SIAM J. CONTROL OPTIM.
Vol. 56, No. 5, pp. 3731–3765

OPTIMAL ESTIMATION VIA NONANTICIPATIVE RATE DISTORTION FUNCTION AND APPLICATIONS TO TIME-VARYING GAUSS–MARKOV PROCESSES*

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

Abstract. In this paper, we develop finite-time horizon causal filters for general processes taking values in Polish spaces using the nonanticipative rate distortion function (NRDF). Subsequently, we apply the NRDF to design optimal filters for time-varying vector-valued Gauss–Markov processes, subject to a mean-squared error (MSE) distortion. Unlike the classical Kalman filter design, the developed filters based on the NRDF are characterized parametrically by a dynamic reverse-waterfilling optimization problem obtained via Karush–Kuhn–Tucker conditions. We develop algorithms that provide, in general, tight upper bounds to the optimal solution to the dynamic reverse-waterfilling optimization problem subject to a total and per-letter MSE distortion constraint. Under certain conditions, these algorithms produce the optimal solutions. Further, we establish a universal lower bound on the total and per-letter MSE of any estimator of a Gaussian random process. Our theoretical framework is demonstrated via simple examples.

Key words. causal filters, nonanticipative rate distortion function, mean-squared error distortion, dynamic reverse-waterfilling, universal lower bound

AMS subject classifications. 93E03, 94A34, 90C25, 15A60, 65F10

DOI. 10.1137/17M1116349

1. Introduction.

Motivated by real-time control applications of communication system design, Gorbunov and Pinsker in [1] introduced the so-called nonanticipatory ϵ-entropy of general processes (see [1, Introduction I]). The nonanticipatory ϵ-entropy is equivalent to Shannon's classical rate distortion function (RDF) [2, 3] with an additional causality constraint imposed on the optimal reproduction distribution or estimator. Along the same lines, for a two-sample Gaussian process, Bucy in [4] derived a causal estimator using the distortion rate function[1] subject to a causality constraint. Galdos and Gustafson in [6] applied the classical RDF to design reduced order estimators. Tatikonda, in his Ph.D. thesis [7], applied the nonanticipatory ϵ-entropy, called therein the sequential RDF, and related it to the optimal performance theoretically attainable by causal codes, as defined by Neuhoff and Gilbert in [8]. In addition, Tatikonda in [7] applied the sequential RDF of a scalar-valued Gaussian process described by a discrete recursion, subject to a mean-squared error (MSE) distortion at each time instant, computed by Gorbunov and Pinsker [9, Examples 1, 2] to illustrate by construction how to communicate a scalar-valued Gaussian process,

*Received by the editors February 13, 2017; accepted for publication (in revised form) July 16, 2018; published electronically October 16, 2018.
http://www.siam.org/journals/sicon/56-5/M111634.html
†Department of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden (fstavrou@kth.se).
‡Department of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University, Espoo, Finland (themistoklis.charalambous@aalto.fi).
§Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus (chadcha@ucy.ac.cy).
¶School of Electrical Engineering and Computer Science, University of Ottawa, Ontario, K1N 6N5, Canada (sergey.loyka@ottawa.ca).
[1]The distortion rate function is the dual of the RDF (see [5]).

optimally over a memoryless additive Gaussian noise channel. In [10], the authors showed that a necessary condition to stabilize a controlled process described by a linear discrete recursion driven by a control process and an independent Gaussian process, over a limited-rate communication channel, is "the capacity of the channel, noiseless or noisy, is larger than the sum of logarithms of the absolute values of the unstable eigenvalues of the open-loop control system." Similar conditions are derived by many authors via alternative methods in [11, 12, 13].

In [14], Charalambous, Stavrou, and Ahmed revisited the relation between information theory and filtering theory, using the so-called nonanticipative RDF (NRDF), showed its equivalence to the nonanticipatory ϵ-entropy RDF (see [14, Lemma II.6]), and derived sufficient conditions for the existence of an optimal reproduction distribution of the NRDF. Moreover, in [14], the authors derived the form of the optimal reproduction distribution, under the assumption that the solution to the NRDF is time-invariant. Then, they used this expression to derive a suboptimal causal filter for time-invariant multidimensional partially observed Gaussian processes described by discrete-time recursions, subject to an MSE distortion. The optimal reproduction distribution which minimizes the directed information from one process to another process, subject to a general fidelity criterion of reproduction, is given in [15] and further explained in [16].

In recent years, the NRDF has been applied in many communication-related problems. Derpich and Østergaard in [17] applied the nonanticipatory ϵ-entropy of the scalar Gaussian process subject to an MSE distortion at each time instant to derive several bounds on the optimal performance theoretically attainable by causal and zero-delay codes. The importance of NRDF to the joint design of an {`encoder, channel, decoder`} operating optimally in real time is investigated in [18]. The simplicity of such joint {`encoder, channel, decoder`} design, operating optimally in real time, is demonstrated by Kourtellaris, Charalambous, and Boutros in [19], first by communicating a binary symmetric Markov process over a binary input-output channel with unit memory on past channel outputs (with symmetry) subject to a transmission cost constraint, and then by reconstructing it subject to an average Hamming distortion.

In [20], Tanaka et al. computed numerically the expression given in [21] of the finite-time and stationary NRDF of a multidimensional fully observed Gauss–Markov process subject to a per-letter and asymptotic MSE distortion, using semidefinite programming. Further, in [20] connections to the minimum data-rate achievable by zero-delay source coding problems are discussed.

1.1. Problem statement. In this paper we investigate the following estimation problem: *given an arbitrary random process, we wish to design an optimal communication system such that at its output, the estimated process satisfies an end-to-end fidelity criterion or the average distortion is below a given level.*

This problem is equivalent to the design of an optimal {`encoder, channel, decoder`} that communicates an arbitrary random process to the output of the decoder, with the specified average distortion level. Formally, the problem can be cast as follows.

PROBLEM 1 (information-based estimation). *Given*

(a) *a random process* $\{X_t : t = 0, \ldots, n\}$ *taking values in complete separable metric spaces* $\{\mathcal{X}_t : t = 0, \ldots, n\}$, *with conditional distribution* $\{\mathbf{P}_{X_t|X_0^{t-1}} : t = 0, \ldots, n\}$, $x_0^{t-1} \triangleq (x_0, x_1, \ldots, x_{t-1})$;

(b) *a distortion function of reproducing* x_t *by* $y_t \in \mathcal{Y}_t \subseteq \mathcal{X}_t, t = 0, 1, \ldots, n$, *defined by a real-valued measurable function* $d_{0,n}(\cdot, \cdot)$

$$d_{0,n}(x_0^n, y^n) \triangleq \sum_{t=0}^{n} \rho_t \left(T^t x_0^n, T^t y^n\right) \in [0, \infty], \tag{1.1}$$

or at each time t, *defined by*

$$d_t\left(x_0^t, y^t\right) \triangleq \rho_t \left(T^t x_0^n, T^t y^n\right), \quad t = 0, \ldots, n, \tag{1.2}$$

where $T^t x_0^n \subseteq (x_0, x_1, \ldots, x_t)$, $T^t y^n \subseteq (y^{-1}, y_0, y_1, \ldots, y_t)$, *is either fixed or nonincreasing with time*[2] *for* $t = 0, 1, \ldots, n$, *and* $y^{-1} \in \mathcal{Y}^{-1}$ *is the initial state,*

we wish to determine an optimal probabilistic {`encoder, channel, decoder`} *to communicate* $\{X_t : t = 0, \ldots, n\}$ *to the output of the decoder or estimator, with end-to-end average distortion that satisfies*

$$\frac{1}{n+1}\mathbf{E}\left\{d_{0,n}(X_0^n, Y^n)\right\} \leq D \quad \forall D \in [0, \infty), \tag{1.3}$$

or at each time t, *the average distortion satisfies*

$$\mathbf{E}\left\{\rho_t \left(T^t X_0^n, T^t Y^n\right)\right\} \leq D_t \ \ \forall D_t \in [0, \infty), \quad t = 0, \ldots, n. \tag{1.4}$$

Regarding application examples, our focus is on Gaussian sources with memory, subject to the total and per-letter MSE distortions (1.3) and (1.4). Apart from the numerical computation of [20], the reverse-waterfilling solution for Problem 1 remains to this date unsolved in the literature.[3] For the analogous classical RDF a similar problem has also remained open for several years (see the discussion in [24]).

The above definition of information-based estimation problem ensures the fidelity criterion (1.3) or (1.4) is met, hence it is fundamentally different from standard estimation techniques, such as MSE, maximum a posteriori, and maximum likelihood. In general, it is known from Shannon's information theory [2] that to achieve such a fidelity criterion, for any $D \in [D_{\min}, \infty] \subseteq [0, \infty]$, we need to design an encoder, a channel whose output is the actual observation process or sensor measurements, and a decoder or estimator that takes as an input the channel outputs and produces the estimated process Y^n of X^n. In Shannon's noiseless source coding theorem [2] the channel is noiseless. However, for the noisy coding theorem the channel is noisy, and the problem is equivalent to the construction of the {`encoder, channel, decoder`}, as shown in Figure 1 (for a thorough discussion on the duality of sources and channels see, e.g., [25]).

Our main objective is to address Problem 1 using information-theoretic measures. By the converse coding theory of causal codes [16], the natural information-theoretic measure to address Problem 1 is the NRDF (see Definition 1.1). Moreover, by data processing inequality, the capacity of the channel in Figure 1 is larger than or equal to the NRDF, and equality holds if the encoder and decoder operate optimally (see, e.g., [18, Theorem 38.3]). Hence, we leverage upon the previous observations to emphasize the connection of the NRDF to Problem 1.

In the next subsection, we describe the fundamental differences between information-based estimation via NRDF and Bayesian estimation theory.

[2]For example $\rho_t(T^t x_0^n, T^t y^n) = \rho(x_t, y_t)$, $t = 0, \ldots, n$, where $\rho(\cdot, \cdot)$ is a distance metric.

[3]We note that it was recently demonstrated via a counterexample in [22] that the reverse-waterfilling algorithms derived in [21, 23] serve as upper bounds to the optimal solution in the asymptotic regime. Hence, these are also suboptimal in the nonasymptotic regime.

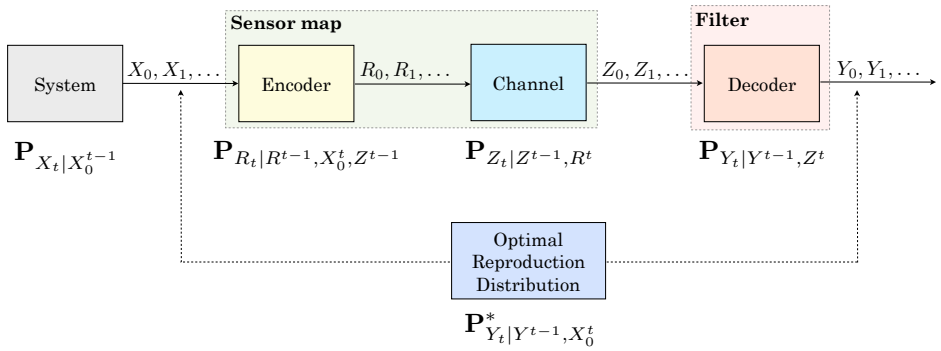

FIG. 1. *Block diagram of Problem* 1 *with probabilistic* {`encoder, channel, decoder`}.

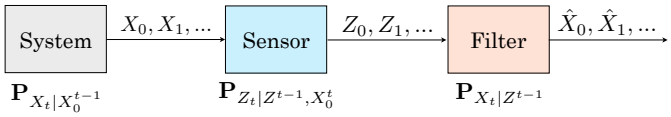

FIG. 2. *Bayesian filtering problem.*

1.2. Relation between Bayesian estimation and estimation using NRDF. In Bayesian filtering [26, 27], one is given a model that generates the unobserved process $X_0^n \triangleq \{X_t : t = 0, \ldots, n\}$, via its conditional distribution $\{\mathbf{P}_{X_t|X_0^{t-1}} : t = 0, \ldots, n\}$, or via discrete-time recursive dynamics, and a model that generates observed data $Z^n \triangleq \{Z^{-1}, Z_0, Z_1, \ldots, Z_n\}$, based on its conditional distribution $\{\mathbf{P}_{Z_t|Z^{t-1},X_0^t} : t = 0, \ldots, n\}$, that is obtained from sensors. At each time t, an estimate of the unobserved process X_t, denoted by $\widehat{X}_t$, is constructed causally, from the observed data Z^{t-1}, for $t = 0, \ldots, n$. Thus, in Bayesian filtering theory, both models which generate the unobserved and observed processes, X_0^n and Z^n, respectively, are given a priori, while at each time t, the estimator is $\widehat{X}_t = g_t(Z^{t-1})$ for some nonanticipative measurable function $g_t(\cdot)$ of the past information Z^{t-1}, often computed recursively, like the Kalman filter. Figure 2 illustrates the block diagram of the Bayesian filtering problem.

On the other hand, in information-based estimation of Problem 1, one is given the distribution $\{\mathbf{P}_{X_t|X_0^{t-1}} : t = 0, \ldots, n\}$ of the process X_0^n and a fidelity criterion, and the objective is to determine the optimal nonanticipative reproduction conditional distribution $\{\mathbf{P}^*_{Y_t|Y^{t-1},X_0^t} : t = 0, \ldots, n\}$ that corresponds to the NRDF, denoted hereinafter by $R^{\mathrm{na}}_{0,n}(D)$, and to realize this distribution by an {`encoder, channel, decoder`} so that the end-to-end MSE distortions (1.3) or (1.4) are met. Thus, in Problem 1, the observation model is constructed by the cascade of the {`encoder, channel`} and the filter is the decoder, which satisfies the end-to-end average distortion (1.3) or (1.4).

1.3. Contributions. The main contributions of this paper are the following:

(R1) A *dynamic recursive expression* for the optimal nonanticipative reproduction conditional distribution, $\{\mathbf{P}^*_{Y_t|Y^{t-1},X_0^t} : t = 0, \ldots, n\}$, which achieves the infimum of the finite-time horizon, NRDF,[4] and some of its properties.

[4]In what follows, when we refer to finite-time horizon NRDF we just say NRDF.

(R2) Applications of (R1) to a *time-varying multidimensional fully observed Gauss–Markov process* X^n with MSE distortion, to derive

(1) a parametric expression of $R_{0,n}^{\mathrm{na}}(D)$ obtained via KKT conditions that is characterized by a time-space reverse-waterfilling;
(2) iterative algorithms that provide, in general, upper bounds to the time-space reverse-waterfilling solution for both distortion constraints (1.3) and (1.4), which perform optimally under certain conditions;
(3) a universal lower bound on the MSE of any causal estimator of the Gauss–Markov process, expressed in terms of the NRDF.

Contribution (R1) generalizes previous work found in [14], in the sense that it holds for any source process conditional distribution $\{\mathbf{P}_{X_t|X_0^{t-1}} : t = 0, \ldots, n\}$, irrespective of whether this is time-varying or Markov, and for any fidelity criterion, such as (1.3). The optimal time-varying reproduction distribution $\{\mathbf{P}^*_{Y_t|Y^{t-1},X_0^t} : t = 0, \ldots, n\}$ of the NRDF is characterized recursively, backward in time, starting at time $t = n$ till time $t = 0$.

Contribution (R2) demonstrates that for time-varying multidimensional fully observed Gauss–Markov processes with MSE distortion, the parametric expression of the NRDF, $R_{0,n}^{\mathrm{na}}(D)$, is characterized via dynamic programming, by a time-space reverse-waterfilling optimization problem. To solve the time-space reverse-waterfilling problem subject to the distortion constraints (1.3) or (1.4), we propose two iterative algorithms which serve, in general, as upper bounds to the optimal value of $R_{0,n}^{\mathrm{na}}(D)$. In some cases, these algorithms perform optimally. The efficiency of these algorithms is exemplified to one numerical simulation where we compare with the optimal numerical solution obtained via semidefinite programming [20]. The Markovian property of the optimal reproduction distribution implies that the optimal distribution is $\{\mathbf{P}^*_{Y_t|Y_{t-1},X_t} : t = 0, \ldots, n\}$. This distribution is realized by an {`encoder, channel, decoder`}, such that the estimation error decays exponentially, under certain conditions. The new recursive estimator is finite-dimensional and ensures the fidelity constraint is met. The time-space reverse-waterfilling implies that given a distortion level, the optimal state estimation is chosen based on an optimal threshold policy, in time and space (dimensions). This is the main fundamental difference compared to the well-known Kalman filter equations. An application of the waterfilling is in sensor selection problems, where the objective is to select, among a set of sensors, only a subset of them to ensure a prespecified estimation error is met.

The universal lower bound on the MSE of any estimator generalizes the well-known bound of a Gaussian random variable (RV) given in [28].

The rest of the paper is structured as follows. In section 2, we introduce the notation used throughout the paper. In section 3, we formulate the NRDF for general processes. In section 4, we describe the form of the optimal nonstationary (time-varying) reproduction distribution of the NRDF. In section 5, we characterize the NRDF for time-varying multidimensional Gauss–Markov processes with MSE distortion, we present examples in the context of realizable filtering theory, and we derive a universal lower bound to the MSE of any estimator in terms of the NRDF. Finally, we draw conclusions and discuss future directions in section 6.

