Stochastic Linear Control Over a Communication Channel

Sekhar Tatikonda, *Member, IEEE*, Anant Sahai, *Member, IEEE*, and Sanjoy Mitter, *Life Fellow, IEEE*

***Abstract*—We examine linear stochastic control systems when there is a communication channel connecting the sensor to the controller. The problem consists of designing the channel encoder and decoder as well as the controller to satisfy some given control objectives. In particular, we examine the role communication has on the classical linear quadratic Gaussian problem. We give conditions under which the classical separation property between estimation and control holds and the certainty equivalent control law is optimal. We then present the sequential rate distortion framework. We present bounds on the achievable performance and show the inherent tradeoffs between control and communication costs. In particular, we show that optimal quadratic cost decomposes into two terms: A full knowledge cost and a sequential rate distortion cost.**

***Index Terms*—Certainty equivalent control, communication constraints, networked control, separation, sequential rate distortion, stochastic linear systems.**

I. Introduction

RECENT advances in technology have led to increased activity in understanding and designing networked control systems. In this paper, we examine a stochastic control problem where there is a communication channel connecting the sensor to the controller. This problem arises when the plant and the controller are geographically separated and there is a band-limited and possibly noisy communication channel connecting them. In addition, communication constraints can arise when there is no major geographical separation between the controller and the plant, but there is a shared communication medium that is being used along with other users in the same local area, or as a part of the larger system. Although we do not explicitly examine the networking issues *per se* in this paper, we believe that a fundamental understanding of the role of communication constraints will be essential to a more complete theory of networked control.

The system we consider consists of a plant, an encoder, a channel, a decoder, and a controller. The plant and the channel are given to us. Our task is to design the encoder, decoder, and controller to satisfy some given control objectives. Since we have a distributed system the choice of information pattern, [26], can have a dramatic effect on what control performance is achievable. We discuss the effect that the choice of information pattern at the encoder has on the communication requirements needed to achieve the control objective. In particular, we examine the role communication has on the classical linear quadratic Gaussian (LQG) problem. To this end we present the sequential rate distortion (SRD) framework. We derive bounds on the achievable performance and show the inherent tradeoffs between control and communication costs. In particular, we show that the optimal LQG cost decomposes into two terms: a full knowledge cost and a sequential rate distortion cost.

There are two classical notions of separation that we examine in this paper. The first notion is the control-theoretic separation between state estimation and control. We present conditions that insure the optimality of the certainty equivalent control law. These build on the work of Bar-Shalom and Tse [3]. The second notion is the information-theoretic separation between source encoder and channel encoder. In particular, in the limit of long delays, it is known that one can, without loss of generality, design the source encoder and the channel encoder separately [11]. This separation is known to hold quite broadly, [25], but, in general, fails for both short delays and for unstable processes. In the limit of large delays, [18] showed that the estimation of unstable processes can be covered by a suitably modified separation theorem, but this information-theoretic result does not extend to the case of limited delay. Since delay is an important issue in control applications we cannot apply the information-theoretic separation results to our problem. To deal with this delay issue, we present the sequential rate distortion framework first introduced in [13] and further developed in [19], [20], and [23].

Bansal and Basar examine the simultaneous design of the measurement and control strategies for a class of stochastic systems [2]. Borkar and Mitter introduced the present problem of LQG control under communication constraints [6]. There they looked at stable systems and noiseless digital channels. They introduced the innovations coding scheme. Unfortunately, their scheme does not work for unstable systems. We generalize these results to unstable systems and noisy channels. Moreover, we give conditions under which our coding scheme is optimal, the separation principle holds, and the certainty equivalent controller is optimal. Nair and Evans [17] examined mean-square stabilizability over a noiseless channel. In contrast, we examine the LQG performance over both noisy and noiseless channels. Reference [7] examines the role of nonclassical information patterns for Markov decision problems. We extend those results to the LQG problem with differing information pattern at the

Manuscript received June 4, 2003; revised December 19, 2003. Recommended by Guest Editors P. Antsaklis and J. Baillieul. This work was supported by the Army Research Office under MURI Grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466, and by the Department of Defense under MURI Grant: Complex Adaptive Networks for Cooperative Control Subaward 03-132.

S. Tatikonda is with Yale University, New Haven, CT 06520 USA (e-mail: sekhar.tatikonda@yale.edu).

A. Sahai is with the University of California, Berkeley, CA 94720 USA.

S. Mitter is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Digital Object Identifier 10.1109/TAC.2004.834430

encoder. The design of optimal sequential quantization schemes for uncontrolled Markov processes was examined in [8].

Recently, there has been a good deal of work studying deterministic systems under communication constraints: [10], [19], [21], [22], [15], [1], [27], and [28]. In [21] and [22], we examined deterministic systems over both noiseless and noisy channels. Our work here differs in that we treat stochastic systems, provide a general separation result, apply the SRD [23] framework to the LQG problem, and discuss conditions under which the plant is matched to the channel. The work in this paper first appeared in preliminary form in [24] and represents a portion of the first author's Ph.D. dissertation [19].

In Section II, we formulate the problem. In Section III, we examine the LQG problem under different information patterns and in different communication channels. In Section IV, we introduce the sequential rate distortion framework and give closed form solutions to the sequential rate distortion function for Gauss–Markov sources. In Section V, we examine some particular scenarios.

II. Problem Formulation

Here, we present the different components of our problem.

a) Plant: We consider the following discrete-time, stochastic, linear system:

$$X_{t+1} = FX_t + GU_t + W_t, \qquad t \geq 0 \tag{1}$$

where $\{X_t\}$ is a $\mathbb{R}^d$-valued state process, $\{U_t\}$ is a $\mathbb{R}^m$-valued control process, and $\{W_t\}$ is an independent and identically distributed (IID) sequence of Gaussian random variables with zero mean and covariance K_W. The initial position X_0 is Gaussian with zero mean and covariance K_{X_0}. Let $F \in \mathbb{R}^{d\times d}, G \in \mathbb{R}^{d\times m}$ and assume (F, G) is a controllable pair. The decoder output process $\{Y_t\}$ is a $\mathbb{R}^d$-valued process. The output Y_t represents the decoder's estimate of the state of the system at time t; see Fig. 1.

Upper case variables, like X, represent random variables and lower case variables, like x, represent particular realizations. Throughout this paper, "log" refers to logarithm base two. Let $x^t \doteq (x_0, \ldots, x_t)$.

b) Channel: The channel input and output alphabet spaces are denoted $\mathcal{A}$ and $\mathcal{B}$, respectively. In this paper, we restrict ourselves to time-invariant memoryless channels which can be modeled as stochastic kernels: $P(dB_t \,|\, a_t)$. Specifically, for each realization of $A_t = a_t$ the conditional probability of B_t given a_t is denoted by $P(dB_t \,|\, a_t)$. At time t the encoder produces a channel input symbol $A_t = a_t$ and the channel outputs the channel output symbol B_t according to the probability $P(dB_t \,|\, a_t)$. The *Shannon capacity* is defined as $C^{\text{cap}} = \sup_{P(dA_t)} I(A_t; B_t)$ where $I(\cdot;\cdot)$ is the mutual information between the channel input and output. See the Appendix for a review of mutual information. In this paper we study two particular channels.

Noiseless digital channel with rate R: The channel input and output alphabets are the same: $\mathcal{A} = \mathcal{B}$. The alphabet size is $|\mathcal{A}| = 2^R$ where R is called the *rate* of the channel. The channel is noiseless and memoryless: $P(dB_t \,|\, a_t) = \delta_{\{B_t = a_t\}}$ (where δ is a Dirac measure.) In this case $C^{\text{cap}} = R$.

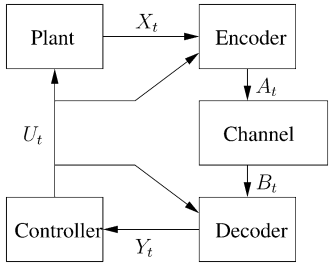


Fig. 1. System.

Memoryless vector Gaussian channel: The channel input and output alphabets are: $\mathcal{A} = \mathcal{B} = \mathbb{R}^d$. The channel is memoryless: $B_t = a_t + V_t$ where $\{V_t\}$ is an IID sequence of Gaussian random variables with zero mean and covariance K_V. The input symbols $\{A_t\}$ satisfy the power constraint: $E(\|A_t\|_2^2) \leq P, \forall t$. This channel is often used as a simplified model of a wireless channel. The supremizing input distribution $P(dA_t)$ can be shown to be a zero mean, vector-valued, Gaussian random variable. Hence, $E(\|A_t\|_2^2) = \text{tr}(K_A)$. See [9] for more details. The capacity is given by

$$C^{\text{cap}} = \max_{\text{tr}(K_A)\leq P} \frac{1}{2} \log \frac{|K_A + K_V|}{|K_V|}$$

where $|K|$ denotes the determinant of K.

c) Information Pattern: Our task is to design an encoder, decoder, and controller to achieve given control objectives. Thus we need to specify the information pattern of each of these components [26]. There are five types of signals: state X_t, channel input A_t, channel output B_t, decoder output Y_t, and control U_t. The following time-ordering represents the causal ordering in which the events of the system occur:

$$X_0, A_0, B_0, Y_0, U_0, \ldots, X_{T-1}, A_{T-1}, B_{T-1}, Y_{T-1}, U_{T-1}. \tag{2}$$

Encoder: We model the encoder at time t as a stochastic kernel: $P(dA_t \,|\, x^t, a^{t-1}, b^{t-1}, y^{t-1}, u^{t-1})$. Deterministic encoders are modeled as Dirac measures. Note that, in general, there are five potential kinds of feedback to the encoder. We will examine three different information patterns.

A) $I_t^A = \{X^t, A^{t-1}, B^{t-1}, Y^{t-1}, U^{t-1}\}$: This information pattern represents full knowledge at the encoder.

B) $I_t^B = \{X^t, A^{t-1}, U^{t-1}\}$: In this information pattern, the channel output and decoder output are not available to the encoder.

C) $I_t^C = \{X_t\}$: In this information pattern, only the current state is available to the encoder.

Information pattern B), along with time-ordering (2), implies that A_t is independent of B^{t-1}, Y^{t-1} conditioned on X^t, A^{t-1}, U^{t-1}. Similarly information pattern C implies A_t is independent of $A^{t-1}, B^{t-1}, Y^{t-1}, U^{t-1}$ conditioned on X_t.

