On the Optimal Reconstruction Kernel of Causal Rate Distortion Function

Charalambos D. Charalambous, Christos K. Kourtellaris and Photios Stavrou

Abstract— **This paper considers source coding of general sources with memory, when causal feedback is available at the decoder. The rate distortion function defined as the infimum of the directed information between source and reconstruction sequences over causal data compression channels, which satisfy a distortion fidelity constraint. The form of the optimal causal data compression Kernel is derived.**

I. Introduction

Over the past few years there has been a renewed interest in communication systems with feedback, and in designing causal encoders and decoders. The emergence of sensor technology stimulated interest in new applications involving control and communication networks, in which source coding and channel coding received significant attention. An important application which involves communication and control analysis and design is that of controlling dynamical systems over finite rate communication channels [1], [2], [3], [4], [5], [6]. These type of control/communication systems are often represented by a dynamical control system whose output is the source which is reproduced at the output of the channel, while the output of the channel is the input to the controller. Hence, dynamical system outputs are reliably reproduced at the decoder, which drives the control law of the dynamical system in real-time. When feedback or side information is available at the communication/control blocks, the encoder and/or decoder, and controller, then causality of the information entering these blocks should be ensured for real-time applications. Since feedback to the control system is applied via a finite rate communication channel (which might be noisy or noiseless), then quantization of the output of the dynamical system is required to reliably communicate the information over a finite rate channel to the controller . One of the fundamental problems often encountered in communication systems and/or control/communication systems is causal lossy compression, in which information should be provided causally to the encoder and/or decoder, and controller. A typical scenario is depicted in Figure II.1. This systems consist of a control system (sensors, actuators, dynamical model) and a communication system (encoders, decoders). The information of the dynamical system output in communicated to the controller via a finite rate communication channel.

The objective of this paper is to investigate the source coding problem, when the decoder has causal feedback for general uncontrolled sources, via the rate distortion function. However, unlike the classical rate distortion theory [7], which employs the mutual information to represent the rate between the source sequence and the reconstruction of source sequence, causality of the decoder requires the use of directed information [8], a variant of the mutual information. Specifically, the optimal classical data compression channel [9], [7] is non-causal. The implications of causality on the data compression channel is explained below.

Consider two sequences, $X^n \triangleq (X_0, X_1, \ldots, X_n)$ denoting the source, and $Y^n \triangleq (Y_0, Y_1, \ldots, Y_n)$, denoting the reconstruction of the source, taking values in $\mathcal{X}_{0,n}$ and $\mathcal{Y}_{0,n}$ respectively. Let $\mathcal{M}_1(\mathcal{X})$ denote the space of probability measure on $\mathcal{X}$.

Shannon's self-mutual information for a given realization $X^n = x^n, Y^n = y^n$ of these sequences is defined by $i(x^n; y^n) \triangleq \log \frac{q(dy^n;x^n)}{\nu(dy^n)}$, where $q(dy^n; x^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ denotes conditional distribution and $\nu(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$, while its average over all realizations, called Shannons mutual information is defined by [10]

$$I(X^n; Y^n) \triangleq E_{P(dx^n,dy^n)}\Big\{ i(x^n; y^n) \Big\} = \int \log \frac{q(dy^n; x^n)}{\nu(dy^n)} q(dy^n; x^n) \mu(dx^n)$$

where $\mu(dx^n) \in \mathcal{M}_1(\mathcal{X}_{0,n})$). The classical source coding or lossy data compression is defined by introducing an average distortion or fidelity constraint associated with a distortion measure $\rho_n : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \rightarrow [0, \infty), n \in \mathbb{N}$. The data compression channel which minimizes the rate of reconstructing X^n by Y^n [9] is given by

$$q^*(dy^n; x^n) = \frac{e^{s\rho_n(x^n,y^n)} \nu^*(dy^n)}{\int_{\mathcal{Y}_{0,n}} e^{s\rho(x^n,z^n)} \nu^*(dz^n)}, \quad s \leq 0 \qquad \text{(I.1)}$$

where $\nu^*(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the marginal of $P^*(dx^n, dy^n) = \mu(dx^n) \otimes q^*(dy^n; x^n) \in \mathcal{M}_1(\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n})$, and $s \leq 0$ is the Largange multiplier associated with the fidelity constraint. Even for single letter distortion $\rho_n(x^n, y^n) = \sum_{i=0}^n \rho_i(x_i, y_i)$ it follows from (I.1) that $q^*(dy^n; x^n) = \times_{j=0}^n q^*(dy_j; y^{j-1}, x^n) \neq \times_{j=0}^n q^*(dy_j; y^{j-1}, x^i)$, hence causality of the data compression fails.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. INFSO-ICT-223844 and the Cyprus Research Promotion Foundation under the grant ARTEMIS-0506/20.

