Structure Theorem for Real–Time Variable–Rate Lossy Source Encoders and Memory–Limited Decoders with 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*—We extend Witsenhausen's structure theorem for real–time source coding, so as to accommodate both variable–rate coding and causal side information at the decoder.**

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 has 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 sought 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).

The scenario where the encoder is also a finite state machine was considered by Gaarder and Slepian in [2]. In [3], Teneketzis considered the case where there is a noisy channel between the encoder and decoder. In this case, unlike this work and [1],[2], the encoder cannot track the decoder's state. 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. When the time horizon and alphabets are finite, there is a finite number of possible encoding, decoding and memory update rules. Theoretically, a brute force search will 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 [4], proposed a search frame that is linear in the communication length and doubly exponential in the alphabet size.

Real time codes are a subclass of causal codes, as defined by Neuhoff and Gilbert [5]. In [5], 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 [5], 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 [6], extended [5] 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 [7].

This work extends [1] in two directions: The first direction is 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. The second direction of extension is in allowing the decoder access to causal side information. Our main result is that a structure theorem, similar to that of Witsenhausen [1], continues to hold in this setting as well.

As mentioned in the previous paragraph, in contrast to [1]–[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

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

In [1], the proof of the structure theorem relied on two lemmas. The proof of the extension of those lemmas to our case is 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).

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 state and prove the main result of this paper. Finally, we conclude this work in Section IV.

II. Preliminaries

We begin with notation conventions. Capital letters represent scalar random variables (RV), 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. We consider a k-th order 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 other alphabets in the sequel, is finite. The SI sequence, $W_1, W_2, ..., W_T$, $W_t \in \mathcal{W}$, is generated by a discrete memoryless channel (DMC), fed by $X_1, X_2, \ldots, X_T$:

$$P(W_1, \ldots, W_T | X_1, \ldots X_T) = \prod_{t=1}^{T} P(W_t | X_t). \quad (1)$$

The probability mass function of X_1, $P(X_1)$ and the transition probabilities, denoted by $P(X_2|X_1)$,$P(X_3|X^2)$,...,$P(X_t|X_{t-k}^{t-1})$, $t = k+1, k+2, \ldots, T$ are known. A variable length encoder is a sequence of functions $\{f_t\}_{t=1}^{T}$. At stage t, an encoder $f_t : \mathcal{X}^t \to \mathcal{Y}_t$ calculates $Y_t = f_t(X^t)$ and noiselessly transmits an entropy coded codeword of Y_t. $\mathcal{Y}_t \subseteq \mathcal{Y}$ is the alphabet used at stage t. Unlike the fixed rate regime in [1],[3], where $\log_2 |\mathcal{Y}_t|$ (rounded up) was the rate of the code of stage t, here the size of the codebook will be one of the outcomes of the optimization.

The encoder structure is not confined *a–priori*, and it is not even limited to be deterministic: at each time instant t, Y_t may be given by an arbitrary (possibly stochastic) function of X^t. 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's state can be divided into two parts. $Z_t^w \in \mathcal{Z}^w$ is updated by:

$$\begin{aligned} Z_1^w &= r_1^w(W_1, Y_1) \\ Z_t^w &= r_t^w(W_t, Y_t, Z_{t-1}^w), \quad t = 2, 3, \ldots, T \end{aligned} \quad (2)$$

and $Z_t^y \in \mathcal{Z}^y$ is updated by

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

Both parts depend on the received compressed data Y_t. However, since Z_t^y does not depend on the SI and the transmission is noiseless, it can be tracked by the encoder. 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

$$\begin{aligned} \hat{X}_1 &= g_1(W_1, Y_1) \\ \hat{X}_t &= g_t(W_t, Y_t, Z_{t-1}^w, Z_{t-1}^y), \quad t = 2, 3, \ldots, T \end{aligned} \quad (4)$$

Since at the beginning of stage t, Z_{t-1}^y 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}^y. We define $\mathcal{A}_t$ to be the set of all uniquely decodable codes for Y_t, i.e all possible length functions $l : \mathcal{Y}_t \to \boldsymbol{Z}_+$ that satisfy Kraft's inequality:

$$\mathcal{A}_t = \left\{ l(\cdot) : \sum_{y \in \mathcal{Y}_t} 2^{-l(y)} \leq 1 \right\} \quad (5)$$

