room for ambiguity, we will use $P(x)$ instead of $P_X(x)$. $\mathbb{1}\{A\}$ will denote the indicator of the event A.

We consider a Markov source producing a random sequence $X_1, X_2, ..., X_T$, $X_t \in \mathcal{X}$, $t = 1, 2, \ldots, T$. The cardinality of $\mathcal{X}$, as well as those of other alphabets in the sequel, is finite. The probability mass function of X_1, $P(x_1)$ and the transition probabilities, denoted by $P(x_t|x_{t-1})$, $t = 2, 3, \ldots, T$ are known.

Let $\mathcal{Y}$ denote the index set $\{1, 2, \ldots, M\}$ for some finite M. A variable–length stochastic encoder is a sequence of functions $\{f_t\}_{t=1}^T$. At stage t, a stochastic encoder uses all the causally available data, (X^t, Y^{t-1}), to choose a probability measure over $\mathcal{Y}$ from which Y_t is drawn. After drawing Y_t, the encoder noiselessly transmits an entropy–coded codeword of Y_t. A deterministic encoder is a stochastic encoder which draws a specific $Y_t \in \mathcal{Y}$ with probability 1 (i.e., Y_t is a deterministic function of (X^t, Y^{t-1})). Unlike the fixed rate regime in [1],[3], where $\log_2 |\mathcal{Y}|$ (rounded up) was the rate of the code at stage t, here the subset of $\mathcal{Y}$ used at each stage, along with the length of the binary representation of Y_t, will be subject to optimization.

The encoder structure is not confined *a–priori*, and at each time instant t, Y_t may be given by an arbitrary (possibly stochastic) function of (X^t, Y^{t-1}) as described above. The decoder, however, is assumed, similarly as in [1] and [3], to be a finite–memory device, defined as follows: At each stage, t, the decoder updates its current state (or memory) and outputs a reproduction symbol $\hat{X}_t$. We assume that the decoder state, Z_t, is updated by

$$\begin{aligned} Z_1 &= r_1(Y_1) \\ Z_t &= r_t(Y_t, Z_{t-1}), \quad t = 2, 3, \ldots, T \end{aligned} \tag{1}$$

Since the transmission is noiseless, Z_t can be tracked by the encoder. Note that this model also includes infinite memory, i.e., $Z_t = Y^t$. The reproduction symbols are produced by a sequence of functions $\{g_t\}$,

$g_t : \mathcal{Y} \times \mathcal{Z} \to \hat{\mathcal{X}}$ as follows

$$\begin{aligned} \hat{X}_1 &= g_1(Y_1) \\ \hat{X}_t &= g_t(Y_t, Z_{t-1}), \quad t = 2, 3, \ldots, T \end{aligned} \tag{2}$$

Since at the beginning of stage t, Z_{t-1} is known to both encoder and decoder, the entropy coder at every stage needs to encode the random variable Y_t given $Z_{t-1} = z_{t-1}$. We define $\mathcal{A}$ to be the set of all instantaneously uniquely decodable codes for $\mathcal{Y}$, i.e., all possible length functions $l : \mathcal{Y} \to \mathbb{Z}_+ \cup \infty$ that satisfy Kraft's inequality:

$$\mathcal{A} = \left\{ l(\cdot) : \sum_{y \in \mathcal{Y}} 2^{-l(y)} \leq 1 \right\}. \tag{3}$$

Note that we allow infinite–length codewords. We will return to this technical issue after properly defining the cost function. The average codeword length at stage t, for a specific decoder state z_{t-1}, will be given by:

$$L_{Y_t|Z_{t-1}}(z_{t-1}) \stackrel{\triangle}{=} \begin{cases} 0 & \text{if } \max_{y_t \in \mathcal{Y}} P(y_t|z_{t-1}) = 1 \\ \min_{l(\cdot) \in \mathcal{A}} \left\{ \sum_{y_t \in \mathcal{Y}} P(y_t|z_{t-1}) l(y) \right\} & \text{otherwise} \end{cases}. \tag{4}$$

i.e., if given $Z_{t-1} = z_{t-1}$, Y_t is deterministically known, there is no need to transmit any information, otherwise $L_{Y_t|Z_{t-1}}(z_{t-1})$ is obtained by designing a Huffman code for the probability distribution $P_{Y_t|Z_{t-1}}(\cdot|z_{t-1})$. Note that for given encoders and state update functions, $L_{Y_t|Z_{t-1}}(z_{t-1})$ is a function of z_{t-1} only. Also, $L_{Y_t|Z_{t-1}}(z_{t-1})$ is discontinuous around 0 in the distribution $P_{Y_t|Z_{t-1}}(\cdot|z_{t-1})$ since if given $Z_{t-1} = z_{t-1}$, Y_t is not deterministically known, then $L_{Y_t|Z_{t-1}}(z_{t-1}) \geq 1$.

The average codeword length of stage t, denoted $L_{Y_t|Z_{t-1}}$, is defined as $\boldsymbol{E} L_{Y_t|Z_{t-1}}(Z_{t-1})$, where the expectation is with respect to Z_{t-1}. Our system model is depicted in Figure 1.

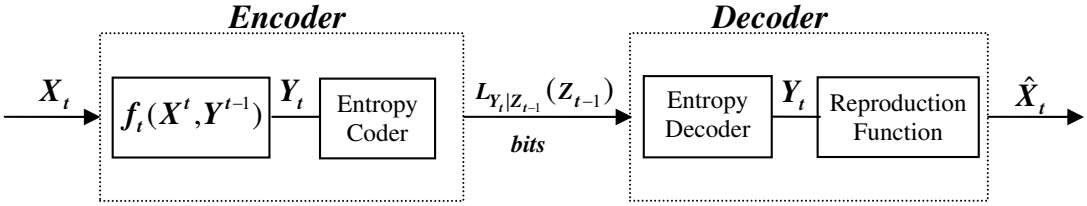

Fig. 1: System model

We are given a sequence of distortion measures $\{\rho_t\}_{t=1}^T$, $\rho_t : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. At each stage, the cost function is a linear combination of the average distortion and codeword length $L_{Y_t|Z_{t-1}}$, i.e.,

$$J_t \stackrel{\triangle}{=} \boldsymbol{E} \left\{ \rho_t(X_t, \hat{X}_t) + \lambda L_{Y_t|Z_{t-1}}(Z_{t-1}) \right\}, \tag{5}$$

where $\lambda > 0$ is a fixed parameter that controls the tradeoff between rate and distortion. Our goal is to

minimize the average cost

$$J \stackrel{\triangle}{=} \frac{1}{T}\sum_{t=1}^{T} J_t. \tag{6}$$

A sequence of encoders, $f_1, \ldots, f_T$, is said to be optimal if for a given sequence of decoders and memory update functions, $f_1, \ldots, f_T$ attains $\inf J$, where the infimum is over the set of all sequences of stochastic encoders, which are functions of all causally available data.

A stage–t encoder is said to be optimal if given the future stages encoders and decoders, it attains $\inf \sum_{i=t}^{T} J_i$, where the infimum is over the set of stochastic stage–t encoders (which are functions of (X^t, Y^{t-1})).

Note that for large enough λ, for some z_{t-1}, the optimal encoders might use only a small subset of $\mathcal{Y}$ (thus attaining higher distortion but smaller overall cost). Technically, this means that there will be a subset $\mathcal{B} \subset \mathcal{Y}$ such that $P(y_t|z_{t-1}) = 0$ if $y_t \in \mathcal{B}^c$. We therefore need that $\mathcal{A}$ will contain good codes for subsets of $\mathcal{Y}$. By allowing infinite length codewords, we make sure that $\mathcal{A}$ contains codes which are uniquely decodable for all subsets of $\mathcal{Y}$ (and satisfy Kraft's inequality for alphabet $\mathcal{Y}$). Needless to say that with this definition, a code for $\mathcal{B} \subset \mathcal{Y}$ will be used iff $P(y_t|z_{t-1}) = 0$ for all $y_t \in \mathcal{B}^c$, where we use $0 \cdot \infty = 0$.

III. Structure Theorems - No Side Information

A. Main results

We start by briefly stating and discussing the main contributions of this section. The proofs of the following theorems are found in the following subsections.

The first contribution of this paper is the following theorem, which basically states that the results of [1] continue to hold in this setting as well.

Theorem I. *For a Markov source and any given sequence of memory update functions $\{r_t\}$, reproduction functions $\{g_t\}$ and distortion measures $\{\rho_t\}$, there exists a sequence of deterministic encoders $Y_t = f_t(X_t, Z_{t-1})$ which is optimal.*

The addition of the variable–rate coding and allowing a larger class of encoders compared to [1], makes the proof of this result considerably more involved than its counterpart in [1], as was discussed at the end of Section I.

While Theorem I covers the infinite decoder memory ($Z_t = Y^t$) setting, in this case, when optimal reproduction functions are used (see Section III-E), we have the following theorem, which refine Theorem I for this case:

Theorem II. *For a Markov source and any sequence of distortion measures $\{\rho_t\}$ and optimal infinite memory decoders, there exists a sequence of deterministic encoders $Y_t = f_t(X_t, P_{X_t|Y^{t-1}}(\cdot|y^{t-1}))$ which is optimal.*

We will show that $P_{X_t|Y^{t-1}}(\cdot|y^{t-1})$ can be recursively updated. Theorem II is a refinement of Theorem I since, in the setup of Theorem II, there is no need to store the whole history of encoder outputs, Y^t, as

the statement of Theorem I, but instead, $P_{X_t|Y^{t-1}}(\cdot|y^{t-1})$ is recursively updated. (given that a probability measure can be stored).

In the remainder of this section, we will prove Theorems I and II, starting with Theorem I. In order to prove Theorem I, we need a few supporting lemmas, as in [1]. In the following two subsections, we state and prove the supporting lemmas and then prove Theorem I in Subsection III-D. Theorem II is proved in Subsection III-E.

B. Two–stage lemma

We start by analyzing a system with two stages only, where the first encoder is known.

Lemma I. *For any two–stage system ($T = 2$), there exists a deterministic second stage encoder $Y_2 = f_2(X_2, Z_1)$, which is optimal.*

Proof: Note that f_1, g_1, g_2, r_1 are fixed, and so, J_1 is unchanged by changing f_2. We need to show that a second stage encoder, that minimizes J_2, can be a deterministic function of (X_2, Z_1). Denote the set of stochastic encoders which are functions of (X_1, X_2, Y_1) by $\{f^s_{X^2Y_1}\}$. For every joint probability measure over the quadruple (X_1, X_2, Y_2, Z_1), J_2 is well defined and our objective is to find the optimal encoder that attains:

$$\inf_{\{f^s_{X^2Y_1}\}} J_2 = \inf_{\{f^s_{X^2Y_1}\}} \boldsymbol{E}\left\{\rho_2(X_2, g_2(Y_2, Z_1)) + \lambda L_{Y_2|Z_1}(Z_1)\right\}. \tag{7}$$

Consider the random quintuple $(X_1, X_2, Y_1, Y_2, Z_1)$ which takes part in the expectation of (7). From the structure of the system, we know that

$$P(x_1, x_2, y_1, y_2, z_1) = P(x_1)P(x_2|x_1)P(y_1|x_1)P(y_2|x_1, x_2, y_1)\mathbb{1}\left\{r_1(y_1) = z_1\right\}, \tag{8}$$

where we used the fact that z_1 is a deterministic function of y_1. Everything but the second stage encoder, which directly affects $P(y_2|x_1, x_2, y_1)$ is fixed. Note that the optimization affects $L_{Y_2|Z_1}(Z_1)$ since

$$L_{Y_2|Z_1}(z_1) = \min_{l(\cdot)\in\mathcal{A}} \sum_{y_2}\sum_{x_2} P(x_2|z_1)P(y_2|x_2, z_1)l(y_2) \tag{9}$$

and $P(y_2|x_2, z_1)$ depends on $P(y_2|x_1, x_2, y_1)$ as we will show shortly.

