OPTIMAL NONANTICIPATIVE ESTIMATION SCHEMES FOR TIME-VARYING GAUSS-MARKOV PROCESSES

PHOTIOS A. STAVROU†, THEMISTOKLIS CHARALAMBOUS‡, CHARALAMBOS D. CHARALAMBOUS∗, AND SERGEY LOYKA§

Abstract. In this paper, we derive recursive filters for time-varying multidimensional Gauss-Markov processes, which satisfy a mean square error fidelity, using the concept of Finite Time Horizon (FTH) Nonanticipative Rate Distortion Function (NRDF) and its connection to real-time realizable filtering theory. Moreover, we derive a universal lower bound on the mean square error of any estimator of time-varying multidimensional Gauss-Markov processes in terms of conditional mutual information. Unlike classical Kalman filters, the proposed filter is constructed from the solution of a reverse-waterfilling problem, which ensures that the mean square error fidelity is met. Our theoretical results are demonstrated via illustrative examples.

Key words. Finite horizon, Kalman filter, nonanticipative rate distortion function, mean square error distortion, reverse-waterfilling, universal lower bound.

1. Introduction. Nonanticipative Rate Distortion Function (NRDF) with respect to a fidelity of reproduction is directly related to Bayesian filtering theory, in which the estimators are constructed to satisfy specific accuracy requirements or fidelity constraints, depending on the choice of the fidelity, such as, average or probabilistic error constraints. Applications include information processing of sensor networks like, for example, in [2] and control over limited capacity communication channels as, for example, in [2, 6]. In such applications, the optimal filters or estimators are required to meet specific performance demands, such as, the mean square estimation error is below a specific level. Specifically, if the objective is to transmit an information process with entropy rate which is higher than the rate supported by the communication link, then the information process should be quantized prior to transmission. The NRDF is the minimum rate of reproducing the information process by another process subject to a fidelity constraint.

In general, the NRDF [1] is directly linked to the design of control-communication schemes, where the controller and estimator architectures process quantized information, by realizing the optimal reproduction conditional distribution of the NRDF by zero-delay or delay constrained quantization schemes (see [7, 9, 11, 12] and references therein). Moreover, the NRDF is necessary for the realization of the compression channel by communication systems processing information causally with zero-delay, see [13] for a specific construction.

The intersection between information theory and filtering theory is first established by Bucy in [14], where the author considered the distortion rate function [7] of a two sample Gaussian process and related this to optimal causal estimation design.
The relation between Nonanticipative Rate Distortion (NRD) theory and real-time realizable filtering theory for general processes is established in [22], where the optimal reproduction distribution under a stationary ergodic assumption is derived, and the connection to real-time realizable filtering theory via a realization scheme utilizing time-invariant partially observable multidimensional Gauss-Markov processes is established. In general, when it comes to the design of controllers or estimators based on information theoretic payoffs, the literature is vast. Some illustrative techniques can be found in the following papers [16–20].

In this paper, we derive the optimal nonstationary reproduction distribution of the FTH NRDF, and then we connect it to real-time realizable filtering theory of time-varying (nonstationary) multidimensional Gauss-Markov processes. The new recursive estimator is finite-dimensional, and involves a time-space reverse-waterfilling, which ensures the fidelity constraint is met. The time-space reverse-waterfilling implies that, given a distortion level, there exists an optimal level which serves as a decision criterion for the estimator whether it should or should not reconstruct the state process in time and space (dimensions). This is the fundamental difference from the well-known Kalman filter equations. In addition, we derive a universal lower bound on the mean square error of any estimator in terms of the FTH NRDF, which is a variant of the Cramér-Rao bound. This bound generalizes the well-known bound for Gaussian Random Variables (see [21,22]) to time-varying multidimensional Gauss-Markov processes.

Before we describe the contributions of this paper in detail, in Section 1.1 we provide a brief introduction on the connection and the differences between Bayesian estimation theory and NRD theory, and in Section 1.2 the connection of NRD theory in zero-delay communication.

1.1. Bayesian Estimation Theory vs. NRD Theory. In classical filtering [23,24], one is given a model that generates the process \( \{X_t : t = 0, \ldots, n\} \), via its conditional distribution \( P_{X_t \mid X_{t-1}}(dx_t \mid x_{t-1}) : t = 0, \ldots, n \), \( x_{t-1} = \{x_0, x_1, \ldots, x_{t-1}\} \), or via discrete-time recursive dynamics, a model that generates observed data obtained from sensors (unobserved process) \( \{Z_t : t = 0, \ldots, n\} \), via its conditional distribution \( P_{Z_t \mid Z_{t-1}, X_t}(dz_t \mid x_{t-1}, x_t) : t = 0, \ldots, n \), while \( \{\hat{X}_t : t = 0, \ldots, n\} \) are the causal estimates of the process \( \{X_t : t = 0, \ldots, n\} \) based on the observed data \( \{Z_t : t = 0, \ldots, n\} \). As a result, in classical filtering theory, both models which generate the unobserved and observed processes, \( \{X_t : t = 0, \ldots, n\} \) and \( \{Z_t : t = 0, \ldots, n\} \), respectively, are given a priori, while the estimator \( \{\hat{X}_t : t = 0, \ldots, n\} \) is a nonanticipative functional of the past information \( Z_{t-1}, t = 0, \ldots, n \), often computed recursively, like the Kalman filter.

Fig. 1.1 illustrates the cascade block diagram of the Bayesian filtering problem.

In filtering based on the NRDF, one is given the process (source) \( \{X_t : t = 0, \ldots, n\} \), which induces the conditional distributions \( \{P_{X_t \mid X_{t-1}}(dx_t \mid x_{t-1}) : t = 0, \ldots, n\} \) and determines the optimal nonanticipative reproduction conditional distribution.

![Bayesian Filtering Problem](image-url)
which minimizes directed information between \( \{X_t : t = 0, \ldots, n\} \) and \( \{Y_t : t = 0, \ldots, n\} \) subject to a fidelity of reproduction between \( \{X_t : t = 0, \ldots, n\} \) and \( \{Y_t : t = 0, \ldots, n\} \), i.e., the mean square error constraint. In this case, the reproduction process \( \{Y_t : t = 0, \ldots, n\} \) of \( \{X_t : t = 0, \ldots, n\} \) is found by realizing the optimal reproduction distribution \( P_{Y_t|Y_{t-1},X_t}(dy_t|y_{t-1},x_t) : t = 0, \ldots, n \) via the cascade of sub-systems shown in Fig. 1.2. The estimator on \( \{X_t : t = 0, \ldots, n\} \) is any function of \( \{Y_t : t = 0, \ldots, n\} \), like the conditional mean \( E\{X_t|Y_{t-1}\} \).

![Diagram of communication system](image)

**Fig. 1.2:** Filtering with fidelity via NRD theory.

The fundamental difference between Bayesian filtering theory and the realizable filtering theory based on NRDF with a prescribed fidelity, lies on the fact that in Bayesian estimation, the sensor map is given a priori, while for the NRDF, this map is identified as a part of the realization of the optimal reproduction conditional distribution \( P_{Y_t|Y_{t-1},X_t}(dy_t|y_{t-1},x_t) : t = 0, \ldots, n \), so that the end-to-end NRDF from \( \{X_t : t = 0, \ldots, n\} \) to \( \{Y_t : t = 0, \ldots, n\} \) subject to a prescribed fidelity constraint is achieved.

### 1.2. Real-Time Communication via Realization of NRDF

The realization scheme in Fig. 1.2 is equivalent to identifying a zero-delay communication scenario, where an **encoder**, **channel**, and **decoder** are constructed, and realize the optimal reproduction conditional distribution \( P_{Y_t|Y_{t-1},X_t}(dy_t|y_{t-1},x_t) : t = 0, \ldots, n \) of the process \( \{X_t : t = 0, \ldots, n\} \) by the process \( \{Y_t : t = 0, \ldots, n\} \), subject to a prescribed fidelity. Fig. 1.3 illustrates the zero-delay communication system via a cascade of sub-systems. Clearly, in Fig. 1.3 \( \{Z_t : t = 0, \ldots, n\} \) is an auxiliary random process which is needed to obtain the filter \( P_{X_t|Z_{t-1}}(dx_t|z_{t-1}) : t = 0, \ldots, n \). The realization scheme shown in Fig. 1.3 is described for zero-delay communication systems for general sources and channels with memory in \([9\text{ Section V}]\). It is also described in \([5]\) and \([8]\) for control over finite capacity communication channels, since this technique allows one to design communication systems, such as, Gaussian systems, without encoding and decoding delays. An application example of the optimal real-time communication for finite alphabet sources and channels with memory is constructed in \([13]\).

### 1.3. Contributions

In this paper, we utilize the concept of FTH NRDF, in the context of filtering theory, to investigate processes governed by time-varying multidimensional Gauss-Markov processes. Towards this end, we summarize the contributions of the paper as follows.
Notation. Let $X_n, Y_n, Z_n, R_n, R_{n-1}, Z_{n-1}$ be measurable spaces. Let $\mathcal{X}_n, \mathcal{Y}_n, \mathcal{Z}_n, \mathcal{R}_n, \mathcal{R}_{n-1}$ be Polish spaces, that include finite, countable, and continuous alphabet. We realize the optimal nonstationary reproduction distribution for the time-varying multidimensional Gaussian processes by an $\{\text{encoder}, \text{channel}, \text{decoder}\}$ in the sense of Fig. 1.3 and we construct an optimal pair of linear encoders and decoders processing information causally with zero-delay.

(R1) We give a closed form expression that achieves the infimum of the FTH NRDF and we identify certain structural properties of the optimal nonstationary reproduction distribution.

(R2) We derive the analytical expression of FTH NRDF for time-varying Gauss-Markov processes subject to a mean squared error distortion, which includes a time-space reverse-waterfilling algorithm.

(R3) We realize the optimal nonstationary reproduction distribution for the time-varying multidimensional Gauss-Markov processes by an $\{\text{encoder}, \text{channel}, \text{decoder}\}$ in the sense of Fig. 1.3 and we construct an optimal pair of linear encoders and decoders processing information causally with zero-delay.

(R4) We derive a universal lower bound on the mean square error for any causal estimator of Gaussian processes in terms of FTH NRDF.

The time-space reverse-waterfilling algorithm of the optimal solution is solved by proposing an iterative algorithm capable of allocating optimally the rate-distortion levels.

The rest of the paper is structured as follows. In Section 2 we introduce NRDF for general processes. In Section 3 we describe the form of the optimal nonstationary (time-varying) reproduction distribution of the FTH NRDF. In Section 4 we concentrate on evaluating the FTH NRDF for time-varying multidimensional Gaussian processes with memory, present examples in the context of realizable filtering theory, and we derive a universal lower bound to the mean square error of any estimator of Gaussian processes based on FTH NRDF. We draw conclusions and discuss future directions in Section 5.

2. FTH NRDF on General Alphabets. In this section, we introduce the definition of FTH NRDF for general processes taking values in Polish spaces (complete separable metric spaces), that include finite, countable, and continuous alphabet spaces.

Notation. Let $\mathbb{N}_0 \triangleq \{0, 1, 2, \ldots\}$ and $\mathbb{N}_0^n \triangleq \{0, 1, 2, \ldots, n\}, n \in \mathbb{N}_0$. Let $(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{N}_0$ and $(\mathcal{Y}_n, \mathcal{B}(\mathcal{Y}_n)) : n \in \mathbb{N}_0$ be measurable spaces, where $\mathcal{X}_n, \mathcal{Y}_n, n \in \mathbb{N}_0$, are Polish spaces, and $\mathcal{B}(\mathcal{X}_n)$ and $\mathcal{B}(\mathcal{Y}_n)$ their Borel $\sigma$–algebras. Points in $\mathcal{X}_0^n \triangleq \times_{i \in \mathbb{N}_0^1} \mathcal{X}_i$ and $\mathcal{Y}_0^n \triangleq \times_{i \in \mathbb{N}_0^n} \mathcal{Y}_i$ are denoted by $x^n \triangleq \{x_0, x_1, \ldots, x_n\} \in \mathcal{X}_0^n$ and $y^n \triangleq \{y_0, y_1, \ldots, y_n\} \in \mathcal{Y}_0^n$, respectively. Let $\mathcal{B}(\mathcal{X}_0^n)$ and $\mathcal{B}(\mathcal{Y}_0^n)$ denote the
\(\sigma\)-algebras with bases over \(A_t \in \mathcal{B}(X_t)\), and \(B_t \in \mathcal{B}(Y_t)\), respectively, for \(t \in \mathbb{N}_0\). The set of probability distributions on any measurable space \((\mathbb{X}, \mathcal{B}(\mathbb{X}))\) is denoted by \(\mathcal{M}(\mathbb{X})\). In Section [3], we denote the time index with “t” and the spatial index with “n”.

Next, we give the definition of a stochastic kernel (stochastic kernels and conditional distributions are identical notions and they are used interchangeably).

**Definition 2.1**. ([23] Stochastic kernel) Let \((\mathbb{X}, \mathcal{B}(\mathbb{X})), (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))\) be measurable spaces in which \(\mathcal{Y}\) is a Polish Space. A stochastic kernel on \(\mathcal{Y}\) given \(\mathbb{X}\) is a mapping \(Q : \mathcal{B}(\mathcal{Y}) \times \mathbb{X} \rightarrow [0, 1]\) satisfying the following two properties:

1. For every \(x \in \mathbb{X}\), the set function \(Q(\cdot | x)\) is a probability measure on \(\mathcal{B}(\mathcal{Y})\);
2. For every \(F \in \mathcal{B}(\mathcal{Y})\), the function \(Q(F | \cdot)\) is \(\mathcal{B}(\mathbb{X})\)-measurable.

The set of such stochastic kernels is denoted by \(\mathcal{Q}(\mathcal{Y}|\mathbb{X})\).

**Source Distribution.** The source is a collection of conditional probability distributions \(\{P_{X_n|X_{n-1}}(x_n | x_{n-1}) : n \in \mathbb{N}_0\}\), i.e., for each \(n \in \mathbb{N}_0\), \(P_{X_n|X_{n-1}}(\cdot | \cdot) \in \mathbb{Q}_n(\mathcal{X}_n|\mathcal{X}_{n-1})\). For \(A_t \in \mathcal{B}(X_t)\), \(t \in \mathbb{N}_0\), we can define the probability distribution \(P_{X_n}(A_0, n) = x_n^{t=0} A_t\) on \(\mathcal{B}(\mathcal{X}_0, n)\) by

\[
P_{X_n}(A_0, n) = \int_{A_0} P_{X_0}(dx_0) \cdots \int_{A_n} P_{X_n|X_{n-1}}(dx_n | x_{n-1}).
\]

Thus, for each \(n \in \mathbb{N}_0\), \(P_{X_n|X_{n-1}}(x_n | x_{n-1}) \in \mathcal{M}(\mathcal{X}_n)\), \(x_n \in \mathcal{X}_{0,n-1}\).

