2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

Directed Information Subject to a Fidelity: Applications to Conditionally Gaussian Processes

Charalambos D. Charalambous, Photios A. Stavrou, Christos K. Kourtellaris and Ioannis Tzortzis

***Abstract*— This paper is concerned with the minimization of directed information over conditional distributions that satisfy a fidelity of reconstructing a conditionally Gaussian random process by another process, causally. This information theoretic extremum problem is directly linked, via bounds to the optimal performance theoretically attainable by non-causal, causal and zero-delay codes of data compression. The application example includes the characterization of causal rate distortion function for conditionally Gaussian random processes subject to a mean-square error fidelity.**

***Index Terms*— Directed information, fidelity, rate distortion function, conditionally Gaussian processes.**

I. Introduction

In classical lossy source coding with fidelity [1], a compression encoder-decoder system is called *non-causal* if for each time n, the reconstruction y_n of the source symbol x_n, depends on past, present, and future symbols $(\dots, x_{-1}, x_0, \dots, x_n, x_{n+1}, \dots,)$ i.e., the reproduction coder is $y_n = f_n^{\rm NC}(x_\infty^\infty)$, for $n = 0, 1, \dots$. The optimal performance theoretically attainable (OPTA) by such non-causal reproduction coders with fidelity less than or equal to $D \in [0, \infty)$, is characterized by the classical rate distortion function (RDF), denoted by $R(D)$ [1].

A lossy compression scheme is called *causal* if for each n, the reproduction coder is $y_n \triangleq f_n^C(x_0, \dots, x_n)$ for $n = 0, 1, \dots$ (for details see [2]). The OPTA by causal codes is characterized by the causal entropy distortion function (EDF) denoted by $r^C(D)$. Zero-delay codes are causal codes with the additional property that each reproduction symbol y_n is entropy-coded separately (rather than in long blocks), and encoding and decoding can be carried out instantaneously.

In recent years, bounds on the OPTA by non-causal and causal codes are developed by revisting Gorbunov and Pinsker's nonanticipative rate distortion function (NRDF) [3]–[5], denoted by $R^{\rm na}(D)$, that is a variant of $R(D)$. These bounds state that [6]

$$R(D) \leq R^{\rm na}(D) \leq r^C(D). \tag{I.1}$$

The NRDF. Consider a random process $X^n \triangleq (X_0, \dots, X_n)$ with induced distribution $\mathbf{P}_{X^n}$, and a measurable distortion function $d : \mathcal{X}^n \times \mathcal{Y}^n \to [0, \infty]$ between $x^n \triangleq (x_0, \dots, x_n)$ and its reproduction $y^n \triangleq (y_0, \dots, y_n)$ defined by

$$d(x^n, y^n) \triangleq \frac{1}{n+1} \sum_{t=0}^{n} \rho_t(T^t x^n, T^t y^n), \tag{I.2}$$

where the dependence on $T^t x^n \subseteq \{x_0, \dots, x_t\}$, and $T^t y^n \subseteq \{y_0, \dots, y_t\}$ is fixed, for $t = 0, 1, \dots, n$, and $\rho_t(\cdot, \cdot)$ may be a distance metric. Define the average fidelity over the joint distributions $\mathbf{P}_{X^n, Y^n}$ with marginal $\mathbf{P}_{X^n}$, by

$$\frac{1}{n+1} \mathbf{E}\{d(X^n, Y^n)\} \leq D, \;\; D \in [0, \infty). \tag{I.3}$$

The finite-time NRDF is defined by minimizing the mutual information between X^n to Y^n, subject to an average fidelity, as follows:

$$R_{0,n}^{\rm na}(D) \triangleq \inf_{\substack{P_{Y^n|X^n} : \frac{1}{n+1}\mathbf{E}\{d(X^n,Y^n)\} \leq D \\ X_{t+1}^n \leftrightarrow X^t \leftrightarrow Y^t, \; t=0,1,\dots,n-1}} I(X^n; Y^n), \tag{I.4}$$

where $X_{t+1}^n \leftrightarrow X^t \leftrightarrow Y^t$ forms a Markov chain, i.e., it is equivalent to conditional independence $\mathbf{P}_{Y^t|X^n} = \mathbf{P}_{Y^t|X^t}, t = 0, 1, \dots, n-1$.

The Classical RDF. This is defined as follows:

$$R_{0,n}(D) \triangleq \inf_{P_{Y^n|X^n} : \frac{1}{n+1}\mathbf{E}\{d(X^n,Y^n)\} \leq D} I(X^n; Y^n). \tag{I.5}$$

Moreover, $R^{\rm na}(D) \triangleq \lim_{n \longrightarrow \infty} \frac{1}{n+1} R_{0,n}^{\rm na}(D)$.

Causal Entropy Distortion Function (EDF). $R^{\rm na}(D)$ is a tight lower bound to the OPTA by causal (and zero-delay) codes, see [6].

Recently, the achievability of the NRDF of vector-valued Gauss-Markov processes was established, using a variable-length causal and zero-delay coding scheme by means of an entropy coded dithered quantizer [7]. Achievability of the NRDF of a discrete or continuous alphabet source can also be established using joint source-channel matching (for details see, e.g., [8], [9]).

The previous analysis demonstrates that the NRDF can be used to provide fundamental performance limitations in delay-constrained communication systems. This in turn can be used to evaluate fundamental limitations of feedback control systems over communication channels [10].

A. Main Contributions

In this paper, we generalize the NRDF to processes $\{X_t : t = 0, \dots, n\}$ which are affected by the reproduction process $\{Y_t : t = 0, \dots, n\}$ via a closed-loop feedback system. In view of this generality, we deal with process X^n that is

C. D. Charalambous, C. K. Kourtellaris and I. Tzotzis are with the Department of Electrical and Computer Engineering, University of Cyprus, Cyprus, *email: {chadcha, kourtellaris.christos,tzortzis.ioannis}@ac.ucy.cy*.