C.D. Charalambous is with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. E-mail:chadcha@ucy.ac.cy

Christos K. Kourtellaris is with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. E-mail:christoskourtellaris@yahoo.gr

Photios Stavrou is with the Department of Electrical Engineering, University of Cyprus, Nicosia, Cyprus. E-mail:fstavr01@ucy.ac.cy

ISBN 978-963-311-370-7

The main objective of this paper is to re-define the rate distortion function and ensure the optimization leads to a causal data compression channel. To further explain the implications of causality and feedback consider the following alternative expressions of self-mutual information.

$$i(x^n;y^n)=\sum_{i=0}^{n}\log\frac{q_i(dy_i;y^{i-1},x^i)}{\nu_i(dy_i;y^{i-1})}+\sum_{i=0}^{n}\log\frac{\hat{q}_i(dx_i;y^{i-1},y^{i-1})}{\mu_i(dx_i;x^{i-1})}$$

where $q_i \in \mathcal{Q}(\mathcal{Y}_i;\mathcal{Y}_{0,i-1}\times\mathcal{X}_{0,i})$, $\nu_i \in \mathcal{Q}(\mathcal{Y}_i;\mathcal{Y}_{0,i-1})$, $\mu_i \in \mathcal{Q}(\mathcal{X}_i;\mathcal{X}_{0,i-1})$, $\hat{q}_i \in \mathcal{Q}(\mathcal{X}_i;\mathcal{Y}_{0,i-1}\times\mathcal{X}_{0,i-1})$. By taking expectation

$$I(X^n;Y^n)=I(X^n\rightarrow Y^n)+I(X^n\leftarrow Y^n)$$

where

$$I(X^n\rightarrow Y^n)\triangleq\sum_{i=0}^{n}I(X^i;Y_i|Y^{i-1})=\sum_{i=0}^{n}\int\log\frac{q_i(dy_i;y^{i-1},x^i)}{\nu_i(dy_i;y^{i-1})}P(dy^i,dx^i)$$

$$I(X^n\leftarrow Y^n)\triangleq\sum_{i=0}^{n}I(Y^{i-1};X_i|X^{i-1})=\sum_{i=0}^{n}\int\log\frac{\hat{q}_i(dx_i;x^{i-1},y^{i-1})}{\mu_i(dx_i;x^{i-1})}P(dy^{i-1},dx^i).$$

Note that the term $I(X^n\rightarrow Y^n)$ is the directed information from X^n to Y^n discussed in [8], and corresponds to the Shannon mutual information restricted to the channel connecting causally X^n to Y^n, while $I(X^n\leftarrow Y^n)$ is the directed information from Y^n to X^n, which corresponds to the Shannon mutual information restricted to the channel connecting causally X^n to Y^n. Thus, the formulation of causal rate distortion function should involve only the term $I(X^n\rightarrow Y^n)$.

The main objectives of this paper are to provide the definition of the causal rate distortion function for general sources, and derive the optimal causal reconstruction kernel.

Previous related work on causal rate distortion is found in [11], where coding theorems are also derived. Recently, the problem is revisited in [12], [13]. An alternative non-information theoretic approach is found in [14], [15] using stochastic optimization methods. The material presented in this paper compliments previous work in causal data compression in the sense that we provide the formulation, as well as the optimal causal reproduction kernel, for general sources.

