2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

On the relation of nonanticipative rate distortion function and filtering theory

Charalambos D. Charalambous and Photios A. Stavrou

***Abstract*—In this paper the relation between nonanticipative rate distortion function (RDF) and Bayesian filtering theory is investigated using the topology of weak convergence of probability measures on Polish spaces. The relation is established via an optimization on the space of conditional distributions of the so-called directed information subject to fidelity constraints. Existence of the optimal reproduction distribution of the nonanticipative RDF is shown, while the optimal nonanticipative reproduction conditional distribution for stationary processes is derived in closed form. The realization procedure of nonanticipative RDF which is equivalent to joint-source channel matching for symbol-by-symbol transmission is described, while an example is introduced to illustrate the concepts.**

I. Introduction

This paper is concerned with the abstract formulation of nonanticipative rate distortion function (RDF) on Polish spaces (complete separable metric spaces) and its relation to filtering theory. In the past, rate distortion (or distortion rate) functions and filtering theory have evolved independently. Specifically, classical RDF addresses the problem of reproduction of a process subject to a fidelity criterion without much emphasis on the realization of the reproduction conditional distribution via nonanticipative operations. On the other hand, filtering theory is developed by imposing real-time realizability on estimators with respect to measurement data.

Historically, the work of R. Bucy [1] appears to be the first to consider the direct relation between distortion rate function and filtering. The work of A. K. Gorbunov and M. S. Pinsker [2] on ϵ-entropy defined via a nonanticipative constraint on the reproduction distribution of the RDF, although not directly related to the realizability question pursued by Bucy, computes the nonanticipative RDF for stationary Gaussian processes via power spectral densities. The objective of this paper is to investigate the connection between nonanticipative RDF and filtering theory for general distortion functions and random processes on abstract Polish spaces using the topology of weak convergence.

The main results discussed in this paper are the following.
(1) Existence of optimal reproduction distribution minimizing directed information using the topology of weak convergence of probability measures on Polish spaces;
(2) Closed form expression of the optimal reproduction conditional distribution for stationary processes;
(3) Realization procedure of the filter;
(4) Example to demonstrate the realization of the filter;
(5) Connection between nonanticipative RDF and joint source-channel coding of symbol-by-symbol transmission [3].

Motivation. This work is motivated by applications in which estimators are desired to have specific accuracy, by control over limited rate communication channel applications [4], [5], and by the desire to provide necessary conditions for symbol-by-symbol or uncoded transmission [3] for sources with memory without anticipation.

First, we give a brief high level discussion on nonanticipative RDF and filtering theory, and discuss their connection. Consider a discrete-time process $X^n \stackrel{\triangle}{=} \{X_0, X_1, \ldots, X_n\} \in \mathcal{X}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathcal{X}_i$, and its reproduction $Y^n \stackrel{\triangle}{=} \{Y_0, Y_1, \ldots, Y_n\} \in \mathcal{Y}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathcal{Y}_i$ where $\mathcal{X}_i$ and $\mathcal{Y}_i$ are Polish spaces.

Bayesian Estimation Theory. In classical filtering, one is given a mathematical model that generates the process X^n, $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \ldots, n\}$, a mathematical model that generates observed data obtained from sensors, say, Z^n, $\{P_{Z_i|Z^{i-1},X^i}(dz_i|z^{i-1}, x^i) : i = 0, 1, \ldots, n\}$, while Y^n are the causal estimates of some function of the process X^n based on the observed data Z^n. The classical Kalman Filter is a well-known example, where $\widehat{X}_i = \mathbb{E}[X_i|Z^{i-1}]$, $i = 0, 1, \ldots, n$, is the conditional mean which minimizes the average least-squares estimation error. Fig. 1 is the block diagram of the filtering problem.

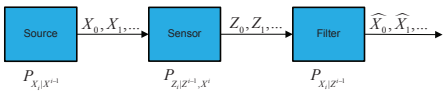

Fig. 1. Filtering problem.

Nonanticipative Rate Distortion Theory and Estimation. In nonanticipative rate distortion theory one is given a distribution for the process X^n, which induces $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \ldots, n\}$, and determines the nonanticipative reproduction conditional distribution $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \ldots, n\}$ which minimizes the directed information from X^n to Y^n subject to distortion or fidelity constraint. The filter $\{Y_i : i = 0, 1, \ldots, n\}$ of $\{X_i : i = 0, 1, \ldots, n\}$ is found by realizing the optimal reproduction distribution

*This work was financially supported by a medium size University of Cyprus grant entitled "DIMITRIS".

The authors are with the Department of Electrical and Computer Engineering (ECE), University of Cyprus, Nicosia, CYPRUS chadcha@ucy.ac.cy, stavrou.fotios@ucy.ac.cy.

$\{P_{Y_i|X^{i-1},X^i}(dy_i|y^{i-1},x^i) : i = 0,1,\ldots,n\}$ via a cascade of sub-systems as shown in Fig. 2. Thus, in nonanticipative rate distortion theory the observation or mapping from $\{X_i : i = 0,1,\ldots,n\}$ to $\{Z_i : i = 0,1,\ldots,n\}$ is part of the realization procedure, while in filtering theory, this mapping is given á priori.

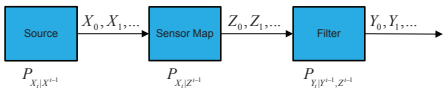

Fig. 2. Filtering via nonanticipative rate distortion function.