P. A. Stavrou is with the Department of Information Science and Engineering, KTH Royal Institute of Technology, Sweden, *emails:fstavrou@kth.se*.

causally affected by previous reproductions Y^{n-1}. Then, we consider the causal RDF, that is defined by the extremum problem of minimizing directed information from X^n to Y^n [11], [12], subject to fidelity. Earlier results on such generalizations are described by the authors in [13]. Here, we additionally derive the characterization of the causal RDF for conditionally Gaussian processes. The results of this paper can be used to analyze and synthesize controllers, when $\{X_t : t = 0, \ldots, n\}$ is the controlled process, which is control over a limited capacity (noisy or noiseless) channel [10], [14], [15] as illustrated in Fig. I.1.

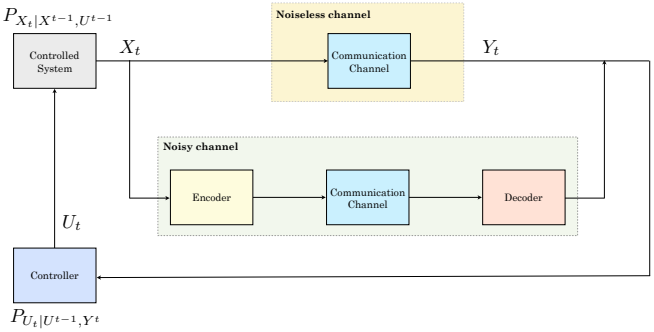

Fig. I.1: Control over limited capacity communication channel.

II. Optimization of Directed Information Subject to Fidelity

Notation: Let $\mathbb{R} = (-\infty, \infty)$, $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$, $\mathbb{Z}_0 = \{0, 1, \ldots\}$ and $\mathbb{Z}_0^n \triangleq \{0, 1, \ldots, n\}$. A measurable space is denoted by $\{(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{Z}\}$, where $\mathcal{X}_n, n \in \mathbb{Z}$ are the alphabet spaces, and these are Polish spaces (complete separable metric spaces), and $\mathcal{B}(\mathcal{X}_n)$ are the Borel $\sigma-$algebras of subsets of $\mathcal{X}_n$. Given a random variable (RV) $X : (\Omega, \mathcal{F}) \longmapsto (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we denote the distribution induced by X on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ by $\mathbf{P}_X(dx) \equiv \mathbf{P}(dx)$. We denote the set of such probability distributions by $\mathcal{M}(\mathcal{X})$. Given another RV $Y : (\Omega, \mathcal{F}) \longmapsto (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, we denote the conditional distribution of RV Y given $X = x$ by $\mathbf{P}_{Y|X}(dy|X = x) \equiv \mathbf{P}_{Y|X}(dy|x)$. We equivalently describe such conditional distributions by stochastic kernels or transition functions[1]. We denote the set of such stochastic kernels by $\mathcal{K}(\mathcal{Y}|\mathcal{X})$. $\langle \cdot, \cdot \rangle$ denotes inner product of elements of linear spaces. For a square matrix $\Sigma \in \mathbb{R}^{d \times d}$, we denote $\Sigma \succ 0$ (respectively, $\Sigma \succeq 0$) a symmetric positive-definite matrix (respectively, positive-semidefinite matrix). The statement $\Sigma \succeq \Sigma'$ means that $\Sigma - \Sigma'$ is positive semidefinite.

We consider the joint sequence of random processes $X^n \triangleq \{X_0, X_1, \ldots, X_n\}$ and $Y^n \triangleq \{Y^{-1}, Y_0, Y_1, \ldots, Y_n\}$ defined on the product measurable space $(\mathcal{X}^n \times \mathcal{Y}^n, \mathcal{B}(\mathcal{X}^n \times \mathcal{Y}^n))$. The process X^n is called the information process while the process $Y_0^n = \{Y_0, \ldots, Y_n\}$ is called its reproduction, and $Y^{-1} = y^{-1}$ is the initial data. These are generated by two sequences of conditional distributions as follows.

[1]For details on stochastic kernels see for example [16].

Information Process. The conditional distribution of information process is defined by

$$\mathcal{P}_{0,n} \triangleq \left\{ \mathbf{P}_{X_t|X^{t-1},Y^{t-1}} = P_t(dx_t|x^{t-1}, y^{t-1}) \in \mathcal{M}(\mathcal{X}_t) : t \in \mathbb{Z}_0^n \right\}.$$

For $t = 0$ we let $\mathbf{P}_{X_0|X^{-1},Y^{-1}} = P_0(dx_0|y^{-1})$.

Reproduction Process. The conditional distribution of reproduction process is defined by

$$\mathcal{Q}_{0,n} \triangleq \left\{ \mathbf{P}_{Y_t|Y^{t-1},X^t} = Q_t(dy_t|y^{t-1}, x^t) \in \mathcal{M}(\mathcal{Y}_t) : t \in \mathbb{Z}_0^n \right\}.$$

For $t = 0$ we let $\mathbf{P}_{Y_0|Y^{-1},X^0} = Q_0(dy_0|y^{-1}, x_0)$, where y^{-1} is the initial data, and $\mathbf{P}_{Y^{-1}}(dy^{-1}) = \mu(dy^{-1})$ is fixed.