Let $\{f^s_{X_2Z_1}\}$ denote the subset of stochastic encoders which are functions of (X_2, Z_1). Also, let $\{f^d_{X_2Z_1}\}$ denote the subset of *deterministic* encoders which are functions of (X_2, Z_1). Since Z_1 is a function of Y_1, $\{f^d_{X_2Z_1}\} \subset \{f^s_{X_2Z_1}\} \subset \{f^s_{X^2Y_1}\}$. We prove Lemma I in two steps. First, we show that it is enough to search in the (infinite) subset $\{f^s_{X_2Z_1}\}$. In the second step, we show that among $\{f^s_{X_2Z_1}\}$, the optimal encoder is a member of $\{f^d_{X_2Z_1}\}$.

Step 1: We rewrite (7) as follows:

$$\begin{aligned}\inf_{\{f^s_{X^2Y_1}\}} J_2 &= \inf_{\{f^s_{X^2Y_1}\}} \sum_{x_2,y_2,z_1} P(x_2,y_2,z_1)\left[\rho_2(x_2,g_2(y_2,z_1)) + \lambda L_{Y_2|Z_1}(z_1)\right] \\ &= \inf_{\{f^s_{X^2Y_1}\}} \sum_{x_2,y_2,z_1} P(x_2,y_2,z_1)\times \\ &\quad \left[\rho_2(x_2,g_2(y_2,z_1)) + \lambda \min_{l(\cdot)\in\mathcal{A}_t} \sum_{y_2'}\sum_{x_2'} P(x_2'|z_1)P(y_2'|x_2',z_1)l(y_2')\right]. \end{aligned} \tag{10}$$

Now, given that the first stage encoder and decoder are known, $P(x_2,z_1)$ is well defined since

$$\begin{aligned} P(x_2,z_1) &= \sum_{x_1,y_1} P(x_1,x_2,y_1,z_1) \\ &= \sum_{x_1,y_1} P(x_1)P(x_2|x_1)P(y_1|x_1)\mathbb{1}\left\{r_1(y_1)=z_1\right\} \end{aligned} \tag{11}$$

and $P(x_1), P(x_2|x_1), \mathbb{1}\left\{r_1(y_1)=z_1\right\}$ are determined by the known source and first stage next–state function, $P(y_1|x_1)$ is directly determined by the first stage encoder. Also, by the Bayes rule, we have, for any second stage encoder:

$$P(x_2,y_2,z_1) = P(y_2|x_2,z_1)P(x_2,z_1). \tag{12}$$

Therefore,

$$\begin{aligned}\inf_{\{f^s_{X^2Y_1}\}} J_2 &= \inf_{\{f^s_{X^2Y_1}\}} \sum_{x_2,y_2,z_1} P(y_2|x_2,z_1)P(x_2,z_1)\times \\ &\quad \left[\rho_2(x_2,g_2(y_2,z_1)) + \lambda \min_{l(\cdot)\in\mathcal{A}_t} \sum_{y_2'}\sum_{x_2'} P(x_2'|z_1)P(y_2'|x_2',z_1)l(y_2')\right]. \end{aligned} \tag{13}$$

The only term that is affected by the optimization is $P(y_2|x_2,z_1)$. Observe that by (8), we have

$$\begin{aligned} P(y_2|x_2,z_1) &= \sum_{x_1,y_1} \frac{P(x_1,x_2,y_1,y_2,z_1)}{P(x_2,z_1)} \\ &= \sum_{x_1,y_1} \frac{P(x_1)P(x_2|x_1)P(y_1|x_1)P(y_2|x_1,x_2,y_1)\mathbb{1}\left\{r_1(y_1)=z_1\right\}}{P(x_2,z_1)}. \end{aligned} \tag{14}$$

From (13), (14), it is evident that the role of the second stage encoder in a two stage system is to select $P_{y_2|x_2,z_1}(\cdot|x_2,z_1)$ for every (x_2,z_1) so as to minimize the cost. To see this, note that every $f_2 \in \{f^s_{X^2Y_1}\}$ is mapped by (14) (through $P(y_2|x_1,x_2,y_1)$ for every (x_1,x_2,y_1)) to a point on the simplex of probability measures on $\mathcal{Y}$ for every (x_2,z_1). Namely, every $f_2 \in \{f^s_{X^2Y_1}\}$ is mapped to $\hat{f}_2 \in \{f^s_{X_2Z_1}\}$ and the optimization is affected only by $\hat{f}_2$. If instead of using a specific f_2 we will use $\hat{f}_2$ that results from it through (14), the joint probability $P(x_2,y_2,z_1)$ will remain the same and therefore, also the second stage cost. Also note that we cannot gain anything from optimizing only over $\{f^s_{X_2Z_1}\}$ and not $\{f^s_{X^2Y_1}\}$ since $\{f^s_{X_2Z_1}\}$ is completely covered by $\{f^s_{X^2Y_1}\}$ through (14). Therefore, since the optimization over $\{f^s_{X^2Y_1}\}$

is mapped to an optimization over $\{f^s_{X_2Z_1}\}$, we have

$$\inf_{\{f^s_{X^2Y_1}\}} J_2 = \inf_{\{f^s_{X_2Z_1}\}} J_2 \tag{15}$$

which completes the fist step of the proof.

Step 2: To complete the proof of the two–stage lemma, we need to show that it is enough to search in the finite space of *deterministic* encoders which are functions of (X_2, Z_1). Observe that the set of stochastic encoders is a convex set. The extreme points of this set (the points that are not convex combinations of other points) are deterministic encoders, namely, the set $\{f^d_{X_2Z_1}\}$. To complete the proof, we use the following lemma, proved in Appendix A.

Lemma II. *The stage t loss function is concave in $\{f^s_{X_tZ_{t-1}}\}$.*

Using Lemma II, we conclude that since we minimize a concave function over a convex set, the minimizer will be one of the extreme points of the set, i.e., a member of $\{f^d_{X_2Z_1}\}$. We thus showed that

$$\inf_{\{f^s_{X^2Y_1}\}} J_2 = \min_{\{f^d_{X_2Z_1}\}} J_2 \tag{16}$$

This completes the proof of the two–stage lemma. Note that no assumptions on the statistics of the source were made in the proof and therefore, the two–stage lemma holds for any source.■

Discussion:

1. Observe that the actual optimal encoding function for each (x_2, z_1) depends on the encoder of the first stage through $P(x_2, z_1)$ (which also governs $P(x_2|z_1)$), as seen from (13). This is true in general and not only in a two–stage system. The joint distribution $P_{X_t,Z_{t-1}}(\cdot,\cdot)$ can be thought of as the state of the system, governed by the choices of previous encoders (note however, that this state is static in the sense that it is not influenced by the actual realization of the source sequence). Therefore, the role of the stage t encoder, besides greedily minimizing the stage t cost (given the state $P_{X_t,Z_{t-1}}(\cdot,\cdot)$), is to control the future states so that they will allow minimal costs in future stages. This is true for all but the last encoder, which does not affect future cost, as seen for the second stage encoder in a two stage system. We will come back to this issue in Subsection III-E when we deal with infinite memory decoders and apply tools of stochastic control.

2. It is not surprising that the optimal second stage cost is attained by a deterministic encoder. Since the second stage is the last stage, the last encoder does not affect future costs and therefore, instead of using a convex combination of deterministic encoders (i.e., a stochastic encoder), use only the one with the best performance. However, in a system with more stages, it is not immediately clear that deterministic encoders in intermediate stages are optimal. In fact, this is also true for the first stage of a two stage system. We saw that the first stage affects the second stage cost through $P(x_2, z_1)$. Specifically, it affects the second stage cost through $P(x_2|z_1)$, (as seen in (10)) which is non linear in the first stage encoder $P(y_1|x_1)$ since

$$P(x_2|z_1) = \frac{\sum_{x_1,y_1} P(x_1,x_2)P(y_1|x_1)\mathbb{1}\left\{r_1(y_1)=z_1\right\}}{\sum_{x_1,x_2,y_1} P(x_1,x_2)P(y_1|x_1)\mathbb{1}\left\{r_1(y_1)=z_1\right\}}. \tag{17}$$

If the first–stage encoder is deterministic, there is only a finite number of possible $P_{X_2|Z_1}(\cdot,\cdot)$. Assume

that we use a stochastic first–stage encoder, f_1^s. Although by Lemma II, f_1^s is sub–optimal for J_1, can it allow us to reach a $P'_{X_2|Z_1}(\cdot,\cdot)$, unreachable by deterministic encoders, that will be favorable in terms of J_2 and yield a lower overall cost? We show in the sequel that the answer is negative and that the optimal first–stage encoder is deterministic as well. We will show that the *stage t cost is a concave functional in the choices of the previous stages encoders*. The proof of the last statement is much more involved than the proof of Lemma II and it is discussed in the next subsection. In [1], [3], the stage t distortion is linear in the choice of the encoders at all previous stages (since the expectation is linear and the non–linear element of the codeword length was not present). Therefore, there was no loss of optimality in a-priori confining the encoders to be deterministic. We further address this issue in the following subsection which deal with a more complex system.

Corollary I. *In any T–stage system ($T \geq 2$) there exists a deterministic last stage encoder $Y_T = f_T(X_T, Z_{T-1})$, which is optimal.*

Proof: Let $\hat{X}_1 \stackrel{\triangle}{=} (X_1, X_2, \ldots, X_{T-1}), \hat{X}_2 \stackrel{\triangle}{=} X_T, \hat{Z}_1 = Z_{T-1}$, where $\hat{Z}_1$ is calculated recursively according the the encoding functions that operate on $\hat{X}_1$ and the resulting $Y_1, \ldots, Y_{T-1}$. We now apply the two–stage lemma to this system to conclude that the last stage encoder is a deterministic function of (X_t, Z_{T-1}). ■

C. Three–stage lemma

Lemma III. *In a three-stage system ($T = 3$) with a Markov source, if the third–stage encoder is a deterministic function of (X_3, Z_2), then there exists a deterministic second stage encoder $Y_2 = f_2(X_2, Z_1)$, which is optimal.*

Proof of Lemma III: We define, as in Subsection III-B, $\{f^s_{X^2Y_1}\}$ to be the set of all possible stochastic second–stage encoders. Let $\{f^s_{X_2Z_1}\} \subset \{f^s_{X^2Y_1}\}$ be the set that contains all stochastic second stage encoders that are functions of (X_2, Z_1) and finally, let $\{f^d_{X_2Z_1}\} \subset \{f^s_{X_2Z_1}\}$ denote the set of deterministic encoders which are functions of (X_2, Z_1). Since the first–stage is fixed, J_1 is unaffected by changing the second stage encoder. Our goal is to jointly optimize $(J_2 + J_3)$ with respect to the second stage encoder and show that

$$\inf_{\{f^s_{X^2Y_1}\}} (J_2 + J_3) = \min_{\{f^d_{X_2Z_1}\}} (J_2 + J_3)\,. \tag{18}$$

Since the third stage encoder is known, the expected third stage cost for any second stage encoder is given by

$$\begin{aligned} J_3 &= \boldsymbol{E}\left\{\rho(X_3, g(Y_3, Z_2)) + L_{Y_3|Z_2}(Z_2)\right\} \\ &= \sum_{x_3,y_3,z_2} P(x_3,z_2)P(y_3|x_3,z_2)\left[\rho(x_3, g(y_3,z_2)) + L_{Y_3|Z_2}(z_2)\right] \\ &= \sum_{x_3,y_3,z_2} P(x_3,z_2)\mathbb{1}\left\{f_3(x_3,z_2) = y_3\right\}\rho(x_3, g(y_3,z_2)) + \sum_{z_2} P(z_2) \min_{l(\cdot)\in\mathcal{A}} \sum_{x_3,y_3} \mathbb{1}\left\{f_3(x_3,z_2) = y_3\right\} P(x_3|z_2) l(y_3). \end{aligned} \tag{19}$$

The second–stage encoder affects the last expression through $P(x_3, z_2)$ (and thus also through $P(z_2)$ and