The precise problem formulation necessitates the definitions of distortion function or fidelity, and directed information.

The distortion function or fidelity constraint [6] between x^n and its reproduction y^n, is a measurable function $d_{0,n} : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \to [0,\infty]$ defined by

$$d_{0,n}(x^n,y^n) \stackrel{\triangle}{=} \frac{1}{n+1}\sum_{i=0}^{n} \rho_{0,i}(x^i,y^i).$$

Directed information from a sequence of Random Variables (RV's) $X^n \stackrel{\triangle}{=} \{X_0,X_1,\ldots,X_n\} \in \mathcal{X}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^{n}\mathcal{X}_i$, to another sequence $Y^n \stackrel{\triangle}{=} \{Y_0,Y_1,\ldots,Y_n\} \in \mathcal{Y}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^{n}\mathcal{Y}_i$ is often defined via [7], [8][4]

$$I(X^n \to Y^n) \stackrel{\triangle}{=} \sum_{i=0}^{n} I(X^i;Y_i|Y^{i-1})$$
$$= \sum_{i=0}^{n}\int \log\Big(\frac{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)}{P_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}\Big)P_{X^i,Y^i}(dx^i,dy^i)$$
$$\equiv \mathbb{I}_{X^n\to Y^n}(P_{X_i|X^{i-1},Y^{i-1}}, P_{Y_i|Y^{i-1},X^i} : i = 0,1,\ldots,n).$$

In this paper, it is assumed that $\forall\ i = 0,1,\ldots,n$,

$$P_{X_i|X^{i-1},Y^{i-1}}(dx_i|x^{i-1},y^{i-1}) = P_{X_i|X^{i-1}}(dx_i|x^{i-1}).$$

The above assumption states that the process $\{X_i : i = 0,1,\ldots,n\}$ is conditionally independent of $Y^{i-1} = y^{i-1}$ given knowledge of $X^{i-1} = x^{i-1}$, and it is implied by the following conditional independence, $P_{Y_i|Y^{i-1},X^\infty}(dy_i|y^{i-1},x^\infty) = P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) - a.s.,\ \forall\ i = 0,1,\ldots,n$. The last assumption implies that the reproduction of Y_i does not depend on future values $X_{i+1}^{\infty} \stackrel{\triangle}{=} \{X_{i+1},X_{i+2},\ldots,X_\infty\}$. Given a sequence of source distributions $\{P_{X_i|X^{i-1}}(\cdot|\cdot) : i = 0,1,\ldots,n\}$ and a sequence of reproduction conditional distributions $\{P_{Y_i|Y^{i-1},X^i}(\cdot|\cdot,\cdot) : i = 0,1,\ldots,n\}$ define the joint distribution $P_{X^n,Y^n}(dx^n,dy^n) = P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \otimes P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)$. The nonanticipative RDF is a special case of directed information defined by

$$I_{P_{X^n}}(X^n \to Y^n)$$
$$= \mathbb{I}_{X^n\to Y^n}(P_{X_i|X^{i-1}}, P_{Y_i|Y^{i-1},X^i} : i = 0,1,\ldots,n).$$

Nonanticipative RDF. The nonanticipative RDF is defined by

$$R_{0,n}^{na}(D) \stackrel{\triangle}{=} \inf_{\substack{P_{Y_i|Y^{i-1},X^i}(\cdot|\cdot,\cdot),\\ i=0,1,\ldots,n:\\ \mathbb{E}\{d_{0,n}(X^n,Y^n)\le D\}}} I_{P_{X^n}}(X^n \to Y^n). \quad (1)$$

The definition of the nonanticipative RDF is consistent with [9] in which nonanticipation is defined via the Markov chain (MC) $X_{n+1}^{\infty} \leftrightarrow X^n \leftrightarrow Y^n$, e.g., $P_{Y^n|X^\infty}(dy^n|x^\infty) = P_{Y^n|X^n}(dy^n|x^n)$. Therefore, by finding the solution of (1), then one can realize it via a channel from which one can construct an optimal filter via nonanticipative operations as in Fig. 2. One can view the sensor map as consisting of an encoder and a channel, thus draw relations to symbol-by-symbol and uncoded transmission in information theory [3].

This paper is organized as follows. Section II discusses the formulation on abstract spaces. Section III establishes existence of optimal minimizing distribution, and Section IV derives the optimal minimizing distribution for stationary processes. Section V describes the realization of nonanticipative RDF, while Section VI provides an example.

II. Abstract Formulation

The source and reproduction alphabets are sequences of Polish spaces [10]. Probability distributions on any measurable space $(\mathcal{Z},\mathcal{B}(\mathcal{Z}))$ are denoted by $\mathcal{M}_1(\mathcal{Z})$. For $(\mathcal{X},\mathcal{B}(\mathcal{X})),(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ measurable spaces, the set of conditional distributions $P_{Y|X}(\cdot|X=x)$ is denoted by $\mathcal{Q}(\mathcal{Y};\mathcal{X})$ and these are equivalent to stochastic kernels on $(\mathcal{Y},\mathcal{B}(\mathcal{Y}))$ given $(\mathcal{X},\mathcal{B}(\mathcal{X}))$.

Given the process distributions $P_{X^n}(dx^n)$ and $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) : i = 0,1,\ldots,n\}$ the following probability distributions are defined.