We define the directed information [12] from X^n to Y_0^n conditional on $Y^{-1} = y^{-1}$ as follows.

$$I(X^n \to Y_0^n|y^{-1}) = \mathbf{E}^Q_{y^{-1}}\Big\{ \sum_{t=0}^n \log\Big(\frac{Q_t(\cdot|Y^{t-1}, X^t)}{\Pi_t^Q(\cdot|Y^{t-1})}(Y_t) \Big) \Big\},$$

where the superscript notation $\mathbf{E}^Q_{y^{-1}}$ indicates the dependence of the joint distribution on $Q \equiv \{Q_0(\cdot|\cdot,\cdot), \ldots, Q_n(\cdot|\cdot,\cdot)\}$. The conditional distribution of $\{Y_t : t \in \mathbb{Z}_0^n\}$ is defined by

$$\Pi_t^Q(dy_t|y^{t-1}) \triangleq \int_{\mathcal{X}^t} Q_t(dy_t|y^{t-1}, x^t) \otimes P_t(dx_t|x^{t-1}, y^{t-1}) \otimes \mathbf{P}^Q(dx^{t-1}|y^{t-1}),\ t \in \mathbb{Z}_0^n. \quad \text{(II.1)}$$

Given a measurable distortion function $d : \mathcal{X}^n \times \mathcal{Y}_0^n \longmapsto [0, \infty]$ of reproducing x^n by y_0^n, we define the fidelity of such reproductions as follows.

$$\mathcal{Q}_{0,n}(D) \triangleq \Big\{ Q_t(\cdot|y^{t-1}, x^t) \in \mathcal{M}(\mathcal{Y}_t), t \in \mathbb{Z}_0^n : \frac{1}{n+1}\mathbf{E}^Q_{y^{-1}}\left(d(X^n, Y_0^n)\right) \le D \Big\},\ D \ge 0. \quad \text{(II.2)}$$

Next, we give the definition of the extremum problem of directed information that we refer to as *causal* RDF.

Definition 1: (Causal RDF)
The extremum problem of directed information subject to a fidelity is defined by

$$R_{0,n}^C(D) \triangleq \inf_{\mathcal{Q}_{0,n}(D)} \mathbf{E}^Q_{y^{-1}}\Big\{ \sum_{t=0}^n \log\Big(\frac{Q_t(\cdot|Y^{t-1}, X^t)}{\Pi_t^Q(\cdot|Y^{t-1})}(Y_t) \Big) \Big\}. \quad \text{(II.3)}$$

We note that the above extremum problem is a convex optimization problem [17, Theorems 5,6]. Sufficient conditions for existence of an optimal solution are provided in [17, Lemma 12, Theorem 14].

In the next theorem, we generalize well-known results of mutual information (see e.g., [18, Theorem 1.8.6]) to directed information.

Theorem 1: Suppose the following holds. For a fixed $Y^{-1} = y^{-1}$,

(A1) $X^{g,n} \triangleq \{X_0^g, X_1^g, \ldots, X_n^g\}$, $X_t^g \in \mathbb{R}^k, t = 0, \ldots, n$ is jointly Gaussian;

(A2) $Y_0^n \triangleq \{Y_0, \ldots, Y_n\}$ and $Y_0^{g,n} \triangleq \{Y_0^g, Y_1^g, \ldots, Y_n^g\}$ are RVs such that $(X^{g,n}, Y_0^n) \triangleq$

$\{X_0^g, X_1^g, \ldots, X_n^g, Y_0, Y_1, \ldots, Y_n\}$ and $(X^{g,n}, Y_0^{g,n}) \triangleq \{X_0^g, X_1^g, \ldots, X_n^g, Y_0^g, Y_1^g, \ldots, Y_n^g\}$, $(Y_t, Y_t^g) \in \mathbb{R}^{p+p}, t = 0, \ldots, n$ are continuous RVs, the covariance of $(X^{g,n}, Y_0^n)$ and $(X^{g,n}, Y_0^{g,n})$ is $\Gamma \triangleq \{\gamma_{i,j}\}$ and that $(X^{g,n}, Y_0^{g,n})$ is Gaussian.

Then for a fixed $Y^{-1} = y^{-1}$ the following inequality holds.

$$I(X^{g,n} \to Y^{g,n}|y^{-1}) \leq I(X^{g,n} \to Y_0^n|y^{-1}). \quad \text{(II.4)}$$

Proof: The derivation employs the variational equality of directed information given in [17, Theorem 19, Part B] to the left and right hand side of (II.4). By Gaussianity assumptions (A1), (A2) the result follows. ∎

III. Sequential Characterization of Optimal Reproduction Dsitribution and Information Structures

In this section, we describe the form of the optimal minimizer that achieves the infimum in (II.3), provided this is attained. Moreover, we state certain structural results regarding $R_{0,n}^C(D)$.

First, we state certain properties regarding the convexity and continuity of $R_{0,n}^C(D)$.

1) $R_{0,n}^C(D)$ is a convex, non-increasing function of $D \in [0, \infty)$.

2) If $R_{0,n}^C(D) < \infty$, then $R_{0,n}^C(D)$ is continuous on $[0, \infty)$.

Note that **1)** is similar to the classical RDF. Also, for **2)** recall that a bounded and convex function is continuous. Since $R_{0,n}^C(D)$ is non-increasing, it is bounded outside the neighbourhood of $D = 0$ and it is also continuous on $(0, \infty)$. In other words, if $R_{0,n}^C(D) < \infty$, then, $R_{0,n}^C(D)$ is bounded and hence continuous on $[0, \infty)$.

Since (II.3) is a convex optimization problem, then, assuming existence of an interior point to the fidelity constraint, it can be reformulated using Lagrange duality theorem [19, Theorem 1, pp. 224-225] as an unconstrained problem. Next, we give the backward recursions of the optimal reproduction distribution of (II.3). We remark that the following recursions were first announced in [13].