2. Notation. $\mathbb{R} \triangleq (-\infty, \infty)$, $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$, $\mathbb{N} \triangleq \{1, 2, \ldots\}$, $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$, $\mathbb{N}_0^n \triangleq \{0, 1, \ldots, n\}$. For any matrix $A \in \mathbb{R}^{p\times m}$, we denote its transpose by A^{T}. We denote the trace of a square matrix $A \in \mathbb{R}^{p\times p}$ by $\mathrm{trace}(A)$ and by $\mathrm{diag}\{A\}$, the matrix having A_{ii}, $i = 1, \ldots, p$, on its diagonal and zero elsewhere. The set of symmetric positive semidefinite matrices $A \in \mathbb{R}^{p\times p}$ is denoted by $\mathcal{S}_+^{p\times p}$ and its subset

of positive definite matrices by $\mathcal{S}_{++}^{p\times p}$. The statement $A \succeq A'$ (resp., $A \succ A'$) means that $A - A'$ is symmetric positive semidefinite (resp., definite). $\{(\mathcal{X}_n, \mathcal{B}(\mathcal{X}_n)) : n \in \mathbb{Z}\}$ denotes a measurable space, where $\mathcal{X}_n$ is a complete separable metric space or Polish space, and $\mathcal{B}(\mathcal{X}_n)$ is the Borel σ-algebra of subsets of $\mathcal{X}_n$. Points in the product space $\mathcal{X}^{\mathbb{Z}} \triangleq \times_{n\in\mathbb{Z}}\mathcal{X}_n$ are denoted by $x_{-\infty}^{\infty} \triangleq (\ldots, x_{-1}, x_0, x_1, \ldots) \in \mathcal{X}^{\mathbb{Z}}$, and their restrictions to finite coordinates for any $(m, n) \in \mathbb{N}_0 \times \mathbb{N}_0$ by $x_m^n \triangleq (x_m, \ldots, \ldots, x_n) \in \mathcal{X}_m^n,\ n \geq m$. $\mathcal{B}(\mathcal{X}^{\mathbb{Z}}) \triangleq \otimes_{t\in\mathbb{Z}}\mathcal{B}(\mathcal{X}_t)$ denotes the σ-algebra on $\mathcal{X}^{\mathbb{Z}}$ generated by cylinder sets $\{\mathbf{x} = (\ldots, x_{-1}, x_0, x_1, \ldots) \in \mathcal{X}^{\mathbb{Z}} : x_j \in A_j,\ j \in \mathbb{Z}\}, A_j \in \mathcal{B}(\mathcal{X}_j), j \in \mathbb{Z}$. Thus, $\mathcal{B}(\mathcal{X}_m^n)$ denote the σ-algebras of cylinder sets in $\mathcal{X}_m^n$, with bases over $A_j \in \mathcal{B}(\mathcal{X}_j)$, $j \in \{m, m+1, \ldots, n\}$, $(m, n) \in \mathbb{Z}\times\mathbb{Z}$. Given an RV $X : (\Omega, \mathcal{F}) \longmapsto (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we denote by[5] $\mathbf{P}_X(dx) \equiv \mathbf{P}(dx)$ the distribution induced by X on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. $\mathcal{M}(\mathcal{X})$ denotes the set probability distributions on $\mathcal{X}$. Given another RV $Y : (\Omega, \mathcal{F}) \longmapsto (\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ we denote by $\mathbf{P}_{Y|X}(dy|X = x) \equiv \mathbf{P}(dy|x)$ the conditional distribution of RV Y for a fixed $X = x$. Such conditional distributions are equivalently described by stochastic kernels or transition functions [29] $\mathbf{K}(\cdot|\cdot)$ on $\mathcal{B}(\mathcal{Y}) \times \mathcal{X}$, mapping $\mathcal{X}$ into $\mathcal{M}(\mathcal{Y})$ (space of distributions), i.e., $x \in \mathcal{X} \longmapsto \mathbf{K}(\cdot|x) \in \mathcal{M}(\mathcal{Y})$, and such that for every $A \in \mathcal{B}(\mathcal{Y})$, the function $\mathbf{K}(A|\cdot)$ is $\mathcal{B}(\mathcal{X})$-measurable. We denote the set of such stochastic kernels by $\mathcal{Q}(\mathcal{Y}|\mathcal{X})$.

3. NRDF on general alphabets. In this section, we introduce the definition of NRDF from the definition of relative entropy, using general processes which take values in Polish spaces (complete separable metric spaces), that include finite, countable, and continuous alphabet spaces. Throughout, we assume there is a complete probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \in \mathbb{N}_0^n\}, \mathbb{P})$ with complete filtration $\{\mathcal{F}_t : t \in \mathbb{N}_0^n\}$ on which all processes are defined.

Source distribution. The process $X_0^n \triangleq (X_0, X_1, \ldots, X_n)$ is described by the collection of conditional probability distributions $\mathbf{P}_{X_t|X_0^{t-1}}(\cdot|x_0^{t-1}), x_0^{t-1} \in \mathcal{X}_0^{n-1}, t \in \mathbb{N}_0^n$. For each $t \in \mathbb{N}_0^n$, we let $\mathbf{P}_{X_t|X_0^{t-1}}(\cdot|\cdot) \equiv P_t(\cdot|\cdot) \in \mathcal{Q}_t(\mathcal{X}_t|\mathcal{X}_0^{t-1})$, and for $t = 0$, we set $\mathbf{P}_{X_0|X_0^{-1}} = P_0(dx_0)$. We define the probability distribution $\mathbf{P}_{X_0^n}(\cdot) \equiv P_{0,n}(\cdot)$ on $\mathcal{X}_0^n$ by

(3.1)
$$P_{0,n}(A_{0,n}) \triangleq \int_{A_0} P_0(dx_0) \ldots \int_{A_n} P_n\left(dx_n|x_0^{n-1}\right), \quad A_t \in \mathcal{B}(\mathcal{X}_t),\ A_{0,n} = \times_{t=0}^n A_t.$$

Thus, for each $n \in \mathbb{N}_0$, $P_{0,n}(\cdot) \in \mathcal{M}(\mathcal{X}_0^n)$.

Reproduction distribution. The reproduction process $Y^n \triangleq (Y^{-1}, Y_0, Y_1, \ldots, Y_n)$ of $X_0^n \triangleq (X_0, X_1, \ldots, X_n)$ is described by the collection of conditional distributions $\mathbf{P}_{Y_t|Y^{t-1},X_0^t}(\cdot|y^{t-1}, x_0^t), (y^{t-1}, x_0^t) \in \mathcal{Y}^{t-1} \times \mathcal{X}_0^t,\ t \in \mathbb{N}_0^n$, where Y^{-1} is the initial state with fixed distribution $\mathbf{P}_{Y^{-1}} \equiv \mu(dy^{-1})$. For each $t \in \mathbb{N}_0$, we let $\mathbf{P}_{Y_t|Y^{t-1},X_0^t}(\cdot|\cdot,\cdot) = Q_t(\cdot|\cdot,\cdot) \in \mathcal{Q}_t(\mathcal{Y}_t|\mathcal{Y}^{t-1} \times \mathcal{X}_0^t)$, and for $t = 0$, $\mathbf{P}_{Y_0|Y^{-1},X_0} = Q_0(dy_0|y^{-1}, x_0)$. We define the family of conditional probability distributions on $\mathcal{Y}_0^n$ parametrized by $(y^{-1}, x_0^n) \in \mathcal{Y}^{-1} \times \mathcal{X}_0^n$ as follows:

$$\overrightarrow{Q}_{0,n}\left(B_{0,n}\big|y^{-1}, x_0^n\right) \triangleq \int_{B_0} Q_0\left(dy_0|y^{-1}, x_0\right) \ldots$$
(3.2)
$$\int_{B_n} Q_n\left(dy_n|y^{n-1}, x_0^n\right), \quad B_t \in \mathcal{B}(\mathcal{Y}_t),\ B_{0,n} = \times_{t=0}^n B_t.$$

[5] The subscript notation is often omitted when it is clear from the arguments of the distribution.

We note that the family of probability distributions $\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)$ parametrized by $(y^{-1},x_0^n)\in\mathcal{Y}^{-1}\times\mathcal{X}_0^n$ satisfies the following *consistency condition* (CC):

CC. For any $F\in\mathcal{B}(\mathcal{Y}_0^n)$, then, $\overrightarrow{Q}_{0,n}(F|y^{-1},x_0^n)$ is a $\mathcal{B}(\mathcal{Y}^{-1})\otimes\mathcal{B}(\mathcal{X}_0^n)$-measurable function of $(y^{-1},x_0^n)\in\mathcal{Y}^{-1}\times\mathcal{X}_0^n$.

Moreover, from [30], for any family of conditional distributions

$$\mathbf{P}_{Y_0^n|Y^{-1},X_0^n}(\cdot|y^{-1},x_0^n)\equiv\mathbf{P}(\cdot|y^{-1},x_0^n)$$

on $\mathcal{Y}_0^n$ parametrized by $(y^{-1},x_0^n)\in\mathcal{Y}^{-1}\times\mathcal{X}_0^n$ that satisfies CC there exists a sequence of stochastic kernels $Q_t(\cdot|\cdot,\cdot)\in\mathcal{Q}_t(\mathcal{Y}_t|\mathcal{Y}^{t-1}\times\mathcal{X}_0^t), t\in\mathbb{N}_0^n$, such that $\mathbf{P}(B_{0,n}|y^{-1},x_0^n)$ is defined by the right-hand side of (3.2). We define the set of probability distributions on $\mathcal{Y}_0^n$ conditioned on $(Y^{-1},X_0^n)=(y^{-1},x_0^n)\in\mathcal{Y}^{-1}\times\mathcal{X}_0^n$ that satisfies CC by

$$\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)\triangleq\left\{\mathbf{P}\left(\cdot\middle|y^{-1},x_0^n\right)\in\mathcal{M}(\mathcal{Y}_0^n):\ \text{such that CC holds}\right\}. \tag{3.3}$$

Thus, for each $n\in\mathbb{N}_0$, $\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)$, $(y^{-1},x_0^n)\in\mathcal{Y}^{-1}\times\mathcal{X}_0^n$. Given a $P_{0,n}(\cdot)\in\mathcal{M}(\mathcal{X}_0^n)$, a $\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)$, and a fixed distribution $\mu(dy^{-1})$, we define the following distributions:

- *The joint distribution* on $\mathcal{X}_0^n\times\mathcal{Y}_0^n$ given $Y^{-1}=y^{-1}$ that is defined by

$$\begin{aligned}\mathbf{P}^{\overrightarrow{Q}}\left(A_{0,n}\times B_{0,n}\middle|y^{-1}\right)&\triangleq\left(P_{0,n}\otimes\overrightarrow{Q}_{0,n}\right)\left(\times_{t=0}^n(A_t\times B_t)\middle|y^{-1}\right)\\&=\int_{A_0}P_0(dx_0)\int_{B_0}Q_0\left(dy_0\middle|y^{-1},x_0\right)\dots\\&\quad\int_{A_n}P_n\left(dx_n\middle|x_0^{n-1}\right)\int_{B_n}Q_n\left(dy_n\middle|y^{n-1},x_0^n\right).\end{aligned} \tag{3.4}$$

- *The marginal distribution* on $\mathcal{Y}_0^n$ given $Y^{-1}=y^{-1}$ that is defined by

$$\begin{aligned}\Pi_{0,n}^{\overrightarrow{Q}}\left(B_{0,n}\middle|y^{-1}\right)&\triangleq\int_{B_{0,n}}\int_{\mathcal{X}_0^n}\left(P_{0,n}\otimes\overrightarrow{Q}_{0,n}\right)\left(dx_0^n,dy_0^n\middle|y^{-1}\right)\\&=\int_{B_{0,n}}\Pi_0^{\overrightarrow{Q}}\left(dy_0\middle|y^{-1}\right)\dots\Pi_n^{\overrightarrow{Q}}\left(dy_n\middle|y^{n-1}\right).\end{aligned}$$

- *The product probability distribution* $\overrightarrow{\Pi}_{0,n}^{\overrightarrow{Q}}(\cdot|y^{-1}):\mathcal{B}(\mathcal{X}_0^n)\otimes\mathcal{B}(\mathcal{Y}_0^n)\longmapsto[0,1]$ conditioned on $Y^{-1}=y^{-1}$ is defined by

$$\begin{aligned}&\overrightarrow{\Pi}_{0,n}^{\overrightarrow{Q}}\left(A_{0,n}\times B_{0,n}\middle|y^{-1}\right)\triangleq\left(P_{0,n}\times\Pi_{0,n}^{\overrightarrow{Q}}\right)\left(\times_{t=0}^n(A_t\times B_t)\middle|y^{-1}\right)\\&\quad=\int_{A_0}P_0(dx_0)\int_{B_0}\Pi_0^{\overrightarrow{Q}}\left(dy_0\middle|y^{-1}\right)\dots\int_{A_n}P_n\left(dx_n\middle|x_0^{n-1}\right)\int_{B_n}\Pi_n^{\overrightarrow{Q}}(dy_n|y^{n-1}).\end{aligned}$$

Using the distributions above, we define the relative entropy between the joint distribution $\mathbf{P}^{\overrightarrow{Q}}(dx_0^n,dy_0^n|y^{-1})$ and the product distribution $\overrightarrow{\Pi}_{0,n}^{\overrightarrow{Q}}(dx_0^n,dy_0^n|y^{-1})$, av-

eraged over the initial distribution $\mu(dy^{-1})$, as follows:

$$\mathbb{D}\left(P_{0,n}\otimes\overrightarrow{Q}_{0,n}||\overrightarrow{\Pi}_{0,n}^{\overrightarrow{Q}}\right)=\int_{\mathcal{X}_0^n\times\mathcal{Y}^n}\log\left(\frac{P_{0,n}(\cdot)\otimes\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)}{P_{0,n}(\cdot)\otimes\Pi_{0,n}^{\overrightarrow{Q}}(\cdot|y^{-1})}(x_0^n,y_0^n)\right)$$

$$\times P_{0,n}\left(dx_0^n\right)\otimes\overrightarrow{Q}_{0,n}\left(dy_0^n|y^{-1},x_0^n\right)\otimes\mu\left(dy^{-1}\right)\in[0,\infty] \tag{3.5}$$

$$\overset{(a)}{=}\int_{\mathcal{X}_0^n\times\mathcal{Y}^n}\log\left(\frac{\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)}{\Pi_{0,n}^{\overrightarrow{Q}}(\cdot|y^{-1})}(y_0^n)\right)$$

$$\times P_{0,n}\left(dx_0^n\right)\otimes\overrightarrow{Q}_{0,n}\left(dy_0^n|y^{-1},x_0^n\right)\otimes\mu(dy^{-1})$$

$$\overset{(b)}{=}\sum_{t=0}^{n}\int_{\mathcal{X}_0^t\times\mathcal{Y}^t}\log\left(\frac{Q_t(\cdot|y^{t-1},x_0^t)}{\Pi_t^{\overrightarrow{Q}}(\cdot|y^{t-1})}(y_t)\right)$$

$$\times Q_t\left(dy_t|y^{t-1},x_0^t\right)\otimes P_t\left(dx_t|x^{t-1}\right)\otimes\mathbf{P}^{\overrightarrow{Q}}\left(dx^{t-1},dy^{t-1}\right)$$

$$=\sum_{t=0}^{n}I\left(X^t;Y_t|Y^{t-1}\right) \tag{3.6}$$

$$\equiv\mathbb{I}_{0,n}\left(P_{0,n},\overrightarrow{Q}_{0,n}\right), \tag{3.7}$$

where (a), (b) are due to the chain rule of relative entropy (see [30]), and $I(X_0^t;Y_t|Y^{t-1})$ is the conditional mutual information between X^t and Y_t, conditioned on Y^{t-1}. In (3.7) the notation $\mathbb{I}_{0,n}(\cdot,\cdot)$ indicates the functional dependence on $\{P_{0,n},\overrightarrow{Q}_{0,n}\}$ (the dependence on $\mu(dy^{-1})$ is omitted).

Using the previous formulation, the following functional properties hold:

(P1) By [30, Theorem 5] the set of distributions $\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)$ is convex.

(P2) By [30, Theorem 6], $\mathbb{I}_{0,n}(P_{0,n},\cdot)$ is a convex functional of $\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x_0^n)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)$.

We define the NRDF using the above definition of relative entropy as follows.

Definition 3.1 (NRDF).

(1) *Given the distortion function* (1.1) *of reproducing* x_t *by* $y_t, t=0,1,\ldots,n$, *define the set of reproduction distributions that satisfy the fidelity criterion by*

$$\overrightarrow{\mathcal{Q}}_{0,n}(D)\triangleq\left\{\overrightarrow{Q}_{0,n}\left(\cdot|y^{-1},x_0^n\right)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n):\frac{1}{n+1}\mathbf{E}_{\mu}^{\overrightarrow{Q}}\left\{d_{0,n}(X_0^n,Y^n)\right\}\leq D\right\},\ D\geq 0,$$

where $\mathbf{E}_{\mu}^{\overrightarrow{Q}}\{\cdot\}$ *indicates that the joint distribution is induced by* $\{P_{0,n}(dx^n),\overrightarrow{Q}_{0,n}(dy_0^n|y^{-1},x^n),\mu(dy^{-1})\}$ *defined by* (3.4). *The* NRDF *is defined by*

$$R_{0,n}^{\mathrm{na}}(D)\triangleq\inf_{\overrightarrow{Q}_{0,n}(dy_0^n|y^{-1},x_0^n)\in\overrightarrow{\mathcal{Q}}_{0,n}(D)}\mathbb{I}_{0,n}\left(P_{0,n},\overrightarrow{Q}_{0,n}\right),\quad D\geq 0. \tag{3.8}$$

(2) *Given the distortion function* (1.2), *define (similarly to* (1)*) the fidelity criterion by*

$$\overrightarrow{\mathcal{Q}}_{0,n}(D_0,D_1,\ldots,D_n)$$

$$\triangleq\left\{\overrightarrow{Q}_{0,n}\left(\cdot|y^{-1},x_0^n\right)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n):\mathbf{E}_{\mu}^{\overrightarrow{Q}}\left\{d_t\left(X_0^t,Y^t\right)\right\}\leq D_t\ \forall t\in\mathbb{N}_0^n\right\},$$

where $D_t \in [0,\infty)\ \forall t \in \mathbb{N}_0^n$. *The* NRDF *with average distortion at each time is defined by*

$$R_{0,n}^{\mathrm{na}}(D_0, D_1, \ldots, D_n) \triangleq \inf_{\overrightarrow{Q}_{0,n}(dy_0^n|y^{-1},x_0^n) \in \overrightarrow{\mathcal{Q}}_{0,n}(D_0,D_1,\ldots,D_n)} \mathbb{I}_{0,n}\left(P_{0,n}, \overrightarrow{Q}_{0,n}\right) \tag{3.9}$$

for $D_t \in [D_t^{\min}, D_t^{\max}] \subseteq [0,\infty]$ *for* $t = 0,\ldots,n$, *similarly as above.*

Next, we state some properties of the NRDF in Definition 3.1.

(P3) By (P1) the set $\overrightarrow{\mathcal{Q}}_{0,n}(D)$ is a convex subset of $\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}_0^n)$.

(P4) By (P2) and (P3) the NRDF defined by (3.8) is a convex optimization problem.

It should be mentioned that sufficient conditions for existence of an optimal reproduction distribution $\overrightarrow{Q}_{0,n}^*(dy_0^n|y^{-1},x_0^n)$ that achieves the infimum of the NRDF defined by (3.8) are identified in [14, Theorems III.3, III.4] by means of weak*-convergence and compactness of probability measures in appropriate function spaces. Similar conditions are also identified in [30, Lemma 12, Theorem 14] using weak-convergence and compactness of probability measures via Prohorov's theorems.

For completeness, in the next remark we discuss the precise relation between the NRDF and nonanticipatory ϵ-entropy [1] and their fundamental differences with respect to Shannon's definition of classical RDF [3].

