On the Rate Distortion Theory for Causal Video Coding

En-hui Yang, Lin Zheng, Da-ke He, and Zhen Zhang

***Abstract*—Causal video coding is considered from an information theoretic point of view, where video source frames $X_1, X_2, \cdots, X_N$ are encoded in a frame by frame manner, the encoder for each frame X_k, $k = 1, \cdots, N$, can use all previous frames and all previous encoded frames while the corresponding decoder can use only all previous encoded frames, and each frame X_k itself is modeled as a source $X_k = \{X_k(i)\}_{i=1}^{\infty}$. A novel computation approach is proposed to analytically characterize, numerically compute, and compare the minimum total rate of causal video coding $R_c(D_1, \cdots, D_N)$ required to achieve a given distortion (quality) level $D_1, \cdots, D_N \geq 0$. Specifically, we first show that for jointly stationary ergodic sources $X_1, X_2, \cdots, X_N$, $R_c(D_1, \cdots, D_N)$ is equal to the infimum of the nth order total rate distortion function $R_{c,n}(D_1, \cdots, D_N)$ over all n, where $R_{c,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. We then present an iterative algorithm for computing $R_{c,n}(D_1, \cdots, D_N)$ and demonstrate the convergence of the algorithm to the global minimum. The global convergence of the algorithm further enables us to establish a single-letter characterization of $R_c(D_1, \cdots, D_N)$ in a novel way when the N sources are an independent and identically distributed (IID) vector source. With the help of the algorithm, we also demonstrate a surprising result (dubbed the more and less coding theorem)—under some conditions on source frames and distortion, the more frames need to be encoded and transmitted, the less amount of data has to be actually sent. Predictive video coding, where each encoder and its corresponding decoder can use only all previous encoded frames, is also investigated.**

I. Introduction

Consider a causal video coding model shown in Figure 1, where X_k, $k = 1, 2, \cdots, N$, represents a video frame, S_k and $\hat{X}_k$ represent respectively its encoded frame and reconstructed frame, all frames X_k, $k = 1, 2, \cdots, N$, are encoded in a frame by frame manner, and the encoder for X_k can use all previous frames X_j, $j = 1, 2, \cdots, k-1$, and all previous encoded frames S_j, $j = 1, 2, \cdots, k-1$, while the corresponding decoder can use only all previous encoded frames. The model is causal because the encoder for X_k is not allowed to access to future frames in the encoding order. In the special case where the encoder for each X_k is further restricted to enlist help only from all previous encoded frames S_j, $j = 1, 2, \cdots, k-1$, causal video coding reduces to predictive video coding.

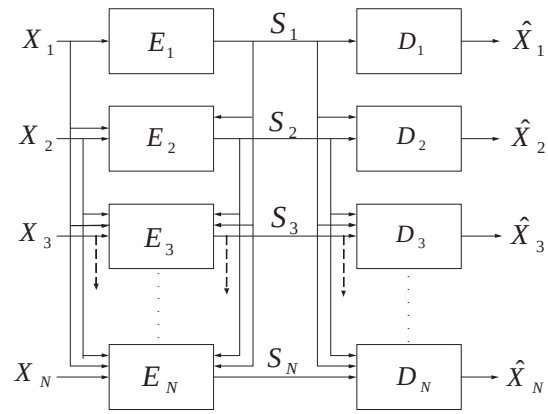

Fig. 1. Causal video coding model

All MPEG-series and H-series video coding standards [1], [6] proposed so far fall into the above causal video coding model (strictly speaking, into the predictive video coding model); the differences among these different video coding standards lie in how information available to the encoder of each frame X_k is used to generate S_k. When $N = 2$, the causal coding model is the same as the sequential coding model of correlated source proposed in [3]. When $N = 3$, the causal coding model is also called the C-C model in [4]. However, when $N > 2$, which is a typical case in MPEG-series and H-series video coding, the causal coding model considered here is quite different from sequential coding.

It is expected that a future video coding standard will continue to fall into the causal video coding model shown in Figure 1. To provide some design guidance for a future video coding standard, in this paper, we aim at investigating from an information theoretic point of view how each frame in the causal model should be encoded so that collectively the total rate is minimized subject to a given distortion (quality) level $D_1, ...D_N \geq 0$.

We model each frame X_k itself as a source $X_k = \{X_k(i)\}_{i=1}^{\infty}$ taking values in an alphabet $\mathcal{X}_k$. Together, the N frames then form a vector source $(X_1, X_2, \cdots, X_N) = \{X_1(i), X_2(i), \cdots, X_N(i)\}_{i=1}^{\infty}$ taking values in the product alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$. The sources $X_1, X_2, \cdots, X_N$ are said to be (first-order) Markov if for any $1 < j \leq N$, X_j is the output of a memoryless channel in response to input X_{j-1}; in this case, we say $X_1 \to X_2 \to \cdots \to X_N$

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grants RGPIN203035-02 and RGPIN203035-06, and by the Canada Research Chairs Program.