II. Problem Formulation

In this section, we introduce the set up of the problem on abstract alphabets (Polish spaces) and a discrete time set $\mathbb{N}^n\triangleq\{0,1,\ldots,n\}$, $n\in\mathbb{N}\triangleq\{0,1,2,\ldots\}$. All processes are defined on a complete probability space $(\Omega,\mathcal{F}(\Omega),\mathbb{P})$ with filtration $\{\mathcal{F}_t\}_{t\geq0}$. The source and reconstruction alphabets are sequences of Polish spaces $\{\mathcal{X}_t:t=0,1,\ldots,n\}$ and $\{\mathcal{Y}_t:t=0,1,\ldots,n\}$, respectively, (e.g., $\mathcal{Y}_t,\mathcal{X}_t$ are complete separable metric spaces). Moreover, the abstract alphabets are associated with their corresponding measurable spaces $(\mathcal{X}_t,\mathcal{B}(\mathcal{X}_t))$ and $(\mathcal{Y}_t,\mathcal{B}(\mathcal{Y}_t))$ (e.g., $\mathcal{B}(\mathcal{X}_t)$ is a Borel $\sigma-$algebra of subsets of the set $\mathcal{X}_t$ generated by closed sets). Thus, sequences of source and reproduction of the source alphabets are identified with the product measurable spaces $(\mathcal{X}_{0,n},\mathcal{B}(\mathcal{X}_{0,n})\triangleq\times_{k=0}^{n}(\mathcal{X}_k,\mathcal{B}(\mathcal{X}_k))$, and $(\mathcal{Y}_{0,n},\mathcal{B}(\mathcal{Y}_{0,n})\triangleq\times_{k=0}^{n}(\mathcal{Y}_k,\mathcal{B}(\mathcal{Y}_k))$, respectively. The source is a random process denoted by $X^n\triangleq\{X_t:t=0,1,\ldots,n\}$, $X:\mathbb{N}^n\times\Omega\mapsto\mathcal{X}_t$, and the reconstruction of the source is another process denoted by $Y^n\triangleq\{Y_t:t=0,1,\ldots,n\}$, $Y:\mathbb{N}^n\times\Omega\mapsto\mathcal{Y}_t$, where the subscript denotes the time evolution of the processes.

Probability measures on any measurable space $(\mathcal{Z},\mathcal{B}(\mathcal{Z}))$ are denoted by $\mathcal{M}_1(\mathcal{Z})$.

Next, we introduce the definition of conditional independence.

Conditional Independence: Conditionally independent Random Variables (R.V's) are denoted by $(X,Y)\perp Z$ or equivalently $Y\leftrightarrow Z\leftrightarrow X$ form a Markov chain.

Conditional distributions are identified by stochastic kernels as defined below.

Definition 2.1: Given the measurable spaces $(\mathcal{X},\mathcal{B}(\mathcal{X})),(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$, a stochastic Kernel on $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ conditioned on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ is a mapping $q:\mathcal{B}(\mathcal{Y})\times\mathcal{X}\rightarrow[0,1]$ satisfying the following two properties:

1) For every $x\in\mathcal{X}$, the set function $q(\cdot;x)$ is a probability measure (possibly finitely additive) on $\mathcal{B}(\mathcal{Y})$;

2) For every $A\in\mathcal{B}(\mathcal{Y})$, the function $q(A;\cdot)$ is $\mathcal{B}(\mathcal{X})$-measurable.

The set of all stochastic Kernels on $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ conditioned on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ are denoted by $\mathcal{Q}(\mathcal{Y};\mathcal{X})$.

The definition of stochastic kernel can be used to define a causal and non-causal rate distortion channels (reproduction kernels) as follows.

Definition 2.2: Given the measurable spaces $(\mathcal{X}_{0,n},\mathcal{B}(\mathcal{X}_{0,n}))$ and $(\mathcal{Y}_{0,n},\mathcal{B}(\mathcal{Y}_{0,n})),n\in\mathbb{N}$, and their product spaces, data compression channels are defined as follows.

1. *Causal Data Compression Channel.* A causal data compression channel is a sequence of stochastic kernels $\{q_j(dy_j;y^{j-1},x^j)\in\mathcal{Q}(\mathcal{Y}_j;\mathcal{Y}_{0,j-1}\times\mathcal{X}_{0,j}):j\in\mathbb{N}^n\}$

2. *Non-Causal Data Compression Channel.* A non-causal data compression channel is a stochastic kernel $q_{0,n}(dy^n;x^n)\in\mathcal{Q}(\mathcal{Y}_{0,n};\mathcal{X}_{0,n})$.