Remark 1 (RDF and nonanticipatory ϵ-entropy). Consider a distribution $P_{0,n}(\cdot) \in \mathcal{M}(\mathcal{X}_0^n)$ and a reproduction distribution $\mathbf{P}(dy_0^n|y^{-1},x_0^n) \triangleq Q_{0,n}^{\mathrm{nc}}(dy_0^n|y^{-1},x_0^n) \in \mathcal{M}(\mathcal{Y}_0^n)$, $(y^{-1},x_0^n) \in \mathcal{Y}^{-1} \times \mathcal{X}_0^n$ which does not satisfy the CC. Then, the conditional distribution on $\mathcal{Y}_0^n$ given $Y^{-1} = y^{-1}$ and the joint distribution on $\mathcal{X}_0^n \times \mathcal{Y}_0^n$ are introduced as follows:

$$\Pi_{0,n}^{Q^{\mathrm{nc}}}\left(dy_0^n\middle|y^{-1}\right) = \int_{\mathcal{X}_0^n} Q_{0,n}^{\mathrm{nc}}\left(dy_0^n\middle|y^{-1},x_0^n\right) \otimes P_{0,n}\left(dx_0^n\right), \tag{3.10}$$

$$\mathbf{P}^{Q^{\mathrm{nc}}}\left(dx_0^n, dy_0^n\middle|y^{-1}\right) = P_{0,n}\left(dx_0^n\right) \otimes Q_{0,n}^{\mathrm{nc}}\left(dy_0^n\middle|y^{-1},x_0^n\right). \tag{3.11}$$

Define the set of conditional distributions that satisfy the fidelity criterion by

$$\mathcal{Q}_{0,n}^{\mathrm{nc}}(D) \triangleq \left\{Q_{0,n}^{\mathrm{nc}}\left(dy_0^n\middle|y^{-1},x_0^n\right) \in \mathcal{M}(\mathcal{Y}_0^n) : \frac{1}{n+1}\mathbf{E}_\mu^{Q^{\mathrm{nc}}}\left\{d_{0,n}(X_0^n,Y^n)\right\} \le D\right\},\ D \ge 0.$$

The classical RDF [3] is defined by

$$R_{0,n}(D) \triangleq \inf_{Q_{0,n}^{\mathrm{nc}}(dy_0^n|y^{-1},x_0^n) \in \mathcal{Q}_{0,n}^{\mathrm{nc}}(D)} I\left(X_0^n; Y_0^n\middle|Y^{-1}\right), \tag{3.12}$$

where $I(X_0^n;Y_0^n|Y^{-1})$ is the conditional mutual information defined by

$$\begin{aligned} I\left(X_0^n; Y_0^n\middle|Y^{-1}\right) &\triangleq \int_{\mathcal{X}_0^n \times \mathcal{Y}^n} \log\left(\frac{Q_{0,n}^{\mathrm{nc}}(\cdot|y^{-1},x_0^n)}{\Pi_{0,n}^{Q^{\mathrm{nc}}}(\cdot|y^{-1})}(y_0^n)\right) \\ &\quad \times P_{0,n}\left(dx_0^n\right) \otimes Q_{0,n}^{\mathrm{nc}}\left(dy_0^n\middle|y^{-1},x_0^n\right) \otimes \mu\left(dy^{-1}\right) \qquad (3.13) \\ &\equiv \mathbb{I}_{0,n}^{\mathrm{nc}}(P_{0,n}, Q_{0,n}^{\mathrm{nc}}). \qquad (3.14) \end{aligned}$$

By Bayes' rule we have the decomposition $Q_{0,n}^{\mathrm{nc}}(dy_0^n|y^{-1},x_0^n) = \otimes_{t=0}^n Q_t^{\mathrm{nc}}(dy_t|y^{t-1},x_0^n)$. Therefore, in general, the solution to the classical RDF cannot be used to construct causal estimators, because for each t, the reproduction distribution $Q_t^{\mathrm{nc}}(dy_t|y^{t-1},x_0^n)$

depends on futures symbols $(x_{t+1}, \ldots, x_n)$. In view of this technicality, Gorbunov and Pinsker in [1] introduced the nonanticipatory ϵ-entropy, defined as follows:

$$R^{\varepsilon}_{0,n}(D) \triangleq \inf_{\mathcal{Q}^{\mathrm{nc}}_{0,n}(D):Q^{\mathrm{nc}}_{0,t}(dy^t_0|y^{-1},x^n_0)=Q^{GP}_{0,t}(dy^t_0|y^{-1},x^t_0),\ t\in\mathbb{N}^n_0} I\left(X^n_0;Y^n_0\big|Y^{-1}\right). \tag{3.15}$$

The extra conditional independence condition $Q^{\mathrm{nc}}_{0,t}(dy^t_0|y^{-1},x^n_0) = Q^{GP}_{0,t}(dy^t_0|y^{-1},x^t_0)$, $t \in \mathbb{N}^n_0$, that is imposed in the definition of classical RDF (3.12) implies CC. This follows from the following equivalent statements of conditional independence shown in [31, Lemma 6.2]:

MC1. $Q^{\mathrm{nc}}_{0,n}(dy^n_0|y^{-1},x^n_0) = \overrightarrow{Q}_{0,n}(dy^n_0|y^{-1},x^n_0) = \otimes^n_{t=0} Q_t(dy_t|y^{t-1},x^t_0)\ \ \forall n \in \mathbb{N}_0$;

MC2. $Q^{\mathrm{nc}}_t(dy_t|y^{t-1},x^t_0,x^n_{t+1}) = Q_t(dy_t|y^{t-1},x^t_0)$ for each $t \in \mathbb{N}^{n-1}_0\ \ \forall n \in \mathbb{N}_0$;

MC3. $P_t(dx_{t+1}|x^t_0,y^t) = P_t(dx_{t+1}|x^t_0)$ for each $t \in \mathbb{N}^{n-1}_0\ \ \forall n \in \mathbb{N}_0$;

MC4. $Q^{\mathrm{nc}}_{0,t}(dy^t_0|y^{-1},x^t_0,x^n_{t+1}) = \overrightarrow{Q}_{0,t}(dy^t_0|y^{-1},x^t_0)$ for each $t \in \mathbb{N}^{n-1}_0\ \ \forall n \in \mathbb{N}_0$.

Since MC1–MC4 are equivalent statements, then it can be shown that the NRDF defined by (3.8) is equivalent to the nonanticipatory ϵ-entropy defined by (3.15), that is, $R^{\mathrm{na}}_{0,n}(D) = R^{\varepsilon}_{0,n}(D)$.

4. Optimal nonstationary reproduction distribution. In this section, we describe the form of the optimal nonstationary (time-varying) reproduction distribution that achieves the infimum in (3.8) (assuming it exists).

First, we introduce the finite-time horizon nonanticipative distortion rate function, hereinafter denoted by $D_{0,n}(R^{\mathrm{na}})$, defined as

$$D_{0,n}(R^{\mathrm{na}}) = \inf_{\overrightarrow{Q}_{0,n}(dy^n_0|y^{-1},x^n_0):\ \frac{1}{n+1}\mathbb{I}_{0,n}(P_{0,n},\overrightarrow{Q}_{0,n})\le R^{\mathrm{na}}} \mathbf{E}^{\overrightarrow{Q}}_{\mu}\left\{d_{0,n}(X^n_0,Y^n)\right\},\ R^{\mathrm{na}} \in [0,\infty). \tag{4.1}$$

Next, we state certain important properties of $R^{\mathrm{na}}_{0,n}(D)$ that follow directly from properties (P3), (P4) (following *mutatis mutandis* the derivation in [32, Theorem 7.1, p. 45]), and we do the same for $D_{0,n}(R^{\mathrm{na}})$.

(P5) $R^{\mathrm{na}}_{0,n}(D)$ and $D_{0,n}(R^{\mathrm{na}})$ are nonincreasing functions of $D \in [0,\infty)$ and $R^{\mathrm{na}} \in [0,\infty)$, respectively, and the function $R^{\mathrm{na}}_{0,n}(D)$ is convex in $D \in [0,\infty)$.

(P6) $R^{\mathrm{na}}_{0,n}(D)$ is continuous on $D \in (0,\infty)$, and if $R^{\mathrm{na}}_{0,n}(0) < \infty$, then it is continuous on $D \in [0,\infty)$.

Note that (P6) follows from the fact that a bounded and convex function is continuous; hence by the nonincreasing property in (P5), $R^{\mathrm{na}}_{0,n}(D)$ is bounded outside the neighborhood of $D = 0$ and continuous on $(0,\infty)$. Moreover, if $R^{\mathrm{na}}_{0,n}(0) < \infty$, then $R^{\mathrm{na}}_{0,n}(D)$ is bounded and hence continuous on $[0,\infty)$.

Assume an optimal reproduction distribution $\overrightarrow{Q}^*_{0,n}(dy^n_0|y^{-1},x^n_0)$ that achieves the infimum of the NRDF defined by (3.8) exists. If, in addition, there exists an interior point in the set $\overrightarrow{\mathcal{Q}}_{0,n}(D)$, then the NRDF is a convex optimization problem that can be reformulated using the Lagrange duality theorem [33, Theorem 1, pp. 224–225], as an unconstrained problem as follows:

$$\begin{aligned} R^{\mathrm{na}}_{0,n}(D) = \sup_{s\le 0} \inf_{\overrightarrow{Q}_{0,n}(\cdot|y^{-1},x^n_0)\in\mathcal{M}^{\mathrm{CC}}(\mathcal{Y}^n_0)} \\ \left\{\mathbb{I}_{0,n}\left(P_{0,n},\overrightarrow{Q}_{0,n}\right) - s\Big(\mathbf{E}^{\overrightarrow{Q}}_{\mu}\left\{d_{0,n}(X^n_0,Y^n)\right\} - D(n+1)\Big)\right\}. \end{aligned} \tag{4.2}$$

In what follows, we state a theorem that generalizes the result of [14, section IV], which was developed under the assumption that the optimal reproduction distributions $Q^*_t(dy_t|y^{t-1},x^t_0) = Q^*(dy_t|y^{t-1},x^t_0)\ \forall t \in \mathbb{N}^n_0$ are identical, or that the joint

process $\{(X_t, Y_t) : \forall t \in \mathbb{N}_0^n\}$ is stationary. The next theorem computes the elements $Q_t^*(dy_t|y^{t-1}, x_0^t) = Q^*(dy_t|y^{t-1}, x_0^t)\ \forall t \in \mathbb{N}_0^n$ recursively moving backward in time. This result is applied in the subsequent analysis to compute the NRDF, $R_{0,n}^{\mathrm{na}}(D)$, of time-varying multidimensional Gauss–Markov processes.

THEOREM 4.1 (optimal nonstationary reproduction distributions). *Suppose there exists* $\overrightarrow{Q}_{0,n}^*(\cdot|y^{-1}, x_0^n) \in \overrightarrow{\mathcal{Q}}_{0,n}(D)$, *which solves* (3.8) *for* $D \in [D_{\min}, D_{\max}]$, *identity* (4.2) *holds, and* $\mathbb{I}_{0,n}(P_{0,n}, \overrightarrow{Q}_{0,n})$ *is Gâteaux differentiable in every direction of* $\{Q_t(\cdot|y^{t-1}, x_0^t) : t \in \mathbb{N}_0^n\}$ *for a fixed* $P_{0,n}(\cdot) \in \mathcal{M}(\mathcal{X}_0^n)$ *and* $\mu(dy^{-1}) \in \mathcal{M}(\mathcal{Y}^{-1})$. *Then the following hold:*

(1) *The optimal nonstationary reproduction distributions denoted by*

$$\{Q_t^*(\cdot|y^{t-1}, x_0^t) \in \mathcal{M}(\mathcal{Y}_t) : t \in \mathbb{N}_0^n\}$$

are given by the following recursive equations backward in time.

For $t = n$,

$$Q_n^*\left(dy_n\middle|y^{n-1}, x_0^n\right) = \frac{e^{s\rho_n(T^n x_0^n, T^n y^n)}\Pi_n^{\overrightarrow{Q}^*}(dy_n|y^{n-1})}{\int_{\mathcal{Y}_n} e^{s\rho_n(T^n x_0^n, T^n y^n)}\Pi_n^{\overrightarrow{Q}^*}(dy_n|y^{n-1})}. \tag{4.3}$$

For $t = n-1, n-2, \ldots, 0$,

$$Q_t^*\left(dy_t\middle|y^{t-1}, x_0^t\right) = \frac{e^{s\rho_t(T^t x_0^n, T^t y^n) - g_{t,n}(x_0^t, y^t)}\Pi_t^{\overrightarrow{Q}^*}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t(T^t x_0^n, T^t y^n) - g_{t,n}(x_0^t, y^t)}\Pi_t^{\overrightarrow{Q}^*}(dy_t|y^{t-1})}, \tag{4.4}$$

where $s < 0$ *is the Lagrange multiplier, and* $\Pi_t^{Q^*}(\cdot|y^{t-1}) \in \mathcal{M}(\mathcal{Y}_t)$ *and* $g_{t,n}(x_0^t, y^t)$ *are defined by*

$$\begin{aligned} &g_{n,n}(x_0^n, y^n) = 0,\\ &g_{t,n}\left(x_0^t, y^t\right) \triangleq -\int_{\mathcal{X}_{t+1}} P_{t+1}\left(dx_{t+1}\middle|x_0^t\right)\\ &\qquad\times \log\left(\int_{\mathcal{Y}_{t+1}} e^{s\rho_{t+1}(T^{t+1}x_0^n, T^{t+1}y^n) - g_{t+1,n}(x_0^{t+1}, y^{t+1})}\Pi_{t+1}^{\overrightarrow{Q}^*}\left(dy_{t+1}\middle|y^t\right)\right). \end{aligned}$$

(2) *The* NRDF *is given by*

$$\begin{aligned} R_{0,n}^{\mathrm{na}}(D) &= sD(n+1)\\ &\quad - \sum_{t=0}^{n}\int_{\mathcal{X}_0^t\times\mathcal{Y}^{t-1}}\left\{\int_{\mathcal{Y}_t} g_{t,n}\left(x_0^t, y^t\right) Q_t^*\left(dy_t\middle|y^{t-1}, x_0^t\right)\right.\\ &\qquad\left. + \log\left(\int_{\mathcal{Y}_t} e^{s\rho_t(T^t x_0^n, T^t y^n) - g_{t,n}(x_0^t, y^t)}\Pi_t^{\overrightarrow{Q}^*}\left(dy_t\middle|y^{t-1}\right)\right)\right\}\\ &\quad \otimes P_t\left(dx_t\middle|x_0^{t-1}\right)\otimes\left(P_{0,t-1}\otimes\overrightarrow{Q}_{0,t-1}^*\right)\left(dx_0^{t-1}, dy_0^{t-1}\middle|y^{-1}\right)\otimes\mu\left(dy^{-1}\right). \end{aligned} \tag{4.5}$$

(3) *If* $R_{0,n}^{\mathrm{na}}(D) > 0$, *then* $s < 0$, *and*

$$\frac{1}{n+1}\sum_{t=0}^{n}\int_{\mathcal{X}_0^t\times\mathcal{Y}^t}\rho_t\left(T^t x_0^n, T^t y^n\right)\left(P_{0,t}\otimes\overrightarrow{Q}_{0,t}^*\right)\left(dx_0^t, dy_0^t\middle|y^{-1}\right)\otimes\mu\left(dy^{-1}\right) = D. \tag{4.6}$$

Proof. We outline the derivation. The minimization over $\{Q_t(\cdot|y^{t-1}, x_0^t) : t \in \mathbb{N}_0^n\}$ in (4.2) is a nested optimization problem. Hence, we apply dynamic programming, backward in time. Then, we carry out the infimum starting at the last stage over $Q_n(\cdot|y^{n-1}, x_0^n) \in \mathcal{M}(\mathcal{Y}_n)$ and sequentially move backward in time to determine $Q_n^*(\cdot|y^{n-1}, x_0^n), Q_{n-1}^*(\cdot|y^{n-2}, x_0^{n-1}), \ldots, Q_0^*(\cdot|y^{-1}, x_0)$, by performing the Gâteaux differential at each direction of $Q_n(\cdot|y^{n-1}, x_0^n), Q_{n-1}(\cdot|y^{n-2}, x_0^{n-1}), \ldots, Q_0(\cdot|y^{-1}, x_0)$. ∎

By utilizing Theorem 4.1, then, for a given distribution $P_{0,n}(\cdot) \in \mathcal{M}(\mathcal{X}_0^n)$, we can identify the dependence of the optimal nonstationary reproduction distribution on past and present symbols of the information process $\{X_t : t \in \mathbb{N}_0^n\}$, called the information structures (IS) of the optimal nonstationary reproduction distribution of (3.8).

IS of the optimal nonstationary reproduction distribution:

(IS1) The dependence of $Q_n^*(dy_n|y^{n-1}, x_0^n)$ on $x^n \in \mathcal{X}_0^n$ is determined by the dependence of $\rho_n(T^n x_0^n, T^n y^n)$ on $x_0^n \in \mathcal{X}_0^n$ as follows:

(IS1.1) If $\rho_t(T^t x_0^n, T^i y^n) = \bar{\rho}(x_t, y^t)\ \forall t \in \mathbb{N}_0^n$, then, at $t = n$, $Q_n^*(dy_n|y^{n-1}, x_0^n) = Q_n^*(dy_n|y^{n-1}, x_n)$, while for $t = n-1, n-2, \ldots, 0$, the dependence of $Q_t^*(dy_t|y^{t-1}, x_0^t)$ on $x_0^t \in \mathcal{X}_0^t$ is determined from the dependence of $g_{t,n}(x_0^t, y^t)$ on $x_0^t \in \mathcal{X}_0^t$.

(IS1.2) If $P_t(dx_t|x_0^{t-1}) = P_t(dx_t|x_{t-1-L}^{t-1})$, where L is a nonnegative finite integer, and $\rho_t(T^t x_0^n, T^t y^n) = \bar{\rho}(x_{t-N}^t, y_t)$, where N is a nonnegative finite integer $\forall t \in \mathbb{N}_0^n$, then, $Q_t^*(dy_t|\ y^{t-1}, x_0^t) = Q_t^{J,*}(dy_t|y^{t-1}, x_{t-J}^t)\ \forall t \in \mathbb{N}_0^n$, where $J = \max\{N, L\}$.

If $L = N = 1$, i.e., $\rho_t(T^t x_0^n, T^t y^n) = \rho_t^{SL}(x_t, y_t)$, $\forall t \in \mathbb{N}_0^n$, and the source is Markov, then $Q_t^*(dy_t|y^{t-1}, x_0^t) = Q_t^{1,*}(dy_t|y^{t-1}, x_t)\ \forall t \in \mathbb{N}_0^n$, and NRDF is characterized by the following optimization problem:

$$R_{0,n}^{\mathrm{na}}(D) \triangleq \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^1(D)} \mathbf{E}_\mu^{\overrightarrow{Q}^1} \left\{ \sum_{t=0}^n \log \left(\frac{Q_t^1(\cdot|Y^{t-1}, X_t)}{\Pi_t^{\overrightarrow{Q}^1}(\cdot|Y^{t-1})}(Y_t) \right) \right\} \tag{4.7}$$

$$\equiv \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^1(D)} \sum_{t=0}^n I\left(X_t; Y_t \big| Y^{t-1}\right), \tag{4.8}$$

where the transition probability distribution of Y_t given $Y^{t-1} = y^{t-1}$ is given by

$$\Pi_t^{\overrightarrow{Q}^1}\left(dy_t \big| y^{t-1}\right) = \int_{\mathcal{X}_t} Q_t^1\left(dy_t \big| y^{t-1}, x_t\right) \otimes \mathbf{P}^{\overrightarrow{Q}^1}\left(dx_t \big| y^{t-1}\right)\ \forall t \in \mathbb{N}_0^n, \tag{4.9}$$

and the fidelity criterion is defined by

$$\overrightarrow{\mathcal{Q}}_{0,n}^1(D) \triangleq \left\{ Q_t^1\left(dy_t \big| y^{t-1}, x_t\right), t \in \mathbb{N}_0^n :\ \frac{1}{n+1} \mathbf{E}_\mu^{\overrightarrow{Q}^1} \left\{ \sum_{t=0}^n \rho_t^{SL}(X_t, Y_t) \right\} \leq D \right\}. \tag{4.10}$$

(IS2) If $g_{t,n}(x_0^t, y^t) = \hat{g}_{t,n}(x_0^t, y^{t-1}), \forall t \in \mathbb{N}_0^n$, then, the optimal reproduction distribution (4.4) reduces to

$$Q_t^*\left(dy_t \big| y^{t-1}, x_0^t\right) = \frac{e^{s\rho_t(T^t x_0^n, T^t y^n)} \Pi_t^{\overrightarrow{Q}^*}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t(T^t x_0^n, T^t y^n)} \Pi_t^{\overrightarrow{Q}^*}(dy_t|y^{t-1})}\ \forall t \in \mathbb{N}_0^n. \tag{4.11}$$

(IS3) If $g_{t,n}(x_0^t, y^t) = \hat{g}_{t,n}(x_0^t, y^{t-1})$, $\rho_t(T^t x_0^n, T^t y^n) = \rho_t^{SL}(x_t, y_t)$ $\forall t \in \mathbb{N}_0^n$, and X_0^n is Markov, i.e., $P_t(dx_t|x_0^{t-1}) = P_t(dx_t|x_{t-1})$ $\forall t \in \mathbb{N}_0^n$, then (4.11) reduces to

(4.12)
$$Q_t^*\left(dy_t\middle|y^{t-1}, x_0^t\right) = Q_t^{1,*}\left(dy_t\middle|y^{t-1}, x_t\right) \frac{e^{s\rho_t^{SL}(x_t,y_t)}\overrightarrow{\Pi}_t^{1,*}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s\rho_t^{SL}(x_t,y_t)}\overrightarrow{\Pi}_t^{1,*}(dy_t|y^{t-1})} \quad \forall t \in \mathbb{N}_0^n,$$

which is Markov in X_0^n. Again, we cannot determine from the above characterization, whether at each t, the optimal reproduction distribution $Q_t^{M,*}(dy_t|y^{t-1}, x_t)$ depends on limited memory on past reproductions y^{t-1}.