Stage $t = n$**:** $Q_n^*(dy_n|y^{n-1}, x^n)$ is given by

$$Q_n^*(dy_n|y^{n-1}, x^n) = \frac{e^{s\rho_n(T^n x^n, T^n y^n)} \Pi_n^{Q^*}(dy_n|y^{n-1})}{\int_{\mathcal{Y}_n} e^{s\rho_n(T^n x^n, T^n y^n)} \Pi_n^{Q^*}(dy_n|y^{n-1})}, \quad \text{(III.1)}$$

where

$$\Pi_n^{Q^*}(dy_n|y^{n-1}) \triangleq \int_{\mathcal{X}^n} Q_n^*(dy_n|y^{n-1}, x^n) \otimes P_n(dx_n|x^{n-1}, y^{n-1}) \otimes \mathbf{P}^{Q^*}(dx^{n-1}|dy^{n-1}) \quad \text{(III.2)}$$

and $s \in (-\infty, 0]$ is the Lagrange multiplier of the average distortion.

Stages $t \in \{n-1, \ldots, 0\}$**:** $Q_t^*(dy_t|y^{t-1}, x^t)$ is given by

$$Q_t^*(dy_t|y^{t-1}, x^t) = \frac{e^{s\rho_t(T^t x^n, T^t y^n) - G_{t,t+1}^{Q^*}(x^t, y^t)} \Pi_t^{Q^*}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t(T^t x^n, T^t y^n) - G_{t,t+1}^{Q^*}(x^t, y^t)} \Pi_t^{Q^*}(dy_t|y^{t-1})}, \quad \text{(III.3)}$$

where $\Pi_t^{Q^*}(dy_t|y^{t-1})$ is given by (III.2) with n replaced by t, and $G_{t,t+1}^{Q^*}(x^t, y^t)$ is given by

$$G_{n,n+1}^{Q^*}(x^n, y^n) = 0, \quad \text{(III.4a)}$$

$$G_{t,t+1}^{Q^*}(x^t, y^t) = -\int_{\mathcal{X}_{t+1}} P_{t+1}(dx_{t+1}|x^t, y^t) \log\left(\int_{\mathcal{Y}_{t+1}} e^{s\rho_{t+1}(T^{t+1}x^n, T^{t+1}y^n) - G_{t+1,t+2}^{Q^*}(x^{t+1}, y^{t+1})} \Pi_{t+1}^{Q^*}(dy_{t+1}|y^t) \right), \ t = n-1, \ldots, t = 0. \quad \text{(III.4b)}$$

Next, we apply the above recursions to identify the information structure of $Q_t^*(dy_t|y^{t-1}, x^t)$, i.e., its dependence on x^t, for $t \in \mathbb{Z}_0^n$.

a) Markov Information Process and Single Letter Distortion with Respect to $\{X_t : t \in \mathbb{Z}_0^n\}$*:* Consider the following case:

$$P_t(dx_t|x^{t-1}, y^{t-1}) = P_t^1(dx_t|x_{t-1}, y^{t-1}), \quad \text{(III.5)}$$

$$\rho_t(T^t x^n, T^t, y^n) = \rho_t^0(x_t, y^t), \ t \in \mathbb{Z}_0^n. \quad \text{(III.6)}$$

By the above recursions we deduce that

$$Q_t^*(dy_t|y^{t-1}, x^t) = Q_t^o(dy_t|y^{t-1}, x_t), \ t \in \mathbb{Z}_0^n, \quad \text{(III.7)}$$

i.e., the optimal reproduction distribution is Markov with respect to $\{X_t : t = 0, \ t \in \mathbb{Z}_0^n\}$. Hence, the probability distribution on $\{Y_t : \ t \in \mathbb{Z}_0^n\}$ is given by

$$\Pi_t^{Q^o}(dy_t|y^{t-1}) = \int_{\mathcal{X}_t} Q_t^o(dy_t|y^{t-1}, x_t) \otimes \mathbf{P}^{Q^o}(dx_t|y^{t-1}).$$

b) Limited Memory Markov Information Process and Distortion Function: Consider the following case:

$$P_t(dx_t|x^{t-1}, y^{t-1}) = P_t^{M,L}(dx_t|x_{t-L}^{t-1}, y_{t-M}^{t-1},), \quad \text{(III.8)}$$

$$\rho_t(T^t x^n, T^t, y^n) = \rho_t^{N,L}(x_{t-N}^t, y_{t-K}^t), \ t \in \mathbb{Z}_0^n, \quad \text{(III.9)}$$

where $\{M, L, N, K\}$ are finite non-negative integers. Then, similar to a), from the above recursions we can obtain

$$Q_t(dy_t|y^{t-1}, x^t) = Q_t^J(dy_t|y^{t-1}, x_{t-J}^t), \ J = \max\{N, L\}.$$

and $\Pi_t^{Q^J}(dy_t|y^{t-1}) = \int_{\mathcal{X}_{t-J}^t} Q_t(dy_t|y^{t-1}, x_{t-J}^t) \otimes \mathbf{P}^Q(dx_{t-J}^t|y^{t-1})$.

c) Assume $G_{t,t+1}^Q(x^t, y^t) = \hat{G}_{t,t+1}^Q(x^t, y^{t-1}), \ \forall t \in \{0, \ldots, n\}$*:* Then, the optimal reproduction distribution reduces to

$$Q_t^1(dy_t|y^{t-1}, x^t) = \frac{e^{s\rho_t(T^t x^n, T^t y^n)} \Pi_t^{Q^1}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t(T^t x^n, T^t y^n)} \Pi_t^{Q^1}(dy_t|y^{t-1})}.$$

Unfortunately, no further reduction of the information structure can be identified from the above recursions. Such reduction, can be identified via the variational characterizations of directed information [17, Theorem 19].

IV. Conditionally Gaussian State Space Models

In this section we derive the optimal reproduction distribution of conditionally Gaussian information processes $\{X_t : t \in \mathbb{Z}_0^n\}$ where the distortion function is the weighted squared-error (WSE).