**Reproduction Distribution.** The reproduction distribution is specified by a collection of conditional probability distributions \(\{Q_{Y_n|Y_{n-1}, X_n}(\cdot | y_{n-1}, x_n) : n \in \mathbb{N}_0\}\), i.e., for each \(n \in \mathbb{N}_0\), \(Q_{Y_n|Y_{n-1}, X_n}(\cdot | \cdot, \cdot) \in \mathbb{Q}_n(\mathcal{Y}_n|\mathcal{Y}_{0,n-1} \times \mathcal{X}_n)\). For \(B_t \in \mathcal{B}(Y_t)\), \(t \in \mathbb{N}_0\), \(B_{0, n} = x_n^{t=0} B_t\), we define the family of conditional probability distributions \(Q_{Y_n|X_n}(C_{0, n}|x_n)\), on \(\mathcal{B}(\mathcal{Y}_0, n)\) (parametrized by \(x_n \in \mathcal{X}_0, n\)) by

\[
Q_{Y_n|X_n}(C_{0, n}|x_n) \triangleq \int_{B_0} Q_{Y_0|Y_{n-1}, X_n}(dy_0 | y_{n-1}, x_n) \cdots \int_{B_n} Q_{Y_n|Y_{n-1}, X_n}(dy_n | y_{n-1}, x_n).
\]

Thus, for each \(n \in \mathbb{N}_0\), \(Q_{Y_n|X_n}(\cdot | x_n) \in \mathcal{M}(\mathcal{Y}_0, n)\), \(x_n \in \mathcal{X}_0, n\).

Given \(P_{X_n}(\cdot) \in \mathcal{M}(\mathcal{X}_0, n)\) and \(Q_{Y_n|X_n}(\cdot | x_n) \in \mathcal{M}(\mathcal{Y}_0, n)\), and a fixed \(Y_{n-1} = y_{n-1}\) (we can also take the distribution \(P_{Y_{n-1}}(dy_{n-1})\) to be fixed), we define the joint distribution on \(\mathcal{X}_0 \times \mathcal{Y}_0\) by\(^3\)

\[
P_{X_n, Y_n}(A_0, n \times B_{0, n}) \triangleq (P_{X_n} \otimes Q_{Y_n|X_n})(\times_{t=0}^n (A_t \times B_t)) = \int_{A_0} P_{X_0}(dx_0) \cdots \int_{A_n} P_{X_n|X_{n-1}}(dx_n | x_{n-1}) \int_{B_{0, n}} Q_{Y_n|X_{n-1}, X_n}(dy_n | y_{n-1}, x_n),
\]

the marginal distribution on \(\mathcal{Y}_0, n\) by \(P_{Y_n}(B_{0, n}) \triangleq (P_{X_n} \otimes Q_{Y_n|X_n})(\times_{t=0}^n (A_t \times B_t))\) and the product probability distribution \(\prod_{0,n} : \mathcal{B}(\mathcal{X}_0, n) \otimes \mathcal{B}(\mathcal{Y}_0, n) \rightarrow [0, 1]\) by

\[
\prod_{0, n}(A_0, n \times B_{0, n}) \triangleq (P_{X_n} \times P_{Y_n})(\times_{t=0}^n (A_t \times B_t)) = \int_{A_0} P_{X_0}(dx_0) \int_{B_0} P_{Y_0|Y_{n-1}}(dy_0 | y_{n-1}) \cdots \int_{A_n} P_{X_n|X_{n-1}}(dx_n | x_{n-1}) \int_{B_n} P_{Y_n|Y_{n-1}}(dy_n | y_{n-1}).
\]

\(^3\)The notation \(\otimes\) denotes the compound probability distributions and the product space of \(\sigma\)-algebras.
Nonanticipative Rate Distortion Function (NRDF). The distortion function or fidelity of reproducing $x_t$ by $y_t$, $t = 0, 1, \ldots, n$, is a measurable function $d_{0,n} : X_{0,n} \times Y_{0,n} \rightarrow [0, \infty]$, $d_{0,n}(x^n, y^n) \triangleq \sum_{t=0}^n \rho_t(T^t x^n, T^t y^n)$, where the dependence on $T^t x^n \subseteq \{x_0, x_1, \ldots, x_t\}$, $T^t y^n \subseteq \{y_0, y_1, \ldots, y_t\}$, $t \in \mathbb{N}_0^n$ is either fixed or non-increasing with time. The fidelity set of reproductions is the set of conditional distributions given by (2.2) and satisfying the fidelity

$$\overline{Q}_{0,n}(D) \triangleq \left\{ \overline{Q}_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}(Y_{0,n}) : \frac{1}{n+1} \mathbb{E}\left\{ d_{0,n}(x^n, y^n) \right\} \leq D \right\}, \quad D \geq 0$$

where

$$\mathbb{E}\left\{ d_{0,n}(x^n, y^n) \right\} \triangleq \int_{X_{0,n} \times Y_{0,n}} d_{0,n}(x^n, y^n) P_{X^n,Y^n}(dx^n, dy^n).$$

The information theoretic measure associated with the FTH NRDF is a special case of directed information \[26,27\], defined via relative entropy

\begin{equation}
I_{P_X^n}(X^n \rightarrow Y^n) \triangleq \mathbb{D}(P_{X^n} \otimes \overline{Q}_{Y^n|X^n} \| \overline{P}_{0,n}) \tag{2.3}
\end{equation}

\begin{equation}
= \int_{X_{0,n} \times Y_{0,n}} \log \left( \frac{\overline{Q}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \right) P_{X^n,Y^n}(dx^n, dy^n) \tag{2.4}
\end{equation}

\begin{equation}
\equiv \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \overline{Q}_{Y^n|X^n}). \tag{2.5}
\end{equation}

In (2.5) the notation $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot, \cdot)$ indicates the functional dependence of $I_{P_X^n}(X^n \rightarrow Y^n)$ on $(P_{X^n}, \overline{Q}_{Y^n|X^n})$. It is shown in [28] that the set of distributions $\overline{Q}_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}(Y_{0,n})$ is convex, and $\mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \cdot)$ is a convex function of $\overline{Q}_{Y^n|X^n}(\cdot|x^n) \in \mathcal{M}(Y_{0,n})$.

Next, we give the precise definition of FTH NRDF.

**Definition 2.2. (FTH NRDF)** The FTH NRDF is defined by

\begin{equation}
R_{0,n}^{\alpha}(D) \triangleq \inf_{\overline{Q}_{Y^n|X^n} \in \overline{Q}_{0,n}(D)} \frac{1}{n+1} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \overline{Q}_{Y^n|X^n}). \tag{2.6}
\end{equation}

We note that FTH NRDF is an equivalent notion to Gorbunov and Pinsker’s definition of nonanticipatory $\epsilon$–entropy \[7, Section III.B\] and to the sequential RDF \[8\].

3. Optimal Nonstationary Reproduction Distribution. In this section, we describe the form of the optimal nonstationary (time-varying) reproduction distribution that achieves the infimum in (2.6) (the question of existence is addressed in [28]).

First, we state the following properties regarding the convexity and continuity of the FTH NRDF $R_{0,n}^{\alpha}(D)$, that are necessary for the development of our results.

1) $R_{0,n}^{\alpha}(D)$ is a convex, non-increasing function of $D \in [0, \infty)$.
2) If $R_{0,n}^{\alpha}(D) < \infty$, then $R_{0,n}^{\alpha}(D)$ is continuous on $[0, \infty)$.

Note that 1) is similar to the one derived in \[7\, Lemma IV.4\]. Also, for 2) recall that a bounded and convex function is continuous. Since $R_{0,n}^{\alpha}(D)$ is non-increasing, it is bounded outside the neighbourhood of $D = 0$ and it is also continuous on $[0, \infty)$. In other words, if $R_{0,n}^{\alpha}(D) < \infty$ then $R_{0,n}^{\alpha}(D)$ is bounded and hence continuous on $[0, \infty)$. Moreover, since $R_{0,n}^{\alpha}(D)$ is convex and non-increasing then its inverse function, $D(R_{0,n}^{\alpha})$, exists and it is convex, non-increasing function of $R_{0,n} \in [0, \infty)$. This is called Finite-Time Horizon (FTH) Nonanticipative Distortion Rate Function (NDRF) and is given
where the dual theorem [29, Theorem 1, pp. 224-225] as an unconstrained problem as follows:

\[
R_{0,n}^{\alpha}(D) = \sup_{s \leq 0} Q_{Y_0|X_0}^*(\cdot|\cdot) \in M(Y_0) \left\{ \frac{1}{n+1} \mathbb{I}_{X_n \rightarrow Y_n}(P_{X_0}^*, \hat{Q}_{Y_0|X_0}) - s \frac{1}{n+1} \mathbb{E}\left\{ d_{0,n}(x^n, y^n) \right\} \right\}.
\]

The FTH NRDF defined by (2.6) is a convex optimization problem, and thus, if there exists an interior point in the set \( \hat{Q}_{0,n}(D) \), it can be reformulated using Lagrange duality theorem [29, Theorem 1, pp. 224-225] as an unconstrained problem as follows:

\[
D(R_{0,n}^{\alpha}) = \inf_{t \in \mathbb{R}^+} t E_{X_0}(X_n \rightarrow Y_n) \leq R_{0,n}^{\alpha} \left\{ \frac{1}{n+1} \mathbb{E}\left\{ \|X_t - Y_t\|^2 \right\} \right\},
\]

The FTH NRDF, \( R_{0,n}^{\alpha}(D) \), of time-varying multidimensional Gauss-Markov processes.

**Theorem 3.1.** (Optimal nonstationary reproduction distributions)

Suppose there exists a \( \hat{Q}_{Y_0|X_0}^*(\cdot|\cdot) \in \hat{Q}_{0,n}(D) \), which solves (2.6), and that \( \mathbb{I}_{X_n \rightarrow Y_n} \) is Gâteaux differentiable in every direction of \( \{Q_{Y_t|Y_{t-1},X_t}(\cdot|y^{t-1}, x^t) : t \in \mathbb{N}_0 \} \) for a fixed \( P_{X_0}^*(\cdot) \in M(Y_0) \). Then, the following hold:

1. The optimal reproduction distributions denoted by \( \{Q_{Y_t|Y_{t-1},X_t}^*(\cdot|y^{t-1}, x^t) \in M(Y_t) : t \in \mathbb{N}_0 \} \) are given by the following recursive equations backwards in time:

\[
Q_{Y_t|Y_{t-1},X_t}^*(dy_t|y^{t-1}, x^t) = \frac{e^{sp_t(T^tx^n,T^ty^n)} P_{Y_t|Y_{t-1}}^*(dy_t|y^{t-1})}{\int_{Y_t} e^{sp_t(T^tx^n,T^ty^n)} P_{Y_t|Y_{t-1}}^*(dy_t|y^{t-1})}.
\]

For \( t = n \):

\[
Q_{Y_n|Y_{n-1},X_n}^*(dy_n|y^{n-1}, x^n) = \frac{e^{sp_n(T^tx^n,T^ty^n)} P_{Y_n|Y_{n-1}}^*(dy_n|y^{n-1})}{\int_{Y_n} e^{sp_n(T^tx^n,T^ty^n)} P_{Y_n|Y_{n-1}}^*(dy_n|y^{n-1})}.
\]

For \( t = n-1, n-2, \ldots, 0 \):

\[
Q_{Y_t|Y_{t-1},X_t}^*(dy_t|y^{t-1}, x^t) = \frac{e^{sp_t(T^tx^n,T^ty^n)} P_{Y_t|Y_{t-1}}^*(dy_t|y^{t-1})}{\int_{Y_t} e^{sp_t(T^tx^n,T^ty^n)} P_{Y_t|Y_{t-1}}^*(dy_t|y^{t-1})}.
\]

where \( s < 0 \), \( P_{Y_t|Y_{t-1}}^*(\cdot|y^{t-1}) \in M(Y_t) \) and \( g_{t,n}(x^t, y^t) \) is given by

\[
g_{t,n}(x^t, y^t) = -\int_{X_{t+1}} P_{X_{t+1}|X_t}(dx_{t+1}|x^t)
\]

\[
\log \left( \int_{Y_{t+1}} e^{sp_{t+1}(T^tx^{t+1}n,T^{t+1}y^n)} - g_{t+1,n}(x^{t+1}, y^{t+1}) \right) P_{Y_{t+1}|Y_t}^*(dy_{t+1}|y^t).
\]

(2) The FTH NRDF is given by

\[
R_{0,n}^{\alpha}(D) = sD - \frac{1}{n+1} \sum_{t=0}^{n} \int_{X_0 \times X_{0:t-1}} \left\{ \int_{Y_t} g_{t,n}(x^t, y^t) Q_{Y_t|Y_{t-1},X_t}^*(dy_t|y^{t-1}, x^t) \right\}
\]

\[
+ \frac{1}{n+1} \log \left( \int_{Y_t} e^{sp_t(T^tx^n,T^ty^n)} - g_{t,n}(x^t, y^t) P_{Y_t|Y_{t-1}}^*(dy_t|y^{t-1}) \right) \}
\]

\[
\otimes P_{X_t|X_{t-1}}^*(dx_{t-1}) \otimes P_{X_{t-1}|y^{t-1}}(dx_{t-1}, dy_{t-1}).
\]
If \( R_{0,n}^a(D) > 0 \) then \( s < 0 \), and

\[
\frac{1}{n+1} \sum_{t=0}^{n} \int_{X_{0,t} \times Y_{0,t}} \rho_t(T^tx^n, T^ty^n) P_{X_t,Y_t} = D.
\]

**Proof.** The sequence of minimizations over \( \{ Q_{Y_t|Y_{t-1},X_t} : t \in \mathbb{N}_0 \} \) corresponds to a nested optimization problem. Hence, we can introduce the dynamic programming recursive equations. Then, we carry out the infimum starting at the last stage over \( Q_{Y_n|Y_{n-1},X_n}(|y^{n-1}, x^n) \in \mathcal{M}(Q_n) \) and sequentially move backward in time to determine \( Q_{Y_n|Y_{n-1},X_n}^*, Q_{Y_{n-1}|Y_{n-2},X_{n-1}}^*, \ldots, Q_{Y_0|X_0}^* \). The procedure is lengthy and tedious, hence it is omitted due to space limitations. \( \square \)

From the above theorem, for a given distribution \( P_{X_0}(\cdot) \in \mathcal{M}(X_{0,n}) \), we can identify the dependence of the optimal nonstationary reproduction distribution on past and present symbols of the information process \( \{ X_t : t \in \mathbb{N}_0 \} \), but not its dependence on past reproduction symbols. In what follows, we give certain properties of the information structure of the optimal nonstationary reproduction distribution that achieves the infimum in (2.6).

**Information structure of the optimal nonstationary reproduction distribution.**

(1) The dependence of \( Q_{Y_n|Y_{n-1},X_n}^*(dy_n|y^{n-1}, x^n) \) on \( x^n \) in \( X_{0,n} \) is determined by the dependence of \( \rho_t(T^tx^n, T^ty^n) \) on \( x^n \) in \( X_{0,n} \) as follows:

(1.1) If \( \rho_t(T^tx^n, T^ty^n) = \bar{\rho}(x_n, y^n) \) then \( Q_{Y_n|Y_{n-1},X_n}^*(dy_n|y^{n-1}, x^n) = Q_{Y_{n-1}|Y_{n-2},X_{n-1}}^*(dy_{n-1}|y^{n-2}, x_{n-1}) \) \( \vdots \) \( Q_{Y_1|X_1}^*(dy_1|y^0, x_1) \) \( \Rightarrow \) \( Q_{Y_n|Y_{n-1},X_n}^*(dy_n|y^{n-1}, x^n) = \bar{\rho}(x_n, y^n) \).

