A novel sequential RDF to compute partially observable Markov processes with MSE distortion

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

17-08-2020 / 24-08-2020

CITATION

Stavrou, Fotios; Skoglund, Mikael (2020): A novel sequential RDF to compute partially observable Markov processes with MSE distortion. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.12816155.v1>

DOI

[10.36227/techrxiv.12816155.v1](https://doi.org/10.36227/techrxiv.12816155.v1)

A Novel Sequential RDF to Compute Partially Observable Markov Processes with MSE Distortion

Photios A. Stavrou and Mikael Skoglund

Abstract—We develop a new sequential rate distortion function to compute lower bounds on the average length of all causal prefix free codes for partially observable multivariate Markov processes with mean-squared error distortion constraint. Our information measure is characterized by a variant of causally conditioned directed information and is utilized in various application examples. First, it is used to optimally characterize a finite dimensional optimization problem for jointly Gaussian processes and to obtain the corresponding optimal linear encoding and decoding policies. Under the assumption that all matrices commute by pairs, we show that our problem can be cast as a convex program which achieves its global minimum. We also derive sufficient conditions which ensure that our assumption holds. We then solve the KKT conditions and derive a new reverse-waterfilling algorithm that we implement. If our assumption is violated, one can still use our approach to derive sub-optimal (upper bound) waterfilling solutions. For scalar-valued Gauss-Markov processes with additional observation noise, we derive a new closed form solution and we compare it with known results in the literature. For partially observable time-invariant Markov processes driven by additive i.i.d. system noise only, we recover using an alternative approach and thus strengthening a recent result by Kostina and Hassibi in [1, Theorem 9] whereas for time-invariant and spatially i.i.d. Markov processes driven by additive noise process we also derive new analytical lower bounds.

I. INTRODUCTION

Nonanticipatory ϵ -entropy was introduced in [2], [3] motivated by real-time communication with minimal encoding and decoding delays. This entity is shown to be a tight lower bound on causal codes for scalar processes [4] whereas for vector processes it provides a tight lower bound at high rates on causal codes and on the average length of all causal prefix free codes [5] (also termed zero-delay coding).

Inspired by the usefulness of nonanticipatory- ϵ entropy in real-time communication, Tatikonda *et al.* in [6] reinvented the same measure under the name sequential rate distortion function (RDF)¹ to study a linear fully observable Gaussian closed-loop control system over a memoryless communication channel subject to rate constraints. In particular, the authors of [6] used the sequential RDF subject to a pointwise mean-squared error (MSE) distortion constraint to describe a lower bound on the minimum cost of control for scalar-valued Gaussian processes and a suboptimal lower bound for the

multivariate case obtained by means of a reverse-waterfilling algorithm [8, 10.3.3].²

Tanaka *et al.* in [10] revisited the estimation/communication part of the problem introduced by Tatikonda *et al.* and showed that the specific description of the sequential RDF is semidefinite representable. Around the same time, Stavrou *et al.* in [11] solved the general KKT conditions that correspond to the rate distortion characterization of the optimal estimation problem in [6] and proposed a dynamic reverse-waterfilling characterization (for both pointwise and total MSE distortions) that computes optimally the KKT conditions as long as all dimensions of the multidimensional setup are active, which is the case at high rates regime. In addition, in [11] they found the optimal linear coding policies (by means of a linear forward test-channel realization) that achieve the specific rate distortion characterization thus filling a gap created in [2, Theorem 5]. Recently, the optimal realization therein was used as a benchmark in [12] to derive bounds on a zero delay multiple description source coding problem with feedback.

Kostina and Hassibi in [1] revisited the framework of [6] to derive bounds on rate-cost tradeoffs for time-invariant multivariate fully observable processes. One major result therein is a lower bound in the case where the system is driven by additive i.i.d. zero mean noise. For scalar-valued Markov processes, this bound generalizes naturally known results obtained in closed form, see, e.g., [3, Eq. (1.43)], [4, Theorem 3] beyond additive i.i.d. Gaussian zero mean noise processes. Recently the authors of [13] used a state augmentation technique to extend the characterization of the Gaussian nonanticipatory ϵ -entropy derived in [3] to nonstationary multivariate Gaussian autoregressive models of any finite order.

The extension of the framework in [6] to linear partially observable Gaussian control systems under noisy or noiseless communication channels was independently studied in [14] (see also [15, §VII]) and [1, §II]. In [14], [15, §VII], the authors choose to minimize the rate-cost tradeoffs via a multi-letter characterization where *the objective is cast by the directed information [16] from the observations process to the output of the decoder/controller when the system's encoder and decoder/controller are allowed to have access to previous decoder/controller signals via noiseless feedback.* Then, they showed that *the optimal system's coding policies can be realized by a pre-Kalman filter (KF) approach to reduce the partially observable system to a full observable one, a sensor design that uses a SDP representable problem and a*

P. A. Stavrou and M. Skoglund received funding by the KAW Foundation and the Swedish Foundation for Strategic Research.

P. A. Stavrou and M. Skoglund are with the Department of Information Science and Engineering, KTH Royal Institute of Technology, Sweden, email: {fstavrou, skoglund}@kth.se.

¹In the literature this information measure can also be found under the name nonanticipative RDF [7].

²The sub-optimality of the lower bound obtained in [6] for multivariate Gaussian processes was recently identified in [1], [9].

post-KF algorithm. Unfortunately, although it is pointed out in [15, Theorem 2] that an optimal policy to solve their problem numerically exists if and only if their SDP characterization is feasible, and in addition, its solution coincides with the rate-cost characterization, such conditions are never provided. In [1] the authors choose a rate-cost characterization similar to [15] and derived lower and upper bounds on the optimal solution of their characterization for jointly Gaussian processes. *This means that like in [15] an additional pre-KF algorithm is required to transform the partially observable problem to a fully observable one.*

A. Contributions

In this paper, we study a new sequential RDF reminiscent of the information measure introduced in [2], [3] to address problems modeled by partially observable Markov processes with MSE distortion. Our new information measure is defined via a *variant of causally conditioned directed information* [17, Chapter 3] *between the observations process and the output of the decoder when both the encoder and the decoder have access to the previous outputs of the decoder via noiseless feedback whereas the decoder is also allowed to have access to previous observations signals* (see Theorem 1, Definition 1). Armed with this major result, we obtain the following additional major results.

(R1) We show that the description of our bound is, in general, a lower bound on the description of the information measure utilized in [1], [15] to compute performance limitations on the communication/estimation part of a partially observable time-invariant LQG closed-loop control system (see Proposition 1); **(R2)** For partially observable Gaussian processes we completely characterize our problem as a finite dimensional optimization (see Theorems 2, 3) and show that the optimal solution can be realized by means of *only one* KF algorithm (see Lemma 2). We also derive the optimal linear policies including the identification of the realization coefficients that surprisingly show that the best policy at the *decoder's output follows a first-order Markov process* whereas the decoder end up being independent of the previous observations signals (see Theorems 2, 3);

(R3) We convexify the characterization obtained in **(R2)** under the assumption that all matrices in the system commute by pairs (see Theorem 4), that is, they are simultaneously diagonalizable by an orthogonal matrix.³ Then, we solve the corresponding KKT conditions [18] to obtain a new reverse-waterfilling algorithm (see Theorem 5) that we also implement (see Algorithm 1);

(R4) We give sufficient conditions that meet our assumption in **(R3)** (see Proposition 2);

(R5) For scalar-valued time-invariant Gauss-Markov processes with additional observation noise we derive a new closed form solution (see Corollary 1);

(R6) We extend our results to Markov processes driven by additive i.i.d. zero mean noise (not necessarily Gaussian): first, by recovering for a partially observable system without observations noise, the bound of [1, Theorem 9] using an

alternative method that leverages Minkowski's determinant inequality [19] and standard entropy power inequalities (EPI) [20]; second, by deriving two analytical solutions for a time-invariant multivariate fully observable system with additive i.i.d. observations noise, when each dimension is statistically i.i.d.. The last result is obtained under the assumption that *either the system model or the observations model are driven by additive i.i.d. zero mean noise process but not both.*

Discussion of our results. **(R1)** suggests that improvements (i.e., further lower bounds) can be obtained when minimizing the estimation/communication cost in the celebrated quantized partially observable LQG closed-loop control problem provided the control-theoretic separation principle holds [21]. Our result in **(R2)** provides a smaller system realization with significantly reduced complexity compared to [1], [14], [15] where two KF algorithms are required. Moreover, identifying the realization coefficients is important because they can lead to practical achievability schemes to bound the optimal performance theoretically attainable by causal codes [22] or the minimum average length of all causal prefix-free codes, see, e.g., [1], [4], [5], [23]. If the assumption of commuting matrices does not hold in **(R3)**, then, our reverse-waterfilling solution offers an elegant sub-optimal (upper bound) solution. In such case, the tightness of our bound depends on the structure of the given matrices and the dimensionality of the partially observable system. We note that the sub-optimality of our algorithm for arbitrarily chosen matrices comes with the advantage of an algorithmic approach that is extremely fast and easily adaptable to high dimensional systems when implemented (see, e.g., [24] for more details). Our result in **(R5)** is compared with an existing lower bound obtained in [25, Appendix A] using [1] to show via simulation experiments that the latter is not tight, in general, with respect to its optimal solution. This also means that our lower bound is the first exact lower bound obtained for scalar-valued Gauss-Markov processes with additional observations noise (i.e., it has an actual implementable realization). Our result in **(R5)** is also used to recover the known result obtained for scalar-valued Gauss-Markov processes derived in numerous papers, see, e.g. [3, Eq. (1.43)], [4, Theorem 3], [6, Equation (14)], and to compute analytically the RL gap due to having additional observations noise (see Proposition 3). To obtain our results in **(R6)**, we use the fact that the optimal linear policies obtained in **(R2)** are also the best linear policies for additive i.i.d. zero mean noise processes and correspond to a lower bound on the minimum average length of all causal prefix free codes (see Proposition 4).

Notation. We let $\mathbb{R} = (-\infty, \infty)$, $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, $\mathbb{N}_0 = \{0, 1, \dots\}$, $\mathbb{N}_0^n = \{0, 1, \dots, n\}$, $n \in \mathbb{N}_0$. Let $\mathcal{X}$ be a finite dimensional Euclidean space and $\mathcal{B}(\mathcal{X})$ the Borel σ -field of $\mathcal{X}$. A random variable (RV) defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a map $\mathbf{x} : \Omega \mapsto \mathcal{X}$, where $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a measurable space. We denote a sequence of RVs by $\mathbf{x}_r^t \triangleq (\mathbf{x}_r, \mathbf{x}_{r+1}, \dots, \mathbf{x}_t)$, $(r, t) \in \mathbb{Z} \times \mathbb{Z}$, $t \geq r$, and their realizations by $\mathbf{x}_r^t \in \mathcal{X}_r^t \triangleq \times_{k=r}^t \mathcal{X}_k$, for simplicity. If $r = -\infty$ and $t = -1$, we use the notation $\mathbf{x}_{-\infty}^{-1} = \mathbf{x}^{-1}$, and if $r = 0$, we use the notation $\mathbf{x}_0^t = \mathbf{x}^t$. The distribution of the RV $\mathbf{x}$ on $\mathcal{X}$ is denoted by $\mathbf{P}(dx)$. The conditional distribution of

³More details are given in the supplementary material of Appendix A.

a RV $\mathbf{y}$ given $\mathbf{x} = x$ is denoted by $\mathbf{P}(dy|x)$. The transpose and covariance of a random vector $\mathbf{x}$ are denoted by $\mathbf{x}^\top$ and $\Sigma_{\mathbf{x}}$. We denote the determinant, trace, diagonal and diagonal elements of a square matrix S by $|S|$, $\text{trace}(S)$, $\text{diag}(S)$ and $[\cdot]_{ii}$. We denote the transpose of a square or rectangular matrix S by $S^\top$. We denote the eigenvalues of a square matrix $S \in \mathbb{R}^{p \times p}$ by $\{\mu_{S,i}\}_{i=1}^p$. The notation $\Sigma_S \succ 0$ (resp. $\Sigma_S \succeq 0$) denotes a positive definite (resp. positive semi-definite) matrix. The notation $A \succ B$ (resp. $A \succeq B$) means $A - B \succ 0$ (resp. $A - B \succeq 0$). We denote a $p \times p$ identity matrix by I_p . $R^G(D)$ denotes the Gaussian version of a specific RDF and $h^G(\mathbf{x})$ (resp. $h^G(\mathbf{x}|\mathbf{y})$) denotes the Gaussian differential entropy (resp. conditional Gaussian differential entropy) of a distribution $\mathbf{P}(dx)$ (resp. $\mathbf{P}(dx|\mathbf{y})$). $N(\cdot)$ denotes the entropy power (EP) of a RV or a random vector $\mathbf{x}$. The expectation operator is denoted by $\mathbb{E}\{\cdot\}$; $\|\cdot\|$ denotes Euclidean norm; $[\cdot]^+ \triangleq \max\{0, \cdot\}$.

II. PROBLEM STATEMENT

We consider the “online” or zero delay source coding setup of Fig. 1. In this setting, the “hidden” $\mathbb{R}^p$ -valued Gaussian source is modeled by a discrete-time time-invariant Markov process as follows

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{x}_0 = \bar{\mathbf{x}}, \quad t \in \mathbb{N}_0, \quad (1)$$

where $A \in \mathbb{R}^{p \times p}$ is a deterministic matrix, $\mathbf{x}_0 \in \mathbb{R}^p \sim (0; \Sigma_{\mathbf{x}_0})$ is the initial state, $\mathbf{w}_t \in \mathbb{R}^p \sim (0; \Sigma_{\mathbf{w}})$, $\Sigma_{\mathbf{w}} \succ 0$ is an i.i.d. sequence, independent of $\mathbf{x}_0$.

