European Journal of Control (2001)7:122–131

European
Journal of
Control

Control with Limited Information

Sanjoy K. Mitter

Department of Electrical Engineering and Computer Science, The Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139 USA

A long-standing open conceptual problem has been the following: How does "information" interact with control of a system, in particular feedback control, and what is the value of "information" in achieving performance objectives for the system through the exercise of control? In answering this question we have to remember that in contrast to a variety of communication settings, the issue of time-delay is of primary importance for control problems, especially control of systems which are unstable. We discuss various issues arising from these fundamental questions.

Keywords: Control with communication; Information theory and control; Quantisation; Sequential rate distortion

1. Introduction

A long-standing open conceptual problem has been the following: How does "information" interact with control of a system, in particular feedback control, and what is the value of "information" in achieving performance objectives for the system through the exercise of control? In answering this question we have to remember that in contrast to a variety of communication settings, the issue of time-delay is of primary importance for control problems, especially control of systems which are unstable.

The theoretical basis for modern digital communications is undoubtedly Information Theory as developed by Shannon. This theory tells us in a precise way the fundamental limitation to reliable communication over a noisy channel. The crowning achievement of this theory is the Noisy Channel Coding Theorem, which identifies the channel in terms of the invariant quantity, called capacity of the channel, and reliable communication can take place if transmission occurs at a rate below capacity and cannot if it occurs at a rate above capacity. This theorem links the input side of the communications problem via the notion of capacity with the output side, namely, the ability to decode with arbitrarily small probability of error. This theorem can be extended to the rate of distortion context as Shannon himself did. One can do no better than quote Shannon to illuminate this situation:

> *Duality of a Source and a Channel.* There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is enhanced if we consider channels in which there is a "cost" associated with the difference input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity. Thus input letter i might have cost a_i and we wish to find the capacity with the side condition $\sum_i P_i a_i \leq a_i$, say, where P_i is the probability of using input letter i. This problem amounts, mathematically, to *maximizing* a mutual information under variation of the P_i with a linear inequality as constraint. The solution of this problem leads to a capacity cost function $C(a)$ for the channel. It can be shown readily that this function is *concave* downward. Solving this problem corresponds, in a sense, to finding a source that is just right for the channel and the desired cost.

Correspondence and offprint requests to: S.K. Mitter, Department of Electrical Engineering and Computer Science, The Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139 USA. Fax: 617-253-3578; Email: mitter@mit.edu.

Received 18 May 2001; Accepted 28 May 2001.
Recommended by Martins de Carvalho, J.M.G. Sá da Costa and António Dourado.

In a somewhat dual way, evaluating the rate distortion function $R(d)$ for a source amount, mathematically, to *minimizing* a mutual information under variation of the $q_i(j)$, again with a linear inequality acts as constraint. The solution leads to a function $R(d)$ which is *convex* downward. Solving this problem corresponds to finding a channel that is just right for the source and the allowed distortion level. This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus, we may have knowledge of the past and cannot control it; we may control the future but have no knowledge of it.

One of the many fundamental contributions which Shannon made which in fact renders the enunciation of the Noisy Channel Coding Theorem possible, is to think "digitally" (to use the work of a modern sage of Media technology), that is, to reduce everything to bits, a common currency in which everything can be evaluated. As we shall see later whether all bits are identical is an issue that we will have to face when dealing with the development of an Information Theory for sources which are decidedly non-stationary and non-ergodic.

A corresponding all embracing theory for control in the presence of uncertainty does not exist. The issue of fundamental limitation is far more complicated here since it is unclear that the dynamical systems which we wish to modify to behave in prescribed ways through control can be characterized through a simple invariant quantity like capacity. Even the invariants of a linear multivariable time-invariant system are the Knonecker invariants which tell us what Jordan forms we can reach through coordinate changes and linear constant feedback [24]. The nearest thing to fundamental limitation of control systems analogous to Shannon theory are the Bode inequalities, the irreducible error in the linear Quadratic Gaussian problem and characterization of performance limitations of control of linear time-invariant systems where the performance measure is sensitivity and this can be characterized through an H^{∞}-norm.

Nevertheless, control systems, even complex systems are being built where sensors, actuators and controllers are being linked through noisy communications channels and a theory which unifies systems theory and a theory of information is badly needed. A diagram of such a system is shown in Fig. 1. The design problem now is to design the estimator, the coders and decoders and the controller to meet specified closed-loop design objectives. We immediately see that this is a far more complex problem than point to point communications. It is totally unclear whether the control part of the problem can be "separated" from the communications part of the problem. This problem is distributed and the issue of information structure, namely, what information is available when and where, is actually a design issue and must be understood.

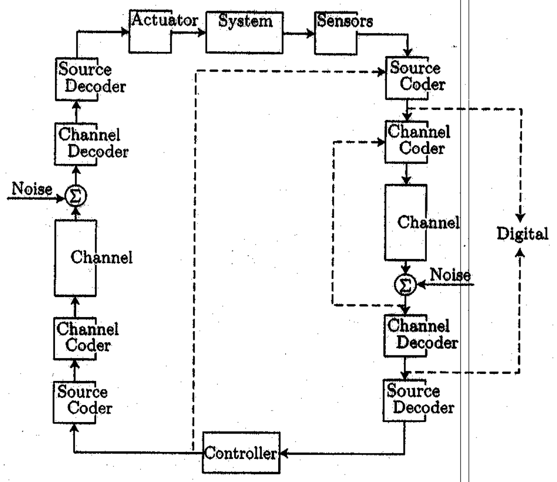

Fig. 1. Closed-loop system. *Note*: Dashed arrows show some of the potential feedback paths. Controller may be viewed as a channel.