En-hui Yang and Lin Zheng are with the Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Emails: ehyang@uwaterloo.ca, l9zheng@uwaterloo.ca

Da-ke He is with SlipStream Data Inc., a subsidiary of Research in Motion, Waterloo, Ontario N2L 3W8, Canada. Email: dhe@rim.com

Zhen Zhang is with the Department of Electrical Engineering-Systems, University of Southern California, Los Angeles. Email: zhzhang@usc.edu

forms a Markov chain. Let $\hat{X}_k = \{\hat{X}_k(i)\}_{i=1}^{\infty}$ denote the reconstruction of $X_k = \{X_k(i)\}_{i=1}^{\infty}$ drawn from an alphabet $\hat{\mathcal{X}}_k$. The distortion between $\hat{X}_k$ and X_k is measured by a single-letter distortion measure $d_k : \mathcal{X}_k \times \hat{\mathcal{X}}_k \to [0, \infty)$. For convenience, we write $\{X_k(i)\}_{i=1}^{n}$ simply as $X_k(1;n)$ for any k and $n \geq 1$. For any N dimensional vector $V = (V_1, V_2, \cdots, V_N)$, denote $(V_1, \cdots, V_{t-1})$ by V_t^-, and $(V_{t+1}, \cdots, V_N)$ by V_t^+. As such, by $X_k^-(1;n)$ we shall mean that $X_k^-(1;n) = (X_1(1;n), \cdots, X_{k-1}(1;n))$. A similar convention will apply to reconstruction sequences and other vectors.

Formally, We define an order-n causal video code C_n for $X_1, \cdots, X_N$ by using N encoder and decoder pairs:

1) For X_1, an encoder of order n is defined by a function f_1 from $\mathcal{X}_1^n$ to $\{0,1\}^*$, the set of all binary sequences of finite length, satisfying the property that the range of f_1 is a prefix set, and a decoder of order n is defined by a function
$$g_1 : \{0,1\}^* \to \hat{\mathcal{X}}_1^n.$$
The encoded and reconstructed sequences of $X_1(1;n)$ are given respectively by $S_1 = f_1(X_1(1;n))$ and $\hat{X}_1(1;n) = g_1(S_1)$.
2) For X_k, $k = 2, \cdots, N$, an encoder of order n is defined by a function
$$f_k : \quad \mathcal{X}_1^n \times \cdots \times \mathcal{X}_k^n \times \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{k-1 \text{ times}} \to \{0,1\}^*$$
satisfying the property that the range of f_k given any $k-1$ binary sequences is a prefix set, and a decoder of order n is defined by a function
$$g_k : \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{k \text{ times}} \to \hat{\mathcal{X}}_k^n.$$
The encoded and reconstructed sequences of $X_k(1;n)$ are given respectively by $S_k = f_k(X_k^-(1;n), X_k(1;n), S_k^-)$ and $\hat{X}_k(1;n) = g_k(S_k^-, S_k)$.
3) For $k = 1, \cdots, N$, the distortion between $X_k(1;n)$ and $\hat{X}_k(1;n)$ is defined by
$$D_{xk} \triangleq \frac{1}{n} E\left[\sum_{i=1}^{n} d_k(X_k(i), \hat{X}_k(i))\right].$$
4) The rates of N encoders are defined respectively by
$$R_{xk} \triangleq \frac{1}{n} E|S_k|, \ k = 1, \cdots, N$$
where $|S_k|$ denotes the length of the binary sequence S_k.

In a similar manner, one can formally define an order n predictive video code; the only difference lies in the definition of the encoders for X_k, $k = 2, \cdots, N$, where the encoder for X_k, $k \geq 2$, is defined instead by a function
$$f_k : \quad \mathcal{X}_k^n \times \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{k-1 \text{ times}} \to \{0,1\}^*$$
satisfying the property that the range of f_k given any $k-1$ binary sequences is a prefix set, and where the encoded sequence of $X_k(1;n)$ is given by $S_k = f_k(X_k(1;n), S_k^-)$.

Definition 1: A set of rates $(R_1, \cdots, R_N)$ is said to be achievable at distortion level $D_1, \cdots, D_N \geq 0$ by causal coding (predictive coding, respectively) if $\forall \epsilon > 0$, there exists an order n causal (predictive, respectively) code $\{(f_k, g_k)\}_{k=1}^{N}$ for all sufficiently large n such that
$$R_{xk} \leq R_k + \epsilon \text{ and } D_{xk} \leq D_k + \epsilon$$
for $k = 1, \cdots, N$.

Let $\mathcal{R}_c^*(D_1, \cdots, D_N)$ and $\mathcal{R}_p^*(D_1, \cdots, D_N)$ denote the set of all rates $(R_1, \cdots, R_N)$ at distortion level $(D_1, \cdots, D_N)$ achievable by causal coding and predictive coding, respectively. As in the usual video compression applications, we are interested in the minimum total rates $R_c(D_1, \cdots, D_N)$ and $R_p(D_1, \cdots, D_N)$ required to achieve the distortion level $(D_1, \cdots, D_N)$, which are defined respectively by
$$R_c(D_1, \cdots, D_N) \triangleq \min\{R_1 + R_2 + \cdots + R_N : (R_1, \cdots, R_N) \in \mathcal{R}_c^*(D_1, \cdots, D_N)\},$$
and
$$R_p(D_1, \cdots, D_N) \triangleq \min\{R_1 + R_2 + \cdots + R_N : (R_1, \cdots, R_N) \in \mathcal{R}_p^*(D_1, \cdots, D_N)\}.$$

One of our purposes in this paper is to numerically compute, analytically characterize, and compare $R_c(D_1, \cdots, D_N)$ so that deep insights can be gained regarding how each frame should be encoded in order to have a minimum total rate.