Decoder: We model the decoder at time t as a stochastic kernel: $P(dY_t \,|\, b^t, y^{t-1}, u^{t-1})$. The information pattern here along with the time-ordering (2) implies that Y_t is independent of X^t, A^t given B^t, Y^{t-1}, U^{t-1}. As we will see later, the output of the decoder can be taken to be an estimate of the state of the plant.

Controller: We model the controller at time t as a stochastic kernel: $P(dU_t \mid y_t)$. The information pattern here along with the time-ordering (2) implies that U_t is independent of $X^t, A^t, B^t, Y^{t-1}, U^{t-1}$ conditioned on Y_t. Note that we are assuming the controller takes as input only the decoder output. Thus, there is a separation structure between the decoder and the controller. We will give conditions under which this separation structure can be assumed without loss of generality.

Clearly, more information available at the encoder will lead to better control performance. Information pattern A) can be viewed as the best scenario and information pattern C) can be viewed as the worst scenario. There are many information patterns in between these two extremes and information pattern B) represents one such information pattern. As we will see later, the important feature of information pattern A) is that the encoder knows the decoder's state. We will discuss scenarios where this naturally occurs. The important feature of information pattern B) is that the encoder has access to the previous control signals U_t. Here, we are envisioning situations where the encoder is geographically collocated with the plant and, hence, can observe both the state of the plant as well as the control actions applied to the plant. Finally, information pattern C) is useful for modeling situations where the control signals are not observed at the encoder. This information pattern can be used for scenarios where we want to model encoders that need to be simple and memoryless.

d) Interconnection and Performance Objective: In order to consider a performance objective, we need to insure that there is a well-defined joint measure over the variables of interest. We are given the plant and channel. For each encoder, decoder, and controller satisfying the required information patterns we can *interconnect* the different stochastic kernels together to produce the following joint measure:

$$\begin{aligned}&P(dX^T, dA^{T-1}, dB^{T-1}, dY^{T-1}, dU^{T-1})\\ &\quad= \prod_{t=0}^{T-1}\{P(dX_{t+1} \mid x_t, u_t)P(dA_t \mid x^t, a^{t-1}, b^{t-1}, y^{t-1}, u^{t-1})\\ &\qquad\times P(dB_t \mid a_t)P(dY_t \mid b^t, y^{t-1}, u^{t-1})P(dU_t \mid y_t)\}P(dX_0).\end{aligned}$$

Note that this joint measure preserves the dynamics of the plant and the channel as well as maintains the information patterns of the encoder, decoder, and controller. Let $\mathcal{P}$ denote the set of all such joint measures: $P(dX^{T-1}, dA^{T-1}, dB^{T-1}, dY^{T-1}, dU^{T-1})$.

Our performance objective is the LQG cost

$$\limsup_{T\to\infty}\frac{1}{T}E\left[\sum_{t=0}^{T-1}X_t'QX_t + U_t'SU_t\right] \tag{3}$$

where Q and S are positive definite. Our goal is to minimize the LQG cost (3) over the set of all measures in $\mathcal{P}$ that are consistent with the given plant, channel, and information pattern. We are also interested in understanding how the properties of the channel, like rate and power, and the choice of information pattern at the encoder can effect the minimum LQG performance cost.

e) Sequential Rate Distortion: One role of the feedback link from the encoder to the decoder is to convey information aimed at reducing the controller's uncertainty about the state of the system. At time t the evolution of the state is given by $X_t = F^tX_0 + \sum_{i=0}^{t-1}F^{t-i-1}(GU_i + W_i)$. Since the decoder has access to the control signals the uncertainty in X_t is determined by the primitive random variables X_0, and $\{W_t\}$. The decoder's estimate of the state is given by

$$\begin{aligned}&E(X_t \mid b^t, u^{t-1}, y^{t-1})\\ &\quad= F^tE(X_0 \mid b^t, u^{t-1}, y^{t-1})\\ &\qquad+\sum_{i=0}^{t-1}F^{t-i-1}E(W_i \mid b^t, u^{t-1}, y^{t-1}) + \sum_{i=0}^{t-1}F^{t-i-1}Gu_i.\end{aligned}$$

The information transmitted across the channel should only be relevant for determining X_0 and $\{W_t\}$. Hence, it makes sense to examine the uncontrolled dynamics: $X_t = F^tX_0 + \sum_{i=0}^{t-1}F^{t-i-1}W_i$. We will examine the uncontrolled dynamics in Section IV where we introduce the *sequential rate distortion* framework. This turns out to be the appropriate framework for understanding what information, relevant to the control objective, should be transmitted over the communication channel.

III. LQG Performance Objective

Here, we consider conditions that insure the optimality of the certainty equivalent controller. We then show that the LQG cost can be decomposed into a full knowledge cost and a partial knowledge cost.

Recall our goal is to minimize the long term average cost given by (3). Under full state observation, it is well known that the optimal steady state control law is a linear gain of the form $U_t = LX_t$ where

$$L = -(G'PG + S)^{-1}G'PF \tag{4}$$

and P satisfies the Riccati equation

$$P = F'(P - PG(G'PG + S)^{-1}G'P)F + Q. \tag{5}$$

The optimal cost is given by

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}E\left[\sum_{t=0}^{T-1}X_t'QX_t + U_t'SU_t\right] &= E(W'PW)\\ &= \mathrm{tr}(PK_W).\end{aligned} \tag{6}$$

Furthermore, these results continue to hold for the LQ problem where the process disturbances $\{W_t\}$ in (1) are no longer Gaussian but are uncorrelated with zero mean and common covariance matrix K_W. That is the separation result continues to hold if only second-order statistics are specified. These results are standard and can be found in [5].

The addition of a communication channel converts the fully observed LQG control problem shown previously into a partially observed LQ control problem. The structure of the observation is determined by both the communication channel and the choice of encoder and decoder.

A. Certainty Equivalence

We now present conditions that insure the optimality of the certainty equivalent control law: $U_t = LY_t$. Where $Y_t = E(X_t \mid B^t, U^{t-1})$ is the decoder's estimate of the state of the plant. To that end, we discuss the *no dual effect* property [3].

Fix a memoryless channel $P(dB_t \mid a_t)$, an information pattern I_t for the encoder, and an encoder $P(dA_t \mid I_t)$. Let the state estimation error be $\Delta_t = X_t - E(X_t \mid B^t, U^{t-1})$. We know that for any control sequence $\{u_t\}$ we have $X_t = F^t X_0 + \sum_{i=0}^{t-1} F^{t-1-i}(Gu_i + W_i)$. Define $\bar{X}_t = X_t - \sum_{i=0}^{t-1} F^{t-1-i} G u_i$ to be the uncontrolled state. Similarly, let $\bar{B}_t$ denote the channel output when all the controls are set to zero. Specifically for each particular realization of the primitive random variables: $X_0, \{W_t\}$, and $\{V_t\}$ (the latter is included if the channel is the Gaussian channel) B_t represents the channel output under the control sequence $\{u_t\}$ and $\bar{B}_t$ represents the channel output when all the controls are set to zero.

The control has *no dual effect* if for all $\{u_t\}$ and $\forall t$

$$E[\Delta_t \Delta_t' \mid B^t, u^{t-1}] = E[\Delta_t \Delta_t' \mid \bar{B}^t] \quad P-a.s.$$

The condition of no dual effect essentially states that the error covariance is independent of the control signals applied. The term "dual" comes from the control's dual role: Effecting state evolution and probing the system to reduce state uncertainty. If the control has no dual effect, the latter probing property will not be in effect.

Proposition 3.1: The optimal control law for system (1) is a certainty equivalent control law if and only if the control has no dual effect.

Proof: Bar-Shalom and Tse [3] prove this result for the case when the output measurement model, $P(dB_t \mid x_t)$, has the following functional form $B_t = g_t(x_t, V_t)$ where $\{g_t\}$ is a given sequence of functions and $\{V_t\}$ is an IID sequence of random variables that are independent of the W_t process. For our case, depending on the information pattern of the encoder and the memoryless channel, we have a more general output measurement model, $P(dB_t \mid x^t, u^{t-1}, a^{t-1}, b^{t-1}) = \int P(dB_t \mid a_t) P(da_t \mid x^t, u^{t-1}, a^{t-1}, b^{t-1})$, with the following functional form $B_t = g_t(x^t, u^{t-1}, a^{t-1}, b^{t-1}, V_t)$. The proof for this more general output measurement model is a straightforward generalization of the proof given in [3]. □

We now give a necessary condition insuring no dual effect.

Lemma 3.1: If the sigma-fields are nested, $\sigma(\bar{B}^t) \subset \sigma(B^t, U^{t-1})$, and $E[\bar{X}_t \mid \bar{B}^t] = E[\bar{X}_t \mid B^t, U^{t-1}, \bar{B}^t]$ then there is no dual effect.

Proof: Note that

$$\begin{aligned} X_t - E[X_t \mid B^t, U^{t-1}] &= \bar{X}_t - E[\bar{X}_t \mid B^t, U^{t-1}] \\ &= \bar{X}_t - E[\bar{X}_t \mid \bar{B}^t, B^t, U^{t-1}] \\ &= \bar{X}_t - E[\bar{X}_t \mid \bar{B}^t]. \end{aligned}$$

The last two equalities holds by hypothesis. Thus, $\Delta_t = X_t - E[X_t \mid B^t, U^{t-1}]$ is independent of the control sequence applied. □

We now use Lemma 3.1 to give conditions for the control to have no dual effect for the noiseless digital channel and the memoryless Gaussian channel.

Noiseless digital channel: Because the channel is noiseless the encoder information patterns I_t^A and I_t^B are equivalent (they define the same sigma-field.) Also, for information patterns I_t^A and I_t^B, both the encoder and decoder have access to the control signals hence they both can subtract out the effect of the control. Furthermore, as shown in [8], the optimal encoder will have the form of a quantizer applied to the innovation: $B_t = A_t = q_t(X_t - E(X_t \mid b^{t-1}, u^{t-1})) = q_t(\bar{X}_t - E(\bar{X}_t \mid b^{t-1}, u^{t-1}))$, where q_t is a quantizer.

We will show that $\bar{B}_t = B_t$ and $\bar{X}_t \to B^t \to U^{t-1}$ forms a Markov chain and, hence, $E[\bar{X}_t \mid B^t, U^{t-1}] = E[\bar{X}_t \mid \bar{B}^t]$.