The issues that I am raising are actually present in Communications problems where feedback and side information are present. In a conversation I had with Jim Massey at ETH in 1995, he pointed out that Shannon in the first Shannon lecture in 1973 had remarked that real time (time delay) issues and feedback in communication problems were questions which had received inadequate attention in Information Theory.

In light of the above discussion, I wish to raise two questions:

I. Is there a role for Information Theory in a unified theory of Control and Communications?

II. Can Systems Theory contribute to Communications and Information Theory in some non-trivial way?

In my view, the answers to both questions are a Qualified Yes.

This is not the first time that these two questions have been posed. A successful interaction between Systems Theory and Coding Theory is through the work of Willems on the behavioral view of systems [18] and Forney, Massey, Trott, Loeliger, Mittelholzer on codes on Finite Groups (see e.g. [11]). There are also attempts at using rate distortion theory to obtain lower bounds on estimation error for non-linear filtering (see [12,25]). Nevertheless we must proceed with caution. This is best captured by quoting from Hans Witsenhausen [22] who thought deeply about these issues:

> The infimum expected cost achievable in a problem depends upon the prevailing information pattern. Changes in information produce changes in optimal cost. This suggests the idea of measuring information by its effects upon the optimal performance.
>
> Such a measure of information is entirely dependent on the problem at hand and is clearly not additive. The only general property that it is known to possess is that additional information, if available free-of-charge can do no harm though it may be useless. This simple monotonicity property is in sharp contrast with the elaborate results of information transmission theory. The latter deals with an essentially simple problem, because the *transmission* of information is considered independently of its *use*, long periods of transmission and use of channel are assumed and *delays* are ignored. H.S. Witsenhausen: 1971.

In light of the above there is a methodological and theory formation issue which must be addressed. Simply stated, we must pose control questions in an appropriate informational sense and we must situate information theory in a dynamical framework. An elaboration of this viewpoint has been undertaken in the recently completed doctoral thesis of Sekhar Tatikonda [22], Anant Sahai [22] and several papers [4,5,14,19,21].

2. Control in an Information Setting

To make the above ideas more concrete let me consider the following question:

> What is the minimal information needed about the current state of a single-input, discrete time, linear time-invariant unstable system in order to stabilize it?

The question we are asking is really about the optimal coding of the state, that is, coarsest vector quantization, to achieve stability.

This problem is mathematically formulated in terms of the construction of Controlled Quadratic Lyapunov Functions (Quadratic for explicit computations). That is given

$$x(t+1) = Ax(t) + bu(t), \quad t = 0, 1, \ldots, \tag{1}$$

where $x(t) \in X = \mathbb{R}^n$ is the state of the system and $u(t) \in U = \mathbb{R}$ is the control, A is an $n \times n$ matrix, b is an n-vector and we assume that (A, b) is a reachable pair, we are required to find the coarsest quantized feedback control which stabilizes the system. The idea of coarseness (minimal information) is captured as follows:

Given a controlled Lyapunov function

$$V(x) = (x, Px)_{\mathbb{R}^n}, \quad P > 0 \tag{2}$$

find a set

$$U = \{u_i \in R | i \in Z\} \tag{3}$$

and a quantizer $f : X \to U$, with

$$f(x) = -f(-x) \tag{3a}$$

and

$$\Delta V(x) = V(Ax + bf(x)) - V(x) < 0, \quad \forall x \in X, \quad x \neq 0. \tag{3b}$$

f naturally induces a partition on the state space X and we assume that the values of f in U are ordered in the sense that $u_i < u_j, i > j, i, j \in Z$. Let $Q(V) =$ set of all quantizers which solves the stabilization problem. For $g \in Q(V)$ and $0 < \epsilon < 1$, let $N(g[\epsilon])$ denote the number of levels that g assumes in the interval $[\epsilon, (1/\epsilon)]$. Define the quantization density

$$\eta_g = \lim_{\epsilon \to 0} \text{Sup} \frac{N(g[\epsilon])}{-\ln \epsilon}, \tag{4}$$

and

$$f^* = \text{Arg Min}_{g \in Q(V)} \eta_g \tag{5}$$

f^* is defined to be the *coarsest* quantizer corresponding to $V(x)$.

It turns out that the quantization problem can be confined to one preferred direction (one-dimensional), and the optimal quantization is logarithmic with the optimal scaling law ρ^* being given by

$$\rho^* = \frac{\Pi_{\ell \leq i \leq k}[\lambda_i^u] - 1}{\Pi_{i \leq \ell \leq k}[\lambda_i^u] + 1} \tag{6}$$

and λ_i^u, $\ell \leq i \leq k \leq n$ are the strictly unstable eigenvalues of A.