Consider a distribution satisfying $P_t(dx_t|x^{t-1}, y^{t-1}) = P_t(dx_t|x_{t-1}, y^{t-1}),\ t \in \mathbb{Z}_0^n$. The distribution is called *conditionally Gaussian* if it is induced by the conditionally Gaussian state space model (CGSSM)

$$X_{t+1} = f_t(Y^t) + A_t X_t + G_t W_t,\ X_0 = x_0 \in \mathbb{R}^d,\ t \in \mathbb{Z}_0^{n-1}, \tag{IV.1}$$

where $\{(A_t, G_t) : t \in \mathbb{Z}_0^n\}$ are deterministic matrices and W_t is an $\mathbb{R}^k-$dimensional independence Gaussian noise process with $\{W_t \sim \mathcal{N}(0; K_{W_t}) :\ t \in \mathbb{Z}_0^n\}$, and the following holds.

$$\mathbf{P}_{X_{t+1}|X^t,Y^t} = \mathbf{P}\Big\{w_t : f_t(y^t) + A_t x_t + G_t w_t \in dx_{t+1}\Big\} \tag{IV.2}$$

A distortion function $\rho_t(T^t x^n, T^t, y^n) = \rho_t^{\mathrm{M}}(x_t, y_t),\ t \in \mathbb{Z}_0^n$ is called WSE if

$$\rho_t(T^t x^t, T^t y^n) \triangleq \langle x_t - y_t, M_t(x_t - y_t)\rangle \equiv ||x_t - y_t||^2_{M_t},$$
$$M_t = M_t^{\mathrm{T}},\ M_t \succ 0,\ \ t \in \mathbb{Z}_0^n,$$

and *squared-error (SE)* if $M_t = I$ (identity matrix), $t \in \mathbb{Z}_0^n$.

For a CGSSM and a WSE distortion function, the optimal reproduction distribution is of the form $\{Q_t^o(dy_t|y^{t-1}, x_t) : t \in \mathbb{Z}_0^n\}$, which implies

$$R_{0,n}^C(D) = \inf_{\substack{\sum_{t=0}^n \mathbf{E}_{y^{-1}}\{||X_t - Y_t||^2_{M_t}\} \\ \le D(n+1)}} \sum_{t=0}^n I_{y^{-1}}(X_t; Y_t|Y^{t-1}), \tag{IV.3}$$

where the subscript $I_{y^{-1}}(\cdot;\cdot|\cdot)$ means $Y^{-1} = y^{-1}$ is fixed. Before we proceed to state the main theorem of this section, we define the filter estimates and conditional covariances.

$$\widehat{X}_{t|t-1} \triangleq \mathbf{E}_{y^{-1}}\left\{X_t \middle| Y^{t-1}\right\},\ \widehat{X}_{t|t} \triangleq \mathbf{E}_{y^{-1}}\left\{X_t \middle| Y^{t}\right\},$$
$$\Sigma_{t|t-1} \triangleq \mathbf{E}_{y^{-1}}\left\{\left(X_t - \widehat{X}_{t|t-1}\right)\left(X_t - \widehat{X}_{t|t-1}\right)^{\mathrm{T}} \middle| Y^{t-1}\right\},$$
$$\Sigma_{t|t} \triangleq \mathbf{E}_{y^{-1}}\left\{\left(X_t - \widehat{X}_{t|t}\right)\left(X_t - \widehat{X}_{t|t}\right)^{\mathrm{T}} \middle| Y^{t}\right\},\ t \in \mathbb{Z}_0^n.$$

The next theorem states that the optimal reproduction conditional distribution is conditionally Gaussian, and it is related to generalized Kalman filter equations.

Theorem 2: Consider the CGSSM, i.e., (IV.1) and a WSE distortion function. Then the following hold.
(a) Any candidate of optimal reproduction conditional distribution is conditionally Gaussian, denoted by $Q_t^{\mathrm{CG}}(dy_t|y^{t-1}, x_t) : t \in \mathbb{Z}_0^n\}$, and it is induced by the following recursion.

$$Y_t = H_t\left(X_t - \widehat{X}_{t|t-1}\right) + \widehat{X}_{t|t-1} + V_t, \tag{IV.4}$$
$$= f_{t-1}(Y^{t-1}) + H_t\left(X_t - \widehat{X}_{t|t-1}\right) + A_{t-1}\widehat{X}_{t-1|t-1} + V_t$$
$$= \widehat{X}_{t|t-1} + \nu_t,\ \ Y^{-1} = y^{-1}\ t \in \mathbb{Z}_0^n,$$

where
(i) $\{V_t \sim \mathcal{N}(0; K_{V_t}) : t \in \mathbb{Z}_0^n\}$ is an independent Gaussian process, which is independent of $\left(\{W_t : t \in \mathbb{Z}_0^n\}, X_0, Y^{-1}\right)$,
(ii) $\{H_t : t \in \mathbb{Z}_0^n\}$ are deterministic matrices,
(iii) $\{\nu_t : t \in \mathbb{Z}_0^n\}$ is the orthogonal innovations process

$$\nu_t \triangleq Y_t - \mathbf{E}_{y^{-1}}\left\{Y_t \middle| Y^{t-1}\right\},\ t \in \mathbb{Z}_0^n$$
$$= H_t\left(X_t - \widehat{X}_{t|t-1}\right) + V_t,$$
$$\mathbf{E}_{y^{-1}}\left\{\nu_t \middle| Y^{t-1}\right\} = \mathbf{E}_{y^{-1}}\{\nu_t\} = 0,$$
$$\mathbf{E}_{y^{-1}}\left\{\nu_t \nu_t^{\mathrm{T}} \middle| Y^{t-1}\right\} = H_t \Sigma_{t|t-1} H_t^{\mathrm{T}} + K_{V_t} = \mathbf{E}_{y^{-1}}\{\nu_t \nu_t^{\mathrm{T}}\}.$$