Remark 2. Note that the discussion of this section holds, even if the average distortion (1.3) is replaced by (1.4), by simply replacing $s\frac{1}{n+1}(\mathbf{E}_\mu^{\overrightarrow{Q}}\{d_{0,n}(X_0^n, Y^n)\} - D)$ in (4.2) with $\sum_{t=0}^n s_t(\mathbf{E}_\mu^{\overrightarrow{Q}}\{\rho_t(T^t X_0^n, T^t Y^n)\} - D_t)$, where s_t are the Lagrange multipliers for $t \in \mathbb{N}_0^n$.

Remark 3. If the σ-algebra generated by the initial state Y^{-1} is the trivial $\sigma\{Y^{-1}\} = \{\Omega, \emptyset\}$, then in Definition 3.1, the payoff is replaced by

$$\sum_{t=0}^n I\left(X_0^t; Y_t\middle|Y^{t-1}\right) = I(X_0; Y_0) + \sum_{t=1}^n I\left(X_0^t; Y_t\middle|Y_0, Y_1, \ldots, Y_{t-1}\right). \tag{4.13}$$

Hence, all previous material and subsequent material can be specialized accordingly.

In the next section, we use Theorem 4.1 and the above observations to derive $R_{0,n}^{\mathrm{na}}(D)$ for the multidimensional Gauss–Markov processes X_0^n.

5. NRDF of time-varying multidimensional Gauss–Markov processes. In this section, we apply Theorem 4.1 to the following time-varying multidimensional Gauss–Markov process, described in state-space form.

Definition 5.1 (time-varying multidimensional Gauss–Markov process). *The source is a time-varying $\mathbb{R}^p$-valued Gauss–Markov process defined by the recursion*

$$X_{t+1} = A_t X_t + W_t,\ X_0 = x_0,\ \forall t \in \mathbb{N}_0^{n-1}, \tag{5.1}$$

where $A_t \in \mathbb{R}^{p\times p}$ $\forall t \in \mathbb{N}_0^{n-1}$ is a nonrandom matrix. We assume

(G1) *$X_0 \in \mathbb{R}^p$ is Gaussian $\mathcal{N}(0; K_{X_0})$;*

(G2) *$\{W_t : t \in \mathbb{N}_0^n\}$ is an $\mathbb{R}^p$-valued independent and identically distributed (*IID*) Gaussian $\mathcal{N}(0; K_{W_t})$, $K_{W_t} \in S_+^{p\times p}$ sequence, independent of (X_0, Y^{-1}).*

(G3) *The distortion function is the sum of squared errors, defined by $d_{0,n}(x_0^n, y^n) \triangleq \sum_{t=0}^n \rho_t(T^t x_0^n, T^t y^n) = \sum_{t=0}^n \|x_t - y_t\|_2^2$.*

Next, we derive the following results:

(1) the analytical expression of the optimal nonstationary reproduction distribution that achieves the infimum of the NRDF and the characterization of the NRDF subject to an MSE distortion;

(2) a universal lower bound on the total or per-letter MSE of any causal estimator of Gaussian processes;

(3) a realization of the optimal nonstationary reproduction distribution in the sense of Figure 3 that allows us to obtain the optimal filter.

The characterization of the NRDF is parametric and involves reverse-waterfilling recursively in time and space. Such a complete characterization is never reported in the literature, for either of the two average distortions (1.3) and (1.4). Further,

two algorithms are developed to provide tight upper bounds to the optimal reverse-waterfilling solution. These are generalizations of the standard reverse-waterfilling algorithm of the classical RDF of scalar autoregressive Gaussian processes with MSE (given in [3]) and of the NRDF for scalar-valued Gauss–Markov processes with MSE (given in [34]).

5.1. The optimal nonstationary reproduction distribution. Note that by Theorem 4.1 and the Markovian property of (5.1), the optimal nonstationary reproduction distribution given by (4.3)–(4.4) is Markov with respect to X_0^n. Moreover, from (IS1.2), the characterization of the NRDF is optimized over the reproduction distributions $Q_t^1(dy_t|y^{t-1}, x_t)$, $t \in \mathbb{N}_0^n$. The joint distribution of $\{X_0^n, Y_0^n\}$ for a fixed $Y^{-1} = y^{-1}$ is $\mathbf{P}^{Q^1}(dx_0^n, y_0^n|y^{-1}) = \mathbf{P}(dx_0|y^{-1}) \otimes_{t=0}^n (P_t(dx_t|x_{t-1}) \otimes Q_t^1(dy_t|y^{t-1}, x_t))$. Thus, for a fixed $Y^{-1} = y^{-1}$, the characterization $R_{0,n}^{\mathrm{na}}(D)$ given by (4.7) can be expressed in terms of relative entropy as in (3.5) and, additionally, the MSE distortion $\frac{1}{n+1}\mathbf{E}_{y^{-1}}^{Q^1}\{\sum_{t=0}^n \|X_t - Y_t\|_2^2\}$ is determined from the covariance matrix $\{X_0^n, Y_0^n\}$. On the other hand, for a fixed $Y^{-1} = y^{-1}$, once the covariance matrix of $\{X_0^n, Y_0^n\}$ is specified, and since X_0^n is Gaussian, by applying [32, Theorem 1.8.6] the payoff $\sum_{t=0}^n I(X_t; Y_t|Y_0^{t-1}, y^{-1})$ in (4.7) is minimized if the $\{X_0^n, Y_0^n\}$ is jointly Gaussian. Hence, for a fixed $Y^{-1} = y^{-1}$, the infimum in (4.7) over $\mathcal{Q}_{0,n}^1(D)$ is achieved if $\{X_0^n, Y_0^n\}$ is jointly Gaussian. Such jointly Gaussian distributions are induced if the reproduction distributions are restricted to conditionally Gaussian distributions, denoted by $Q_t^1(\cdot|y^{t-1}, x_t) = Q_t^G(\cdot|y^{t-1}, x_t)$, with conditional means that are linear in (x_t, y^{t-1}), and conditional covariances that are independent of (x_t, y^{t-1}) for $\forall t \in \mathbb{N}_0^n$.

Further, on an appropriate probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \in \mathbb{N}_0^n\}, \mathbb{P})$, we can construct a jointly Gaussian distribution $\mathbf{P}^G(dx_0^n, y^n)$, induced by the process X_0^n of Definition 5.1 and the process Y_0^n defined by the recursion

$$Y_t = H_t X_t + g_t\left(Y^{t-1}\right) + V_t^c, \quad Y^{-1} = y^{-1}, \qquad t \in \mathbb{N}_0^n, \tag{5.2}$$

$$g_t\left(Y^{t-1}\right) \triangleq M_t Y^{t-1}, \quad Q_t^G\left(\cdot\middle|y^{t-1}, x_t\right) \sim \mathcal{N}\left(H_t x_t + M_t y^{t-1}; K_{V_t^c}\right), \tag{5.3}$$

where (H_t, M_t) are nonrandom matrices, and $V_t^c \sim \mathcal{N}(0; K_{V_t^c}), K_{V_t^c} \in S_+^{p\times p}$ $\forall t \in \mathbb{N}_0^n$ is an independent sequence of Gaussian vectors that is independent of $\{W_t : t \in \mathbb{N}_0^n\}$ and X_0.

It should be noted that the following hold:

(i) the marginal $\mathbf{P}^G(dx_0^n) = \otimes_{t=0}^n P_t(dx_t|x_{t-1})$ is the distribution induced by the Gauss–Markov process X_0^n of Definition 5.1;

(ii) conditional independence holds, $\mathbf{P}^G(dx_t|x_0^{t-1}, y^{t-1}) = P_t(dx_t|x_{t-1})$, $t \in \mathbb{N}_0^n$;

(iii) $\mathbf{P}^G(dy_t|y^{t-1}, x^t) = Q_t^G(dy_t|y^{t-1}, x_t)$, $t \in \mathbb{N}_0^n$, is a conditionally Gaussian distribution.

Thus, $\{(H_t, M_t, K_{V_t^c}) : t \in \mathbb{N}_0^n\}$ is the parametrization of $\{Q_t^G(\cdot|y^{t-1}, x_t) : t \in \mathbb{N}_0^n\}$.

An alternative approach to show that $\{Q_t^G(dy_t|y^{t-1}, x_t) : t \in \mathbb{N}_0^n\}$ achieves the infimum in (4.7) over $\overrightarrow{\mathcal{Q}}_{0,n}^1(D)$, for fixed $Y^{-1} = y^{-1}$, is to verify that the jointly Gaussian distribution of (X_0^n, Y_0^n) for fixed $Y^{-1} = y^{-1}$, induced by (5.1) and (5.2), satisfies the implicit equations (4.3) and (4.4) of Theorem 4.1. Note that by the Markov property of X_0^n and the distortion function, (IS1.2) implies that the reproduction distribution is of the form $Q_t^1(\cdot|y^{t-1}, x_t)$ $\forall t \in \mathbb{N}_0^n$. Hence, it is sufficient to verify that (4.3) and (4.4) are satisfied for the jointly Gaussian process (X_0^n, Y_0^n) defined by (5.1) and (5.2).

Stage n. Since the exponential term $\rho_n(T^n x_0^n, T^n y^n) \triangleq \|x_n - y_n\|_2^2$ is quadratic in (x_n, y_n), and $\Pi_n^{\overrightarrow{Q}^G}(dx_n|y^{n-1}) = \int_{\mathcal{X}_n} Q_n^G(dx_n|y^{n-1}, x_n) \otimes \mathbf{P}^{\overrightarrow{Q}^G}(dx_n|y^{n-1})$, where

$\mathbf{P}^{\overrightarrow{Q}^G}(dx_n|y^{n-1})$ is conditionally Gaussian, with nonrandom covariance (by the Kalman filter equations), then, the right-hand side of (4.3) is of exponential quadratic form in (x_n, y^{n-1}), and hence the implicit equation (4.3) is satisfied.

Stages $t \in \{n-1, n-2, \ldots, 1, 0\}$. By (4.4), 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 y^{n-1}. Similarly to stage n, the right-hand side of (4.4) is of exponential quadratic form in (x_{n-1}, y^{n-2}). Hence, the implicit equation (4.4) is satisfied at time $t = n-1$. By induction, we deduce that (4.4) is satisfied for the jointly Gaussian process (X_0^n, Y_0^n) defined by (5.1) and (5.2).

By (5.1) and (5.2), for a fixed $Y^{-1} = y^{-1}$, the Gaussian NRDF is characterized by the following optimization problem:

$$R_{0,n}^{\mathrm{na}}(D) \triangleq \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^G(D)} \mathbf{E}_{y^{-1}}^{Q^G} \left\{ \sum_{t=0}^{n} \log \left(\frac{Q_t^G(\cdot|Y^{t-1}, X_t)}{\Pi_t^{Q^G}(\cdot|Y^{t-1})}(Y_t) \right) \right\} \tag{5.4}$$

$$= \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^G(D)} \sum_{t=0}^{n} I\left(X_t; Y_t \middle| Y_0^{t-1}, y^{-1}\right), \tag{5.5}$$

where

$$\overrightarrow{\mathcal{Q}}_{0,n}^G(D) \triangleq \left\{ Q_t^G\left(dy_t \middle| y^{t-1}, x_t\right),\ t \in \mathbb{N}_0^n :\ \frac{1}{n+1} \mathbf{E}_{y^{-1}}^{Q^G} \left\{ \sum_{t=0}^{n} \|X_t - Y_t\|_2^2 \right\} \le D \right\}.$$

Further, we can express $R_{0,n}^{\mathrm{na}}(D)$ in terms of the Kalman filter prediction and correction error as follows. For a fixed $Y^{-1} = y^{-1}$, define the conditional expectations

$$\begin{aligned}
\widehat{X}_{t|t-1} &\triangleq \mathbf{E}_{y^{-1}}^{Q^G}\left\{X_t | \sigma\left\{Y^{t-1}\right\}\right\},\ \Sigma_{t|t-1} \\
&\triangleq \mathbf{E}_{y^{-1}}^{Q^G}\left\{ (X_t - \widehat{X}_{t|t-1})(X_t - \widehat{X}_{t|t-1})^{\mathrm{T}} \middle| \sigma\left\{Y^{t-1}\right\} \right\}, \\
\widehat{X}_{t|t} &\triangleq \mathbf{E}_{y^{-1}}^{Q^G}\left\{X_t | \sigma\left\{Y^{t}\right\}\right\},\ \Sigma_{t|t} \\
&\triangleq \mathbf{E}_{y^{-1}}^{Q^G}\left\{ (X_t - \widehat{X}_{t|t})(X_t - \widehat{X}_{t|t})^{\mathrm{T}} \middle| \sigma\left\{Y^{t}\right\} \right\} \quad \forall t \in \mathbb{N}_0^n,
\end{aligned}$$

where $\sigma\{Z\}$ denotes the σ-algebra (observable events) generated by an RV Z. Since the joint process (X_0^n, Y_0^n) is jointly Gaussian generated by (5.1) and (5.2), it follows from the Kalman filter equations (see [26] with minor modifications) that the conditional covariances satisfy the recursions

$$\Sigma_{t|t-1} = A_{t-1}\Sigma_{t-1|t-1}A_{t-1}^{\mathrm{T}} + K_{W_{t-1}},\ \ \Sigma_{0|-1} \text{ is given},\ t \in \mathbb{N}_1^n, \tag{5.6}$$

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1}H_t^{\mathrm{T}}\left(H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + K_{V_t^c}\right)^{-1} H_t\Sigma_{t|t-1}. \tag{5.7}$$

Thus, the above conditional covariances are independent of the process Y^n and the function $g_t(y^{t-1}) = M_t y^{t-1},\ t \in \mathbb{N}_0^n$. That is,

$$\begin{aligned}
\Sigma_{t|t-1} &= \mathbf{E}_{y^{-1}}^{Q^G}\left\{ (X_t - \widehat{X}_{t|t-1})(X_t - \widehat{X}_{t|t-1})^{\mathrm{T}} \right\}, \\
\Sigma_{t|t} &\triangleq \mathbf{E}_{y^{-1}}^{Q^G}\left\{ (X_t - \widehat{X}_{t|t})(X_t - \widehat{X}_{t|t}) \right\} \quad \forall t \in \mathbb{N}_0^n.
\end{aligned}$$

Note that for a fixed $Y^{-1} = y^{-1}$, from a property of conditional mutual information,

the payoff in $R_{0,n}^{\mathrm{na}}(D)$ satisfies the identity

$$\sum_{t=0}^{n} I\left(X_t;Y_t\middle|Y_0^{t-1},y^{-1}\right)=\sum_{t=0}^{n}\mathbf{E}_{y^{-1}}^{Q^G}\left\{\log\left(\frac{\mathbf{P}_{X_t|Y^t}^{Q^G}(\cdot|y^t)}{\mathbf{P}_{X_t|Y^{t-1}}^{Q^G}(\cdot|y^{t-1})}(X_t)\right)\right\} \tag{5.8}$$

$$=\frac{1}{2}\sum_{t=0}^{n}\log\max\left\{1,\frac{|\Sigma_{t|t-1}|}{|\Sigma_{t|t}|}\right\}, \tag{5.9}$$

where (5.9) is calculated from $I(X_t;Y_t|Y_0^{t-1},y^{-1})=H(X_t|Y_0^{t-1},y^{-1})-H(X_t|Y_0^t,y^{-1})$, in which $H(\cdot|\cdot)$ denote conditional entropies, and using the fact that $\Sigma_{t|t-1},\ \Sigma_{t|t}$, are independent of $Y_0^t, g_t(y^{t-1}),\ t\in\mathbb{N}_0^n$. The MSE is given by

$$\begin{aligned}&\mathbf{E}_{y^{-1}}^{Q^G}\left\{\sum_{t=0}^{n}\|X_t-Y_t\|_2^2\right\}\\&\quad=\mathbf{E}_{y^{-1}}^{Q^G}\left\{\sum_{t=0}^{n}\left\|(I-H_t)X_t-g_t\left(Y^{t-1}\right)\right\|_2^2\right\}+\sum_{t=0}^{n}\mathrm{trace}(K_{V_t^c})\\&\quad\geq\mathbf{E}_{y^{-1}}^{Q^G}\left\{\sum_{t=0}^{n}\left\|(I-H_t)X_t-g_t^*\left(Y^{t-1}\right)\right\|_2^2\right\}\Bigg|_{g_t=g_t^*}+\sum_{t=0}^{n}\mathrm{trace}(K_{V_t^c})\\&\qquad\text{if}\ \ g_t\left(Y^{t-1}\right)=g_t^*\left(Y^{t-1}\right)=(I-H_t)\widehat{X}_{t|t-1}\ \ \forall t\in\mathbb{N}_0^n,\end{aligned} \tag{5.10}$$

where the inequality holds due to mean-square estimation theory. Since by (5.9) the payoff in $R_{0,n}^{\mathrm{na}}(D)$ does not depend on $g_t(\cdot),\ t\in\mathbb{N}_0^n$, then $g_t(Y^{t-1})=g_t^*(Y^{t-1}),\ t\in\mathbb{N}_0^n$, is optimal.

In view of the above discussion, we have the following preliminary characterization of the Gaussian NRDF.