Clearly, $\bar{B}_0 = B_0$. Note that $\bar{X}_0, W_0 \to B_0 \to U_0$ because conditioned on B_0 the control U_0 is independent of $\bar{X}_0$ and W_0. This implies $\bar{X}_1 \to B_0 \to U_0$ because $\bar{X}_1 = F\bar{X}_0 + W_0$. Hence, $\bar{B}_1 = B_1$. Now $\bar{X}_1 \to B^1 \to U_0$ because B_1 is a function of $\bar{X}_1$ and B_0 and is independent of U_0.

By the induction hypothesis, assume that $\bar{B}^t = B^t$ and $\bar{X}_k \to B^k \to U^{k-1}$ for $1 \le k \le t$. Note that $\bar{X}_t, W_t \to B^t, U^{t-1} \to U_t$ because, conditioned on B^t, U^{t-1}, the control U_t is independent of $\bar{X}_t$ and W_t. Now, by the induction hypothesis and the fact that W_t is independent of B^t, U^{t-1}, we have $\bar{X}_t, W_t \to B^t \to U^{t-1}$. Therefore, $\bar{X}_t, W_t \to B^t \to U^t$. This implies $\bar{X}_{t+1} \to B^t \to U^t$ because $\bar{X}_{t+1} = F\bar{X}_t + W_t$. Hence, $\bar{B}_{t+1} = B_{t+1}$. Now, $\bar{X}_{t+1} \to B^{t+1} \to U^t$ because B_{t+1} is a function of $\bar{X}_{t+1}$ and B^t and is independent of U^t. Hence, there is no dual effect.

Memoryless vector Gaussian channel: Here, we restrict ourselves to encoders that are deterministic and linear: $\{A_t = \gamma_{1t} x^t + \gamma_{2t} u^{t-1} + \gamma_{3t} a^{t-1} + \gamma_{4t} b^{t-1}\}$ where the γ's are matrices of appropriate dimension. Some of the γ's may be set to zero depending on the information pattern of the encoder. (We show in the next section that we may choose linear encoders without loss of generality.) Due to the Gaussian channel we have $B_t = a_t + V_t$.

Since the plant dynamics, the encoder, and the channel are linear we can rewrite B_t in terms of the primitive random variables and the controls: $B_t = \eta_{1t} X_0 + \eta_{2t} W^{t-1} + \eta_{3t} V^t + \eta_{4t} U^{t-1}$ where the η's are matrices of appropriate dimension. Similarly, the "uncontrolled" channel output can be written as $\bar{B}_t = \eta_{1t} X_0 + \eta_{2t} W^{t-1} + \eta_{3t} V^t$ with the same η matrices. Hence we can write $\bar{B}^t = B^t - \Gamma_t U^{t-1}$ for appropriate matrices Γ_t. The information in B^t, U^{t-1} relevant to estimating $\bar{X}_t$ is summarized by $\bar{B}^t$. Thus, for linear encoders the control has no dual effect; see [5, Lemma 5.2.1] for more details.

We have shown that for any linear encoder the no dual effect condition holds. We have also shown that no dual effect can hold for certain nonlinear encoders. In particular, it holds for those nonlinear encoders that first remove the effect of the control before applying the nonlinearity. This was done for the noiseless digital channel with information pattern $I_t^A = I_t^B$.

B. Reduction to Fully Observed LQ Model

We now reduce our partially observed control problem into a fully observed LQ control problem. As before let $Y_t = E[X_t \mid B^t, U^{t-1}]$ be the decoder's estimate of the state of

the system and let $\Delta_t = X_t - Y_t$ be the state estimation error. The running cost at time t can be written for every (b^t, u^{t-1})

$$\begin{aligned}
&E[X_t'QX_t + U_t'SU_t \,|\, b^t, u^{t-1}] \\
&\quad = E[(Y_t + \Delta_t)'Q(Y_t + \Delta_t) + U_t'SU_t \,|\, b^t, u^{t-1}] \\
&\quad = E[Y_t'QY_t + U_t'SU_t \,|\, b^t, u^{t-1}] + 2E[Y_t'Q\Delta_t \,|\, b^t, u^{t-1}] \\
&\qquad + E[\Delta_t'Q\Delta_t \,|\, b^t, u^{t-1}] \\
&\quad = E[Y_t'QY_t + U_t'SU_t \,|\, b^t, u^{t-1}] + E[\Delta_t'Q\Delta_t \,|\, b^t, u^{t-1}]
\end{aligned}$$

where the last equation holds because $E[Y_t'Q\Delta_t \,|\, b^t, u^{t-1}] = Y_t'QE[\Delta_t \,|\, b^t, u^{t-1}] = 0$. Note that if the control has no dual effect the last term of the running cost, $E[\Delta_t'Q\Delta_t \,|\, b^t, u^{t-1}]$, will not depend on u^{t-1}.

We now define a new "fully observed" process with the decoder's estimate of the state, Y_t, as the new state

$$\begin{aligned}
Y_{t+1} &= E[X_{t+1} \,|\, B^{t+1}, U^t] \\
&= E[F(Y_t + \Delta_t) + GU_t + W_t \,|\, B^{t+1}, U^t] \\
&= FY_t + GU_t + E[F\Delta_t + W_t \,|\, B^{t+1}, U^t] \\
&= FY_t + GU_t + \bar{W}_t
\end{aligned}$$

with "process disturbance:" $\bar{W}_t = E[F\Delta_t + W_t \,|\, B^{t+1}, U^t]$. Our new system has the dynamics

$$Y_{t+1} = FY_t + GU_t + \bar{W}_t. \tag{7}$$

The term $\bar{W}_t$ represents the new information being transmitted from the encoder to the decoder. To see this, note that $\Delta_{t+1} = X_{t+1} - Y_{t+1} = FX_t + GU_t + W_t - (FY_t + GU_t + \bar{W}_t) = F\Delta_t + W_t - \bar{W}_t$. At time t, the decoder's prediction error in estimating the state at time $t+1$ is $F\Delta_t + W_t$. After receiving B_{t+1}, U_t the error reduces to Δ_{t+1}. The difference, $\bar{W}_t = F\Delta_t + W_t - \Delta_{t+1}$, represents the new information sent to the encoder.

To complete the formulation of the fully observed LQ problem we need to show the process disturbances, $\{\bar{W}_t\}$, in (7) are uncorrelated.

Lemma 3.2: The random variables $\{\bar{W}_t\}$ are uncorrelated.

Proof: We need to show that $E(\bar{W}_t\bar{W}_s') = 0$ for all $s \neq t$. We will prove $E(\bar{W}_t\bar{W}_{t+1}') = 0$. The general case will then follow. First, note that $\bar{W}_t$ is uncorrelated with Δ_{t+1} because the error in estimating $F\Delta_t + W_t$ is uncorrelated with the estimate $\bar{W}_t$. Second, note that $\bar{W}_t$ is uncorrelated with W_{t+1} because the process disturbance at time $t+1$ is independent of everything that occurs before time $t+1$. Now, $E[\bar{W}_t\bar{W}_{t+1}'] = E[\bar{W}_t E[(F\Delta_{t+1} + W_{t+1})' \,|\, B^{t+2}, U^{t+1}]] = E[E[\bar{W}_t(F\Delta_{t+1} + W_{t+1})' \,|\, B^{t+2}, U^{t+1}]] = E[\bar{W}_t(F\Delta_{t+1} + W_{t+1})'] = 0$. Where the second equality follows because $\bar{W}_t$ is measurable with respect to $\sigma(B^{t+1}, U^t)$. □

We can now put all the pieces together. Let $\Lambda_t = \mathrm{cov}(\Delta_t)$ be the covariance of the state estimation error Δ_t. As stated before, the cost is

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T} E\left[\sum_{t=1}^{T} X_t'QX_t + U_t'SU_t\right] \\
&= \limsup_{T\to\infty} \frac{1}{T}\left(\sum_{t=1}^{T} E[Y_t'QY_t + U_t'SU_t] + E[\Delta_t'Q\Delta_t]\right).
\end{aligned} \tag{8}$$

If there is no dual effect then the second term, $\limsup_{T\to\infty}(1/T)(\sum_{t=1}^{T} E[\Delta_t'Q\Delta_t])$, is independent of the control actions chosen. This term represents the cost due to the encoder and channel. We have shown previously that the $\{\bar{W}_t\}$ process is uncorrelated. Therefore, the dynamics (7) with cost given by the first term of (8) is an LQ control problem. Hence, the optimal control law is the certainty equivalent control.

Note that $K_{\bar{W}_t} = F\Lambda_t F' + K_W - \Lambda_{t+1}$. If we further assume that $\Lambda_t = \Lambda$ for all t, then we can directly apply the results described in (4)–(6) to get

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T} E\left[\sum_{t=1}^{T} X_t'QX_t + U_t'SU_t\right] \\
&\quad = \mathrm{tr}(PK_{\bar{W}}) + \mathrm{tr}(Q\Lambda) \\
&\quad = \mathrm{tr}(P(F\Lambda F' + K_W - \Lambda)) + \mathrm{tr}(Q\Lambda) \\
&\quad = \mathrm{tr}(PK_W) + \mathrm{tr}((F'PF - P + Q)\Lambda).
\end{aligned}$$

We see that the optimal cost decomposes into two terms. The first term is the cost under full state observation and the second term is a cost that depends only on Λ the steady state estimation error covariance. Thus, we have reduced the problem of computing the optimal cost to that of minimizing $\mathrm{tr}((F'PF - P + Q)\Lambda)$ for a given information pattern and channel.

We have presented conditions under which the certainty equivalent controller is optimal. In the next section, we discuss the encoder and decoder design.

IV. Sequential Rate Distortion

We first review traditional rate distortion theory and the concept of "separation" between the source encoder and the channel encoder. We then describe the failings of the traditional theory. These failings have to do with the long delays, noncausality, and the failure of separation to hold for traditional rate distortion theory in sequential settings. To correct for these failings we review the notion of source-channel matching [19], [12], and introduce the SRD framework [13], [23].