(2.2) If \( g_{t,n}(x^n, y^n) = g_{t,n}(x^n, y^{n-1}) \), \( t \in \mathbb{N}_0 \) optimal reproduction distribution (3.4) reduces to

\[
Q_{Y_t|Y_{t-1},X_t}^*(dy_t|y^{t-1}, x_t) = \frac{e^{s\rho_t(T^tx^n, T^ty^n)} \rho_t(T^tx^n, T^ty^n) P_{X_t,Y_t}^*}{\int_{Y_t} e^{s\rho_t(T^tx^n, T^ty^n)} P_{X_t,Y_t}^*} (dy_t|y^{t-1}, x_t).
\]

To further understand the dependence of the optimal nonstationary reproduction distributions (3.3), (3.4) on past reproductions, we state an alternative characterization of the nonstationary solution of \( R_{0,n}^a(D) \), as a maximization over a certain class of functions. We use this additional characterization to derive lower bounds on \( R_{0,n}^a(D) \), which are achievable.

**Theorem 3.2. (Characterization of solution of \( R_{0,n}^a(D) \))**

An alternative characterization of \( R_{0,n}^a(D) \) is

\[
R_{0,n}^a(D) = \sup_{s \leq 0} \sup_{\{ x_t : t \in \mathbb{N}_0 \}} \left\{ sD - \frac{1}{n+1} \sum_{t=0}^{n} \int_{X_{0,t} \times Y_{0,t}} g_{t,n}(x^n, y^n) Q_{Y_t|Y_{t-1},X_t}^*(dy_t|y^{t-1}, x_t) \right. \\
+ \log \left( \lambda_t(x^n, y^{t-1}) P_{X_t|X_{t-1}}(dx_t|x^{t-1}) \otimes P_{X_{t-1},Y_{t-1}}(dx^{t-1}, dy^{t-1}) \right) \left. \right\},
\]

(3.8)
where
\[
\Psi^t_s \triangleq \left\{ \lambda_t(x^t, y^{t-1}) \geq 0 : \int_{X_0, t-1} \left( \int_{X_t} e^{s \rho_t(T^t x^n, T^t y^n) - g_{n,n}(x^n, y^n)} \lambda_t(x^t, y^{t-1}) P_{X_t|X_{t-1}}(dx_t|x^{t-1}) \right) \otimes P_{X_{t-1}|Y_{t-1}}(dx_{t-1}|y^{t-1}) \leq 1 \right\}
\]
(3.9)

and \( g_{n,n}(x^n, y^n) = 0 \), and for \( t \in N_{n-1} \),
\[
g_{t,n}(x^t, y^t) = -\int_{X_t+1} P_{X_{t+1}|X^t}(dx_{t+1}|x^t) \log \left( \lambda_{t+1}(x^{t+1}, y^t) \right)^{-1}.
\]

For \( s \in (-\infty, 0] \) a necessary and sufficient condition for \( \{ \lambda_t(\cdot, \cdot) : t = 0, \ldots, n \} \) to achieve the supremum of (3.8) is the existence of a probability distribution \( P_{Y_t|Y_{t-1}}(\cdot|y^{t-1}) \in M(\lambda_t) \) such that
\[
\lambda_t(x^t, y^{t-1}) = \left( \int_{Y_t} e^{s \rho_t(T^t x^n, T^t y^n) - g_{n,n}(x^n, y^n)} P_{Y_t|Y_{t-1}}(dy_t|y^{t-1}) \right)^{-1}, \quad t \in N_0^n.
\]

Proof. See Appendix A.

Without the above characterization it will be very difficult to compute exactly \( R^\alpha_{0,n}(D) \) for a given source (with memory), simply because to solve a rate distortion problem explicitly, one needs to identify the dependence of the optimal reproduction distribution on past reproduction symbols, and in general to find the information structure of the optimal reproduction distribution.

In the next section, we use Theorem 3.2, to derive \( R^\alpha_{0,n}(D) \).

4. FTH NRDF of Time-Varying Multidimensional Gauss-Markov Processes. In this section, we derive the FTH NRDF of time-varying multidimensional Gauss-Markov processes in state-space form, by applying Theorem 3.1 and Theorem 3.2 from Section 3. We show the following.

(1) the analytical expression of the optimal nonstationary reproduction distribution that achieves the infimum of the FTH NRDF and the analytical expression of the FTH NRDF subject to a square error distortion;
(2) a realization of the optimal nonstationary reproduction distribution in the sense of Fig. 1.3 that allows us to obtain the optimal filter;
(3) a universal lower bound on the mean square error of any causal estimator of Gaussian processes based on the closed form expression of FTH NRDF.

The analytical expression of the FTH NRDF is found by developing a time-space algorithm, which is a generalization of the standard reverse waterfilling algorithm derived in [15, Section 10.3.3] for independent Gaussian Random Variables (RV). Toward this, illustrative examples that verify our theory are presented.

Note that, unlike [9], there is an additional complexity here, because we deal with time-varying random processes with memory, leading to a time-space reverse-waterfilling algorithm.

First, we give the optimal nonstationary reproduction distribution that achieves the infimum of the FTH RDF for the time-varying multidimensional Gauss-Markov processes. Next, we realize the optimal nonstationary reproduction distribution with
an encoder, an Additive White Gaussian Noise (AWGN) channel and a decoder part of which is the causal filter. This procedure is illustrated in Fig. 4.1.

**Time-Varying multidimensional Gaussian source.** Consider a time-varying $p$-dimensional Gaussian source described in state space form by

\[ X_{t+1} = A_t X_t + B_t W_t, \quad X_0 = x_0, \quad t \in \mathbb{N}^{n-1}, \]

where $A_t \in \mathbb{R}^{p \times p}$, $B_t \in \mathbb{R}^{p \times k}$, $t \in \mathbb{N}_0^{n-1}$. We assume

\[ (G1) \quad X_0 \in \mathbb{R}^p \) is Gaussian $N(0; \Sigma_0); \]

\[ (G2) \quad \{W_t : t \in \mathbb{N}_0^n\} \) is a $k$-dimensional IID Gaussian $N(0; I_k)$ sequence, independent of $X_0$;

\[ (G3) \quad \text{The distortion function is single letter defined by} \quad d_{0,n}(x^n, y^n) \triangleq \sum_{t=0}^n \rho_t(T^t x^n, T^t y^n) = \sum_{t=0}^n ||x_t - y_t||^2. \]

Information Structure. Recall that by Theorem 3.1 and the Markovian property of (4.1), the optimal nonstationary reproduction distribution given by (3.3)-(3.4) is Markov with respect to the input source symbols, i.e., $\{Q^{*}_{Y_t} | Y_{t-1}, X_t \} \triangleq Q^{*}_{Y_t} | Y_{t-1}, X_t, (dy_t | y^{t-1}, x_t) : t \in \mathbb{N}_0^n\}$ (see the comments below Theorem 3.1 on information structures of optimal reproduction distribution).

Next, by starting from stage $n$ and going backwards, we can show that $\{Q^{*}_{Y_t} | Y_{t-1}, X_t \} \triangleq (dy_t | y^{t-1}, x_t) : t \in \mathbb{N}_0^n\}$ are conditional Gaussian distributions.

Stage $n$. Since the exponential term $||y_n - x_n||^2$ in the RHS of (3.3) is quadratic in $(x_n, y_n)$, and $\{X_t : t \in \mathbb{N}_0^n\}$ is Gaussian, then it follows that a Gaussian distribution $Q_{Y_{n-1} | Y_{n-2}, X_{n-1}}(\cdot | y^{n-1}, x_{n-1})$ for a fixed realization of $(y^{n-1}, x_{n-1})$, and a Gaussian distribution $P_{Y_{n-1} | Y_{n-2}, X_{n-1}}(\cdot | y^{n-1})$ satisfy both the left and right sides of (3.3). This implies that $Q_{Y_{n-1} | Y_{n-2}, X_{n-1}}(\cdot | y^{n-1}, x_{n-1})$ and $P_{Y_{n-1} | Y_{n-2}, X_{n-1}}(\cdot | y^{n-1})$ are both Gaussian for fixed $(y^{n-1}, x_{n-1})$ and $y^{n-1}$.

Stages $t \in \{n-1, n-2, \ldots, 1, 0\}$. By (3.4), evaluated at $t = n-1$, then $g_{n-1,n}(x_{n-1}, y^{n-1})$ will include terms of quadratic form in $x_{n-1}$ and nonlinear in $y^{n-1}$. Repeating this argument recursively, it can be verified that at any time $t \in \mathbb{N}_0^{n-1}$, the optimal reproduction distribution $Q_{Y_t | Y_{t-1}, X_t}(\cdot | y^{t-1}, x_t)$ is conditionally Gaussian with respect to $(x_t, y^{t-1})$, $t \in \mathbb{N}_0^{n-1}$.

Since the optimal reproduction distributions are conditionally Gaussian, then they can be realized using a general equation of the form

\[ Y_t = A_t X_t + \bar{B}_t Y^{t-1} + V^c_t, \quad t \in \mathbb{N}_0^n, \]

where $\bar{A}_t \in \mathbb{R}^{p \times p}$, $\bar{B}_t \in \mathbb{R}^{p \times p}$, and $\{V^c_t : t \in \mathbb{N}_0^n\}$ is an independent sequence of Gaussian vectors $\{N(0; Q_t) : t \in \mathbb{N}_0^n\}$.

Next, we simplify the computation by introducing the following preprocessing at the encoder and decoder associated with channel (4.2) (as shown in Fig. 4.1).

**Preprocessing at Encoder.** Introduce (i) the estimation error $\{K_t : t \in \mathbb{N}_0^n\}$ of $\{X_t : t \in \mathbb{N}_0^n\}$ based on $\{Y_0, \ldots, Y_{t-1}\}$, and (ii) its covariance $\{\Pi_t : t \in \mathbb{N}_0^n\}$, defined by

\[ (4.3) \quad K_t \triangleq X_t - \bar{X}_{t|t-1}, \quad \bar{X}_{t|t-1} \triangleq \mathbb{E}\{X_t | \sigma(Y^{t-1})\}, \quad \Pi_t \triangleq \mathbb{E}\{K_t K_t^T\}, \quad t \in \mathbb{N}_0^n, \]

where $\sigma(Y^{t-1})$ is the $\sigma$-algebra (observable events) generated by the sequence $\{Y^{t-1}\}$. The covariance is diagonalized by introducing a unitary transformation $\{E_t : t \in \mathbb{N}_0^n\}$ such that

\[ (4.4) \quad E_t \Pi_t E_t^T = \Lambda_t, \quad \text{where} \quad \Lambda_t \triangleq diag\{\lambda_{t,1}, \ldots, \lambda_{t,p}\}, \quad t \in \mathbb{N}_0^n. \]
To facilitate the computation, we introduce the scaling process \( \{ \Gamma_t : t \in \mathbb{N}_0 \} \), where \( \Gamma_t \triangleq E_t K_t, \ t \in \mathbb{N}_0 \), has independent Gaussian components but all of the components are correlated.

**Preprocessing at Decoder.** Analogously, we introduce the error process \( \{ \tilde{K}_t : t \in \mathbb{N}_0 \} \) and the scaling process \( \{ \tilde{\Gamma}_t : t \in \mathbb{N}_0 \} \) defined by

\[
(4.5) \quad \tilde{K}_t \triangleq Y_t - \tilde{X}_{t(t-1)}, \ \text{and} \ \tilde{\Gamma}_t \triangleq \Phi_t (\Theta_t E_t K_t + V_t^\gamma), \ t \in \mathbb{N}_0.
\]

Fig. 4.1: Realization of the optimal nonstationary reproduction distribution of multidimensional Gaussian process.

The square error fidelity criterion \( d_{0,n}(\cdot,\cdot) \) is not affected by the above processing of \( \{(X_t,Y_t) : t \in \mathbb{N}_0 \} \), since the preprocessing at both the encoder and decoder does not affect the form of the squared error distortion function, that is,

\[
(4.6) \quad d_{0,n}(X^n,Y^n) = d_{0,n}(K^n,\tilde{K}^n) = \frac{1}{n+1} \sum_{t=0}^{n} ||\tilde{K}_t - K_t||^2
\]

Using basic properties of conditional entropy, it can be shown that the following expressions are equivalent.

\[
R_{0,n}^{n,n}(D) = R_{0,n}^{n,K^n,\tilde{K}^n}(D) \triangleq \inf_{\{P_{K_t|\tilde{K}_{t-1},\tilde{K}_t} : t=0,\ldots,n\}} \sum_{t=0}^{n} I(K_t;\tilde{K}_t|\tilde{K}^{t-1})
\]

\[
(4.7) \quad = R_{0,n}^{n,K^n,\tilde{K}^n}(D) \triangleq \inf_{\{P_{K_t|\tilde{K}_{t-1},\tilde{K}_t} : t=0,\ldots,n\}} \sum_{t=0}^{n} I(K_t;\tilde{K}_t|\tilde{K}^{t-1}).
\]

Next, we derive the main theorem which gives the closed form expression of the FTH NRDF for multidimensional Gaussian processes \( (4.4) \) by considering the realization shown in Fig. 4.1, where \( \{ V_t^\gamma : t \in \mathbb{N}_0 \} \) is Gaussian \( \{ N(0,Q_t) : t \in \mathbb{N}_0 \} \), and \( \{ \Theta_t, \Phi_t : t \in \mathbb{N}_0 \} \) are the matching matrices to be determined.