In the above formation, we have allowed quantizers with a countable number of levels. An equivalent formulation of the problem leads to a method for designing finite quantizers leading to practical stability. For continuous-time systems there is a relation between the optimal sampling time T^* and the optimal quantization scaling law ρ^*:

$$T^* \sum_{i=1}^{k} \lambda_i^u(F) = \ln(1 + \sqrt{2}) \tag{7}$$

and

$$\rho^*(T^*) = \sqrt{2} - 1, \tag{8}$$

where $\lambda_i^u(F)$ are the unstable eigenvalues of the continuous-time system matrix F.

Note that $\rho^* \cdot T^*$ is an invariant of the class of single-input, continuous-time linear time-invariant systems. We may think of the quantized, stabilized feedback system as a symbolic description of the stabilized linear feedback system. It is also an example of a source-coding problem with a non-standard criterion function. For details of above see [10].

The stochastic version of this problem where the quantized stabilization problem is posed for

$$x(t+1) = Ax(t) + bu(t) + w(t), \quad t = 0, 1, \ldots \tag{9}$$

$w(t)$ being white Gaussian noise, is even more interesting. Here one can exhibit the non-linear effects of quantization as a desirable effect as opposed to a source of noise which is undesirable and should be guarded against. A little bit later we shall see the desirable effects of quantization in a different context.

The generalization to the stochastic case of the above quantization results can be obtained by invoking the notion of a storage function in the stochastic case.

Definition. A measurable function $V : \mathbb{R}^d \to \mathbb{R}$ is said to be storage function associated with a supply rate $g \in C(\mathbb{R}^n \times \mathbb{R}; \mathbb{R})$ if it is bounded from below and

$$V(X(t)) + \sum_{s=0}^{t} g(X(s), u(s)), \quad t > 0$$

is a positive $\mathcal{F}_t$-super-martingale for all $(X(\cdot); u(\cdot))$, satisfying (9), where

$$\mathcal{F}_t = \sigma(X(s) | 0 \le s \le t).$$

For our problem the function g is a quadratic function, and the control $u(t)$ is of the form $u(t) = k'x(t)$. The deterministic results can be generalized using the above definition and exploiting the connection of this to value functions for ergodic control problems [8].

3. Distributed Control and Quantization

To demonstrate how quantized controllers may have important advantages, consider the following Stochastic Control problem originally posed by Witsenhausen [14,23]:

> The time horizon $T = \{0, 1, 2\}$. All random variables are scalar. X_0 is a Gaussian random variable with mean zero and variance σ^2. The state transition equations are
>
> $$X_1 = X_0 + u_1, \tag{10}$$
>
> $$X_2 = X_1 + u_2. \tag{11}$$
>
> Here u_1 and u_2 denote control values. The output equations are
>
> $$\begin{aligned} Y_1 &= X_0 \\ Y_2 &= X_1 + W_1 \end{aligned} \tag{12}$$
>
> where W is a zero-mean, unit variance, Gaussian random variable independent of X_0. The cost function to be minimized is
>
> $$\underset{(X_0, W_1)}{\mathbb{E}} [k^2 u_1^2 + X_2^2] \tag{13}$$
>
> where the control policies have the information structure $U_1 = \gamma_1(Y_1)$ and $U_2 = \gamma_2(Y_2)$ where γ_1 and γ_2 are measurable functions of Y_1 and Y_2. Note that this is an example of a distributed control because the controller at stage 2 does not have full access to the past information. Note that if γ_2 could be a function of (Y_1, Y_2, u_1), then the choice of $\gamma_1(Y_1) = 0$ and $\gamma_2(Y_1, Y_2, u_1) = Y_1$ gives us zero cost.

The best affine controller

$$\gamma_1(Y_1) = aY_1 = aX_0 \quad \text{and} \quad \gamma_2(Y_2) = bY_2$$

can be computed, and the expected cost where b is chosen optimally is:

$$k^2 a^2 \sigma^2 + \frac{(1+a)^2 \sigma^2}{1 + (1+a)^2 \sigma^2}.$$

To compute the optimal a, if we define $t = \sigma(1+a)$, we see it is given implicitly by the equation for t:

$$\frac{t}{(1+t^2)^2} = k^2(\sigma - t).$$

For $k = 0.1$, $\sigma = 10$, gives us an $a = -0.0101$ with optimal cost $= 0.99$.

Now consider the following control

$$\begin{aligned} \gamma_1(Y_1) &= -Y_1 + \sigma \operatorname{sgn}(Y_1) \\ \gamma_2(Y_2) &= \sigma \operatorname{sgn}(Y_2). \end{aligned} \tag{14}$$

This control strategy which uses quantization, may be thought of as doing 1-bit quantization followed by Maximum Likelihood decoding. Now, by close inspection we can see that for large σ, the expected cost at the second stage is nearly zero since it is equal to $4\sigma^2 P_e(\sigma)$ where P_e is the probability of decoding error at the second stage. But P_e obviously dies off as $e^{-\sigma^2/2}$ since it is the integral of a tail of a Gaussian random variable. *No integrals need to be computed.* Furthermore, we see that we only needed one simple non-linear element (the sgn function – a comparator) for each controller, making the practical significance of these results clearer. This phenomenon is not something that we need "complicated" non-linearities to take advantage of.