Our approach is computation oriented. Starting with a jointly stationary ergodic vector source $(X_1, X_2, \cdots, X_N)$, we first show in Section II that $R_c(D_1, \cdots, D_N)$ is equal to the infimum of the nth order total rate distortion function $R_{c,n}(D_1, \cdots, D_N)$ over all n, where $R_{c,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. Then we develop an iterative algorithm in Section III to calculate $R_{c,n}(D_1, \cdots, D_N)$, and further show that this algorithm converges to an optimal solution that achieves $R_{c,n}(D_1, \cdots, D_N)$. The global convergence of the algorithm enables us to establish a single-letter characterization of $R_c(D_1, \cdots, D_N)$ in Section IV in the case where the vector source $(X_1, X_2, \cdots, X_N)$ is independent and identically distributed (IID), by comparing $R_{c,n}(D_1, \cdots, D_N)$ with $R_{c,1}(D_1, \cdots, D_N)$ through a novel application of the algorithm. With the help of the algorithm, we further demonstrate in Section V a surprising result dubbed the more and less coding theorem—under some conditions on source frames and distortion, the more frames need to be

encoded and transmitted, the less amount of data has to be actually sent.

For predictive coding, the corresponding problem turns out to be even harder. In Section VI, we show that under the condition that $X_1, X_2, \cdots, X_N$ form a (first-order) Markov chain, predictive coding achieves the same performance as does causal coding. In this case, therefore, all the information theoretic results and our proposed algorithm for causal coding can be applied to predictive coding. When $X_1, X_2, \cdots, X_N$ do not form a (first-order) Markov chain, however, the problem remains open.

II. Minimum Total Rate and Achievable Rate Region: Ergodic Causal Case

Suppose now that $(X_1, X_2, \cdots, X_N)$ is jointly stationary and ergodic. Define $\mathcal{R}_{c,n}(D_1, \cdots, D_N)$ to be the region consisting of rates $(R_1, \cdots, R_N)$ for which there exist auxiliary random variables U_k, $k = 1, 2, \cdots, N-1$, and $\hat{X}_N(1;n)$ such that

$$\begin{aligned} &R_1 \geq \frac{1}{n} I(X_1(1;n); U_1) \\ &R_k \geq \frac{1}{n} I(X_1(1;n), \cdots, X_k(1;n); U_k | U_k^-) \\ &k = 2, 3, \cdots, N-1 \\ &R_N \geq \frac{1}{n} I(X_N(1;n); \hat{X}_N(1;n) | U_N^-) \end{aligned} \tag{1}$$

and the following requirements are satisfied:

(1) $\hat{X}_1(1;n) = g_1(U_1)$ for some deterministic function g_1,

(2) $\hat{X}_k(1;n) = g_k(U_k^-, U_k)$ for some deterministic function g_k, $k = 2, \cdots, N-1$,

(3) for any $1 \leq k \leq N$, $\frac{1}{n} E[d_k(X_k(1;n), \hat{X}_k(1;n))] \leq D_k$, and

(4) the Markov chain conditions $U_k \to (X_k(1;n), X_k^-(1;n), U_k^-) \to X_k^+(1;n)$, $k = 1, \cdots, N-1$, and $X_N^-(1;n) \to (X_N(1;n), U_N^-) \to \hat{X}_N(1;n)$ are met.

Let $\mathcal{R}_c'(D_1, \cdots, D_N) = \bigcup_{n=1}^{\infty} \mathcal{R}_{c,n}(D_1, \cdots, D_N)$. Denote its convex hull closure by $co(\mathcal{R}_c'(D_1, \cdots, D_N))$. Then we have the following results, the proofs of which along with other omitted proofs and other practical video coding settings can be found in the full paper [2].

Theorem 1: For jointly stationary and ergodic sources $X_1, \cdots, X_N$ and any distortion level $D_1, \cdots, D_N \geq 0$, $\mathcal{R}_c^*(D_1, \cdots, D_N) = co(\mathcal{R}_c'(D_1, \cdots, D_N))$.

Theorem 2: For jointly stationary and ergodic sources $X_1, \cdots, X_N$ and any distortion level $D_1, \cdots, D_N \geq 0$,

$$R_c(D_1, \cdots, D_N) = \inf\{R_{c,n}(D_1, \cdots, D_N) : n \geq 1\}$$

where

$$\begin{aligned} &R_{c,n}(D_1, \cdots, D_N) \\ &\triangleq \frac{1}{n} \min_{\{\hat{X}_k(1;n)\}_{k=1}^N} [I(X_1(1;n); \hat{X}_1(1;n)) + \\ &\sum_{t=2}^{N-1} I(X_1(1;n), \cdots, X_t(1;n); \hat{X}_t(1;n) | \hat{X}_t^-(1;n)) + \\ &I(X_N(1;n); \hat{X}_N(1;n) | \hat{X}_N^-(1;n))] \end{aligned} \tag{2}$$

where the minimum is taken over all auxiliary random vectors $\hat{X}_k(1;n)$ satisfying the following two conditions for all $j = 1, \cdots, N$, and $k = 1, \cdots, N-1$: 1) the Markov chains $\hat{X}_k(1;n) \to (X_k(1;n), X_k^-(1;n), \hat{X}_k^-(1;n)) \to X_k^+(1;n)$ and $X_N^-(1;n) \to (X_N(1;n), \hat{X}_N^-(1;n)) \to \hat{X}_N(1;n)$ hold; and 2) $\frac{1}{n} E d_j(X_j(1;n), \hat{X}_j(1;n)) \leq D_j$.

For general stationary ergodic sources $X_1, \cdots, X_N$, Theorem 2 is probably the best result one could hope for in terms of characterizing analytically $R_c(D_1, \cdots, D_N)$. However, its impact on practical video coding will be limited if the optimization problem involved can not be solved by an effective algorithm. To a large extent, this is also true even if $R_c(D_1, \cdots, D_N)$ admits a single letter characterization and for many other multi-user problems. In the following section, we will develop an iterative algorithm to compute $R_{c,n}(D_1, \cdots, D_N)$ defined in (2) for any stationary ergodic sources, and establish its convergence to the global minimum.