(**P1**): The reproduction conditional probability distribution $\overrightarrow{P}_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n};\mathcal{X}_{0,n})$:

$$\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \stackrel{\triangle}{=} \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i).$$

(**P2**): The joint probability distribution $P_{X^n,Y^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n}\times\mathcal{X}_{0,n})$ for $G_{0,n} \in \mathcal{B}(\mathcal{X}_{0,n}) \times \mathcal{B}(\mathcal{Y}_{0,n})$:

$$P_{X^n,Y^n}(G_{0,n}) \stackrel{\triangle}{=} (P_{X^n} \otimes \overrightarrow{P}_{Y^n|X^n})(G_{0,n})$$
$$= \int \overrightarrow{P}_{Y^n|X^n}(G_{0,n,x^n}|x^n) \otimes P_{X^n}(dx^n)$$

where G_{0,n,x^n} is the x^n-section of $G_{0,n}$ at point x^n defined by $G_{0,n,x^n} \stackrel{\triangle}{=} \{y^n \in \mathcal{Y}_{0,n} : (x^n,y^n) \in G_{0,n}\}$

[4]Unless otherwise, integrals with respect to probability distributions are over the spaces on which these are defined.

and $\otimes$ denotes the convolution.
(**P3**): The marginal distribution $P_{Y^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$:

$$P_{Y^n}(F_{0,n}) \stackrel{\triangle}{=} P(\mathcal{X}_{0,n} \times F_{0,n}),\ F_{0,n} \in \mathcal{B}(\mathcal{Y}_{0,n}) = \int \overrightarrow{P}_{Y^n|X^n}(F_{0,n}|x^n) P_{X^n}(dx^n).$$

Define

$$\overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) = \Big\{ \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \in \mathcal{Q}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) : \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \stackrel{\triangle}{=} \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) \Big\}.$$

Directed information (special case) is defined via the Kullback-Leibler distance:

$$\begin{aligned} I_{P_{X^n}}(X^n \to Y^n) &\stackrel{\triangle}{=} \mathbb{D}(P_{X^n,Y^n} || P_{X^n} \times P_{Y^n}) \\ &= \mathbb{D}(P_{X^n} \otimes \overrightarrow{P}_{Y^n|X^n} || P_{X^n} \times P_{Y^n}) \\ &= \int \log \Big(\frac{d(P_{X^n} \otimes \overrightarrow{P}_{Y^n|X^n})}{d(P_{X^n} \times P_{Y^n})} \Big) d(P_{X^n} \otimes \overrightarrow{P}_{Y^n|X^n}) \\ &\equiv \mathbb{I}_{X^n \to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}). \end{aligned} \tag{2}$$

Note that (2) states that directed information is expressed as a functional of $\{P_{X^n}, \overrightarrow{P}_{Y^n|X^n}\}$.
Next, the definition of nonanticipative RDF is given.

Definition 1: (**Nonanticipative RDF**) Suppose $d_{0,n} \stackrel{\triangle}{=} \sum_{i=0}^{n} \rho_{0,i}(x^i, y^i)$ is measurable, and let $\overrightarrow{\mathcal{Q}}_{0,n}(D)$ (assuming is non-empty) denotes the fidelity set

$$\overrightarrow{\mathcal{Q}}_{0,n}(D) \stackrel{\triangle}{=} \Big\{ \overrightarrow{P}_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n}) :\ \ell_{d_{0,n}}(\overrightarrow{P}_{Y^n|X^n}) \stackrel{\triangle}{=} \int d_{0,n}(x^n, y^n) \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \leq D \Big\} \tag{3}$$

where $D \geq 0$. The nonanticipative RDF is defined by

$$R_{0,n}^{na}(D) \stackrel{\triangle}{=} \inf_{\overrightarrow{P}_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}_{0,n}(D)} \mathbb{I}_{X^n \to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}). \tag{4}$$

Clearly, $R_{0,n}^{na}(D)$ is characterized by minimizing $\mathbb{I}_{X^n \to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ over $\overrightarrow{\mathcal{Q}}_{0,n}(D)$.

III. Existence of Reproduction Distribution

In this section, the existence of the minimizing $(n+1)$-fold convolution of conditional distributions in (4) is established by using the topology of weak convergence of probability measures on Polish spaces. First, we state some properties derived in [8].

Theorem 1: [8] Let $\{\mathcal{X}_n :\ n \in \mathbb{N}\}$ and $\{\mathcal{Y}_n :\ n \in \mathbb{N}\}$ be Polish spaces. Then
(1) The set $\overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is convex.
(2) $\mathbb{I}_{X^n \to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ is a convex functional of $\overrightarrow{P}_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ for a fixed $P_{X^n} \in \mathcal{M}_1(\mathcal{X}_{0,n})$.
(3) The set $\overrightarrow{\mathcal{Q}}_{0,n}(D)$ is convex.

Let $BC(\mathcal{Y}_{0,n})$ denotes the set of bounded continuous real-valued functions on $\mathcal{Y}_{0,n}$. We need the following.

Assumption 1: The following conditions are assumed throughout the paper.
(A1) $\mathcal{Y}_{0,n}$ is a compact Polish space, $\mathcal{X}_{0,n}$ is a Polish space;
(A2) for all $h(\cdot) \in BC(\mathcal{Y}_{0,n})$, the function mapping $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1} \mapsto \int_{\mathcal{Y}_n} h(y) P_{Y|Y^{n-1},X^n}(dy|y^{n-1}, x^n) \in \mathbb{R}$ is continuous jointly in the variables $(x^n, y^{n-1}) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n-1}$;
(A3) $d_{0,n}(x^n, \cdot)$ is continuous on $\mathcal{Y}_{0,n}$;
(A4) the distortion level D is such that there exist sequence $(x^n, y^n) \in \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}$ satisfying $d_{0,n}(x^n, y^n) < D$.