Building on the intuition given above, consider the following family of "quantizing" controllers, parameterized by a single number B.

$$\gamma_1^B(y_1) = -y_1 + B\left\lfloor \frac{y_1}{B} + \frac{1}{2} \right\rfloor, \tag{15}$$

$$\gamma_2^B(y_2) = B\left\lfloor \frac{y_2}{B} + \frac{1}{2} \right\rfloor. \tag{16}$$

The first stage takes the input and "quantizes" it into bins of size B. The decoder then just looks to see which bin the value is in. Consider now a series of problems $(k, \sigma)_n$ and non-linear controllers as follows:

$$k_n = \frac{1}{n^2}, \tag{17}$$

$$\sigma_n = n^2, \tag{18}$$

$$B_n = n. \tag{19}$$

For our purposes, the analysis of the performance of these controllers is also simple. The first stage cost is $k^2E((\gamma_1^B(x_0))^2)$ which by inspection can certainly be bounded by $k^2B^2/4$ since the absolute value of the control is clearly bounded above by $B/2$. Since, $k_n^2B_n^2(1/n^2)$, the first stage cost tends to zero in this sequence.

For the second stage, we notice that since the bin size B grows as n while the variance of the observation noise w stays fixed at 1, the second stage cost is zero, unless the noise w has magnitude greater than $B/2 = n/2$. But since w is Gaussian, this tail event happens with a probability that tends to zero as $e^{-n^2/8}$. So, in the limit of large n, the second stage cost is zero as well. Thus

$$\lim_{n\to\infty} E(J_n|\gamma^{B_n}) = 0. \tag{20}$$

But what happens to the affine cost? Examining Equation 13, and substituting, we have:

$$E(J_n|\gamma_{\text{affine}}) = a^2 + \frac{(1+a)^2}{(1/n^4) + (1+a)^2}. \tag{21}$$

Clearly,

$$\lim_{n\to\infty} E(J_n|\gamma_{\text{affine}}) = a^2 + 1. \tag{22}$$

And so, we can see that the minimum cost is achieved by setting a to zero, giving us

$$\lim_{n\to\infty} E(J_n|\gamma_{\text{bestaffine}}) = 1. \tag{23}$$

So, the ratio $\frac{E(J_n|\gamma_{\text{bestaffine}})}{E(J_n|\gamma^{B_n})}$ tends to infinity.

Discussion

We have seen that in the case of this particular information pattern, a non-linear controller can be superior to the best linear one. Can we get any intuition as to why this situation arose?

It seems that since the cost of control in stage 2 is zero, all that mattered at the second stage was how well it could predict x_1. Also, by not penalizing the state and keeping the cost of control in stage 1 low, we were effectively giving the first stage a lot of freedom in setting x_1 and a strong incentive to view the output x_1 purely as a way to communicate over a Gaussian channel with the second stage about the state. This coincidence of the message[1] and the messenger[2] is what is causing this seemingly strange behavior.

Ideally, what we would like is for the message to be simple (i.e. low entropy = informative prior[3]) so that there is less-information for the decoder to try and extract from the signal. However, to get the message across intact, we would like the messenger to have high-energy so that the signal-to-noise ratio is favorable (high mutual information = informative likelihoods[4]). Unfortunately, when we restrict ourselves to affine controllers for this problem, *these two objectives are in direct opposition.* An affine controller implies Gaussian state and for a Gaussian random variable, high energy implies high entropy and low entropy implies low energy.

4. LQG and its Variants (see [4,20,21])

In this section we examine the LGQ problem under communication constraints. In Section 4.1 we state the problem. In Section 4.2 we state a lower bound in terms of the sequential rate distortion function defined in Section 4.2. In Section 4.3 we state upper bounds. Finally we conclude in Section 4.4.

[1] x_1 is exactly what we want to communicate to the second stage.

[2] x_1 is also the input to the "channel"

[3] The intuition involved is that low entropy implies less unpredictability. Less unpredictability means that our prior knowledge is quite strong.

[4] The intuition for the case of signalling is that we want to reduce the effect of the noise. We do this by having a large mutual information between the input and output of the channel. Using the terms of hypothesis-testing, this means that we would like our "likelihood" terms to be strongly discriminating.

4.1. Problem Setup

Throughout this section we consider the following time-invariant system:

$$X_{n+1} = AX_n + BU_n + W_n, \quad \forall n \geq 0, \tag{24}$$

where $\{X_n\}$ is a $\mathbb{R}^d$-valued state process and $\{U_n\}$ is a $\mathbb{R}^m$-valued control process. The sequence $\{W_n\}$ is IID Gaussian $\sim \mathcal{N}(0, K_W)$. And the initial position $X_0 \sim \mathcal{N}(0, K_X)$.

Our goal is to minimize the long-term average cost

$$\limsup_{N\to\infty} \frac{1}{N} E\left[\sum_{n=0}^{N-1} X_n'QX_n + U_n'TU_n\right] \tag{25}$$

where Q is positive semidefinite and T is positive definite.

Under full state observation it is well known that the optimal steady state control law is a linear gain of the form $U_n = LX_n$ where

$$L = -(B'PB + T)^{-1}B'PAX \tag{26}$$

where P satisfies the Riccati equation

$$P = A'(P - PB(B'PB + T)^{-1}B'P)A + Q. \tag{27}$$

Furthermore the optimal cost is

$$E(W'PW) = \mathrm{tr}(PK_W). \tag{28}$$

These standard results can be found in [2].