The average codeword length of stage t will be given by:

$$L(Y_t | Z_{t-1}^y) \stackrel{\triangle}{=} \sum_{z \in \mathcal{Z}^y} P(z) \min_{l(\cdot) \in \mathcal{A}_t} \left\{ \sum_{y \in \mathcal{Y}_t} P(y|z) l(y) \right\}. \quad (6)$$

$L(Y_t|z_{t-1}^y)$ will denote the optimal average codeword length at stage t, for a specific decoder state, z_{t-1}^y. $L(Y_t|z_{t-1}^y)$ is obtained by designing a Huffman code for the probability distribution $P(\cdot|z_{t-1}^y)$. The model is depicted in Figure 1.

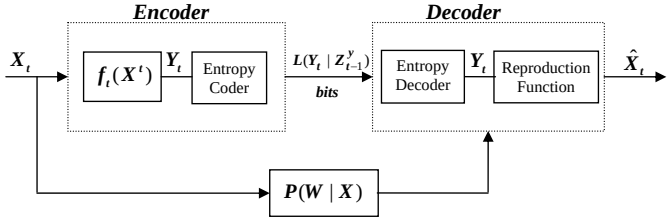

Fig. 1: System model

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 $L(Y_t|Z_{t-1}^y)$, i.e.

$$D(t) \stackrel{\triangle}{=} \boldsymbol{E}\left\{ \rho_t(X_t, \hat{X}_t) \right\} + \lambda L(Y_t | Z_{t-1}^y). \quad (7)$$

where λ is a fixed parameter that controls the tradeoff between rate and distortion. Our goal is to minimize the cumulative cost

$$D \stackrel{\triangle}{=} \sum_{t=1}^{T} D(t). \quad (8)$$

A specific choice of the memory update functions $\{r_t^y\}$,$\{r_t^w\}$ reproduction functions $\{g_t\}$ and encoders $\{f_t\}$ along with their resulting cost D is called a *design*. We say that design A with cumulative cost D^A outperforms design B, with cumulative cost D^B, if $D^A \leq D^B$. We say that the encoder has *memory order* k if Y_t is produced using the decoder's state that is tracked by the encoder and the last k source symbols, i.e $Y_t = f_t(X_{t-k+1}^t, Z_{t-1}^y)$.

In the following section, we state and prove a structure theorem for the system described above.

III. Main Result

The main result of this paper is the following theorem:

Theorem 1. *Under the conditions described in Section II, for any given design of the system, there exists a design with the same decoders, whose encoders are deterministic and have memory order k, that outperforms the original design.*

Remark: We will prove this structure theorem using lemmas regarding the structure of a two and three stage system ($T = 2, 3$) and then use backwards induction as we show in the following subsections. This is the original method used by Witsenhausen [1]. The main difference is in the proofs of the lemmas, which call for more caution. In [1], these lemmas could be proved by optimizing the encoder over the distortion function independently for each possible pair (x_i, z_{i-1}^y). For each such pair, Y_t was chosen as the one that minimizes the distortion $\rho_t(X_t, g_t(Y_t, Z_{t-1}^y))$. Here, even without SI, we cant take this approach since such Y_t (that minimizes the distortion) might entail a large $L(Y_t|Z_{t-1})$. Moreover, $Y_t|Z_{t-1}^y$, is well defined only after all stage t encoding rules are known. Any change for a specific (x_i, z_{i-1}^y) will affect the PMF of $Y_t|Z_{t-1}^y$ (and the code length) and therefore cannot be done independently of the other pairs.

We start by stating and proving the lemmas for the two and three stage systems in the next two subsections. Using these lemmas, Theorem 1 is proved in Section III-C

A. Two stage lemma

Lemma 1. *For a two stage system ($T = 2$), any source, SI sequence and any given design, a design different only in f_2, with f_2 deterministic and having memory order of $k = 1$, will outperform the original design.*