LEMMA 5.2 (preliminary characterization of $R_{0,n}^{\mathrm{na}}(D)$). *Consider the Gauss–Markov process with* MSE *distortion given in Definition* 5.1. *Then, a preliminary characterization of* $R_{0,n}^{\mathrm{na}}(D)$, *for a fixed* $Y^{-1}=y^{-1}$, *is described by the optimization problem:*

$$\begin{aligned}R_{0,n}^{\mathrm{na}}(D)&=\inf_{\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D)}\sum_{t=0}^{n}I\left(X_t;Y_t\middle|Y_0^{t-1},y^{-1}\right)\\&=\inf_{\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D)}\frac{1}{2}\sum_{t=0}^{n}\log\max\left\{1,\frac{|\Sigma_{t|t-1}|}{|\Sigma_{t|t}|}\right\},\end{aligned} \tag{5.11}$$

where the average distortion constraint set is given by

$$\begin{aligned}\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D)\triangleq\Bigg\{&(H_t,K_{V_t^c}),\ t\in\mathbb{N}_0^n:\\&\frac{1}{n+1}\mathbf{E}_{y^{-1}}^{Q^G}\left\{\sum_{t=0}^{n}\|(I-H_t)(X_t-\widehat{X}_{t|t-1})\|_2^2\right\}\leq D\Bigg\}.\end{aligned}$$

$\Sigma_{t|t-1},\ \Sigma_{t|t}$ *satisfy recursions* (5.6), (5.7), *and the realization of the reproduction*

distribution is given by

$$Y_t = H_t\left(X_t - \widehat{X}_{t|t-1}\right) + \widehat{X}_{t|t-1} + V_t^c, \quad Y^{-1} = y^{-1}, \quad t = 0, \ldots, n, \tag{5.12}$$
$$Q_t^G\left(\cdot \middle| y^{t-1}, x_t\right) \sim \mathcal{N}(H_t(x_t - \widehat{x}_{t|t-1}) + \widehat{x}_{t|t-1}; K_{V_t^c}).$$

Moreover, the filter is given by the following recursions:

$$\text{Prediction: } \widehat{X}_{t|t-1} = A_{t-1}\widehat{X}_{t-1|t-1}, \quad \widehat{X}_{0|-1} = \textit{given}, \quad t = 1, \ldots, n, \tag{5.13a}$$

$$\text{Correction: } \widehat{X}_{t|t} = \widehat{X}_{t|t-1} + \Sigma_{t|t-1}H_t^T\left(H_t\Sigma_{t|t-1}H_t^T + K_{V_t^c}\right)^{-1}\tilde{K}_t, \tag{5.13b}$$

$$\text{Innovations: } \tilde{K}_t \triangleq Y_t - \widehat{X}_{t|t-1}, \quad t \in \mathbb{N}_0^n. \tag{5.13c}$$

Proof. The statements of the lemma follow directly from the discussion prior to the lemma and the Kalman filter equations [26] with minor modifications. ∎

Remark 4. It should be noted that the above characterization implies that $R_{0,n}^{\mathrm{na}}(D)$ is an optimization problem over the choices of $\{(H_t, K_{V_t^c}) : t \in \mathbb{N}_0^n\}$ which control $\{\Sigma_{t|t} : t \in \mathbb{N}_0^n\}$. That is, the optimization is over all choices of joint realizations of the process (X^n, Y^n) such that (i)–(iii) hold.

Next, we derive the main theorem to characterize $R_{0,n}^{\mathrm{na}}(D)$ parametrically, via a dynamic reverse-waterfilling optimization, and to determine the structure of the optimal matrices $\{(H_t, K_{V_t^c}) : t \in \mathbb{N}_0^n\}$, and hence of the specific realization of the process (X^n, Y^n) such that (i)–(iii) hold.

THEOREM 5.3 ($R_{0,n}^{\mathrm{na}}(D)$ of multidimensional Gauss–Markov process with MSE distortion). *Consider the Gauss–Markov process with* MSE *distortion given in Definition* 5.1. *Then, the following hold:*

(1) *The infimum over the reproduction distributions of the characterization of the Gaussian* NRDF (5.4) *occurs in the set of Markov distributions in* Y^{n-1}*, that is,*

$$Q_t^G\left(dy_t \middle| y^{t-1}, x_t\right) = Q_t^M(dy_t|y_{t-1}, x_t), \quad t \in \mathbb{N}_0^n, \tag{5.14}$$

and the characterization of the NRDF *for a fixed* $Y_{-1} = y_{-1}$ *is given by*

$$R_{0,n}^{\mathrm{na}}(D) \triangleq \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^M(D)} \mathbf{E}_{y_{-1}}^{Q^M}\left\{\sum_{t=0}^{n} \log\left(\frac{Q_t^M(\cdot|Y_{t-1}, X_t)}{\Pi_t^{Q^M}(\cdot|Y_{t-1})}(Y_t)\right)\right\} \tag{5.15}$$

$$= \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^M(D)} \left\{I(X_0; Y_0|Y_{-1} = y_{-1}) + \sum_{t=1}^{n} I(X_t; Y_t|Y_{t-1})\right\}, \tag{5.16}$$

where

$$\overrightarrow{\mathcal{Q}}_{0,n}^M(D) \triangleq \left\{Q_t^M(dy_t|y_{t-1}, x_t),\ t \in \mathbb{N}_0^n :\ \frac{1}{n+1}\mathbf{E}_{y_{-1}}^{Q^M}\left\{\sum_{t=0}^{n} \|X_t - Y_t\|_2^2\right\} \le D\right\}.$$

Moreover, the distribution $Q_t^G(dy_t|y^{t-1}, x_t) = Q_t^M(dy_t|y_{t-1}, x_t),\ t \in \mathbb{N}_0^n$*, is realized by*

$$Y_t = H_tX_t + (I - H_t)A_{t-1}Y_{t-1} + V_t^c,\ Y_{-1} = y_{-1}, \quad t \in \mathbb{N}_0^n, \tag{5.17}$$

$$H_t \triangleq I - \Delta_t\Lambda_t^{-1}, \quad K_{V_t^c} \triangleq \Delta_t H_t^T \succeq 0, \tag{5.18}$$

$$\Lambda_t = A_{t-1}\Delta_{t-1}A_{t-1}^T + K_{W_{t-1}},\ \Lambda_0 = \textit{given}, \tag{5.19}$$

$$\frac{1}{n+1}\sum_{t=0}^{n} \mathrm{trace}\,(\Delta_t) \le D, \quad \mathrm{trace}\,(\Delta_t) \triangleq \mathbf{E}_{y_{-1}}^{Q^M}\left\{\|X_t - Y_t\|_2^2\right\} \equiv D_t. \tag{5.20}$$

Furthermore, the above realization satisfies

$$\widehat{X}_{t|t-1} = A_{t-1}Y_{t-1}, \quad \widehat{X}_{t|t} = Y_t. \tag{5.21}$$

(2) *The characterization of the Gaussian* NRDF *in* (1) *is equivalent to the following optimization problem:*

$$R^{\mathrm{na}}_{0,n}(D) = \inf_{\Delta_t \in S^{p\times p}_+,\ t\in\mathbb{N}^n_0:\ \sum_{t=0}^n \mathrm{trace}(\Delta_t)\le D} \frac{1}{2}\sum_{t=0}^n \log\left\{\frac{|\Lambda_t|}{|\Delta_t|}\right\} \tag{5.22a}$$

subject to

$$0 \preceq \Delta_t \preceq \Lambda_t,\ t \in \mathbb{N}^n_0, \tag{5.22b}$$

$$\Lambda_t = A_{t-1}\Delta_{t-1}A^T_{t-1} + K_{W_{t-1}},\ \Lambda_0 = \textit{given}, \quad t \in \mathbb{N}^n_1. \tag{5.22c}$$

For the rest of the statements it is assumed that $K_{W_t} \in \mathcal{S}^{p\times p}_{++},\ t \in \mathbb{N}^n_0$.

The Lagrangian functional for the above optimization problem is

$$\begin{aligned}
&\mathcal{L}(\{\Delta_t,\Lambda_t\}^n_{t=0},\theta,\{F^1_t,F^2_t\}^n_{t=0})\\
&\quad= \left(\frac{1}{2}\log|\Lambda_0| - \mathrm{trace}\left(F^2_0\Lambda_0\right)\right)\\
&\qquad+\sum_{t=0}^{n-1}\left\{\frac{1}{2}\log|B_t\Delta_t+I| - \frac{1}{2}\log|\Delta_t| + \mathrm{trace}\left(\left[F^2_t - F^1_t\right]\Delta_t\right)\right.\\
&\qquad\qquad\left.- \mathrm{trace}\left(F^2_{t+1}\left(A_t\Delta_tA^T_t + K_{W_t}\right)\right) + \theta\,\mathrm{trace}(\Delta_t)\right\}\\
&\qquad- \frac{1}{2}\log|\Delta_n| + \theta\,\mathrm{trace}(\Delta_n) + \mathrm{trace}\left(\left[F^2_n - F^1_n\right]\Delta_n\right),\quad t\in\mathbb{N}^n_0,
\end{aligned} \tag{5.23}$$

where $B_t \triangleq A^T_tK^{-1}_{W_t}A_t$, $\theta \in [0,\infty)$ *is a Lagrange multiplier of the* MSE *constraint, and* $F^j_t \in \mathcal{S}^{n\times n}_+,\ j=1,2$, *are the matrix Lagrange multipliers for* $0 \preceq \Delta_t \preceq \Lambda_t,\ t \in \mathbb{N}^n_0$. *The necessary and sufficient conditions for* $\Delta^*_t \in \mathcal{S}^{p\times p}_+,\ t\in\mathbb{N}^n_0$, *to achieve the minimum are given by the following equations:*

$$\left.\frac{\partial\mathcal{L}(\{\Delta_t,\Lambda_t\}^n_{t=0},\theta,\{F^1_t,F^2_t\}^n_{t=0})}{\partial\Delta_t}\right|_{\Delta=\Delta^*,\Lambda=\Lambda^*} = 0, \quad t\in\mathbb{N}^n_0. \tag{5.24}$$

For $t=n$,

$$\Delta^{*,-1}_n = 2\Big(\theta I - F^1_n + F^2_n\Big). \tag{5.25}$$

If $\Delta^*_n \succ 0$, *then*

$$\Delta^{*,-1}_n = 2\left(\theta I + F^2_n\right) \implies \Delta^*_n = \frac{1}{2}\left(\theta I + F^2_n\right)^{-1}. \tag{5.26}$$

For $t = n-1,\dots,0$,

$$\frac{1}{2}(I+B_t\Delta^*_t)^{-1}B_t - \frac{1}{2}\Delta^{*,-1}_t - F^1_t + F^2_t + \theta I - A^T_tF^2_{t+1}A_t = 0.$$

If $\Delta^*_t \succ 0$, *then*

$$(I+B_t\Delta^*_t)^{-1}B_t - \Delta^{*,-1}_t + 2\left(\theta I + F^2_t\right) - 2A^T_tF^2_{t+1}A_t = 0, \tag{5.27}$$

or equivalently

$$\left(-\frac{I}{2}\right)\Delta_t^* + \Delta_t^*\left(-\frac{I}{2}\right) - \Delta_t^* B_t \Delta_t^* + \Upsilon_t^{-1} = 0, \quad t \in \mathbb{N}_0^{n-1}, \tag{5.28}$$

$$\Upsilon_t \triangleq 2\left(\theta I + F_t^2 - A_t^T F_{t+1}^2 A_t\right). \tag{5.29}$$

The complementary slackness conditions are

$$\theta\left(\sum_{t=0}^{n} \operatorname{trace}(\Delta_t^*) - D(n+1)\right) = 0, \theta \geq 0, \tag{5.30a}$$

$$\sum_{t=0}^{n} \operatorname{trace}(\Delta_t^*) \leq D(n+1), \tag{5.30b}$$

$$\Delta_t^* \succeq 0, \Delta_t^* - \Lambda_t^* \preceq 0, \operatorname{trace}\left(F_t^1 \Delta_t^*\right) = 0, F_t^2\left(\Delta_t^* - \Lambda_t^*\right) = 0, F_t^1 \succeq 0, F_t^2 \succeq 0, \tag{5.30c}$$

where $t \in \mathbb{N}_0^n$ *and* Λ_t^* *is given by* (5.22c) *for* Δ_t^*.

Proof. See Appendix A. ∎

Remark 5. Note that (5.28) is of the form of a Riccati equation, with a terminal condition given by (5.26). Thus, it is possible to apply properties of Riccati equations to analyze the solutions of such an equation. ∎

In the next remark, we apply Theorem 5.3 to independent, time-varying vector-valued Gaussian sources, with correlated spatial components.

Remark 6 (application of Theorem 5.3). Consider a source described by an independent, vector-valued zero mean Gaussian process X_t, i.e., $\mathcal{N}(0; K_{X_t}), K_{X_t} \in S_+^{p\times p}$, $t \in \mathbb{N}_0^n$, with correlated spatial components. This is a degenerate version of (5.1) if we set $A_t = 0$. By (5.17), the optimal realization that corresponds to this source process degenerates to $Y_t = H_t X_t + V_t^c,\ t \in \mathbb{N}_0^n$, where H_t is given by (5.18), with $\Lambda_t = K_{X_t}$. The optimization problem in (5.22) degenerates to

$$R_{0,n}^{\mathrm{na}}(D) = \inf_{\Delta_t \in S_+^{p\times p},\ t\in\mathbb{N}_0^n:\ \sum_{t=0}^{n} \operatorname{trace}(\Delta_t) \leq D} \frac{1}{2}\sum_{t=0}^{n} \log\left\{\frac{|K_{X_t}|}{|\Delta_t|}\right\}, \tag{5.31a}$$

subject to

$$0 \preceq \Delta_t \preceq K_{X_t},\ t \in \mathbb{N}_0^n. \tag{5.31b}$$

By the KKT conditions of Theorem 5.3, then

$$\Delta_t^{*,-1} = 2\left(\theta I + F_t^2\right) \implies \Delta_t^* = \frac{1}{2}\left(\theta I + F_t^2\right)^{-1}, \tag{5.32}$$

where $F_t^1 = 0$ because if Δ_t^* is singular, then it gives infinite value of mutual information. By (5.32), we deduce that Δ_t^* and F_t^2 have spectral representations with the same unitary matrix, hence $\Delta_t^* F_t^2 = F_t^2 \Delta_t^*,\ t \in \mathbb{N}_0^n$, i.e., they commute. Let $F_t^2 \triangleq U \Lambda_{F_t^2} U^T$, where $\Lambda_{F_t^2}$ is a diagonal matrix with entries the eigenvalues of matrix F_t^2. Then, Δ_t^* can be written as $\Delta_t^* = U \Lambda_{\Delta_t^*} U^T$. Note that unitary matrix U is a design parameter. Complementary slackness condition (5.30c), i.e., $F_t^2\left(\Delta_t^* - K_{X_t}\right) = 0$, can be written as

$$U \Lambda_{F_t^2} U^T\left(U \Lambda_{\Delta_t^*} U^T - K_{X_t}\right) = 0 \Leftrightarrow \Lambda_{F_t^2} \Lambda_{\Delta_t^*} = \Lambda_{F_t^2} U^T K_{X_t} U. \tag{5.33}$$

As it can be seen in (5.33), since U is a design parameter, one can choose unitary matrix U such that $U^T K_{X_t} U$ is diagonal (i.e., if $K_{X_t} = V \Lambda_{K_{X_t}} V^T$, then $U = V^T$). As a result, Δ_t^*, K_{X_t} and F_t^2 have spectral representations with the same unitary matrix and commute. Next, we analyze the feasible set of solutions that correspond to the optimization problem (5.31) if $K_{X_t} \in S_{++}^{p \times p}$.

(i) $0 \prec K_{X_t} \preceq \Delta_t$ ($0 \prec K_{X_t} \prec \Delta_t$ included). In this case, the rate is zero.

(ii) $0 \prec \Delta_t \prec K_{X_t}$. Since F_t^2, Δ_t^*, and K_{X_t} commute, then $\mu_{K_{X_t,i}} - \mu_{\Delta_t^*,i} = \mu_{K_{X_t}-\Delta_t^*,i} > 0$, where $\mu_{K_{X_t}-\Delta_t^*,i}$ is the ith-eigenvalue of $K_{X_t} - \Delta_t^*$. As a result, $\mu_{F_t^2,i} = 0$ and from (5.32) we deduce that $\Delta_t^* = \frac{1}{2\theta} I$.

(iii) $0 \prec \Delta_t \preceq K_{X_t}$ ($0 \prec K_{X_t} \prec \Delta_t$ excluded). Since F_t^2, Δ_t^*, and K_{X_t} commute, $\mu_{K_{X_t,i}} - \mu_{\Delta_t^*,i} = \mu_{K_{X_t}-\Delta_t^*,i} \geq 0$. Hence, if $\mu_{K_{X_t}-\Delta_t^*,i} = 0$ (which also implies that $\mu_{F_t^2,i} > 0$), $\mu_{K_{X_t,i}} = \mu_{\Delta_t^*,i}$. Otherwise, if $\mu_{K_{X_t}-\Delta_t^*,i} > 0$ (i.e., $\mu_{F_t^2,i} = 0$), then $\mu_{K_{X_t,i}} > \mu_{\Delta_t^*,i}$.

(iv) $0 \prec \Delta_t \not\preceq K_{X_t}$ and $0 \prec K_{X_t} \not\preceq \Delta_t$. In these cases, one can show that there exist Δ_t with a lower rate that lies within the other cases.

Using the previous analysis on the complementary slackness conditions, and (5.32), we observe that $\mu_{\Delta_t^*,i} = \min_i\{\frac{1}{2\theta}, \mu_{K_{X_t,i}}\}$ for each t and $\forall i$. Hence, the solution of the optimization problem of (5.31) is precisely the solution of the classical reverse-waterfilling algorithm, i.e.,

$$R_{0,n}^{\mathrm{na}}(D) = \sum_{t=0}^{n} R_t^{na}(\Delta_t^*) = \frac{1}{2} \sum_{t=0}^{n} \sum_{i=1}^{p} \log\left(\frac{\mu_{K_{X_t},i}}{\mu_{\Delta_t^*,i}}\right), \tag{5.34}$$

where

$$R_t^{na}(\Delta_t^*) = \frac{1}{2} \log\left(\frac{|K_{X_t}|}{|\Delta_t^*|}\right), \tag{5.35}$$

$$\mu_{\Delta_t^*,i} = \begin{cases} \frac{1}{2\theta} & \text{if } \frac{1}{2\theta} < \mu_{K_{X_t},i}, \\ \mu_{K_{X_t},i} & \text{if } \frac{1}{2\theta} \geq \mu_{K_{X_t},i}, \end{cases} \text{ at each } t \text{ and all } i, \tag{5.36}$$

and θ is chosen such that the distortion constraint is satisfied.

In the following example, we consider the optimization problem of (5.31a) in Remark 6 to derive a closed form solution for $R_{0,n}^{na}$ when $n = 1$. To do so, we first apply the KKT conditions at the last time-step of the optimization problem (5.31a) and then we move sequentially backward in time.