The observation process is modeled by the discrete-time time-invariant process

$$\mathbf{z}_t = C\mathbf{x}_t + \mathbf{n}_t, \quad t \in \mathbb{N}_0, \quad (2)$$

where $C \in \mathbb{R}^{m \times p}$ is a deterministic full row rank matrix, $\mathbf{n}_t \in \mathbb{R}^m \sim (0; \Sigma_{\mathbf{n}})$, $\Sigma_{\mathbf{n}} \succeq 0$, is an i.i.d. sequence, independent of $\{\mathbf{w}_t : t \in \mathbb{N}_0\}$, $\mathbf{x}_0$.

System’s operation: At every time step $t \in \mathbb{N}_0$, the hidden vector source $\mathbf{x}_t$ is conveyed with an additional noise $\mathbf{n}_t$ at the *encoder* who observes the impair measurement $\mathbf{z}_t$ (provided $\mathbf{z}^{t-1}$ are already observed) and produces a binary codeword $\mathbf{m}_t$ of length l_t (in bits) from a predefined set of codewords $\mathcal{M}_t$ of at most countable number of codewords. The codewords are transmitted across an instantaneous noiseless digital channels to a decoder. Upon receiving $\mathbf{m}^t$, the *decoder* immediately produces an estimate $\mathbf{y}_t$ of the observations sample $\mathbf{z}_t$, under the assumption that $\mathbf{y}^{t-1}$ is already reproduced. The analysis of the noiseless digital channel is restricted to the class of prefix-free binary codes $\mathbf{m}_t$.

Fig. 1. System model.

Zero-delay source coding: Formally, the zero-delay source coding problem of Fig. 1 can be explained as follows. Define the input and output alphabet of the noiseless digital channel by $\mathcal{B} = \{1, 2, \dots, B\}$ where $B = \max_t |\mathcal{M}_t|$ (that is allowed to be infinite). The elements in $\mathcal{B}$ enumerate the codewords of $\mathcal{M}_t$. The encoder is specified by the sequence of functions $\{f_t : t \in \mathbb{N}_0\}$ with $f_t : \mathcal{B}^{t-1} \times \mathcal{Z}^t \rightarrow \mathcal{B}_t$. At time $t \in \mathbb{N}_0$, the output of the encoder is a message $m_t = f_t(m^{t-1}, z^t)$ with $m_0 = f_0(z_0)$ which is transmitted through a noiseless channel to the decoder. The decoder is specified by the sequence of measurable functions $\{g_t : t \in \mathbb{N}_0\}$ with $g_t : \mathcal{B}^t \rightarrow \mathcal{Y}_t$. For each $t \in \mathbb{N}_0$, the decoder generates $y_t = g_t(m^t)$ with $y_0 = g_0(m_0)$ assuming y^{t-1} is already generated. The design in Fig. 1 is required to yield the long-term average distortion $\limsup_{n \rightarrow \infty} \frac{1}{n+1} \mathbb{E}\{d(\mathbf{x}^n, \mathbf{y}^n)\} \leq D$, where $D \in (D_{\min}, D_{\max}] \subseteq (0, \infty]$ is a pre-specified distortion level, with $d(x^n, y^n) \triangleq \sum_{t=0}^n \|\mathbf{x}_t - \mathbf{y}_t\|^2$. The rate at the encoder is given by the long term average codeword length of all instantaneous codes, denoted by $\limsup_{n \rightarrow \infty} \frac{1}{n+1} \sum_{t=0}^n \mathbb{E}\{l_t\}$. We denote by $\mathbf{L}_n \triangleq \sum_{t=0}^n l_t$ the accumulated number of bits received by the decoder at the time it reproduces the estimate $\mathbf{y}_n$.

Performance. The performance of the multi-input multi-output (MIMO) system in Fig. 1 can be cast by the following optimization problem:

$$\begin{aligned} R_{\text{ZD}, \text{in}}^{\text{op}}(D) &\triangleq \inf_{\substack{m_t = f_t(m^{t-1}, z^t), \quad t \in \mathbb{N}_0 \\ y_t = g_t(m^t)}} \limsup_{n \rightarrow \infty} \frac{1}{n+1} \mathbb{E}\{\mathbf{L}_n\}. \\ \text{s. t. } &\limsup_{n \rightarrow \infty} \frac{1}{n+1} \mathbb{E}\left\{\sum_{t=0}^n \|\mathbf{x}_t - \mathbf{y}_t\|^2\right\} \leq D \end{aligned} \quad (3)$$

III. A NEW LOWER BOUND

In this section, we derive a new information measure, that gives a new lower bound on (3). To do it, we first prove a novel data processing theorem that corresponds to the system of Fig. 1.

First, we write the data processing of information for the MIMO system of Fig. 1 in terms of its joint distribution. In particular, the joint distribution induced by the joint process $\{(\mathbf{z}_t, \mathbf{m}_t, \mathbf{y}_t) : t \in \mathbb{N}_0^n\}$ admits the following decomposition:

$$\begin{aligned} \mathbf{P}(dy^n, dm^n, dz^n) &= \otimes_{t=0}^n \mathbf{P}(dy_t, dm_t, dz_t | y^{t-1}, m^{t-1}, z^{t-1}) \\ &= \otimes_{t=0}^n \mathbf{P}(dy_t | y^{t-1}, z^t, m^t) \otimes \mathbf{P}(dm_t | m^{t-1}, z^t, y^{t-1}) \\ &\quad \otimes \mathbf{P}(dz_t | z^{t-1}, y^{t-1}, m^{t-1}) \\ &\stackrel{(a)}{=} \otimes_{t=0}^n \mathbf{P}(dy_t | y^{t-1}, z^{t-1}, m^t) \otimes \mathbf{P}(dm_t | m^{t-1}, z^t, y^{t-1}) \\ &\quad \otimes \mathbf{P}(dz_t | z^{t-1}, y^{t-1}), \end{aligned} \quad (4)$$

where (a) stems from the fact that we assume in our system the following conditional independence constraints:

$$\mathbf{P}(dz_t | z^{t-1}, y^{t-1}, m^{t-1}) = \mathbf{P}(dz_t | z^{t-1}, y^{t-1}), \quad (5)$$

$$\mathbf{P}(dy_t | y^{t-1}, z^t, m^t) = \mathbf{P}(dy_t | y^{t-1}, z^{t-1}, m^t). \quad (6)$$

Remark 1: (Trivial initial information) In (4) we assume that the joint distribution $\mathbf{P}(dz^{-1}, dm^{-1}, dy^{-1})$ generates trivial information.

The following data processing theorem, is a main result of this paper.

Theorem 1: (Data processing theorem) Provided the decomposition of the joint distribution in (4) holds, the system in Fig. 1 admits the following data processing inequalities:

$$I(\mathbf{z}^n \rightarrow \mathbf{y}^n | \mathbf{z}^{n-1}) \stackrel{(ii)}{\leq} I(\mathbf{z}^n \rightarrow \mathbf{m}^n | \mathbf{z}^{n-1}, \mathbf{y}^{n-1}) \stackrel{(i)}{\leq} \sum_{t=0}^n \mathbb{E}\{\mathbf{1}_t\}, \quad (7)$$

where

$$I(\mathbf{z}^n \rightarrow \mathbf{y}^n | \mathbf{z}^{n-1}) = \sum_{t=0}^n I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{z}^{t-1}, \mathbf{y}^{t-1}),$$

$$I(\mathbf{z}^n \rightarrow \mathbf{m}^n | \mathbf{z}^{n-1}, \mathbf{y}^{n-1}) = \sum_{t=0}^n I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}),$$

and $I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) < \infty$, $I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) < \infty$, $\forall t$.

Proof: We first prove (i).

$$\begin{aligned} \sum_{t=0}^n \mathbb{E}\{\mathbf{1}_t\} &\geq \sum_{t=0}^n H(\mathbf{m}_t | \mathbf{m}^{t-1}) \\ &\stackrel{(a)}{\geq} \sum_{t=0}^n H(\mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) \\ &\stackrel{(b)}{\geq} \sum_{t=0}^n \left[H(\mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) \right. \\ &\quad \left. - H(\mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^t, \mathbf{y}^{t-1}) \right] \\ &\stackrel{(c)}{=} \sum_{t=0}^n I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \\ &\equiv I(\mathbf{z}^n \rightarrow \mathbf{m}^n | \mathbf{y}^{n-1}, \mathbf{z}^{n-1}), \end{aligned}$$

where (a) follows because conditioning reduces entropy [8]; (b) follows because of the non-negativity of the discrete entropy [8]; (c) follows by construction.

Next, we prove (ii). Here we show that $I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) - I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \geq 0$, $\forall t$. This is shown as follows:

$$\begin{aligned} &I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) - I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \\ &\stackrel{(d)}{=} I(\mathbf{z}_t; \mathbf{m}_t, \mathbf{y}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) - I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \\ &\stackrel{(e)}{=} I(\mathbf{z}_t; \mathbf{m}^t | \mathbf{z}^{t-1}, \mathbf{y}^t) \geq 0, \end{aligned} \quad (8)$$

where (d) follows from an adaptation of [26, Lemma 3.3] to processes, i.e., $I(\mathbf{z}_t; \mathbf{m}_t, \mathbf{y}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) = I(\mathbf{z}_t; \mathbf{m}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) + I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{m}^t, \mathbf{z}^{t-1}, \mathbf{y}^{t-1})$ and the second term is zero because of the conditional independence constraint (6); (e) follows by the chain rule of conditional mutual information (again an adaptation of [26, Theorem 3.4]) which decomposes the conditional mutual information in two different ways, i.e.,

$$\begin{aligned} &I(\mathbf{z}_t; \mathbf{m}^t, \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \\ &= I(\mathbf{z}_t; \mathbf{m}^{t-1} | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) + I(\mathbf{z}_t; \mathbf{m}_t, \mathbf{y}_t | \mathbf{m}^{t-1}, \mathbf{z}^{t-1}, \mathbf{y}^{t-1}) \\ &= I(\mathbf{z}_t; \mathbf{m}^t | \mathbf{y}^t, \mathbf{z}^{t-1}) + I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}), \end{aligned}$$

where $I(\mathbf{z}_t; \mathbf{m}^{t-1} | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) = 0$ due to the conditional independence in (5).

From (8) we obtain $\sum_{t=0}^n I(\mathbf{z}_t; \mathbf{m}^t | \mathbf{z}^{t-1}, \mathbf{y}^t) \geq 0$ (due to non-negativity of conditional mutual information [8]). This completes the derivation. ■

We stress the following technical remark on Theorem 1.

Remark 2: (Technical remark) (1) One can derive precisely the same result with Theorem 1 by not allowing the observations process $\{\mathbf{z}_t : t \in \mathbb{N}_0\}$ to have access to $\{\mathbf{y}^{t-1} : t \in \mathbb{N}_0\}$, $\mathbf{y}^{-1} = \emptyset$, i.e., in (5) we have $\mathbf{P}(dz_t | z^{t-1}, y^{t-1}) = \mathbf{P}(dz_t | z^{t-1})$. In such case, the information measures in (7) will be denoted as follows $I(\mathbf{z}^n \rightarrow \mathbf{y}^n | \mathbf{z}^{n-1}) \equiv I(\mathbf{z}^n; \mathbf{y}^n | \mathbf{z}^{n-1})$, $I(\mathbf{z}^n \rightarrow \mathbf{m}^n | \mathbf{z}^{n-1}, \mathbf{y}^{n-1}) \equiv I(\mathbf{z}^n; \mathbf{m}^n | \mathbf{z}^{n-1}, \mathbf{y}^{n-1})$.

(2) Theorem 1 is a non-trivial generalization of [15, Lemma 1]. Compared to [15, Lemma 1] that assumes noiseless feedback information at both the encoder and the decoder here we additionally assume knowledge of all previous observation signals at the decoder.

The next lemma plays an important role in some of our results in the sequel.

Lemma 1: (Inequality bound) The following bound holds,

$$I(\mathbf{z}^n \rightarrow \mathbf{y}^n | \mathbf{z}^{n-1}) \leq I(\mathbf{z}^n \rightarrow \mathbf{y}^n), \quad (9)$$

where $I(\mathbf{z}^n \rightarrow \mathbf{y}^n) \triangleq \sum_{t=0}^n I(\mathbf{z}^t; \mathbf{y}_t | \mathbf{y}^{t-1})$ [16] assuming $I(\mathbf{z}^t; \mathbf{y}_t | \mathbf{y}^{t-1}) < \infty$, $I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) < \infty$, $\forall t$.

Proof: It suffices to show that $I(\mathbf{z}^t; \mathbf{y}_t | \mathbf{y}^{t-1}) - I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \geq 0$, $\forall t$. Observe that by definition we obtain

$$\begin{aligned} &I(\mathbf{z}^t; \mathbf{y}_t | \mathbf{y}^{t-1}) - I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \\ &\stackrel{(a)}{=} h(\mathbf{y}_t | \mathbf{y}^{t-1}) - h(\mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \stackrel{(b)}{\geq} 0, \end{aligned} \quad (10)$$

where (a) follows by definition under the assumption that both information measures are well defined; (b) follows because conditioning reduces (differential) entropy. ■

Next, we show how to construct the new sequential RDF.