Our problem differs from the standard LQG result because we have a communication channel between the sensor and the controller. See Fig. 2. The channel has an input alphabet $\mathcal{A}$ and an output alphabet $\mathcal{B}$. The channel is defined by a sequence of stochastic kernels $\{Q(dB_n|a^n, b^{n-1})\}_{n=0}^{\infty}$. In this section we treat two channels. The digital noiseless channel with a rate $\mathbb{R}$ in which $\mathcal{A} = \mathcal{B} = \Sigma$ for some finite set Σ. For the digital channel we have $b_n = a_n$. And the additive white Gaussian noise channel with power constraint P in which $\mathcal{A} = \mathcal{B} = \mathbb{R}^d$. For this channel $B_n = A_n + V_n$ where V_n is a zero mean Gaussian with covariance K_V. The rate is $\mathbb{R} = \frac{1}{2}\log_2(1+P)$.

We now quickly define the encoder, decoder, and the controller.

Encoder: The encoder at time n is a map

$$\mathcal{E}_n : \mathbb{R}^{d(n+1)} \times \mathcal{A}^n \times \mathcal{B}^n \times \mathbb{R}^{mn} \to \mathcal{A}$$

that takes

$$(x^n, a^{n-1}, b^{n-1}, u^{n-1}) \mapsto a_n.$$

Note that the encoder is allowed to be a function of the past controls and past channel outputs.

Decoder: The decoder at time n is a map

$$\mathcal{D}_n : \mathcal{B}^{n+1} \times \mathbb{R}^{dn} \times \mathbb{R}^{mn} \to \mathbb{R}^d$$

that takes

$$(b^n, \hat{x}^{n-1}, u^{n-1}) \mapsto \hat{x}_n.$$

The output of the decoder is some estimate of the state. In the sequel the output will be the conditional expectation of the current state given the channel outputs and controls.

Controller: The controller at time n is a map

$$\mathcal{C}_n : \mathbb{R}^d \to \mathbb{R}^m$$

that takes

$$\hat{x}_n \mapsto u_n.$$

Note that the encoder has available to it all the information that the decoder has. This is called equi-memory.

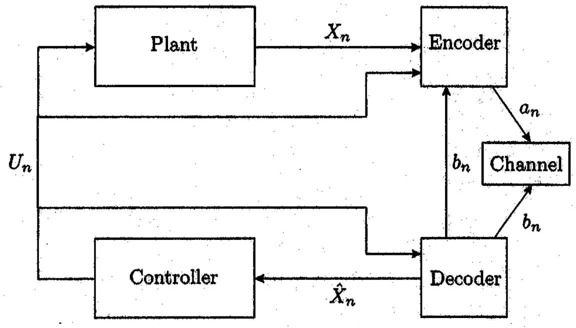

Fig. 2. System.

Note also that the controller is a function of the state estimate. Thus we are imposing a certainty equivalent structure on the controller.

Define $\hat{X}_n = E(X_n | B^n, \hat{X}^{n-1}, U^{n-1})$. And define $e_n = X_n - \hat{X}_n$.

Lemma 4.1. The error e_n is independent of all control choices U^{n-1} for all n.

Proof. This is known. See [2]. First note that

$$\begin{aligned} e_{n+1} &= X_{n+1} - \hat{X}_{n+1} \\ &= AX_n + BU_n + W_n - E(AX_n + BU_n \\ &\quad + W_n | B^{n+1}, \hat{X}^n, U^n) \\ &= AX_n + BU_n + W_n - E(A\hat{X}_n + Ae_n + BU_n \\ &\quad + W_n | B^{n+1}, \hat{X}^n, U^n) \\ &= Ae_n + W_n - E(Ae_n + W_n | B^{n+1}, \hat{X}^n, U^n). \end{aligned}$$

We prove this by induction. First note that $e_0 = X_0 - E(X_0 | B_0)$. Now $e_1 = Ae_0 + W_0 - E(Ae_0 + W_0 | B^1, \hat{X}_0, U_0)$. This is independent of U_0. Assume that e_n is independent of U^{n-1}. Then by the induction hypothesis $e_{n+1} = Ae_n + W_n - E(Ae_n + W_n | B^{n+1}, \hat{X}^n, U^n)$ must be independent of U^n. □

Lemma 4.2. If the error is independent of the controls then the certainty equivalent controller is optimal.

Proof. This can be found in [22].

The following steps follow from Borkar-Mitter [4]. The running cost can be written as

$$\begin{aligned} E(X_n'QX_n + U_n'TU_n) &= E(\hat{X}_n'Q\hat{X}_n + U_n'TU_n) \\ &\quad + E(e_n'Qe_n). \end{aligned}$$

Furthermore the evolution of $\hat{X}_n$ can be written as

$$\hat{X}_{n+1} = A\hat{X}_n + BU_n + \hat{W}_n,$$

where $\hat{W}_n = Ae_n + W_n - e_{n+1}$.