$P(x_3|z_2)$) since

$$P(x_3, z_2) = \sum_{x_2,y_2,z_1} P(x_2, x_3, y_2, z_1, z_1)$$
$$= \sum_{x_2,y_2,z_1} P(x_2, z_1)P(y_2|x_2, z_1)P(x_3|x_2)\mathbb{1}\left\{r_2(y_2, z_1) = z_2\right\} \tag{20}$$

where $P(x_2, z_1)$ is the result of the first–stage and we used the fact that the source is Markov and that z_2 is a deterministic function of (y_2, z_1). Therefore, as we saw in Subsection III-B, the optimization affects the third–stage only through $P(y_2|x_2, z_1)$ for all (x_2, y_2, z_1). We saw in (13) that the second stage cost can be written as:

$$J_2 = \sum_{x_2,y_2,z_1} P(y_2|x_2, z_1)P(x_2, z_1)\times$$
$$\left[\rho_2(x_2, g_2(y_2, z_1)) + \lambda \min_{l(\cdot)\in\mathcal{A}_t} \sum_{y_2'} \sum_{x_2'} P(x_2'|z_1)P(y_2'|x_2', z_1)l(y_2')\right]. \tag{21}$$

where $P(x_2, z_1)$ and thus $P(x_2|z_1)$ are the result of the first–stage encoder. We see that the optimization in the l.h.s of (18) affects both the second and third stage costs only through the conditional probabilities $P(y_2|x_2, z_1)$, for all (x_2, y_2, z_1). Repeating the arguments used in the proof of Lemma I, instead of using a specific $f_2 \in \{f^s_{X^2Y_1}\}$, we can use $\hat{f}_2 \in \{f^s_{X_2Z_1}\}$ that results from it through (14), to draw Y_2. Since $P(x_2, y_2, z_1)$ will remain the same,

$$P(x_3, z_2) = \sum_{x_2,y_2,z_1} P(x_2, z_1)P(y_2|x_2, z_1)P(x_3|x_2)\mathbb{1}\left\{r_2(y_2, z_1) = z_2\right\} \tag{22}$$

will also remain the same and $(J_2 + J_3)$ will not be affected by this step. We therefore have

$$\inf_{\{f^s_{X^2Y_1}\}} (J_2 + J_3) = \inf_{\{f^s_{X_2Z_1}\}} (J_2 + J_3). \tag{23}$$

As in the two stage lemma, we need to show that it is enough to search in the space of deterministic encoders, $\{f^d_{X_2Z_1}\}$. Here, we have to show that both the second stage cost *and* the third stage cost are concave in $\{f^s_{X_2Z_1}\}$. We know that the second stage cost is concave in $\{f^s_{X_2Z_1}\}$ from Lemma II. The following lemma asserts that the third stage cost is concave in $\{f^s_{X_2Z_1}\}$.

Lemma IV. *The third stage cost, J_3, is concave functional of $\{f^s_{X_2Z_1}\}$.*

The proof Lemma IV is much more involved than the proof of Lemma II and can be found in Appendix B.

Using lemma IV, we conclude that $(J_2 + J_3)$, which is the sum of two concave functionals, is concave in $\{f^s_{X_2Z_1}\}$. Therefore, the minimizer will be one of the extreme points of the convex set of $\{f^s_{X_2Z_1}\}$, namely, a member of $\{f^d_{X_2Z_1}\}$. We showed that

$$\inf_{\{f^s_{X_2Z_1}\}} (J_2 + J_3) = \min_{\{f^d_{X_2Z_1}\}} (J_2 + J_3) \tag{24}$$

Using (23), we arrive at (18) which completes the proof of Lemma III. ■

D. Proof of Theorem I

With the two– and three–stage lemmas, we can prove Theorem I by using the method of [1], used for fixed rate encoding. Theorem I is proven by backward induction. First apply Corollary I to any system to conclude that the optimal f_T is a deterministic function of (X_T, Z_{T-1}). Now assume that the last m encoders $f_{T-m+1}, ..., f_T$ are deterministic functions of $(X_{T-m+1}, Z_{T-m}), \ldots, (X_T, Z_{T-1})$, respectively. We will show that the encoder at time $(T-m)$ also has this structure and continue backwards until $t = 2$. The first encoder is trivially a function of X_1 and by lemma IV (with Z_0 as a constant) it is also deterministic. Let

$$\begin{aligned}
\hat{X}_1 &= (X_1, X_2, ..., X_{T-m-1}),\\
\hat{Y}_1 &= (Y_1, Y_2, ..., Y_{T-m-1}),\\
\hat{Z}_1 &= \hat{r}_1(\hat{Y}_1),\\
\hat{X}_2 &= X_{T-m},\\
\hat{Y}_2 &= Y_{T-m},\\
\hat{Z}_2 &= r_{T-m}(\hat{Y}_2, \hat{Z}_1),\\
\hat{X}_3 &= (X_{T-m+1}, X_{T-m+2}, ..., X_T),\\
\hat{Y}_3 &= (Y_{T-m+1}, Y_{T-m+2}, ..., Y_T),
\end{aligned} \tag{25}$$

where $\hat{Z}_1$ is recursively calculated from $\hat{Y}_1$ and it represents the state of the decoder after $(T-m-1)$ stages. Using this new notation, the encoder that produces $\hat{Y}_3$ is a deterministic function of $(\hat{X}_3, \hat{Z}_2)$ (since, by assumption, the last m encoders have the desired structure). The source is Markov since $\hat{X}_3$ is independent of $\hat{X}_1$ given $\hat{X}_2$ (since the original source is Markov). Now, by the three–stage lemma, $\hat{Y}_2 = Y_{T-m} = f_{T-m}(\hat{X}_2, \hat{Z}_1^y) = f_{T-m}(X_{T-m}, Z_{T-m-1})$. Thus, the induction step is proved. This completes the proof of Theorem I.■

Remark: Theorem I can be extended to a k-order Markov source using Witsenhausen's method [1]. Namely, for a k-order Markov source, define $\tilde{X}_1 = (X_1, X_2, \ldots, X_k)$, $\tilde{X}_2 = (X_2, X_3, \ldots, X_{k+1})$ and so on. Now, $\tilde{X}_t$ is a Markov source. Using Theorem I, we can conclude that the optimal encoder is a function of the last k source symbols and the state of the decoder.

E. Infinite memory decoder - proof of Theorem II

In this section, we deal with the case where the decoder has infinite memory, i.e., $Z_t = Y^t$. The memory update functions $\{r_t\}$ in this case are only appending the new received index Y_t to Z_{t-1}. Note that this scenario is covered by Theorem I, however, in this case we can be more specific regarding the role of Y^t at the encoder. While Theorem I was true for any decoding rule, Theorem II is true only for the optimal reproduction function. We define the *Bayes Envelope* as

$$B(P_{X_t|Y^t}) \stackrel{\triangle}{=} \min_{\hat{x}_t} \sum_{x_t} P(x_t|y^t)\rho_t(x_t, \hat{x}_t). \tag{26}$$

The minimizer of the last expression is called the *Bayes–response* and will be denoted by $\hat{X}_{Bayes}(P_{X_t|Y^t})$. Clearly, $\hat{X}_{Bayes}(P_{X_t|Y^t})$ is a function of $P_{X_t|Y^t}(\cdot|y^t)$ and the cost function, ρ_t. The fact that the optimal

reproduction function is the Bayes–response was shown in many places, for example [3],[8, Lemma 3].

When infinite memory is available, we can use tools from Markov decision processes (MDP's) in order to derive a structure theorem. In Appendix C, we provide a brief background on MDP's. By Theorem I, we know that we can confine the discussion to deterministic encoders without loss of optimality. We need to show that our original problem can be represented as a MDP. The proof of Theorem II will follow immediately from Theorem A.1, given in Appendix C. In order to show that we have an MDP, as we discuss in Appendix C, we need to show that:

- We can a find a sequence of deterministic functions $\{\gamma_t\}$, along with two finite spaces, $\mathcal{S}, \mathcal{A}$, such that the average cost, defined by (5),(6), can be written as $J = \frac{1}{T}\boldsymbol{E}\sum_{t=1}^{T}\gamma_t(s_t, a_t)$, where $s_t \in \mathcal{S}, a_t \in \mathcal{A}$ are the system state and the action taken by the decision maker at stage t, respectively.
- The next state is chosen according to $P(s_{t+1}|s^t, a^t) = P(s_{t+1}|s_t, a_t)$, i.e., the state is Markov conditioned on a_t.

We define our state as $s_t = P_{X_t|Y^{t-1}}(\cdot|y^{t-1})$ and our actions $a_t : \mathcal{X} \to \mathcal{Y}$. We note that for every history x^{t-1}, the general deterministic encoder (which is a function of x^t) is a mapping from x_t to y_t. Our action, a_t, is this mapping. Since there is only a finite number of mappings from X_t to Y_t, our action space is finite. Our state space is also finite. This is true since we consider only deterministic encoders, from which there is only a finite number. Therefore, at each stage, there is only a finite number of possible $P_{X_t|Y^{t-1}}$. This means that the cardinality of the state alphabet, grows with the time horizon T. Note however, that the decoder's state alphabet, $\mathcal{Z}_t = \mathcal{Y}^t$, grows as well in this case. We start by showing that the cost function can be written as a function of the current state and action. Treating the codeword length first:

$$\begin{aligned}
L_{Y_t|Y^{t-1}}(y^{t-1}) &= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t} P(y_t|y^{t-1}) l(y_t) \\
&= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t, x_t} P(y_t, x_t|y^{t-1}) l(y_t) \\
&= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t, x_t} P(x_t|y^{t-1}) P(y_t|x_t, y^{t-1}) l(y_t) \\
&= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t, x_t} P(x_t|y^{t-1}) \mathbb{1}\left\{a_t(x_t) = y_t\right\} l(y_t) \\
&\triangleq \alpha_t(s_t, a_t),
\end{aligned} \tag{27}$$

where the equation preceding the last one is true since we know the function from x_t to y_t. We now move on to the average distortion. We first show that the optimal reproduction function, $\hat{X}_{Bayes}(P_{X_t|Y^t})$, is a function of (a_t, s_t, y_t). To see this note that

$$\begin{aligned}
P(x_t|y^t) &= \frac{P(x_t, y_t|y^{t-1})}{\sum_{x_t} P(x_t, y_t|y^{t-1})} \\
&= \frac{P(x_t|y^{t-1}) \mathbb{1}\left\{a_t(x_t) = y_t\right\}}{\sum_{x_t} P(x_t|y^{t-1}) \mathbb{1}\left\{a_t(x_t) = y_t\right\}} \\
&\triangleq f(s_t, a_t, x_t, y_t).
\end{aligned} \tag{28}$$

Therefore, the optimal reproduction function, which is a function of $P_{X_t|Y^t}(\cdot|y^t)$, is a function of (s_t, a_t, y_t),

i.e., $\hat{X}_{Bayes}(P_{X_t|Y^t}) = g_t^*(s_t, a_t, y_t)$. Using this notation we have

$$\begin{aligned}
\boldsymbol{E}\left[\rho(X_t, g_t^*(s_t, a_t, Y_t))\middle| Y^{t-1} = y^{t-1}\right] &= \sum_{x_t, y_t} P(x_t, y_t|y^{t-1})\rho(x_t, g_t^*(s_t, a_t, y_t)) \\
&= \sum_{x_t, y_t} P(x_t|y^{t-1})P(y_t|x_t, y^{t-1})\rho(x_t, g_t^*(s_t, a_t, y_t)) \\
&= \sum_{x_t, y_t} P(x_t|y^{t-1})\mathbb{1}\left\{a_t(x_t) = y_t\right\}\rho(x_t, g_t^*(s_t, a_t, y_t)) \\
&\triangleq \beta_t(s_t, a_t). \qquad (29)
\end{aligned}$$

Denoting $\beta_t(s_t, a_t) + \lambda\alpha_t(s_t, a_t) = \gamma_t(s_t, a_t)$, our optimality criterion can be written as $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{E}\gamma_t(s_t, a_t)$.