Hidden Source Distribution. The hidden source distribution $\mathbf{x}_t$ satisfies conditional independence

$$\mathbf{P}(dx_t | x^{t-1}, z^{t-1}, y^{t-1}) \triangleq \mathbf{P}(dx_t | x_{t-1}), \quad t \in \mathbb{N}_0^n. \quad (11)$$

At $t = 0$ we assume $\mathbf{P}(dx_0)$. Also, by Bayes' rule we obtain $\mathbf{P}(dx^n) \triangleq \otimes_{t=0}^n \mathbf{P}(dx_t | x_{t-1})$.

Observations Distribution. The observations process $\mathbf{z}_t$ satisfy conditional independence

$$\mathbf{P}(dz_t | z^{t-1}, x^t, y^{t-1}) \triangleq \mathbf{P}(dz_t | x_t), \quad t \in \mathbb{N}_0^n. \quad (12)$$

At $t = 0$ we have $\mathbf{P}(dz_0 | x_0)$. Also, by Bayes' rule we obtain $\bar{\mathbf{P}}(dz^n | x^n) \triangleq \otimes_{t=0}^n \mathbf{P}(dz_t | x_t)$.

Clearly, from (11) and (12), we can define the joint distribution of $\{(\mathbf{x}_t, \mathbf{z}_t) : t \in \mathbb{N}_0^n\}$ conditioned on $\mathcal{Y}^{t-1}$ by

$$\begin{aligned} \mathbf{P}(dx^n, dz^n | y^{n-1}) &= \otimes_{t=0}^n \mathbf{P}(dx_t, dz_t | x^{t-1}, z^{t-1}, y^{t-1}) \\ &= \otimes_{t=0}^n \mathbf{P}(dx_t | x_{t-1}) \otimes \mathbf{P}(dz_t | x_t). \end{aligned} \quad (13)$$

In addition, from (13), we can define the $\mathcal{Z}^n$ -marginal distribution parametrized by $\mathcal{Y}^{n-1}$ as follows

$$\bar{\mathbf{P}}(dz^n | y^{n-1}) \triangleq \otimes_{t=0}^n \mathbf{P}(dz_t | z^{t-1}, y^{t-1}), \quad (14)$$

where $\mathbf{P}(dz_t|z^{t-1}, y^{t-1}) = \int_{\mathcal{X}_t} \mathbf{P}(dz_t|x_t) \otimes \mathbf{P}(dx_t|z^{t-1}, y^{t-1})$. We assume that at $t = 0$, $\mathbf{P}(dz_0|z^{-1}, y^{-1}) = \mathbf{P}(dz_0)$.

Reproduction or “test-channel”. The reproduction conditional distributions, known as test-channels, satisfy conditional independence

$$\mathbf{P}(dy_t|y^{t-1}, x^t, z^t) = \mathbf{P}(dy_t|y^{t-1}, z^t), \quad t \in \mathbb{N}_0^n. \quad (15)$$

At $t = 0$, no initial state information is assumed, hence $\mathbf{P}(dy_0|y^{-1}, z^0) = \mathbf{P}(dy_0|z_0)$. The conditional distributions $\{\mathbf{P}(dy_t|y^{t-1}, z^t) : t \in \mathbb{N}_0\}$ in (15), uniquely define the family of conditional distributions on $\mathcal{Y}^n$ parametrized by $z^n \in \mathcal{Z}^n$, given by

$$\vec{\mathbf{Q}}(dy^n|z^n) \triangleq \otimes_{t=0}^n \mathbf{P}(dy_t|y^{t-1}, z^t),$$

and vice-versa. From (14) and (15), we can uniquely define the joint distribution of $\{(\mathbf{z}_t, \mathbf{y}_t) : t \in \mathbb{N}_0^n\}$ by

$$\mathbf{P}(dy^n, dz^n) = \vec{\mathbf{P}}(dz^n|y^{n-1}) \otimes \vec{\mathbf{Q}}(dy^n|z^n). \quad (16)$$

In addition, from (16), we can define the $\mathcal{Y}^n$ -marginal distribution parametrized by $\mathcal{Z}^{n-1}$ as follows

$$\vec{\mathbf{P}}(dy^n||z^{n-1}) \triangleq \otimes_{t=0}^n \mathbf{P}(dy_t|y^{t-1}, z^{t-1}), \quad (17)$$

where

$$\mathbf{P}(dy_t|y^{t-1}, z^{t-1}) = \int_{\mathcal{Z}_t} \mathbf{P}(dy_t|y^{t-1}, z^t) \otimes \mathbf{P}(dz_t|z^{t-1}, y^{t-1}). \quad (18)$$

Given the above construction of distributions we obtain the following variant of causally conditioned directed information [17]:

$$\begin{aligned} I(\mathbf{z}^n \rightarrow \mathbf{y}^n || \mathbf{z}^{n-1}) &\stackrel{(a)}{=} \sum_{t=0}^n \mathbb{E} \left\{ \log \left(\frac{d\mathbf{P}(\cdot | \mathbf{y}^{t-1}, \mathbf{z}^t)}{d\mathbf{P}(\cdot | \mathbf{y}^{t-1}, \mathbf{z}^{t-1})}(\mathbf{y}_t) \right) \right\} \\ &\stackrel{(b)}{=} \sum_{t=0}^n I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}), \end{aligned} \quad (19)$$

where (a) is due to chain rule of relative entropy using the Radon-Nykodym derivative; (b) follows by definition.

Next, we formally define the non-trivial generalization of the sequential RDF defined in [2].

Definition 1: (Sequential RDF for partially observable Markov systems) For a given hidden and observation processes that induce (11) and (12), the following variant of sequential RDF subject to an average total MSE distortion constraint both in finite time and in the asymptotic limit, can be defined as follows:

$$R_{[0,n],\text{in}}(D) \triangleq \inf_{\substack{\mathbf{P}(dy_t|y^{t-1}, z^t): t \in \mathbb{N}_0^n \\ \frac{1}{n+1} \mathbb{E}\{d(\mathbf{x}^n, \mathbf{y}^n)\} \leq D}} I(\mathbf{z}^n \rightarrow \mathbf{y}^n || \mathbf{z}^{n-1}), \quad (20)$$

$$R_{\text{in}}(D) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n+1} R_{[0,n],\text{in}}(D). \quad (21)$$

provided the limit takes a finite value, whereas $d(x^n, y^n) \triangleq \sum_{t=0}^n \{||\mathbf{x}_t - \mathbf{y}_t||^2\}$.

Next, we state some useful properties of the Definition 1 that can be extracted from known results.

Remark 3: (Comments on (20)) (1) Similar to what we have already discussed in Remark 2, (1), it can be shown that the objective function in (20) can be replaced by causally conditioned mutual information.

(2) It is easy to show that (20) is convex with respect to the test channel following for instance [27].

(3) If the joint process $\{(\mathbf{x}_t, \mathbf{z}_t) : t \in \mathbb{N}_0\}$ is jointly Gaussian and, $\{(\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t) : t \in \mathbb{N}_0\}$ is also jointly Gaussian, then, (20) achieves a smaller value (see, e.g., [28, Theorem 1]).

(4) The implicit expressions of the optimal minimizing distribution $\mathbf{P}^*(dy_t|y^{t-1}, z^t)$ obtained backward in time, can be found using dynamic programming [29]. The solution will extend the result of fully observable processes derived in [11, Theorem 4.1] to partially observable processes. Although the information measure introduced in Definition 1 can be quit handy in addressing problems related to partially observable Markov decision processes [30], we will not pursue that goal in this paper.

A. Comparison of Definition 1 to other information measures

Recall that the problem statement of §II was also studied in [14] (see also [15]), [1, §II.B] for jointly Gaussian processes. In both works, the authors considered the following information measure to compute a lower bound on (3):

$$\bar{R}_{[0,n],\text{in}}(D) \triangleq \inf_{\substack{\mathbf{P}(dy_t|y^{t-1}, z^t): t \in \mathbb{N}_0^n \\ \frac{1}{n+1} \mathbb{E}\{d(\mathbf{x}^n, \mathbf{y}^n)\} \leq D}} I(\mathbf{z}^n \rightarrow \mathbf{y}^n), \quad (22)$$

$$\bar{R}_{\text{in}}(D) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n+1} \bar{R}_{[0,n],\text{in}}(D). \quad (23)$$

Unfortunately, the utility of (22), (23) increases the difficulty for a characterization that yields analytical tractability because the pay-off in (22) admits a multi-letter expression with respect to the observation process $\{\mathbf{z}_t : t \in \mathbb{N}_0^n\}$ and one needs to transform this problem to a similar one that admits a single-letter characterization of the observed signals in order to compute it. For Gaussian processes this can be done optimally (with an increased computational complexity) if we transform the partially observable Gauss-Markov process to a fully observable one using a pre-KF algorithm from the system's process to the observations process that estimates the hidden state of the Gauss-Markov process. The estimate of such KF serves as a sufficient statistic of the partially observable process (see, e.g., [14]).

The next result is immediate from Lemma 1.

Proposition 1: (Comparison to similar bounds) For the system model in (1), (2) the following bounds hold:

$$R_{[0,n],\text{in}}(D) \leq \bar{R}_{[0,n],\text{in}}(D), \quad R_{\text{in}}(D) \leq \bar{R}_{\text{in}}(D). \quad (24)$$

Proposition 1 is fundamental as it shows that our sequential RDF defined in Definition 1 is a lower bound on the description utilized in both [1], [14], [15] to compute the communication cost for partially observable controlled Markov system driven by additive Gaussian noise process.

IV. NEW RESULTS FOR JOINTLY GAUSSIAN PROCESSES

In this section, we use Definition 1 to derive a general finite-dimensional characterization for multivariate partially

observable Gaussian processes with MSE distortion constraint. To solve the problem in the asymptotic limit, we assume that all involved matrices commute by pairs (sufficient conditions for this assumption to hold are also provided). Then, the corresponding optimization problem becomes convex and can be solved optimally using KKT conditions [18]. The solution of KKT conditions reveals a new non-trivial reverse-waterfilling algorithm.

First, we need the following helpful lemma which is a non-trivial generalization of classical KF algorithm [31], [32] and a generalization of a recent result in [11].

Lemma 2: (Realization of $\{\mathbf{P}^*(dy_t|y^{t-1}, z^t) : t \in \mathbb{N}_0^n\}$) Suppose that the joint process $\{\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t\} : t \in \mathbb{N}_0^n\}$ is jointly Gaussian. Then, the following statements hold.

(1) Any $\{\mathbf{P}^*(dy_t|y^{t-1}, z^t) : t \in \mathbb{N}_0^n\}$ is realized by

$$\mathbf{y}_t = H_t(\mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}) + \hat{\mathbf{x}}_{t|t-1} + \mathbf{v}_t, \quad t \in \mathbb{N}_0^n, \quad (25)$$

where $\hat{\mathbf{z}}_{t|t-1} \triangleq \mathbb{E}\{\mathbf{z}_t|\mathbf{y}^{t-1}, \mathbf{z}^{t-1}\}$, $\hat{\mathbf{x}}_{t|t-1} \triangleq \mathbb{E}\{\mathbf{x}_t|\mathbf{y}^{t-1}, \mathbf{z}^{t-1}\}$, $\{\mathbf{v}_t \in \mathbb{R}^p \sim \mathcal{N}(0; \Sigma_{\mathbf{v}_t}) : t \in \mathbb{N}_0^n\}$ is an independent Gaussian process independent of $\{\{\mathbf{w}_t, \mathbf{n}_t\} : t \in \mathbb{N}_0^n\}$ and $\mathbf{x}_0$, and $\{H_t \in \mathbb{R}^{p \times m} : t \in \mathbb{N}_0^n\}$ are time-varying deterministic matrices (to be designed). Moreover, the innovations process $\{\mathbf{I}_t \in \mathbb{R}^p : t \in \mathbb{N}_0^n\}$ of (25) is the orthogonal process given by

$$\mathbf{I}_t = \mathbf{y}_t - \mathbb{E}\{\mathbf{y}_t|\mathbf{y}^{t-1}, \mathbf{z}^{t-1}\} = H_t(\mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}) + \mathbf{v}_t, \quad (26)$$

where $\mathbf{I}_t \sim \mathcal{N}(0; \Sigma_{\mathbf{I}_t})$, $\Sigma_{\mathbf{I}_t} = H_t C \Sigma_{t|t-1} (H_t C)^T + H_t \Sigma_{\mathbf{n}} H_t^T + \Sigma_{\mathbf{v}_t}$ with $\Sigma_{t|t-1} \triangleq \mathbb{E}\{(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1})(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1})^T | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}\}$. (2) Let $\hat{\mathbf{x}}_{t|t} \triangleq \mathbb{E}\{\mathbf{x}_t | \mathbf{z}^t, \mathbf{y}^t\}$, $\Sigma_{t|t} \triangleq \mathbb{E}\{(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t})(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t})^T | \mathbf{y}^t, \mathbf{z}^t\}$. Then, $\{\hat{\mathbf{x}}_{t|t-1}, \Sigma_{t|t-1} : t \in \mathbb{N}_0^n\}$ satisfy the following generalized discrete-time KF recursions:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t} &= \hat{\mathbf{x}}_{t|t-1} + \mathbf{k}_t \mathbf{I}_t, \\ \hat{\mathbf{x}}_{t|t-1} &= A_t \hat{\mathbf{x}}_{t-1|t-1}, \\ \Sigma_{t|t-1} &= A_t \Sigma_{t-1|t-1} A_t^T + \Sigma_{\mathbf{w}_t}, \\ \mathbf{k}_t &= \Sigma_{t|t-1} (H_t C)^T \Sigma_{\mathbf{I}_t}^{-1} \text{ (Kalman Gain)}, \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - \Sigma_{t|t-1} (H_t C)^T \Sigma_{\mathbf{I}_t}^{-1} H_t C \Sigma_{t|t-1}, \end{aligned} \quad (27)$$

where $\Sigma_{t|t} = \Sigma_{t|t}^T \succeq 0$ and $\Sigma_{t|t-1} = \Sigma_{t|t-1}^T \succ 0$.