Given a source $\{P(dX_t \,|\, x^{t-1})\}$ and a channel $\{P(dB_t \,|\, a_t)\}$ our job is to design an encoder and decoder that can transmit the source over the channel while maintaining a given end-to-end distortion criterion; see Fig. 2. For our source, we consider the Gauss–Markov system (1) with the controls $\{U_t\}$ set to zero

$$X_{t+1} = FX_t + W_t, \qquad t \geq 0. \tag{9}$$

We will focus on the *weighted squared error distortion measure*. For $x, y \in \mathbb{R}^d$, let

$$d_M(x, y) = \|x - y\|_M^2 = (x - y)'M(x - y)$$

where the weight matrix M is positive definite.

The channel capacity theorem [11, Th. 5.6.4] states that any rate $R < C^{\mathrm{cap}}$ can be transmitted over the channel with probability of decoding error decreasing to zero exponentially with T the number of channel uses. Specifically, there exists a channel encoder that can transmit 2^{TR} messages, equivalently TR bits, by using the channel T times and incurring a small probability of decoding error.

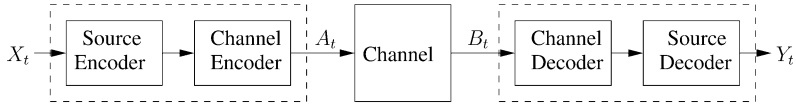


Fig. 2. Source-channel coding.

The source-channel separation theorem gives us conditions under which we can separate the encoder into two pieces: a source encoder that encodes the source into bits and a channel encoder that converts those bits into channel symbols. Similarly the decoder is split into two pieces: a channel decoder and a source decoder; see Fig. 2.

The *rate distortion function* for horizon T is $R_T(D) \doteq \inf_{P\in\mathcal{F}_T}(1/T)I(X^{T-1};Y^{T-1})$ where $\mathcal{F}_T = \{\{P(dY_t|x^{T-1},y^{t-1})\}_{t=0}^{T-1}\colon E[(1/T)\sum_{t=0}^{T-1} d_M(X_t,Y_t)] \leq D\}$.

The next theorem follows from [11, Th. 9.6.3].

Theorem 4.1: We are given a Gauss–Markov process as described in (9), $D > 0$, and a channel with capacity C^{cap}. If $R_T(D) > C^{\text{cap}}$, then there does not exist a coding scheme that achieves distortion D over T channel uses. If $R_T(D) < C^{\text{cap}}$ for T sufficiently large then there exists a coding scheme that can achieve a distortion arbitrarily close to D.

This theorem shows that there exists a coding scheme that achieves the rate distortion bound. Specifically, we can construct a source encoder that quantizes our source into $R_T(D)$ bits. We can then design independently a channel encoder that will transmit these $R_T(D)$ bits over a noisy channel with small probability of channel decoding error. Hence, one can design the source encoder and the channel encoder separately so as to achieve the required end-to-end distortion [11].

There is another coding scheme that can sometimes achieve the end-to-end distortion bound over a channel. Note that the definition of the rate distortion function involves an infimization over stochastic kernels $\{P(dY_t\,|\,x^{T-1},y^{t-1})\}$. These stochastic kernels can be viewed as channels connecting the source $\{X_t\}$ to the reproduction $\{Y_t\}$. If the true communication channel equals the rate distortion infimizing channel, then we say the channel is *matched* to the source [19], [12]. Specifically, for all t, we have $A_t = X_t, B_t = Y_t$, and $P(dB_t\,|\,a^{T-1},b^{t-1}) = P(dY_t\,|\,x^{T-1},y^{t-1})$. Note that the communication channel may need to have memory.

In the source-channel matching case, we do not need to separately design the source encoder and channel encoder. In fact, we can send the source over the channel in an uncoded fashion. Here, there is no explicit separation between the source and channel encoding.

A. Causality and Sequential Rate Distortion

Unfortunately, under the source-channel coding scheme, there is a delay of at least $2T-t-1$ time units in computing the reconstruction Y_t of X_t. Specifically, at time t, we observe X_t. It then takes $T-1-t$ steps to observe the remainder of the source, $X_{t+1},\ldots,X_{T-1}$. At this point the channel input symbols, A^{T-1}, are produced. It takes another T time steps until all the channel output symbols B^{T-1} are received. Then the decoder can compute the estimate Y_t of X_t. The time-ordering is

$$X_0,\ldots,X_{T-1},Y_0,\ldots,Y_{T-1}. \tag{10}$$

In our control situation, this delay is unacceptable. The natural *causal* time-ordering we require is

$$X_0,Y_0,X_1,\ldots,X_{T-1},Y_{T-1}. \tag{11}$$

In time-ordering (10), the random variable Y_0 is the $T+1$th event whereas in time-ordering (11) Y_0 is the second event.

Similarly, in the source-channel matching scheme there is a delay. Note that the rate distortion infimizing stochastic kernel in Theorem 4.1 can be viewed as a channel that factors: $P(dY^{T-1}\,|\,x^{T-1}) = \prod_{t=0}^{T-1} P(dY_t\,|\,x^{T-1},y^{t-1})$. Under the time-ordering (10) this is a causal channel. Under the time-ordering (11), this is a noncausal channel because the reconstruction Y_t depends on future X's.

We now define the *sequential rate distortion function.* This function involves an optimization of the mutual information over all causal, with respect to time-ordering (11) channels.

Definition 4.1: The *sequential rate distortion function* is

$$R_T^{\text{SRD}}(D) = \inf_{P\in\mathcal{F}_T^{\text{SRD}}} \frac{1}{T} I(X^{T-1};Y^{T-1})$$

with $\mathcal{F}_T^{\text{SRD}} = \{\{P(dY_t\,|\,x^t,y^{t-1})\}_{t=0}^{T-1}\colon E[d_M(X_t,Y_t)] \leq D, \forall t\}$. Note that in the definition we are expected to achieve a distortion $E[d_M(X_t,Y_t)] \leq D$ at each time step t.

We provide a necessary condition on the channel capacity to achieve a given end-to-end distortion causally. To that end, we state the data-processing inequality [9].

Lemma 4.1: Let $X \to A \to B \to Y$ be a Markov chain. Then, $I(X;Y) \leq I(A;B)$.

Proposition 4.1: A necessary condition on a given memoryless channel with capacity C^{cap} to achieve the distortion D causally as described in Definition 4.1 is $R_T^{\text{SRD}}(D) \leq C^{\text{cap}}$.

Proof: Let $P(dX^{T-1},dA^{T-1},dB^{T-1},dY^{T-1})$ be any joint measure which satisfies the distortion and channel requirements. Then

$$\begin{aligned} R_T^{\text{SRD}}(D) &\leq \frac{1}{T} I(X^{T-1};Y^{T-1}) \\ &\leq \frac{1}{T} I(A^{T-1};B^{T-1}) \leq C^{\text{cap}} \end{aligned}$$

where the second inequality follows from the data-processing inequality. □

Finding sufficient conditions on the channel to achieve the SRD end-to-end distortion is more difficult. The source-channel separation principle does not generally hold in situations where delay is an issue. But there do exist channels with capacity $C^{\text{cap}} = R_T^{\text{SRD}}(D)$ over which an end-to-end distortion D can

be achieved causally [23]. This can occur when the communication channel is matched to the sequential rate distortion infimizing channel $\{P(Y_t \,|\, x^t, y^{t-1})\}$.

Definition 4.2: We say a channel $\{P(dB_t \,|\, a^t, b^{t-1})\}$ is *matched* to the SRD infimizing channel $\{P(dY_t \,|\, x^t, y^{t-1})\}$ if there exists an encoder $\{Q(dA_t \,|\, x^t, a^{t-1}, b^{t-1})\}$ and a decoder $\{Q(dY_t \,|\, b^t, y^{t-1})\}$ such that

$$P(Y_t \,|\, X^t, Y^{t-1}) = Q(Y_t \,|\, X^t, Y^{t-1}) \quad Q-a.s.$$

Here, $Q(dX^{T-1}, dA^{T-1}, dB^{T-1}, dY^{T-1})$ is the joint measure determined by interconnecting the source $P(dX^{T-1})$, the encoder $\{Q(dA_t \,|\, x^t, a^{t-1}, b^{t-1})\}$, the channel $\{P(dB_t \,|\, a^t, b^{t-1})\}$, and the decoder $\{Q(dY_t \,|\, b^t, y^{t-1})\}$.

In words, a communication channel is matched to the SRD infimizing channel if we can find an encoder and decoder for this communication channel such that the source-to-reconstruction behavior behaves like the SRD infimizing channel.

B. Computations

Here, we compute the sequential rate distortion function and determine the structure of the infimizing channel law for the Gauss–Markov source described in (9). We first review the solution for a single Gaussian source. Then, we describe the sequential situation.

1) Gaussian Source: Let our source be $X \sim \mathcal{N}(0, K_X)$. First, we present a simplifying lemma.

Lemma 4.2: Let $\tilde{X} = M^{1/2}X, \tilde{Y} = M^{1/2}Y$. Then, 1) $I(\tilde{X};\tilde{Y}) = I(X;Y)$, and 2) $\|\tilde{x} - \tilde{y}\|_I^2 = \|x - y\|_M^2$ where I is the identity matrix. Let U be the unitary matrix that diagonalizes $K_X = U'\Lambda U$. Let $\hat{X} = UX, \hat{Y} = UY$. Then, 3) $I(\hat{X};\hat{Y}) = I(X;Y)$, and 4) $\|\hat{x} - \hat{y}\|_I^2 = \|x - y\|_I^2$.

Proof: (1) holds because mutual information is invariant under injective transformations (see Appendix.) To see (2) note $\|\tilde{x} - \tilde{y}\|_I^2 = (M^{1/2}x - M^{1/2}y)'(M^{1/2}x - M^{1/2}y) = (x - y)'M(x-y) = \|x-y\|_M^2$. Both (3) and (4) hold because mutual information and squared error distortion with weight matrix I are invariant under unitary transformations. □

Thus, without loss of generality we can restrict our attention to the case where $M = I$ and the source covariance K_X is diagonal. In practice, the encoder would preprocess the observation by applying $M^{1/2}$ and U to it. Similarly the decoder would postprocess its output by $M^{-(1/2)}$ and U'.

Let $K_X = \mathrm{diag}[\lambda_1, \ldots, \lambda_d]$ and $M = I$. Then, [4], [9]

$$R(D) = \frac{1}{2}\sum_{i=1}^{d} \log \frac{\lambda_i}{\delta_i} \quad \text{where}$$
$$\delta_i = \begin{cases} \eta, & \text{if } \eta \le \lambda_i \\ \lambda_i, & \text{if } \eta > \lambda_i \end{cases} \tag{12}$$

where η is chosen so that $\sum_{i=1}^{d} \delta_i = D$. This is the so called water-filling solution. It is useful to view $R_i \doteq (1/2)\log(\lambda_i/\delta_i)$ as the rate, or channel capacity, allocated to reconstructing the ith component of X.