We move on to show that the state sequence is Markov conditioned on the action, namely, $P(s_{t+1}|s^t, a^t) = P(s_{t+1}|s_t, a_t)$. We start by noting that $s_{t+1} = P_{X_{t+1}|Y^t}(\cdot|y^t)$ is a function of (a_t, s_t, y_t). For every x_{t+1}, we have

$$\begin{aligned}
P(x_{t+1}|y^t) &= \frac{\sum_{x_t} P(x_{t+1}, x_t, y_t|y^{t-1})}{\sum_{x_t, x_{x+1}} P(x_{t+1}, x_t, y_t|y^{t-1})} \\
&= \frac{\sum_{x_t} P(x_t|y^{t-1})P(x_{t+1}|x_t, y^{t-1})P(y_t|x_t, x_{t+1}, y^{t-1})}{\sum_{x_t, x_{x+1}} P(x_{t+1}, x_t, y_t|y^{t-1})} \\
&= \frac{\sum_{x_t} P(x_t|y^{t-1})P(x_{t+1}|x_t)\mathbb{1}\left\{a_t(x_t) = y_t\right\}}{\sum_{x_t, x_{x+1}} P(x_{t+1}, x_t, y_t|y^{t-1})} \\
&= \frac{\sum_{x_t} P(x_t|y^{t-1})P(x_{t+1}|x_t)\mathbb{1}\left\{a_t(x_t) = y_t\right\}}{\sum_{x_t, x_{x+1}} P(x_t|y^{t-1})P(x_{t+1}|x_t)\mathbb{1}\left\{a_t(x_t) = y_t\right\}} \\
&\triangleq f(a_t, s_t, x_{t+1}, y_t). \qquad (30)
\end{aligned}$$

Therefore, $s_{t+1} = h(a_t, s_t, y_t)$, for a function h that uses (30) for every x_{t+1}. Now,

$$\begin{aligned}
P(s_{t+1} = \nu|s^t, a^t) &= \sum_{y_t, x_t} \mathbb{1}\left\{h(a_t, s_t, y_t) = \nu\right\}\mathbb{1}\left\{a_t(x_t) = y_t\right\}P(x_t|y^{t-1}) \\
&= P(s_{t+1} = \nu|s_t, a_t), \qquad (31)
\end{aligned}$$

since the current prior on x_t is given. We showed that our system can be represented as an MDP. By invoking Theorem A.1, we know that the optimal action at each stage, a_t, is a deterministic function of the state. Namely, The mapping from x_t to y_t can be chosen deterministically as a function of $P_{X_t|Y^{t-1}}(\cdot|y^{t-1})$. Therefore, Y_t is a deterministic function of $(X_t, P_{X_t|Y^{t-1}}(\cdot|y^{t-1}))$, which concludes the proof of Theorem II. Since the state can be recursively calculated (see eq. (30)), the encoder does not need to store y^{t-1} but rather a probability measure (a vector in $\mathbb{R}^{|\mathcal{Y}|}$).

IV. Markov Memory Update Functions

A. Preliminaries and main result

In Section III, we showed that for given memory update, distortion and reproduction functions, there is no loss of optimality if the encoders use the current source symbol and the state of the decoder, which

they track. We will refer to this class of encoders as *tracking encoders*. In the overall optimization of the system, there is still the task of finding the best memory update and reproduction functions at each stage. When the memory update functions and encoders are fixed, as we discussed in Section III-E, the reproduction function should output the $\hat{X}_t$ that minimizes the average distortion for a given (Y_t, Z_{t-1}), i.e., the Bayes response of $P_{X_t|Y_t,Z_{t-1}}$. This is simple since the reproduction function has no influence on the future costs (cost to go) and it affects only the present distortion (in a way, for the same reasons, the two–stage lemma was simpler than the three–stage lemma). However, similarly to the encoders, the memory update function at stage t affects all future costs. In this section, we show that for a "small" cost at each stage, one can take Markov memory update functions, defined as sliding windows over the received symbols at the decoder and avoid the search for the $|\mathcal{Z}|$–states optimal memory update functions. The extra cost is a function of $|\mathcal{Z}|$ and the sliding window size only and it vanishes as the window size is increased.

Let

$$\Delta_{|\mathcal{Z}|} = \min_{\{r_t\}\{f_t\},\{g_t\}} \boldsymbol{E}\frac{1}{T}\left\{\sum_{t=1}^{T}\left[\rho(X_t, g_t(Y_t, Z_{t-1})) + \lambda L_{Y_t|Z_{t-1}}(Z_{t-1})\right]\right\} \tag{32}$$

where the minimization is over all next state functions $\{r_t\}$ with a state set of size $|\mathcal{Z}|$ and all decoders and tracking encoders that use. Note that we choose here the whole sequence of next–state functions, encoders and decoders for $t = 1, 2, \ldots, T$. We say that the state is Markov of length l, if $Z_t = \{Y_{t-l}, \ldots, Y_{t-1}\}$, i.e., a sliding window of length l on the encoder outputs. Let

$$\tilde{\Delta}_l = \min_{\{\tilde{f}_t\},\{\tilde{g}_t\}} \boldsymbol{E}\frac{1}{T}\sum_{t=1}^{T}\left[\rho(X_t, \tilde{g}_t(Y_t, Y_{t-l}^{t-1})) + \lambda L_{Y_t|Y_{t-l}^{t-1}}(Y_{t-l}^{t-1})\right]. \tag{33}$$

where here, the minimization is with respect to all decoders and tracking encoders that use a Markov state of length l.

Theorem III. *For any source statistics, when considering only tracking encoders, we have for any l that divides T:*

$$\Delta_{\mathcal{Z}} \geq \tilde{\Delta}_l - \lambda\frac{\log|\mathcal{Z}|}{l} \tag{34}$$

The significance of this theorem is more conceptual than operational. The system on the r.h.s might require more memory than the system on the l.h.s. and the search for the optimal encoders becomes more complex as l increases. However, the system on the r.h.s is conceptually simpler and analytically more tractable since the memory structure is simple.

Combining Theorem III with Theorem I we have the following theorem:

Theorem IV. *For a Markov source, there exists a system with deterministic encoders $Y_t = f_t(X_t, Z_{t-1})$ and Markov memory update functions with a performance loss no greater than $\lambda\frac{\log|\mathcal{Z}|}{l}$ per source symbol, compared to the optimal system.*

Theorem III can be extended to the case where instead of our Lagrangian cost function, we would look

for the minimal average distortion subject to an average length constraint. Let

$$\begin{aligned}\Delta_{\mathcal{Z}}(R) &\stackrel{\triangle}{=} \min_{\{f_t\},\{g_t\},\{r_t\}} \boldsymbol{E}\left\{\frac{1}{T}\sum_{t=1}^{T}\rho(X_t, g_t(Y_t, Z_{t-1}))\right\} \\ &\qquad s.t \qquad \boldsymbol{E}\left\{\frac{1}{T}\sum_{t=1}^{T}L_{Y_t}(Z_{t-1})\right\} < R \\ \tilde{\Delta}_l(R) &\stackrel{\triangle}{=} \min_{\{f_t\},\{g_t\}} \boldsymbol{E}\left\{\frac{1}{T}\sum_{t=1}^{T}\rho(X_t, g_t(Y_{t-l}^{t}))\right\} \\ &\qquad s.t \qquad \boldsymbol{E}\left\{\frac{1}{T}\sum_{t=1}^{T}L_{Y_t}(Y_{t-l}^{t-1})\right\} < R \end{aligned} \tag{35}$$

where the minimization is over all tracking encoders that use X_t and the decoder's state, reproduction functions and state update functions (in $\Delta_{\mathcal{Z}}(R)$ only). We have the following theorem:

Theorem V. *In the constrained setting, for any l that divides T we have*

$$\Delta_{\mathcal{Z}}(R) \geq \tilde{\Delta}_l\left(R + \frac{\log|\mathcal{Z}|}{l}\right) \tag{36}$$

Note that here we do not have a theorem in the spirit of Theorem IV since we did not show that in this case, tracking encoders are optimal.

In the next subsection, we prove Theorem III. Theorem IV is a direct consequence of Theorem I and Theorem III combined. Theorem V is proven exactly in the same manner as Theorem III and its proof is therefore, omitted. Theorem III is valid even without taking expectations in (32),(33) and therefore, it is also valid for individual sequences (see [13]). Theorems III–V will also hold in the setting of the Section V, where SI is available to the decoder.

B. Proof of Theorem III:

The ideas in the proof rely on some ideas from [13]. Fix the optimal encoders, state update and reproduction functions of $\Delta_{|\mathcal{Z}|}$. We start by focusing on the codeword length element of $\Delta_{|\mathcal{Z}|}$, using the fact that conditioning reduces the length element (see Appendix D), we have

$$\begin{aligned} R &\stackrel{\triangle}{=} \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{E}L_{Y_t|Z_{t-1}}(Z_{t-1}) \\ &\geq \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{E}L_{Y_t|Y_{t-l}^{t-1},Z_{t-1}}(Y_{t-l}^{t-1}, Z_{t-1}) \end{aligned} \tag{37}$$

Since we will always deal with the expected codeword length, in order to simplify the notation, we will use from now $\boldsymbol{E}L_{Y_t|Z_{t-1}}(Z_{t-1}) \stackrel{\triangle}{=} L_{Y_t|Z_{t-1}}$ (as defined in Section II). We now add conditioning on $Z_0, Z_l, Z_{2l}, \ldots$ which will further reduce the last expression. Z_0 is added to the first l summands of (37), Z_l to the summands indexed by $l+1, \ldots, 2l$, and so on. This conditioning makes the conditioning on Z_{t-1} redundant since if we know the state in the past and the encoder outputs up to the present, we know

the current state as well. We continue by assuming that l divides T:

$$\begin{aligned} R &\geq \frac{1}{T}\sum_{t=1}^{T} L_{Y_t|Y_{t-l}^{t-1},Z_{t-1}} \\ &\geq \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+1}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} \end{aligned} \tag{38}$$

Now there are two types of terms:

1) (Y_{jl+1}, Z_{jl}) appear together in the conditioning.
2) Y_{jl+1} is conditioned on Z_{jl} and the previous block: $Y_{j(l-1)+1}^{jl}$.

We rewrite the sum of (38) as two sums, pertaining to the above two types:

$$\begin{aligned} R &\geq \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+1}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} \\ &= \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} + \frac{1}{T}\sum_{j=0}^{T/l-1} L_{Y_{jl+1}|Y_{j(l-1)+1}^{jl},Z_{jl}}. \end{aligned} \tag{39}$$

We now use the following inequality, which is proved in Appendix D:

$$L_{Y_{jl+1}|Y_{j(l-1)+1}^{jl},Z_{jl}} \geq L_{Y_{jl+1},Z_{jl}|Y_{j(l-1)+1}^{jl}} - \log|\mathcal{Z}| \tag{40}$$

Substituting (40) in (39), we have:

$$\begin{aligned} R &\geq \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} + \frac{1}{T}\sum_{j=0}^{T/l-1} L_{Y_{jl+1}|Y_{j(l-1)+1}^{jl},Z_{jl}} \\ &\geq \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} + \frac{1}{T}\sum_{j=0}^{T/l-1} L_{Y_{jl+1},Z_{jl}|Y_{j(l-1)+1}^{jl}} - \frac{\log|\mathcal{Z}|}{l} \\ &\geq \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} L_{Y_t|Y_{t-l}^{t-1},Z_{jl}} + \frac{1}{T}\sum_{j=0}^{T/l-1} L_{Y_{jl+1},Z_{jl}|Y_{j(l-1)+1}^{jl},Z_{j(l-1)}} - \frac{\log|\mathcal{Z}|}{l}. \end{aligned} \tag{41}$$

Regarding the distortion element of $\Delta_{|\mathcal{Z}|}$, we have:

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E}\rho(X_t, g_t(Y_t, Z_{t-1})) \\ &= \frac{1}{T}\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} \boldsymbol{E}\rho(X_t, g_t(Y_t, Z_{t-1})) + \frac{1}{T}\sum_{j=0}^{T/l-1} \boldsymbol{E}\rho(X_{jl+1}, g_t(Y_{jl+1}, Z_{jl})) \\ &\geq \frac{1}{T}\min_{\{g_t\}}\left\{\sum_{j=0}^{T/l-1}\sum_{t=jl+2}^{jl+l} \boldsymbol{E}\rho(X_t, g_t(Y_{t-l}^{t}, Z_{jl})) + \sum_{j=0}^{T/l-1} \boldsymbol{E}\rho(X_{jl+1}, g(Y_{j(l-1)+1}^{jl+1}, Z_{jl}, Z_{j(l-1)}))\right\}. \end{aligned} \tag{42}$$