(3) The characterization of $R_{[0,n],\text{in}}^{\text{seq},G}(D)$ that achieves (25) is given by

$$R_{[0,n],\text{in}}^G(D) = \inf_{\substack{H_t \in \mathbb{R}^{p \times m}, \Sigma_{\mathbf{v}_t} \succeq 0 \\ \Sigma_{t|t} \succeq 0, \Sigma_{t|t-1} \succ 0 \\ \frac{1}{n+1} \sum_{t=0}^n \text{trace}(\mathcal{G}) \leq D}} \frac{1}{2} \sum_{t=0}^n \left[\log \frac{|\Sigma_{\mathbf{I}_t}|}{|\Sigma_{\mathbf{v}_t}|} \right]^+, \quad (28)$$

where

$$\mathcal{G} = (I_p - H_t C) \Sigma_{t|t-1} (I_p - H_t C)^T + H_t \Sigma_{\mathbf{n}} H_t^T + \Sigma_{\mathbf{v}_t},$$

for some $D \in [0, \infty]$.

Proof: (1) Since the joint process is assumed to be jointly Gaussian, then, $\{\mathbf{P}^*(dy_t|y^{t-1}, z^t) : t \in \mathbb{N}_0^n\}$ is conditionally Gaussian, and we can obtain the orthogonal realization

$$\mathbf{y}_t = H_t \mathbf{z}_t + R_t(\mathbf{z}^{t-1}, \mathbf{y}^{t-1}) + \mathbf{v}_t, \quad t \in \mathbb{N}_0^n, \quad (29)$$

where $R_t(\mathbf{z}^{t-1}, \mathbf{y}^{t-1}) \triangleq G_{t-1} \mathbf{z}^{t-1} + \Gamma_{t-1} \mathbf{y}^{t-1}$, $\mathbf{P}^*(\cdot | y^{t-1}, z^t) \sim \mathcal{N}(H_t \mathbf{z}_t + G_{t-1} \mathbf{z}^{t-1} + \Gamma_{t-1} \mathbf{y}^{t-1}; \Sigma_{\mathbf{v}_t})$, with $\{H_t : t \in \mathbb{N}_0^n\}$ being deterministic matrices and (G_{t-1}, Γ_{t-1}) deterministic matrices of appropriate dimensions. For such realization, $I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1})$ does not depend on $R_t(\cdot, \cdot)$, $\forall t \in \mathbb{N}_0^n$. Moreover,

$$\begin{aligned} \mathbb{E}\{\|\mathbf{x}_t - \mathbf{y}_t\|^2\} &= \mathbb{E}\{\|(I_p - H_t C)\mathbf{x}_t - R_t(\mathbf{z}^{t-1}, \mathbf{y}^{t-1})\|^2\} \\ &\quad + \text{trace}(H_t \Sigma_{\mathbf{n}} H_t^T + \Sigma_{\mathbf{v}_t}) \\ &\stackrel{(*)}{\geq} \mathbb{E}\{\|(I_p - H_t C)\mathbf{x}_t - R_t^*(\mathbf{z}^{t-1}, \mathbf{y}^{t-1})\|^2\} \\ &\quad + \text{trace}(H_t \Sigma_{\mathbf{n}} H_t^T + \Sigma_{\mathbf{v}_t}), \end{aligned}$$

where $(*)$ holds with equality if $R_t(\cdot, \cdot) = R_t^*(\cdot, \cdot) = (I_p - H_t C)\hat{\mathbf{x}}_{t|t-1}$, $\forall t \in \mathbb{N}_0^n$. (2) This follows from the discrete-time KF equations. (3) The characterization that achieves (25) is obtained from (1), (2) as follows. First note that by definition, we have

$$I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) = h^G(\mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) - h^G(\mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^t). \quad (30)$$

The first term in (30) is computed as follows:

$$h^G(\mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1}) \stackrel{(a)}{=} \frac{1}{2} \log(2\pi e)^p |\Sigma_{\mathbf{I}_t}|, \quad (31)$$

where (a) follows from the fact that $\mathbf{P}(dy_t | y^{t-1}, z^{t-1}) \sim \mathcal{N}(H_t(\mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}) + \mathbf{v}_t; \Sigma_{\mathbf{I}_t})$. The second term in (30) is computed as follows:

$$h^G(\mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^t) \stackrel{(b)}{=} \frac{1}{2} \log(2\pi e)^p |\Sigma_{\mathbf{v}_t}|, \quad (32)$$

where (b) stems from the fact that $\mathbf{P}(dy_t | y^{t-1}, z^t) \sim \mathcal{N}(H_t(\mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1}) + \hat{\mathbf{x}}_{t|t-1}; \Sigma_{\mathbf{v}_t})$. Incorporating both (31), (32) in (30), we obtain the objective in (28). Finally, the MSE distortion constraint follows from (1). This completes the proof. ■

The next theorem gives the general finite dimensional characterization of (20) including the feasible set of solutions for time-varying partially observable Gaussian processes with average total MSE distortion constraint. It also reveals the optimal linear Gaussian test-channel distribution (forward test-channel realization) that corresponds to this problem. The derivation relies on the identification of the optimization variables $(H_t, \Sigma_{\mathbf{v}_t})$ of (25).

Theorem 2: (Characterization of (20) for jointly Gaussian processes) Let $\Lambda_t = \Sigma_{t|t-1}$ and $\Delta_t = \Sigma_{t|t}$. Then, the characterization of (20) for the system (1), (2) with total MSE distortion constraint is the following

$$\begin{aligned} R_{[0,n],\text{in}}^G(D) &= \inf \frac{1}{2} \sum_{t=0}^n \left[\log \frac{|\Lambda_t|}{|\Delta_t|} + \log \frac{|\Lambda_t|}{|\Lambda_t + Q - \Lambda_t \Delta_t^{-1} Q|} \right], \\ \text{s.t. } 0 &\prec \Lambda_t (\Lambda_t + Q)^{-1} Q \prec \Delta_t \\ 0 &\prec \Delta_t \preceq \Lambda_t \\ \frac{1}{n+1} \sum_{t=0}^n \text{trace}(\Delta_t) &\leq D \end{aligned} \quad (33)$$

for some $D \in [D_{\min}, D_{\max}] \subset (0, D_{\max}]$, where $Q \triangleq C^\dagger \Sigma_n C^{\dagger\top} \succeq 0$ such that $\text{rank}(Q) = \text{rank}(\Sigma_n)$. Moreover, the above characterization, is achieved by a linear Gaussian “test channel” $\mathbf{P}^*(dy_t|y_{t-1}, z_t)$ of the form

$$\mathbf{y}_t = H_t \mathbf{z}_t + (I_p - H_t C) A \mathbf{y}_{t-1} + \mathbf{v}_t, \quad t \in \mathbb{N}_0^n, \quad (34)$$

with $\mathbf{y}_{-1} = 0$, and

$$\begin{aligned} H_t C &\triangleq I_p - \Delta_t \Lambda_t^{-1}, \quad H_t = C^\dagger - \Delta_t \Lambda_t^{-1} C^\dagger \\ C^\dagger &= C^\top (C C^\top)^{-1}, \quad \Sigma_{\mathbf{v}_t} \triangleq \Delta_t (H_t C)^\top - H_t \Sigma_n H_t^\top \succeq 0, \end{aligned} \quad (35)$$

where $C^\dagger$ is the generalized inverse matrix of C .⁴

Proof: From MSE estimation theory we know that the MSE inequality $\sum_{t=0}^n \mathbb{E} \{ \|\mathbf{x}_t - \mathbf{y}_t\|^2 \} \geq \sum_{t=0}^n \mathbb{E} \{ \|\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\|^2 \}$ holds for all $(H_t, \Sigma_{\mathbf{v}_t})$, $t \in \mathbb{N}_0^n$, and it is achieved if and only if $\hat{\mathbf{x}}_{t|t} = \mathbf{y}_t$. Sufficient conditions for the latter to hold are (i) $\hat{\mathbf{x}}_{t|t-1} \equiv \mathbb{E} \{ \mathbf{y}_t | \mathbf{y}^{t-1}, \mathbf{z}^{t-1} \}$ and (ii) $\mathbf{k}_t = I_p$. Note that (i) holds by the general KF algorithm in Lemma 2. The choice of (35) satisfies (ii) hence a smaller distortion for a given rate is achieved and also the Markov realization in (34) holds. Moreover, by substituting in the pay-off of (28) the scalings of (35) (without $\frac{1}{2} \sum_{t=0}^n$) and after some matrix algebra, we obtain

$$\begin{aligned} &\left[\log \frac{|\Lambda_t|}{|\Delta_t|} \right. \\ &\quad \left. + \log \frac{|I_p - \Lambda_t^{-1} \Delta_t|}{|I_p - \Lambda_t^{-1} \Delta_t - (\Delta_t^{-1} - \Lambda_t^{-1}) Q (I_p - \Lambda_t^{-1} \Delta_t)|} \right]^+ \end{aligned} \quad (36)$$

where we have set $Q \triangleq C^\dagger \Sigma_n C^{\dagger\top}$ which is a positive semidefinite matrix of rank dictated by the rank of Σ_n . Observe that (36) can be reformulated as follows

$$(36) \stackrel{(a)}{=} \left[\log \frac{|\Lambda_t|}{|\Delta_t|} + \log \frac{|I_p|}{|I_p - (\Delta_t^{-1} - \Lambda_t^{-1}) Q|} \right]^+ \quad (37)$$

$$\stackrel{(b)}{=} \left[\log \frac{|\Lambda_t|}{|\Delta_t|} + \log \frac{|\Lambda_t|}{|\Lambda + Q - \Lambda \Delta^{-1} Q|} \right]^+, \quad (38)$$

where (a) follows by dividing both the numerator and denominator of the second term in (37) by $I_p - \Lambda_t^{-1} \Delta_t$; (b) follows by multiplying both the numerator and denominator of the second term in (38) by Λ_t .

We ensure that both terms in (38) are well defined if the linear matrix inequality $\Delta_t \preceq \Lambda_t$, $\forall t$ holds and additionally for the second term only if the non-linear matrix inequality $0 \prec \Lambda_t (\Lambda_t + Q)^{-1} Q \prec \Delta_t$, $\forall t$ is imposed (otherwise that term may give a matrix with non-positive real eigenvalues, hence the solution will not be well-defined). Additionally, a sufficient condition for existence of a finite solution to both terms in (38) is to restrict $\Delta_t \succ 0$. In fact that condition can be made more rigorous upon observing that $\Lambda_t (\Lambda_t + Q)^{-1} Q \prec \Delta_t$, $\forall t$. The previous analysis is described by the pair of linear and non-linear matrix inequalities in (33). ■

In what follows we point out some interesting technical observations that stem from the result of Theorem 2.

Remark 4: (Technical comments on Theorem 2) (1) The linear and non-linear matrix inequality constraints in (33) can be compactly written as $0 \prec \Lambda_t (\Lambda_t + Q)^{-1} Q \prec \Delta_t \preceq \Lambda_t$, $\forall t$. (2) Theorem 2 shows that the optimal minimizer in (20) for jointly Gaussian processes with MSE distortion admits a Markov realization, i.e., $\mathbf{P}^*(dy_t|y^{t-1}, z^t) = \mathbf{P}^*(dy_t|y_{t-1}, z_t)$. This implies that the output of the decoder $\{\mathbf{y}_t : t \in \mathbb{N}_0\}$ is modeled as a first-order Markov process and the corresponding rate-distortion characterization (20) is simplified to the following finite state rate distortion expression:

$$R_{[0,n],\text{in}}^G(D) \triangleq \inf_{\substack{\mathbf{P}(dy_t|y_{t-1}, z_t): t \in \mathbb{N}_0^n \\ \frac{1}{n+1} \mathbb{E} \{ \|\mathbf{x}_t - \mathbf{y}_t\|^2 \} \leq D}} \sum_{t=0}^n I(\mathbf{z}_t; \mathbf{y}_t | \mathbf{y}_{t-1}). \quad (39)$$

The description in (39) demonstrates that the additional information via the previous observation signals $\mathbf{z}^{t-1}$ at the decoder is surprisingly redundant. Moreover, (39) corresponds to the best linear coding policies as long as the additive noise processes are zero-mean, uncorrelated and white (this conditions are satisfied in our setup), because then, the KF algorithm becomes the best linear MMSE estimator, see, e.g., [31, §3.2] or [32, p. 130].

(3) Clearly, the result of Theorem 2 can be trivially reformulated to the case of pointwise MSE distortion, i.e., $\text{trace}(\Delta_t) \leq D_t$, $\forall t$.

(4) The first term in the objective function of (33) is precisely the objective of the “fully observable” Gauss-Markov process with MSE distortion constraint, see, e.g., [10], [11], [5]. Interestingly, if we assume in the second term of the objective function that $Q = 0$ (null matrix) which means that $\Sigma_n = 0$, then, the problem recovers the known expression for fully observable Gauss-Markov processes.

Next, we restrict Theorem 2 to its time-invariant characterization.