III. An Iterative Algorithm

In this section, an iterative algorithm is proposed to calculate $R_{c,n}(D_1, \cdots, D_N)$ defined in (2), which serves three purposes in this paper: first, it allows us to do numerical calculations; second, the global convergence of this algorithm provides a completely different approach to establish a single-letter characterization of $R_c(D_1, .., D_N)$ when the N sources are IID; and third, it allows us to do comparisons and gain deep insights into $R_c(D_1, .., D_N)$.

Without loss of generality, we consider the case of $N = 3$ and denote three sources by $\{X(i)\}_{i=1}^n, \{Y(i)\}_{i=1}^n$, and $\{Z(i)\}_{i=1}^n$, which in turn will be written as X^n, Y^n, and Z^n respectively to simplify our notation for describing the iterative algorithm.

Let $p_{X^nY^n}$ and $p_{X^nY^nZ^n}$ denote joint distributions of random vectors (X^n, Y^n) and (X^n, Y^n, Z^n), respectively; and let $p(X^n)$ denote the marginal distribution of X^n. If there is no ambiguity, subscripts in distributions will be omitted. For example, we may write $p(x)$ instead of $p_X(x)$. In order to find the random variables $\hat{X}^n, \hat{Y}^n$ and $\hat{Z}^n$ that achieve $R_{c,n}(D_1, D_2, D_3)$, we try to find transition probability and probability functions $p_{\hat{X}^n|X^n}$, $p_{\hat{Y}^n|\hat{X}^nY^nX^n}$, $p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}$ and

$q_{\hat{X}^n\hat{Y}^n\hat{Z}^n}$ that minimize

$$\begin{aligned}
&F_{s,n}(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}, q_{\hat{X}^n\hat{Y}^n\hat{Z}^n}) \\
&\triangleq \sum_{x^n,y^n,z^n,\hat{x}^n,\hat{y}^n,\hat{z}^n} p(x^n,y^n,z^n)p(\hat{x}^n|x^n) \times \\
&p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n) \times \\
&\log[\frac{p(\hat{x}^n|x^n)p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n)}{q(\hat{x}^n\hat{y}^n\hat{z}^n)}] + \\
&\alpha \sum_{x^n,\hat{x}^n} p(x^n)p(\hat{x}^n|x^n)d_1(x^n,\hat{x}^n) + \\
&\beta \sum_{x^n,y^n,\hat{x}^n,\hat{y}^n} p(x^n,y^n)p(\hat{x}^n|x^n)p(\hat{y}^n|\hat{x}^ny^nx^n)d_2(y^n,\hat{y}^n) + \\
&\nu \sum_{x^n,y^n,z^n,\hat{x}^n,\hat{y}^n,\hat{z}^n} p(x^n,y^n,z^n)p(\hat{x}^n|x^n) \times \\
&p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n)d_3(z^n,\hat{z}^n)
\end{aligned} \tag{3}$$

where $s \triangleq (\alpha, \beta, \nu)$ denotes the standard Lagrange multiplier, and the base of the logarithm is e. For brevity, we shall denote $(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n})$ by $\mathbf{P}_n$, and $q_{\hat{X}^n\hat{Y}^n\hat{Z}^n} = q_{\hat{X}^n}q_{\hat{Y}^n|\hat{X}^n}q_{\hat{Z}^n|\hat{X}^n\hat{Y}^n}$ by $\mathbf{Q}_n$. Write $F_{s,n}(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}, q_{\hat{X}^n\hat{Y}^n\hat{Z}^n})$ accordingly as $F_{s,n}(\mathbf{P}_n, \mathbf{Q}_n)$. When there is no ambiguity, the superscript or subscript n will be dropped. The iterative algorithm works as follows.

Step 1: Initialize $i = 0$ and set $\mathbf{Q}^{(0)} \triangleq q^{(0)}_{\hat{X}\hat{Y}\hat{Z}}$ as a joint distribution function over $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$ and $\hat{\mathcal{Z}}$.

Step 2: Fix $\mathbf{Q}^{(i)}$. Find $\mathbf{P}^{(i+1)} \triangleq (p^{(i+1)}_{\hat{X}|X}, p^{(i+1)}_{\hat{Y}|\hat{X}YX}, p^{(i+1)}_{\hat{Z}|\hat{X}\hat{Y}Z})$ such that

$$\mathbf{P}^{(i+1)} \triangleq \arg\min_{\mathbf{P}} F_s(\mathbf{P}, \mathbf{Q}^{(i)}) \tag{4}$$

where the minimum is taken over all transition probability functions $\mathbf{P} = (p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z})$. $\mathbf{P}^{(i+1)}$ can be further derived as follows

$$p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z) = \frac{q^{(i)}(\hat{z}|\hat{x}\hat{y})e^{-\nu d_3(z,\hat{z})}}{\Delta^{(i)}(z,\hat{x},\hat{y})} \tag{5}$$

where $\Delta^{(i)}(z,\hat{x},\hat{y}) \triangleq \sum_{\hat{z}} q^{(i)}(\hat{z}|\hat{x}\hat{y})e^{-\nu d_3(z,\hat{z})}$;

$$p^{(i+1)}(\hat{y}|\hat{x}yx) = \frac{q^{(i)}(\hat{y}|\hat{x})e^{-\beta d_2(y,\hat{y})}}{\Lambda^{(i)}(x,y,\hat{x})} \times e^{\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})} \tag{6}$$

where $\Lambda^{(i)}(x,y,\hat{x}) \triangleq \sum_{\hat{y}} q^{(i)}(\hat{y}|\hat{x}) \times e^{-\beta d_2(y,\hat{y})}e^{\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})}$; and

$$p^{(i+1)}(\hat{x}|x) = \frac{q^{(i)}(\hat{x})e^{-\alpha d_1(x,\hat{x})}}{\Gamma^{(i)}(x)} \times e^{\sum_y p(y|x)\log\Lambda^{(i)}(x,y,\hat{x})} \tag{7}$$

where $\Gamma^{(i)}(x) \triangleq \sum_{\hat{x}} q^{(i)}(\hat{x})e^{-\alpha d_1(x,\hat{x})} \times e^{\sum_y p(y|x)\log\Lambda^{(i)}(x,y,\hat{x})}$.