In the last inequality, we used the fact that with the same encoders (same $\{Y_t\}$), optimal decoders that use more data will do at least as well as the original decoders. Note that in the above derivation, (Y_{jl+1}, Z_{jl})

always appear together. Therefore, we set for all $j = 0, 1, \ldots, n/l - 1$ $Y'_{jm+1} = (Y_{jm+1}, Z_{jl})$ and for all other indexes we set $Y'_t = Y_t$. Using this notation, we have for (41):

$$R \geq \frac{1}{T}\sum_{t=1}^{n} L_{Y'_t}(Y'^{t-1}_{t-l}) - \frac{\log|\mathcal{Z}|}{l}. \tag{43}$$

and for (42)

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E}\rho(X_t, g_t(Y_t, Z_{t-1})) \geq \min_{\{g_t\}} \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E}\rho(X_t, g_t(Y'^{t}_{t-l})). \tag{44}$$

and each Y'_t is a function of X_t, Y'^{t-1}_{t-l}. Note that although the size of the alphabet of Y' is now $|\mathcal{Y}| \times |\mathcal{Z}|$, the size of the alphabet was not a constraint on the system and was introduced so it will be convenient to define $\mathcal{A}$. The fact that it is now larger does not change any of the results obtained in the previous sections. We have

$$\Delta_{\mathcal{Z}} \geq \min_{\{g_t\}} \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{E}\rho(X_t, g_t(Y'^{t}_{t-l})) + \lambda\left(\frac{1}{T}\sum_{t=1}^{n} \boldsymbol{E}L_{Y'_t|Y'^{t-1}_{t-l}}(Y'^{t-1}_{t-l}) - \frac{\log|\mathcal{Z}|}{l}\right). \tag{45}$$

The r.h.s of the above equation was calculated with the optimal encoders of the l.h.s. with a scheme that appends the original decoder state once every block. This is, of course, only one of the possible schemes for Markovian states and therefore if we optimize the r.h.s over all encoders that use a Markovian state of length l we get

$$\Delta_{\mathcal{Z}} \geq \tilde{\Delta}_l - \lambda\frac{\log|\mathcal{Z}|}{l} \tag{46}$$

■

V. Side Information at the Decoder

A. Preliminaries and main result

In this section, we assume that the decoder has access to SI. The SI sequence, $W_1, W_2, ..., W_T$, $W_t \in \mathcal{W}$, is generated by a discrete memoryless channel (DMC), fed by $X_1, X_2, \ldots, X_T$:

$$P(w_1, \ldots, w_T|x_1, \ldots x_T) = \prod_{t=1}^{T} P(w_t|x_t).$$

For simplicity, we assume that $P(w|x) > 0$ for all $X \in \mathcal{X}$ and $W \in \mathcal{W}$. Our results, however, will continue to hold without this assumption with minor changes to the length function (see [14]). The SI is used both in the reproduction function and in the state update function. We assume that the state now consists of two sub–states. The first, $Z^y_t \in \mathcal{Z}^y$, is independent of the SI and is updated as in Section II. The second, $Z^w_t \in \mathcal{Z}^w$, is updated by

$$\begin{aligned} Z^w_1 &= r^w_1(W_1, Y_1), \\ Z^w_t &= r^w_t(W_t, Y_t, Z^w_{t-1}), \qquad t = 2, 3, \ldots, T. \end{aligned} \tag{47}$$

The reproduction symbols are produced by a sequence of functions $\{g_t\}$, $g_t : \mathcal{Y} \times \mathcal{W} \times \mathcal{Z}^w \times \mathcal{Z}^y \to \hat{\mathcal{X}}$ as follows:

$$\begin{aligned} \hat{X}_1 &= g_1(W_1, Y_1), \\ \hat{X}_t &= g_t(W_t, Y_t, Z^w_{t-1}, Z^y_{t-1}), \qquad t = 2, 3, \ldots, T. \end{aligned} \tag{48}$$

Since Z^w_t is not known at the encoder, it cannot be used by the variable–length encoder and thus the cost function is now given by

$$J_t = \boldsymbol{E}\left\{\rho_t(X_t, g_t(W_t, Y_t, Z^w_{t-1}, Z^y_{t-1})) + \lambda L_{Y_t}(Z^y_{t-1})\right\} \tag{49}$$

Let $B_t \stackrel{\triangle}{=} P_{Z^w_t|X^t}(\cdot|X^t)$ and $b_t \stackrel{\triangle}{=} P_{Z^W_t|X^t}(\cdot|x^t)$, i.e., $b_t \in \mathbb{R}^{\mathcal{Z}^w}$ is a probability measure over the sub–state of the decoder, Z^w_t, which is not known to the encoder. Note that since the decoder does not have access to x^t, b_t is not known to the decoder. Our system model with SI is depicted in Figure 2.

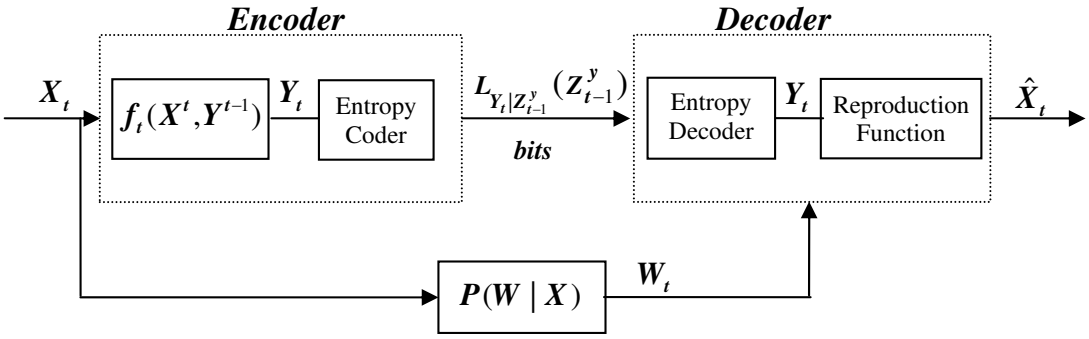

Fig. 2: System model with SI.

The following two theorems are the contribution of this section.

Theorem VI. *For a Markov source and any given sequence of memory update functions $\{r_t\}$, reproduction functions $\{g_t\}$ and distortion measures $\{\rho_t\}$, there exists a sequence of deterministic encoders $Y_t = f_t(B_{t-1}, X_t, Z^y_{t-1})$, which is optimal.*

The last theorem basically states that the results of [3] continue to hold in this setting as well.

As in Section III, When $Z^y_t = Y^t$, we have the counterpart of Theorem II for the SI setting when the optimal reproduction functions are used:

Theorem VII. *For a Markov source and any given sequence of SI memory update functions $\{r^w_t\}$ and distortion measures $\{\rho_t\}$, when $Z^y_t = Y^t$ and the optimal reproduction function are used, there exists a sequence of deterministic encoders $Y_t = f_t(P_{X_t, Z^w_{t-1}|Y_{t-1}}(\cdot, \cdot|y^{t-1}), X_t)$ which is optimal.*

Note that unlike the result of Theorem VI, in the setting of Theorem VII, the encoder does not need to store B_{t-1} which is a function of X^{t-1}. Instead it stores the joint conditional probability measure of (X_t, Z^w_{t-1}), which is a function of Y^{t-1}. There is no contradiction between the theorems since the setting of Theorem VII is different both in the use of the optimal reproduction functions and in the SI independent sub–state of the decoder.

The proof of Theorem VI follows the lines of the proof of Theorem I after Lemmas I-IV are extended to the setting of this section. The changes to Lemmas II, IV are quite simple (roughly speaking, instead of x_t write (b_{t-1}, x_t) everywhere in the proof). The extension of the two– and three–stage lemmas (Lemmas I,III) is more involved and is given in the next two subsections. After these lemmas will be proven, the remainder of the proof is the same as in the previous section and therefore, will be omitted. Theorem VII is proved in Subsection V-C.

B. Theorem VI proof outline

We redefine B_t, b_t to be $B_t \stackrel{\triangle}{=} P_{Z_t^w|X^t,Y^t}(\cdot|X^t, Y^t)$ and $b_t \stackrel{\triangle}{=} P_{Z_t^W|X^t,Y^t}(\cdot|x^t, y^t)$. Since Theorem VI states that the encoders can be deterministic, the conditioning on Y^t in the definition of B_t is redundant since the sequence of encoder outputs Y^t is a deterministic function of the source symbols X^t. However, in the proof of Theorem VI, since we are allowing stochastic encoders a-priori, Y^t adds information to X^t and therefore, this conditioning is needed. We precede the proof of this theorem with a short discussion regarding its significance. Since b_{t-1} is a deterministic function of (x^{t-1}, y^{t-1}), one may argue that this theorem does not simplify the structure of the general encoder, which is, anyway, a function of x^t, y^{t-1}. However, it turns out that the encoder can update b_t recursively using only the data that is available to it at each stage (i.e., X_t, Y_t, B_{t-1}). To see why this is true, observe that

$$\begin{aligned} b_t(z) = P(z_t^w = z|x^t, y^t) &= P(r_t^w(y_t, w_t, z_{t-1}^w) = z|x^t, y^t) \\ &= \sum_{w_t, z_{t-1}^w : r_t^w(y_t, w_t, z_{t-1}^w) = z} P(w_t, z_{t-1}^w|x^t, y^t) \\ &= \sum_{w_t, z_{t-1}^w : r_t^w(y_t, w_t, z_{t-1}^w) = z} P(w_t|x^t, y^t) P(z_{t-1}^w|w_t, x^t, y^t) \\ &= \sum_{w_t, z_{t-1}^w : r_t^w(y_t, w_t, z_{t-1}^w) = z} P(w_t|x_t) P(z_{t-1}^w|x^{t-1}, y^{t-1}) \\ &\stackrel{\triangle}{=} h(b_{t-1}, x_t, y_t, z) \end{aligned} \tag{50}$$

Since this is true for any $z \in \mathcal{Z}^w$, we showed that b_t is a function of (b_{t-1}, x_t, y_t). Therefore, the encoder can recursively update b_t at the end of each encoding stage using its knowledge of (b_{t-1}, x_t) and its last output y_t.

1) Two-stage lemma: We start by analyzing a system with only two stages, where the first encoder is known.

Lemma V. *In a two–stage system ($T = 2$), there exists a deterministic second–stage encoder, $Y_2 = f_2(B_1, X_2, Z_1^y)$ which is optimal.*

Proof of Lemma V: Note that J_1 is unchanged by changing the second stage encoder. Denote the set of stochastic encoders which are functions of (X_1, X_2, Y_1) by $\{f^s_{X^2Y_1}\}$. The minimization of J_2 can be

written as

$$J_2^* = \inf_{\{f^s_{X^2Y_1}\}} \boldsymbol{E}\left\{\boldsymbol{E}\left[\rho_2(X_2, g_t(W_2, Y_2, Z_1^w, z_1^y)) + \lambda L_{Y_2|Z_1^y}(Z_1^y)|X_1, X_2, Y_1, Y_2, Z_1^y\right]\right\}$$
$$= \inf_{\{f^s_{X^2Y_1}\}} \boldsymbol{E}\left\{\lambda L_{Y_2|Z_1^y}(Z_1^y) + \boldsymbol{E}\left[\rho_2(X_2, g_t(W_2, Y_2, Z_1^w, Z_1^y))|X_1, X_2, Y_1, Y_2, Z_1^y\right]\right\}. \tag{51}$$