Theorem 3: (Time-invariant characterization of (33)) Suppose that $D \in [D_{\min}, D_{\max}] \subset (0, D_{\max}]$. Moreover, let the conditionally Gaussian distribution $\mathbf{P}^*(dy_t|y_{t-1}, z_t)$ to be time-invariant and the corresponding distribution $\mathbf{P}^*(dy_t|y_{t-1})$ to have a unique invariant distribution. Then, if $R_{\text{in}}^G(D) < \infty$, its characterization is given by

$$\begin{aligned} R_{\text{in}}^G(D) &= \inf \frac{1}{2} \left[\log \frac{|\Lambda|}{|\Delta|} + \log \frac{|\Lambda|}{|\Lambda + Q - \Lambda \Delta^{-1} Q|} \right], \\ \text{s.t. } &0 \prec \Lambda (\Lambda + Q)^{-1} Q \prec \Delta \\ &0 \prec \Delta \preceq \Lambda \\ &\text{trace}(\Delta) \leq D \end{aligned} \quad (40)$$

where (Λ, Δ) are the time-invariant values of (Λ_t, Δ_t) , respectively. Moreover, (40) is achieved by a time-invariant realization of the form

$$\mathbf{y}_t = H \mathbf{z}_t + (I_p - H C) A \mathbf{y}_{t-1} + \mathbf{v}_t, \quad (41)$$

and

$$\begin{aligned} H C &\triangleq I_p - \Delta \Lambda^{-1} \succeq 0, \quad H = C^\dagger - \Delta \Lambda^{-1} C^\dagger \\ C^\dagger &= C^\top (C C^\top)^{-1}, \quad \Sigma_{\mathbf{v}} \triangleq \Delta (H C)^\top - H \Sigma_n H^\top \succeq 0. \end{aligned} \quad (42)$$

Proof: From the sub-additivity property of $R_{[0,n],\text{in}}^G(D)$ in Definition 1, the limit $R_{\text{in}}^G(D)$ always exists although it

⁴See the supplementary material in Appendix A (Definition 5 and Lemma 4) for the definition of matrix $C \in \mathbb{R}^{m \times p}$.

may be infinite. If $R_{\text{in}}^G(D) = +\infty$ there is nothing to prove. However, by restricting the distributions $\mathbf{P}^*(dy_t|y_{t-1}, z_t)$ and $\mathbf{P}^*(dy_t|y_{t-1})$ to be time-invariant, and if $R_{\text{in}}^G(D) < \infty$, then, its solution is given by (40). (41) follows because we have assumed time-invariant $\mathbf{P}^*(dy_t|y_{t-1}, z_t)$. ■

Clearly, Theorems 2, 3 provide not only the characterization of the optimization problem that corresponds to the system (1), (2) driven by additive i.i.d. Gaussian noise process with MSE distortion, but also the feasible set of solutions that ensure a finite value for the corresponding optimization problem. Unfortunately, in general, the characterization in Theorem 3 (see also Theorem 2) forms a non-convex optimization problem (because we cannot ensure that the product of the specific positive semidefinite matrices will preserve a symmetric (positive-semidefinite) structure (i.e., the set of involved matrices does not belong in the positive semidefinite cone [18])). This means that without certain sufficient conditions which will guarantee that the problem is convex, the solution will not likely achieve its global minimum value [18]. In the next theorem, we give sufficient conditions to ensure that the optimization problem of Theorem 3 when solved achieves its minimum value (i.e., it is convex).

Theorem 4: (A convex characterization of (20) for time-invariant partially observable Gaussian processes with MSE distortion) Suppose that the square matrices (Λ, Δ, Q) commute by pairs.⁵ Then, the characterization of (33) is simplified to the following convex optimization problem:

$$\begin{aligned} R_{\text{in}}^G(D) = \inf \frac{1}{2} \sum_{i=1}^p & \left[\log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Delta,i}} \right) + \right. \\ & \left. + \log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Lambda,i} + \mu_{Q,i} - \mu_{\Lambda,i}\mu_{\Delta^{-1},i}\mu_{Q,i}} \right) \right], \quad (43) \\ \text{s.t.} \quad & 0 < \frac{\mu_{\Lambda,i}\mu_{Q,i}}{\mu_{\Lambda,i} + \mu_{Q,i}} < \mu_{\Delta,i}, \quad \forall i \\ & 0 < \mu_{\Delta,i} \leq \mu_{\Lambda,i}, \quad \forall i \\ & \sum_{i=1}^p \mu_{\Delta,i} \leq D \end{aligned}$$

where $\mu_{\Lambda,i} = \mu_{A^2,i}\mu_{\Delta,i} + \mu_{\Sigma_w,i}$, for some $D \in [D_{\min}, D_{\max}] \subset (0, D_{\max}]$.

Proof: Note that by assumption of the theorem, i.e., (Λ, Δ, Q) commute by pairs, then, they are also simultaneously diagonalizable by an orthogonal matrix $U \in \mathbb{R}^{p \times p}$.⁶ It is simple to show that by writing the spectral representation (eigenvalue decomposition) of each of the above matrices and performing simple matrix algebra, we can simplify the complex structure of (40) to the one in (43). Clearly, the objective as a function of the variable $\mu_{\Delta,i} > \frac{\mu_{\Lambda,i}\mu_{Q,i}}{\mu_{\Lambda,i} + \mu_{Q,i}}, \forall i$ is differentiable and continuous in its domain. Moreover, it can be shown that its second derivative (with respect to $\mu_{\Delta,i}$) is non-negative hence the objective is convex with respect to $\mu_{\Delta,i}$'s. ■

⁵For details on this concept, see the supplementary material in Appendix A, Definitions 2, 3.

⁶For details see the supplementary material in Appendix A, Theorem 8.

Next, we derive sufficient conditions that meet the assumption of Theorem 4. This requires identification of the structure of the square matrices (A, Q, Σ_w) .

Proposition 2: (Sufficient conditions for optimality of Theorem 4) Suppose that (A, Q, Σ_w) satisfy one of the following matrix structures:

- (1) A is symmetric, $\Sigma_w = \sigma_w^2 I_p$ (scalar matrix), where $\sigma_w^2 > 0$, and $Q = qI_p$ (scalar matrix), where $q \geq 0$.
 - (2) $A = \alpha I_p$ (scalar matrix), $\Sigma_w = \sigma_w^2 I_p$ (scalar square matrix), where $\sigma_w^2 > 0$, and $Q \succeq 0$.
 - (3) $A = \alpha I_p$ (scalar matrix), $\Sigma_w \succ 0$, and $Q = qI_p$ (scalar square matrix) where $q \geq 0$.
 - (4) All (A, Σ_w, Q) have only diagonal elements with $\Sigma_w \succ 0$ and $Q \succeq 0$;
 - (5) (A, Σ_w, Q) have precisely the same matrix structure.
- Then, (A, Q, Σ_w, Δ) , commute by pairs and consequently Δ and Λ commute.

Proof: We only prove (1) because the other cases follow similarly. Under the matrix structure of (1), (A, Σ_w, Q) commute by pairs, because the last two matrices are scalar matrices that commute with every other matrix of the same dimensions. This means that all (A, Σ_w, Q) are simultaneously diagonalizable by an orthogonal matrix U . Clearly, (Σ_w, Q) commute with Δ because they are scalar matrices. It remains to show that A commutes with Δ . We write the spectral representation of $A = U \text{diag}(\{\mu_{A,i}\}_{i=1}^p) U^T$ and $\Delta = V \text{diag}(\{\mu_{\Delta,i}\}_{i=1}^p) V^T$. The two matrices commute if $U \text{diag}(\{\mu_{A,i}\}_{i=1}^p) U^T V \text{diag}(\{\mu_{\Delta,i}\}_{i=1}^p) V^T = V \text{diag}(\{\mu_{\Delta,i}\}_{i=1}^p) V^T U \text{diag}(\{\mu_{A,i}\}_{i=1}^p) U^T$. But, Δ is a design matrix (variable) therefore we can always choose the design parameter $V = U$ (i.e, eigenvectors alignment) ensuring that (A, Δ) commute. Finally since (A, Σ_w, Q, Δ) commute by pairs then Λ and Δ commute as well. Cases (2), (3), (4), (5) follow similarly. ■

Remark 5: (Technical comments) Although the sufficient conditions derived in Proposition 2 are restrictive compared to the general available structure of matrices (A, Q, Σ_w) , they maintain optimality of the convex optimization problem in Theorem 4 and achieve the minimum possible rates. Unfortunately, if one deviates from such conditions and provided that the chosen structure of matrices (A, Q, Σ_w) will not yield commuting by pairs and simultaneous diagonalization, then, Theorem 4 will be, in general, a sub-optimal solution to the general characterization of Theorem 3. The tightness of the sub-optimal solution depends on the dimensionality of the problem and the corresponding matrices structure.

The next result is a main result of this paper.

Theorem 5: (Reverse-waterfilling solution of (43)) The parametric solution of (43) is

$$\begin{aligned} R_{\text{in}}^G(D) = \frac{1}{2} \sum_{i=1}^p & \left[\log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Delta,i}} \right) + \right. \\ & \left. + \log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Lambda,i} + \mu_{Q,i} - \mu_{\Lambda,i}\mu_{\Delta^{-1},i}\mu_{Q,i}} \right) \right], \quad (44) \end{aligned}$$

such that $\mu_{\Lambda,i} = \mu_{A^2,i}\mu_{\Delta,i} + \mu_{\Sigma_w,i}$, $\forall i$, and $\mu_{\Delta,i}$ is computed

based on the following reverse-waterfilling algorithm

$$\mu_{\Delta,i} = \begin{cases} \mu_{\Delta^*,i} & \text{if } \mu_{\Delta^*,i} < \mu_{\Delta^*,i} < \mu_{\Lambda,i}, \forall i, \\ \mu_{\Lambda,i} & \text{if } \mu_{\Delta^*,i} \geq \mu_{\Lambda,i} \end{cases}, \quad (45)$$

with $\sum_{i=1}^p \mu_{\Delta,i} = D$, and $D > D_{\min} = \sum_{i=1}^p \mu_{\Delta^*,i}$ where

$$\mu_{\Delta^*,i} \triangleq \frac{\sqrt{v^2 + 4\mu_{A^2,i}\mu_{\Sigma_{\mathbf{w}},i}\mu_{Q,i}} - v}{2\mu_{A^2,i}}, \quad \mu_{A,i} \neq 0, \quad \forall i, \quad (46)$$

with $v \triangleq \mu_{\Sigma_{\mathbf{w}},i} + (1 - \mu_{A^2,i})\mu_{Q,i}$, $\mu_{Q,i} \neq \infty$, and $\mu_{\Delta^*,i} > \mu_{\Delta^*,i}$ is the solution that achieves the highest rates obtained from the third degree polynomial equation

$$C1\mu_{\Delta^*,i}^3 + C2\mu_{\Delta^*,i}^2 + C3\mu_{\Delta^*,i} - C4 = 0, \quad (47)$$

where

$$\begin{aligned} C1 &\triangleq 2\mu_{A^4,i}\theta \\ C2 &\triangleq 2\mu_{A^2,i}\theta(v + \mu_{\Sigma_{\mathbf{w}},i}) \\ C3 &\triangleq \mu_{A^2,i}(v - 2\mu_{\Sigma_{\mathbf{w}},i}) + 2\theta\mu_{\Sigma_{\mathbf{w}},i}(v - \mu_{A^2,i}\mu_{Q,i}) \\ C4 &\triangleq \mu_{\Sigma_{\mathbf{w}},i} [2\theta\mu_{\Sigma_{\mathbf{w}},i}\mu_{Q,i} + \mu_{A^2,i}\mu_{Q,i} + \mu_{\Sigma_{\mathbf{w}},i} + \mu_{Q,i}] \end{aligned} \quad (48)$$

Proof: The solution is obtained using KKT conditions [18, Chapter 5.5.3]. First, we introduce the augmented (unconstrained) Lagrange functional of (43) as follows:

$$\begin{aligned} J(\{\mu_{\Delta,i}\}_{i=1}^p, \theta, \{f_i^1\}_{i=1}^p, \{f_i^2\}_{i=1}^p) \\ = \frac{1}{2} \sum_{i=1}^p \left[\log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Delta,i}} \right) + \right. \\ \left. + \log \left(\frac{\mu_{\Lambda,i}}{\mu_{\Lambda,i} + \mu_{Q,i} - \mu_{\Lambda,i}\mu_{\Delta^{-1},i}\mu_{Q,i}} \right) \right] \\ + \theta \left(\sum_{i=1}^p \mu_{\Delta,i} - D \right) + \sum_{i=1}^p f_i^1 \left(\frac{\mu_{\Lambda,i}\mu_{Q,i}}{\mu_{\Lambda,i} + \mu_{Q,i}} - \mu_{\Delta,i} \right) \\ + \sum_{i=1}^p f_i^2 (\mu_{\Delta,i} - \mu_{\Lambda,i}), \end{aligned} \quad (49)$$

where $\mu_{\Delta,i} \geq 0$ is the primal variable, $\theta \geq 0, \{(f_i^1 \geq 0, f_i^2 \geq 0)\}, \forall i$, are the dual variables (Lagrange multipliers) responsible for the distortion constraint, the quadratic inequality constraint $\frac{\mu_{\Lambda,i}\mu_{Q,i}}{\mu_{\Lambda,i} + \mu_{Q,i}} < \mu_{\Delta,i}, \forall i$, and the linear inequality constraint $\mu_{\Delta,i} \leq \mu_{\Lambda,i}, \forall i$, respectively. Note that the optimization problem in (43) under the specific constraints is convex (the objective is convex w.r.t. to $\mu_{\Delta,i}$, the quadratic inequality is continuously differentiable convex function under the assumption that $\mu_{\Delta,i} \geq 0$, and the other two inequality constraints are affine). Hence Slater's condition are satisfied and the KKT conditions are also sufficient for global optimal-

ity. The KKT conditions for this problem are as follows:

$$\left. \frac{\partial J(\{\mu_{\Delta,i}\}_{i=1}^p, \theta, \{f_i^1\}_{i=1}^p, \{f_i^2\}_{i=1}^p)}{\partial \mu_{\Delta,i}} \right|_{\mu_{\Delta,i} = \mu_{\Delta^*,i}} = 0, \quad \forall i, \quad (50)$$

$$\theta \left(\sum_{i=1}^p \mu_{\Delta^*,i} - D \right) = 0, \quad f_i^1 \left(\frac{\mu_{\Lambda^*,i}\mu_{Q,i}}{\mu_{\Lambda^*,i} + \mu_{Q,i}} - \mu_{\Delta^*,i} \right) = 0, \quad (51)$$

$$f_i^2 (\mu_{\Delta^*,i} - \mu_{\Lambda^*,i}) = 0, \quad \forall i \quad (52)$$

$$\sum_{i=1}^p \mu_{\Delta^*,i} \leq D, \quad \frac{\mu_{\Lambda^*,i}\mu_{Q,i}}{\mu_{\Lambda^*,i} + \mu_{Q,i}} \leq \mu_{\Delta^*,i}, \quad \mu_{\Delta^*,i} \leq \mu_{\Lambda^*,i} \quad (53)$$

$$f_i^1 \geq 0, \quad f_i^2 \geq 0, \quad \theta \geq 0, \quad (54)$$

where (50) is the stationarity condition; (51), (52) are the complementary slackness conditions; (53) are the primal feasibility conditions and (54) are the dual feasibility conditions.