Step 3: Fix $\mathbf{P}^{(i+1)}$. Find $\mathbf{Q}^{(i+1)}$ such that

$$\mathbf{Q}^{(i+1)} \triangleq \arg\min_{\mathbf{Q}} F_s(\mathbf{P}^{(i+1)}, \mathbf{Q}) \tag{8}$$

where the minimum is taken over all joint distribution functions $\mathbf{Q}$ over $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$ and $\hat{\mathcal{Z}}$. (8) is solved by the following equation: for any $(\hat{x},\hat{y},\hat{z}) \in \hat{\mathcal{X}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Z}}$,

$$\begin{aligned}
q^{(i+1)}(\hat{x}\hat{y}\hat{z}) &= \sum_{x,y,z} p(xyz)p^{(i+1)}(\hat{x}|x) \times \\
&p^{(i+1)}(\hat{y}|\hat{x}yx)p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z).
\end{aligned} \tag{9}$$

Step 4: Increase i by 1. Record $F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)})$ as $F_s^{(i)}$.

Step 5: Repeat Steps 2–4 until $F_s^{(i)} - F_s^{(i+1)}$ is smaller than a prescribed threshold.

For brevity, let $\mathbf{Q}(\mathbf{P})$ denote the joint probability function over $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$ and $\hat{\mathcal{Z}}$ obtained from $\mathbf{P}$ through (9). The following theorem shows that the sequence $\{(\mathbf{Q}^{(i-1)}, \mathbf{P}^{(i)}) : i \geq 1\}$ obtained by our iterative algorithm converges to a quadruple of distributions that achieve

$$F_s^* \triangleq \inf F_s(p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z}, q_{\hat{X}\hat{Y}\hat{Z}})$$

where the infimum is taken over all possible $p_{\hat{X}|X}$, $p_{\hat{Y}|\hat{X}YX}$, $p_{\hat{Z}|\hat{X}\hat{Y}Z}$, and $q_{\hat{X}\hat{Y}\hat{Z}}$.

Theorem 3: There exists a $\mathbf{P}^* = (p^*_{\hat{X}|X}, p^*_{\hat{Y}|\hat{X}YX}, p^*_{\hat{Z}|\hat{X}\hat{Y}Z})$ such that as $i \to \infty$, $\mathbf{P}^{(i)} \to \mathbf{P}^*, \mathbf{Q}^{(i)} \to \mathbf{Q}^* = \mathbf{Q}(\mathbf{P}^*)$, and $F_s(\mathbf{P}^*, \mathbf{Q}^*) = F_s^*$.

Remark 1: The above iterative algorithm can be easily extended to the case of $N > 3$, and Theorem 3 remains valid. By setting $\nu = 0$, it also reduces to the case of $N = 2$.

Remark 2: The iterative algorithm can be further extended to work for coupled distortion measures (as defined in [3]) $d'_k : \mathcal{X}_k \times \hat{\mathcal{X}}_k \times \hat{\mathcal{X}}_{k-1} \times \cdots \times \hat{\mathcal{X}}_1 \to [0,\infty)$, $k = 2, \cdots, N$, where the distortion $d'_k(X_k, \hat{X}_k|\hat{X}_k^-)$ depends not only on $(X_k, \hat{X}_k)$ but also on $(\hat{X}_1, \cdots, \hat{X}_{k-1})$. Its global convergence as expressed in Theorem 3 is still guaranteed.

Remark 3: It is worthwhile to point out that the Blahut-Arimoto algorithm [5] can not be applied directly to compute $R_{c,n}(D_1, \cdots, D_N)$ since the corresponding optimization problem has N Markov chain conditions in addition to the standard distortion constraints, which makes the problem become a non-convex optimization problem. Although there are many other ways to derive iterative procedures, their global convergence can not be guaranteed.

IV. Single-letter Characterization: IID Causal Case

Suppose now that $(X_1, \cdots, X_N)$ is IID. In this case, both $\mathcal{R}^*_c(D_1, \cdots, D_N)$ and $R_c(D_1, \cdots, D_N)$ have their single-letter characterizations, as shown in the following theorems[1].

[1]When $N = 2$, Theorems 4 and 5 reduce to Theorems 1 and 3 in [3], respectively. However, the proofs in [3] are incomplete due to the invalid claim of the Markov condition made in the proofs therein; as such formulas therein can not be extended to the case of $N > 2$.

Theorem 4: For an IID vector source, $\mathcal{R}_c^*(D_1,\cdots,D_N) = co(\mathcal{R}_{c,1}(D_1,\cdots,D_N))$.

Theorem 5: For an IID vector source, $R_c(D_1,\cdots,D_n) = R_{c,1}(D_1,\cdots,D_N)$.

Theorem 4 can be proved via the standard converse proof technique in multi-user information theory by introducing right auxiliary random variables. To prove Theorem 5, we will apply our iterative algorithm in a novel way, not for numerical calculation, but for analytically comparing $R_{c,n}(D_1,\cdots,D_N)$ with $R_{c,1}(D_1,\cdots,D_N)$.