The error covariance is given by

$$\Lambda_{X|Y} = \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_d \end{bmatrix} = K_X \begin{bmatrix} 2^{-2R_1} & & \\ & \ddots & \\ & & 2^{-2R_d} \end{bmatrix}.$$

For $(D/d) \le \min_i \lambda_i$ the formula (12) reduces to $R(D) = (d/2)\log(d|\Lambda|^{1/d}/D)$. Specifically, the distortion accrued on each component is the same: $\Lambda_{X|Y} = \mathrm{diag}[(D/d), \ldots, (D/d)]$. The rate R_i used for each component may be different though.

We now characterize the infimizing channel. Following [11], we first define the *backward channel*, $P(dX \,|\, y)$, before defining the *forward channel*, $P(dY \,|\, x)$. The backward channel is given by $X = y + \Delta$ where the error Δ is zero mean with covariance $\Lambda_{X|Y}$. The channel output Y is Gaussian with zero mean and covariance $K_X - \Lambda_{X|Y}$. Thus, we can compute the joint Gaussian measure $P(dX, dY) = P(dX \,|\, y)P(dY)$. The forward channel is given by $Y = Hx + Z$ where $H = E(YX')E(XX')^{-1}$ and Z is a zero mean Gaussian vector with covariance $E(YY') - E(YX')E(XX')^{-1}E(XY')$. The optimal channel has the property that $E[X \,|\, Y] = Y$.

There are many ways to *realize* the channel $Y = Hx + Z$. One such way is to let Γ be an invertible transformation such that $\Gamma H\Gamma^{-1} = I$. Let $B = \Gamma Y, A = \Gamma X$ and $N = \Gamma Z$. For this case, we say the channel $B = a + N$ with power constraint $E[A'A] \le \mathrm{tr}(\Gamma K_X \Gamma')$ is *matched* to the source X.

2) SRD for Gauss-Markov Sources: We now compute the sequential rate distortion function for the Gauss–Markov source described in (9). We first present some results on the structure of the optimal sequential rate distortion infimizing channel in Definition 4.1. In the previous section, we showed that the static Gaussian source is matched to a memoryless Gaussian channel. Here, we show that the source (9) is matched to a Gaussian channel with memory. Furthermore, this Gaussian channel can be realized over a memoryless channel with feedback [23].

See the Appendix for the proof of the following.

Lemma 4.3: The optimal SRD infimizing channel for the Gauss–Markov source (9) is a Gaussian channel of the form $P(dY_t \,|\, x_t, y^{t-1})$.

This lemma states that the channel output at time t only depends on the current channel input x_t and not on the previous channel inputs. Specifically, it has the form

$$Y_t = \alpha_t x_t + \beta_t y^{t-1} + Z_t \tag{13}$$

where $\alpha_t \in \mathbb{R}^{d\times d}$, $\beta_t \in \mathbb{R}^{d\times(t-1)d}$, and $\{Z_t\}$ is an independent sequence of Gaussian random variables. This channel has memory due to its dependence on the past channel outputs.

This channel can be realized over a memoryless Gaussian channel with feedback. To see this, note

$$Y_t = \alpha_t(x_t - E[X_t \,|\, y^{t-1}]) + \alpha_t E[X_t \,|\, y^{t-1}] + \beta_t y^{t-1} + Z_t.$$

The channel outputs y^{t-1} are available to the encoder via feedback. The encoder first computes the scaled innovation $A_t = \alpha_t(x_t - E[X_t \,|\, y^{t-1}])$ and then transmits A_t over the channel $B_t = a_t + Z_t$. The decoder upon receiving B_t can then compute Y_t. The memoryless "A-B" channel has power constraint: $P_t = E(A_t'A_t) = \mathrm{trace}(\alpha_t \Lambda_{X_t|Y^{t-1}} \alpha_t')$ where $\Lambda_{X_t|Y^{t-1}} = E[(X_t - E[X_t|y^{t-1}](X_t - E[X_t|y^{t-1}])']$ is the error covariance in estimating X_t from Y^{t-1}.

For this realization, the total mutual information separates into a sum of mutual information terms: $I(X^{T-1}, Y^{T-1}) = \sum_{t=0}^{T-1} I(A_t; B_t)$. To see this, first note that $I(X^{T-1}, Y^{T-1}) =$

$\sum_{t=0}^{T-1} I(X^t; Y_t|Y^{t-1}) = \sum_{t=0}^{T-1} I(X_t; Y_t|Y^{t-1})$. The second equality follows because we can restrict ourselves to channels of the form (13). Now, each addend can be written as

$$\begin{aligned} I(X_t; Y_t|Y^{t-1}) &= h(Y_t|Y^{t-1}) - h(Y_t|X_t, Y^{t-1}) \\ &= h(B_t|Y^{t-1}) - h(B_t|A_t, Y^{t-1}) \\ &= h(B_t) - h(B_t|A_t) = I(A_t; B_t). \end{aligned}$$

The second equality comes from the fact differential entropy is invariant under translations. The third equality follows by noting A_t is independent of Y^{t-1}. Since $B_t = A_t + Z_t$ we may conclude that B_t is independent of Y^{t-1}.

We are now ready to compute the SRD function and the SRD infimizing channel for the Gauss–Markov source (9). We first address the scalar case. Let $\{B_t = a_t + V_t\}$ be a sequence of scalar memoryless Gaussian channels that realize, under noiseless feedback, the sequential rate distortion infimizing channel. Let $R(t) = I(A_t; B_t)$ be the mutual information for the tth channel use.

As we have shown, one can reconstruct a scalar Gaussian source with variance Λ over a matched Gaussian channel of capacity R within a distortion $\Lambda 2^{-2R}$. For the scalar Gauss–Markov source, we know that the encoder first computes the innovation: $X_t - E(X_t|Y^{t-1}) = FX_{t-1} + W_{t-1} - FY_{t-1} = F\Delta_{t-1} + W_{t-1}$ where the error is $\Delta_t = X_t - Y_t$. It then scales this innovation and transmits it over the "$A_t - B_t$" channel. The distortion at time t is $D_t = E(\Delta_t^2)$. Hence, the variance of the innovation is $F^2 D_{t-1} + K_W$. Then, the reconstruction has distortion

$$\begin{aligned} &D_t = (F^2 D_{t-1} + K_W) 2^{-2R(t)} t \geq 1 \quad \text{and} \\ &D_0 = K_{X_0} 2^{-2R(0)}. \end{aligned}$$

To compute $R_T^{\mathrm{SRD}}(D)$ we require each $D_t = D$ hence $R(0) = \max\{0, (1/2)\log(K_{X_0}/D)\}$ and $R(t) = \max\{0, (1/2)\log(F^2 + (K_W/D))\}$ for $t \geq 1$. Now $R_T^{\mathrm{SRD}}(D) = (1/T)\sum_{t=0}^{T-1} R(t)$. Thus

$$\lim_{T\to\infty} R_T^{\mathrm{SRD}}(D) = \max\left\{0, \frac{1}{2}\log\left(F^2 + \frac{K_W}{D}\right)\right\}. \quad (14)$$

If the source is unstable $|F| > 1$, then by Proposition 4.1 and (14) a necessary condition to insure a bounded distortion at each time step is that the channel capacity be at least $\log|F|$.

We now treat the d-dimensional vector valued Gauss–Markov source. Let $\Lambda_t = E(\Delta_t \Delta_t')$ be the covariance of the distortion error Δ_t. At time t the encoder first computes the innovation, $F\Delta_{t-1} + W_t$ which has covariance $F\Lambda_{t-1}F' + K_W$. Let Φ_t be the unitary matrix that diagonalizes $F\Lambda_{t-1}F' + K_W$. That is $\Phi_t(F\Lambda_{t-1}F' + K_W)\Phi_t' = \mathrm{diag}[\lambda_1(t), \ldots, \lambda_d(t)]$. By our discussion, for the single Gaussian source we see that for a distortion D we need a rate: $\sum_{i=1}^d R_i(t) = (1/2)\sum_{i=1}^d \log(\lambda_i(t)/\delta_i(t))$ where

$$\delta_i(t) = \begin{cases} \eta(t) & \text{if } \eta(t) \leq \lambda_i(t) \\ \lambda_i(t) & \text{if } \eta(t) > \lambda_i(t) \end{cases} \quad (15)$$

and $\eta(t)$ is chosen so that $\sum_{i=1}^d \delta_i(t) = D$. The error covariance at time t is then given by

$$\Lambda_t = \Phi_t' \Phi_t (F\Lambda_{t-1}F' + K_W)\Phi_t' \times \begin{bmatrix} 2^{-2R_1(t)} & & \\ & \ddots & \\ & & 2^{-2R_d(t)} \end{bmatrix} \Phi_t \quad (16)$$

and that $\Phi_t \Lambda_t \Phi_t' = \mathrm{diag}[\delta_1(t), \ldots, \delta_d(t)]$. The interpretation of (16) is as follows. First the covariance of the innovation $F\Lambda_{t-1}F' + K_W$ is diagonalized. Each component is then transmitted at a rate of $R_i(t)$. The decoder receives the channel output and then computes the state estimate. This state estimate has error covariance given by Λ_t.