Proof: Note that $f_1, g_1, g_2, r_1^y, r_1^w$ are fixed and $D(1)$ is unchanged by changing f_2. We need to show that f_2 that minimizes $D(2)$ can be a deterministic function of X_2, Z_1^y. For every joint probability measure over the 6–tuple $(X_1, X_2, W_2, Y_2, Z_1^w, Z_1^y)$, $D(2)$ is well defined and our goal is to minimize:

$$D(2) = \boldsymbol{E}\left\{\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y))\right\} + \lambda L(Y_2|Z_1^y) \quad (9)$$

with respect to the second stage encoder. Note that the expectation is over $(X_1, X_2, W_2, Y_2, Z_1^w, Z_1^y)$ in the first expression only. The second expression (which is a constant) contains the expectation in its definition. Let us look at the 6–tuple of RV's $(X_1, X_2, W_2, Y_2, Z_1^w, Z_1^y)$. From the structure of the system we know that

$$\begin{aligned} P(X_1, X_2, W_2, Y_2, Z_1^w, Z_1^y) = P(X_1)P(X_2|X_1)P(Y_1|X_1)\times \\ P(Z_1^y|X_1)P(Z_1^w|X_1)P(Y_2|X_1, X_2)P(W_2|X_2) \end{aligned} \quad (10)$$

Everything but the second stage encoder $P(Y_2|X_1, X_2)$ is fixed. Our objective is to find

$$\min_{P(Y_2|X_1,X_2)} \boldsymbol{E}\left\{\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y))\right\} + \lambda L(Y_2|Z_1^y) \quad (11)$$

Note that the minimization affects $L(Y_2|Z_1^y)$ since its inner minimization depends on $P(Y_2|Z_1^y)$ which in turn depends on $P(Y_2|X_1, X_2)$. We rewrite the expression in (11):

$$\begin{aligned} &\min_{P(Y_2|X_1,X_2)} \boldsymbol{E}\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) + \lambda L(Y_2|Z_1^y) \\ &= \min_{P(Y_2|X_1,X_2)} \sum_{x_1,x_2,w_2,y_2,z_1^w,z_1^y} P(x_1, x_2, w_2, y_2, z_1^w, z_1^y)\times \\ &\{\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y)) + \lambda L(Y_2|z_1^y\}. \end{aligned} \quad (12)$$

For any $P(Y_2|X_1, X_2)$ and given z_1^y, $L(Y_2|z_1^y)$ does not depend on x_1 and obviously, $\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y))$ is not a function of x_1. Therefore the term in brackets does not depend on x_1. Define

$$\begin{aligned} &d_2'(x_2, y_2, z_1^w, z_1^y) \stackrel{\triangle}{=} \\ &\left[\sum_{w_2} P(w_2|x_2)\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y))\right] + \lambda L(Y_2|z_1^y) \end{aligned} \quad (13)$$

Also note that z_1^y, z_1^w does not play a role in the optimization. Therefore,

$$\begin{aligned} &\min_{P(Y_2|X_1,X_2)} \boldsymbol{E}\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) + \lambda L(Y_2|Z_1^y) \\ &= \sum_{z_1^w,z_1^y} \min_{P(Y_2|X_1,X_2)} \sum_{x_2,y_2} P(x_2, y_2, z_1^w, z_1^y) d_2'(x_2, y_2, z_1^w, z_1^y) \end{aligned} \quad (14)$$

Now, given that the first stage encoder and decoder are known, $P(x_2, z_1^w, z_1^y)$ is well defined. Also, since the encoder does not have access to the side information sequence, we have, for any second stage encoder:

$$P(X_2, Y_2, Z_1^w, Z_1^y) = P(Y_2|X_2, Z_1^y)P(X_2, Z_1^w, Z_1^y) \quad (15)$$

For a given z_1^y and any second stage encoder, $L(Y_2|z_1^y)$ is given by

$$L(Y_2|z_1^y) = \min_{l(y_2)\in\mathcal{A}_t} \sum_{y_2} \sum_{x_2,z_1^w} \frac{P(x_2, z_1^w, z_1^y)}{P(z_1^y)} P(y_2|x_2, z_1^y) l(y_2) \quad (16)$$

Substituting (15)–(16) back into (14) we have

$$\begin{aligned} &\min_{P(Y_2|X_1,X_2)} \boldsymbol{E}\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) + \lambda L(Y_2|Z_1^y) \\ &= \sum_{z_1^w,z_1^y} P(z_1^w, z_1^y) \min_{P(Y_2|X_1,X_2)} \Bigg\{ \Bigg[\sum_{x_2} P(x_2, z_1^w, z_1^y)\times \\ &\sum_{y_2} P(y_2|x_2, z_1^y)\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y)) \Bigg] \\ &+ \lambda \min_{l\in\mathcal{A}_t} \sum_{y_2', x_2'} P(y_2'|x_2', z_1^y) P(x_2, z_1^w|z_1^y) l(y_2') \Bigg\} \end{aligned} \quad (17)$$

Note that, for any $P(Y_2|X_2, Z_1^y)$, there can be more than one possible second stage encoder which will yield the same r.h.s in (15) (there is at least one such encoder: $P(y_2|x_1, x_2) = P(y_2|x_2, z_1^y)$ for all x_1 with $P(z_1^y, x_1) > 0$ for all x_2, y_2, z_1^y).