Sketch of the Proof of Theorem 5: The global convergence of our iterative algorithm enables us to verify whether a pair $(\mathbf{P},\mathbf{Q})$ is a solution to $F_{s,n}^*$ by checking if $(\mathbf{P},\mathbf{Q})$ is a stationary point of our iterative algorithm. This in turn allows us to verify whether a product form of a solution to $R_{c,1}(D_1,\cdots,D_N)$ will be a solution to $R_{c,n}(D_1,\cdots,D_N)$.

Without loss of generality, once again, we consider the case of $N=3$ and denote three sources by $X, Y,$ and Z. Since the vector source (X,Y,Z) is IID, we have $p_{X^n} = \prod_{i=1}^n p_{X_i}$, $p_{X^nY^n} = \prod_{i=1}^n p_{X_iY_i}$, and $p_{X^nY^nZ^n} = \prod_{i=1}^n p_{X_iY_iZ_i}$. Suppose that $\mathbf{P}_1 = (p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z})$ and the corresponding $\mathbf{Q}_1 = q_{\hat{X}\hat{Y}\hat{Z}}$ obtained through (5),(6),(7), and (9) are the solution to $F_{s,1}^*$ for the case of $n=1$. This implies that if $\mathbf{Q}_1^{(0)}$ is set to be $\mathbf{Q}_1$ in our iterative algorithm for the case of $n=1$, then $\mathbf{P}_1^{(1)}$ obtained in Step 2 of our iterative algorithm will remain to be $\mathbf{P}_1$. Now consider the n-fold product $\mathbf{P}_n$ of $\mathbf{P}_1$ for any $n>1$. That is, $\mathbf{P}_n = (p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n})$ is defined by

$$p(\hat{x}^n|x^n) = \prod_{j=1}^n p(\hat{x}_j|x_j) \tag{10}$$

$$p(\hat{y}^n|\hat{x}^ny^nx^n) = \prod_{j=1}^n p(\hat{y}_j|\hat{x}_jy_jx_j) \tag{11}$$

$$p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n) = \prod_{j=1}^n p(\hat{z}_j|\hat{x}_j\hat{y}_jz_j). \tag{12}$$

Using $\mathbf{P}_n$ we initialize $\mathbf{Q}_n^{(0)} = q_{\hat{X}^n\hat{Y}^n\hat{Z}^n}^{(0)}$ as follows

$$\begin{aligned} & q^{(0)}(\hat{x}^n\hat{y}^n\hat{z}^n) \\ = & \sum_{x^n,y^n,z^n} p(x^n,y^n,z^n)p(\hat{x}^n|x^n) \times \\ & p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n) \\ = & \prod_{j=1}^n q(\hat{x}_j\hat{y}_j\hat{z}_j). \end{aligned} \tag{13}$$

Fix $\mathbf{Q}_n^{(0)}$. It can be verified from (5),(6) and (7) that $\mathbf{P}_n^{(1)}$ obtained in Step 2 of the iterative algorithm has the following form

$$\begin{aligned} \mathbf{P}_n^{(1)} &= (p^{(1)}(\hat{x}^n|x^n), p^{(1)}(\hat{y}^n|\hat{x}^ny^nx^n), p^{(1)}(\hat{z}^n|\hat{x}^n\hat{y}^nz^n)) \\ &= \left(\prod_{j=1}^n p(\hat{x}_j|x_j), \prod_{j=1}^n p(\hat{y}_j|\hat{x}_jy_jx_j), \right. \\ & \left. \prod_{j=1}^n p(\hat{z}_j|\hat{x}_j\hat{y}_jz_j)\right). \end{aligned} \tag{14}$$

Checking (14) against (10), (11) and (12) , we see that $\mathbf{P}_n^{(1)} = \mathbf{P}_n$, which indicates that $(\mathbf{P}_n, \mathbf{Q}_n^{(0)})$ is indeed a stationary point. It thus follows from the global convergence of the iterative algorithm that

$$\begin{aligned} F_{s,n}^* &= F_{s,n}(\mathbf{P}_n, \mathbf{Q}_n^{(0)}) \\ &= \sum_{j=1}^n F_{s,1}(\mathbf{P}_1, \mathbf{Q}_1) \\ &= nF_{s,1}^*. \end{aligned} \tag{15}$$

Since (15) holds for any (s,n), we conclude that for any $\mathbf{D} = (D_1,D_2,D_3)$ and any $n>1$,

$$R_{c,n}(D_1,\cdots,D_N) = R_{c,1}(D_1,\cdots,D_N).$$

which, together with Theorem 2, in turn implies that

$$R_c(D_1,\cdots,D_N) = R_{c,1}(D_1,\cdots,D_N).$$

This completes the proof of Theorem 5.

Remark 4: Note that the characterization of the minimum total rate does not involve any auxiliary random variables other than $\hat{X}_j, j=1,2,...,N$ taking values from the reconstruction alphabets $\hat{\mathcal{X}}_j, j=1,2,...,N$. This is in contrast with the achievable region.

V. More and Less Coding Theorem

To gain deep insights into how each frame in the causal video coding should be efficiently encoded, in this section, we compare $R_c(D_1,\cdots,D_N)$ among different values of N.

To be specific, we will compare the case of $N=3$ with that of $N=2$. Let $R_c^{X_1X_2X_3}(D_1,D_2,D_3)$ denote the minimum total rate required to encode three source frames X_1, X_2 and X_3 at the distortion level $D_1,D_2,D_3 \geq 0$, and $R_c^{X_2X_3}(D_2,D_3)$ denote the minimum total rate required to encode two source frames X_2 and X_3 at the distortion level D_2 and D_3. When X_1, X_2, and X_3 are jointly stationary ergodic and form a (first-order) Markov chain, we have the following result.