Note that in (15), for low enough distortion, specifically $(D/d) \leq \min_i \lambda_i((D/d)FF' + K_w)$, we get for each t that $\Phi\Lambda_t\Phi' = \mathrm{diag}[(D/d), \ldots, (D/d)]$ where Φ diagonalizes $(D/d)FF' + K_W$. Here, $\lambda_i((D/d)FF' + K_w)$ represents the ith eigenvalue of $(D/d)FF' + K_w$. Thus, (16) becomes

$$\begin{aligned} \Phi\Lambda_t\Phi' &= \frac{D}{d} I \\ &= \Phi\left(\frac{D}{d}FF' + K_W\right)\Phi' \begin{bmatrix} 2^{-2R_1(t)} & & \\ & \ddots & \\ & & 2^{-2R_d(t)} \end{bmatrix}. \end{aligned}$$

Hence, the rate required on the ith component is

$$\begin{aligned} R_i(t) &= \frac{1}{2}\log\frac{\lambda_i\left(\frac{D}{d}FF' + K_W\right)}{\frac{D}{d}} \\ &= \frac{1}{2}\log\lambda_i\left(FF' + \frac{d}{D}K_W\right). \end{aligned}$$

The total rate required at time t is

$$\sum_{i=1}^d \frac{1}{2}\log\lambda_i\left(FF' + \frac{d}{D}K_W\right) = \frac{1}{2}\log\left|FF' + \frac{d}{D}K_W\right|.$$

We have reduced the sequential rate distortion problem for the Gauss–Markov process to a standard rate distortion problem for a single Gaussian source. For D small enough to satisfy the previous condition, we have

$$\lim_{T\to\infty} R_T^{\mathrm{SRD}}(D) = \frac{1}{2}\log\left|FF' + \frac{d}{D}K_W\right|. \quad (17)$$

Note that $(1/2)\log|FF' + (d/D)K_W| \geq \sum_{\lambda(F)} \max\{0, \log|\lambda(F)|\}$. This result relating the channel rate to the eigenvalues of the open-loop system is related to the stability requirements presented in [21]. In particular, $\sum_{\lambda(F)} \max\{0, \log|\lambda(F)|\}$ is the minimum rate required to insure bounded state estimation error.

For the Gauss–Markov source we have shown that if the communication channel is matched to the SRD infimizing channel then the SRD distortion can be achieved. In many situations the communication channel is not matched to the SRD infimizing channel. In the case when the channel is Gaussian but not matched to the source, Lee and Peterson [14] derive the optimal linear encoder and decoder. For the unmatched case, though, the optimal encoder and decoder are generally

nonlinear and difficult to characterize. The problem of joint source-channel coding with small latencies is currently an active area of research.

C. Noiseless Digital Channel and Operational SRD

We have computed the SRD function for the Gauss–Markov source. We have also shown that if the communication channel is a memoryless Gaussian channel with noiseless feedback matched to the source then the SRD function can be achieved. Here we examine the situation where the communication channel is a digital noiseless channel and hence not matched to the source. In this case the encoder must quantize the information it wants to transmit (as opposed to scaling and rotating as it did in the case of the Gaussian channel with noiseless feedback.)

Definition 4.3: A *sequential rate distortion quantizer* is a sequence of measurable functions q_t such that q_t: $\mathbb{R}^{(t+1)\times d} \times \mathbb{R}^{t\times d} \to \mathbb{R}^d$ where the range of each function is at most countable. Specifically q_t takes $(x^t, y^{t-1}) \mapsto y_t$.

In the following, the superscript "o" represents "operational."

Definition 4.4: The *operational sequential rate distortion* is

$$R_T^{\mathrm{SRD},o}(D) = \inf_{(q_0,\ldots,q_{T-1})\in\mathcal{F}_T^o} \frac{1}{T} H(Y^{T-1})$$

where $\mathcal{F}_T^o = \{(q_0,\ldots,q_{T-1})\colon E[d_M(X_t,Y_t)] \le D, \forall t\}$.

It is easy to see that $R_T^{\mathrm{SRD},o}(D) \ge R_T^{\mathrm{SRD}}(D)$. Specifically let $q_0,\ldots q_{T-1}$ be any sequential quantizer such that $E[d_M(X_t,Y_t)] \le D, \forall t$. Then $(1/T)H(Y^{T-1}) \ge (1/T)I(X^{T-1};Y^{T-1}) \ge R_T^{\mathrm{SRD}}(D)$.

Computing $R_T^{\mathrm{SRD},o}$ can be difficult. However, some structural properties are known. In [8], it is shown that the optimal quantizer for the Gauss–Markov source is itself Markov. Unfortunately, designing optimal sequential quantizers is difficult. In the limit of high rate (low distortion) more can be said. The quantizer can be chosen to be a uniform quantizer (see Appendix.) Also, the difference between the operational SRD function and the SRD function becomes negligible. Thus, the SRD function is a useful approximate measure for the rate required to achieve a given distortion [23], [16]. See the Appendix for a proof of the following.

Proposition 4.2: $\lim_{D\to 0}(R_T^{\mathrm{SRD},o}(D)/R_T^{\mathrm{SRD}}(D)) = 1\ \forall T$.

In summary, $R_T^{\mathrm{SRD},o}(D) \ge R_T^{\mathrm{SRD}}(D)$ with the difference becoming negligible as $D \to 0$. The sequential rate distortion function R_T^{SRD} can be considered a relaxation of the operational SRD function where the relaxation comes from extending the class of deterministic quantizers to the class of random quantizers. In the SRD framework the random quantizers are represented by the infimizing channel laws. The randomization can be viewed as coming from the noise in the infimizing channel law. Hence for the matched case the randomization comes from the true channel noise.

V. Examples

In this section, we apply the results of Sections III and IV to four different scenarios. We compute or bound the cost in each scenario.

A. Noiseless Digital Channel

As stated before, for the noiseless digital channel, the information patterns I_t^A and I_t^B are equivalent. For these two information patterns let $R_T^{\mathrm{SRD},o}(D)$ be the operational SRD function for the uncontrolled dynamics $X_{t+1} = FX_t + W_t$ with weight matrix $M = F'PF - P + Q$.

Recall that the optimal LQG cost has the form: $\mathrm{tr}(PK_W) + \mathrm{tr}(M\Lambda)$. The cost $\mathrm{tr}(PK_W) + D$ can be achieved over a noiseless digital channel with rate $\limsup_{T\to\infty} R_T^{\mathrm{SRD},o}(D)$. As discussed earlier $\lim_{D\to 0} R_T^{\mathrm{SRD},o}(D) = R_T^{\mathrm{SRD}}(D)$, hence, $\limsup_{T\to\infty} R_T^{\mathrm{SRD}}(D)$ is a good approximation to the channel rate needed in the low distortion regime.

There is a tradeoff between the channel rate and the control performance. Notice that if we are interested in engineering the total cost to within some percentage of the optimal value, then beyond a certain point it is no longer worth trying to lower D by increasing the quality of communication since the full knowledge cost, $\mathrm{tr}(PK_W)$, will dominate the total cost.

B. Memoryless Gaussian Channel With Information Pattern A

In this case, the encoder has access to the past channel outputs. There are many ways to realize this in practice. One possibility is that there is a direct noiseless feedback path for the memoryless Gaussian channel. Another possibility is that encoder can infer the channel output from the observation of the previous control signals and knowledge of the controller policy and decoder policy. The encoder also has access to the previous controls as well.

Let $R_T^{\mathrm{SRD}}(D)$ be the SRD function for the uncontrolled process $X_{t+1} = FX_t + W_t$ with weight $M = F'PF - P + Q$. Then the LQG cost $\mathrm{tr}(PK_W) + D$ can be achieved over a memoryless Gaussian channel if that Gaussian channel is matched to the SRD infimizing channel law.

Following Lemma 4.2, we can redefine the dynamics. Let $\tilde{X}_t = M^{1/2}X_t, \tilde{W}_t = M^{1/2}W_t$, and $\tilde{F} = M^{1/2}FM^{-(1/2)}$. Then $\tilde{X}_{t+1} = \tilde{F}\tilde{X}_t + \tilde{W}_t$.

We first examine the scalar case. For Gaussian channels with capacity $R > \max\{0, \log|\tilde{F}|\}$, we have shown that the steady state distortion is $D = (K_{\tilde{W}}/2^{2R} - \tilde{F}^2) = (K_W(F^2P - P + Q)/2^{2R} - F^2)$. Thus, the optimal LQG cost equals

$$PK_W + \frac{K_W(F^2P - P + Q)}{2^{2R} - F^2}.$$

Note that if $R < \log|F|$ then the LQG cost equals infinity.

For the vector case, we can get an explicit solution in the low distortion regime. To achieve an LQG cost of $\mathrm{tr}(PK_W) + D$, we need a memoryless Gaussian channel matched to the source with capacity $\lim_{T\to\infty} R_T^{\mathrm{SRD}}(D)$. Recall from before that for D small enough we have

$$\begin{aligned}\lim_{T\to\infty} R_T^{\mathrm{SRD}}(D) &= \frac{1}{2}\log\left|\tilde{F}\tilde{F}' + \frac{d}{D}K_{\tilde{W}}\right| \\ &= \frac{1}{2}\log|M||FM^{-1}F' + \frac{d}{D}K_W|.\end{aligned}$$

In the case when the Gaussian channel is not matched to the source the expression $\mathrm{tr}(PK_W) + D$ is a lower

bound on the optimal LQG cost for a channel with capacity $(1/2)\log|M||FM^{-1}F' + (d/D)K_W|$.

C. Memoryless Gaussian Channel With Information Pattern B

In this case, the encoder has access to the past controls. We are assuming that the encoder is collocated with the plant and, hence, can observe both the state and the control actions applied to the plant.

Computing the SRD function for this information pattern is difficult because encoder does not know the decoder's state. One potential suboptimal encoding scheme consists of first computing the encoder's innovation and then transmitting that across the Gaussian channel. This *innovations coding* scheme was introduced in [6].

At time t the encoder computes the innovation using the encoder's information: $X_t - E(X_t \,|\, A^{t-1}, U^{t-1})$. Specifically, $X_0 - E(X_0) = X_0$ and for $t \geq 1$ we have $X_t - E(X_t \,|\, A^{t-1}, U^{t-1}) = W_t$. Note that the innovation computed by the encoder is different than the innovation computed by the decoder: $X_t - E(X_t \,|\, B^{t-1}, U^{t-1})$. This is because the encoder does not have access to the channel outputs B^{t-1}.

This innovation coding scheme only works for systems with a stable F matrix. To see this, we examine the scalar case. Let the Gaussian channel have capacity R. Then, at time zero, the estimation error at the decoder is $\Lambda_0 = K_{X_0}2^{-2R}$. For general times t, the decoder upon receiving the channel output B_t will compute the estimate of the state as follows: $Y_t = FY_{t-1} + GU_t + E(W_t \,|\, b_t)$. Hence, the error variance satisfies the recursion $\Lambda_{t+1} = F^2\Lambda_t + K_W 2^{-2R}$. If F is unstable then $\lim_{t\to\infty}\Lambda_t = \infty$. In the case when F is stable we can compute the steady state error variance: $\Lambda = (K_W 2^{-2R}/1 - F^2)$. Hence, the LQG cost is

$$PK_W + \frac{(F^2P - P + Q)K_W 2^{-2R}}{1 - F^2}.$$

Note that even over a noiseless digital channel this innovations coding scheme will not be stable if F is unstable. This is because the innovations W_t are quantized. This quantization error leads to errors in the decoder's state estimate. The new information sent to the decoder will never correct these previous errors. Because F is unstable, this error will grow.