Note that the objective in (43) has a finite solution if $\frac{\mu_{\Lambda^*,i}\mu_{Q,i}}{\mu_{\Lambda^*,i} + \mu_{Q,i}} < \mu_{\Delta^*,i}$ which implies by (51) that $f_i^1 = f_i^{1,*} = 0, \forall i$, whereas if $f_i^2 > 0$, i.e., $\mu_{\Lambda^*,i} = \mu_{\Delta^*,i}$ by (52), the solution is zero. Hence without loss of optimality we can also take $f_i^2 = f_i^{2,*} = 0, \forall i$.

Next, we proceed to solve (50) (without $\sum_{i=1}^p$).

$$\begin{aligned} \left. \frac{\partial J(\{\mu_{\Delta,i}\}_{i=1}^p, \theta, \{f_i^1\}_{i=1}^p, \{f_i^2\}_{i=1}^p)}{\partial \mu_{\Delta,i}} \right|_{\substack{\mu_{\Delta,i} = \mu_{\Delta^*,i} \\ f_i^1 = f_i^{1,*} \\ f_i^2 = f_i^{2,*}}} \\ = \frac{1}{2} \left[\frac{\mu_{A^2,i}\mu_{\Delta^*,i} - \mu_{\Sigma_{\mathbf{w}},i}}{\mu_{\Delta^*,i}(\mu_{A^2,i}\mu_{\Delta^*,i} + \mu_{\Sigma_{\mathbf{w}},i})} \right. \\ \left. - \frac{\mu_{A^2,i}\mu_{\Delta^*,i} + \mu_{\Sigma_{\mathbf{w}},i}\mu_{Q,i}}{\mu_{\Delta^*,i} \left(\mu_{A^2,i}\mu_{\Delta^*,i} + (1 - \mu_{A^2,i})\mu_{Q,i} - \frac{\mu_{\Sigma_{\mathbf{w}},i}\mu_{Q,i}}{\mu_{\Delta^*,i}} \right)} \right] \\ + \theta = 0 \\ \implies 2\mu_{A^4,i}\theta\mu_{\Delta^*,i} + 2\mu_{A^2,i}\theta(v + \mu_{\Sigma_{\mathbf{w}},i})\mu_{\Delta^*,i} \\ + [\mu_{A^2,i}(v - 2\mu_{\Sigma_{\mathbf{w}},i}) + 2\theta\mu_{\Sigma_{\mathbf{w}},i}(v - \mu_{A^2,i}\mu_{Q,i})] \mu_{\Delta^*,i} \\ - \mu_{\Sigma_{\mathbf{w}},i} [2\theta\mu_{\Sigma_{\mathbf{w}},i}\mu_{Q,i} + \mu_{A^2,i}\mu_{Q,i} + \mu_{\Sigma_{\mathbf{w}},i} + \mu_{Q,i}] = 0, \quad \forall i, \end{aligned} \quad (55)$$

where (55) follows after some algebra and by setting $v \triangleq \mu_{\Sigma_{\mathbf{w}},i} + (1 - \mu_{A^2,i})\mu_{Q,i}$. Hence we obtain (47), (48). From the constraints of the optimization problem, the solution of (55) must satisfy

$$\frac{\mu_{\Lambda^*,i}\mu_{Q,i}}{\mu_{\Lambda^*,i} + \mu_{Q,i}} - \mu_{\Delta^*,i} < 0, \quad (56)$$

which corresponds to a quadratic inequality with two boundary points of which one is negative hence rejected because $\mu_{\Delta^*,i} > 0$ (by definition). The latter also implies that the quadratic inequality in (56) is active only when (46) holds $\forall i$. Hence, the third-degree polynomial equation in (55) should satisfy the minimum distortion criterion of (46) at each i . If this criterion is satisfied by all three solutions, then, we pick the one that achieves higher rates due to Lagrange duality theorem [33]. The problem is solved once we check whether (44) is non-negative via the criterion in (45). This completes the derivation. ■

Next, we point out a technical comment and certain special cases for Theorem 5.

Remark 6: (Comments on Theorem 5) (1) The reverse-waterfilling solution of Theorem 5 relies on the third degree polynomial equation (47). That polynomial equation is known to have one or three real-valued solutions and it can be solved numerically using simple commands in Matlab or in Python. (2) If in (44), we take $\mu_{Q,i} = 0$, for some i , then for that i , it can be easily shown that our solution will recover as a special case, the reverse-waterfilling algorithm obtained for the “fully observable” time-invariant multidimensional Gauss-Markov processes subject to a MSE distortion constraint derived in [24, Proposition 1]. (3) Due to the pay-off in (44) that is greater than the corresponding one obtained for fully observable processes (only the first term appears in that case), the reverse-waterfilling solution of Theorem 5 will always be an upper bound on the reverse-waterfilling obtained for fully observable multivariate processes in [24, Proposition 1] for the rate-distortion region that these are both defined.

In Fig. 2, we illustrate a plot obtained using the iterative method of Algorithm 1. For this simulation, we consider $A = \text{diag}(1.1, 1.1, 1.1) \in \mathbb{R}^{3 \times 3}$, $\Sigma_w = \text{diag}(1, 1, 1) \in \mathbb{R}^{3 \times 3} \succ 0$, $Q = \text{diag}(2.8360, 0.0112, 0) \in \mathbb{R}^{3 \times 3}$. Note that $D_{\min} = 1.38$ whereas $\sum_{\mu_{A,i} > 1} \log_2 \mu_{A,i} = 0.41$ (all dimensions have magnitude greater than one). We note that for this particular example, the solution is optimal because the matrix structure is included in the classes of matrices derived in Proposition 2.

Algorithm 1 Implementation of (45)

Initialize: number of spatial components p ; error tolerance ϵ_1, ϵ_2 where $\epsilon_1 \geq \epsilon_2$; nominal minimum and maximum value $\theta^{\min} = 0$ and $\theta^{\max}$ (sufficiently large); pick an initial variance for $\mu_{\Lambda,1}$; pick the matrix structure of (A, Σ_w, Q) based on (1), (2), and their corresponding eigenvalues $\{(\mu_{A,i}, \mu_{\Sigma_w,i}, \mu_{Q,i}) : i \in \mathbb{N}_1^p\}$; Choose distortion level $D > D_{\min} = \sum_{i=1}^p \mu_{\Delta_{\min}^*}$.
Set $\theta = \theta^{\max}$; flag = 0.

while flag = 0 **do**

 Compute $\mu_{\Delta,i} \forall i$ as follows:

for $i = 1 : p$ **do**

 Compute $\mu_{\Delta^*,i}$ according to (46), (47), (48).

 Compute $\mu_{\Delta,i}$ according to (45).

end for

if $\sum_{i=1}^p \mu_{\Delta,i} - D \geq \epsilon_1$ **then**

 Set $\theta^{\min} = \theta$.

else

 Set $\theta^{\max} = \theta$.

end if

if $\theta^{\max} - \theta^{\min} \geq \epsilon_2$ **then**

 Compute $\theta = \frac{(\theta^{\min} + \theta^{\max})}{2}$.

else

 flag $\leftarrow 1$

end if

end while

Output: $\{\mu_{\Delta,i} : i \in \mathbb{N}_1^p\}$, $\{\mu_{\Lambda,i} : i \in \mathbb{N}_1^p\}$, for a given distortion level D .

Comments on Algorithm 1. Algorithm 1 utilizes the *bisection method* to converge to a certain value for a given error tolerance. This is done by picking nominal values for θ (i.e., $\theta^{\min}, \theta^{\max}$). Finally, the convergence of θ affects the convergence of $\sum_{t=1}^p \mu_{\Delta,i} \rightarrow D$. Our iterative scheme does not guarantee that both $(\theta, \sum_{t=1}^p \mu_{\Delta,i} \rightarrow D)$ will converge concurrently.

Fig. 2. Numerical simulation based on Algorithm 1.

A. Analytical solution for scalar-valued processes

Next, we focus on providing the optimal analytical solution of (40) for scalar-valued Gaussian processes with additional observations noise to elucidate further insights on this problem and to easily compare with known results in the literature. In this case, Theorem 4 gives the optimal solution when solved (i.e., commutativity is inherent in scalar-valued processes). For simplicity, we denote the scalar-values of $(A, \Delta, \Lambda, \Sigma_n, \Sigma_w)$ by $(\alpha, \delta, \lambda, \sigma_n^2, \sigma_w^2)$ instead of their corresponding eigenvalue representation, i.e., $\mu_{\bullet,1}$.

Fig. 3. Pictorial view of (57), (58), (60). The effect of the additional noise process between (57), (58) is shown by the RL gap.

Corollary 1: (Analytical solution of (40) for scalar-valued processes) Consider the scalar-valued version of the system in (1), (2). Then, for $D > D_{\min} = \frac{\sqrt{v^2 + 4\alpha^2 c^2 \sigma_w^2 \sigma_n^2} - v}{2\alpha^2 c^2}$, $\alpha \neq 0$, $c \neq 0$, where $v \triangleq c^2 \sigma_w^2 + \sigma_n^2 (1 - \alpha^2)$, the solution of (40) for is as follows:

$$R_{\text{in}}^G(D) = \frac{1}{2} \left[\log \left(\frac{\lambda}{D} \right) + \log \left(\frac{\lambda}{\lambda + \frac{\sigma_n^2}{c^2} (1 - \frac{\lambda}{D})} \right) \right], \quad (57)$$

where $\lambda = \alpha^2 D + \sigma_w^2$, and $D_{\max} = \lambda$.

Proof: The solution follows from Theorem 5 when $p = 1$. In particular, complementary slackness conditions in (51), (52) ensure that $\delta = D$ (uniform distortion allocation) and specify the value of $D_{\min}$ and $D_{\max}$. ■

Next, we remark how one can recover the well-known result of sequential or nonanticipative RDF obtained for scalar-valued time-invariant Gauss-Markov processes (without additional observation noise process).

Remark 7: (Sequential RDF) If in (57) we assume $\sigma_n^2 = 0$, then, $R_{\text{in}}^G(D) \equiv R^G(D)$ where

$$R^G(D) = \frac{1}{2} \log \left(\alpha^2 + \frac{\sigma_w^2}{D} \right), \quad (58)$$

with $D_{\min} > 0$ and $D_{\max} = \lambda$. This special case corresponds precisely to the sequential or nonanticipative RDF derived for scalar-valued time-invariant Gauss-Markov sources with MSE distortion (without observation noise process) derived in [3, Eq. (1.43)], [6, Eq. (14)], [4, Thm 3].

Using Remark 7 and the optimal closed form expression of (57), we can compute the RL gap due to having the additional observation noise. This result is stated in the following proposition.

Proposition 3: (RL Gap due to the additional observation noise) The RL gap between (57) and (58) is given by

$$RL \equiv R_{\text{in}}^G(D) - R^G(D) = \begin{cases} \frac{1}{2} \log \left(\frac{\lambda}{\lambda + \frac{\sigma_n^2}{c^2} (1 - \frac{\lambda}{D})} \right), & \text{if } D \in (D_{\min}, D_{\max}) \\ 0 & \text{if } D = D_{\max} \end{cases}, \quad (59)$$

where $(D_{\min}, D_{\max})$ are obtained in Corollary 1.

In the next remark, we state the steady-state solution of a lower bound on the optimal $\bar{R}_{\text{in}}^G(D)$ obtained via [1] (see [25, Appendix A, Eq. (103)]) for scalar-valued Gauss-Markov processes with additional observations noise.