(b) $\{\widehat{X}_{t|t-1}, \widehat{X}_{t|t} : t \in \mathbb{Z}_0^n\}$ are the solutions to the discrete-time filtering recursive equations:

$$\widehat{X}_{t+1|t} = f_t(Y^t) + A_t\widehat{X}_{t|t-1} + \Psi_{t|t-1}\left(Y_t - \widehat{X}_{t|t-1}\right)$$
$$= f_t(Y^t) + A_t\widehat{X}_{t|t-1} + \Psi_{t|t-1}\nu_t, \quad \widehat{X}_{0|-1},$$
$$\widehat{X}_{t+1|t+1} = f_t(Y^t) + A_t\widehat{X}_{t|t} + \overline{\Psi}_{t+1|t}\nu_{t+1}, \tag{IV.5}$$

$$\Sigma_{t+1|t} = A_t\Sigma_{t|t-1}A_t^{\mathrm{T}} + G_t K_{W_t} G_t^{\mathrm{T}} - A_t\Sigma_{t|t-1}H_t^{\mathrm{T}}$$
$$\left[H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + K_{V_t}\right]^{-1}\left(A_t\Sigma_{t|t-1}H_t^{\mathrm{T}}\right)^{\mathrm{T}},\ \Sigma_{0|-1}, \tag{IV.6}$$

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1}H_t^{\mathrm{T}}\left[H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + K_{V_t}\right]^{-1}$$
$$\left(\Sigma_{t|t-1}H_t^{\mathrm{T}}\right)^{\mathrm{T}},\ \Sigma_{0|0}, \tag{IV.7}$$

$$\Psi_{t|t-1} \triangleq A_t\overline{\Psi}_{t|t-1}, \tag{IV.8}$$
$$\overline{\Psi}_{t|t-1} \triangleq \Sigma_{t|t-1}H_t^{\mathrm{T}}\left[H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + K_{V_t}\right]^{-1}.$$

(c) $R_{0,n}^C(D)$ is the solution of the matrix optimization problem:

$$R_{0,n}^C(D) = \inf_{\mathcal{Q}_{[0,n]}^{H,K_V}(D)} \sum_{t=0}^n I_{y^{-1}}(X_t; Y_t|Y^{t-1}) \tag{IV.9}$$

where

$$\sum_{t=0}^n I_{y^{-1}}(X_t; Y_t|Y^{t-1}) \tag{IV.10}$$
$$= \sum_{i=0}^n \Big\{H_{ent}(\nu_t|Y^{-1} = y^{-1}) - H_{ent}(V_t)\Big\} \tag{IV.11}$$
$$= \frac{1}{2}\sum_{i=0}^n \log\frac{|H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + K_{V_t}|}{|K_{V_t}|} \tag{IV.12}$$
$$= \sum_{i=0}^n \Big\{H_{ent}(X_t|Y_0^{t-1}, Y^{-1} = y^{-1})$$
$$- H_{ent}(X_t|Y_0^t, Y^{-1} = y^{-1})\Big\} \tag{IV.13}$$
$$= \frac{1}{2}\sum_{i=0}^n \log\frac{|\Sigma_{t|t-1}|}{|\Sigma_{t|t}|}, \tag{IV.14}$$

$H_{ent}(\cdot)$ denotes differential entropy, and the average distortion constraint is

$$\mathcal{Q}^{H,K_V}_{[0,n]}(D) \triangleq \Bigg\{ H_t \in \mathbb{R}^{d\times d},\ K_{V_t} \in \mathbb{R}^{d\times d} \succeq 0, t \in \mathbb{Z}_0^n : \mathbf{E}_{y^{-1}}\Big\{ \sum_{t=0}^{n} ||X_t - Y_t||^2_{M_t} \Big\} = \sum_{t=0}^{n} \mathrm{trace}\Big((I-H_t)\,\Sigma_{t|t-1} (I-H_t)^{\mathrm{T}} M_t + K_{V_t} M_t \Big) \le (n+1)D \Bigg\}. \tag{IV.15}$$

Proof: We outline the proof. (a) The fact that the optimal reproduction is conditionally Gaussian follows from the recursions. Moreover, by properties of conditional mutual information we can substract $f_{t-1}(y^{t-1})$ from X_t and Y_t without affecting the optimization problem, and hence by Theorem 1 and (III.7), an reproduction process that induces such a conditionally Gaussian distribution is

$$Y_t = H_t X_t + g_t(Y^{t-1}) + V_t, \forall t \in \mathbb{N}_0^n, g_t(Y^{t-1}) = \Gamma_t Y^{t-1}, \tag{IV.16}$$

where $\{V_t : t \in \mathbb{Z}_0^n\}$ is independent Gaussian, and $\{H_t : t \in \mathbb{Z}_0^n\}$ is deterministic. Moreover, $\{W_t : t \in \mathbb{Z}_0^n\}$ is an independent process, and $\{V_t : t \in \mathbb{Z}_0^n\}$ is independent of $\{W_t : t \in \mathbb{Z}_0^n\}$, X_0, and Y^{-1}. From (IV.16) then $\sum_{t=0}^n I_{y^{-1}}(X_t; Y_t|Y^{t-1})$ does not depend on $g_t(\cdot)$, $\forall t \in \mathbb{Z}_0^n$. Since

$$\begin{aligned}\mathbf{E}_{y^{-1}}\left\{ \sum_{t=0}^{n} ||X_t - Y_t||^2_{\mathrm{M}_t} \right\} &= \sum_{t=0}^{n} \mathrm{trace}(K_{V_t} M_t), \\ &= \mathbf{E}_{y^{-1}}\left\{ \sum_{t=0}^{n} ||(1-H_t)X_t - g_t(Y^{t-1})||^2_{\mathrm{M}_t} \right\},\end{aligned} \tag{IV.17}$$

then, by mean-square estimation theory, a smaller average distortion occurs when $g_t(Y^{t-1}) = (1-H_t)\widehat{X}_{t|t-1}$, $\forall t \in \mathbb{Z}_0^n$. The constraint occurs on the boundary, due to strict convexity and decreasing property of $R^C_{0,n}(D)$. This completes the derivation of (IV.4), and the items in (a). (b) The equations under (b) follow from a derivation similar to the Kalman filter [20]. (c) All equations under (c) are obtained from (a) and (b). ∎