Example 1 (closed form solution of a memoryless $\mathbb{R}^2$-valued time-varying Gaussian process). We consider a memoryless $\mathbb{R}^2$-valued time-varying Gaussian correlated process $\{X_t : t = 0, 1\}$ with covariance matrix $K_{X_t} \in \mathcal{S}_{++}^{2\times 2}$, $t = 0, 1$, given by

$$K_{X_t} = \begin{bmatrix} \sigma_{X_{t,1}}^2 & \sigma_{X_{t,12}}^2 \\ \sigma_{X_{t,12}}^2 & \sigma_{X_{t,2}}^2 \end{bmatrix}. \tag{5.37}$$

Full-rank solution. We consider the case for which $\Delta_t^* \prec K_{X_t}$, $t = 0, 1$. From the complementary slackness conditions, this means that $\mu_{\Delta_t,i} - \mu_{K_{X_t},i} < 0$[6] and $\mu_{F_t^2,i} = 0$, for $i = 1, 2$, $t = 0, 1$. Upon solving (5.32) we obtain

$$\Delta_t^* = U_t \begin{bmatrix} \frac{1}{2\theta} & 0 \\ 0 & \frac{1}{2\theta} \end{bmatrix} U_t^{\mathrm{T}}, \tag{5.38}$$

[6]Where $\mu_{\bullet,i}$ denotes the ith-eigenvalue of $\bullet$ matrix.

where $U_t \in \mathbb{R}^{2\times 2}$ is the unitary matrix that diagonalizes Δ_t^*. This means that $\text{trace}(\Delta_t^*) = \frac{1}{\theta}$, and using the distortion condition we obtain that $\theta = 1/D$. Then, the objective function, given in (5.31a), becomes

$$R_{0,1}^{\text{na}}(D) = \frac{1}{2}\sum_{t=0}^{1}\sum_{i=1}^{2}\log\left(\frac{2\mu_{K_{X_t},i}}{D}\right) \tag{5.39}$$

with $D \in (0, 2\min\{\mu_{K_{X_t,i}}\})$, $i = 1, 2$, $t = 0, 1$.

Rank-deficient solution. In this example, we will consider one of the many possible cases (other cases can be solved likewise):

at $t = 1$: $\mu_{\Delta_1^*,1} - \mu_{K_{X_1},1} < 0$, which in turn means $\mu_{F_1^2,1} = 0$ and $\mu_{\Delta_1^*,2} - \mu_{K_{X_1},2} = 0$ which means that $\mu_{F_1^2,2} > 0$;

at $t = 0$: $\mu_{\Delta_0^*,1} - \mu_{K_{X_0},1} = 0$, which in turn means $\mu_{F_0^2,1} > 0$ and $\mu_{\Delta_0^*,2} - \mu_{K_{X_0},2} = 0$ which means that $\mu_{F_0^2,2} > 0$.

at $t = 1$: for this case, we obtain from (5.32) that

$$\Delta_1^* = U_1\begin{bmatrix}\frac{1}{2\theta} & 0\\ 0 & \frac{1}{2(\theta+\mu_{F_1^2,2})}\end{bmatrix}U_1^{\text{T}}, \tag{5.40}$$

where $U_1 \in \mathbb{R}^{2\times 2}$ is the unitary matrix that diagonalizes Δ_1^* and the eigenvalues of Δ_1^* are given in a decreasing order, i.e., $\mu_{\Delta_1^*,1} \geq \mu_{\Delta_1^*,2}$, and then, using the objective function of (5.31a) evaluated at $t = 1$, we obtain

$$R_1^{\text{na}}(\Delta_1^*) \overset{\text{(b)}}{=} \frac{1}{2}\log\left(\frac{\mu_{K_{X_1},1}}{\frac{1}{2\theta}}\right) + \underbrace{\frac{1}{2}\log\left(\frac{\mu_{K_{X_1},2}}{\frac{1}{2(\theta+\mu_{F_1^2,2})}}\right)}_{=0} \overset{\text{(c)}}{=} \frac{1}{2}\log\left(\frac{\mu_{K_{X_1},1}}{\frac{1}{2\theta}}\right), \tag{5.41}$$

where (b) follows from (5.40) and (c) from the fact that $\mu_{\Delta_1^*,2} = \mu_{K_{X_1},2}$. Therefore, at $t = 1$, $\text{trace}(\Delta_1^*) = \frac{1}{2\theta} + \mu_{K_{X_1},2}$.

at $t = 0$: for this case, we obtain from (5.32) that

$$\Delta_0^* = U_0'\begin{bmatrix}\frac{1}{2(\theta+\mu_{F_0^2,1})} & 0\\ 0 & \frac{1}{2(\theta+\mu_{F_0^2,2})}\end{bmatrix}U_0'^{\text{T}}, \tag{5.42}$$

where $U_0' \in \mathbb{R}^{2\times 2}$ is the unitary matrix that diagonalizes Δ_0^*. Since $\mu_{\Delta_0^*,i} = \mu_{K_{X_0},i}$, $i = 1, 2$, $R_0^{\text{na}}(\Delta_0^*) = 0$ and $\text{trace}(\Delta_0^*) = \mu_{K_{X_0},1} + \mu_{K_{X_0},2}$.

Overall solution: invoking the distortion constraint, we get

$$\frac{1}{2\theta} = 2D - (\mu_{K_{X_0},1} + \mu_{K_{X_0},2} + \mu_{K_{X_1},2}), \tag{5.43}$$

and the objective function, given in (5.31a), becomes

$$R_{0,1}^{\text{na}}(D) = \frac{1}{2}\log\left(\frac{\mu_{K_{X_1},1}}{2D - (\mu_{K_{X_0},1} + \mu_{K_{X_0},2} + \mu_{K_{X_1},2})}\right), \tag{5.44}$$

with $2D \in (\mu_{K_{X_0},1}+\mu_{K_{X_0},2}+\mu_{K_{X_1},2}, \sum_{t=0}^{1}\sum_{i=1}^{2}\mu_{K_{X_t},i}\})$, $i = 1, 2$, $t = 0, 1$.

Next, we approximate the characterization of the NRDF in (5.22) in terms of a dynamic time-space reverse-waterfilling.

PROPOSITION 5.4 (upper bound to (5.22)). *An upper bound of the solution to the characterization* (5.22) *is given by*

$$\Delta_t \triangleq \min\left\{\Delta_t^\dagger, \Lambda_t\right\}, \tag{5.45}$$

where $\min\{\Delta_t^\dagger, \Lambda_t\} \triangleq S_t \operatorname{diag}(\min\{\mu_{\Delta_t^\dagger,i}, \mu_{\Lambda_t,i}\})S_t^{-1}$, $S_t \in \mathbb{R}^{p\times p}$ *is a nonsingular matrix,* $\mu_{\bullet,i}$ *is the ith eigenvalue of matrix* $\bullet$, $\Delta_t^\dagger \in \mathcal{S}_{++}^{p\times p}, t \in \mathbb{N}_0^n$ *is the unique solution of*

$$\left(-\frac{I}{2}\right)\Delta_t^\dagger + \Delta_t^\dagger\left(-\frac{I}{2}\right) - \Delta_t^\dagger B_t \Delta_t^\dagger + \frac{1}{2\theta}I = 0 \ \forall t \in \mathbb{N}_0^{n-1}, \tag{5.46a}$$

$$\Delta_n^\dagger = \frac{1}{2\theta}I, \tag{5.46b}$$

Λ_t *is given by* (5.22c) *using* Δ_{t-1}, *with* $\theta \in (0,\infty)$ *chosen to satisfy*

$$\frac{1}{n+1}\sum_{t=0}^{n}\operatorname{trace}(\Delta_t) = D. \tag{5.47}$$

This upper bound gives the optimal solution for the case which $\Delta_t \prec \Lambda_t \ \forall t \in \mathbb{N}_0^n$.

Proof. See Appendix B. ∎

Based on the solution provided in Proposition 5.4, we propose Algorithm 1 for solving the problem numerically to a good approximation.

Next, we consider the case in which there is an MSE distortion constraint at each time. We refer to it as *pointwise* MSE *distortion*. It can be shown that the problem can be treated as a special case of Lemma 5.2 and Theorem 5.3 (where the constraint is for the total MSE distortion).

COROLLARY 5.5 ($R_{0,n}^{\mathrm{na}}(D)$ of multidimensional Gauss–Markov process with pointwise MSE distortion). *Consider the Gauss–Markov process in Definition* 5.1, *with the total distortion constraint* $\frac{1}{n+1}\mathbf{E}_\mu^Q\{\sum_{t=0}^n \|X_t - Y_t\|_2^2\} \le D$ *replaced by a pointwise* MSE *distortion constraint defined by*

$$\mathbf{E}_\mu^Q\left\{\|X_t - Y_t\|_2^2\right\} \le D_t, \quad D_t \in [0,\infty), \quad t \in \mathbb{N}_0^n.$$

Then, the following hold:

(1) *All statements of Lemma* 5.2 *hold with the characterization of the Gaussian* NRDF *denoted by* $R_{0,n}^{\mathrm{na}}(D_0,\dots,D_n)$ *and the corresponding optimization problem to be*

$$\begin{aligned} R_{0,n}^{\mathrm{na}}(D_0,\dots,D_n) &= \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D_0,\dots,D_n)} \sum_{t=0}^{n} I\left(X_t;Y_t\middle|Y^{t-1}\right) \\ &= \inf_{\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D_0,\dots,D_n)} \frac{1}{2}\sum_{t=0}^{n}\log\max\left\{1,\frac{|\Lambda_t|}{|\Delta_t|}\right\}, \end{aligned} \tag{5.49}$$

Algorithm 1 Dynamic reverse-waterfilling algorithm of Proposition 5.4.

Initialize: number of time-steps n; distortion level D ($D \le D_{0,n}^{\max}$); error tolerance ϵ; nominal minimum value $\theta^{\min} \approx 0$; initial variance $\Lambda_0 = \Sigma_{X_0}$ of the initial state X_0, values of A_t and K_{W_t} of (5.1).
Set $\theta = 1/2D$; flag $= 0$.
while flag $= 0$ **do**
  Compute $\Delta_t \ \forall\ t$ as follows:
  for $t = 0 : n$ **do**
    Compute $\Delta_t^\dagger$ according to (5.46a), (5.46b).
    Compute Δ_t according to (5.45).
    if $t < n$ **then**
      Compute Λ_{t+1} according to (5.22c).
    end if
  end for
  if $|\frac{1}{n+1}\sum_{t=0}^{n} \operatorname{trace}(\Delta_t) - D| \le \epsilon$ **then**
    flag $\leftarrow 1$
  else
    Readjust θ as follows:

$$\theta \leftarrow \max\left\{\theta^{\min}, \theta - \gamma\left(D - \frac{1}{n+1}\sum_{t=0}^{n}\operatorname{trace}(\Delta_t)\right)\right\}, \tag{5.48}$$

    where $\gamma \in (0,1]$ is a proportionality gain; its choice affects the rate of convergence.
  end if
end while
Output: Δ_t, Λ_t, for $D\ \forall t \in \mathbb{N}_0^n$.

where

$$\begin{aligned}\overrightarrow{\mathcal{Q}}_{0,n}^{H_t,K_{V_t^c}}(D_0,\dots,D_n) &\triangleq \Big\{(H_t, K_{V_t^c}),\ t \in \mathbb{N}_0^n : \\ &\quad \mathbf{E}_\mu^{Q^G}\left\{\|(I-H_t)(X_t - \widehat{X}_{t|t-1})\|_2^2\right\} = \operatorname{trace}(\Delta_t) \le D_t,\ t \in \mathbb{N}_0^n\Big\},\end{aligned} \tag{5.50}$$

and Λ_t, Δ_t, *satisfy recursions* (5.6), (5.7), (5.12)–(5.13b).

(2) *Theorem* 5.3(1) *holds with* $\overrightarrow{\mathcal{Q}}_{0,n}^{M}(D)$ *replaced by* $\overrightarrow{\mathcal{Q}}_{0,n}^{M}(D_0,\dots,D_n)$, *defined by*

$$\overrightarrow{\mathcal{Q}}_{0,n}^{M}(D_0,\dots,D_n) \triangleq \Big\{Q_t^M(dy_t|y_{t-1},x_t),\ t \in \mathbb{N}_0^n :\ \mathbf{E}_\mu^{Q^M}\left\{\|X_t - Y_t\|_2^2\right\} \le D_t,\ t \in \mathbb{N}_0^n\Big\}. \tag{5.51}$$

(3) *Theorem* 5.3(2) *holds with* $R_{0,n}^{\mathrm{na}}(D)$ *replaced by*

$$R_{0,n}^{\mathrm{na}}(D_0,\dots,D_n) = \inf_{\Delta_t \in S_+^{p\times p},\ \operatorname{trace}(\Delta_t)\le D_t,\ t=0,\dots,n} \frac{1}{2}\sum_{t=0}^{n} \log \max\left\{1, \frac{|\Lambda_t|}{|\Delta_t|}\right\} \tag{5.52}$$

for some $D_t \in [0,\infty),\ t \in \mathbb{N}_0^n$.

For the rest of the statements it is assumed that $K_{W_t} \in \mathcal{S}_{++}^{p\times p},\ t \in \mathbb{N}_0^n$.

Then the analogue of Theorem 5.3(2) *holds, as follows.*

The Lagrangian functional for (5.52) *is*

(5.53)
$$\begin{aligned}
&\mathcal{L}^{LL}(\{\Delta_t,\Lambda_t\}_{t=0}^n,\ \{\theta_t\}_{t=0}^n,\{F_t^1,F_t^2\}_{t=0}^n)\\
&\quad=\left(\frac{1}{2}\log|\Lambda_0|-\operatorname{trace}\left(F_0^2\Lambda_0\right)\right)-\sum_{t=0}^{n}\theta_t D_t\\
&\qquad+\sum_{t=0}^{n-1}\left\{\frac{1}{2}\log|K_{W_t}|+\frac{1}{2}\log|B_t\Delta_t+I|-\frac{1}{2}\log|\Delta_t|+\operatorname{trace}(\left[F_t^2-F_t^1\right]\Delta_t)\right.\\
&\qquad\qquad\left.-\operatorname{trace}\left(F_{t+1}^2\left(A_t\Delta_tA_t^T+K_{W_t}\right)\right)+\theta_t\operatorname{trace}(\Delta_t)\right\}\\
&\qquad-\frac{1}{2}\log|\Delta_n|+\theta_n\operatorname{trace}(\Delta_n)+\operatorname{trace}\left(\left[F_n^2-F_n^1\right]\Delta_n\right),\quad B_t\triangleq A_t^TK_{W_t}^{-1}A_t,
\end{aligned}$$

where $\theta_t\in[0,\infty)$, $t\in\mathbb{N}_0^n$, *are the Lagrange multipliers for the* MSE *constraint, and* $F_t^j\in\mathcal{S}_+^{p\times p},j=1,2$, *are the matrix Lagrange multipliers for* $0\preceq\Delta_t\preceq\Lambda_t$, $t\in\mathbb{N}_0^n$. *The necessary and sufficient conditions for* $\Delta_t^*\in\mathcal{S}_+^{p\times p}$, $t\in\mathbb{N}_0^n$, *to achieve the minimum are given by the following equations.*

For $t=n$,

$$\Delta_n^{*,-1}=2\left(\theta_nI-F_n^1+F_n^2\right).$$

If $\Delta_n^*\succ 0$, *then*

$$\Delta_n^{*,-1}=2\left(\theta_nI+F_n^2\right). \tag{5.54}$$

For $t\in\mathbb{N}_0^{n-1}$,

$$\frac{1}{2}(I+B_t\Delta_t^*)^{-1}B_t-\frac{1}{2}\Delta_t^{*,-1}-F_t^1+F_t^2+\theta_tI-A_t^TF_{t+1}^2A_t=0.$$

If $\Delta_t^*\succ 0$, *then*

$$(I+B_t\Delta_t^*)^{-1}B_t-\Delta_t^{*,-1}+2\left(\theta_tI+F_t^2\right)-2A_t^TF_{t+1}^2A_t=0. \tag{5.55}$$

Equations (5.54) *and* (5.55) *are precisely as the ones in* (5.26) *and* (5.27) *with* θ *replaced by* θ_t, $t\in\mathbb{N}_0^n$. *In addition, the complementary slackness conditions are*

$$\theta_t\left(\operatorname{trace}(\Delta_t^*)-D_t\right)=0,\operatorname{trace}(\Delta_t^*)\le D_t,\theta_t\ge 0,\quad t\in\mathbb{N}_0^n, \tag{5.56a}$$

$$\Delta_t^*\succeq 0,\Delta_t^*-\Lambda_t^*\preceq 0,F_t^1\succeq 0,F_t^2\succeq 0,\operatorname{trace}(F_t^1\Delta_t^*)=0,F_t^2\left(\Delta_t^*-\Lambda_t^*\right)=0. \tag{5.56b}$$

Proof. This follows directly from Lemma 5.2 and Theorem 5.3. ∎

Similar to Proposition 5.4, in what follows we approximate the characterization of the NRDF obtained in (5.52) in terms of a dynamic time-space reverse-waterfilling.

PROPOSITION 5.6 (upper bound on the characterization of (5.52)). *The solution to the characterization* (5.52) *is given by* Δ_t *of* (5.45), *where* $\Delta_t^\dagger\in\mathcal{S}_{++}^{p\times p},t\in\mathbb{N}_0^n$, *is the unique solution of*

$$\left(-\frac{I}{2}\right)\Delta_t^\dagger+\Delta_t^\dagger\left(-\frac{I}{2}\right)-\Delta_t^\dagger B_t\Delta_t^\dagger+\frac{1}{2\theta_t}I=0\ \forall t\in\mathbb{N}_0^{n-1}, \tag{5.57a}$$

$$\Delta_n^\dagger=\frac{1}{2\theta_n}I, \tag{5.57b}$$

Λ_t is given by (5.22c) *using* Δ_{t-1}, *with* $\theta \in (0,\infty)$ *chosen to satisfy*

$$\operatorname{trace}(\Delta_t) = D_t \text{ for each } t. \tag{5.58}$$

This upper bound gives the optimal solution for the case which $\Delta_t \prec \Lambda_t,\ \forall t \in \mathbb{N}_0^n$.

Proof. The proof is similar to the one of Proposition 5.4, hence we omit it. ∎

Based on the solution provided in Proposition 5.6, we propose Algorithm 2 for solving the problem numerically to a good approximation.

Algorithm 2 Dynamic reverse-waterfilling algorithm of Proposition 5.6.

Initialize: number of time-steps n; distortion levels $D = (D_0, \dots, D_t)$ ($D_t \leq D_t^{\max}\ \forall t \in \mathbb{N}_0^n$); error tolerance ϵ; nominal minimum value $\theta_t^{\min} \approx 0\ \forall t$; initial variance $\Lambda_0 = \Sigma_{X_0}$ of the initial state X_0, values of A_t and K_{W_t} of (5.1).
Set $\theta_t = 1/2D_t\ \forall t \in \mathbb{N}_0^n$;
for $t = 0$:length(D) **do**
  flag $= 0$.
  while flag $= 0$ **do**
    Compute $\Delta_t\ \forall t$ as follows:
    if $t < n$ **then**
      Compute $\Delta_t^\dagger$ according to (5.57a) and Δ_t according to (5.45).
      Compute Λ_{t+1} according to (5.22c).
    else
      Compute $\Delta_n^\dagger$ according to (5.57b) and Δ_n according to (5.45).
    end if
    if $|\operatorname{trace}(\Delta_t) - D_t| \leq \epsilon$ **then**
      flag $\leftarrow 1$
    else
      Readjust θ_t as follows:

$$\theta_t \leftarrow \max\left\{\theta_t^{\min}, \theta_t - \gamma_t(D_t - \operatorname{trace}(\Delta_t))\right\}, \tag{5.59}$$

      where $\gamma_t \in (0,1]\ \forall t$ is a proportionality gain; its choice affects the rate of convergence at each t.
    end if
  end while
end for
Output: Δ_t, Λ_t for each D_t, $t \in \mathbb{N}_0^n$.