Theorem 6: If X_1, X_2, and X_3 form a (first-order) Markov chain, and X_1 is stationary ergodic, then

$$R_c^{X_1X_2X_3}(D_1,D_2,D_3) \geq R_c^{X_2X_3}(D_2,D_3)$$

for any $D_1,D_2,D_3 \geq 0$.

Theorem 6 is what one would expect and consistent with our intuition. Let us now look at the case where X_1, X_2, and X_3 do not form a (first-order) Markov chain. Define

$$D_{1,max} \triangleq \min\{D_1 : R_{X_1}(D_1) = 0\},$$
$$\text{and } D_{2,max} \triangleq \min\{D_2 : R_{X_2}(D_2) = 0\}$$

where $R_X(D)$, for any source X, is the classical rate distortion function of X.

Theorem 7 (More and less coding theorem): Suppose that (X_1, X_2, X_3) is an IID vector source, and X_1, X_2, and X_3 do not form a (first-order) Markov chain. Then for any D_2 and D_3 satisfying $D_2 < D_{2,max}$ and $R_c^{X_2X_3}(D_2, D_3) > 0$ and for which the transitional probability functions to auxiliary random variables achieving $R_c^{X_2X_3}(D_2, D_3)$ satisfy some mild conditions, there is a $D_1^* < D_{1,max}$ such that

$$R_c^{X_1X_2X_3}(D_1, D_2, D_3) < R_c^{X_2X_3}(D_2, D_3)$$

for any $D_1 > D_1^*$.

Theorem 7 is really surprising and counter intuitive. It says that whenever the conditions specified in Theorem 7 are met, the more source frames need to be encoded and transmitted, the less amount of data has to be actually sent!

Remark 5: One needs the conditions specified in Theorem 7 to rule out some corner cases such as the case where X_1, X_2, and X_3 do not form a (first-order) Markov chain, but X_2 and X_3 are independent. The conditions are nonetheless really minor and satisfied by most sources which do not form a (first-order) Markov chain, as shown in the following examples.

Example 1: Suppose that $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \hat{\mathcal{X}}_1 = \hat{\mathcal{X}}_2 = \hat{\mathcal{X}}_3 = \{0, 1\}$, and that the Hamming distortion measure is used. Let $p_{X_1}(0) = 1/3$, $p_{X_2|X_1}(0|1) = p_{X_2|X_1}(1|0) = 3/5$, and $p_{X_3|X_1X_2} =$

$$\begin{pmatrix} & X_1X_2 & 00 & 01 & 10 & 11 \\ X_3 & & & & & \\ 0 & & 0.97 & 0.03 & 0.03 & 0.97 \\ 1 & & 0.03 & 0.97 & 0.97 & 0.03 \end{pmatrix}.$$

It is easy to see that X_1, X_2 and X_3 do not form a Markov chain. We consider the following three cases.

Case 1: $D_1 = 0.3100 < D_{1,max}$, and $D_3 = 0.1500$.
Case 2: $D_2 = 0.2000 < D_{2,max}$, and $D_3 = 0.1500$.
Case 3: $D_2 = 0.2200 < D_{2,max}$, and $D_3 = 0.2300$.

For Case 1, Figure 2 shows the rate-distortion curves of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_2. Over the interval of D_2 shown in Figure 2, it is clear that $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ is always strictly less than $R_c^{X_2X_3}(D_2, D_3)$.

For Case 2, Figure 3 shows $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 with fixed $D_2 < D_{2,max}$ and D_3. It is observed that the critical point at which $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ meets $R_c^{X_2X_3}(D_2, D_3)$ is the intersection of the two curves. Denote this critical point by D_1^*. Then it is clear that when $D_1 > D_1^*$, $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ is indeed strictly less than $R_c^{X_2X_3}(D_2, D_3)$.

When we assign different values to $D_2 < D_{2,max}$ and D_3, we observe the same phenomenon, as shown again in Figure 4 for Case 3.

Let us now look at another example with a different joint distribution.

Example 2: Suppose that $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \hat{\mathcal{X}}_1 = \hat{\mathcal{X}}_2 = \hat{\mathcal{X}}_3 = \{0, 1\}$, and that the Hamming distortion measure is used.

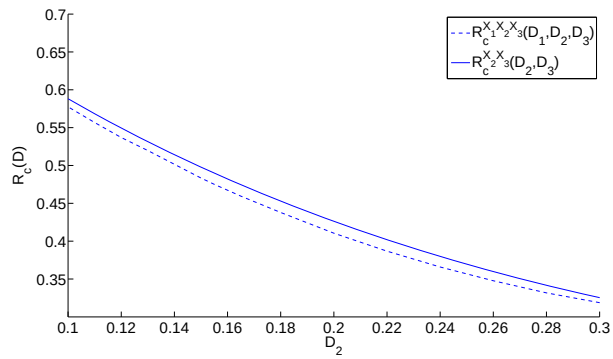

Fig. 2. Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_2 for fixed $D_1 = 0.3100$ and $D_3 = 0.1500$.

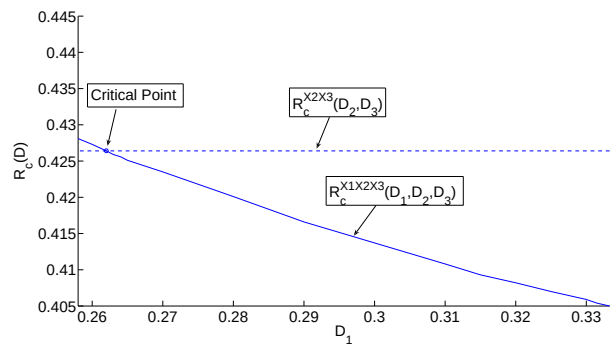

Fig. 3. Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.2000$ and $D_3 = 0.1500$ in example 1.