Theorem 3: (Characterization of causal RDF)
Consider the CGSSM, i.e., (IV.1) and a WSE distortion function. Then the following hold.
(a) Define (H_t, K_{V_t}) by

$$H_t \triangleq I - \Sigma_{t|t}\Sigma^{-1}_{t|t-1} \succeq 0, \quad \Sigma_{t|t} \succeq 0,\ \Sigma_{t|t-1} \succeq 0, \tag{IV.18}$$

$$K_{V_t} \triangleq \Sigma_{t|t} H_t^{\mathrm{T}} \succeq 0, \quad t = 0, \ldots, n. \tag{IV.19}$$

Then, K_{V_t} is a covariance matrix, i.e., $K_{V_t} = K^{\mathrm{T}}_{V_t}$ that necessarily implies $\Sigma_{t|t}\Sigma_{t|t-1} = \Sigma_{t|t-1}\Sigma_{t|t}$, i.e, they commute and $\Sigma_{t|t}H_t^{\mathrm{T}} = H_t^{\mathrm{T}}\Sigma_{t|t}$. Further, $(\widehat{X}_{t|t}, \widehat{X}_{t|t-1})$ and the reproduction process Y_t given in Theorem 2, satisfy the following equations:

$$\widehat{X}_{t|t} = Y_t, \quad \widehat{X}_{t|t-1} = f_{t-1}(Y^{t-1}) + A_{t-1}Y_{t-1} \tag{IV.20}$$

$$Y_t = f_{t-1}(Y^{t-1}) + H_t\Big(X_t - A_{t-1}Y_{t-1}\Big) + A_{t-1}Y_{t-1} + V_t. \tag{IV.21}$$

and $(H_t, K_{V_t}, \Sigma_{t|t}, \Sigma_{t|t-1})$ are simultaneously diagonalizable matrices, i.e., they have the same eigenvectors.
(b) The characterization of the causal RDF is

$$R^C_{0,n}(D) = \frac{1}{2}\inf_{\Sigma_{t|t}\succeq 0} \frac{1}{n+1}\sum_{t=0}^{n}\left[\log\frac{|\Sigma_{t|t-1}|}{|\Sigma_{t|t}|}\right]^+, \tag{IV.22a}$$

$$\text{s.t. } 0 \preceq \Sigma_{t|t} \preceq \Sigma_{t|t-1} \tag{IV.22b}$$

$$\Sigma_{t|t-1} = A_{t-1}\Sigma_{t-1|t-1}A^{\mathrm{T}}_{t-1} + G_{t-1}K_{W_{t-1}}G^{\mathrm{T}}_{t-1},\ \Sigma_{0|-1} \tag{IV.22c}$$

$$\frac{1}{n+1}\sum_{t=0}^{n}\mathrm{trace}(M_t\Sigma_{t|t}) \le D \tag{IV.22d}$$

where $[x]^+ = \{0, x\}$ for $D \in [0, \infty)$. Equivalently, the choice (H_t, K_{V_t}) defined by (IV.18) and (IV.19) gives the achievable lower bound (IV.22a) on $R^C_{0,n}(D)$ defined (IV.9).

Proof: (a) The first part follows from the fact that if A, B are two symmetric matrices then AB is symmetric if and only if A, B commute, i.e, $AB = BA$. For the rest, we apply Theorem 2. (a) This follows directly from the Kalman filter equations of Theorem 2, by substituting (IV.18) and (IV.19). (b) By mean-squared estimation theory, then the following inequality holds:

$$\begin{aligned}(n+1)D &\ge \mathbf{E}_{y^{-1}}\left\{ \sum_{t=0}^{n} ||X_t - Y_t||^2_{\mathrm{M}_t} \right\} \\ &\ge \mathbf{E}_{y^{-1}}\left\{ \sum_{t=0}^{n} ||X_t - \widehat{X}_{t|t}||^2_{\mathrm{M}_t} \right\},\end{aligned} \tag{IV.23}$$

i.e., $\forall (H_t, K_{V_t})$, $t \in \mathbb{Z}_0^n$, where $\widehat{X}_{t|t}$ is given by (IV.5). Now, we turn our attention to the pay-off $\sum_{t=0}^n I_{y^{-1}}(X_t; Y_t|Y^{t-1})$ in (IV.9). By (IV.14) and substituting (IV.7), we obtain

$$\begin{aligned} I_{y^{-1}}(X_t; Y_t|Y^{t-1}) &= \frac{1}{2}\log\frac{|\Sigma_{t|t-1}|}{|\Sigma_{t|t}|} \\ &= -\frac{1}{2}\log|I - H_t^T\Big[H_t\Sigma_{t|t-1}H_t^T + K_{V_t}\Big]^{-1}H_t\Sigma_{t|t-1}| \qquad \text{(IV.24)} \\ &= -\frac{1}{2}\log|I - H_t\Sigma_{t|t-1}H_t^T\Big[H_t\Sigma_{t|t-1}H_t^T + K_{V_t}\Big]^{-1}| \qquad \text{(IV.25)} \\ &\overset{(\alpha)}{=} -\frac{1}{2}\log|I - U_t\Lambda_t U_t^T\Big[U_t\Lambda_t U_t^T + K_{V_t}\Big]^{-1}| \qquad \text{(IV.26)} \\ &= -\frac{1}{2}\log|I - \Lambda_t\Big[\Lambda_t + U_t^T K_{V_t}U_t\Big]^{-1}| \qquad \text{(IV.27)} \\ &\overset{(\beta)}{\ge} -\frac{1}{2}\log\prod_i\Big(1 - \lambda_{t,i}\Big[\lambda_{t,i} + \Big(U_t^T K_{V_t}U_t\Big)_{ii}\Big]^{-1}\Big) \qquad \text{(IV.28)}\end{aligned}$$