The following remark is a direct consequence of Theorem 5.3 and illustrates the connection between $R_{0,n}^{\mathrm{na}}(D)$ and $D_{0,n}(R^{\mathrm{na}})$ given by (4.1).

Remark 7. From Theorem 5.3, the NRDF of the Gaussian process (5.1) with total MSE distortion is given by

$$R_{0,n}^{\mathrm{na}}(D) = \frac{1}{2}\sum_{t=0}^{n}\sum_{i=1}^{p}\log\left\{\max\left(1, \frac{\mu_{\Lambda_t,i}}{\mu_{\Delta_t,i}}\right)\right\} \overset{\text{(a)}}{=} \sum_{t=0}^{n}\sum_{i=1}^{p} R_{t,i}^{\mathrm{na}}(\mu_{\Delta_t,i}), \tag{5.60}$$

where (a) follows if we let

$$R_{t,i}^{\mathrm{na}}(\mu_{\Delta_t,i}) \triangleq \frac{1}{2}\log\left\{\max\left(1, \frac{\mu_{\Lambda_t,i}}{\mu_{\Delta_t,i}}\right)\right\},\ t \in \mathbb{N}_0^n,\ i = 1, \dots, p. \tag{5.61}$$

By (5.61) we obtain

$$\mu_{\Delta_t,i} = \mu_{\Lambda_t,i} e^{-2R^{\mathrm{na}}_{t,i}},\ t \in \mathbb{N}_0^n,\ i = 1, \ldots, p. \tag{5.62}$$

Utilizing (5.22a), we have

$$D = \frac{1}{n+1} \sum_{t=0}^{n} \mu_{\Delta_t} = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \mu_{\Delta_t,i},\ \mu_{\Delta_t} \triangleq \sum_{i=1}^{p} \mu_{\Delta_t,i}. \tag{5.63}$$

Substituting (5.62) into (5.63) we obtain

$$D_{0,n}(R^{\mathrm{na}}) = \frac{1}{n+1} \sum_{t=0}^{n} \mu_{\Delta_t} = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \mu_{\Lambda_t,i} e^{-2R^{\mathrm{na}}_{t,i}}. \tag{5.64}$$

A similar result to Remark 7 holds when we consider the pointwise MSE distortion. This is obvious, hence we omit it.

5.2. Universal lower bound on MSE. Next, we utilize the parametric expressions of the full rank solution of the Gaussian NRDF given in Theorem 5.3(2) to derive a lower bound on the total and pointwise MSE given in terms of conditional mutual information $I(X^n; Y_0^n | Y^{-1})$.

THEOREM 5.7 (universal lower bound on total MSE). *Let $\{X_t :\ t \in \mathbb{N}_0^n\}$ be the multidimensional Gauss–Markov process given by* (5.1) *and let $\{\widetilde{Y}_t :\ t \in \mathbb{N}_0^n\}$ be any estimator (not necessarily Gaussian) of $\{X_t :\ t \in \mathbb{N}_0^n\}$. The total* MSE *is bounded below by*

$$\frac{1}{n+1} \sum_{t=0}^{n} \mathbf{E}\left\{\|X_t - \widetilde{Y}_t\|_2^2\right\} \geq \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \mu_{\Lambda_t,i} e^{-2I(X_{t,i}; \widetilde{Y}_{t,i} | \widetilde{Y}_{t-1,i})}. \tag{5.65}$$

Proof. Let $D = \frac{1}{n+1} \sum_{t=0}^{n} \mathbf{E}\{\|X_t - \widetilde{Y}_t\|_2^2\}$, where

$$\mathbf{E}\left\{\|X_t - \widetilde{Y}_t\|_2^2\right\} = \sum_{i=1}^{p} \mu_{\Delta_t,i} \text{ with } D \in [0, \infty).$$

Since, in general, $R^{\mathrm{na}}_{t,i} \leq I(X_{t,i}; \widetilde{Y}_{t,i} | \widetilde{Y}_{t-1,i}),\ t \in \mathbb{N}_0^n,\ i = 1, \ldots, p$, then by (5.64), we obtain

$$\begin{aligned} \frac{1}{n+1} \sum_{t=0}^{n} \mathbf{E}\left\{\|X_t - \widetilde{Y}_t\|_2^2\right\} &= D_{0,n}(R^{\mathrm{na}}) = \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \mu_{\Lambda_t,i} e^{-2R^{\mathrm{na}}_{t,i}} \\ &\geq \frac{1}{n+1} \sum_{t=0}^{n} \sum_{i=1}^{p} \mu_{\Lambda_t,i} e^{-2I(X_{t,i}; \widetilde{Y}_{t,i} | \widetilde{Y}_{t-1,i})}, \end{aligned} \tag{5.66}$$

which is the desired result. This completes the proof. ∎

In the next corollary, we specialize the result of Theorem 5.7 to pointwise MSE distortion.

COROLLARY 5.8 (universal lower bound on pointwise MSE). *Let $\{X_t :\ t \in \mathbb{N}_0^n\}$ be the multidimensional Gauss–Markov process given by* (5.1) *and let $\{\widetilde{Y}_t :\ t \in \mathbb{N}_0^n\}$ be any estimator (not necessarily Gaussian) of $\{X_t :\ t \in \mathbb{N}_0^n\}$. The pointwise* MSE *is bounded below by*

$$\mathbf{E}\left\{\|X_t - \widetilde{Y}_t\|_2^2\right\} \geq \sum_{i=1}^{p} \mu_{\Lambda_t,i} e^{-2I(X_{t,i}; \widetilde{Y}_{t,i} | \widetilde{Y}_{t-1,i})} \text{ for each } t \in \mathbb{N}_0^n. \tag{5.67}$$

Proof. The proof is a special case of the derivation in Theorem 5.7. ∎

It should be noted that if we set $\widetilde{Y}_t = \widehat{X}_{t|t-1} = A_{t-1}Y_{t-1}$ in Theorem 5.7, then we have the lower bound (5.65).

In the next remark, we relate degenerated versions of the lower bound given by (5.65) to existing results in the literature.

Remark 8 (relations to existing results).

(a) (See [32, Theorem 5.8.1], [35].) Let $X = (X_1, \ldots, X_p)$ be an $\mathbb{R}^p$-valued Gaussian vector with distribution $X \sim \mathcal{N}(0; \Gamma_X)$ and let $Y = (Y_1, \ldots, Y_p)$ be its reproduction vector. Then, for any $D > 0$,

$$R(D) \triangleq \inf_{Q(dy|x):\mathbf{E}\|X-Y\|_2^2 \le D} I(X;Y) = \frac{1}{2}\sum_{i=1}^{p} \log\left\{\max\left(1, \frac{\lambda_i}{\xi}\right)\right\}, \tag{5.68}$$

where $\{\lambda_i : \ i = 1, \ldots, p\}$ are the eigenvalues of Γ_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 (5.68) is based on the reverse-waterfilling method (see [32, Lemma 5.8.2]). The above results are also obtained as a special case of Remark 6 if we assume an IID sequence $\{X_t : \ t \in \mathbb{N}_0^n\}$.

(b) Assume $X \sim N(0; \sigma_X^2)$. By [32, Theorem 1.8.7] the following holds:

$$\begin{aligned} R(D) &= \min_{Q(dy|x):\ \mathbf{E}\|X-Y\|_2^2 \le D} I(X;Y) = \frac{1}{2}\log\left\{\max\left(1, \frac{\sigma_X^2}{D}\right)\right\}, \ D \ge 0, \\ D(R) &= \min_{Q(dy|x):\ I(X;Y) \le R} \mathbf{E}\left\{\|X-Y\|_2^2\right\} = \sigma_X^2 e^{-2R}. \end{aligned}$$

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

$$Y = \left(1 - \frac{D}{\sigma_X^2}\right) X + V^c, \ V^c \sim \mathcal{N}\left(0; D\left(1 - \frac{D}{\sigma_X^2}\right)\right). \tag{5.69}$$

Note that (5.69) is a degenerated version of (5.17) assuming the model of (5.1) generates IID sequence $\{X_t : \ t \in \mathbb{N}_0^n\}$ as in (a), and the connection to Theorem 5.3 is established by setting $H_t = 1 - \frac{D}{\sigma_X^2}$, $\widehat{X}_{t|t-1} = 0$, and $\tilde{V}_t^c \sim N(0;1)$.

(c) (Lower bound on MSE [32, 1.8.8], [28].) Given a Gaussian RV $X \sim N(0; \sigma_X^2)$, then for any real-valued RV $\widetilde{Y}$(not necessarily Gaussian) the MSE is bounded below by

$$\mathbf{E}\|X - \widetilde{Y}\|_2^2 \ge \sigma_X^2 e^{-2I(X;\bar{Y})}. \tag{5.70}$$

The RDF of the Gaussian RV $X \sim \mathcal{N}(0; \sigma_X^2)$ and the lower bound in (5.70) are utilized in [28, 32] to derive optimal coding and decoding schemes for trans-

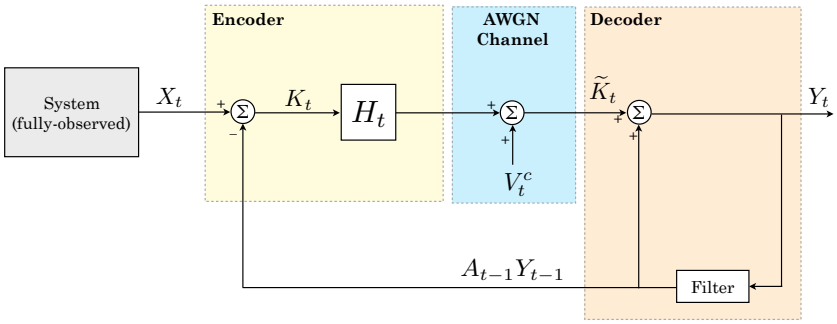

FIG. 3. *Realization of the optimal nonstationary reproduction of* $R^{\mathrm{na}}_{0,n}(D)$ *given by* (5.17) *that corresponds to the time-varying Gauss–Markov model of* (5.1) *subject to a total or pointwise* MSE *distortion.*

mitting 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 an IID Gaussian process. Although we do not pursue such problems in this paper, we note that Theorems 5.3 and 5.7 are necessary in order to derive optimal coding schemes for additive Gaussian channels with memory (including additive Gaussian memoryless channels).

5.3. Realization of (5.17) via an {`encoder, channel, decoder`}. In this section, we exemplify the relation between information-based estimation via NRDF and the fact that the latter can also be seen as a realization of an {`encoder, channel, decoder`} processing information optimally with zero-delay.

Realization with feedback. A realization of (5.17) by an {`encoder, channel, decoder`} with feedback is shown in Figure 3.

Feedback encoder. This is an innovations encoder which introduces the estimation error $\{K_t :\ t \in \mathbb{N}_0^n\}$, $K_t \triangleq X_t - \widehat{X}_{t|t-1}$, where $\{X_t :\ t \in \mathbb{N}_0^n\}$ is the Gaussian source process and $\widehat{X}_{t|t-1} = A_{t-1}Y_{t-1}$ is the a priori estimate of the filter; $\{H_t :\ t \in \mathbb{N}_0^n\}$ is a scaling matrix to be determined and has the structure of (5.18).

Feedback channel. This is an additive Gaussian noise channel of the form

$$\tilde{K}_t = H_t K_t + V_t^c,\ V_t^c \sim \mathcal{N}(0;\Delta_t H_t^{\mathrm{T}}),\ t \in \mathbb{N}_0^n, \tag{5.71}$$

where $\tilde{K}_t$ is the innovations process given by (5.13c).

Decoder. This is the a posteriori estimate of the filter $\widehat{X}_{t|t} = Y_t,\ t \in \mathbb{N}_0^n$. Specifically, the decoder introduces the innovations process $\{\tilde{K}_t :\ t \in \mathbb{N}_0^n\}$ and the a priori estimate of the filter $\widehat{X}_{t|t-1}$ that both added a result into the a posteriori estimate $\widehat{X}_{t|t}$ which is Y_t at each time instant t.

Realization without feedback. A realization of (5.17) by an {`encoder, channel, decoder`} without feedback can be derived as well.

5.4. Examples. In this section, we numerically compute the Gaussian NRDF of the *time-varying* Gauss–Markov process (5.1), using the reverse-waterfilling solution of Algorithms 1 and 2 that corresponds to Propositions 5.4 and 5.6. For Algorithm 2 we also give an example where we assume $\Delta_t \prec \Lambda_t\ \forall t$ and compare with the numerical solution obtained via SDP in [20, equation (19)]. Moreover, we give the closed form expression of a memoryless two-dimensional time-varying Gaussian source that corresponds to the optimization problem of Remark 6.

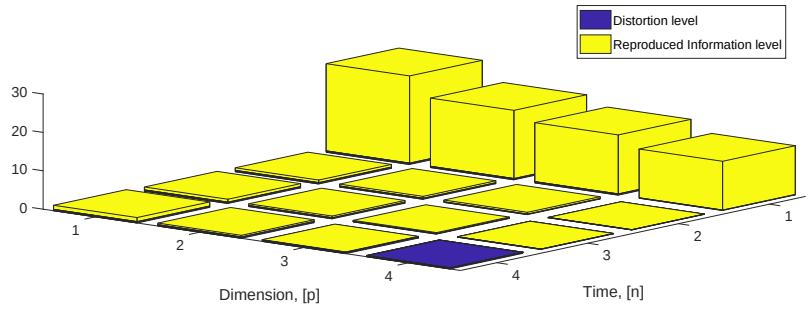

FIG. 4. *Dynamic reverse-waterfilling subject to a total* MSE *distortion for* $n = 3$ *time units.*

Example 2 (nonasymptotic regime subject to total MSE distortion). For this example, we choose $p = 4$, i.e., a four-dimensional source X_t, and a time horizon $n = 3$ for which we pick random values as entries of matrices $\{(A_t, K_{W_t}) : \ t = 0, \ldots, 3\}$ in the range $(0, 1)$, while K_{W_t} is chosen to be diagonal. We also choose the initial value of the covariance matrix of (5.22c) to pick up random values as entries Σ_{X_0} (hence, $\Lambda_0 = \Sigma_{X_0}$). We choose the distortion level $D = 1$. We run Algorithm 1 for error tolerance $\epsilon = 10^{-9}$ and an initial $\theta = \theta_0$ to start our iterations (a good starting point is $\theta_0 = \frac{1}{2D}$). Then, we proceed as follows:

(1) At $t = 0$, using (5.46a), we evaluate $\Delta_0^\dagger$. Then, from (5.45) we evaluate Δ_0. Next, from (5.22c), we evaluate $\Lambda_1 = A_0 \Delta_0 A_0^{\mathrm{T}} + K_{W_0}$.
(2) At $t = 1$, using (5.46a), we evaluate $\Delta_1^\dagger$ and subsequently, Δ_1 and Λ_2.
(3) Similarly, the procedure is repeated until $t = n = 3$. At $n = 3$ we evaluate $\Delta_3^\dagger$ using (5.46b).
(4) At the end, for the given value of θ, we check if $|\frac{1}{n+1} \sum_{t=0}^{n} \mathrm{trace}(\Delta_t) - D| \leq \epsilon$. If it does, we stop the iterations and the last evaluated value of θ is used to find the solution of Δ_t's and subsequently Λ_t's that when diagonalized give the desired waterlevels.
(5) If the approximation criterion $|\frac{1}{n+1} \sum_{t=0}^{n} \mathrm{trace}(\Delta_t) - D| \leq \epsilon$ is not satisfied, we update θ using (5.48); in this example we set $\gamma = 0.1$. We repeat the previous procedure (steps (1)–(4)) with the new value of θ $\forall t$.

The final value of the reverse-waterfilling solution is found after 362 iterations and it is shown in Figure 4. Then, the solution of (5.22) obtained via Algorithm 1 gives

$$\frac{1}{4} R_{0,3}^{\mathrm{na}}(D) = \frac{1}{2} \frac{1}{4} \sum_{t=0}^{3} \log \frac{|\Lambda_t|}{|\Delta_t|} = 4.8983 \text{ bits}.$$

Example 3 (nonasymptotic regime subject to pointwise MSE distortion). For this example, we choose $p = 3$, i.e., a three-dimensional source X_t, and a time horizon $n = 3$ for which we pick random values as entries of matrices $\{(A_t, K_{W_t}) : \ t = 0, \ldots, 3\}$ in the range $(0, 1)$ with K_{W_t} being diagonal. We also choose the initial value of the covariance matrix of (5.22c) to pick up random values as the entry of Σ_{X_0} (hence, $\Lambda_0 = \Sigma_{X_0}$). We choose distortion levels $(D_0, D_1, D_2, D_3) = (0.4, 1.3, 0.1, 0.4)$. We run Algorithm 2 for error tolerance $\epsilon = 10^{-9}$ and initial $\theta_t = \theta_{t_0}$, $t = 0, \ldots, 3$, to start our iterations (a good starting point is $\theta_0 = \frac{1}{2D_t}$ $t = 0, \ldots, 3$). Then, we proceed as follows:

(1) At $t = 0$, using (5.57a), we evaluate $\Delta_0^\dagger$. Then, from (5.45) we compute Δ_0 and from (5.22c) we evaluate $\Lambda_1 = A_0 \Delta_0 A_0^{\mathrm{T}} + K_{W_0}$.

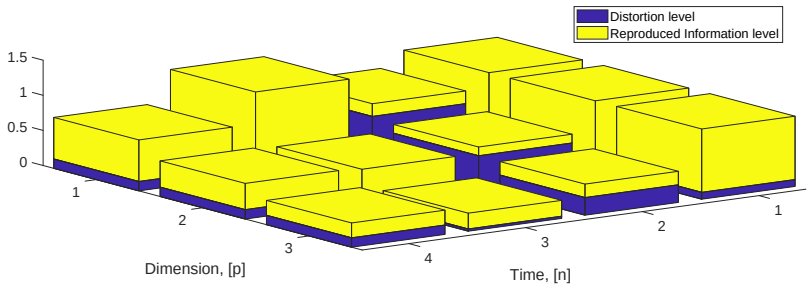

FIG. 5. *Dynamic reverse-waterfilling subject to a pointwise* MSE *distortion in time-domain for* $n = 3$ *time units.*

(2) At $t = 1$, using (5.57a), we evaluate $\Delta_1^\dagger$, and subsequently, Δ_1 and Λ_2.
(3) Similarly, the procedure is repeated until $t = n = 3$, where we compute $\Delta_3^\dagger$ from (5.57b).
(4) At the end, for each t and for the given value of θ_t, we check if $|\operatorname{trace}(\Delta_t) - D_t| \leq \epsilon$. If it does, we stop the iterations to find that particular θ_t; the last evaluated value of θ_t is then used to find the solution of matrix Δ_t via (5.57a) and (5.57b) for each t, which in turn when diagonalized gives the desired waterlevels.
(5) If the approximation criterion $|\operatorname{trace}(\Delta_t) - D_t| \leq \epsilon$ is not satisfied, we update θ_t using (5.59); in this example we choose $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.3, 0.1, 0.95, 0.4)$. We repeat the previous procedure (steps (1)–(4)) with the new value of θ_t for the specific t. This procedure is repeated $\forall t$.