Thus, a causal data compression channel is a sequence of conditional distributions for Y_j given $Y^{j-1}=y^{j-1}$ and $X^j=x^j$ denoted by $P_{Y_j|Y^{j-1},X^j}(dy_j|Y^{j-1}=y^{j-1},X^j=$

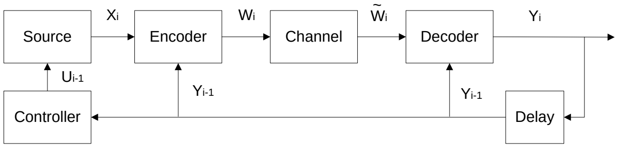

Fig. II.1. Control/Communication System with Feedback

$x^j), j = 0, 1, \ldots, n$. On the other hand, a non-causal data compression channel is given by $P_{Y^n|X^n}(dy^n|X^n = x^n)$. Since by chain rule $P_{Y^n|X^n}(dy^n|X^n = x^n) = \bigotimes_{j=0}^n P_{Y_j|Y^{j-1},X^n}(dy_j|Y^{j-1} = y^{j-1}, X^n = x^n)$, it is clear that in classical rate distortion theory the reconstruction of $Y_j = y_j$ depends on future values of the source sequence, namely, $(X_{j+1} = x_{j+1}, \ldots, X_n = x_n)$ in addition to the past reconstructions $Y^{j-1} = y^{j-1}$, and past and present source symbols $X^j = x^j$. Thus, classical rate distortion theory is not restricted to causal relations between source and reconstruction sequences.

III. Rate Distortion Function

The goal of this section is to formulate the rate distortion function subject to a causal constraint on the reconstruction Kernel, and then to derive the optimal casual reconstruction kernel. Next, starting with the abstract formulation of the mutual information between sequences X^n and Y^n defined via relative entropy, we will define the classical rate distortion function, and then deduce from it the formulation of the causal rate distortion function.

Given the source $\hat{q}_i \in \mathcal{Q}(\mathcal{X}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i-1})$, $i = 0, 1, ..., n$, and a causal stochastic kernel $q_i \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i})$ $i = 0, 1, ..., n$ the joint measure $P(dx^n, dy^n)$ and the marginal measures, $\nu_{0,n}(dy^n) \in \mathcal{M}_1(\mathcal{Y}_{0,n})$, $\mu_{0,n}(dx^n) \in \mathcal{M}_1(\mathcal{X}_{0,n})$ are uniquely defined with probability one. Hence, the directed information is also defined via

$$I(X^n \rightarrow Y^n) = \sum_{i=0}^n \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log\Big(\frac{q_i(dy_i; y^{i-1}, x^i)}{\nu_i(dy_i; y^{i-1})}\Big) P(dx^i, dy^i) \tag{III.2}$$

where $P(dx^n, dy^n) = \otimes_{i=0}^n q_i(dy_i; y^{i-1}x^i) \otimes \hat{q}_i(dx_i; x^{i-1}y^{i-1})$ Hence, the following rate distortion definition.

Definition 3.1: Let $\rho_j : \mathcal{X}_{0,j} \times \mathcal{Y}_{0,j} \rightarrow [0, \infty)$ be a sequence of $\mathcal{B}(\mathcal{X}_{0,j}) \times \mathcal{B}(\mathcal{Y}_{0,j})$ measurable distortion functions continuous in the second argument, $0 \leq j \leq n$, and let $\overrightarrow{Q}_{0,n}(D)$ denote the average distortion function defined by

$$\overrightarrow{Q}_{0,n}(D) = \Big\{ q_i(dy_i; y^{i-1}x^i), 0 \leq j \leq n : \frac{1}{n+1} \sum_{j=0}^n \int_{\mathcal{X}_{0,j}} \int_{\mathcal{Y}_{0,j}} \rho_j(x^j, y^j) P(dx^j, dy^j) \leq D \Big\}$$

The causal rate distortion function is defined by

$$\overrightarrow{R}_{0,n}(D) = \inf_{\{q_i(dy_j; y^{j-1}, x^j)\}_{j=0}^n \in \overrightarrow{Q}_{0,n}(D)} \frac{1}{n+1} I(X^n \rightarrow Y^n)$$

The rate distortion function is made precise by first identifying the appropriate spaces on which existence of solution to $\overrightarrow{R}_{0,n}(D)$ is sought, and equivalence between the constrained and unconstrained problems is shown. The claim is that at least for the case when the source is independent of the reconstruction sequence, e.g, $\hat{q}_i(dx_i; x^{i-1}y^{i-1}) = \mu_i(dx_i; x^{i-1})$, P-a.s., $\forall i$, then the weak* convergence found in [16] is sufficient to establish the above mentioned results. However, in the current paper this issue will not be discussed. Rather, the subsequent results will be based on the following assumption.