Note that since $\mathcal{Y}_{0,n}$ is assumed to be a compact Polish space, then by [10] probability measures on $\mathcal{Y}_{0,n}$ are weakly compact. Moreover, the following weak compactness result can be obtained.

Lemma 1: Suppose Assumption 1 **(A1)**, **(A2)** hold. Then
(1) The set $\overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ is weakly compact.
(2) Under the additional conditions **(A3)**, **(A4)** the set $\overrightarrow{\mathcal{Q}}_{0,n}(D)$ is a closed subset of $\overrightarrow{\mathcal{Q}}(\mathcal{Y}_{0,n}; \mathcal{X}_{0,n})$ (hence compact).

Proof: The derivation is found in [11]. ■

The previous results follow from Prohorov's theorem that relates tightness and weak compactness. The next theorem establishes existence of the minimizing reproduction distribution for (4); it follows from Lemma 1 and the lower semicontinuity of $\mathbb{I}_{X^n \to Y^n}(P_{X^n}, \cdot)$ with respect to $\overrightarrow{P}_{Y^n|X^n}$ [11].

Theorem 2: (**Existence**) Suppose the conditions of Lemma 1 hold. Then $R_{0,n}^{na}(D)$ has a minimum.

Proof: The derivation is found in [11]. ■

IV. Optimal Reproduction of Nonanticipative RDF

In this section the form of the optimal reproduction conditional distribution is derived under a stationarity assumption. We introduce the following main assumption.

Assumption 2: (**Stationarity**) The $(n+1)$-fold convolution conditional distribution $\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i)$, is the convolution of stationary conditional distributions.

The consequence of Assumption 2, which holds for stationary processes and a single letter distortion function, is that the Gateaux differential of $\mathbb{I}_{X^n \to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ is done in only one direction $\overrightarrow{P}_{Y^n|X^n} - \overrightarrow{P}^0_{Y^n|X^n}$ via $\overrightarrow{P}^{\epsilon}_{Y^n|X^n} \stackrel{\triangle}{=} \overrightarrow{P}_{Y^n|X^n} + \epsilon\big(\overrightarrow{P}_{Y^n|X^n} - \overrightarrow{P}^0_{Y^n|X^n}\big)$, $\epsilon \in [0,1]$, since under Assumption 2, the functionals $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) \in \mathcal{Q}(\mathcal{Y}_i; \mathcal{Y}_{0,i-1} \times \mathcal{X}_{0,i}) :\ i = 0, 1, \ldots, n\}$ are identical.

Theorem 3: Suppose Assumption 2 holds and $\mathbb{I}_{P_{X^n}}(\overrightarrow{P}_{Y^n|X^n}) \stackrel{\triangle}{=} \mathbb{I}_{X^n\to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ is well defined for every $\overrightarrow{P}_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}_{0,n}(D)$ possibly taking values from the set $[0,\infty]$. Then $\overrightarrow{P}_{Y^n|X^n} \to \mathbb{I}_{P_{X^n}}(\overrightarrow{P}_{Y^n|X^n})$ is Gateaux differentiable at every point in $\overrightarrow{\mathcal{Q}}_{0,n}(D)$, and the Gateaux derivative at the point $\overrightarrow{P}^0_{Y^n|X^n}$ in the direction $\overrightarrow{P}_{Y^n|X^n} - \overrightarrow{P}^0_{Y^n|X^n}$ is given by

$$\begin{aligned}
&\delta\mathbb{I}_{P_{X^n}}(\overrightarrow{P}^0_{Y^n|X^n}, \overrightarrow{P}_{Y^n|X^n} - \overrightarrow{P}^0_{Y^n|X^n}) \\
&= \int \log\left(\frac{\overrightarrow{P}^0_{Y^n|X^n}(dy^n|x^n)}{P^0_{Y^n}(dy^n)}\right) \\
&\otimes (\overrightarrow{P}_{Y^n|X^n} - \overrightarrow{P}^0_{Y^n|X^n})(dy^n|x^n)P_{X^n}(dx^n)
\end{aligned}$$

where $P^0_{Y^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n})$ is the marginal measure corresponding to $\overrightarrow{P}^0_{Y^n|X^n} \otimes P_{X^n} \in \mathcal{M}_1(\mathcal{Y}_{0,n}\times\mathcal{X}_{0,n})$.

Proof: The proof is similar to the one in [12] (although it is more involved). ■

The constrained problem defined by (4) can be reformulated as an unconstrained problem using Lagrange multipliers [11]

$$\begin{aligned}
R^{na}_{0,n}(D) = \sup_{s\le 0} \inf_{\substack{\overrightarrow{P}_{Y^n|X^n} \\ \in\overrightarrow{Q}(\mathcal{Y}_{0,n};\mathcal{X}_{0,n})}} &\Big\{\mathbb{I}_{X^n\to Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \\
- s(\ell_{d_{0,n}}(\overrightarrow{P}_{Y^n|X^n}) - D(n+1))\Big\},\ & s\in(-\infty,0]. \quad (5)
\end{aligned}$$

The above observations yield the following theorem.