The final value of the reverse-waterfilling solution is found after $(t = 0, t = 1, t = 2, t = 3) = (540, 543, 2715, 431)$ iterations and it is shown in Figure 5. Then, the solution of (5.52) obtained via Algorithm 2 gives

$$\frac{1}{4} R_{0,3}^{\mathrm{na}}(D_0, D_1, D_2, D_3) = \frac{1}{2}\frac{1}{4}\sum_{t=0}^{3} \log \frac{|\Lambda_t|}{|\Delta_t|} = 3.5066 \text{ bits}.$$

Example 4 (comparison of Algorithm 2 to SDP solution when $\Delta_t \prec \Lambda_t$). Consider the time-invariant version of (5.1), i.e., $A_t \equiv A$, $\Sigma_{W_t} \equiv \Sigma_W$, $X_0 \sim \mathcal{N}(0; \Sigma_{X_0})$, and distortion levels $(D_0, D_1, D_2, D_3) = (1, 0.2, 2, 0.5)$, and we choose

$$(A, \Sigma_W) = \left(\begin{bmatrix} 0.5 & 0.3 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

We run Algorithm 2 following the procedure described in Example 3, for $n = 2$, error tolerance $\epsilon = 10^{-9}$, and an initial $\theta = \frac{1}{2D_t}$, $t = 0, 1, 2, 3$. Then, we use the same parameters (A, Σ_W) in the SDP algorithm of [20, section IV, equation (19)] that provides the optimal numerical solution for $R_{0,3}^{\mathrm{na}}(D_0, D_1, D_2, D_3)$. In Figure 6 we illustrate a comparison between Algorithm 2 and the SDP for each $R_t(D_t)$ at each $t = 0, 1, 2, 3$. According to this, the computation via Algorithm 2 gives precisely the same numerical result as the one obtained via SDP for each D_t. Then, the solution

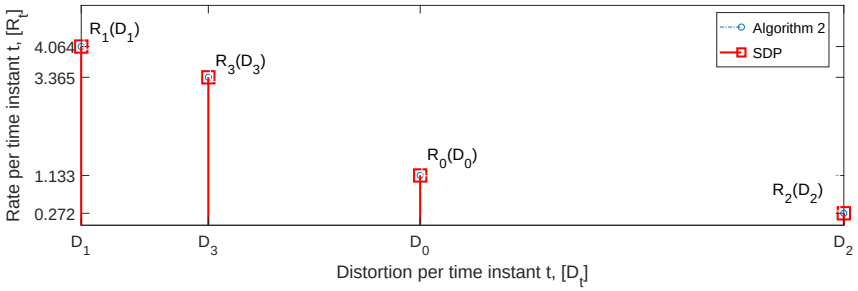

FIG. 6. *Comparison of Algorithm* 2 *to the optimal solution obtained via SDP for* $\Delta_t \prec \Lambda_t \ \forall t$.

of (5.52) obtained via Algorithm 2 gives

$$\frac{1}{4}R^{\mathrm{na}}_{0,3}(D_0, D_1, D_2, D_3) = \frac{1}{2}\frac{1}{4}\sum_{t=0}^{3}\log\frac{|\Lambda_t|}{|\Delta_t|} = 2.2086 \text{ bits}.$$

6. Conclusions and future directions. In this paper, we derived information-based causal filters via nonanticipative rate distortion theory in the finite-time horizon. Unlike classical Kalman filters, the new information-based causal filters are characterized by a time-space reverse-waterfilling algorithm. We developed iterative algorithms to compute the dynamic reverse-waterfilling optimization problem subject to a total and per-letter MSE distortion constraint. These algorithms provide tight upper bounds to the optimal solution, although in some cases these perform optimally. Further, we established a universal lower bound on the total and pointwise MSE of any estimator of a Gaussian random process. Our theoretical framework is demonstrated via several numerical experiments.

Future directions and open problems include the following:

(1) Remove the condition that K_{W_t} is full rank.
(2) Analyze the per unit time limit $\lim_{n\longrightarrow\infty}\frac{1}{n+1}R^{\mathrm{na}}_{0,n}(D)$.
(3) Compute closed-form expressions of $R^{\mathrm{na}}_{0,n}(D)$ for specific examples.
(4) Characterize $R^{\mathrm{na}}_{0,n}(D)$ for autoregressive Gaussian models with memory K, such as

$$X_t = \sum_{j=1}^{K} A_{t,j}X_{t-j} + W_t, \tag{6.1}$$

and to Gaussian sources governed by partially observed processes.
(5) Develop schemes to compute optimally the resulting time-space reverse-waterfilling optimization problems of Theorem 5.3 and Corollary 5.5.
(6) Generalize the results to controlled systems.

Appendix A. Proof of Theorem 5.3. (1) The realization is given in [9, Theorem 5] without determining the optimal structure of matrices $(H_t, K_{V_t^c})$. One may recognize that the choice of $(H_t, K_{V_t^c})$ given in (5.18) ensures that the dynamics of Δ_t are decoupled from the dynamics of Λ_t, since the Kalman filter gain in Lemma 5.2, (5.13), satisfies $\Lambda_t H_t^{\mathrm{T}}(H_t\Lambda_t H_t^{\mathrm{T}} + K_{V_t^c})^{-1} = I$. This can be used to show an achievable lower bound, when the reproduction distribution satisfies (5.14).

(2) Equation (5.22) follows directly from (1) by evaluating the NRDF. For the rest, we apply the KKT conditions [36, Chapter 5.5.3] to the optimization problem

(5.22). Define the augmented Lagrange functional as follows:

(A.1)
$$\mathcal{L}(\{\Delta_t,\Lambda_t\}_{t=0}^n,\ \theta,\ \{F_t^1,F_t^2\}_{t=0}^n)$$
$$= \frac{1}{2}\sum_{t=0}^n \log\frac{|\Lambda_t|}{|\Delta_t|}$$
$$+\theta\left(\sum_{t=0}^n \operatorname{trace}(\Delta_t) - D(n+1)\right) - \sum_{t=0}^n \operatorname{trace}\left(F_t^1\Delta_t\right) + \sum_{t=0}^n \operatorname{trace}\left(F_t^2\left(\Delta_t-\Lambda_t\right)\right),$$

where $\theta\in[0,\infty)$ is a Lagrange multiplier for the distortion constraint $\sum_{t=0}^n \operatorname{trace}(\Delta_t) \le D(n+1)$, and $F_t^j\in\mathcal{S}_+^{p\times p},\ j=1,2$, are the Lagrange multiplier matrices responsible for $\Delta_t\in S_+^{p\times p},\Lambda_t\succeq\Delta_t,\ t\in\mathbb{N}_0^n$. We write the first right-hand-side term of (A.1) as follows:

$$\frac{1}{2}\sum_{t=0}^n \log\frac{|\Lambda_t|}{|\Delta_t|} = \frac{1}{2}\sum_{t=0}^n\left(\log|\Lambda_t| - \log|\Delta_t|\right)$$
$$= \frac{1}{2}\log|\Lambda_0| + \frac{1}{2}\sum_{t=1}^n \log\left|A_{t-1}\Delta_{t-1}A_{t-1}^{\mathrm{T}} + K_{W_{t-1}}\right| - \frac{1}{2}\sum_{t=0}^n\log|\Delta_t|$$
$$= \frac{1}{2}\left\{\underbrace{\log|\Lambda_0|}_{\text{initial step}} + \sum_{t=0}^{n-1}\log\frac{|A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}|}{|\Delta_t|} - \underbrace{\log|\Delta_n|}_{\text{final step}}\right\}. \tag{A.2}$$

Also, we write the last right-hand-side term of (A.1) as follows:

(A.3)
$$\sum_{t=0}^n \operatorname{trace}\left(F_t^2\left(\Delta_t-\Lambda_t\right)\right)$$
$$= \operatorname{trace}\left(F_0^2\left(\Delta_0-\Lambda_0\right)\right) + \sum_{t=1}^n \operatorname{trace}\left(F_t^2\left(\Delta_t - A_{t-1}\Delta_{t-1}A_{t-1}^{\mathrm{T}} - K_{W_{t-1}}\right)\right)$$
$$= \sum_{t=0}^n \operatorname{trace}\left(F_t^2\Delta_t\right) - \operatorname{trace}\left(F_0^2\Lambda_0\right) - \sum_{t=1}^n \operatorname{trace}\left(F_t^2\left(A_{t-1}\Delta_{t-1}A_{t-1}^{\mathrm{T}}+K_{W_{t-1}}\right)\right)$$
$$= -\operatorname{trace}\left(F_0^2\Lambda_0\right) + \operatorname{trace}(F_n^2\Delta_n) + \sum_{t=0}^{n-1}\operatorname{trace}\left(F_t^2\Delta_t\right)$$
$$-\sum_{t=0}^{n-1}\operatorname{trace}\left(F_{t+1}^2\left(A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}\right)\right).$$

Hence, using (A.2) and (A.3), the augmented Lagrange functional can be reformulated

as follows:

$$\begin{aligned}(\text{A.4})\qquad &\mathcal{L}(\{\Delta_t,\Lambda_t\}_{t=0}^n,\ \theta,\ \{F_t^1,F_t^2\}_{t=0}^n)\\ &\quad=\left(\frac{1}{2}\log|\Lambda_0|-\operatorname{trace}(F_0^2\Lambda_0)\right)\\ &\qquad-\frac{1}{2}\log|\Delta_n|+\theta\operatorname{trace}(\Delta_n)+\operatorname{trace}(\left[F_n^2-F_n^1\right]\Delta_n)-\theta(n+1)D\\ &\qquad+\sum_{t=0}^{n-1}\Bigg\{\frac{1}{2}\log\frac{|A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}|}{|\Delta_t|}+\operatorname{trace}(\left[F_t^2-F_t^1\right]\Delta_t)\\ &\qquad\qquad-\operatorname{trace}\left(F_{t+1}^2\left(A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}\right)\right)+\theta\operatorname{trace}(\Delta_t)\Bigg\}.\end{aligned}\tag{A.5}$$

Recall by assumption $K_{W_t}\in S_{++}^{p\times p}$, and define $B_t\triangleq A_t^{\mathrm{T}}K_{W_t}^{-1}A_t$. Then, $\log\frac{|A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}|}{|\Delta_t|}$ in (A.5) is expressed as follows:

$$\begin{aligned}\log\frac{|A_t\Delta_tA_t^{\mathrm{T}}+K_{W_t}|}{|\Delta_t|}&=\log\left|K_{W_t}\left(K_{W_t}^{-1}A_t\Delta_tA_t^{\mathrm{T}}+I\right)\right|-\log|\Delta_t|\\ &=\log|K_{W_t}|+\log\left|K_{W_t}^{-1}A_t\Delta_tA_t^{\mathrm{T}}+I\right|-\log|\Delta_t|\\ &\overset{(a)}{=}\log|K_{W_t}|+\log\left|A_t^{\mathrm{T}}K_{W_t}^{-1}A_t\Delta_t+I\right|-\log|\Delta_t|\\ &=\log|K_{W_t}|+\log|B_t\Delta_t+I|-\log|\Delta_t|,\end{aligned}\tag{A.6}$$

where (a) is due to Sylvester's determinant identity [37, Corollary 18.1.2]. Substituting (A.6) into (A.5), the Lagrange functional is given by (5.23).

By the KKT conditions, ${\Delta_t}^*\in\mathcal{S}_+^{p\times p}, t\in\mathbb{N}_0^n$, achieves the minimum if (5.24)–(5.30b) hold.

We remark that the terms in (5.30a) are the complementary slackness conditions, the terms in (5.30c) are the primal feasibility conditions, and the terms in (5.30b) are the dual feasibility conditions. Note that the affine (linear) constraints $\sum_{t=0}^n\operatorname{trace}(\Delta_t)\le D(n+1)$ and $\Delta_t-\Lambda_t\preceq 0$ satisfy Slater's conditions (see, e.g., [36, Chapter 5.5.3]) and since the problem is convex it turns out that the KKT conditions are necessary and sufficient conditions for global optimality.

By (5.24), performing the derivative of the Lagrangian (5.23), we obtain (5.25)–(5.27). Since the problem is convex the MSE (5.30b) is satisfied with equality, hence $\theta>0$, because $F_t^1,\ F_t^2$ are positive semidefinite, and thus, if $\theta=0$, then ${\Delta_t}^*$ is not full rank, and $|{\Delta_t}^*|=0,\ t\in\mathbb{N}_0^n$, which gives infinite rate. Thus, for any $D>0$, then ${\Delta_t}^*\in\mathcal{S}_{++}^{p\times p}$, which implies $F_t^1=0$, while $F_t^2\in\mathcal{S}_+^{p\times p}$. This completes the proof. ∎

Appendix B. Proof of Proposition 5.4. In the solution of the Riccati equation (5.28), we omit F_t^2 and F_{t+1}^2 and, hence, this corresponds to the optimal solution when $\Delta_t\prec\Lambda_t$ and therefore, based on the complementary slackness conditions $F_t^2=0\ \forall t\in\mathbb{N}_0^n$. However, for the cases for which $\Delta_t\prec\Lambda_t$ does not hold, this solution is not necessarily the optimal one, and as a result, it serves as an upper bound. This completes the proof. ∎

REFERENCES

[1] A. K. Gorbunov and M. S. Pinsker, *Nonanticipatory and prognostic epsilon entropies and message generation rates*, Probl. Inf. Transm., 9 (1973), pp. 184–191.

[2] C. E. Shannon, *Coding Theorems for a Discrete Source with a Fidelity Criterion*, in Claude Elwood Shannon: Collected Papers, N. J. A. Sloane and A. D. Wyner, eds., Wiley, New York, 1993, pp. 325–350.

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

[4] R. S. Bucy, *Distortion rate theory and filtering*, IEEE Trans. Inform. Theory, 28 (1982), pp. 336–340.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, New York, 2006.

[6] J. I. Galdos and D. E. Gustafson, *Information and distortion in reduced-order filter design*, IEEE Trans. Inform. Theory, 23 (1977), pp. 183–194.

[7] S. C. Tatikonda, *Control Under Communication Constraints*, Ph.D. thesis, MIT, Cambridge, MA, 2000.

[8] D. Neuhoff and R. Gilbert, *Causal source codes*, IEEE Trans. Inform. Theory, 28 (1982), pp. 701–713.

[9] A. K. Gorbunov and M. S. Pinsker, *Prognostic epsilon entropy of a Gaussian message and a Gaussian source*, Probl. Inf. Transm., 10 (1974), pp. 93–109.

[10] S. Tatikonda and S. Mitter, *Control under communication constraints*, IEEE Trans. Automat. Control, 49 (2004), pp. 1056–1068.

[11] W. S. Wong and R. W. Brockett, *Systems with finite communication bandwidth constraints* II: *Stabilization with limited information feedback*, IEEE Trans. Automat. Control, 44 (1999), pp. 1049–1053.

[12] N. Elia, *When Bode meets Shannon: Control-oriented feedback communication schemes*, IEEE Trans. Automat. Control, 49 (2004), pp. 1477–1488.

[13] C. D. Charalambous and A. Farhadi, *LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels*, Automatica J. IFAC, 44 (2008), pp. 3181–3188.

[14] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed, *Nonanticipative rate distortion function and relations to filtering theory*, IEEE Trans. Automat. Control, 59 (2014), pp. 937–952.

[15] C. D. Charalambous and P. A. Stavrou, *Optimization of directed information and relations to filtering theory*, in Proceedings of the European Control Conference, 2014, pp. 1385–1390.

[16] P. A. Stavrou, C. K. Kourtellaris, and C. D. Charalambous, *Information nonanticipative rate distortion function and its applications*, in Coordination Control of Distributed Systems, J. H. van Schuppen and T. Villa, eds., Springer, New York, 2015, pp. 317–324.

[17] M. S. Derpich and J. Østergaard, *Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources*, IEEE Trans. Inform. Theory, 58 (2012), pp. 3131–3152.

[18] C. K. Kourtellaris, C. D. Charalambous, and P. A. Stavrou, *Nonanticipative duality of sources and channels with memory and feedback*, in Coordination Control of Distributed Systems, J. H. van Schuppen and T. Villa, eds., pp. 325–335, Springer, New York, 2015.

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

[20] T. Tanaka, K. K. K. Kim, P. Parrilo, and S. Mitter, *Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs*, IEEE Trans. Automat. Control, 62 (2017), pp. 1896–1910.

[21] S. Tatikonda, A. Sahai, and S. Mitter, *Stochastic linear control over a communication channel*, IEEE Trans. Automat. Control, 49 (2004), pp. 1549–1561.

[22] P. A. Stavrou, T. Tanaka, and S. Tatikonda, *The Time-Invariant Multidimensional Gaussian Sequential Rate-Distortion Problem Revisited*, arXiv:1711.09853; IEEE Trans. Automat. Control, to appear.

[23] P. A. Stavrou, T. Charalambous, and C. D. Charalambous, *Filtering with fidelity for time-varying Gauss-Markov processes*, in Proceedings of the IEEE Conference on Decision and Control, 2016, pp. 5465–5470.

[24] R. M. Gray and T. Hashimoto, *Rate-distortion functions for nonstationary Gaussian autoregressive processes*, in Proceedings of the Data Compression Conference, 2008, pp. 53–62.

[25] C. D. Charalambous, C. K. Kourtellaris, and P. A. Stavrou, *On Shannon's duality of a source and a channel and nonanticipative communication and communication for control*, in Coordination Control of Distributed Systems, J. H. van Schuppen and T. Villa, eds., Springer, New York, 2015, pp. 291–305.
[26] P. E. Caines, *Linear Stochastic Systems*, Wiley Ser. Probab. Stat. 100, Wiley, New York, 1988.
[27] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer, New York, 1995.
[28] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes* II*: Applications*, 2nd ed., Springer, New York, 2001.
[29] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.
[30] C. D. Charalambous and P. A. Stavrou, *Directed information on abstract spaces: Properties and variational equalities*, IEEE Trans. Inform. Theory, 62 (2016), pp. 6019–6052.
[31] P. A. Stavrou, *Extremum Problems of Directed Information*, Ph.D. thesis, University of Cyprus, 2016.
[32] S. Ihara, *Information Theory—for Continuous Systems*, World Scientific, River Edge, NJ, 1993.
[33] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
[34] P. A. Stavrou, T. Charalambous, and C. D. Charalambous, *Finite-time nonanticipative rate distortion function for time-varying scalar-valued Gauss-Markov sources*, IEEE Control Syst. Lett., 2 (2018), pp. 175–180.
[35] R. E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1987.
[36] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, 2004.
[37] D. A. Harville, *Matrix Algebra from a Statistician's Perspective*, Springer, New York, 1997.