**Theorem 4.1.** \( (R_{0,n}^{n,n}(D) \text{ of time-varying multidimensional Gauss-Markov processes}) \)
The FTH NRDF, \(R^{na}_{0,n}(D)\), of the Gauss-Markov process (4.1), is given by

\[
R^{na}_{0,n}(D) = \frac{1}{2n+1} \sum_{i=0}^{n} \sum_{i=1}^{P} \log \left( \frac{\lambda_{t,i}}{\delta_{t,i}} \right), \quad \delta_{t,i} \leq \lambda_{t,i}, \ t \in \mathbb{N}_0^n, \ i = 1, \ldots, p,
\]

where \(\Lambda_t \triangleq \text{diag}\{\lambda_{t,1}, \ldots, \lambda_{t,p}\} = E_t \Pi_t E_t^T\).

\[
\Pi_t \triangleq E\left\{ \left( X_t - \mathbb{E}\{X_t|\sigma\{Y^{t-1}\}\} \right) \left( X_t - \mathbb{E}\{X_t|\sigma\{Y^{t-1}\}\} \right)^T \right\}
\]

\[
\delta_{t,i} \triangleq \begin{cases} 
\xi, & \text{if } \xi \leq \lambda_{t,i}, \\
\lambda_{t,i}, & \text{if } \xi > \lambda_{t,i}, \ t \in \mathbb{N}_0^n, \ i = 1, \ldots, p
\end{cases}
\]

and \(\xi\) is chosen such that \(\frac{1}{n+1} \sum_{i=0}^{n} \sum_{i=1}^{p} \delta_{t,i} = D\). Moreover, a realization of the optimal time-varying (nonstationary) reproduction distribution \(\{P^*_t|Y^{t-1},X_t(dy_t)|Y^{t-1},x_t) : t \in \mathbb{N}_0^n\}\) is shown in Fig. 4.1 and is given by

\[
Y_t = E_t^T H_t E_t (X_t - \hat{X}_{t|t-1}) + E_t^T \Phi_t V_t^c + \hat{X}_{t|t-1}, \ H_t \triangleq \Phi_t \Theta_t,
\]

\[
\Pi_t \triangleq \mathbb{E}\{ (X_t - \mathbb{E}\{X_t|\sigma\{Y^{t-1}\}\}) (X_t - \mathbb{E}\{X_t|\sigma\{Y^{t-1}\}\})^T \}, \ \delta_{t,i} \triangleq \begin{cases} 
\xi, & \text{if } \xi \leq \lambda_{t,i}, \\
\lambda_{t,i}, & \text{if } \xi > \lambda_{t,i}, \ t \in \mathbb{N}_0^n, \ i = 1, \ldots, p
\end{cases}
\]

where the error \(X_t - \mathbb{E}\{X_t|Y^{t-1}\}\) is Gaussian \(N(0; \Pi_t)\), \(\hat{X}_{t|t-1} \triangleq \mathbb{E}\{X_t|Y^{t-1}\}\), and \(\Pi_t\) are given by the Kalman filter equations

\[
\hat{X}_{t+1|t} = A_t \hat{X}_{t|t-1} + A_t \Pi_t (E_t^T H_t E_t)^\top M_t^{-1} (Y_t - \hat{X}_{t|t-1}), \ \hat{X}_{0|0} = \mathbb{E}\{X_0|Y^{0-1}\}, \ t \in \mathbb{N}_0^n
\]

\[
\Pi_{t+1} = A_t \Pi_t A_t^\top - A_t \Pi_t (E_t^T H_t E_t)^\top M_t^{-1} (E_t^T H_t E_t) \Pi_t A_t^\top + B_t B_t^\top, \ \Pi_0 = \bar{\Pi}_0, \ t \in \mathbb{N}_0^n
\]

\[
M_t = E_t^T (H_t E_t)^\top \Pi_t + M_t (E_t^T H_t E_t)^\top E_t^T, \ t \in \mathbb{N}_0^n
\]

where

\[
\eta_{t,i} = 1 - \frac{\delta_{t,i}}{\lambda_{t,i}}, \ H_t \triangleq \text{diag}\{\eta_{t,1}, \ldots, \eta_{t,p}\}, \ \Delta_t = \text{diag}\{\delta_{t,1}, \ldots, \delta_{t,p}\}, \ t \in \mathbb{N}_0^n, \ i = 1, \ldots, p,
\]

\[
Q_t \triangleq \text{Cov}(V_t^c), \ \Phi_t \triangleq \sqrt{H_t \Delta_t Q_t^{-1}}, \ t \in \mathbb{N}_0^n.
\]

In addition, the processes \(\{Y_t : t \in \mathbb{N}_0^n\}\), \(\{K_t : t \in \mathbb{N}_0^n\}\), and \(\{\Upsilon_t : t \in \mathbb{N}_0^n\}\) generate the same information, i.e., \(\sigma\{Y^t\} = \sigma\{K^t\} = \sigma\{\Upsilon^t\}, t \in \mathbb{N}_0^n\).

Proof. See Appendix B. □

Remark 1. (Comments on Theorem 4.1)
(1) The main feature of Theorem 4.1 is the time-space reverse-waterfilling property \([4.8]-[4.11]\), which states that if the reproduction error \(\delta_{t,i}\) is above the eigenvalue \(\lambda_{t,i}\) of the error covariance \(\Pi_{t}\), then the time-space component \(X_{t,i}\) is not reconstructed by \(Y_{t,i}\) for \(t\in\mathbb{N}_{0}\), \(i=1,\ldots,p\). The behavior of \(\delta_{t,i}\) is described by the reverse-waterfilling expression \((4.11)\), and the level \(\xi\) depends on \(D\), i.e., the overall fidelity of the error.

(2) Theorem 4.1 states that \(\{Y_t : t \in \mathbb{N}_0\}\) is the compressed version of \(\{X_t : t \in \mathbb{N}_0\}\), and the process \(\{\tilde{X}_{t+1|t} : t \in \mathbb{N}_0\}\), given by \((4.14)-(4.16)\), is the estimator of \(\{X_t : t \in \mathbb{N}\}\) based on the compressed data \(\{Y_t : t \in \mathbb{N}_0\}\). In addition, the time-space reverse-waterfilling is part of the estimation algorithm. This is a variant of the Kalman filter.

The following remark, is a direct consequence of Theorem 4.1, and illustrates the connection between \(R_{0,n}^{a}(D)\) and \(D(R_{0,n}^{a})\) given by \((4.1)\).

Remark 2. (The solution of \(D(R_{0,n}^{a})\) for time-varying multidimensional Gauss-Markov processes)

From Theorem 4.1 the FTH NRDF of the Gaussian process \((4.1)\) is given by

\[
R_{0,n}^{a}(D) = \frac{1}{2} \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \log \left\{ \max \left( 1, \frac{\lambda_{t,i}}{\delta_{t,i}} \right) \right\} \equiv \frac{1}{2} \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} R_{t,i}^{a}(\delta_{t,i})
\]

where \((a)\) follows if we let

\[
R_{t,i}^{a}(\delta_{t,i}) \triangleq \frac{1}{2} \log \left\{ \max \left( 1, \frac{\lambda_{t,i}}{\delta_{t,i}} \right) \right\}, \ t \in \mathbb{N}_0, \ i = 1,\ldots,p.
\]

Therefore, we can determine the FTH NDRF via \((4.19)\) by solving with respect to the time-space component \(\delta_{t,i}\). Toward this, we obtain

\[
\delta_{t,i} = \lambda_{t,i} e^{-2R_{t,i}^{a}}, \ t \in \mathbb{N}_0, \ i = 1,\ldots,p.
\]

Moreover, by Theorem 4.1 we have that

\[
D = \frac{1}{n+1} \sum_{t=0}^{n} \delta_{t} = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i}, \text{ where } \delta_{t} = \sum_{i=1}^{p} \delta_{t,i}.
\]

Substituting \((4.21)\) into \((4.22)\) we obtain

\[
D(R_{0,n}^{a}) = \frac{1}{n+1} \sum_{t=0}^{n} \delta_{t} = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \lambda_{t,i} e^{-2R_{t,i}^{a}}.
\]

Next, we utilize the closed form expressions of the FTH NRDF and FTH NDRF evaluated for time-varying multidimensional Gauss-Markov processes to derive a lower bound on the mean square error given in terms of directed information \(I_{P_X^n(X^n \rightarrow Y^n)}\).

Theorem 4.2. (Universal lower bound on mean square error)

Let \(\{X_t : t \in \mathbb{N}\}\) be the multidimensional Gauss-Markov process given by \((4.1)\) and

\[
\text{(4.21)} \quad X_{t,i} \text{ is the time-space component of the vector process } \{X_t : t \in \mathbb{N}_0\},
\]

\[
\text{(4.23)} \quad Y_{t,i} \text{ is the time-space component of the vector process } \{Y_t : t \in \mathbb{N}_0\}.
\]
let \( \{\tilde{Y}_t : t \in \mathbb{N}_0^p \} \) be any estimator (not necessarily Gaussian) of \( \{X_t : t \in \mathbb{N}_0^p \} \). The mean square error is bounded below by

\[
\frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E} \left\{ \|X_t - \tilde{Y}_t\|_2^2 \right\} \geq \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \lambda_{t,i} e^{-2I(X_t;i; \tilde{Y}_t,i; \tilde{Y}^{t-1,i})}
\]

**Proof.** Let \( D = \frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E} \left\{ \|X_t - \tilde{Y}_t\|_2^2 \right\} \) where

\[
\mathbb{E} \left\{ \|X_t - \tilde{Y}_t\|_2^2 \right\} = \sum_{i=1}^{p} \delta_{t,i} \triangleq \sum_{i=1}^{p} \mathbb{E} \left\{ \|X_{t,i} - \tilde{Y}_{t,i}\|_2^2 \right\} \text{ with } D \in [0, \infty).
\]

Since, in general, \( R_{t,i}^{na} \leq I(X_{t,i}; \tilde{Y}_{t,i}; \tilde{Y}^{t-1,i}) \), \( t \in \mathbb{N}_0^p \), \( i = 1, \ldots, p \), then by Remark 2 (4.23), we obtain

\[
\frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E} \left\{ \|X_t - \tilde{Y}_t\|_2^2 \right\} = D(R_{0,na}^{na}) = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \lambda_{t,i} e^{-2R_{t,i}^{na}}
\]

\[
\geq \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \lambda_{t,i} e^{-2I(X_{t,i}; \tilde{Y}_{t,i}; \tilde{Y}^{t-1,i})},
\]

which is the desired result. This completes the proof.\[\Box\]

Notice that from Remark 1 (2), if we substitute \( \tilde{Y}_t = \hat{X}_{t|t-1} \) in Theorem 4.2 then we have the lower bound (4.24).

In the next remark, we discuss existing degenerated versions of our lower bound which illustrate its generality and we draw the relation of Theorems 4.1 and 4.2 to the existing literature.

**Remark 3. **(Relations to existing results)

(a) [24] Theorem 5.8.1, [30] Let \( X = (X_1, \ldots, X_p) \) be a \( p \)-dimensional Gaussian vector with distribution \( X \sim N(0; \Gamma_X) \) and \( Y = (Y_1, \ldots, Y_p) \) be its reproduction vector. Then, for any \( D > 0 \),

\[
R(D) \triangleq \inf_{Q_{Y|X}(dy|x) : \mathbb{E}[||X-Y||_2^2] \leq D} I(X;Y) = \frac{1}{2} \sum_{i=1}^{p} \log \left\{ \max \left( 1, \frac{\lambda_i}{\xi} \right) \right\}
\]

where \( \{\lambda_i : i = 1, \ldots, p\} \) are the eigenvalues of \( \Gamma_X \) and \( \xi > 0 \) is a constant uniquely determined by \( \sum_{i=1}^{p} \min \{\lambda_i, \xi\} = D \). Note that the solution of classical RDF in (4.26) is based on reverse-waterfilling method (see [24] Lemma 5.8.2). The above results are also obtained from Theorem 4.1 if we assume model (4.1) generates an independent and identically distributed sequence \( \{X_t : t \in \mathbb{N}_0^p\} \) (by setting \( A_t = 0, B_t = I \)). In this case, \( \hat{X}_{t|t-1} = E X_t = 0 \) and \( \Pi_t = E X_t X_t^T = \Gamma_X \).

(b) Assume \( X \sim N(0; \sigma_X^2 I) \). Then it is known that

\[
R(D) = \min_{Q_{Y|X}(dy|x) : \mathbb{E}[||X-Y||_2^2] \leq D} I(X;Y) = \frac{1}{2} \log \left\{ \max \left( 1, \frac{\sigma_X^2}{D} \right) \right\}, D \geq 0,
\]

\[
D(R) = \min_{Q_{Y|X}(dy|x) : I(X;Y) \leq R} \mathbb{E}[||X-Y||_2^2] \sigma_X^2 e^{-2R}.
\]

The realization scheme to achieve the classical RDF or the distortion rate function is the following.

\[
Y = \left(1 - \frac{D}{\sigma_X^2}\right) X + V, V \sim N \left(0; D(1 - \frac{D}{\sigma_X^2})\right).
\]
This result can be found in [21, Theorem 1.8.7]. Note that (4.27) is a degenerated version of (4.12) assuming the model (4.1) generates independent and identically distributed sequence \( \{X_t: t \in \mathbb{N}_0^n\} \) as in (a), and the connection to Theorem 4.1 is established by setting \( E_t = 1, H_t = 1 - \frac{D}{\sigma_X^2}, X_{t|t-1} = 0, \Phi_t = H_tD \) and \( V_t \sim N(0; 1) \).

(c) (Lower bound on mean square error [21, 1.8.8], [22]) Given a Gaussian RV \( X \sim N(0; \sigma_X^2) \), then for any real valued RV \( \hat{Y} \) (not necessarily Gaussian) the mean square error is bounded below by

\[
\mathbb{E}\|X - \hat{Y}\|^2 \geq \sigma_X^2 e^{-2I(X; \bar{Y})}. \tag{4.28}
\]

The RDF of the Gaussian RV \( X \sim N(0; \sigma_X^2) \) and the lower bound in (4.28), are utilized in [21, 22] to derive optimal coding and decoding schemes for transmitting a Gaussian message \( \theta \sim N(0; \sigma_\theta^2) \) over an AWGN channel with feedback, \( Y_t = X_t(\theta, Y_{t-1}) + V_t^c, t \in \mathbb{N}_0^n \), where \( \{V_t^c: t \in \mathbb{N}_0^n\} \) is IID Gaussian process. Although we do not pursue such problems in this paper, we note that Theorems 4.1 and 4.2 are necessary in order to derive optimal coding schemes for additive Gaussian channels with memory (including additive Gaussian memoryless channels).

4.1. Examples. In what follows, we demonstrate examples where we numerically compute the FTH NRDF of time-varying Gauss-Markov processes using Theorem 4.1. For these examples, the utility of the reverse waterfilling algorithm is necessary even when the process elements are scalar (i.e., \( p = 1 \)). For process elements in higher dimensions (i.e., \( p \geq 2 \)), the complexity of the problem increases, since the reverse waterfilling algorithm must be solved both in time and space units. We overcome this obstacle by proposing an iterative algorithmic technique that allocates information of the Gaussian process and distortion levels optimally.

Remark 4. (Relations to existing results)

Note that the examples presented here are fundamentally different from the examples discussed in [7, Section IV.C] because here we deal with time-space aspects of the reverse waterfilling algorithm, while in [7, Section IV.C] there is only the space aspect of the reverse waterfilling algorithm. For the case of time-invariant Gauss-Markov processes, a closed form expression is given when the process elements are scalar and standard reverse waterfilling [17, Section 10.3.3.] is used when the process elements are vectors, in the same way as it is done for parallel independent sources. The fundamental difference between stationary and non-stationary Gauss-Markov processes is that the latter, for both scalar and vector cases requires an iterative approach to solve the problem, where in each iteration a new time-step is included.