Theorem 4: (**Optimal Reproduction Distribution**) Suppose the Assumption 2 holds and consider $d_{0,n}(x^n,y^n) \stackrel{\triangle}{=} \sum_{i=0}^n \rho(T^ix^n, T^iy^n)$. Then
(1) The infimum in (5) is attained at $\overrightarrow{P}^*_{Y^n|X^n} \in \overrightarrow{\mathcal{Q}}_{0,n}(D)$ given by[5]

$$\begin{aligned}
&\overrightarrow{P}^*_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \\
&= \otimes_{i=0}^n \frac{e^{s\rho(T^ix^n,T^iy^n)}P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho(T^ix^n,T^iy^n)}P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})} \qquad (6)
\end{aligned}$$

where $s\le 0$ and $P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1}) \in \mathcal{Q}(\mathcal{Y}_i;\mathcal{Y}_{0,i-1})$.
(2) The nonanticipative RDF is given by

$$\begin{aligned}
R^{na}_{0,n}(D) = sD(n+1) - \sum_{i=0}^n \int \log\Big(\int_{\mathcal{Y}_i} e^{s\rho(T^ix^n,T^iy^n)} \\
P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})\Big)\overrightarrow{P}^*_{Y^{i-1}|X^{i-1}}(dy^{i-1}|x^{i-1})\otimes P_{X^i}(dx^i).
\end{aligned}$$

If $R^{na}_{0,n}(D) > 0$ then $s<0$ and

$$\sum_{i=0}^n \int \rho(T^ix^n,T^iy^n)\overrightarrow{P}^*_{Y^i|X^i}(dy^i|x^i)P_{X^i}(dx^i) = (n+1)D.$$

[5]Due to stationarity assumption $P_{Y_i|Y^{i-1}}(\cdot|\cdot) = P(\cdot|\cdot)$ and $P^*_{Y_i|Y^{i-1},X^i}(\cdot|\cdot,\cdot) = P^*(\cdot|\cdot,\cdot)$

Proof: The derivation is found in [11]. ■

Remark 1: Note that if the distortion function satisfies $\rho(T^ix^n, T^iy^n) = \rho(x_i, T^iy^n)$ then for $i = 0,1,\ldots,n$

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) = P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x_i)$$

that is, the reproduction kernel is Markov in X^n.

V. Realization of Nonanticipative RDF

The realization of the nonanticipative RDF (optimal reproduction conditional distribution) is equivalent to the sensor mapping as shown in Fig. 2 which produces the auxiliary random process $\{Z_i : i\in\mathbb{N}\}$ which is used for filtering. This is equivalent to identifying a communication channel, an encoder and a decoder such that the reproduction from the sequence X^n to the sequence Y^n matches the nonanticipative rate distortion minimizing reproduction kernel. Fig. 3 illustrates the cascade subsystems that realize the nonanticipative RDF.

Definition 2: (**Realization**) Given a source $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0,\ldots,n\}$, a channel $\{P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i) : i = 0,\ldots,n\}$ is a realization of the optimal reproduction distribution (6) if there exists a pre-channel encoder $\{P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) : i = 0,\ldots,n\}$ and a post-channel decoder $\{P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) : i = 0,\ldots,n\}$ such that

$$\begin{aligned}
\overrightarrow{P}^*_{Y^n|X^n}(dy^n|x^n) &\stackrel{\triangle}{=} \otimes_{i=0}^n P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \\
&= \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \quad (7)
\end{aligned}$$

where (7) is generated from the joint distribution

$$\begin{aligned}
&P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n) \\
&= \otimes_{i=0}^n P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) \\
&\quad\otimes P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i) \\
&\quad\otimes P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) \\
&\quad\otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1}).
\end{aligned}$$

The filter is given by $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i = 0,\ldots,n\}$ or by $\{P_{X_i|Y^{i-1}}(dx_i|y^{i-1}) : i = 0,\ldots,n\}$.

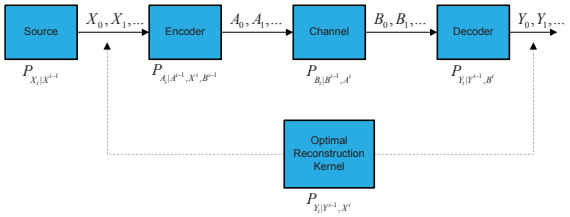

Fig. 3. Realizable nonanticipative rate distortion function.

Clearly, $\{B_i : i = 0,1,\ldots,n\}$ is an auxiliary random process which is needed to obtain the filter $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i = 0,\ldots,n\}$. If we further ensure that there exists (D,P) such that $R^{na}(D) \stackrel{\triangle}{=}$

$\lim_{n\to\infty}\frac{1}{n+1}R^{na}_{0,n}(D)=C(P)$, where $C(P)$ is the capacity of the channel with power level P, then the realization of Fig. 3 is equivalent to symbol-by-symbol transmission in which the source is matched to the channel, e.g., real-time transmission of information.