However, from (17), it is clear that all second stage encoders that result in the same marginal $P(Y_2|X_2, Z_1^y)$ will yield the same second stage average cost $D(2)$. Now, every possible $P(Y_2|X_2, Z_1^y)$ has at least one possible $P(Y_2|X_1, X_2)$ that leads to it. Since every $P(Y_2|X_1, X_2)$ results in some $P(Y_2|X_2, Z_1^y)$, minimizing over $P(Y_2|X_2, Z_1^y)$ will yield the same (optimal) cost function for the second stage. Therefore, we have

$$\begin{aligned}&\min_{P(Y_2|X_1,X_2)} \boldsymbol{E}\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) + \lambda L(Y_2|Z_1^y) \\ &= \min_{P(Y_2|X_2,Z_1^y)} \boldsymbol{E}\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) + \lambda L(Y_2|Z_1^y)\end{aligned} \tag{18}$$

and we showed that its enough to search for second stage encoding rules that depend only on X_2, Z_1^y and still receive the optimal second stage cost. Now, since $L(Y_2|Z_1^y)$ is a concave functional of $P(Y_2|X_2, Z_1^y)$ (see Appendix), the second stage cost function is concave in $P(Y_2|X_2, Z_1^y)$. This means that the minimizing $P(Y_2|X_2, Z_1^y)$ is one of the extreme points of the simplex consisting of all possible $P(Y_2|X_2, Z_1^y)$. Since the extreme points of the simplex represent a deterministic choice of y_2 given x_2, z_1^y, this means that an optimal second stage encoder is a deterministic function of (X_2, Z_1^y). ■

B. Three stage lemma

Lemma 2. *For a three stage system ($T = 3$), first order Markov source and SI as described in Section II, any design in which f_3 has memory order of $k = 1$ will be outperformed by a design different only in f_2, where f_2 in the new design is deterministic and with memory order of $k = 1$.*

Proof: The first stage encoder and decoder, and therefore the first stage cost, are fixed. We optimize the second stage encoder given a decoder and a memory update function. The minimum cost of the last two stages is given by the minimization of:

$$\begin{aligned}D(2) + D(3) = \boldsymbol{E}\{&\rho_2(X_2, g_2(W_2, Y_2, Z_1^w, Z_1^y)) \\ &+\lambda L(Y_2|Z_1^y) + \rho_3(X_3, g_3(W_3, Y_3, Z_2^w, Z_2^y)) \\ &+\lambda L(Y_3|Z_2^y)\}\end{aligned} \tag{19}$$

with respect to $P(Y_2|X_1, X_2)$.
Since we know the second stage decoder and the third stage encoder/decoder pair

$$P(X_3, W_3, Y_3, Z_2^w, Z_2^y|X_2, W_2, Y_2, Z_1^w, Z_1^y) \tag{20}$$

is well defined regardless of the second stage encoder. Therefore we can calculate the conditional expectation of the third stage cost given $X_2, W_2, Y_2, Z_1^w, Z_1^y$.

$$\begin{aligned}&d_3(x_2, w_2, y_2, z_1^w, z_1^y) \stackrel{\triangle}{=} \\ &\sum_{x_3,w_3,z_2^w,z_2^y} P(x_3, w_3, z_2^w, z_2^y|x_2, w_2, y_2, z_1^w, z_1^y)\times \\ &\{\rho_3(x_3, g_3(w_3, y_3, z_2^w, z_2^y)) + \lambda L(Y_3|z_2^y)\}\end{aligned} \tag{21}$$