Focusing on the inner conditional expectation, we have

$$\begin{aligned}&\boldsymbol{E}\left[\rho_2(X_2, g_t(W_2, Y_2, Z_1^w, Z_1^y))|X_1, X_2, Y_1, Y_2, Z_1^y\right] = \\ &= \sum_{w_2, z_1^w} P(w_2|X_2)P(z_1^w|X_1, Y_1)\rho_2(X_2, g_t(w_2, Y_2, z_1^w, Z_1^y)) \\ &\stackrel{\triangle}{=} \hat{\rho}_2(B_1, X_2, Y_2, Z_1^y)\end{aligned} \tag{52}$$

where $B_1 \stackrel{\triangle}{=} P_{z_1^w|x_1,y_1}(\cdot|X_1, Y_1)$ is a probability measure on Z_1^w that represents the encoder's belief on the decoder's unknown state. Note that B_1 is a deterministic function of (X_1, Y_1) and the modified distortion measure (52) depends on (X_1, Y_1) only through B_1. Combining (51) and (52) we have

$$J_2^* = \inf_{\{f^s_{X^2Y_1}\}} \boldsymbol{E}\left\{\hat{\rho}_2(X_2, Y_2, B_1, Z_1^y) + \lambda L_{Y_2|Z_1^y}(Z_1^y)\right\} \tag{53}$$

where the expectation is with respect to $P(b_1, x_2, y_2, z_1^y)$. Consider the quadruple of RV's (B_1, X_2, Y_2, Z_1^y). We have

$$P(b_1, x_2, y_2, z_1^y) = P(b_1, x_2, z_1^y)P(y_2|b_1, x_2, z_1^y). \tag{54}$$

While $P(b_1, x_2, z_1^y)$, which depends on the first stage design and the source, remains fixed in the optimization in (53) (it can be thought of a state of the system, governed by the choice of the first stage design), $P(y_2|b_1, x_2, z_1^y)$ depends on the second stage encoder since

$$\begin{aligned}&P(y_2|b_1, x_2, z_1^y) = \\ &\frac{\sum_{x_1,y_1} P(x_1, x_2, y_1)P(y_2|x_1, x_2, y_1)P(b_1|x_1, y_1)P(z_1^y|y_1)}{P(b_1, x_2, z_1^y)}\end{aligned} \tag{55}$$

in the last expression, $P(b_1|x_1, y_1) = 1$ for all x_1, y_1 that yield the same specific conditional distribution, b_1, over Z_1^w and zero otherwise. $P(y_2|x_1, x_2, y_1)$ is governed by the second stage stochastic encoder, which maps (x_1, x_2, y_1) to a probability measure on $\mathcal{Y}$. Let us now look at the expectation in (53):

$$\begin{aligned}&\boldsymbol{E}\left\{\hat{\rho}_2(X_2, Y_2, B_1, Z_1^y) + L_{Y_2|Z_1^y}(Z_1^y)\right\} = \\ &\sum_{b_1, x_2, y_2, z_1^y} P(b_1, x_2, z_1^y)P(y_2|b_1, x_2, z_1^y)\Bigg\{\hat{\rho}_2(x_2, y_2, b_1, z_1^y) + \\ &\min_{l(\cdot)\in\mathcal{A}} \sum_{b_1', x_2', y_2'} P(y_2'|b_1', x_2', z_1^y)P(b_1', x_2'|z_1^y)l(y_2')\Bigg\}.\end{aligned} \tag{56}$$

As in the proof of Lemma I, from (56) we see that the optimization will be affected by the choice of the second stage encoder through $P(y_2'|b_1', x_2', z_1^y)$ for all (b_1', x_2', z_1^y). Denote the subset of stochastic second

stage encoders that are functions of (b_1, x_2, z_1^y) by $\{f^s_{B_1X_2Z_1^y}\}$. Since (b_1, z_1^y) are functions of (x_1, y_1), $\{f^s_{B_1X_2Z_1^y}\} \subset \{f^s_{X_1X_2Y_1}\}$. From (55) we see that every specific $f_2 \in f^s_{X_1X_2Y_1}$ is mapped to some specific $\hat{f}_2 \in f^s_{B_1X_2Z_1^y}$. Since the optimization is affected only by $P(y_2|b_1, x_2, z_1^y)$, if instead of using a specific f_2, we would use $\hat{f}_2$ that result from it, we would not change the joint probability of the quadruple (B_1, X_2, Y_2, Z_1^y) and thus the second stage cost will not be changed. Therefore we conclude that

$$\inf_{\{f^s_{X_1X_2Y_1}\}} J_2 = \inf_{\{f^s_{B_1X_2Z_1^y}\}} J_2. \tag{57}$$

To complete the proof, we need to show that it is enough to search in the finite subset of deterministic functions of (b_1, x_2, z_1^y), which we denote by $\{f^d_{B_1X_2Z_1^y}\}$. This is done by repeating the arguments we used below (15) in the end of the proof of Lemma I. ■

2) Three-stage Lemma:

Lemma VI. *In a three-stage system $(T = 3)$ with a Markov source, if the third–stage encoder is a deterministic function of (B_2, X_3, Z_2^y), then there exists a deterministic second stage encoder $Y_2 = f_2(B_1, X_2, Z_1^y)$ which is optimal.*

Proof of Lemma VI: We define, as in we did in Subsection V-B1, $\{f^s_{X_1X_2Y_1}\}$ to be the set of all possible stochastic second stage encoders. Let $\{f^s_{B_1X_2Z_1^y}\} \subset \{f^s_{X_1X_2Y_1}\}$ be the subset that contains all stochastic second stage encoders that are functions of (B_1, X_2, Z_1^y) and finally, let $\{f^d_{B_1X_2Z_1^y}\} \subset \{f^s_{B_1X_2Z_1^y}\}$ denote the set of deterministic encoders which are functions of (B_1, X_2, Z_1^y). Since the first stage is fixed, J_1 is unaffected. Our goal is to jointly optimize $(J_2 + J_3)$ with respect to the second stage encoder and show that

$$\inf_{\{f^s_{X_1X_2Y_1}\}} (J_2 + J_3) = \min_{\{f^d_{B_1X_2Z_1^y}\}} (J_2 + J_3)\,. \tag{58}$$

We start by focusing on the third stage cost.

$$\begin{aligned} J_3 &= \boldsymbol{E}\left\{\rho_3(X_3, g_3(Y_3, W_3, Z_2^w, Z_2^y)) + \lambda L_{Y_3|Z_2^y}(Z_2^y)\right\} \\ &= \boldsymbol{E}\left\{ \lambda L_{Y_3|Z_2^y}(Z_2^y) + \boldsymbol{E}\left\{ \rho_3(X_3, g_3(Y_3, W_3, Z_2^w, Z_2^y)) \middle| X^3, Y^2, Z_2^y \right\} \right\} \end{aligned} \tag{59}$$

Focusing on the inner expectation of (59), we have

$$\begin{aligned} &\boldsymbol{E}\left\{ \rho_3(X_3, g_3(Y_3, W_3, Z_2^w, Z_2^y)) \middle| X^3, Y^2, Z_2^y \right\} \\ &= \boldsymbol{E}\left\{ \rho_3(X_3, g_3(Y_3, W_3, Z_2^w, Z_2^y)) \middle| X^3, Y^3, Z_2^y \right\} \\ &= \sum_{w_3, z_2^w} P(w_3|X_3) P(z_2^w|X^2, Y^2) \rho_3(X_3, g_3(Y_3, w_3, z_2^w, Z_2^y)) \\ &\triangleq \hat{\rho}_3(B_2, X_3, Y_3, Z_2^y) \end{aligned} \tag{60}$$

where the first equality is true since Y_3 is a function of (B_2, X_3, Z_2^y) and B_2 is a deterministic function of (X^2, Y^2). Therefore,

$$J_3 = \sum_{b_2,x_3,y_3,z_2^y} P(b_2, x_3, z_2^y) P(y_3|b_2, x_3, z_2^y) \times \left[\hat{\rho}_3(b_2, x_3, y_3, z_2^y) + \lambda L_{Y_3|Z_2^y}(z_2^y)\right]. \tag{61}$$

In the last expression, $P(y_3|b_2, x_3, z_2^y)$ will not be affected by the optimization of the second stage encoder since, under the assumptions of Lemma VI, the third encoder is a fixed deterministic function of (b_2, x_3, z_2^y) (i.e., $P(y_3|b_2, x_3, z_2^y) = \mathbb{1}\{f_3(b_2, x_3, z_2^y) = y_3\}$). Thus, the second stage encoder affects the last expression only through $P(b_2, x_3, z_2^y)$ since

$$P(b_2, x_3, z_2^y) = \sum_{b_1,x_2,y_2} P(b_1, x_2, z_1^y) P(y_2|b_1, x_2, z_1^y) \times P(x_3|x_2) \mathbb{1}\{h(b_1, x_2, y_2) = b_2\} \mathbb{1}\{r_2^y(z_1^y, y_2) = z_2^y\}, \tag{62}$$

where $h(b_1, x_2, y_2)$ was defined in (50) and we used the fact that the source is Markov. As in subsection V-B1, $P(b_1, x_2, z_1^y)$ is the result of the first stage design and the source. Note that $\mathbb{1}\{h(b_1, x_2, y_2) = b_2\}$, $\mathbb{1}\{r_2^y(z_1^y, y_2) = z_2^y\}$ are not affected by the choice of the second stage encoder since they represent known deterministic functions of (b_1, x_2, y_2) and (z_1^y, y_2) respectively. Focusing on the third stage average codeword length for $Z_2^y = z_2^y$, we have

$$\begin{aligned} L_{Y_3|Z_2^y}(z_2^y) &= \min_{l(\cdot)\in\mathcal{A}} \sum_{b_2,x_3,y_3} P(b_2, x_3, y_3|z_2^y) l(y_3) \\ &= \min_{l(\cdot)\in\mathcal{A}} \sum_{b_2,x_3,y_3} P(b_2, x_3|z_2^y) P(y_3|b_2, x_3, z_2^y) l(y_3). \end{aligned} \tag{63}$$

Again, the second–stage encoder affects only $P(b_2, x_3, z_2^y)$ (and thus $P(b_2, x_3|z_2^y)$). In (55),(56) we showed that the second–stage encoder affect the second stage cost only through $P(y_2|b_1, x_2, z_1^y)$. In (61),(62),(63) we showed that the third stage cost depends on the second stage cost only through $P(y_2|b_1, x_2, z_1^y)$. Therefore, we conclude that the optimization of the second stage encoder affects $(J_2 + J_3)$ only through $P(y_2|b_1, x_2, z_1^y)$. Repeating the arguments we used in the proof of Lemma V, if we use $\hat{f}_2 \in f^s_{B_1 X_2 Z_1^y}$ that result from a specific $f_2 \in f^s_{X_1 X_2 Y_1}$ (through (55)) instead of using that specific f_2, we would not change the joint probability of $(B^2, X^3, Y^3, Z_1^y, Z_2^y)$ and therefore will not change the value of $(J_2 + J_3)$. Therefore we have

$$\min_{\{f^s_{X_1 X_2 Y_1}\}} (J_2 + J_3) = \min_{\{f^s_{B_1 X_2 Z_1^y}\}} (J_2 + J_3). \tag{64}$$

From here, the same arguments we used after (23) in the proof of Lemma III will complete the proof.

C. Infinite memory decoder - proof of Theorem VII

As in the case without SI, when $Z_t^y = Y^t$, we can use the tools of MDP to derive a structure theorem. We will need to redefine the state to $s_t = P_{X_t, Z_{t-1}^w|Y^{t-1}}(\cdot, \cdot|y^{t-1})$. The action is defined in the same manner as in the case without SI, i.e., $a_t : \mathcal{X} \to \mathcal{Y}$. The optimal reproduction function, $\hat{x}_t^* = g^*(w_t, y^t, z_{t-1}^w)$ is

the Bayes response to $P_{X_t|W_t,Y^t,Z_{t-1}^w}(\cdot|w_t,y^t,z_{t-1}^w)$:

$$\hat{x}_t^* = \arg\max_{\hat{x}} \sum_{x_t} P(x_t|w_t,y^t,z_{t-1}^w)\rho_t(x_t,\hat{x}). \tag{65}$$

As in Subsection III-E, in order to use the tools of MDP, we need to show that we can write the cost function as a function of (s_t,a_t) and that the state is conditionally Markov, given a_t.