VI. Example

Consider the following discrete-time partially observed linear Gauss-Markov system described by

$$\begin{cases} X_{t+1}=AX_t+BW_t,\ X_0=X\in\mathbb{R}^n,\ t\in\mathbb{N} \\ Y_t=CX_t+GV_t,\ t\in\mathbb{N} \end{cases} \tag{8}$$

where $X_t\in\mathbb{R}^m$ is the state (unobserved) process of information source (plant), and $Y_t\in\mathbb{R}^p$ is the partially measurement (observed) process. Assume that (C,A) is detectable and $(A,\sqrt{BB^{tr}})$ is stabilizable, $(G\neq 0)$. The state and observation noises $\{(W_t,V_t):t\in\mathbb{N}\}$, $W_t\in\mathbb{R}^k$ and $V_t\in\mathbb{R}^p$, are Gaussian IID processes with zero mean and identity covariances are mutually independent, and independent of the Gaussian RV X_0, with parameters $N(\bar{x}_0,\bar{V}_0)$.

The objective is to reconstruct $\{Y_t:\ t\in\mathbb{N}\}$ from $\{\tilde{Y}_t:\ t\in\mathbb{N}\}$ using single letter distortion. First, we compute

$$R^{na}_{0,n}(D)=\inf_{\overrightarrow{P}_{\tilde{Y}^n|Y^n}\in\overrightarrow{\mathcal{Q}}_{0,n}(D)}\frac{1}{n+1}\mathbb{I}_{X^n\to Y^n}(P_{Y^n},\overrightarrow{P}_{\tilde{Y}^n|Y^n})$$

and then realize the optimal reproduction distribution. According to Theorem 4, the optimal reproduction is given by

$$\overrightarrow{P}^{*}_{\tilde{Y}^n|Y^n}(d\tilde{y}^n|y^n)=\otimes_{t=0}^{n}\frac{e^{s||\tilde{y}_t-y_t||^2}P_{\tilde{Y}_t|\tilde{Y}^{t-1}}(d\tilde{y}_t|\tilde{y}^{t-1})}{\int_{\mathcal{Y}_t}e^{s||\tilde{y}_t-y_t||^2}P_{\tilde{Y}_t|\tilde{Y}^{t-1}}(d\tilde{y}_t|\tilde{y}^{t-1})} \tag{9}$$

where $s\leq 0$. Hence, from (9) it follows that $P_{\tilde{Y}_t|\tilde{Y}^{t-1},Y^t}=P_{\tilde{Y}_t|\tilde{Y}^{t-1},Y_t}(d\tilde{y}_t|\tilde{y}^{t-1},y_t)-a.a.$, that is, the reproduction is Markov with respect to the process $\{Y_t:\ t\in\mathbb{N}\}$, and $\{(X_t,Y_t):\ t\in\mathbb{N}\}$ is jointly Gaussian, hence it follows that $P_{\tilde{Y}_t|\tilde{Y}^{t-1},Y_t}(\cdot|\tilde{y}^{t-1},y_t)$ is Gaussian. Hence, it has the general form

$$\tilde{Y}_t=\bar{A}Y_t+\bar{B}\tilde{Y}^{t-1}+\bar{Z}_t,\ t\in\mathbb{N} \tag{10}$$

where $\bar{A}_t\in\mathbb{R}^{p\times p}$, $\bar{B}_t\in\mathbb{R}^{p\times tp}$, and $\{\bar{Z}_t:\ t\in\mathbb{N}\}$ is an independent sequence of Gaussian vectors.

The nonanticipative RDF is given by [11]

$$R^{na}_{0,n}(D)=\frac{1}{n+1}\sum_{t=0}^{n}\sum_{i=1}^{p}\log\left(\frac{\lambda_{t,i}}{\delta_{t,i}}\right) \tag{11}$$

where $\{\xi_t:\ t\in\mathbb{N}\}$ are such that

$$\delta_{t,i}\stackrel{\triangle}{=}\begin{cases}\xi_t & \text{if}\quad \xi_t\leq\lambda_{t,i}\\ \lambda_{t,i} & \text{if}\quad \xi_t>\lambda_{t,i}\end{cases},\ t\in\mathbb{N},\ i=1,\ldots,p$$

and $\{\xi_t:\ t\in\mathbb{N}\}$ satisfies $\sum_{i=1}^{p}\delta_{t,i}=D$. Define $\Delta_t\stackrel{\triangle}{=}diag\{\delta_{t,i},\ldots,\delta_{t,p}\}$.

We realize (10) and (11) via a scalar additive Gaussian noise (AGN) channel with feedback defined by

$$B_t=A_t+Z_t,\ Var(Z_t)=Q,\ t\in\mathbb{N} \tag{12}$$

where the encoder is a mapping $A_t=\Phi_t(Y_t,\tilde{Y}^{t-1})$ with power $P_t\stackrel{\triangle}{=}E\{(A_t)^2\}$. Hence, the capacity of (12) is $C(P)\stackrel{\triangle}{=}\lim_{n\to\infty}\frac{1}{n+1}I(A^n\to B^n)=\lim_{n\to\infty}\frac{1}{2}\frac{1}{n+1}\sum_{t=0}^{n}\log\left(1+E\{(A_t)^2\}Var(Z_t)^{-1}\right)=\frac{1}{2}\log(1+\frac{P}{Q})$.