Remark 8: (A lower bound on $\bar{R}_{\text{in}}^G(D)$) A lower bound on $\bar{R}_{\text{in}}^G(D)$ was recently derived in [25, Appendix A, Eq. (103)] via [1], for time-invariant scalar-valued Gauss-Markov processes with additional observations noise. This bound is described as follows

$$\bar{R}_{\text{in}}^G(D) \geq \frac{1}{2} \log \left(\frac{\lambda - D'_{\min}}{D - D'_{\min}} \right), \quad (60)$$

where $D'_{\max} = D_{\max} = \lambda$, and $D'_{\min}$ can be computed by finding the steady-state a posteriori error variance solution of a pre-KF algorithm imposed between the system process $\{\mathbf{x}_t : t \in \mathbb{N}_0\}$ and the output of the observations process $\{\mathbf{z}_t : t \in \mathbb{N}_0\}$. One can easily obtain after some calculations that $D'_{\min} = D_{\min}$.

Simulation study. In Fig. 3, we provide an illustrative example where we compare (57), (58) and (60). In addition, we illustrate the RL gap of Proposition 3. Interestingly, our simulations show that the lower bound that can be obtained via [1] achieves lower rates compared to the exact solution of our bound. Hence from Proposition 1, this means that the bound in (60) is not tight, in general, with respect to its optimal solution $\bar{R}_{\text{in}}^G(D)$ because if it was it would have been an upper bound on our solution. Clearly, (57), (60) coincide if $\sigma_n^2 = 0$.

V. ANALYTICAL BOUNDS BEYOND GAUSSIAN PROCESSES

In this section, we derive analytical lower bounds for additive i.i.d. zero mean noise processes. First, note that a subclass of $R_{\text{in}}(D)$ in (21) that only considers the best linear coding policies (among all linear policies for causal MMSE decoding) is already obtained from the analysis of jointly Gaussian processes in §IV (see Remark 4). Recall that this description is defined in (39). In what follows we denote the specific class of linear coding policies by $R_{\text{in}}^{\text{linear}}(D)$.

Next, we state the following proposition.

Proposition 4: (Data-Rate Bounds) For the system model described by (1), (2) with MSE distortion, the following inequalities hold

$$R_{\text{in}}(D) \stackrel{(a)}{\leq} R_{\text{in}}^{\text{linear}}(D) \leq R_{\text{ZD},\text{in}}^{\text{op}}(D) \quad (61)$$

where (a) holds with equality if (1), (2) are jointly Gaussian and the optimal minimizer, i.e., $\mathbf{P}^*(dy_t|y^{t-1}, z^t)$ of $R_{\text{in}}(D)$ is conditionally Gaussian. Note that (a) holds trivially at $D_{\max}$ for stable processes.

Proof: The derivation is similar to [34, Lemma 4], hence we omit it. ■

The next result strengthens a recent result obtained in [1, Theorem 9] beyond fully observable systems driven by additive i.i.d. zero mean noise processes.

Theorem 6: Suppose that the system model of (1), (2), is described by the following special case:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t, \quad \mathbf{z}_t = C\mathbf{x}_t, \quad (62)$$

where $A \in \mathbb{R}^{p \times p}$ is full rank, $\mathbf{w}_t \sim (0; \Sigma_{\mathbf{w}})$, $\Sigma_{\mathbf{w}} \succ 0$ is an i.i.d. non-Gaussian process with $h(\mathbf{w}_t) > -\infty$ and $C \in \mathbb{R}^{m \times p}$ is a full row rank matrix with $m \leq p$. Then,

$$R_{\text{ZD},\text{in}}^{\text{op}}(D) \geq \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} + \frac{pN(\mathbf{w})}{D} \right), \quad (63)$$

where $N(\mathbf{w}) = \frac{1}{2\pi e} 2^{\frac{2}{p} h(\mathbf{w}_t)}$ and $D \in (0, |A^T A|^{\frac{1}{p}} D + pN(\mathbf{w})]$.

Proof: For the model in (62), (40) simplifies to

$$R_{\text{in}}^{\text{linear}}(D) = \inf_{\substack{0 \prec \Delta \preceq \Lambda \\ \text{trace}(\Delta) \leq D}} \frac{1}{2} \log \frac{|\Lambda|}{|\Delta|}, \quad (64)$$

because $\Sigma_n = 0$ and its rank dictates the rank of Q . Now, the objective in (64) can be bounded as follows:

$$\begin{aligned} \frac{1}{2} \log \frac{|\Lambda|}{|\Delta|} &= \frac{p}{2} \log |A\Delta A^T + \Sigma_{\mathbf{w}}|^{\frac{1}{p}} - \frac{p}{2} \log |\Delta|^{\frac{1}{p}} \\ &\stackrel{(a)}{\geq} \frac{p}{2} \log \left(|A\Delta A^T|^{\frac{1}{p}} + |\Sigma_{\mathbf{w}}|^{\frac{1}{p}} \right) - \frac{p}{2} \log |\Delta|^{\frac{1}{p}} \\ &\stackrel{(b)}{=} \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} |\Delta|^{\frac{1}{p}} + |\Sigma_{\mathbf{w}}|^{\frac{1}{p}} \right) - \frac{p}{2} \log |\Delta|^{\frac{1}{p}} \\ &\stackrel{(c)}{=} \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} + |\Sigma_{\mathbf{w}}|^{\frac{1}{p}} |\Delta|^{-\frac{1}{p}} \right) \\ &\stackrel{(d)}{\geq} \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} + |\Sigma_{\mathbf{w}}|^{\frac{1}{p}} \frac{p}{\text{trace}(\Delta)} \right) \\ &\stackrel{(e)}{\geq} \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} + |\Sigma_{\mathbf{w}}|^{\frac{1}{p}} \frac{p}{D} \right) \\ &\stackrel{(f)}{\geq} \frac{p}{2} \log \left(|A^T A|^{\frac{1}{p}} + \frac{pN(\mathbf{w})}{D} \right), \end{aligned} \quad (65)$$

where (a) follows from Theorem 9; (b) follows from standard properties of square matrices of the same size; (c) follows by properties of log-determinant functions; (d) follows from the reverse application of (78); (e) follows because $\text{trace}(\Delta) \leq D$; (f) follows from (78). Note that (65) is well defined if $D \in (0, |A^T A|^{\frac{1}{p}} D + pN(\mathbf{w})]$. Finally, from Proposition 4, we obtain that $R_{\text{ZD},\text{in}}^{\text{op}}(D)$ is bounded from below by (65). This completes the derivation. ■

Next, we describe some technical remarks on Theorem 6.

Remark 9: (Technical remarks) (1) Theorem 6 can be derived from Theorem 2 using similar arguments, that is, the fact that $\text{trace}(\Delta_t) \mapsto \log\left(|A^T A|^{\frac{1}{p}} + \frac{pN(\mathbf{w})}{\text{trace}(\Delta_t)}\right)$ is convex and decreasing, then, applying Jensen's inequality [8, Theorem 2.6.2].

(2) We note that inequalities (a), (d), (e) hold with equality if $(A, \Sigma_{\mathbf{w}}, \Delta)$ are scalar matrices. In particular, (a) holds with equality because $A\Delta A^T$ will be proportional to $\Sigma_{\mathbf{w}}$, (d) because $|\Delta|^{-\frac{1}{p}} = \frac{p}{\text{trace}(\Delta)}$ and (e) because allocating equally Δ , i.e., $\Delta = \text{diag}(\frac{D}{p}, \dots, \frac{D}{p}) \in \mathbb{R}^{p \times p}$ is optimal (see Theorem 5). Finally, (f) holds with equality if $\{\mathbf{w}_t : t \in \mathbb{N}_0\}$ is Gaussian (see (78)).

In the next theorem, we derive analytical lower bounds to multidimensional time-invariant Gauss-Markov processes with additional additive i.i.d. observations noise process with spatially i.i.d. components.

Theorem 7: Suppose that for the system model of (1), (2), is described by the following special case:

$$\begin{aligned} \mathbf{x}_{t+1} &= A\mathbf{x}_t + \mathbf{w}_t \\ \mathbf{z}_t &= C\mathbf{x}_t + \mathbf{n}_t, \end{aligned} \quad (66)$$

where $A = \alpha I_p$, $\{\mathbf{w}_t : t \in \mathbb{N}_0\}$ is a zero mean i.i.d. process with $h(\mathbf{w}_t) > -\infty$ independent of $(\mathbf{x}_0, \{\mathbf{n}_t : t \in \mathbb{N}_0\})$, with $\Sigma_{\mathbf{w}} = \sigma_{\mathbf{w}}^2 I_p$ and $\sigma_{\mathbf{w}}^2 > 0$, $C = cI_p \in \mathbb{R}^{p \times p}$ is a square matrix, with $c \neq 0$, and $\{\mathbf{n}_t : t \in \mathbb{N}_0\}$ is a zero-mean i.i.d. process independent of $(\mathbf{x}_0, \{\mathbf{w}_t : t \in \mathbb{N}_0\})$ with $h(\mathbf{n}_t) > -\infty$, with $\Sigma_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I_p$ and $\sigma_{\mathbf{n}}^2 > 0$. Then, the following statements hold.

(1) If $\mathbf{w}_t \sim \mathcal{N}(0; \Sigma_{\mathbf{w}})$, and $\mathbf{n}_t \sim (0; \Sigma_{\mathbf{n}})$, then,

$$\begin{aligned} R_{\text{ZD},\text{in}}^{\text{op}}(D) &\geq \frac{p}{2} \log\left(\alpha^2 + \frac{p\sigma_{\mathbf{w}}^2}{D}\right) \\ &\quad + \frac{p}{2} \log\left(\frac{\lambda}{\lambda + \frac{N(\mathbf{n})}{c^2} \left(1 - \frac{\lambda p}{D}\right)}\right), \end{aligned} \quad (67)$$

where $N(\mathbf{n}) = \frac{1}{2\pi e} 2^{\frac{2}{p}h(\mathbf{n}_t)}$, $\frac{D_{\max}}{p} = \lambda$, $\lambda = \alpha^2 \frac{D}{p} + \sigma_{\mathbf{w}}^2$ and

$$\frac{D}{p} > \frac{\sqrt{v^2 + 4\alpha^2 c^2 \sigma_{\mathbf{w}}^2 N(\mathbf{n})} - v}{2\alpha^2 c^2}, \quad \alpha \neq 0, \quad c \neq 0, \quad (68)$$

with $v \triangleq c^2 \sigma_{\mathbf{w}}^2 + N(\mathbf{n})(1 - \alpha^2)$.

(2) If $\mathbf{w}_t \sim (0; \Sigma_{\mathbf{w}})$, and $\mathbf{n}_t \sim \mathcal{N}(0; \Sigma_{\mathbf{n}})$, then,

$$\begin{aligned} R_{\text{ZD},\text{in}}^{\text{op}}(D) &\geq \frac{p}{2} \log\left(\alpha^2 + \frac{pN(\mathbf{w})}{D}\right) \\ &\quad + \frac{p}{2} \log\left(\frac{\lambda}{\lambda + \frac{\sigma_{\mathbf{n}}^2}{c^2} \left(1 - \frac{\lambda p}{D}\right)}\right), \end{aligned} \quad (69)$$

where $N(\mathbf{w}) = \frac{1}{2\pi e} 2^{\frac{2}{p}h(\mathbf{w}_t)}$, $\frac{D_{\max}}{p} = \lambda$, $\lambda = \alpha^2 \frac{D}{p} + N(\mathbf{w})$ and

$$\frac{D}{p} > \frac{\sqrt{v^2 + 4\alpha^2 c^2 N(\mathbf{w}) \sigma_{\mathbf{n}}^2} - v}{2\alpha^2 c^2}, \quad \alpha \neq 0, \quad c \neq 0 \quad (70)$$

with $v \triangleq c^2 N(\mathbf{w}) + \sigma_{\mathbf{n}}^2(1 - \alpha^2)$.

Proof: First, observe that (40) simplifies to Theorem 4, because $(A, \Sigma_{\mathbf{w}}, \Sigma_{\mathbf{n}}, C)$ commute by pairs (this is a special case of the classes discussed in Proposition 2), hence $(A, \Sigma_{\mathbf{w}}, \Sigma_{\mathbf{n}}, C, \Delta, \Lambda)$ are simultaneously diagonalizable by an orthogonal matrix. In addition, since all dimensions are i.i.d., then, allocating equally the distortion across the dimensions is optimal. Indeed, using the KKT conditions of Theorem 5 we can obtain an analytical solution of Theorem 4 of the following form:

$$\begin{aligned} R_{\text{in}}^{\text{linear}}(D) &= \\ \frac{p}{2} \log\left(\alpha^2 + \frac{p\sigma_{\mathbf{w}}^2}{D}\right) &+ \frac{p}{2} \log\left(\frac{\lambda}{\lambda + \frac{\sigma_{\mathbf{n}}^2}{c^2} \left(1 - \frac{\lambda p}{D}\right)}\right), \end{aligned} \quad (71)$$

where $\lambda = \alpha^2 \frac{D}{p} + \sigma_{\mathbf{w}}^2$ and

$$\frac{D}{p} > \frac{\sqrt{v^2 + 4\alpha^2 c^2 \sigma_{\mathbf{w}}^2 \sigma_{\mathbf{n}}^2} - v}{2\alpha^2 c^2}, \quad \alpha \neq 0, \quad c \neq 0, \quad (72)$$

with $v \triangleq c^2 \sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{n}}^2(1 - \alpha^2)$.