Example 1. Consider the following two-dimensional Gauss-Markov process

\[
\begin{pmatrix}
X_{t+1,1} \\
X_{t+1,2}
\end{pmatrix}
= \begin{bmatrix}
-\alpha_t & 1 \\
-\beta_t & 0
\end{bmatrix}
\begin{pmatrix}
X_{t,1} \\
X_{t,2}
\end{pmatrix}
+ \begin{bmatrix}
\sigma_{W_{t,1}} & 0 \\
0 & \sigma_{W_{t,2}}
\end{bmatrix}
\begin{pmatrix}
W_{t,1} \\
W_{t,2}
\end{pmatrix}
\quad t = 0, 1, 2, \quad i = 1, 2,
\]

where \( W_{t,i} \sim N(0; 1) \), \( \sigma_{W_{t,1}}, \sigma_{W_{t,2}} \sim N(0; \sigma_{W_{t,i}}^2) \) and \( \{A_t, B_t\} \) are time varying matrices.

This example corresponds to (4.1) for \( p = k = n = 2 \). For this example, we choose
Algorithm 1 Rate distortion allocation algorithm: The vector case

Initialize:
The number of time-steps \( n \); the number of channels \( p \) the distortion level \( D \); the error tolerance \( \epsilon \); the initial covariance matrix \( \Pi_0 \) of the error process \( K_0 \), the state-space matrices \( A_t \) and \( B_t \) of the time-varying multidimensional Gauss-Markov process \( X_t \) given by (4.1).

Set \( \xi = D \); flag = 0.

while flag = 0 do
  Compute \( \delta_{t,i} \) \( \forall t,i \) as follows:
  for \( t = 0 : n \) do
    Perform Singular Value Decomposition: \( [E_t, \Lambda_t] = \text{SVD}(\Pi_t) \)
    \( \Delta_t \) is computed according to (4.11).
    Use \( A_t, B_t \) and \( \Delta_t \) to compute \( \Pi_{t+1} \) according to (4.15).
  end for
  if \( \left| \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i} - D \right| \leq \epsilon \) then
    flag ← 1
  else
    Re-adjust \( \xi \) as follows:
    \( \xi \leftarrow \xi + \beta(\frac{D}{\Pi_{t+1}} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i}) \), where \( \beta \in (0, 1] \) is a proportionality gain and affects the rate of convergence.
  end if
end while

the distortion level \( D = 3 \) and consider the following matrices \( \{A_t, B_t\} \):

\[
A_0 = \begin{bmatrix} -0.5 & 1 \\ -0.4 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},
A_1 = \begin{bmatrix} -0.4 & 1 \\ -0.5 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 1.4 \end{bmatrix},
A_2 = \begin{bmatrix} -0.9 & 1 \\ -0.5 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1.2 & 0 \\ 0 & 1.3 \end{bmatrix}.
\]

The initial covariance matrix of the error process \( K_t \) is

\[
\Pi_0 = \begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{bmatrix}.
\]

Recall that the covariance matrix of the error process \( K_t \) given by (4.15) is simplified to

\[
(4.30) \quad \Pi_{t+1} = A_t E_t \{ \text{diag}\{\delta_{t,1}, \delta_{t,2}\} \} E_t A_t^T + B_t B_t^T, \quad t = 0, 1, 2, \quad \Pi_0 = \Pi_0
\]

and \( \delta_{t,i} \) given by (4.11) becomes

\[
(4.31) \quad \delta_{t,i} = \min\{\lambda_{t,1}, \xi\}, \quad t = 0, 1, 2, \quad i = 1, 2.
\]

Now let’s implement Algorithm [4] for error tolerance \( \epsilon = 10^{-3} \). We choose an initial \( \xi = \xi_0 \) to start our iterations. A good starting point is \( \xi_0 = D \). For \( \Pi_0 \) we perform
Singular Value Decomposition (SVD) and we obtain the unitary matrix

$$E_0 = \begin{bmatrix}
-0.7882 & -0.6154 \\
-0.6154 & 0.7882
\end{bmatrix}$$

and the eigenvalues in a diagonal matrix that correspond to the levels of the noise $\lambda_{0,1}$ and $\lambda_{0,2}$, i.e.,

$$\Lambda_0 = \begin{bmatrix}
0.7562 & 0 \\
0 & 0.3438
\end{bmatrix}.$$  

For $\xi = \xi_0 = D = 3$ and $(\lambda_{0,1}, \lambda_{0,2}) = (0.76, 0.34)$ we compute $\Delta_0$ using (4.31). Hence,

$$\Delta_0 = \Lambda_0 = \begin{bmatrix}
0.7562 & 0 \\
0 & 0.3438
\end{bmatrix}.$$  

Using $A_0$, $B_0$, $\Delta_0$ and $E_0$ we compute $\Pi_1$ using (4.30)

$$\Pi_1 = \begin{bmatrix}
1.4500 & 0.0400 \\
0.0400 & 1.0960
\end{bmatrix},$$

and the procedure of (a) computing the SVD of $\Pi_1$, (b) computing $\Delta_1$ is repeated as it is done for $\Pi_0$. Similarly, the procedure is repeated for all $t = 0, 1, \ldots, n$. At the end, for the given $\xi$ we check if $\left| \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i} - D \right| \leq \epsilon$. If it does, we stop the iterations and the last $\xi$ is the level we want. If not, we update $\xi$ as $\xi \leftarrow \xi + \beta(D - \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i})$ and we repeat the procedure for all $t$ again. For this example, the final reverse waterfilling is found in 9 iterations and it is shown in Figure 4.2.

![Reverse waterfilling in time-space: vector case](image)

Fig. 4.2: Reverse waterfilling in time-space for n=2 time units and p=2 space units.
By (4.8) we compute the FTH NRDF:

\[
R^\alpha_{n,0}(D) = \frac{1}{2} \frac{1}{n+1} \sum_{t=0}^{2} \sum_{i=1}^{2} \log \left( \frac{\lambda_{i,t}}{\delta_{i,t}} \right)
\]

\[
= \frac{1}{6} \left\{ \log \left( \frac{4.4728}{1.5612} \right) + \log \left( \frac{1.8740}{1.5612} \right) + \log \left( \frac{0.070}{1.5612} \right) + \log \left( \frac{1.1943}{1.1943} \right) 
+ \log \left( \frac{2.5666}{1.5612} \right) + \log \left( \frac{1.9938}{1.5612} \right) \right\}
\]

\[
= 0.6330 \text{ bits/source symbol}
\]

In the next corollary, we degrade the results derived in Theorem 4.1 to the case of time-varying scalar Gauss-Markov processes. This corollary emphasizes on the fact that even in its simplest form, i.e., when \( p = 1 \), the evaluation of FTH NRDF for time-varying Gauss-Markov processes can only be evaluated numerically by utilizing algorithmic methods. Note that in the sequel, when we refer to the scalar Gaussian process, for simplicity we will not make use of the dimension subscript, that is, \( \lambda_{i,1} \equiv \lambda_{i}, \delta_{i,1} \equiv \delta_{i}, \eta_{i,1} \equiv \eta_{i} \) etc.

**Corollary 4.3.** \( R^\alpha_{n,0}(D) \) of time-varying scalar Gauss-Markov processes

This corresponds to (4.1) by setting \( p = k = 1, A_{t} = \alpha_{t}, B_{t} = \sigma_{W_{t}}, \) i.e., \( \sigma_{W_{t}}W_{t} \sim N(0; \sigma_{W_{t}}) \), giving

\[
X_{t+1} = \alpha_{t}X_{t} + \sigma_{W_{t}}W_{t}, \quad W_{t} \sim N(0; 1), \quad X_{0} \sim N(0; \sigma_{X_{0}}^{2}), \quad t = 0, 1, \ldots, n
\]

where \( \{\alpha_{t}, \sigma_{W_{t}} : t = 0, 1, \ldots, n\} \) are time varying. Then \( \sigma_{X_{t}}^{2} \triangleq \text{Var}(X_{t}), \) satisfies \( \sigma_{X_{t}}^{2} = \alpha_{t}^{2}\sigma_{X_{t-1}}^{2} + \sigma_{W_{t}}^{2}, \quad \sigma_{X_{0}}^{2} = \sigma_{0}^{2}, \quad t \in \mathbb{N}_{0} \).

In this case, by Theorem 4.1 and (4.8) we obtain

\[
R^\alpha_{0,n}(D) = \frac{1}{2} \frac{1}{n+1} \sum_{t=0}^{n} \log \left( \frac{\lambda_{t}}{\delta_{t}} \right)
\]

where

\[
\delta_{t} \triangleq \left\{ \begin{array}{ll}
\xi & \text{if} \quad \xi \leq \lambda_{t} \\
\lambda_{t} & \text{if} \quad \xi > \lambda_{t}
\end{array} \right., \quad t = 0, \ldots, n
\]

with \( \xi \) fixed such that \( \frac{1}{n+1} \sum_{t=0}^{n} \delta_{t} = D, \delta_{t} = \min_{t} \{\lambda_{t}, \xi\} \) and \( \Pi_{t} = \Lambda_{t} = \lambda_{t}, \) (i.e., \( E_{t} = 1 \)), \( H_{t} = \eta_{t} = 1 - \frac{\delta_{t}}{\lambda_{t}}, \) \( t = 0, \ldots, n. \)

By (4.16), we obtain

\[
M_{t} = \lambda_{t}H_{t}^{2} + H_{t}\delta_{t} = H_{t}(\lambda_{t}H_{t} + \delta_{t}) = H_{t}(\lambda_{t}(1 - \frac{\delta_{t}}{\lambda_{t}}) + \delta_{t}) = \lambda_{t,1}H_{t}.
\]

Also, by (4.15), we obtain

\[
\lambda_{t+1} = \alpha_{t}^{2}\lambda_{t} - \alpha_{t}^{2}\lambda_{t}^{2}H_{t}^{2}M^{-1} + \sigma_{W_{t}}^{2} \overset{(a)}{=} \alpha_{t}^{2}\lambda_{t} - \alpha_{t}^{2}\lambda_{t}^{2}H_{t}^{2}H_{t}^{-1}\lambda_{t}^{-1} + \sigma_{W_{t}}^{2}
\]

\[
= \alpha_{t}^{2}\lambda_{t} - \alpha_{t}^{2}\lambda_{t}H_{t} + \sigma_{W_{t}}^{2} = \alpha_{t}^{2}\lambda_{t} - \alpha_{t}^{2}\lambda_{t}(1 - \delta_{t}) + \sigma_{W_{t}}^{2} = \alpha_{t}^{2}\lambda_{t} + \sigma_{W_{t}}^{2}, \quad \lambda_{0} = \sigma_{X_{0}}^{2}
\]

where \((a)\) follows from (4.35).
Algorithm 2: Rate distortion allocation algorithm: The scalar case

Initialize:
The number of time-steps $n$; the distortion level $D$; the error tolerance $\epsilon$; the initial variance $\bar{\lambda}_0 = \sigma^2_{X_0}$ of the initial state $X_0$, the values $a_t$ and $\sigma^2_{W_t}$ of the time-varying scalar Gauss-Markov process $X_t$ given by (4.32).

Set $\xi = D$; flag = 0.

while flag = 0 do
    Compute $\delta_t \forall t$ as follows:
    for $t = 0$ : $n$ do
        $\delta_t$ is computed according to (4.34).
        Use $a_t$ and $\sigma^2_{W_t}$ to compute $\lambda_{t+1}$ according to (4.36).
    end for
    if $|\frac{1}{n+1} \sum_{t=0}^{n} \delta_t - D| \leq \epsilon$ then
        flag $\leftarrow 1$
    else
        Re-adjust $\xi$ as follows:
        $\xi \leftarrow \xi + \beta (D - \frac{1}{n+1} \sum_{t=0}^{n} \delta_t)$, where $\beta \in (0, 1]$ is a proportionality gain and affects the rate of convergence.
    end if
end while

Similarly to Algorithm 1, we structure Algorithm 2 for rate distortion allocation.

Example 2. For this example, we choose the distortion level $D = 2$ and use the following $\{a_t^2, \sigma^2_{W_t}\}$:

$$(a_0^2, \sigma^2_{W_0}) = (1, 1), \quad (a_1^2, \sigma^2_{W_1}) = (0.2, 1.3), \quad (a_2^2, \sigma^2_{W_2}) = (1.8, 0.7).$$

The initial variance is $\sigma_{X_0} = 1$. Hence, $\lambda_0 = \sigma_{X_0} = 1$.

Now let’s implement Algorithm 2 for error tolerance $\epsilon = 10^{-3}$. We choose an initial $\xi = \xi_0$ to start our iterations. A good starting point is $\xi_0 = D$. Using (4.34), $\delta_0 = \min\{1, 2\} = 1$. Then, using (4.36), $\lambda_1 = a_0^2 \delta_0 + \sigma^2_{W_0}$ and thus $\delta_1$ is computed. Similarly, the procedure is repeated for all $t = 0, 1, \ldots, n$. At the end, for the given $\xi$ we check if $|\frac{1}{n+1} \sum_{t=0}^{n} \delta_t - D| \leq \epsilon$. If it does, we stop the iterations and the last $\xi$ is the level we want. If not, we update $\xi$ as $\xi \leftarrow \xi + \beta (D - \frac{1}{n+1} \sum_{t=0}^{n} \delta_t)$ and we repeat the procedure for all $t$ again.

For this example, the final reverse waterfilling is found after 15 iterations and it is shown in Figure 4.2.

By (4.33), we compute the FTH NRDF:

$$R_{0,2}^a(D) = \frac{1}{2} \frac{1}{2 + 1} \sum_{t=0}^{2} \log \left( \frac{\lambda_t}{\delta_t} \right) = \frac{1}{6} \left\{ \log \left( \frac{3.8093}{2.1367} \right) + \log \left( \frac{3.1367}{2.1367} \right) + \log \left( \frac{1.7273}{1.7273} \right) \right\}$$

$$= 0.2314 \ \text{bits/source symbol}$$

5. Conclusions and Future Directions. In this paper, we derived the optimal reproduction conditional distribution of the FTH NRDF and drew its connection to real-time realizable filtering theory. Then, we derived the optimal filter for time-varying Gaussian random processes using the solution of the FTH NRDF subject to
a mean square error fidelity. Further, we established a universal lower bound on the
mean square error of any estimator of a Gaussian random process.

Our future work focuses on investigating a similar structure with respect to time-
varying multidimensional partially observable Gaussian processes.

**Appendix A. Proof of Theorem 3.2.**

First, we show the validity of our theorem for the time instant \( t = n \). Without
abuse of notation, we denote \( P_{X_{t+n} \mid X_{t}}(dx_t \mid x_{t-1}^{T_t}) \equiv P(dx_t \mid x_{t-1}^{T_t}) \),
\( Q_{Y_{t+n} \mid Y_{t-1}, X_{t}}(dy_t \mid y_{t-1}^{T_t}, x_{t}^{T_t}) \equiv Q(dy_t \mid y_{t-1}^{T_t}, x_{t}^{T_t}) \), and \( P_{Y_{t+n} \mid X_{t}}(dy_t) \equiv P(dy_t) \). Let \( s \leq 0 \), and \( \lambda_t \in \Psi^*_n \) and \( \overrightarrow{Q}_{Y_{t+n} \mid X^n}(\cdot \mid x^n) \in \overrightarrow{Q}_{0,n}(D) \) be given.