But in this case we have a full state observation LQG problem with state process $\hat{X}_n$ and running cost $E(\hat{X}_n'Q\hat{X}_n + U_n'TU_n)$. Thus the optimal control law is given by (26). □

Assume that $\mathrm{var}(e_n) = D$ for all n. Then the optimal cost for the original problem is

$$\begin{aligned} &\limsup_{n\to\infty} \frac{1}{N} E\left[\sum_{n=0}^{N-1} X_n'QX_n + U_n'TU_n\right] \\ &= \limsup_{n\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} E(\hat{X}_n'Q\hat{X}_n + U_n'TU_n) + E(e_n'Qe_n) \\ &= \mathrm{tr}(TK_{\tilde{W}}) + \mathrm{tr}(QD) \\ &= \mathrm{tr}(TK_W) + \mathrm{tr}((A'PA - P + Q)D). \end{aligned}$$

Note that the optimal cost decomposes into two terms. The first term is the full state cost and the second term depends only on D the state estimation error covariance. Thus we have reduced the problem of computing the optimal cost to that of minimizing $\mathrm{tr}((A'PA - P + Q)D)$ over a given channel.

4.2. Lower Bound

We can lower bound the $\mathrm{tr}((A'PA - P + Q)D)$ term by treating the problem as a sequential rate distortion problem for the source $X_{n+1} = AX_n + W_n$ with squared error distortion metric and a weight matrix $A'PA - P + Q = M$.

Lower bounds can be determined by the sequential rate distortion calculation.

Sequential rate distortion theory (see [1,17] and the references cited there) has a key role to play in this problem. In our situation, we consider the process

$$X(k+1) = AX(k) + W(k)$$

with $(W(k))$ white Gaussian noise, the sequential rate distortion problem is defined as follows:

$$\begin{aligned} D_{N,\mathrm{Seq.}}(R, M) = \mathrm{In}\, f_{P(\hat{X}_1^N | X_1^N)} \times \frac{1}{N} E\Bigg(\sum_{k=1}^{N} [(X(k) \\ -\hat{X}(k), M(X(k) - \hat{X}(k)))]\Bigg) \end{aligned}$$

where M is positive definite, subject to the rate constraint $(1/N)I(\hat{X}_1^N; X_1^N) \le R$, where $I(\hat{X}_1^N; X_1^N)$ is the mutual information between $\hat{X}_1^N$ and X_1^N, and where the minimization is carried out over all $P(\hat{X}_1^N | X_1^N)$ which are causal, that is of the form $\Pi_{k=1}^n P(\hat{X}_k | X_k)$. The rate distortion function is

$$D\mathrm{Seq.}(R, M) = \lim_{N\to\infty} D_{N,\mathrm{Seq.}}(R, M).$$

For simplicity, consider the scalar case. A surprising result is that there is a minimum rate, $R > \log_2 A$, required to stabilize the system. In this case

$$D\mathrm{Seq.}(R, M) = \frac{MK_W}{2^{2R} - A^2},$$

where Σ_W is the variance of W. For special channels, including the AWGN with equi-memory, $D\mathrm{Seq.}(R,M)$ can be achieved. The structure of the minimizing conditional law suggests that the optimal structure of the encoder is predictive.

Note that this lower bound holds completely independent of the encoder, decoder and controller. The question then becomes when can we achieve the sequential rate distortion lower bound. We can achieve it if the channel we are given is "matched" to the source.

4.3. Upper Bound

If the channel is matched to the source $X_{n+1} = AX_n + W_n$ then we can achieve the sequential rate distortion value. If the channel is a digital channel then $\mathrm{tr}((A'PA - P + Q)D)$ will equal the operational sequential rate distortion bound.

4.4. Conclusions

Equi-memory is a strong condition and generally requires at least one noiseless link. One needs a way for the decoder to communicate to the encoder. Now, in a general way, we can consider the plant as a channel. Consider the scalar case and change the quadratic cost on U_k to a hard constraint: $E(U_k^2) \leq P_2$. Assume that the encoder has noisy observations of U_k. Let

$$R_2 = \frac{1}{2}\log_2\left(1 + \frac{B^2P_2}{K_W}\right).$$

A necessary condition for well-posedness is that $R_2 \geq \log_2 A$.

Returning to the original average cost problem, if there is no cost on control, then the equi-memory assumption can be dispensed with. Otherwise, there is a fundamental tradeoff between control energy and capacity required from the encoder to the decoder. Sub-optimal schemes which are optimal in the high-rate regime can be designed when the equi-memory assumption cannot be justified.

A more general view of this problem where the state process is a controlled Markov Chain has been considered in [6]. In other work, we have shown how the optimal sequential quantization of Markov sources can be viewed as a partially observed stochastic control problem [6].

5. Towards a Dynamical View of Information Theory

The discussion in the previous section raises a new problem in Information Theory:

> How can one reliably transmit an unstable source over a noisy channel through appropriate source and channel coding and decoding at the receiving end?

More precisely, given a scalar discrete-time finite-state Markov source $(X(t))$ given by

$$X(t+1) = aX(t) + W(t), \quad a > 1, t = 0, 1, \ldots$$

and $(W(t))_{t\geq 0}$ is additive white Gaussian noise or bounded noise with finite support and a memoryless channel (or additive white Gaussian noise channel) is it possible to design encoders and decoders within a specified finite end to end delay constraint so that the output of the decoder $(\hat{X}(t))$ achieves a desired mean-squared performance $\mathrm{Sup}_{t\geq 0}\, E(X(t) - \hat{X}(t))^2 \leq K$?