Assumptions 3.2: Appropriate conditions are assumed so that the constrained problem $\overrightarrow{R}_{0,n}(D)$ is equivalent to the unconstrained problem, specifically, it is assumed that

$$\overrightarrow{R}_{0,n}(D) = \sup_{s \leq 0} \inf_{\{q_j(dy_j; y^{j-1}, x^j) \in \mathcal{Q}(\mathcal{Y}_j; \mathcal{Y}_{0,j-1} \times \mathcal{X}_j)\}_{j=0}^n} = \Big\{ \frac{1}{n+1} I(X^n \rightarrow Y^n) - s\Big(\frac{1}{n+1} \int_{\mathcal{X}_{0,j}} \int_{\mathcal{Y}_{0,j}} \rho_j(x^j, y^j) P(dx^j, dy^j) - D\Big)\Big\}$$

Next, the main result of the paper are presented.

Theorem 3.3: Suppose Assumptions 3.2 hold.

The optimal reconstruction Kernel which achieves the infimum of the rate distortion function is given by

$$q_t^*(dy_t; x^t, y^{t-1}) = \frac{e^{s\rho_t(x^t, y^t)} \nu_t^*(dy_t; y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t(x^t, y^t)} \nu_t^*(dy_t; y^{t-1})}, \tag{III.3}$$

where $s \leq 0$ and denotes the solution of $s = \frac{d}{dD} \overrightarrow{R}_n(D)$. The information rate distortion function is given by

$$\overrightarrow{R}_n(D) = sD - \frac{1}{n+1} \sum_{t=0}^n \int_{\mathcal{X}_{0,t} \times \mathcal{Y}_{0,t-1}} \log \Big(\int_{\mathcal{Y}_t} e^{s\rho_t(x^t, y^t)} \nu_t^*(dy_t; y^{t-1})\Big) \hat{q}_t(dx_t; y^{t-1}, x^{t-1}) \bigotimes_{i=0}^{t-1} q_i^*(dy_i; x^i, y^{i-1}) \otimes \hat{q}_i(dx_i; y^{i-1}, x^{i-1}) \tag{III.4}$$

Proof: Consider $\{q_t(dy_t; x^t, y^{t-1}) : t = 0, .., n\} \in \overrightarrow{Q}(D)$. then

$$\frac{1}{n+1} \int_{\mathcal{Y}_0 \times \mathcal{X}_0} \Big(\sum_{t=0}^n \rho_t(x^t, y^t)\Big) P(dx^n, dy^n) \leq D \tag{III.5}$$

which is equivalent to

$$\frac{1}{n+1} \sum_{t=0}^n \int_{\mathcal{Y}_{0,t} \times \mathcal{X}_{0,t}} \rho_t(x^t, y^t) \bigotimes_{i=0}^t q_i(dy_i; x^i, y^{i-1}) \otimes \hat{q}_i(dx_i; y^{i-1}, x^{i-1}) = \frac{1}{n+1} \sum_{t=0}^n \int_{\mathcal{X}_{0,t} \times \mathcal{X}_{0,t}} \rho_t(x^t, y^t) q_t(dy_t; x^t, y^{t-1}) \otimes \hat{q}_t(dx_t; y^{t-1}, x^{t-1}) \bigotimes_{i=0}^{t-1} q_i(dy_i; x^i, y^{i-1}) \otimes \hat{q}_i(dx_i; y^{i-1}, x^{i-1}) \leq D. \tag{III.6}$$

Define the relative entropy of measure P with respect to measure Q as $\mathbb{D}(P||Q) = \int \log\left(\frac{dP}{dQ}\right)dP$ if $P << Q$ and ∞ otherwise. Then

$$\frac{1}{n+1}I(X^n \rightarrow Y^n) = \frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\times\mathcal{Y}_{0,t}}$$
$$\log\frac{q_t(dy_t;x^t,y^{t-1})}{\nu_t(dy_t;y^{t-1})}P(dx^t,dy^t) = \frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\times\mathcal{Y}_{0,t}}$$
$$\log\frac{q_t(dy_t;x^t,y^{t-1})}{\nu_t(dy_t;y^{t-1})}\bigotimes_{i=0}^{t} q_i(dy_i;x^i,y^{i-1})$$
$$\otimes\hat{q}_i(dx_i;y^{i-1},x^{i-1}) = \frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\times\mathcal{Y}_{0,t-1}} \tag{III.7}$$
$$\mathbb{D}(q_t(.;x^t,y^{t-1})||\nu_t(.;y^{t-1}))\hat{q}_t(dx_t;y^{t-1},x^{t-1})$$
$$\bigotimes_{i=0}^{t-1} q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^i,x^{i-1})$$