Here, when we wrote $P(X_3|X_2)$, we used the fact that the source is Markov. This, alongside the requirement that the third stage encoder has memory order of $k = 1$, ensures that $d_3(x_2, w_2, y_2, z_1^w, z_1^y)$ is not a function of x_1. We rewrite (19):

$$\begin{aligned}D(2) + D(3) = &\min_{P(Y_2|X_1,X_2)} \sum_{x_1,x_2,y_2,z_1^w,z_1^y} P(x_1, x_2, y_2, z_1^w, z_1^y) \\ &\times \{\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y)) + \lambda L(Y_2|z_1^y) \\ &+d_3(x_2, w_2, y_2, z_1^w, z_1^y)\}\end{aligned} \tag{22}$$

Let

$$\begin{aligned}\rho_2'(x_2, y_2, z_1^w, z_1^y) = &\rho_2(x_2, g_2(w_2, y_2, z_1^w, z_1^y)) \\ &+ d_3(x_2, w_2, y_2, z_1^w, z_1^y).\end{aligned} \tag{23}$$

Substituting this into (22) and using the fact that, as in the two stage lemma case, given x_2, z_1^w, z_1^y, the term in brackets in (22) is not a function of x_1 we have

$$\begin{aligned}D(2) + D(3) = &\min_{P(Y_2|X_1,X_2)} \sum_{x_2,y_2,z_1^w,z_1^y} P(x_2, y_2, z_1^w, z_1^y)\times \\ &\{\rho_2'(x_2, y_2, z_1^w, z_1^y) + \lambda L(Y_2|z_1^y)\}\end{aligned} \tag{24}$$

Note the similarity of the last expression to (12). Although here we have a different distortion measure, both here and in (12) the distortion measures are functions on x_2, y_2, z_1^w, z_1^y. Therefore, the same steps we used to prove the two stage lemma after (12), can be used here and the three stage lemma will be proven. ■

Remark: From looking back at the definition of $d_3(x_2, w_2, y_2, z_1^w, z_1^y)$ in (21), it is clear that the three stage lemma will continue to hold if the third stage encoder would have memory order of $k = 2$, since the resulting conditional expectation in (21) still would not be a function of x_1. However, this is not needed in the proof of Theorem 1.

C. Proof of Theorem 1

With the two and three stage lemmas, we can prove Theorem 1 by using the method of [1], used for fixed rate encoding. We will prove Theorem 1 for a first order Markov source. The extension to a k-th order Markov source is straight forward by repacking k source symbols each time into a new super symbol in a "sliding window" manner: $\hat{X}_t = (X_t, X_{t+1}, \ldots, X_{t+k-1})$. The SI is produced from these super symbols. The resulting source (with super symbols) is a first order Markov source for which the proof we give here can be applied. An encoder with memory order $k = 1$ with the super symbols has memory order k with the original source.

We need one last auxiliary lemma before we can prove Theorem 1.

Lemma 3. *For any source statistics, SI and any design, one can replace the last encoder, f_T, with a deterministic encoder having memory order $k = 1$ and the new design will outperform the original design.*

Proof: Let $X_1' = X^{T-1}, W_1' = W^{T-1}$. We now have a two stage system with source symbols (X_1', X_T) and SI (W_1', W_T). Now, by the the two stage lemma, the second stage encoder in the new two stage system. has memory order $k = 1$ for any first stage encoding and decoding rules. ■

The main theorem is proven by backward induction. First apply the last lemma to any design to conclude that an optimal f_T has memory order $k = 1$. Now assume that the last m encoders (i.e $f_{T-m+1}, ..., f_T$) have memory order $k = 1$. We will show that the encoder at time $T-m$ also has this structure and continue backwards until $t = 2$. The first encoder, trivially, has memory order $k = 1$. To prove the induction step, define

$$\begin{aligned}
\hat{X}_1 &= (X_1, X_2, ..., X_{T-m-1}) \\
\hat{W}_1 &= (W_1, W_2, ..., W_{T-m-1}) \\
\hat{Y}_1 &= (Y_1, Y_2, ..., Y_{T-m-1}) \\
\hat{Z}_1^w &= \hat{r}_1^w(\hat{Y}_1, \hat{W}_1) \\
\hat{Z}_1^y &= \hat{r}_1^y(\hat{Y}_1) \\
\hat{X}_2 &= X_{T-m} \\
\hat{W}_2 &= W_{T-m} \\
\hat{Y}_2 &= Y_{T-m} \\
\hat{Z}_2^w &= r_{T-m}^w(W_{T-m}, Y_{T-m}, Z_{T-m-1}^w) \\
\hat{Z}_2^y &= r_{T-m}^y(Y_{T-m}, Z_{T-m-1}^w) \\
\hat{X}_3 &= (X_{T-m+1}, X_{T-m+2}, ..., X_T) \\
\hat{W}_3 &= (W_{T-m+1}, W_{T-m+2}, ..., W_T) \\
\hat{Y}_3 &= (Y_{T-m+1}, Y_{T-m+2}, ..., Y_T)
\end{aligned} \tag{25}$$