A similar question was posed by Berger [1] for the Wiener process and an information transmission theorem for this case has been an open problem for many years. In Anant Sahai's thesis [19] a solution to this problem is presented. The solution requires a dynamical view of Information Theory, since the message, the Markov Source, is not given at time -1, but unfolds in time and a little thought will make it clear that block coding of any kind will not work in this situation. Indeed, all coding and decoding operations must be causal, causality suitably defined. In this problem, the separation of source and channel coding is no longer obvious and separation has to be imposed by a new definition of channel capacity.

$$\begin{aligned} &C_{\mathrm{anytime}}(\alpha) \\ &\quad = \mathrm{Sup}\{R | \exists (K > 0, \mathrm{Rate}(\epsilon, D^a) = R)\ \forall d > 0 \\ &\qquad\qquad P_{\mathrm{error}}(\epsilon, D^a md) \leq K \cdot 2^{-\alpha D}\} \end{aligned}$$

In the above ϵ denotes the encoder, D^a the anytime decoder and d the prescribed delay.

This definition should be contrasted with the classical operational definition of capacity. The exponent α is related to the error exponents corresponding to block coding and convolutional coding. Appropriate source and channel coding theorems for this problem with the above definition of capacity are proved in [19]. I want to emphasize that the total end to end distortion problem has to be considered for this situation and the dynamical view which I have referred to is an essential element in the solution to this problem.

The discussions in the previous section and this section provides evidence as to why the sequential (zero delay) rate distortion problem is an important problem. Although it would be too much to expect that a Shannon-like rate distortion theorem would be true in this causal situation (see [17], for example), it is still important to characterize in a precise way the gap between the non-causal rate distortion function and the causal rate distortion function. This has been carried out in [7].

6. Control and Communication as Interconnection of Probabilistic Systems

In [18], Willems has proposed a definition of Dynamical Systems and a methodology for control which consists of interconnecting two dynamical systems to obtain a desired behavior. We propose here a

generalization of this view to Probabilistic Systems. This methodology has been exploited in the thesis of Tatikonda [20].

Let Ω be a finite set and Ω^Z be the set of sequences thought of as a compact, metrizable space. Let $\mathcal{P}(\Omega^Z)$ be the convex, weak* compact set of all probability measures on Ω^Z.

The shift

$$\sigma : \Omega^Z \longrightarrow \Omega^z$$

induces a continuous affine transformation

$$\hat{\sigma} : P(\Omega^Z) \longrightarrow \mathcal{P}(\Omega^Z).$$

Definition 6.1. A random system is a convex subset $\mathcal{S} \subset \mathcal{P}(\Omega^Z)$. If $\hat{\sigma}\mathcal{S} \subseteq \mathcal{S}$, we say that $\mathcal{S}$ is a shift-invariant random system.

We have a natural embedding.

$$\begin{aligned} j : \Omega^Z &\longrightarrow P(\Omega^Z), \\ &: x \longmapsto \delta_x. \end{aligned}$$

Definition 6.2. A random system $\mathcal{S}$ is said to be complete iff $\mu \in \mathcal{P}(\Omega^Z)$ with $\mu|_I \in \mathcal{S}|_I$ $\forall I$ finite intervals of $Z \Rightarrow \mu \in \mathcal{S}$.

Definition 6.3. A random system $\mathcal{S}$ is said to be L-complete iff in the above definition we can restrict I to be $I = [t, t+L]$.

Proposition 6.4. (Topological Characterization). Let $\mathcal{S}$ be a random system over Ω^Z. The following are equivalent:

(i) $\mathcal{S}$ is complete and $\mathcal{S}|_I$ is closed in $\mathcal{P}(\Omega^I)$ $\forall I$ finite interval.
(ii) $\mathcal{S}$ is closed in the weak*-topology.

This is a natural generalization of the work of Willems for deterministic systems.

j is a continuous map. We use the notation:

$$\mathcal{P}_\delta(\Omega^Z) := j(\Omega^Z).$$

Definition 6.5 (*Fagnani*). A random system $\mathcal{S}$ is said to be *deterministic* if $\exists \tilde{\mathcal{S}} \subseteq \Omega^Z$ s.t.

$$\mathcal{S} = \text{Conv.}(j\tilde{\mathcal{S}}).$$

If $I \subseteq Z$, we have a *projection*

$$\begin{aligned} \Pi_I : \Omega^Z &\longrightarrow \Omega^I \\ &: x \longmapsto \Pi_I(x) = x|_I \end{aligned}$$

If $\mathcal{S}$ is a random system, we let

$$\mathcal{S}|_I = \{\tilde{\Pi}_I \mu | \mu \in \mathcal{S}\}$$

where $\tilde{\Pi}_I$ is the projection (induced) on $\mathcal{P}(\Omega^Z)$. $\mathcal{S}|_I$ is a convex subset of $\mathcal{P}(\Omega^I)$.

Let $\mathcal{S} \subset \mathcal{P}(\Omega^Z)$ be a complete $\tilde{\sigma}$-invariant random system. $\mathcal{S}$ is therefore completely determined by a convex set

$$\mathcal{S}|_{[0,1]} \subseteq \mathcal{P}(\Omega \times \Omega) \hookrightarrow \mathbb{R}^{|\Omega|^2}$$

In the deterministic case $\mathcal{S}|_{[0,1]}$ is essentially equivalent to specifying a directed graph, consisting of 0 and 1 along the arcs.