Minimizing (III.7) subject to fidelity constraint (III.6) is addressed via the Lagrangian method. The Lagrangian associated with constraint optimization problem $\overrightarrow{R}_n(D)|_R$, is defined by

$$L(\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n, s) \triangleq \Big\{\frac{1}{n+1}I(X^n\rightarrow Y^n)$$
$$-s\Big(\frac{1}{n+1}\int_{\mathcal{Y}_{0,n}\times\mathcal{X}_{0,n}}\Big(\sum_{t=0}^{n}\rho_t(x^t,y^t)\Big)P(dx^n,dy^n) - D\Big)\Big\}$$

where $s \leq 0$ is the Lagrangian multiplier associated with the fidelity constraint. The dual functional is

$$L(\{q_t^*(dY_t;x^t,y^{t-1})\}_{t=0}^n, s^*)$$
$$= \sup_{s\leq 0}\inf_{\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n} L(\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n, s)$$

Notice that under 3.2 $\overrightarrow{R}_n(D) = \sup_{s\leq 0}\inf_{\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n} L(\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n, s)$. Consider the Lagrangian

$$L(\{q_t(dy_t;x^t,y^{t-1})\}_{t=0}^n, s) = sD + \frac{1}{n+1}\sum_{t=0}^{T-1}\int_{\mathcal{X}_{0,t}\otimes\mathcal{Y}_{0,t}}$$
$$\log\left(\frac{q_t(dy_t;x^t,y^{t-1})}{q_t^*(dy_t;x^t,y^{t-1})}\frac{q_t^*(dy_t;x^t,y^{t-1})}{\nu_t(dy_t;y^{t-1})}\right)q_t(dy_t;x^t,y^{t-1})\otimes$$
$$\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$-\frac{s}{n+1}\sum_{t=0}^{n}\int_{\mathcal{Y}_{0,t}\otimes\mathcal{X}_{0,t}}\rho_t(x^t,y^t)q_t(dy_t;x^t,y^{t-1})$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$= sD + \frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\otimes\mathcal{Y}_{0,t-1}}\mathbb{D}\Big(\frac{q_t(\cdot;x^t,y^{t-1})}{q_t^*(\cdot;x^t,y^{t-1})}\Big)$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$+\frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\otimes\mathcal{Y}_{0,t}}\log\left(\frac{q_t^*(dy_t;x^t,y^{t-1})}{\nu_t(dy_t;y^{t-1})}\right)q_t(dy_t;x^t,y^{t-1})$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$-\frac{s}{n+1}\sum_{t=0}^{n}\int_{\mathcal{Y}_{0,t}\otimes\mathcal{X}_{0,t}}\rho_t(x^t,y^t)q_t(dy_t;x^t,y^{t-1})$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$\geq sD + \frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_{0,t}\otimes\mathcal{Y}_{0,t}}\log\left(\frac{q_t^*(dy_t;x^t,y^{t-1})}{\nu_t(dy_t;y^{t-1})}\right)q_t(dy_t;x^t,y^{t-1})$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^{i-1},x^{i-1})$$
$$-\frac{s}{n+1}\sum_{t=0}^{n}\int_{\mathcal{Y}_{0,t}\otimes\mathcal{X}_{0,t}}\rho_t(x^t,y^t)q_t(dy_t;x^t,y^{t-1})$$
$$\otimes\hat{q}_t(dx_t;y^{t-1},x^{t-1})\bigotimes_{i=0}^{t-1}q_i(dy_i;x^i,y^{i-1})\otimes\hat{q}_i(dx_i;y^{i-1}x^{i-1})$$

where the last inequality holds because $\mathbb{D}(\frac{q(.;x^t,y^{t-1})}{q^*(.;x^t,y^{t-1})}) \geq 0, P-a.s.$. Moreover, the above inequality is achieved by letting $\{q_t(dy_t;x^t,y^{t-1}) = q_t^*(dy_t;x^t,y^{t-1})\}_{t=0}^n$ given bx (III.3), while $s \leq 0$, is found by the rate distortion constraint. Substituting (III.3) into $L(\{q_t(dy_t;x^t,y^t)\}_{t=0}^n, s)$ yields (III.4). Taking the derivative with respect to D of the rate distortion function gives s.