where

$$\hat{r}_1^y(\hat{Y}_1) = r_{T-m-1}^y(Y_{T-m-1}, r_{T-m-2}^y(\dots r_2^y(Y_2, r_1^y Y_1))\dots) \tag{26}$$

and $\hat{r}_1^w(\hat{Y}_1, \hat{W}_1)$ is defined likewise. In words, $\hat{Z}_1^w, \hat{Z}_1^y$ are the states of the decoder after $T - m - 1$ stages. Using this new notation, the encoder that produces $\hat{Y}_3$ has memory order of $k = 1$, since its a function of $\hat{X}_3, \hat{Z}_2^y$ (since, by assumption, the last m encoders in the original notation are of memory order of $k = 1$). 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, $\hat{Y}_2 = Y_{T-m} = f_{T-m}(\hat{X}_2, \hat{Z}_1^y) = f_{T-m}(X_{T-m}, Z_{T-m-1}^y)$. Thus, the induction step is proved.

IV. Conclusions

We showed that the results of [1] can be extended to both include variable rate coding and SI at the decoder. Following the same steps we used, this result can be further extended to the case the decoder has SI with some lookahead (i.e at time t it sees W_t^{t+l} where l is the lookahead). The results of [1] and some of the results in [3] are obtained in the special case of $\lambda = 0$, i.e the instantaneous rate is not taken into account in the cost function if we set $\mathcal{Y}_t$ to be the encoder output alphabet size at stage t of [1].

Appendix

This appendix show that $L(Y_2|Z_1^y)$ is concave in $P(Y_2|X_2, Z_1^y)$. Let

$$P(Y_2|X_2, Z_1^y) = \alpha P_1(Y_2|X_2, Z_1^y) + (1-\alpha)P_2(Y_2|X_2, Z_1^y)$$

for some $0 \le \alpha \le 1$. Focusing on the inner minimization of the definition of $L(Y_2|Z_1^y)$ (see (6)) we have

$$\begin{aligned}
&\min_{l(\cdot)\in\mathcal{A}_2} \left\{ \sum_{y_2\in\mathcal{Y}} P(y_2|z_1^y) l(y_2) \right\} \\
&= \min_{l(\cdot)\in\mathcal{A}_2} \left\{ \sum_{y_2\in\mathcal{Y}} \sum_{x_2\in\mathcal{X}} [\alpha P_1(y_2|x_2, z_1^y) \right. \\
&\quad \left. + (1-\alpha)P_2(y_2|x_2, z_1^y)]\, P(x_2|z_1^y) l(y_2) \right\} \\
&\ge \alpha \min_{l(\cdot)\in\mathcal{A}_2} \left\{ \sum_{y_2\in\mathcal{Y}} \sum_{x_2\in\mathcal{X}} P_1(y_2|x_2, z_1^y) P(x_2|z_1^y) l(y_2) \right\} \\
&+ (1-\alpha) \min_{l(\cdot)\in\mathcal{A}_2} \left\{ \sum_{y_2\in\mathcal{Y}} \sum_{x_2\in\mathcal{X}} P_2(y_2|x_2, z_1^y) P(x_2|z_1^y) l(y_2) \right\}
\end{aligned} \tag{27}$$

Let $L_P(Y_2|Z_1^y)$, $L_{P_1}(Y_2|Z_1^y)$, $L_{P_2}(Y_2|Z_1^y)$ denote the length function calculated with $P(Y_2|X_2, Z_1^y)$, $P_1(Y_2|X_2, Z_1^y)$, $P_2(Y_2|X_2, Z_1^y)$ respectively. Substituting this into the definition of $L(Y_2|Z_1^y)$ we have:

$$L_P(Y_2|Z_1^y) \ge \alpha L_{P_1}(Y_2|Z_1^y) + (1-\alpha)L_{P_2}(Y_2|Z_1^y) \tag{28}$$

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] 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.

[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. 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.

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

[6] 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.

[7] 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.