Definition 6.6. Given two random systems $\mathcal{S}_1 \subseteq \mathcal{P}(\Omega^Z)$ and $\mathcal{S}_2 \subseteq \mathcal{P}(\Omega^Z)$ its *interconnection* is the random system

$$\mathcal{S}_1 \cap \mathcal{S}_2 \subseteq \mathcal{P}(\Omega^Z).$$

The problem of control is then the following: Given a random system $\mathcal{S}_\text{u} \subseteq \mathcal{P}(\Omega^Z)$ and a desired behavior $\mathcal{S}_\text{d} \subseteq \mathcal{P}(\Omega^Z)$, find a controller $\mathcal{S}_\text{c} \subseteq \mathcal{P}(\Omega^Z)$ such that

$$d(\mathcal{S}_\text{d}, \mathcal{S}_\text{u} \cap \mathcal{S}_\text{c})$$

is minimized, where d is an appropriate distance measure between two weak*-closed convex sets in the space of all probability measures $\mathcal{P}(\Omega^Z)$.

7. Conclusions

In this paper I have suggested that a unified view of Control and Communication is badly needed if we are to make progress towards a science of distributed systems, where subsystems are linked via Communication channels. I have given examples to show how informational questions need to be asked in a control context and how control questions need to be asked in an informational context. The subject is clearly in its infancy and much needs to be done.

Acknowledgements

This research has been supported by the NSF-KDI Grant ECS-9873451 and by the Army Research Office under Grant DAAD 19-00-1-0466.

Note

This is an expanded version of the author's plenary talk at the International Symposium on Information Theory 2000, Sorrento, Italy. It represents joint work with V. Borkar, N. Elia, A. Sahai and S. Tatikonda. This paper could not have been written without their important and essential contributions.

References

1. Berger T. Rate distortion theory. Prentice Hall, Englewood Cliffs New Jersey 1971
2. Bertsekas DP. Dynamic programming and optimal control. 2nd edn. Athena Press, Belmont MA 2001
3. Bertsekas DP, Shreve S. Stochastic optimal control: the discrete-time case. Academic Press, New York 1978
4. Borkar VS, Mitter SK. LQG control with communications constraints. In: Communications, control, computation and signal processing. A tribute to Thomas Kailath; Pautraj A, Roychowdhury V, Shaperr C. (eds) Kluwer Academic Publishers, pp 365–373, 1997
5. Borkar VS, Mitter SK, Tatikonda S. Optimal sequential vector quantization of markov sources. SIAM J Cont Optim (to appear)
6. Borkar VS, Mitter SK, Tatikonda S. Markov control problems with communications constraints Comm Inf Sys 2001 1(1): 16–33
7. Borkar VS, Mitter SK, Sahai A, Tatikonda S. Sequential source coding: an optimization viewpoint, preprint, LIDS, MIT, 1998 IEEE Trans Inf Th (submitted)
8. Borkar VS, Mitter SK. On stochastic dissipativeness (to appear)
9. Dobrushin RL. A general formulation of the basic Shannon Theorem in Information Theory. Uspekhi Math. Nauk, USSR 1959; 14(6): 3–103
10. Elia N, Mitter SK. Stabilization of linear systems with quantized controls. IEEE Trans Control Autom (to appear)
11. Forney GD, Trott MD. The dynamics of group codes: state spaces, Trellis diagrams and canonical Encoders. IEEE Trans Inf Th 1993; 39: 1491–1513
12. Galdos J. Information and distortion in filtering theory, Ph.D. Thesis, MIT, 1975
13. Massey JL. Causality, feedback and directed information. Proceedings of the ISIT-90, pp 303–305
14. Mitter SK, Sahai A. Information and Control: Witsenhausen Revisited. Lecture Notes in Control and Information Sciences, 241, Springer, 1998
15. Mitter SK. Lecture at Allerton Conference, 1998
16. Mitter SK. Lecture at IMA Workshop on Codes, Graphs and Systems, August, 1999
17. Neuhoff DL, Gilbert RK. Causal source codes. IEEE Trans Inf 1982; Th. 28(5): 701–713
18. Polderman JW, Willems JC. Introduction to mathematical systems theory, Springer, 1997
19. Sahai A. Any-time information theory, Ph.D. Thesis, MIT, December, 2000
20. Tatikonda S. Control under communication constraints. Ph.D. Thesis, MIT, September, 2000
21. Tatikonda S, Sahai A, Mitter SK. Control of LQG systems under communications constraints, forthcoming. See also Proceedings of the ACC, 1999; Proceedings of the CDC, 1998
22. Witsenhausen HS. Separation of estimation and control for discrete-time systems. Proceedings of the IEEE 1971; 59(11): 1557–1566
23. Witsenhausen HS. A counter-example in stochastic optimal control. SIAM J Control 1968; 6(1): 131–147
24. Wonham WM. Linear multivariable control. Springer-Verlag, New York, Berlin
25. Zakai M, Ziv J. Lower and upper bounds on optimal filtering errors for certain diffusion processes. IEEE Trans Inf Th 1972; IT-18.