Remark 3.4: Notice that the optimal reconstruction kernel (III.3) is causal, hence decoding can be done without waiting to receive the entire sequence x^n before the symbol $y_i, i \leq n$ is reconstructed. The general solution also deals with applications in which the source is independent of the reconstruction, that is $\hat{q}_j(dx_j;x^{j-1},y^{j-1}) = \mu_j(dx_j;x^{j-1})$, P-a.s $\forall i$. in the above expressions. This situation corresponds to Figure II.1 in which there is no feedback between the decoder output and the source. Suppose the source is independent of the reconstruction sequence, that is $\hat{q}_j(dx_j;x^{j-1},y^{j-1}) = \mu_j(dx_j;x^{j-1})$, P-a.s $\forall i$. Then the function spaces considering in [16] are applicable, existence of solution can be shown using weak* convergence, and the constraint and unconstraint problems are equivalent (e.g, Assumptions 3.2 hold).

IV. CONCLUSION AND FUTURE WORK

This paper investigates optimal causal data compression for general sources. The optimal reconstruction kernel is derived, which depends causally on the source output sequence. The material are fundamental and may expanded for the analysis and design of control systems in which feedback is accessible only through a limited rate communication channel.

Future work will investigate 1) conditions under which constrained and unconstrained problems are equivalent, 2) operational meaning of the causal rate distortion function, 3) and new examples of optimal quantization derived via the optimal causal reconstruction kernel.

REFERENCES

[1] K. Tsumura and J. Maciejowski, “Stabilizability of SISO Control Systems under Constraints of Channel Capacities”, in *Proceedings of 42nd IEEE Conference on Decesion and Control*, Hawaii, pp. 193–197, December 2003.

[2] S. Tatikonda and S. Mitter, Control Over Noisy Channels, *IEEE Transactions on Automatic Control*, vol. 49, No. 7, pp. 1196–1201, 2004.

[3] S. Tatikonda and S. Mitter, Control Under Communication Constraints, *IEEE Transactions on Automatic Control*, vol. 49, No. 7, pp. 1056–1068, 2004.

[4] G. N. Nair, R. J. Evans, I. M. Y. Mareels and W. Moran, Topological Feedback Entropy and Nonlinear Stabilization, *IEEE Transations on Automatic Control*, vol. 49, No. 9, pp. 1585–1597, 2004.

[5] K. Li and J. Baillieal, Robust Quantization for Digital Finite Communication Bandwidth (DFCB) Control, *IEEE Transactions on Automatic Control*, vol. 49, No. 9, pp. 1573–1584, 2004.

[6] Charalambos D. Charalambous and Alireza Farhadi, LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels, *Automatica*, vol. 44, pp. 3181–3188, 2008.

[7] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Printice-Hall, 1971.

[8] J. Massey, "Causality, Feedback and Directed Information", in the *in Proceedings of the 1990 IEEE International Symposium on Information Theory and its Applications*, pp. 303–305, Nov.27–30, Hawaii, U.S.A.1990.

[9] I. Csiszar, On an extremum problem of information theory, *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, pp. 57-71, 1974.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[11] S. Tatikonda, *Control Under Communication Constraint*, Ph.D. Thesis, Department of Electrical Enginnering and Computer Science, MIT., September 2000.

[12] R. Venkataramanan, S.S. Pradhan, Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source, *IEEE Transactions on Information Theory*, Vol. 53, No. 6, June 2007.

[13] R. Venkataramanan, Information-theoretic results on communication problems with feed-forward and feedback, Ph.D Thesis, University of Michigan, 2008

[14] D. Teneketzis, On the Structure of Optimal Real-Time Encoders and Decoders in Noisy Communication, *IEEE Transactions on Information Theory*, Vol. IT-52, No. 9, pp.4017-4035, 2006.

[15] J.C. Walrand, P. Varaiya, Optimal causal coding - decoding problems, *IEEE Transactions on Information Theory* Vol. 29, No. 6, pp. 814-819, 1983.

[16] F. Rezaei, N.U. Ahmed and C.D. Charalambous,Rate Distortion Theory for General Sources With Potential Application to Image Compression, *International Journal of Applied Mathematical Sciences*, ISSN 0973-0176 Vol. 3 No. 2, pp. 141-165,2006.