Then, using the fact that

\[
\frac{1}{n+1} \sum_{t=0}^{n} \int_{x_{0,t} \times y_{0,t}} \rho_t(T^t x^n, T^t y^n) \otimes_{j=0}^t \left( P(dx_j \mid x_{j-1}^{T_j}) \otimes Q(dy_j \mid y_{j-1}^{T_j}, x_{j}^{T_j}) \right) \leq D
\]

gives

\[
\frac{1}{n+1} \mathbb{I}_{X \rightarrow Y^n}(P_{X^n}, \overrightarrow{Q}_{Y^n \mid X^n}) - sD
\]

\[
+ \frac{1}{n+1} \sum_{t=0}^{n} \int_{x_{0,t} \times y_{0,t}} g_{t,n}(x_t, y_t) \otimes_{j=0}^t \left( P(dx_j \mid x_{j-1}^{T_j}) \otimes Q(dy_j \mid y_{j-1}^{T_j}, x_{j}^{T_j}) \right)
\]

\[
- \frac{1}{n+1} \sum_{t=0}^{n} \int_{x_{0,t} \times y_{0,t-1}} \log \left( \lambda_t(x_t, y_t^{-1}) \right) P(dx_t \mid x_{t-1}^{T_t}) \otimes_{j=0}^{t-1} \left( P(dx_j \mid x_{j-1}^{T_j}) \otimes Q(dy_j \mid y_{j-1}^{T_j}, x_{j}^{T_j}) \right)
\]
where (a) follows from the inequality \( \log x \geq 1 - \frac{1}{x}, \ x > 0 \), and (b) follows from (3.9).
Hence, we obtain

\[(A.1)\]

\[R_{0,n}^{\alpha_n}(D) \geq \]

\[
sup_{s \leq 0} \sup_{\lambda \in \Psi_n} \left\{ sD - \frac{1}{n + 1} \sum_{t=0}^{n} \int_{\mathcal{Y}_0 \times \mathcal{Y}_0} g_{t,n}(x^t, y^t) \otimes t = 0 \left( P(dx_j | x^{j-1}) \otimes Q(dy_j | y^{j-1}, x^t) \right) \right. 

\[+ \frac{1}{n + 1} \sum_{t=0}^{n} \int_{\mathcal{Y}_0 \times \mathcal{Y}_0, t} \log \left( \lambda_t(x^t, y^{t-1}) \right) P(dx_t | x^{t-1}) \otimes t = 0 \left( P(dx_j | x^{j-1}) \otimes Q(dy_j | y^{j-1}, x^t) \right) \right\}.

However, equality in \((c)\) holds if

\[
\lambda_t(x^t, y^{t-1}) = \left( \int_{\mathcal{Y}_n} e^{t \rho_c(T^n x^n \times T^n y^n) - g_{t,n}(x^t, y^t)} P^*(dy_k | y^{t-1}) \right)^{-1}, \quad t \in N_0^n.
\]

The recursive procedure for \(t = n - 1, \ldots, 0\) is identical because at each time instant \(t\), the term \(g_{t,n}(x^t, y^t)\) contains the previous terms of \(\lambda_{t+1}\). As a result, we have every term of \(\lambda_t\) for each \(t \in N_0^n\).

**Appendix B. Proof of Theorem 4.1**

The derivation is based on the fact that the realization scheme of Fig. 4.1 generally an upper bound on the FTH NRDF \(R_{0,n}^{\alpha_n}(D)\) of the Gaussian processes, and this realization gives \((4.8)\). The achievability of this upper bound and, hence, its optimality, is established by evaluating the lower bound in \((4.8)\) which is done recursively moving backward in time, utilizing the expression we obtained in Theorem 3.2

**Upper Bound.** First, consider the realization of Fig. 4.1 Define \(H_t : t \in N_0^n\) as in \((4.17)\). By Fig. 4.1

\[(B.1)\]

\[
\tilde{K}_t = E_t^T H_t E_t (X_t - \mathbb{E}\{X_t | \sigma(Y^{t-1})\}) + E_t^T \Phi_t V_t^c = E_t^T H_t E_t K_t + E_t^T \Phi_t V_t^c, \quad t \in N_0^n,
\]

where \(\{V_t^c : t \in N_0^n\}\) is an independent Gaussian zero mean process with covariance \(\text{cov}(V_t^c) = Q_t = \text{diag}\{q_{1,1}, \ldots, q_{p,p}\}\), and \(\{\Phi_t : t \in N_0^n\}\) is to be determined.

Next, we show that by letting \(\Phi_t = \sqrt{H_t \Delta_t Q_t^{-1}}\), and \(\Delta_t \triangleq \text{diag} \{\delta_{t,1}, \ldots, \delta_{t,p}\}\), then

\[
\Pi_t = \mathbb{E}\{K_t K_t^T\}, \quad \text{and also} \quad \frac{1}{n+1} \mathbb{E}\left\{ \sum_{t=0}^{n} ||X_t - Y_t||^2 \right\} = \frac{1}{n+1} \mathbb{E}\left\{ \sum_{t=0}^{n} ||K_t - \tilde{K}_t||^2 \right\} = D.
\]

Clearly, by \((4.3)\), \((4.5)\), \((B.1)\)

\[
\frac{1}{n+1} \sum_{t=0}^{n} \mathbb{E}\left\{ (X_t - Y_t)^T (X_t - Y_t) \right\} = \frac{1}{n+1} \sum_{t=0}^{n} \text{trace} \left( \mathbb{E}\left\{ (K_t - \tilde{K}_t) (K_t - \tilde{K}_t)^T \right\} \right)
\]

\[= \frac{1}{n+1} \sum_{t=0}^{n} \text{trace} \left\{ (K_t - E_t^T H_t E_t K_t - E_t^T \Phi_t V_t^c) (K_t - E_t^T H_t E_t K_t - E_t^T \Phi_t V_t^c)^T \right\}
\]

\[= \frac{1}{n+1} \sum_{t=0}^{n} \text{trace} \left\{ ((I - E_t^T H_t E_t) K_t - E_t^T \Phi_t V_t^c) ((I - E_t^T H_t E_t) K_t - E_t^T \Phi_t V_t^c)^T \right\}
\]

\[= \frac{1}{n+1} \sum_{t=0}^{n} \text{trace} \left\{ (I - E_t^T H_t E_t) \Pi_t (I - E_t^T H_t E_t)^T + E_t^T \Phi_t Q_t \Phi_t^T E_t \right\}
\]
The covariance of \( \tilde{B} \) and \( \tilde{C} \) corresponds to processes. The covariance of the Gaussian zero mean term 

\[
E \phi_1, \phi_2 \Rightarrow \phi_3, \phi_4 \] 

Next, we compute the entropies appearing in (B.3) from the covariances of the corresponding processes. The covariance of the Gaussian zero mean term 

\[
E_t \phi_1 V_t^c, t \in \mathbb{N}_0
\]

is given by

\[
E \left\{ (E_t^T \phi_1 V_t^c)(E_t \phi_1 V_t^c)^T \right\} = E_t^T \phi_1 E \{ V_t^c V_t^c \} \phi_1^T E_t = E_t^T \phi_1 Q_t \phi_1^T E_t
\]

(B.4)

The covariance of \( \tilde{K} \), \( t \in \mathbb{N}_0 \) is given by

\[
E \left\{ \tilde{K} \right\} = E \left\{ (E_t^T H_t E_t K_t + E_t^T \phi_1 V_t^c)(E_t^T H_t E_t K_t + E_t^T \phi_1 V_t^c)^T \right\}
\]

\[
= E_t^T H_t E_t \mathbb{E} \{ K_t \} E_t + E_t^T \phi_1 \mathbb{E} \{ V_t^c V_t^c \} \phi_1^T E_t
\]

\[
= E_t^T H_t E_t \Pi_t E_t H_t E_t + E_t^T \sqrt{H_t \Delta_t Q_t^{-1} Q_t \sqrt{H_t \Delta_t Q_t^{-1}}} E_t
\]

\[
= E_t^T \left( \text{diag} \{ \eta_t^2 \lambda_{t,1}, \ldots, \eta_t^2 \lambda_{t,p} \} + \text{diag} \{ \eta_{t,1} \delta_t, \ldots, \eta_{t,p} \delta_t \} \right) E_t
\]

(B.5)

where \( \text{(a)} \) holds by setting \( \Phi_t \) as in (4.18). By (4.7), the FTH NRDF can be written as follows:

\[
R_{\theta,n,K_n}^n (D) \leq \frac{1}{n+1} \sum_{t=0}^n I(K_t; \tilde{K}_t | \tilde{K}_t^{t-1})
\]

\[
= \frac{1}{n+1} \sum_{t=0}^n \left\{ H(\tilde{K}_t | \tilde{K}_t^{t-1}) - H(\tilde{K}_t | \tilde{K}_t^{t-1}, K_t) \right\}
\]

\[
\leq \frac{1}{n+1} \sum_{t=0}^n \left\{ H(\tilde{K}_t) - H(\tilde{K}_t | \tilde{K}_t^{t-1}, K_t) \right\}
\]

\[
\leq \frac{1}{n+1} \sum_{t=0}^n \left\{ H(\tilde{K}_t) - H(\tilde{K}_t | K_t) \right\}
\]

\[
\leq \frac{1}{n+1} \sum_{t=0}^n \left\{ H(\tilde{K}_t) - H(E_t^T \phi_1 V_t^c) \right\},
\]

where \( \text{(b)} \) follows from the fact that conditioning reduces entropy (see also [7], Lemma V.1, Remark V.2), \( \text{(c)} \) follows again from the fact that \( \tilde{K}_t = E_t^T H_t E_t K_t + E_t^T \phi_1 V_t^c \) is a memoryless Gaussian channel, and \( \text{(d)} \) follows from the orthogonality of \( \tilde{K}_t \) and \( V_t^c \). Actually, by [7], Lemma V.1, Remark V.2, it can be shown that the inequalities \( \text{(b)}, \text{(c)}, \text{(d)} \) are equalities.

Next, we compute the entropies appearing in (B.3) from the covariances of the corresponding processes. The covariance of the Gaussian zero mean term \( E_t^T \phi_1 V_t^c, t \in \mathbb{N}_0 \) is given by

\[
E \left\{ (E_t^T \phi_1 V_t^c)(E_t \phi_1 V_t^c)^T \right\} = E_t^T \phi_1 E \{ V_t^c V_t^c \} \phi_1^T E_t
\]

(B.4)
Using (B.5) we obtain the first term of (B.3) as follows:

\[ \sum_{t=0}^{n} H(\tilde{K}_t) = \frac{1}{2} \sum_{t=0}^{n} \log \left\{ (2\pi e) \times \prod_{i=1}^{p} (\lambda_{t,i} - \delta_{t,i})^+ \right\} \]

\[ = \frac{1}{2} \sum_{t=0}^{n} \sum_{i=1}^{p} \log \left\{ (2\pi e) (\lambda_{t,i} - \delta_{t,i})^+ \right\}. \]  

(B.6)

Also, by (B.4), we obtain the second term in (B.3) as follows:

\[ \sum_{t=0}^{n} H(E_t^T \Phi_t V_t^c) = \frac{1}{2} \sum_{t=0}^{n} \log \left\{ (2\pi e)^{\text{diag} \{ \eta_t, \delta_t \}} \right\} \]

\[ = \frac{1}{2} \sum_{t=0}^{n} \sum_{i=1}^{p} \log \left\{ (2\pi e) (\eta_{t,i} \delta_{t,i}) \right\}. \]

(B.7)

By using (B.6) and (B.7) in (B.3) we have the following upper bound:

\[ R_{na,K_n,\tilde{K}_n}^{\text{FTH}}(D) \leq \frac{1}{n+1} \sum_{t=0}^{n} I(K_t; \tilde{K}_t|\tilde{K}_{t-1}) \]

\[ \leq \frac{1}{2} \frac{1}{n+1} \sum_{i=0}^{n-1} \sum_{i=1}^{p} \log \left\{ \frac{(\lambda_{t,i} - \delta_{t,i})^+}{\eta_{t,i} \delta_{t,i}} \right\} \]

\[ = \frac{1}{2} \frac{1}{n+1} \sum_{i=0}^{n-1} \sum_{i=1}^{p} \log \left\{ \frac{\lambda_{t,i}}{\delta_{t,i}} \right\}, \]  

(B.8)

where \( \delta_{t,i} = \min\{\xi, \lambda_{t,i}\}, \ t \in \mathbb{N}_0, i = 1, \ldots, p, \) and \( \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \delta_{t,i} = D. \) Note that if \( \delta_{t,i} = \lambda_{t,i}, \) for \( t \in \mathbb{N}_0 \) and \( i = 1, \ldots, p, \) then no data are reproduced.

**Lower Bound.** Here, we apply Theorem 3.2 recursively, to obtain a lower bound for the FTH NRDF \( R_{0,n,K_n,\tilde{K}_n}^{\text{FTH}}(D) = R_{0,n,K_n,\tilde{K}_n}^{\text{FTH}}(D), \) which is precisely (4.8).