where (α) is obtained by using the singular value decomposition $H_t\Sigma_{t|t-1}H_t^T = U_t\Lambda_t U_t^T, \Lambda_t = \mathrm{diag}\{\lambda_{t,1}, \ldots, \lambda_{t,d}\}, U_tU_t^T = I$, (β) is due to Hadamard's

inequality, i.e., for any non-negative square matrix A, with diagonal entries $(A)_{ii}$, then $|A| \leq \prod_i (A)_{ii}$, and equality holds if and only A is diagonal. The choice (IV.18) and (IV.19) achieves the inequality (IV.23), since $\widehat{X}_{t|t} = Y_t$, i.e., by (a). Moreover, the choice (IV.18) and (IV.19) implies $(H_t, K_{V_t}, \Sigma_{t|t}, \Sigma_{t|t-1})$ are simultaneously diagonalizable, i.e., $U_t^T K_{V_t} U_t$ is diagonal, and hence the lower bound (IV.28) is also achieved. A direct consequence of substituting (IV.18) and (IV.19) into (IV.6) is (IV.22c). Also, for the causal RDF we can show that $\mathbf{E}_{y^{-1}} \left\{ \sum_{t=0}^{n} ||X_t - Y_t||^2_{M_t} \right\} = (n+1)D$. Hence, by the choice of (IV.18) and (IV.19) the inequalities (IV.23) and (IV.28) turn into equalities, and, thus, (b) is shown. ■

In the next remark we comment on the optimal solution of the characterization of causal RDF in Theorem 3.

Remark 1: (Optimal Solution of (IV.22))
To the best of our knowledge Theorem 3 generalizes recent results established in [21] for scalar-valued Gaussian processes X^n. It is also a generalization of a similar result obtained in [3] for scalar-valued Gaussian processes (that do not depend on past reproductions), to multidimensional Gaussian processes X^n that depend on past reproductions, i.e., defined by (IV.1).
It should be noted that characterization of Theorem 3, (a), that the that matrices commute, and hence they have the same eigenvectors, is important and should be incorporated in any attempt to derive necessary and sufficient conditions for the optimization problem of Theorem 3, (b), using convex optimization methods. It is also noted that (IV.22) can be formulated and solved numerically, using the semidefinite programming (SDP) approach of [5, Theorem 1] (although the authors did not use the extra knowlege of Theorem 3, (a) to arrive to (IV.22), and hence did not incorporate Theorem 3, (a) in their SDP algorithm).

Finally, we make the following conjecture.

Conjecture 1: Theorem 3 extends to $A_t = A_t(Y^t), G_t = G_t(Y^t), t \in \mathbb{Z}_0^{n-1}$, with appropriate changes.

V. Conclusion

In this paper we analyzed causal RDF, when the information process depends on past reproductions. Further, we derived the characterization of causal RDF for conditionally Gaussian state space models that depends on past reproductions subject a total mean-squared reproduction error. Our framework includes controlled information processes, since we considered conditionally Gaussian processes.

References

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

[3] A. K. Gorbunov and M. S. Pinsker, "Nonanticipatory and prognostic epsilon entropies and message generation rates," *Problems Inf. Transmiss.*, vol. 9, no. 3, pp. 184–191, July-Sept. 1973.

[4] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed, "Nonanticipative rate distortion function and relations to filtering theory," *IEEE Trans. Autom. Control*, vol. 59, no. 4, pp. 937–952, April 2014.

[5] T. Tanaka, K. K. K. Kim, P. A. Parrilo, and S. K. Mitter, "Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1896–1910, April 2017.

[6] M. S. Derpich and J. Østergaard, "Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3131–3152, May 2012.

[7] P. A. Stavrou, J. Østergaard, C. D. Charalambous, and M. S. Derpich, "An upper bound to zero-delay rate distortion via Kalman filtering for vector Gaussian sources," in *Information Theory Workshop*, November 2017, pp. 534–538.

[8] M. Gastpar, "To code or not to code," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne (E.P.F.L), 2002.

[9] C. K. Kourtellaris, C. D. Charalambous, and J. J. Boutros, "Nonanticipative transmission for sources and channels with memory," in *IEEE International Symposium on Information Theory*, June 2015, pp. 521–525.

[10] S. Tatikonda, A. Sahai, and S. Mitter, "Stochastic linear control over a communication channel," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1549–1561, Sept. 2004.

[11] H. Marko, "The bidirectional communication theory–A generalization of information theory," *IEEE Transactions on Communications*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.

[12] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Nov. 27-30 1990, pp. 303–305.

[13] C. D. Charalambous and P. A. Stavrou, "Optimization of directed information and relations to filtering theory," in *European Control Conference (ECC)*, June 2014, pp. 1385–1390.

[14] E. I. Silva, M. S. Derpich, and J. Østergaard, "A framework for control system design subject to average data-rate constraints," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1886–1899, Aug 2011.

[15] E. I. Silva, M. S. Derpich, J. Østergaard, and M. A. Encina, "A characterization of the minimal average data rate that guarantees a given closed-loop performance level," *IEEE Transactions on Automatic Control*, vol. 61, no. 8, pp. 2171–2186, Aug 2016.

[16] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, Inc., New York, 1997.

[17] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: Properties and variational equalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6019 – 6052, 2016.

[18] S. Ihara, *Information theory - for Continuous Systems*. World Scientific, 1993.

[19] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, 1969.

[20] P. E. Caines, *Linear Stochastic Systems*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1988.

[21] P. A. Stavrou, T. Charalambous, and C. D. Charalambous, "Finite-time nonanticipative rate distortion function for time-varying scalar-valued Gauss-Markov sources," *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 175–180, Jan 2018.