Let $p_{X_1}(0) = 1/10$, $p_{X_2|X_1}(0|1) = p_{X_2|X_1}(1|0) = 1/10$, and $p_{X_3|X_1X_2} =$

$$\begin{pmatrix} & X_1X_2 & 00 & 01 & 10 & 11 \\ X_3 & & & & & \\ 0 & & 0.80 & 0.05 & 0.1 & 0.92 \\ 1 & & 0.20 & 0.95 & 0.9 & 0.08 \end{pmatrix}.$$

Once again, X_1, X_2 and X_3 do not form a Markov chain. Fix $D_2 = 0.0988 < D_{2,max}$ and $D_3 = 0.0911$. Figure 5 shows the two rate distortion curves $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1. The same phenomenon is revealed as in example 1 .

For all cases shown in Examples 1 and 2, in comparison with $R_c^{X_2X_3}(D_2, D_3)$, when we include X_1 in the encoding and transmission, we not only get the reconstruction of X_1 free at the receiver end, but are also able to reduce the the number of bits to be transmitted. In other words, we can achieve a double gain.

VI. Predictive Video Coding

In this section, we investigate the rate performance of predictive video coding. We have the following result.

Theorem 8: If the jointly stationary and ergodic sources $X_1, \cdots, X_N$ form a (first-order) Markov chain, then

$$R_c(D_1, \cdots, D_N) = R_p(D_1, \cdots, D_N)$$

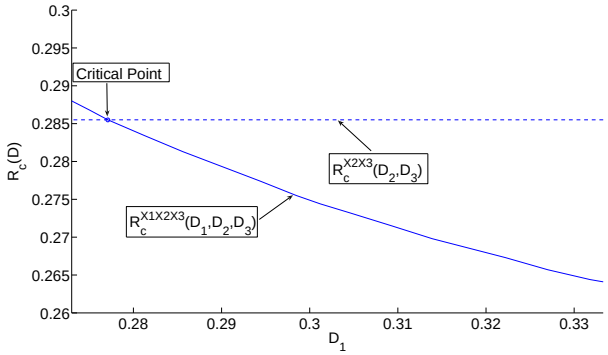

Fig. 4. Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.2200$ and $D_3 = 0.2300$ in example 1.

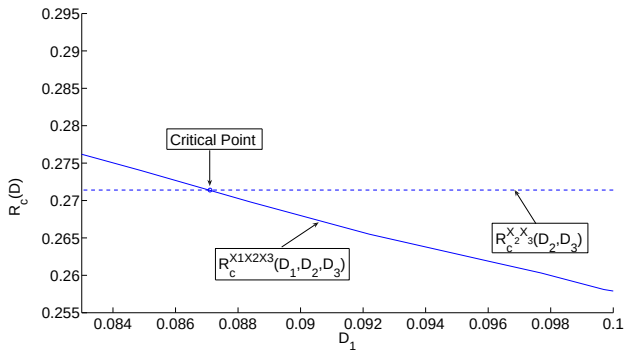

Fig. 5. Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.0988$ and $D_3 = 0.0911$ in example 2.

for any $D_1, \cdots, D_N \geq 0$.

Theorem 8 implies that all the information theoretic results and the computation algorithm on causal video coding can be directly applied to predictive video coding when sources form a (first-order) Markov chain. However, when $X_1, \cdots, X_N$ do not form a (first-order) Markov chain, the problems of characterizing, computing, and comparing $R_p(D_1, \cdots, D_N)$ are still open.

VII. Conclusion

In this paper, we investigated the causal coding model of source frames $X_1, \cdots, X_N$ from an information theoretic point of view. An iterative algorithm was proposed to numerically compute the minimum total rate $R_c(D_1, \cdots, D_N)$ for jointly stationary ergodic sources at distortion level $D_1, \cdots, D_N \geq 0$, and analytically characterize $R_c(D_1, \cdots, D_N)$ for IID sources $(X_1, \cdots, X_N)$. The global convergence of the algorithm was also demonstrated. The algorithm also gives an optimal solution for bit allocation among different frames. By comparing $R_c(D_1, \cdots, D_N)$ among different values of N with the help of the algorithm, we further established a surprising more and less coding theorem—under some conditions on source frames and distortion, the more frames need to be coded and transmitted, the less amount of data has to be sent! Predictive coding was also investigated.

References

[1] Iain E.G. Richardson, *H.264 and MPEG-4 Video Compression*, New York: NY: Wiley, 2003.

[2] E.-h. Yang, Lin Zheng, and Da-ke He, and Zhen Zhang "The rate distortion theory for causal video coding: Characterization, computation algorithm, and comparison," *in preparation*.

[3] H. Viswanathan, and T. Berger, "Sequential coding of correlated sources," *IEEE Trans. Inform. Theory*, vol. 46, pp. 236–246, Jan. 2000.

[4] Nan Ma, and Prakash Ishwar, "On delayed sequential coding of correlated sources," *arXiv: cs/0701197v2 [CS.IT]*, Sep. 30 2008.

[5] R. E. Blahut, "Computation of channel capacity and rate-distortion function," *IEEE Trans. Inform. Theory*, vol. 18, pp. 460–473, 1972.

[6] E.-h. Yang and X. Yu, "Rate distortion optimization for H.264 inter-frame video coding: A general framework and algorithms," *IEEE Trans. on Image Processing*, Vol.16, No.7, pp. 1774–1784, July 2007.