Realization of the nonanticipative RDF. The realization is based on the block diagram of Fig. 4. The encoder $\Phi_t(\cdot,\cdot)$ consists of a pre-encoder which produces the Gaussian innovation process $\{K_t:\ t\in\mathbb{N}\}$, defined by

$$K_t\stackrel{\triangle}{=}Y_t-E\Big\{Y_t|\sigma\{\tilde{Y}^{t-1}\}\Big\},\ t\in\mathbb{N} \tag{13}$$

whose covariance is defined by $\Lambda_t\stackrel{\triangle}{=}E\{K_tK_t^{tr}\}$. The decoder consists of a pre-decoder $\{\tilde{K}_t:\ t\in\mathbb{N}\}$ which is defined by

$$\tilde{K}_t\stackrel{\triangle}{=}\tilde{Y}_t-E\Big\{Y_t|\sigma\{\tilde{Y}^{t-1}\}\Big\},\ t\in\mathbb{N}. \tag{14}$$

Let $\{E_t:\ t\in\mathbb{N}\}$ be the unitary matrix such that

$$E_t\Lambda_tE_t^{tr}=diag\{\lambda_{t,1},\ldots\lambda_{t,p}\},\ t\in\mathbb{N}. \tag{15}$$

Define $\Gamma_t\stackrel{\triangle}{=}E_tK_t$ and let $\{\tilde{\Gamma}_t:\ t\in\mathbb{N}^n\}$ denote its reproduction.

Thus, the pre-encoder can be further scalled by $\Gamma_t=E_tK_t$, and Γ_t is compressed by $A_t=\mathcal{A}_t\Gamma_t$ and sent through the AGN channel with feedback, after which the received signal is decompressed by $\tilde{\Gamma}_t=\mathcal{B}_tB_t$ in the pre-decoder. By the knowledge of the channel output at the decoder, the mean square estimator $\hat{X}_t$ is generated at the decoder (and encoder because $\hat{X}_t\stackrel{\triangle}{=}E\{X_t|\sigma\{\tilde{Y}_{t-1}\}\}$). The complete design is illustrated in Fig. 4. We can design

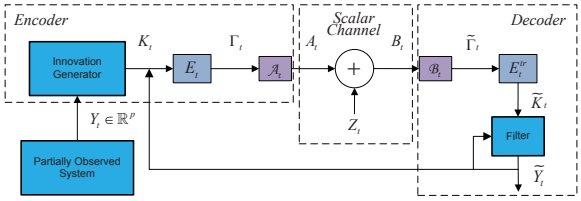

Fig. 4. Design of the discrete-time communication system with scalar additive Gaussian noise (AGN) channel.

$\{(\mathcal{A}_t,\mathcal{B}_t):\ t\in\mathbb{N}\}$ by

$$\mathcal{A}_t=\left[\sqrt{\frac{\alpha_1P_t}{\lambda_{t,1}}},\ldots,\sqrt{\frac{\alpha_pP_t}{\lambda_{t,p}}}\right],\ t\in\mathbb{N}$$

$$\mathcal{B}_t=\left[\sqrt{\alpha_1P_t\lambda_{t,1}},\ldots,\sqrt{\alpha_pP_t\lambda_{t,p}}\right]^{tr},\ t\in\mathbb{N}$$

where $\sum_{i=1}^{p} \alpha_i = 1$, $i = 1, \ldots, p$. Note that $H_t \stackrel{\triangle}{=} \mathcal{B}_t \mathcal{A}_t$.

Decoder. From Fig. 4,

$$\tilde{K}_t = E_t^{tr} \tilde{\Gamma}_t = E_t^{tr} H_t E_t K_t + E_t^{tr} \mathcal{B}_t Z_t, \ t \in \mathbb{N}.$$

The reproduction of Y_t is given by the sum of $\tilde{K}_t$ and $C\hat{X}_t$ as follows.

$$\begin{aligned} \tilde{Y}_t &= E_t^{tr} H_t E_t K_t + E_t^{tr} \mathcal{B}_t Z_t + C\hat{X}_t, \ t \in \mathbb{N}. \\ &= E_t^{tr} H_t E_t C(X_t - \hat{X}_t) + C\hat{X}_t \\ &\quad + (E_t^{tr} H_t E_t G V_t + E_t^{tr} \mathcal{B}_t Z_t) \end{aligned}$$

where $\{V_t : \ t \in \mathbb{N}\}$ and $\{Z_t : \ t \in \mathbb{N}\}$ are independent Gaussian vectors. The desired distortion is achieved as follows

$$\begin{aligned} &E\Big\{(Y_t - \tilde{Y}_t)^{tr}(Y_t - \tilde{Y}_t)\Big\} \\ &= Tr\Big\{E_t^{tr}\Big((I - H_t) diag(\lambda_{t,1}, \ldots, \lambda_{t,p})(1 - H_t)^{tr} \\ &+ (\mathcal{B}_t Q \mathcal{B}_t^{tr})\Big) E_t\Big\} = \sum_{i=0}^{p} \delta_{t,i} = D. \end{aligned} \tag{16}$$

Thus, from (16), $\{\delta_{t,i}\}_{i=1}^{p}$ are eigenvalues of the matrix

$$T_t \stackrel{\triangle}{=} (I - H_t) diag(\lambda_{t,1}, \ldots, \lambda_{t,p})(1 - H_t)^{tr} + (\mathcal{B}_t Q \mathcal{B}_t^{tr})$$

and we can calculate $\{a_i\}_{i=1}^{p}$ and P_t in terms of $\{\lambda_{t,i}, \delta_{t,i}\}_{i=1}^{p}$ and Q.