The optimal reproduction function is a function of $P_{X_t|W_t,Y^t,Z_{t-1}^w}(\cdot|w_t,y^t,z_{t-1}^w)$. Note that

$$\begin{aligned} P(x_t|w_t,y^t,z_{t-1}^w) &= \frac{P(w_t,x_t,y_t,z_{t-1}^w|y^{t-1})}{\sum_{x_t'} P(w_t,x_t',y_t,z_{t-1}^w|y^{t-1})} \\ &= \frac{P(x_t,z_{t-1}^w|y^{t-1})P(w_t|x_t)\mathbb{1}\left\{a_t(x_t)=y_t\right\}}{\sum_{x_t'} P(x_t',z_{t-1}^w|y^{t-1})P(w_t|x_t')\mathbb{1}\left\{a_t(x_t')=y_t\right\}} \\ &\triangleq f(s_t,a_t,w_t,x_t,y_t,z_{t-1}^w) \end{aligned} \tag{66}$$

Therefore, the optimal reproduction function is a function of $(s_t,a_t,w_t,y_t,z_{t-1}^w)$, which we denote by $g_t^*(s_t,a_t,w_t,y_t,z_{t-1}^w)$. We now move on to show that the cost function can be written as a function of the state and action. As in Subsection III-E, we deal with the distortion and codeword length elements of the cost separately. Treating the expected codeword length first we have:

$$\begin{aligned} L_{Y_t|Y^{t-1}}(y^{t-1}) &= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t} P(y_t|y^{t-1})l(y_t) \\ &= \min_{l(\cdot)\in\mathcal{A}} \sum_{x_t,y_t,z_{t-1}^w} P(x_t,y_t,z_{t-1}^w|y^{t-1})l(y_t) \\ &= \min_{l(\cdot)\in\mathcal{A}} \sum_{y_t,x_t,z_{t-1}^w} P(x_t,z_{t-1}^w|y^{t-1})\mathbb{1}\left\{a_t(x_t)=y_t\right\}l(y_t) \\ &\triangleq \alpha_t(s_t,a_t), \end{aligned} \tag{67}$$

When using the optimal reproduction function, the average distortion is given by:

$$\begin{aligned} &\boldsymbol{E}\left[\rho(X_t,g_t^*(s_t,a_t,Y_t,W_t,Z_{t-1}^w))\middle|Y^{t-1}=y^{t-1}\right] \\ &= \sum_{w_t,x_t,y_t,z_{t-1}^w} P(x_t,y_t,z_{t-1}^w|y^{t-1})\rho(x_t,g_t^*(s_t,a_t,w_t,y_t,z_{t-1}^w)) \\ &= \sum_{w_t,x_t,y_t,z_{t-1}^w} P(x_t,z_{t-1}^w|y^{t-1})P(w_t|x_t)\mathbb{1}\left\{a_t(x_t)=y_t\right\}\rho(x_t,g_t^*(s_t,a_t,w_t,y_t,z_{t-1}^w)) \\ &\triangleq \beta_t(s_t,a_t). \end{aligned} \tag{68}$$

Denoting $\beta_t(s_t,a_t)+\lambda\alpha_t(s_t,a_t)=\gamma_t(s_t,a_t)$, our optimality criterion can be written as $\frac{1}{T}\sum_{t=1}^T \boldsymbol{E}\gamma_t(s_t,a_t)$.

We move on to show that the state process is Markov conditioned on the action, namely, $P(s_{t+1}|s^t,a^t) = P(s_{t+1}|s_t,a_t)$. We start by noting that $s_{t+1} = P_{X_{t+1},Z_t^w|Y^t}(\cdot,\cdot|y^t)$ is a function of (s_t,a_t,y_t). For every

(x_{t+1}, z_t^w) we have

$$\begin{aligned}
P(x_{t+1}, z_t^w|y^t) &= \frac{\sum_{w_t,x_t,z_{t-1}^w} P(x_t, x_{t+1}, w_t, y_t, z_{t-1}^w, z_t^w|y^{t-1})}{\sum_{w_t,x_t,x_{t+1},z_{t-1}^w} P(x_t, x_{t+1}, w_t, y_t, z_{t-1}^w, z_t^w|y^{t-1})} \\
&= \frac{\sum_{w_t,x_t,z_{t-1}^w} P(x_t, z_{t-1}^w|y^{t-1})P(x_{t+1}|x_t)P(w_t|x_t)\mathbb{1}\left\{a_t(x_t) = y_t\right\}\mathbb{1}\left\{z_t^w = r_t^w(w_t, y_t, z_{t-1}^w)\right\}}{\sum_{w_t,x_t,x_{t+1},z_{t-1}^w} P(x_t, z_{t-1}^w|y^{t-1})P(x_{t+1}|x_t)P(w_t|x_t)\mathbb{1}\left\{a_t(x_t) = y_t\right\}\mathbb{1}\left\{z_t^w = r_t^w(w_t, y_t, z_{t-1}^w)\right\}} \\
&= f(s_t, a_t, x_{t+1}, y_t, z_t^w) \qquad (69)
\end{aligned}$$

and therefore, $s_{t+1} = h(s_t, a_t, y_t)$ for a function, h, that uses (69) for every pair (x_{t+1}, z_t^w). Now,

$$\begin{aligned}
P(s_{t+1} = \nu|s^t, a^t) &= \sum_{y_t,x_t,z_{t-1}^w} \mathbb{1}\left\{h(a_t, s_t, y_t) = \nu\right\}\mathbb{1}\left\{a_t(x_t) = y_t\right\}P(x_t, z_{t-1}^w|y^{t-1}) \\
&= P(s_{t+1} = \nu|s_t, a_t). \qquad (70)
\end{aligned}$$

We showed that our system can be represented as a MDP. By invoking Theorem A.1, we know that the optimal action at each stage, a_t is a deterministic function of the state. Namely, the mapping from x_t to y_t can be chosen deterministically as a function of $P_{X_t,Z_{t-1}^w|Y^{t-1}}(\cdot,\cdot|y^{t-1})$. Therefore, Y_t is a deterministic function of $(X_t, P_{X_t,Z_{t-1}^w|Y^{t-1}}(\cdot,\cdot|y^{t-1}))$, which concludes the proof of Theorem VII. By (69), the encoder does not need to store y^{t-1}, but rather a probability measure (a vector in $\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Z}^w|}$).

VI. Conclusion

This work extended the setting of [1] to include both variable rate coding and SI. It was shown that structure theorems, in the spirit of [1] and [3], continue to hold in this setting as well. These theorems are further refined when the decoder has infinite memory. We were able to show that the cost function is concave in the choices of past encoders (Lemmas II, IV) and therefore, the optimal encoders are deterministic. It was also shown that in order to simplify the overall system optimization, one can use sliding–window next–state functions and the excess loss incurred by this suboptimal choice vanishes as the window size increases (Theorems III, IV). However, in the finite horizon setting we investigated, the window size is always upper bounded by the time horizon.

Extensions to this work would include investigating the infinite horizon setting. While Theorem III carries over verbatim to the infinite horizon setting, it is not necessarily true for the other theorems, which were proved using dynamic programming. Another extension would be to investigate the constrained setting (briefly mentioned in Theorem V). In this case, relying on results from constrained MDPs, we do not expect the optimal encoders to be deterministic (see [15]).

Appendix

A. Proof of Lemma II

We start by focusing on the average codeword length element of the cost function and show that $L_{Y_t}(Z_{t-1})$ is concave in $\{f^s_{X_t,Z_{t-1}}\}$. For $0 \leq \alpha \leq 1$ and $f_1, f_2 \in \{f^s_{X_t,Z_{t-1}}\}$, let

$$f_\alpha = \alpha f_1 + (1-\alpha) f_2.$$

This means that for any (x_t, z_{t-1}) we have

$$P_\alpha(y_t|x_t, z_{t-1}) = \alpha P_1(y_t|x_t, z_{t-1}) + (1-\alpha)P_2(y_t|x_t, z_{t-1}) \tag{A.1}$$

Let $L_{Y_t}^{f_\alpha}(Z_{t-1})$, $L_{Y_t}^{f_1}(Z_{t-1})$, $L_{Y_t}^{f_2}(Z_{t-1})$ denote the length function calculated with f_α, f_1, f_2 respectively. We have

$$\begin{aligned}
&L_{Y_t}^{f_\alpha}(z_{t-1}^y) = \min_{l(\cdot)\in\mathcal{A}} \left\{ \sum_{y_t} P_\alpha(y_t|z_{t-1}) l(y_t) \right\} \\
&= \min_{l(\cdot)\in\mathcal{A}} \left\{ \sum_{y_t, x_t} [\alpha P_1(y_t|x_t, z_{t-1}) \right. \\
&\left. + (1-\alpha)P_2(y_t|x_t, z_{t-1})]\, P(x_t|z_{t-1}) l(y_t) \right\} \\
&\geq \alpha \min_{l(\cdot)\in\mathcal{A}} \left\{ \sum_{y_t, x_t} P_1(y_t|x_t, z_{t-1}) P(x_t|z_{t-1}) l(y_t) \right\} + (1-\alpha) \min_{l(\cdot)\in\mathcal{A}} \left\{ \sum_{y_t, x_t} P_2(y_t|x_t, z_{t-1}) P(x_t|z_{t-1}) l(y_t) \right\} \\
&= \alpha L_{Y_t}^{f_1}(z_{t-1}) + (1-\alpha) L_{Y_t}^{f_2}(z_{t-1})
\end{aligned} \tag{A.2}$$

where we used the fact that the sum of minima is smaller than the minimum of a sum. Since the distortion part of the cost is linear in $P(y_t|x_t, z_{t-1})$ (through the expectation), we have that the overall stage t cost function is concave in $\{f_{X_t, Z_{t-1}}^s\}$.

B. Proof of Lemma IV

Fix any third stage encoder which is a deterministic function of (X_3, Z_2). We showed in (19) that the second stage encoder affects J_3 only through $P(x_3, z_2)$ (and thus also through $P(z_2)$ and $P(x_3|z_2)$). Let $f_1, f_2 \in \{f_{X_2 Z_1}^s\}$ be two second stage stochastic encoders which are functions of (X_2, Z_1). Let

$$\begin{aligned}
P_\gamma(x_3, z_2) &= \sum_{x_2, y_2, z_1} P(x_2, z_1) \left[\gamma P_1(y_2|x_2, z_1) + (1-\gamma)P_2(y_2|x_2, z_1)\right] P(z_2|z_1, y_2) P(x_3|x_2), \\
&= \gamma P_1(x_3, z_2) + (1-\gamma) P_2(x_3, z_2),
\end{aligned} \tag{A.3}$$

where $P_1(x_3, z_2), P_2(x_3, z_2)$ are calculated with $P_1(y_2|x_2, z_1), P_2(y_2|x_2, z_1)\}$ that result from f_1, f_2 respectively. Similarly, for $i = 1, 2, \gamma$, define $P_i(z_2)$, and $P_i(x_3|z_2)$ as the marginal and conditional distribution, respectively, resulting from the probability measures in (A.3). We now show that $P_\gamma(x_3|z_2)$ can be written as a convex combination of $P_1(x_3|z_2), P_2(x_3|z_2)$.

$$\begin{aligned}
P_\gamma(x_3|z_2) &= \frac{P_\gamma(x_3, z_2)}{P_\gamma(z_2)} \\
&= \frac{\gamma P_1(x_3, z_2) + (1-\gamma) P_2(x_3, z_2)}{\sum_{x_3'} \gamma P_1(x_3', z_2) + (1-\gamma) P_2(x_3', z_2)} \\
&= \frac{\gamma P_1(x_3, z_2)}{\sum_{x_3'} \gamma P_1(x_3', z_2) + (1-\gamma) P_2(x_3', z_2)} + \frac{(1-\gamma) P_2(x_3, z_2)}{\sum_{x_3'} \gamma P_1(x_3', z_2) + (1-\gamma) P_2(x_3', z_2)} \\
&= \alpha \frac{P_1(x_3, z_2)}{\sum_{x_3'} P_1(x_3', z_2)} + \beta \frac{P_2(x_3, z_2)}{\sum_{x_3'} P_2(x_3', z_2)}
\end{aligned} \tag{A.4}$$

with,

$$\alpha = \frac{\gamma \sum_{x_3'} P_1(x_3', z_2)}{\sum_{x_3'} \gamma P_1(x_3', z_2) + (1-\gamma) P_2(x_3', z_2)} = \frac{\gamma P_1(z_2)}{P_\gamma(z_2)}$$
$$\beta = \frac{(1-\gamma) \sum_{x_3'} P_2(x_3', z_2)}{\sum_{x_3'} \gamma P_1(x_3', z_2) + (1-\gamma) P_2(x_3', z_2)} = \frac{(1-\gamma) P_2(z_2)}{P_\gamma(z_2)}. \tag{A.5}$$