It remains to investigate separately the cases (1) and (2). To do it we rewrite (71) as follows:

$$\widehat{R}_{\text{in}}^{\text{linear}}(D) = \frac{p}{2} \log\left(\alpha^2 + \frac{p\sigma_{\mathbf{w}}^2}{D}\right) - \frac{p}{2} \log(1 + \widehat{f}), \quad (73)$$

where the domain of the second right hand term is well-defined for $\widehat{f} \in (-1, 0]$ (this follows from the constraints of the optimization problem) such that

$$\widehat{f} \triangleq \frac{p\sigma_{\mathbf{n}}^2}{c^2} \left(\frac{1}{\alpha^2 D + p\sigma_{\mathbf{w}}^2} - \frac{1}{D} \right). \quad (74)$$

Note that $-\frac{p}{2} \log(1 + \widehat{f})$ is a convex and decreasing function with respect to $\widehat{f} \in (-1, 0]$.

(1) For this case, we use EPI (in (78)) to obtain $\frac{\text{trace}(\Sigma_{\mathbf{n}})}{p} = \sigma_{\mathbf{n}}^2 \geq N(\mathbf{n})$ which further means that

$$\widehat{f}' \triangleq \frac{pN(\mathbf{n})}{c^2} \left(\frac{1}{\alpha^2 D + p\sigma_{\mathbf{w}}^2} - \frac{1}{D} \right) \in (-1, 0], \quad (75)$$

with minimum distortion constrained by (68). Since $\sigma_{\mathbf{n}}^2 \geq N(\mathbf{n})$, it means that $\widehat{f} \leq \widehat{f}'$ because the latter takes negative values closer to zero. This in turn implies that $\log(1 + \widehat{f}) \leq \log(1 + \widehat{f}')$ hence, $-\log(1 + \widehat{f}) \geq -\log(1 + \widehat{f}')$ and the result follows. It can be trivially observed in (67) that $\frac{D_{\max}}{p} = \lambda$.

(2) follows using similar arguments to (1) and we omit it. ■

Remark 10: (Scalar-case) If in Theorem 7 we assume Gaussian processes, then, (67), (69) bounds are tight. Moreover, if in Theorem 7 we assume $p = 1$, we obtain analogous lower bounds for scalar-valued Markov processes with additive i.i.d. non-Gaussian processes reminiscent to Corollary 1.

VI. CONCLUSIONS AND ONGOING RESEARCH

In this paper, we introduced a novel sequential RDF to compute lower bounds on the average length of all causal prefix free codes (zero-delay codes) for a partially observable multivariate Markov system driven by additive i.i.d. zero mean noise processes with a MSE distortion constraint. When our setup is characterized by jointly Gaussian processes, we provided a complete characterization in finite time and in the asymptotic limit, and a non-trivial computational approach by means of a generalized reverse-waterfilling algorithm that solves the problem optimally in the asymptotic limit, under the assumption that all matrices of the problem commute by pairs. For scalar-valued Gaussian processes with additional observations noise, we found an exact explicit solution. Finally, we derived analytical results that consider setups that go beyond additive zero mean i.i.d. Gaussian processes.

As ongoing research activity we will extend our framework to non-stationary partially observable systems with various distortion constraints, revisit the celebrated LQG closed-loop control problem for linear time-invariant partially observable Gaussian systems (see e.g., [21]) and re-derive the separation principle for noiseless channels based on our new information measure. We expect that in comparison to [21, Fig. 3], the additional pre-KF algorithm in the realization of the predictive coding scheme will not be necessary. This will significantly reduce the system's complexity.

REFERENCES

- [1] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4525–4540, Nov 2019.
- [2] A. K. Gorbunov and M. S. Pinsker, "Nonanticipatory and prognostic epsilon entropies and message generation rates," *Problems Inf. Transmiss.*, vol. 9, no. 3, pp. 184–191, July-Sept. 1972.
- [3] —, "Prognostic epsilon entropy of a Gaussian message and a Gaussian source," *Problems Inf. Transmiss.*, vol. 10, no. 2, pp. 93–109, Apr.-June 1972, translation from Problemy Peredachi Informatsii, vol. 10, no. 2, pp. 5–25, April-June 1974.
- [4] M. S. Derpich and J. Østergaard, "Improved upper bounds to the causal quadratic rate-distortion function for Gaussian stationary sources," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3131 – 3152, May 2012.
- [5] P. A. Stavrou, J. Østergaard, and C. D. Charalambous, "Zero-delay rate distortion via filtering for vector-valued Gaussian sources," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 841–856, Oct 2018.
- [6] S. Tatikonda, A. Sahai, and S. Mitter, "Stochastic linear control over a communication channel," *IEEE Trans. Autom. Control*, vol. 49, pp. 1549 – 1561, 2004.
- [7] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed, "Nonanticipative rate distortion function and relations to filtering theory," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 937–952, 2014.
- [8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.
- [9] P. A. Stavrou, T. Tanaka, and S. Tatikonda, "The time-invariant multidimensional Gaussian sequential rate-distortion problem revisited," *IEEE Transactions on Automatic Control*, vol. 65, no. 5, pp. 2245–2249, 2020.
- [10] T. Tanaka, K. K. Kim, P. A. Parrilo, and S. K. Mitter, "Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1896–1910, April 2017.
- [11] P. A. Stavrou, T. Charalambous, C. D. Charalambous, and S. Loyka, "Optimal estimation via nonanticipative rate distortion function and applications to time-varying Gauss-Markov processes," *SIAM J. on Control Optim.*, vol. 56, no. 5, pp. 3731–3765, 2018.
- [12] A. J. Fuglsig and J. Østergaard, "Zero-delay multiple descriptions of stationary scalar Gauss-Markov sources," *Entropy*, vol. 21, no. 12, 2019.
- [13] C. D. Charalambous, C. Kourtellis, T. Charalambous, and J. H. van Schuppen, "Generalizations of nonanticipative rate distortion function to multivariate nonstationary Gaussian autoregressive processes," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 8190–8195.
- [14] T. Tanaka, "Zero-delay rate-distortion optimization for partially observable Gauss-Markov processes," in *Proc. IEEE Conf. Decision Control*, Dec 2015, pp. 5725–5730.
- [15] T. Tanaka, P. M. Esfahani, and S. K. Mitter, "LQG control with minimum directed information: Semidefinite programming approach," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 37–52, Jan 2018.
- [16] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Nov. 27–30 1990, pp. 303–305.
- [17] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), 1998.
- [18] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.
- [19] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501–1518, 1991.
- [20] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [21] S. Yüksel, "Jointly optimal LQG quantization and control policies for multi-dimensional systems," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1612–1617, June 2014.
- [22] D. Neuhoff and R. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 701 – 713, 1982.
- [23] A. Khina, V. Kostina, A. Khisti, and B. Hassibi, "Tracking and control of Gauss-Markov processes over packet-drop channels with acknowledgments," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 2, pp. 549–560, 2019.
- [24] P. A. Stavrou and M. Skoglund, "Asymptotic reverse-waterfilling algorithm for certain classes of vector Gauss-Markov processes," *diva.org*, 2020.
- [25] V. Kostina, "Fundamental limitations in distributed tracking," in *IEEE International Symposium on Information Theory*, 2020. [Online]. Available: <https://arxiv.org/abs/1910.02534v1>
- [26] A. Wyner, "A definition of conditional mutual information for arbitrary ensembles," *Information and Control*, vol. 38, no. 1, pp. 51 – 59, 1978.
- [27] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: Properties and variational equalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6019–6052, Nov 2016.
- [28] C. D. Charalambous, P. A. Stavrou, C. K. Kourtellis, and I. Tzortzis, "Directed information subject to a fidelity. applications to conditionally Gaussian processes," in *Proc. European Control Conf.*, June 2018, pp. 3071–3076.
- [29] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2005.
- [30] V. Krishnamurthy, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press, 2016.
- [31] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.
- [32] D. Simon, *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. Wiley-Interscience, 2006.
- [33] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, 1969.
- [34] P. A. Stavrou, J. Østergaard, and M. Skoglund, "Bounds on the sum-rate of MIMO causal source coding systems with memory under spatio-temporal distortion constraints," *Entropy*, vol. 22, no. 8, 2020.
- [35] D. A. Harville, *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, 1997.
- [36] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 33–55, 2011.

APPENDIX A
SUPPLEMENTARY MATERIAL I

In what follows, we state a few important definitions, theorems and lemmas that we use throughout the paper.

Definition 2: (Commuting matrices) [35, p. 5] Two $p \times p$ matrices A, B commute if $AB = BA$. More generally, the collection of $p \times p$ matrices $(A_1, \dots, A_k)$ *commute by pairs* if $A_i A_j = A_j A_i$, for $j > i$, $i = 1, 2, \dots, k$.

Definition 3: (Product of symmetric matrices) [35, Section 1.3 (6)] The production AB of two *square symmetric* matrices A, B is itself *symmetric* if and only if A and B commute.

Definition 4: (Orthogonal matrix) [35, Section 8.4 (a)] A square (say $p \times p$) matrix U is said to be *orthogonal* if $U^T U = U U^T = I_p$.

Theorem 8: (Commuting matrices are simultaneously diagonalizable) [35, Theorem 21.13.1] If a collection of $p \times p$ matrices $(A_1, \dots, A_k)$ *commute by pairs*, then they can be simultaneously diagonalized with an orthogonal matrix $U \in \mathbb{R}^{p \times p}$, namely, there exists an orthogonal matrix U and diagonal matrices, say $D_1, \dots, D_k$ such that for $i = 1, \dots, k$ $U^T A_i U = D_i$. If a collection of $p \times p$ matrices $(A_1, \dots, A_k)$ are *simultaneously diagonalizable*, then, they *commute by pairs*.

Definition 5: (Right inverse matrix) [35, Section 8.1] A *right inverse* of a rectangular matrix $U \in \mathbb{R}^{m \times p}$ is said to be a matrix $U^\dagger \in \mathbb{R}^{m \times p}$ such that $U U^\dagger = I_m$.

The next lemma provides a necessary and sufficient condition for existence of a right inverse matrix.

Lemma 3: [35, Lemma 8.1.1] A rectangular matrix $U \in \mathbb{R}^{m \times p}$ has a right inverse if and only if it is full row rank, i.e., $\text{rank}(U) = m$.

The next lemma gives a nice structure of the right inverse matrix for rectangular full row rank matrices.

Lemma 4: [35, Lemma 9.2.9] If a rectangular matrix $U \in \mathbb{R}^{m \times p}$ has a full row rank, then, $U^\dagger = U^T (U U^T)^{-1}$.

The structure in Lemma 4 is a special case of the generalized inverse matrix that satisfies Penrose conditions [35, Theorem 20.1.1]. This special class of inverse matrices is called Moore-Penrose (pseudo)-inverse matrix.

The following definition of EP and EPI were introduced in the landmark paper by Shannon [20] (see also [36] and the references therein).

Definition 6: (Entropy Power Inequalities) Let $\mathbf{x} \in \mathbb{R}^p$ be any random p -dimensional vector with probability density $\bar{p}(x)$ (w.r.t. to the Lebesgue measure). Such vector has differential entropy defined by

$$h(\mathbf{x}) \triangleq \int_{\mathbb{R}^p} \bar{p}(x) \log_2 \bar{p}(x) dx \quad (76)$$

provided the integral exists (it is finite). Then, the EP of $\mathbf{x} \in \mathbb{R}^p$ is defined as follows [20]

$$N(\mathbf{x}) \triangleq \frac{1}{2\pi e} 2^{\frac{2}{p} h(\mathbf{x})}. \quad (77)$$

The entropy power satisfies the following EPI:

$$N(\mathbf{x}) \stackrel{(a)}{\leq} |\Sigma_{\mathbf{x}}|^{\frac{1}{p}} \stackrel{(b)}{\leq} \sigma_{\mathbf{x}}^2, \quad (78)$$

where $\sigma_{\mathbf{x}}^2 = \frac{1}{p} \text{var}(\mathbf{x})$ is the per-component (dimension) variance or equivalently, $\sigma_{\mathbf{x}}^2 = \frac{1}{p} \text{trace}(\Sigma_{\mathbf{x}})$, $\Sigma_{\mathbf{x}} \succ 0$; (a) holds with equality if and only if $\mathbf{x}$ is Gaussian; (b) holds with equality if $\mathbf{x}$ is white. Note that $N(\mathbf{x})$ is the power of a white Gaussian random vector having the same entropy as $\mathbf{x}$. Next, we prove some scaling properties of entropy power.

Proposition 5: (Scaling property of EP for square full rank matrix) Suppose that $U \in \mathbb{R}^{p \times p}$ is full rank and $\mathbf{x} \in \mathbb{R}^p$. Then, EP scales as

$$N(U\mathbf{x}) = \text{abs}(|U|)^{\frac{2}{p}} N(\mathbf{x}) \quad (79)$$

Proof: By definition of EP, $N(U\mathbf{x}) = \frac{1}{2\pi e} 2^{\frac{2}{p} h(U\mathbf{x})} \stackrel{(*)}{=} \frac{1}{2\pi e} 2^{\frac{2}{p} (h(\mathbf{x}) + \log_2(\text{abs}|U|))} = \text{abs}(|U|)^{\frac{2}{p}} N(\mathbf{x})$, where $(*)$ follows from standard properties of differential entropy [8, Eq. (8.71)]. This completes the proof. ■

The next theorems are used to derive some of our results in this paper.

Theorem 9: (Minkowski's determinant inequality) [19, Theorem 5] Suppose that $(A, B) \in \mathbb{R}^{p \times p}$ are non-negative semi-definite matrices. Then, the following inequality holds:

$$|A + B|^{\frac{1}{p}} \geq |A|^{\frac{1}{p}} + |B|^{\frac{1}{p}}, \quad (80)$$

with equality if and only if A is proportional to B .