The decoder is $\tilde{Y}_t = \tilde{K}_t + C\hat{X}_t$, where $\{\hat{X}_t : \ t \in \mathbb{N}\}$ is obtained from the modified Kalman filter as follows.

$$\begin{aligned} \hat{X}_{t+1} &= A\hat{X}_t + A\Sigma_t (E_t^{tr} H_t E_t C)^{tr} M_t^{-1} (\tilde{Y}_t - C\hat{X}_t), \hat{X}_0 = \bar{x}_0 \\ \Sigma_{t+1} &= A\Sigma_t A^{tr} - A\Sigma_t (E_t^{tr} H_t E_t C)^{tr} M_t^{-1} (E_t^{tr} H_t E_t C)\Sigma_t A \\ &\quad + BB_t^{tr}, \ \Sigma_0 = \bar{\Sigma}_0 \end{aligned}$$

where

$$\begin{aligned} M_t &= E_t^{tr} H_t E_t C \Sigma_t (E_t^{tr} H_t E_t C)^{tr} \\ &\quad + E_t^{tr} H_t E_t G G^{tr} (E_t^{tr} H_t E_t)^{tr} + E_t^{tr} \mathcal{B}_t Q \mathcal{B}_t^{tr} E_t. \end{aligned}$$

Infinite Horizon. As $t \to \infty$, under the assumption that the linear system is stabilizable and detectable, we have

$$\begin{aligned} \Sigma_\infty &= A\Sigma_\infty A^{tr} \\ &\quad - A\Sigma_\infty (E_\infty^{tr} H_\infty E_\infty C)^{tr} M_\infty^{-1} (E_\infty^{tr} H_\infty E_\infty C)\Sigma_\infty A \\ &\quad + BB_\infty^{tr} \end{aligned}$$

where

$$\begin{aligned} M_\infty &= E_\infty^{tr} H_\infty E_\infty C \Sigma_\infty (E_\infty^{tr} H_\infty E_\infty C)^{tr} \\ &\quad + E_\infty^{tr} H_\infty E_\infty G G^{tr} (E_\infty^{tr} H_\infty E_\infty)^{tr} + E_\infty^{tr} \mathcal{B}_\infty Q \mathcal{B}_\infty^{tr} E_t \end{aligned}$$

and E_∞ is the unitary matrix that diagonalizes Λ_∞ by

$$E_\infty \Lambda_\infty E_\infty^{tr} = diag(\lambda_{\infty,1}, \ldots, \lambda_{t,p}).$$

Also,

$$\delta_{\infty,i} \triangleq \left\{ \begin{array}{ll} \xi_\infty & \text{if} \ \ \xi_\infty \leq \lambda_{\infty,i} \\ \lambda_{\infty,i} & \text{if} \ \ \xi_\infty > \lambda_{\infty,i} \end{array} \right. , \ i = 1, \ldots, p$$

satisfying $\sum_{i=1}^{p} \delta_{\infty,i} = D$. Define $\Delta_\infty \stackrel{\triangle}{=} diag(\delta_{\infty,1}, \ldots, \delta_{\infty,p})$.

Finally, we show matching of the source to the channel.

$$\begin{aligned} R^{na}(D) &\stackrel{\triangle}{=} \lim_{t \to \infty} \frac{1}{n+1} R_{0,n}^{na}(D) \\ &= \lim_{n \to \infty} \frac{1}{2} \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \log\Big(\frac{\lambda_{t,i}}{\delta_{t,i}}\Big) = \frac{1}{2} \sum_{i=1}^{p} \log\Big(\frac{\lambda_{\infty,i}}{\delta_{\infty,i}}\Big) \\ &= \frac{1}{2} \log \frac{|\Lambda_\infty|}{|\Delta_\infty|} = \frac{1}{2} \log(1 + \frac{P}{Q}) = C(P). \end{aligned}$$

Thus, for a given (D, P), $C(P) = R^{na}(D)$ is the minimum capacity under which there exists a realizable filter for the data reproduction of $\{Y_t : \ t \in \mathbb{N}\}$ by $\{\tilde{Y}_t : \ t \in \mathbb{N}\}$ ensuring an average distortion equal to D. This is precisely the so-called source-channel matching with symbol-by-symbol transmission.

REFERENCES

[1] R. S. Bucy, "Distortion rate theory and filtering," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 336–340, Mar. 1982.

[2] A. K. Gorbunov and M. S. Pinsker, "Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes," *Problems of Information Transmission*, vol. 27, no. 4, pp. 361–365, 1991.

[3] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[4] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, July 2004.

[5] G. N. Nair and R. J. Evans, "Stabilizability of Stochastic Linear Systems with Finite Feedback Data Rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.

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

[7] J. L. Massey, "Causality, feedback and directed information," in *International Symposium on Information Theory and its Applications (ISITA '90)*, Nov. 27-30 1990, pp. 303–305.

[8] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: properties and extremum problems," in *IEEE International Symposium on Information Theory (ISIT)*, July 1-6 2012, pp. 518–522, an extended version is submitted in *IEEE Transactions on Information Theory* and it is available online at http://arxiv.org/abs/1302.3971.

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

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

[11] P. A. Stavrou and C. D. Charalambous, "Nonanticipative rate distortion function and filtering theory: A weak convergence approach," *submitted to Systems and Control Letters*, 2013. [Online]. Available: http://arxiv.org/abs/1212.6643

[12] F. Rezaei, N. U. Ahmed, and C. D. Charalambous, "Rate distortion theory for general sources with potential application to image processing," *International Journal of Applied Mathematical Sciences*, vol. 3, no. 2, pp. 141–165, 2006.