We now examine the vector case. At each time step we only transmit information about W_t. This is nothing more than the standard rate distortion problem discussed before. Let $\hat{W}_t$ be the decoder's reconstruction of W_t and let $\tilde{W}_t = W_t - \hat{W}_t$ be the error in reconstruction. Let Φ be a unitary matrix that diagonalizes K_W. Specifically, $\Phi K_W \Phi' = \mathrm{diag}[\lambda_1(K_W), \ldots, \lambda_d(K_W)]$. Let $R_1, \ldots, R_d$ be a choice of rates for each component. Then

$$K_{\tilde{W}} = \Phi' \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_d \end{bmatrix} \Phi$$
$$= \Phi'\Phi K_W \Phi' \begin{bmatrix} 2^{-2R_1} & & \\ & \ddots & \\ & & 2^{-2R_d} \end{bmatrix} \Phi$$

where $\delta_i = \lambda_i(K_W)2^{-2R_i}$. If all the $\delta_i = (D/d)$ then error covariance $K_{\tilde{W}} = (D/d)I$. The rate required to achieve this distortion for the reconstruction of W_t is given by $R(D) = (d/2)\log(d|K_W|^{1/d}/D)$.

The state estimation error covariance evolves as: $\Lambda_{t+1} = F\Lambda_t F' + K_{\tilde{W}}$. For F stable this Lyapunov equation has the steady state solution $\Lambda = \sum_{t\geq 0} F^t K_{\tilde{W}}(F')^t$. In the case when all the $\delta_i = (D/d)$ we get $\Lambda = (D/d)\sum_{t\geq 0} F^t(F')^t$. See [6] for more details on this innovations encoding scheme including a discussion on coding versus delay.

For the noiseless digital channel with information pattern A) and B) and for the memoryless Gaussian channel with information pattern A) we are able to treat unstable systems. However, for the memoryless Gaussian channel with information B), we are not able to treat unstable systems. This difference arises for information pattern B) because the encoder is using different information than the decoder's information to base its decisions. This asymmetry in information can cause the filters employed by the encoder and the decoder to lose synchronization with each other. In the stable case such errors in synchronization will decay away. However, in the unstable case they can grow in an unbounded fashion. The problem with encoder's employing information pattern B) and transmitting over a noisy channel is that the encoder is not made aware of the channel noise corrupting the signal received by the decoder.

The failure of innovations coding in the unstable case does not mean that it is impossible to design control systems faced with information pattern B). More complicated coding is necessary to insure synchrony between the encoder and decoder. It is unclear what the fundamental penalty is in terms of rate for this situation.

For the cases treated so far, the channel input symbols, A_t, transmitted over the channel have been uncorrelated. This is consistent with the source coding idea that we should only transmit new information. But due to channel noise the encoder may lose synchronization with the decoder. Hence, the natural idea is to add error correction, or redundancy, to the channel coding. In the next section, we discuss this in the context of information pattern C).

D. Memoryless Gaussian Channel With Information Pattern C

Here, the encoder's information pattern at time t consists only of X_t. The only linear encoder is one that scales X_t and transmits it across the channel: $B_t = a_t + Z_t$ where $A_t = \alpha_t X_t$ for some gain α_t. Note that this is the standard LQG problem with linear observation equation except that the α_t's need to be chosen so as to satisfy the channel power constraint.

In the previous settings, the channel input symbols $\{A_t\}$ have been uncorrelated. They represented "new" information about the source. We saw that for information pattern B) if the system is unstable and there is channel noise then we can lose synchronization due to the fact that the encoder no longer knows the decoder's state. In such settings, it makes sense to add redundancy to the channel input symbols $\{A_t\}$. For the classical LQG problem the input symbols $A_t = \alpha_t X_t$ are correlated because the $\{X_t\}$ process is correlated.

Assume we have (1) with a linear observation equation of the form $B_t = \alpha X_t + V_t$ where (F, α) is an observable pair. As is well known, the optimal control law is the certainty equivalent control law as described in (4)–(6). Let K_X be the steady state covariance of the process $\{X_t\}$ under this control law. Then, in

steady–state the capacity of the memoryless Gaussian channel should be at least $(1/2)\log(|\alpha K_X \alpha' + K_Z|/|K_Z|)$.

We have a new interpretation for the classical LQG problem with linear observation model. The linear observation model adds redundancy to the information about the innovation to combat channel noise. This is required because the encoder is not able to compute the decoder's state estimate.

APPENDIX

The mutual information between two random variables X and Y is defined as

$$I_{P_{X,Y}}(X;Y) \doteq \begin{cases} \int \log \frac{dP}{dQ} dP, & \text{else} \end{cases}$$

where $Q_{X,Y} = P_X \times P_Y$ and (dP/dQ) is the Radon–Nikodym derivative.

If X is a discrete random variable then its entropy is defined as: $H(X) \doteq -\sum_i P(X = x_i) \log P(X = x_i)$ and its conditional entropy is defined as: $H(X|Y) \doteq -\int (\sum_i P(X = x_i|y) \log P(X = x_i|y)) p(dy)$. If X is a random variable admitting a density, p_X, then its differential entropy is defined as: $h(X) \doteq -\int p_X(x) \log p_X(x)\, dx$ and its conditional differential entropy is defined as: $h(X|Y) \doteq -\int (\int p_{X|Y}(x|y) \log p_{X|Y}(x|y)\, dx) p(dy)$.

The following useful properties can be found in [9].

a)

$$I(X;Y) \geq 0 \quad \text{and} \quad I(X;Y) = \begin{cases} H(Y) - H(Y|X), & \text{if } Y \text{ is a discrete random variable} \\ h(Y) - h(Y\,|\,X), & \text{if } Y \text{ admits a density for each } x. \end{cases}$$

This implies conditioning reduces entropy.

b) $I(X;Z\,|\,Y) = I(X,Y;Z) - I(Y;Z)$.

c) $X \to Y \to Z$ is a Markov chain if and only if $I(X;Z\,|\,Y) = 0$.

d) Mutual information is invariant under injective transformations. In particular, let f be an injective function. Then, $I(X_1;X_2) = I(f(X_1);f(X_2))$.

e) The differential entropy of a d-dimensional Gaussian random variable $X \sim N(0,K)$ is $h(X) = (1/2)\log(2\pi e)^d |K|$ where $|K|$ is the determinant of K.

f) The differential entropy is invariant under translations $h(X + c) = h(X)$.

Lemma 4.3: The optimal SRD infimizing channel for the Gaussian source (9) is a Gaussian channel of the form $P(dY_t\,|\,x_t, y^{t-1})$.

Proof: We first show the optimal channel is Gaussian. Let $G(dX^T)$ be a jointly Gaussian source admitting a density g. Let $\{P(dY_t\,|\,x^t, y^{t-1})\}$ be a channel. Call the resulting joint measure $P(dX^T, dY^T)$. Let $G(dX^T, dY^T)$ be a jointly Gaussian measure with the same second-order properties as $P(dX^T, dY^T)$. Then

a) $\{G(dY_t\,|\,x^t, y^{t-1})\}$ is a Gaussian channel;

b) $G(dX^T, dY^T)$ has the same independence properties as $P(dX^T, dY^T)$;

c) $I_G(X^T;Y^T) \leq I_P(X^T;Y^T)$.

Part a) follows because $G(dX^T, dY^T)$ is jointly Gaussian. Part b) follows from noting that independence or conditional independence of some random variables implies that those same random variables are uncorrelated or conditionally uncorrelated. $G(dX^T, dY^T)$ has the same second-order properties as $P(dX^T, dY^T)$ thus it inherits the same independence properties. For c), note

$$\begin{aligned} &I_P(X^T;Y^T) - I_G(X^T;Y^T) \\ &\quad = \int p_{X^T,Y^T}(x^T,y^T) \log \frac{p_{X^T,Y^T}(x^T,y^T)}{p_{X^T}(x^T) p_{Y^T}(y^T)}\, dx^T\, dy^T \\ &\qquad - \int g_{X^T,Y^T}(x^T,y^T) \log \frac{g_{X^T,Y^T}(x^T,y^T)}{g_{X^T}(x^T) g_{Y^T}(y^T)}\, dx^T\, dy^T \\ &\quad = \int p_{X^T,Y^T}(x^T,y^T) \log \frac{p_{X^T\,|\,Y^T}(x^T\,|\,y^T)}{g_{X^T}(x^T)}\, dx^T\, dy^T \\ &\qquad - \int p_{X^T,Y^T}(x^T,y^T) \log \frac{g_{X^T\,|\,Y^T}(x^T\,|\,y^T)}{g_{X^T}(x^T)}\, dx^T\, dy^T \\ &\quad = \int p_{Y^T}(y^T) D(P_{X^T\,|\,y^T}\,|\,G_{X^T\,|\,y^T})\, dy^T \\ &\quad \geq 0. \end{aligned}$$

The second equality follows because G has the same second-order properties as P and because $P(dX^T) = G(dX^T)$.

For the weighted squared error distortion measure we see that the distortion is the same under P and G. Thus, for the Gaussian source the best channel that minimizes mutual information while maintaining a given squared error distortion is the Gaussian channel.

Now, we show the optimal Gaussian channel can be chosen of the form $P(dY_t\,|\,x_t, y^{t-1})$. Let $G(dY_t\,|\,x^t, y^{t-1})$ be any Gaussian channel. Once given the Gauss–Markov source (9) we can determine the joint measure $G(dX^{T-1}, dY^{T-1})$. Now define a new channel: $Q(Y_t\,|\,x_t, y^{t-1}) = G(Y_t\,|\,x_t, y^{t-1})$ $G-a.s.$ As before once given the Gauss–Markov source (9) we can determine the joint measure $Q(dX^{T-1}, dY^{T-1})$.