Let \( \bar{p}(\cdot) \) and \( \bar{\pi}(\cdot) \) denote the conditional and unconditional densities, respectively. Using the property of \( \{\lambda_t(\cdot, \cdot): t = 0, \ldots, n\} \) corresponding to the fact that \( \lambda_t(k_t, \tilde{k}_t-1) \equiv \lambda_t(k_t, \tilde{k}_t-1), \ t = 0, \ldots, n \) and by Theorem 3.2, an alternative expression for the FTH NRDF, \( R_{0,n,K_n,\tilde{K}_n}^{\text{FTH}}(D) \) is the following:

\[ R_{0,n,K_n,\tilde{K}_n}^{\text{FTH}}(D) = \sup_{s \leq 0} \sup_{\{\lambda_t(k_t, \tilde{k}_t-1)\in \Psi_t: t \in \mathbb{N}_0\}} \left\{ \text{term-(0)+} + \ldots + \text{term-(n-1)+} + \text{term-(n)} \right\} \]

\[ \downarrow \]

Note that \( (\cdot)^+ \equiv \max\{0, \cdot\} \).
where
\[
\text{term-(0)} = -\frac{1}{n+1} \int_{K_0} \left( \int_{K_0} g_{0,n}(\tilde{k}_0) \tilde{p}(\tilde{k}_0|k_0) d\tilde{k}_0 \right) \tilde{p}(k_0) dk_0 + \frac{1}{n+1} \int_{K_0} \log \left( \lambda_0(k_0) \right) \tilde{p}(k_0) dk_0
\]
\[
\text{term-(1)} = -\frac{1}{n+1} \int_{K_1 \times K_0} \left( \int_{K_1} g_{1,n}(\tilde{k}_1) \tilde{p}(\tilde{k}_1|k_0, k_1) d\tilde{k}_1 \right) \tilde{p}(k_1, \tilde{k}_0) dk_1 d\tilde{k}_0
\]
\[
+ \frac{1}{n+1} \int_{K_1 \times K_0} \log \left( \lambda_0(k_1, \tilde{k}_0) \right) \tilde{p}(k_1, \tilde{k}_0) dk_1 d\tilde{k}_0
\]
\[
\vdots
\]
\[
\text{term-(n-2)} = -\frac{1}{n+1} \int_{K_{n-2} \times K_{n-3}} \left( \int_{K_{n-2}} g_{n-2,n}(\tilde{k}^{n-2}) \tilde{p}(\tilde{k}^{n-2}|k_{n-2}, k_{n-3}) d\tilde{k}^{n-2} \right) \tilde{p}(k_{n-2}, \tilde{k}^{n-3}) dk_{n-2} d\tilde{k}^{n-3}
\]
\[
= -\frac{1}{n+1} \int_{K_{n-1} \times K_{n-2}} \left( \int_{K_{n-1}} g_{n-1,n}(\tilde{k}^{n-1}) \tilde{p}(\tilde{k}^{n-1}|k_{n-1}, k_{n-2}) d\tilde{k}^{n-1} \right) \tilde{p}(k_{n-1}, \tilde{k}^{n-2}) dk_{n-1} d\tilde{k}^{n-2}
\]
\[
\text{term-(1)} = -\frac{1}{n+1} \int_{K_{n-1} \times K_{n-2}} \left( \int_{K_{n-1}} g_{n-1,n}(\tilde{k}^{n-1}) \tilde{p}(\tilde{k}^{n-1}|k_{n-1}, k_{n-2}) d\tilde{k}^{n-1} \right) \tilde{p}(k_{n-1}, \tilde{k}^{n-2}) dk_{n-1} d\tilde{k}^{n-2}
\]
\[
\text{term-(n)} = sD + \frac{1}{n+1} \int_{K_n \times K_{n-1}} \log \left( \lambda_n(k_n, \tilde{k}^{n-1}) \right) \tilde{p}(k_n, \tilde{k}^{n-1}) dk_n d\tilde{k}^{n-1}
\]
and
\[
\Psi^t_s \triangleq \left\{ \lambda_t(k_t, \tilde{k}^t) \geq 0 : \int_{K_t} e^{s||k_t - \tilde{k}^t|\tilde{k}^t|} \lambda_t(k_t, \tilde{k}^t) \tilde{p}(k_t|\tilde{k}^t|) dk_t = 1 \right\}, \ t \in \mathbb{N}_0^n
\]
\[
g_{n,n}(\tilde{k}^n) = 0,
\]
\[
g_{n,n}(\tilde{k}^n) = -\int_{K_{t+1}} \log \left( \lambda_{t+1}(k_{t+1}, \tilde{k}^{t+1}) \right) \tilde{p}(k_{t+1}|\tilde{k}^{t+1}) dk_{t+1}, \ t \in \mathbb{N}_0^{n-1}.
\]

Clearly, if \( g_t(n, \tilde{k}^{t}) = \tilde{g}_t(n, \tilde{k}^{t-1}) \), i.e., it is independent of \( \tilde{k}_t \), for \( t \in \mathbb{N}_0^{n-1} \), then by Theorem 3.1 the RHS terms in \( \text{B.9} \) involving \( g_{t,n}(\cdot, \cdot) \), \( t \in \mathbb{N}_0^{n-1} \), will not appear (because the optimal reproduction distribution will not involve such terms).

Since \( g_{n,n}(\cdot, \cdot) = 0 \), by \( \text{B.10} \), \( \text{B.11} \), \( \lambda_n(k_n, \tilde{k}^{n-1}) \) determines \( g_{n-1,n}(\cdot, \cdot) \), \( \lambda_{n-1}(\cdot, \cdot) \) determines \( g_{n-2,n}(\cdot, \cdot) \) so on, and the right side of \( \text{B.9} \) involves supremum over \( \{ \lambda_t(\cdot, \cdot) : t \in \mathbb{N}_0^n \} \), then any choice of \( \{ \lambda_t(\cdot, \cdot) : t \in \mathbb{N}_0^n \} \) gives a lower bound.

The main idea, implemented below, uses the property of distortion function, and the source distribution, to show that \( \{ \lambda_t(\cdot, \cdot) : t \in \mathbb{N}_0^n \} \) can be chosen so that \( g_{t,n}(\tilde{k}^t) = \tilde{g}_t(n, \tilde{k}^{t-1}) \), \( t \in \mathbb{N}_0^{n-1} \), giving a lower bound which is achievable, and that the optimal reproduction distribution is of the form
\[
\tilde{p}(\tilde{k}_t|k^{t-1}, k_t) = \frac{e^{s||k_t - \tilde{k}^t||^2} \tilde{p}(\tilde{k}_t|k^{t-1})}{\int_{K_t} e^{s||k_t - \tilde{k}^t||^2} \tilde{p}(\tilde{k}_t|k^{t-1})}.
\]

Step \( t = n \): The set \( \Psi^t_s \) is defined as follows:
\[
\Psi^t_s \triangleq \left\{ \lambda_n(k_n, \tilde{k}^{n-1}) \geq 0 : \int_{K_n} e^{s||k_n - \tilde{k}^n||^2} \lambda_n(k_n, \tilde{k}^{n-1}) \tilde{p}(k_n|\tilde{k}^{n-1}) dk_n = 1 \right\},
\]
where \( \bar{p}(k_n|\hat{k}^{n-1}) \) denotes the conditional density of \( k_n \) given \( \hat{k}^{n-1} \). Take \( \lambda_n(k_n, \hat{k}^{n-1}) \in \Psi_n \) such that

\[
\lambda_n(k_n, \hat{k}^{n-1}) = \frac{\alpha_n}{\bar{p}(k_n|\hat{k}^{n-1})}
\]

for some \( \alpha_n \) not depending on \( k_n \), and substitute (B.13) into the integral inequality in (B.12) to obtain

\[
\alpha_n \int_{\mathbb{K}_n} e^{|k_n - \hat{k}^{n-1}|^2} dk_n \leq 1.
\]

By change of variable of integration then

\[
\alpha_n \int_{-\infty}^{\infty} e^{|s||z_n|^2} dz_n = \alpha_n \sqrt{\left(-\frac{\pi}{s}\right)^p} = \alpha_n \left(-\frac{\pi}{s}\right)^{\frac{p}{2}} \leq 1.
\]

where “s” is the non-positive Lagrange multiplier. Moreover, \( \alpha_n \) is chosen so that the inequality of (B.14) holds with equality, giving

\[
\alpha_n = \frac{1}{\sqrt{\int_{\mathbb{K}_n} e^{|s||z_n|^2} dz_n}} = \left(-\frac{s}{\pi}\right)^{\frac{p}{2}}, \quad \lambda_n(k_n, \hat{k}^{n-1}) = \frac{\bar{p}(k_n|\hat{k}^{n-1})}{\alpha_n}.
\]

Substituting (B.15) into the term-(n) of (B.9) gives

\[
\text{term-(n)} = sD + \frac{1}{n+1} \log \alpha_n - \frac{1}{n+1} \int_{\mathbb{K}_n \times \hat{\mathbb{K}}^{n-1}} \log \left( \bar{p}(k_n|\hat{k}^{n-1}) \right) \bar{p}(k_n, \hat{k}^{n-1}) dk_n d\hat{k}^{n-1}
\]

\[
= sD + \frac{1}{n+1} \log \left(-\frac{\pi}{s}\right)^{\frac{p}{2}} + \frac{1}{n+1} H(K_n|\hat{K}^{n-1}).
\]

The choice of \( \lambda_n(\cdot, \cdot) \) given by (B.15) determines \( g_{n-1,n}(\cdot) \) given by

\[
g_{n-1,n}(\hat{k}^{n-1}) = -\int_{\mathbb{K}_n} \bar{p}(dk_n|\hat{k}^{n-1}) \lambda_n(k_n, \hat{k}^{n-1})^{-1}(a)
\]

\[
= -\int_{\mathbb{K}_n} \bar{p}(dk_n|\hat{k}^{n-1}) \log \left( \frac{\bar{p}(k_n|\hat{k}^{n-1})}{\alpha_n} \right)
\]

\[
= \log \alpha_n + H(K_n|\hat{K}^{n-1} = \hat{k}^{n-1}), \quad \alpha_n = \left(-\frac{s}{\pi}\right)^{\frac{p}{2}}
\]

\[
\leq \log \alpha_n + H(K_n|\hat{K}^{n-2} = \hat{k}^{n-2})
\]

\[
\left(\begin{array}{c}
(a) \\
(b)
\end{array}\right)
\]

\[
\log \left(-\frac{\pi}{s}\right)^{\frac{p}{2}} + H(K_n|\hat{K}^{n-2} = \hat{k}^{n-2}) \equiv g_{n-1,n}(\hat{k}^{n-2})
\]

where (a) follows from the fact that \( \left( \lambda_n(k_n, \hat{k}^{n-1}) \right)^{-1} = \frac{\bar{p}(k_n|\hat{k}^{n-1})}{\alpha_n} \), and (b) from the fact that conditioning reduces entropy.

When the upper bound in (B.17) is substituted into the second expression of term-(n-1) of (B.9) involving \( g_{n-1,n}(\cdot) \), it gives

\[
- \frac{1}{n+1} \int_{\mathbb{K}_{n-1} \times \hat{\mathbb{K}}^{n-2}} \left( \int_{\mathbb{K}_{n-1}} g_{n-1,n}(\hat{k}^{n-1}) \bar{p}(\hat{k}^{n-2}, \hat{k}_{n-1}) d\hat{k}_{n-1} \right) \bar{p}(k_{n-1}, \hat{k}^{n-2}) dk_{n-1} d\hat{k}^{n-2}
\]

\[
\geq - \frac{1}{n+1} \int_{\mathbb{K}_{n-1} \times \hat{\mathbb{K}}^{n-2}} \left( \int_{\mathbb{K}_{n-1}} g_{n-1,n}(\hat{k}^{n-2}) \bar{p}(\hat{k}^{n-1}) d\hat{k}^{n-1} \right) \bar{p}(k_{n-1}, \hat{k}^{n-2}) dk_{n-1} d\hat{k}^{n-2}.
\]
Step $t = n - 1$: The set $\Psi_{s}^{n-1}$ is defined as follows (using $g_{n-1,n}(\hat{k}^{-1}) = g_{n-1,n}(\hat{k}^{-2})$

given by (B.17) obtained in step $t = n$)

$$
\Psi_{s}^{n-1} \triangleq \left\{ \lambda_{n-1}(k_{n-1},\hat{k}^{-2}) \geq 0 : \right.
\left. \int_{K_{n-1}} e^{s||k_{n-1} - \hat{k}^{-1}||_{2}^{2} - \tilde{g}_{n-1,n}(\hat{k}^{-2})} \lambda_{n-1}(k_{n-1},\hat{k}^{-2}) \tilde{p}(k_{n-1} | \hat{k}^{-2}) dk_{n-1} \leq 1 \right\}.
$$

Take $\lambda_{n-1}(k_{n-1},\hat{k}^{-2}) \in \Psi_{s}^{n-1}$ such that

$$
\lambda_{n-1}(k_{n-1},\hat{k}^{-2}) = \frac{\alpha_{n-1}(\hat{k}^{-2})}{\tilde{p}(k_{n-1} | \hat{k}^{-2})}
$$

for some $\alpha_{n-1}(\hat{k}^{-2})$ not depending on $k_{n-1}$, and substitute (B.19) into the integral

inequality in (B.18) to obtain

$$
\alpha_{n-1}(\hat{k}^{-2}) e^{-\tilde{g}_{n-1,n}(\hat{k}^{-2})} \int_{K_{n-1}} e^{s||k_{n-1} - \hat{k}^{-1}||_{2}^{2}} dk_{n-1} \leq 1.
$$

By change of variable of integration then

$$
\alpha_{n-1}(\hat{k}^{-2}) e^{-\tilde{g}_{n-1,n}(\hat{k}^{-2})} \int_{-\infty}^{\infty} e^{s||z_{n-1}||_{2}^{2}} dz_{n-1} = \alpha_{n-1}(\hat{k}^{-2}) e^{-\tilde{g}_{n-1,n}(\hat{k}^{-2})} \left( \frac{\pi}{s} \right)^{\frac{p}{2}} \leq 1.
$$

Hence,

$$
\alpha_{n-1}(\hat{k}^{-2}) \left( \frac{\pi}{s} \right)^{\frac{p}{2}} \leq e^{\tilde{g}_{n-1,n}(\hat{k}^{-2})}.
$$

Moreover, $\alpha_{n-1}(\cdot)$ is chosen so that the inequality in (B.20) holds with equality, giving

$$
\alpha_{n-1}(\hat{k}^{-2}) = e^{\tilde{g}_{n-1,n}(\hat{k}^{-2})} \left( \frac{\pi}{s} \right)^{\frac{p}{2}} = e^{\log \alpha_{n} + H(K_{n} | \hat{k}^{-2} = \hat{k}^{-2}) - \frac{s}{\pi} \frac{p}{2}}
$$

(B.21)

where (j) holds due to (B.17). Therefore, (B.19) is given by

$$
\lambda_{n-1}(k_{n-1},\hat{k}^{-1}) = \frac{\left( -\frac{s}{\pi} \right)^{\frac{p}{2}} e^{H(K_{n} | \hat{k}^{-2} = \hat{k}^{-2})}}{\tilde{p}(k_{n-1} | \hat{k}^{-2})}.
$$
Substituting (B.22) into the term-(n-1) of (B.9) gives

\[
\text{term-(n-1)} \geq \frac{1}{n+1} \int_{k_{n-1} \times \hat{k}_{n-2}} \left( \int_{\hat{k}_{n-1}} \hat{g}_{n-1,n}(\hat{k}_{n-2}) \hat{p}(k_{n-1}, \hat{k}_{n-2}) d\hat{k}_{n-1} \right) \\
\times \hat{p}(k_{n-1}, \hat{k}_{n-2}) d\hat{k}_{n-1} d\hat{k}_{n-2}
\]

(B.23)

\[
+ \frac{1}{n+1} \int_{k_{n-1} \times \hat{k}_{n-2}} \log \left( \lambda_{n-1}(k_{n-1}, \hat{k}_{n-2}) \right) \hat{p}(k_{n-1}, \hat{k}_{n-2}) d\hat{k}_{n-1} d\hat{k}_{n-2}
\]

\[
\geq -\frac{1}{n+1} \log \left( -\frac{s}{\pi} \right) \frac{2}{p} - \frac{1}{n+1} H(k_{n-1}|\hat{K}_{n-2})
\]

(B.24)

where (c) follows from the fact that \( g_{n-1,n}(\hat{k}_{n-1}) \leq \hat{g}_{n-1,n}(\hat{k}_{n-2}) \) (see (B.17)) and (d) follows by substituting (B.17) and (B.19) into the the second and third expression of (B.23), respectively.