Note that $0 \leq \alpha, \beta \leq 1$ and $\alpha + \beta = 1$. We showed that

$$P_\gamma(x_3|z_2) = \alpha P_1(x_3|z_2) + (1-\alpha) P_2(x_3|z_2) \tag{A.6}$$

We are now ready to prove the lemma. For any given third stage encoder, let $J_3(P_i)$, $i = 1, 2, \gamma$, denote the third stage cost as a function of the joint probability of (X_3, Z_2), where the dependence on the second stage encoder was shown in (20),(A.3). In order to prove the lemma, we need to show that:

$$J_3(P_\gamma) \geq \gamma J_3(P_1) + (1-\gamma) J_3(P_2). \tag{A.7}$$

We now focus on the codeword length element of the cost function. Let $L_3(P_i)$, $i = 1, 2, \gamma$, denote the third stage average codeword length as a function of the joint probability of (X_3, Z_2)

$$\begin{aligned}
L_3(P_\gamma) &= \sum_{z_2} P_\gamma(z_2) \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3} P_\gamma(y_3|z_2) l(y_3) \\
&= \sum_{z_2} P_\gamma(z_2) \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_\gamma(y_3, x_3|z_2) l(y_3) \\
&= \sum_{z_2} P_\gamma(z_2) \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_\gamma(x_3|z_2) P(y_3|x_3, z_2) l(y_3) \\
&= \sum_{z_2} P_\gamma(z_2) \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} \left[\alpha P_1(x_3|z_2) + (1-\alpha) P_2(x_3|z_2)\right] P(y_3|x_3, z_2) l(y_3) && \text{(A.8)} \\
&\geq \sum_{z_2} P_\gamma(z_2) \alpha \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_1(x_3|z_2) P(y_3|x_3, z_2) l(y_3) + \\
&\qquad \sum_{z_2} P_\gamma(z_2)(1-\alpha) \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_2(x_3|z_2) P(y_3|x_3, z_2) l(y_3) && \text{(A.9)} \\
&= \sum_{z_2} P_\gamma(z_2) \frac{\gamma P_1(z_2)}{P_\gamma(z_2)} \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_1(x_3|z_2) P(y_3|x_3, z_2) l(y_3) + \\
&\qquad \sum_{z_2} P_\gamma(z_2) \frac{(1-\gamma) P_2(z_2)}{P_\gamma(z_2)} \min_{l(\cdot)\in\mathcal{A}} \sum_{y_3, x_3} P_2(x_3|z_2) P(y_3|x_3, z_2) l(y_3) && \text{(A.10)} \\
&= \gamma d_3(P_1) + (1-\gamma) d_3(P_2), && \text{(A.11)}
\end{aligned}$$

where in (A.8) we used (A.6), (A.9) is true since the minimum of a sum is greater than the sum of minima and finally, in (A.10) we substituted α given in (A.5). Thus, we showed that the codeword length element

of the cost function is concave in the choice of the second stage encoder. We have

$$\begin{aligned} J_3(P_\gamma) &= \sum_{x_3,y_3,z_2} P(y_3|x_3,z_2)P_\gamma(x_3,z_2)\rho(x_3,g(y_3,z_2)) + \lambda L_3(P_\gamma) \\ &= \sum_{x_3,y_3,z_2} \gamma P(y_3|x_3,z_2)P_1(x_3,z_2)\rho(x_3,g(y_3,z_2)) + \\ &\qquad (1-\gamma)P(y_3|x_3,z_2)P_2(x_3,z_2)\rho(x_3,g(y_3,z_2)) + \lambda L_3(P_\gamma) \\ &\geq \gamma J_3(P_1) + (1-\gamma)J_3(P_2) \end{aligned} \tag{A.12}$$

where in the last step we used (A.11) and the lemma is proven. ■

C. Markov decision processes - short overview

In a Markov decision process, a decision maker is influencing the behavior of a Markov probabilistic system through his actions, as the system evolves in time. Formally, a discrete time, finite horizon Markov decision process is defined by $\{T, S, A, \{P_t(\cdot|s,a)\}, \{\rho_t(s,a)\}\}$, where,

- T is the time horizon, $t = 1, 2, \ldots, T$.
- $\mathcal{S}$ is the state space.
- $\mathcal{A}$ is the action space.
- $P_t(\cdot|s,a)$ is the transition probability to the systems's next state, given the previous system state and action. The transition probabilities obey $P_t(\cdot|s^t,a^t) = P_t(\cdot|s_t,a_t)$, namely, the next state, s_{t+1}, distribution depends on the history only through (s_t,a_t).
- $P_0(\cdot)$ is the probability measure over the initial state.
- $\rho_t(s,a)$ is the cost incurred when at stage t and state s, action a is taken.

In our case, the goal of the decision maker is to minimize the expected average cost $\boldsymbol{E}\frac{1}{T}\sum_{t=1}^{T}\rho_t(S_t,A_t)$. The history of the process at stage t is $h_t = (s_1,a_1,s_2,a_2,\ldots,s_{t-1},a_{t-1},s_t)$, i.e., all previous actions taken by the decision maker and the system states, up to stage t. Note that $h_t = \{h_{t-1},a_{t-1},s_t\}$.

A decision rule, d_t, prescribes the procedure for action selection in a given state at stage t. Decision rules can range from deterministic functions of the current state to randomized functions that depend on the whole history of states and actions, up to stage t. A decision rule that is a deterministic function of the current state will be called a Markovian deterministic (MD) decision rule. A policy specifies the decision rules to be used at all stages, i.e., a policy π is a sequence of decision rules $d_1,\ldots,d_T$. We say that a policy is MD if all its decision rules are MD.

In Sections III-E,V-C, the state space is finite, however, it grows as the system evolves, i.e., at each stage the stage space is $\mathcal{S}_t$. We set $\mathcal{S} = \cup_{t=1}^{T}\mathcal{S}_t$. The action space is the set of deterministic functions $f : \mathcal{X} \to \mathcal{Y}$, which is finite.

We will use the following theorem which is the key to the results of section III-E.

Theorem A.1. *([16, Proposition 4.4.3]): There exist an MD policy which is optimal.*

We outline the proof here for completeness.

Proof of Theorem A.1 (outline): Define for policy π, $u_t^\pi(h_t) = \boldsymbol{E}\left\{\sum_{i=t}^{T}\rho_i(s_i,a_i)|h_t\right\}$, where the actions

a_i are prescribed by the policy π. Note that

$$u_t^\pi(h_t) = \rho_t(s_t, a_t) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a_t) u_{t+1}^\pi(h_t, j, a_t) \tag{A.13}$$

Let $u_t^*(h_t) = \inf_\pi u_t^\pi(h_t)$. We start by showing the $u_t^*(h_t)$ depends on the history only through s_t. We will use backwards induction. Note that $u_T^*(h_T) = \min_{a \in \mathcal{A}} \rho(s_T, a)$, so the claim is valid for the last stage. Now assume that the claim is valid for $n = t+1, t+2, \ldots, T$. We have

$$\begin{aligned} u_t^*(h_t) &= \min_{a \in \mathcal{A}} \left\{ \rho_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a) u_{t+1}^*(h_t, j, a) \right\} \\ &= \min_{a \in \mathcal{A}} \left\{ \rho_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a) u_{t+1}^*(j) \right\} \end{aligned} \tag{A.14}$$

where the last equation is due to the induction hypothesis. Since the term in brackets depends on the history only through s_t, the induction step is proven. Now, define the decision rule at each stage for every $s_t \in \mathcal{S}$ as the minimizer of (A.14). By construction, this decision rule is MD and the policy constructed from these decision rules is optimal.

D. Properties of the length function

Let $\mathcal{W}, \mathcal{Y}, \mathcal{Z}$ be finite alphabets.

1) Conditioning reduces the length: We have

$$\begin{aligned} L_{Y|Z} &= \sum_{z \in \mathcal{Z}} P(z) \min_{l(\cdot) \in \mathcal{A}} \sum_{y \in \mathcal{Y}} P(y|z) l(y) \\ &= \sum_{z \in \mathcal{Z}} P(z) \min_{l(\cdot) \in \mathcal{A}} \sum_{y \in \mathcal{Y}} \sum_{w \in \mathcal{W}} P(y|w, z) P(w|z) l(y) \\ &\geq \sum_{z \in \mathcal{Z}} \sum_{w \in \mathcal{W}} P(z) P(w|z) \min_{l(\cdot) \in \mathcal{A}} \sum_{y \in \mathcal{Y}} P(y|w, z) l(y) \\ &= L_{Y|W,Z} \end{aligned} \tag{A.15}$$

where the inequality is true since the minimum of a sum is greater than the sum of minima.

2) Proving that $L_{Y|W,Z} \geq L_{Y,Z|W} - \log|\mathcal{Z}|$: The intuition behind this is simple: given W, the average optimal code length for the pair (Y, Z) can not be larger than coding Z separately and concatenating a codeword that describes Y and is decodable when Z is known. The optimal scheme for coding the pair can not be worse, otherwise, this scheme can be used. To see this mathematically, for each $y \in \mathcal{Y}, z \in \mathcal{Z}$, let $l^*(y), l^*(z)$ be the length functions optimized for the distributions $P(y|w, z)$ and $P(z|w)$ respectively.

Using the fact that $L_{Z|W} < \log|\mathcal{Z}|$ we have:

$$\begin{aligned} L_{Y|Z,W} + \log|\mathcal{Z}| &\geq \sum_{w,z} P(w,z) \sum_{y} P(y|w,z) l^*(y) + \sum_{w} P(w) \sum_{z} P(z|w) l^*(z) \\ &= \sum_{w} P(w) \sum_{z} P(z|w) \left[\left(\sum_{y} P(y|w,z) l^*(y) \right) + l^*(z) \right] \\ &= \sum_{w} P(w) \sum_{z} P(z|w) \sum_{y} P(y|w,z) [l^*(y) + l^*(z)] \\ &\geq \sum_{w} P(w) \min_{l \in \tilde{\mathcal{A}}} \sum_{z,y} P(y,z|w) l(y,z) \\ &= L_{Y,Z|W} \end{aligned} \tag{A.16}$$

where $\tilde{\mathcal{A}}$ is defined as in Section II with $\mathcal{Y} \times \mathcal{Z}$ replacing $\mathcal{Y}$.

REFERENCES

[1] H. S. Witsenhausen, "The structure of real time source coders," *Bell Systems Technical Journal*, vol. 58, no. 6, pp. 1338–1451, July 1979.

[2] J. C. Walrand and P. Varaiya, "Optimal causal coding–decoding problems," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 814–820, November 1983.

[3] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4017–4035, September 2006.

[4] A. D. Wyner and J. Ziv, "The rate–distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, January 1976.

[5] V. S. Borkar, S. K. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of Markov sources," *SIAM Journal on Control and Optimization*, vol. 40, no. 1, pp. 135–148, 2001. [Online]. Available: http://link.aip.org/link/?SJC/40/135/1

[6] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 167–186, March 1982.

[7] S. K. Gorantla and T. P. Coleman, "Information-theoretic viewpoints on optimal causal coding-decoding problems," *CoRR*, vol. abs/1102.0250, 2011.

[8] H. Asnani and T. Weissman, "On real time coding with limited lookahead," *CoRR*, vol. abs/1105.5755, 2011.

[9] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5317–5338, November 2009.

[10] D. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 701–713, September 1982.

[11] T. Weissman and N. Merhav, "On causal source codes with side information," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 4003–4013, November 2005.

[12] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 609–625, March 2003.

[13] N. Merhav and J. Ziv, "On the Wyner-Ziv problem for individual sequences," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 867–873, March 2006.

[14] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1329–1339, September 1996.

[15] E. Altman, *Constrained Markov Decision Processes*, ser. Stochastic Modeling Series. Chapman and Hall/CRC, 1999.

[16] M. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, ser. Wiley series in probability and statistics. Wiley-Interscience, 1994.