We will prove $Q(dX_t, dY^t) = G(dX_t, dY^t)$. It is straightforward to verify that $q(x_0, y_0) = g(x_0, y_0)$ holds for all (x_0, y_0). Assume the result holds for all $t \leq k$. We now prove the induction step. For any (x_{k+1}, y^{k+1}), we have

$$\begin{aligned} &q(x_{k+1}, y^{k+1}) \\ &\quad = q(y_{k+1}\,|\,x_{k+1}, y^k) \int q(x_{k+1}\,|\,x_k, y^k) q(x_k, y^k)\, dx_k \\ &\quad = q(y_{k+1}\,|\,x_{k+1}, y^k) \int g(x_{k+1}\,|\,x_k) g(x_k, y^k) dx_k \\ &\quad = g(x_{k+1}, y^{k+1}). \end{aligned}$$

Thus, the distortion under Q is the same as under G. We now show that the mutual information under Q is less than or equal to that under G: $I_G(X^t;Y_t\,|\,Y^{t-1}) \geq I_G(X_t;Y_t\,|\,Y^{t-1}) = I_Q(X_t;Y_t\,|\,Y^{t-1})$. □

Proposition 4.2: For any T the limit $\lim_{D\to 0}(R_T^{\mathrm{SRD},o}(D)/R_T^{\mathrm{SRD}}(D)) = 1$.

Proof: We sketch the proof for the scalar source given in (9): $X_{t+1} = FX_t + W_t$. Following [16], let Q_Δ represent a scalar quantizer defined as follows: $Q_\Delta(x) = k\Delta + (\Delta/2)$ if $k\Delta \leq x < (k+1)\Delta, \forall k = 0, \pm 1, \pm 2, \ldots$. Let $Q_\Delta(X^{T-1}) = (Q_\Delta(X_0), Q_\Delta(X_1), \ldots, Q_\Delta(X_{T-1}))$ denote

the componentwise application of Q_Δ to each component of X^{T-1}. Note that the entropy $(1/T)H(Q_\Delta(X^{T-1}))$ is an upper bound to $R_T^{\mathrm{SRD},o}(\Delta^2/4)$ because the squared error distortion under the quantizer Q_Δ is at most $\Delta^2/4$.

Now, for X^{T-1} with finite-differential entropy it can be shown that [16]

$$\lim_{\Delta\to 0}\left(\frac{1}{T}H(Q_\Delta(X^{T-1}))+\log\Delta\right)=\frac{1}{T}h(X^{T-1}). \quad (18)$$

By the previous results, we know that for D small enough $R_T^{\mathrm{SRD}}(D) = (1/2T)\log(K_{X_0}/D) + (T-1/2T)\log(F^2 + (K_W/D))$. Since X^{T-1} is jointly Gaussian, we can write the differential entropy as

$$\begin{aligned}\frac{1}{T}h(X^{T-1}) &= \frac{1}{T}\left(h(X_0)+\sum_{t=1}^{T-1}h(X_t\,|\,x_{t-1})\right)\\ &= \frac{1}{2}\log 2\pi e+\frac{1}{2T}\log K_{X_0}+\frac{T-1}{2T}\log K_W.\end{aligned}$$

Using (18) with $D=(\Delta^2/4)$, we get

$$\begin{aligned}&\lim_{\Delta\to 0}\left(\frac{1}{T}\right)H(Q_\Delta(X^{T-1}))\\ &\quad=\lim_{\Delta\to 0}\frac{1}{T}h(X^{T-1})-\log\Delta\\ &\quad=\lim_{D\to 0}\frac{1}{2}\log 2\pi e+\frac{1}{2T}\log K_{X_0}\\ &\qquad+\frac{T-1}{2T}\log K_W-\frac{1}{2}\log 4D\\ &\quad=\lim_{D\to 0}\frac{1}{2}\log\frac{\pi e}{2}+\frac{1}{2T}\log\frac{K_{X_0}}{D}+\frac{T-1}{2T}\log\frac{K_W}{D}\\ &\quad\overset{(a)}{=}\lim_{D\to 0}\frac{1}{2}\log\frac{\pi e}{2}+\frac{1}{2T}\log\frac{K_{X_0}}{D}\\ &\qquad+\frac{T-1}{2T}\log\left(F^2+\frac{K_W}{D}\right)\\ &\quad=\lim_{D\to 0}\frac{1}{2}\log\frac{\pi e}{2}+R_T^{\mathrm{SRD}}(D)\end{aligned}$$

where a) follows because $\lim_{D\to 0}\log(K_W/F^2D+K_W)=0$. Thus $\lim_{D\to 0}((1/T)H(Q_{2\sqrt{D}}(X^{T-1}))/R_T^{\mathrm{SRD}}(D))=1$. See [16] and [19] for details. □

ACKNOWLEDGMENT

The authors would like to thank V. Borkar and N. Elia for many helpful discussions.

REFERENCES

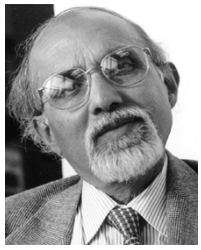
[1] J. Baillieul, "Feedback designs in information-based control," presented at the Workshop on Stochastic Theory and Control, Lawrence, KS, Oct. 2001.
[2] R. Bansal and T. Basar, "Simultaneous design of measurement and control strategies for stochastic systems with feedback," *Automatica*, vol. 25, no. 5, pp. 679–694, 1989.
[3] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 494–500, Oct. 1974.
[4] T. Berger, *Rate Distortion Theory*. Upper Saddle River, NJ: Prentice-Hall, 1971.
[5] D. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 2000.
[6] V. Borkar and S. Mitter, "LQG control with communication constraints," in *Communications, Computation, Control, and Signal Processing: A Tribute to Thomas Kailath*. Norwell, MA: Kluwer, 1997.
[7] V. Borkar, S. Mitter, and S. Tatikonda, "Markov control problems under communication constraints," *Commun. Inform. Systems (CIS)*, vol. 1, no. 1, pp. 15–32, Jan. 2001.
[8] ——, "Optimal sequential vector quantization of Markov sources," *SIAM J. Control Optimiz.*, vol. 40, no. 1, pp. 135–148, Jan. 2001.
[9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[10] N. Elia and S. Mitter, "Stabilization of linear systems with limited information," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 1384–1400, Sept. 2001.
[11] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[12] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1147–1158, May 2003.
[13] A. Gorbunov and M. Pinsker, "Prognostic epsilon entropy of a Gaussian message and a Gaussian source," *Trans. Problemy Peredachi Informatsii*, vol. 10, no. 2, pp. 5–25, Apr.–June 1973.
[14] K. H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. 24, pp. 1283–1290, Dec. 1976.
[15] D. Liberzon and R. Brockett, "Quantized feedback stabilization of linear systems," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 1279–1289, July 2000.
[16] T. Linder and R. Zamir, "Causal coding of stationary sources and individual sequences with high resolution," *IEEE Trans. Inform. Theory* , 2004, submitted for publication.
[17] G. Nair and R. Evans, "Mean square stabilisability of stochastic linear systems with data rate constraints," in *Proc. 41st IEEE Conf. Decision and Control*, Dec. 2002, pp. 1632–1637.
[18] A. Sahai, "Anytime information theory," Ph.D. dissertation, Dept. EECS, Mass. Inst. Technol., Cambridge, MA, Feb. 2001.
[19] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Dept. EECS, Mass. Inst. Technol., Cambridge, MA, Aug. 2000.
[20] ——, "The sequential rate distortion function and joint source-channel coding with feedback," in *Proc. 41st Allerton Conf. Communication, Control, and Computing*, Oct. 2003.
[21] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Automat. Contr.*, vol. 49, pp. 1056–1068, July 2004.
[22] ——, "Control over noisy channels," *IEEE Trans. Automat. Contr.*, vol. 49, pp. 1196–1201, July 2004.
[23] ——, "The sequential rate distortion function," *IEEE Trans. Inform. Theory*, 2004, submitted for publication.
[24] S. Tatikonda, A. Sahai, and S. Mitter, "Control of LQG systems under communication constraints," presented at the 37th IEEE Conf. Decision and Control, Tampa, FL, Dec. 1998.
[25] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, vol. 41, pp. 44–54, Jan. 1995.
[26] H. Witsenhausen, "Separation of estimation and control for discrete-time systems," *Proc. IEEE*, vol. 59, pp. 1557–1566, Nov. 1971.
[27] W. Wong and R. Brockett, "Systems with finite communication bandwidth constraints I: State estimation problems," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 1294–1299, Sept. 1997.
[28] ——, "Systems with finite communication bandwidth constraints II: Stabilization with limited information feedback," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 1049–1053, May 1999.

Sekhar Tatikonda (S'92–M'00) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 2000.

From 2000 to 2002, he was a Postdoctoral Fellow in the Computer Science Department at the University of California, Berkeley. He is currently an Assistant Professor of Electrical Engineering at Yale University, New Haven, CT. His research interests include communication theory, information theory, stochastic control, distributed estimation and control, statistical machine learning, and inference.

Anant Sahai (M'02) received the B.S. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1994, and the M.S. and Ph.D. degrees in electrical engineering and computer science, from the Massachusetts Institute of Technology, Cambridge, in 1996 and 2001, respectively.

In 2001, he developed adaptive signal processing algorithms for software radio at Enuvis, San Francisco, CA. He joined the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, as an Assistant Professor in 2002. His current research interests are in control over noisy channels, information theory, and wireless communication.



Sanjoy Mitter (M'68–SM'77–F'79–LF'01) received the Ph.D. degree from the Imperial College of Science and Technology, London, U.K., in 1965.

He joined the Massachusetts Institute of Technology (MIT), Cambridge, in 1969, where he has been a Professor of Electrical Engineering since 1973. He was the Director of the MIT Laboratory for Information and Decision Systems from 1981 to 1999. He was also a Professor of Mathematics at the Scuola Normale, Pisa, Italy, from 1986 to 1996. He has held visiting positions at Imperial College, London, U.K.; University of Groningen, Groningen, The Netherlands; INRIA, Paris, France; Tata Institute of Fundamental Research, Tata, India; and ETH, Zürich, Switzerland. He was the McKay Professor at the University of California, Berkeley, in March 2000, and held the Russell-Severance-Springer Chair in Fall 2003. His current research interests are communication and control in networked environments, the relationship of statistical and quantum physics to information theory, and control and autonomy and adaptiveness for integrative organization.

Dr. Mitter is a Member of the National Academy of Engineering and the winner of the 2000 IEEE Control Systems Award.