The choice of \( \lambda_{n-1}(\cdot, \cdot) \) (given by (B.22)) determines \( g_{n-2,n}(\cdot) \) given by

\[
g_{n-2,n}(\hat{k}_{n-2}) = -\int_{k_{n-1}} \hat{p}(k_{n-1}|\hat{k}_{n-2}) \log \left( \lambda_{n-1}(k_{n-1}, \hat{k}_{n-2}) \right)^{-1}
\]

\[
\equiv -\int_{k_{n-1}} \hat{p}(k_{n-1}|\hat{k}_{n-2}) \log \left( \frac{\hat{p}(k_{n-1}|\hat{k}_{n-2})}{\alpha_{n-1}(k_{n-2})} \right), \quad \alpha_{n-1}(\hat{k}_{n-2}) = \left( -\frac{s}{\pi} \right)^p e^{H(k_{n-1}|\hat{K}_{n-2}=\hat{k}_{n-2})}
\]

\[
= \log \left( \alpha_{n-1}(\hat{k}_{n-2}) \right) - \int_{k_{n-1}} \hat{p}(k_{n-1}|\hat{k}_{n-2}) \log \left( \hat{p}(k_{n-1}|\hat{k}_{n-2}) \right)
\]

\[
= \log \left( -\frac{s}{\pi} \right)^p + H(K_{n-1}|\hat{K}_{n-2} = \hat{k}_{n-2}) + H(K_{n-1}|\hat{K}_{n-2} = \hat{k}_{n-2})
\]

\[
\leq \log \left( -\frac{s}{\pi} \right)^p + H(K_{n-1}|\hat{K}_{n-3} = \hat{k}_{n-3}) + H(K_{n-1}|\hat{K}_{n-3} = \hat{k}_{n-3})
\]

\[
\equiv \hat{g}_{n-2,n}(\hat{k}_{n-3})
\]

where (e) follows from the fact that \( \left( \lambda_{n-1}(k_{n-1}, \hat{k}_{n-2}) \right)^{-1} = \frac{\hat{p}(k_{n-1}|\hat{k}_{n-2})}{\alpha_{n-1}(k_{n-2})} \), and (f) follows from the fact that conditioning reduces entropy.
When the upper bound in (B.25) is substituted into the second expression of term-(n-2) of (B.9) involving \( g_{n-2,n}(\cdot) \), it gives

\[
- \frac{1}{n+1} \int_{\mathcal{K}_{n-2} \times \mathcal{E}_{n-3}} \left( \int_{\mathcal{K}_{n-2}} g_{n-2,n}(\tilde{k}^{n-2}) \bar{p}(\tilde{k}_{n-2}|\tilde{k}^{n-3}, k_{n-2}) d\tilde{k}_{n-2} \right) \bar{p}(k_{n-2}, \tilde{k}^{n-3}) dk_{n-2} d\tilde{k}^{n-3} \\
\geq - \frac{1}{n+1} \int_{\mathcal{K}_{n-2} \times \mathcal{E}_{n-3}} \left( \int_{\mathcal{K}_{n-2}} \bar{g}_{n-2,n}(\tilde{k}^{n-3}) \bar{p}(\tilde{k}_{n-2}|\tilde{k}^{n-3}, k_{n-2}) d\tilde{k}_{n-2} \right) \bar{p}(k_{n-2}, \tilde{k}^{n-3}) dk_{n-2} d\tilde{k}^{n-3}.
\]

**Step t = n - 2:** The set \( \Psi_s^{n-2} \) is defined as follows (using \( g_{n-2,n}(\tilde{k}^{n-2}) = \bar{g}_{n-2,n}(\tilde{k}^{n-3}) \) given by (B.25) obtained in step-n - 1).

(B.26)

\[
\Psi_s^{n-2} \triangleq \left\{ \lambda_{n-2}(k_{n-2}, \tilde{k}^{n-3}) \geq 0 : \int_{\mathcal{K}_{n-2}} e^{s||k_{n-2}-\tilde{k}_{n-2}||^2} \bar{g}_{n-2,n}(\tilde{k}^{n-3}) \lambda_{n-2}(k_{n-2}, \tilde{k}^{n-3}) \bar{p}(k_{n-2}|\tilde{k}^{n-3}) dk_{n-2} \leq 1 \right\}.
\]

Take \( \lambda_{n-2}(k_{n-2}, \tilde{k}^{n-3}) \in \Psi_s^{n-2} \) such that

(B.27)

\[
\lambda_{n-2}(k_{n-2}, \tilde{k}^{n-3}) = \frac{\alpha_{n-2}(\tilde{k}^{n-3})}{\bar{p}(k_{n-2}|\tilde{k}^{n-3})}
\]

for some \( \alpha_{n-2}(\tilde{k}^{n-3}) \) not depending on \( k_{n-2} \), and substitute (B.27) into the integral inequality in (B.26) to obtain

\[
\alpha_{n-2}(\tilde{k}^{n-3}) e^{-\bar{g}_{n-2,n}(\tilde{k}^{n-3})} \int_{\mathcal{K}_{n-2}} e^{s||k_{n-2}-\tilde{k}_{n-2}||^2} dk_{n-2} \leq 1.
\]

By change of variable of integration then

\[
\alpha_{n-2}(\tilde{k}^{n-3}) e^{-\bar{g}_{n-2,n}(\tilde{k}^{n-3})} \int_{-\infty}^{\infty} e^{s||z_{n-2}||^2} dz_{n-2} = \alpha_{n-2}(\tilde{k}^{n-3}) e^{-\bar{g}_{n-2,n}(\tilde{k}^{n-3})} \left( -\frac{\pi}{s} \right)^{\frac{3}{2}} \leq 1.
\]

Hence,

(B.28)

\[
\alpha_{n-2}(\tilde{k}^{n-3}) \left( -\frac{\pi}{s} \right)^{\frac{3}{2}} \leq e^{-\bar{g}_{n-2,n}(\tilde{k}^{n-3})}.
\]

Moreover, \( \alpha_{n-2}(\cdot) \) is chosen so that the inequality in (B.28) holds with equality, giving

(B.29)

\[
\alpha_{n-2}(\tilde{k}^{n-3}) = \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^3 e^{H(K_{n-1}|\tilde{k}^{n-3}=\tilde{k}^{n-3}) - \bar{H}(K_{n-1}|\tilde{k}^{n-3}=\tilde{k}^{n-3})}.
\]

Therefore, (B.27) is given by

(B.30)

\[
\lambda_{n-2}(k_{n-2}, \tilde{k}^{n-3}) = \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^3 e^{H(K_{n-1}|\tilde{k}^{n-3}=\tilde{k}^{n-3}) + \bar{H}(K_{n-1}|\tilde{k}^{n-3}=\tilde{k}^{n-3})} \bar{p}(k_{n-2}|\tilde{k}^{n-3}).
\]
Substituting (B.30) into term-(n-2) of (B.9) gives

\[
\text{Term} - (n-2) : (g) \geq \int_{K_{n-2} \times K_{n-3}} \left( \int_{K_{n-2}} \tilde{g}_{n-2,n}(k_{n-3}) \tilde{p}(k_{n-2},k_{n-3}) d\tilde{k}_{n-2} \right) \\
\tilde{p}(k_{n-2},k_{n-3}) d\tilde{k}_{n-2} d\tilde{k}_{n-3} \\
\int_{K_{n-2} \times K_{n-3}} \log \left( \lambda_{n-2}(k_{n-2}, \tilde{k}_{n-3}) \right) \tilde{p}(k_{n-2}, \tilde{k}_{n-3}) d\tilde{k}_{n-2} d\tilde{k}_{n-3} \\
(h) - \frac{1}{n+1} \log \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^2 - H(K_n|\tilde{K}_{n-3}) - H(K_{n-1}|\tilde{K}_{n-3}) \\
+ \frac{1}{n+1} \int_{K_{n-2} \times K_{n-3}} \log \left( \alpha_{n-2}(\tilde{k}_{n-3}) \right) \tilde{p}(k_{n-2}, \tilde{k}_{n-3}) d\tilde{k}_{n-2} d\tilde{k}_{n-3} \\
- \frac{1}{n+1} \int_{K_{n-2} \times K_{n-3}} \log \left( \tilde{p}(k_{n-2}, k_{n-3}) \right) \tilde{p}(k_{n-2}, \tilde{k}_{n-3}) d\tilde{k}_{n-2} d\tilde{k}_{n-3} \\
= -\frac{1}{n+1} \log \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^2 - \frac{1}{n+1} H(K_n|\tilde{K}_{n-3}) \\
- \frac{1}{n+1} H(K_{n-1}|\tilde{K}_{n-3}) + \frac{1}{n+1} \log \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^3 \\
+ H(K_n|\tilde{K}_{n-3}) + \frac{1}{n+1} H(K_{n-1}|\tilde{K}_{n-3}) + \frac{1}{n+1} H(K_{n-2}|\tilde{K}_{n-3}) \\
= \frac{1}{n+1} \log \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} + \frac{1}{n+1} H(K_n|\tilde{K}_{n-3}) \\
\text{(B.32)}
\]

where (g) follows from the fact that \( g_{n-2,n}(\tilde{k}_{n-2}) \leq \tilde{g}_{n-2,n}(\tilde{k}_{n-3}) \) (see (B.25)), and (h) follows by substituting (B.25) and (B.27) into the the second and third expression of (B.31), respectively.

By applying induction, we obtain the following lower bound for the FTH NRDF.

\[
P_{0,n,K^n,D}^{\Lambda^n} \geq sD + \frac{1}{n+1} \left\{ \left( -\frac{s}{\pi} \right)^{\frac{3}{2}} \right\}^{n+1} \\
+ \frac{1}{n+1} \left\{ H(K_n|\tilde{K}_{n-1}) + H(K_{n-1}|\tilde{K}_{n-2}) + \ldots + H(K_1|\tilde{K}_0) + H(K_0) \right\} \\
= sD + \frac{1}{2n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \log \left( -\frac{s}{\pi} \right) + \frac{1}{n+1} \sum_{t=0}^{n} H(K_i|\tilde{K}^{t-1}) \\
\text{(B.33)}
\]

where (i) follows from the fact that

\[
H(K_i|\tilde{K}^{t-1}) = H(X_t - \mathbb{E}\{X_t|\sigma\{K^{t-1}\}\})|\tilde{K}^{t-1} \\
= H(X_t|\tilde{K}^{t-1}) = H(X_t) = \frac{1}{2} \sum_{t=0}^{n} \log 2\pi e |\Lambda_t|.
\]

Next, we show how to find the Lagrangian multiplier “s” so that the lower bound (B.33) equals \( \frac{1}{2} \sum_{t=0}^{n} \sum_{i=1}^{p} \log \left( \frac{\Lambda_i}{m_i} \right) \). To this end, we need to ensure existence of
some $s < 0$ such that the following identity holds.

$$sD + \frac{1}{2} \frac{1}{2n+1} \sum_{n=0}^{\infty} \sum_{i=1}^{n} \log \left( -\frac{s}{\pi} \right) + \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log 2\pi e |A_i| = \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log \left( \frac{\lambda_{t,i}}{\delta_{t,i}} \right).$$

After some algebra, the previous expression can be simplified into the following expression.

$$\frac{1}{2} \log e^{2t} \left( \frac{1}{\pi t^2} \right) \sum_{n=0}^{\infty} \sum_{i=1}^{n} \log \left( -\frac{s}{\pi} \right) + \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log 2\pi e |A_i| = \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log \left( \frac{\lambda_{t,i}}{\delta_{t,i}} \right).$$

In turn, from the equation above we obtain

$$\frac{1}{2} \log e^{2t} \left( \frac{1}{\pi t^2} \right) \sum_{n=0}^{\infty} \sum_{i=1}^{n} \log e^{2s\delta_{t,i}} \left( -\frac{s}{\pi} \right) + \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log 2\pi e |A_i| \Rightarrow \delta_{t,i} = -\frac{1}{2s}$$

where $\delta_{t,i} = \{ \xi, \lambda_{t,i} \}$. Now, if $\delta_{t,i} = \xi$ then $\delta_{t,i} = -\frac{1}{2s}$ and the FTH NRDF is bounded below by the following expression

$$R^{\text{GNS},\text{F}}(D) \geq \frac{1}{2} \frac{1}{2n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \log \left( \frac{\lambda_{t,i}}{\delta_{t,i}} \right), \quad \frac{1}{n+1} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \delta_{t,i} = D.$$
[6] A. Farhadi and C. D. Charalambous, “Stability and reliable data reconstruction of uncertain dynamic systems over finite capacity channels,” *Automatica*, vol. 46, no. 5, pp. 889–896, 2010.

[7] P. A. Stavrou, C. K. Kourtellaris, and C. D. Charalambous, “Information nonanticipative rate distortion function and its applications,” *submitted to IEEE Transactions on Information Theory*, 2015.

[8] S. C. Tatikonda, *Control Over Communication Constraints*. PhD thesis, Mass. Inst. of Tech. (M.I.T.), Cambridge, MA, 2000.

[9] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed, “Nonanticipative rate distortion function and relations to filtering theory,” *IEEE Transactions on Automatic Control*, vol. 59, pp. 937–952, April 2014.

[10] A. K. Gorbunov and M. S. Pinsker, “Nonanticipatory and prognostic epsilon entropies and message generation rates,” *Problems of Information Transmission*, vol. 9, pp. 184–191, July-Sept. 1973.

[11] T. Tanaka, 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*, 2016 (to appear).

[12] T. Tanaka, “Zero-delay rate-distortion optimization for partially observable Gauss-Markov processes,” in *54th IEEE Conference on Decision and Control (CDC)*, pp. 5725–5730, Dec. 2015.

[13] 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 (ISIT)*, pp. 521–525, June 2015.

[14] R. S. Bucy, “Distortion rate theory and filtering,” *IEEE Transactions on Information Theory*, vol. 28, pp. 330–340, Mar. 1982.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, New Jersey, second ed., 2006.

[16] J. I. Galdos and D. E. Gustafson, “Information and distortion in reduced-order filter design,” *IEEE Transactions on Information Theory*, vol. 23, pp. 183–194, Mar. 1977.

[17] X. Feng, A. Loparo, and Y. Fang, “Optimal state estimation for stochastic systems: An information theoretic approach,” *IEEE Transactions on Automatic Control*, vol. 42, pp. 771–785, June 1997.

[18] S. Tatikonda, A. Sahai, and S. Mitter, “Stochastic linear control over a communication channel,” *IEEE Transactions on Automatic Control*, vol. 49, pp. 1549–1561, Sept. 2004.

[19] D. Teneketzis, “On the structure of optimal real-time encoders and decoders in noisy communication,” *IEEE Transactions on Information Theory*, vol. 52, pp. 4017–4035, Sept 2006.

[20] L. Guo, L. Yin, H. Wang, and T. Chai, “Entropy optimization filtering for fault isolation of nonlinear non-Gaussian stochastic systems,” *IEEE Transactions on Automatic Control*, vol. 54, pp. 804–810, April 2009.

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

[22] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes II: Applications*. Springer-Verlag, 2 ed., 2001.

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

[24] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. Springer-Verlag, Berlin, Heidelberg, New York, 1995.

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

[26] H. Marko, “The bidirectional communication theory—A generalization of information theory,” *IEEE Transactions on Communications*, vol. 21, pp. 1345–1351, Dec. 1973.

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

[28] C. D. Charalambous and P. A. Stavrou, “Directed information on abstract spaces: Properties and variational equalities,” *submitted to IEEE Transactions on Information Theory*, 2015.

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

[30] R. E. Blahut, *Principles and Practice of Information Theory*. in Electrical and Computer Engineering, Reading, MA: Addison-Wesley Publishing Company, 1987.