Information and Control: Witsenhausen Revisited

Sanjoy Mitter, Anant Sahai *

Department of Electrical Engineering and Computer Science

and Laboratory for Information and Decision Systems

Massachusetts Institute of Technology

mitter@lids.mit.edu, sahai@mit.edu

Dedicated to Bruce Francis and Mathukumalli Vidyasgar on the occasion of their fiftieth birthday.

Abstract

The role of information in the context of control is a deep issue. To get at this, we review Witsenhausen's notions of *information patterns* for control problems. While staying in that basic framework, we then use ideas from traditional information theory as we re-examine Witsenhausen's famous "counterexample". In the process, we construct a family of nonlinear "quantizing" control laws that can perform infinitely better than the best linear policies.

1 Introduction

In traditional information theory, a technical notion of information is developed that is independent from the actual use of that information. Aside from its considerable aesthetic appeal, this body of ideas has proven itself to be quite useful in the context of information transmission. However, fundamental to most of the results in information theory is the use of long block lengths and letting sequence lengths tend to infinity as a way of getting the laws of large

*This research supported by U.S. Army Grant PAAL03-92-G-0115, Center for Intelligent Control Systems.

numbers to work to reduce uncertainty. In a control context, the focus is on the present. While there is a sense in which all of feedback control is about trying to reduce uncertainty, a control action must be applied now and we can not afford to wait forever.

In this report, we will attempt to get a handle on the role of information in control by revisiting two classic papers. The first of these is Witsenhausen's 1971 survey paper [4] on the "Separation of Estimation and Control for Discrete Time Systems." Here, we will give the essentials of Witsenhausen's framework for talking about stochastic control problems. The key idea is that of *information patterns* — a formal way of talking about the issue of "who knows what and when do they know it." Using this, we will restate his main assertions on the various forms of separation between estimation and control. Though the language is general, we will quickly find ourselves talking about linear systems with quadratic costs and Gaussian distributions for primitive random variables — the LQG problem.

With the basics behind us, we next consider Witsenhausen's 1968 "Counterexample In Stochastic Optimum Control" [3] which shows how important the information pattern really is to the control problem. It details a deceptively simple 2-stage LQG problem and shows that when you restrict to memoryless control, affine[1] controllers are no longer sufficient to minimize cost. The paper does this by computing the best affine controller and then exhibiting a nonlinear control law which does better.

We then present a simpler family of nonlinear control laws and use them to get something much stronger — a demonstration that the ratio of the cost of the best affine controller and a nonlinear controller can go to infinity! Then, we try to use ideas from information theory to give some intuition as to why the affine controllers are suboptimal. At its heart, the problem seems to boil down to one of communication between stages 1 and 2. We argue that the restriction to affine controllers is suboptimal because it forces a tension between the complexity of the message and the reliability of its transmission. We show how the nonlinear controller is able to circumvent this tension, achieving better performance.

[1] Linear plus constant

2 Separation of Estimation and Control

In Witsenhausen's classic survey paper [4], he sets out to elucidate the relationships between estimation and control for discrete time, Bayesian[2] systems. The fundamental issue stems from the distinction between the control *laws* and the actual realizations of the control *variables* applied to the system. The designer chooses the *laws* to fulfill some objective, and until that choice is made, the control *variables* are still "random variables to be of yet uncertain status."

2.1 Problem Framework

Witsenhausen considers a general finite-horizon distributed discrete time control problem. Time goes from 1 to T, there are M observation posts[3], and K control stations[4] The causal sequence is as follows:

1. Generation of random initial state x_0.
2. Observations of outputs $y_1^1, \cdots, y_1^M = (g_1^1(x_0, w_1^1), \cdots, g_1^M(x_0, w_1^M))$
3. Application of controls $u_1^1, \cdots, u_1^K$
4. Transition to state $x_1 = f_1(x_0, v_1, u_1^1, \cdots, u_1^K)$

and then this continues until the final state x_T is reached.

The uncertainty in the system is modeled by a basic set of independent primitive random variables: $x_0; v_t, w_t^m (t = 1, \cdots, T; m = 1, \cdots, M)$. The v_t enter into the state transition functions f_t and the w_t^m into the observation functions g_t^m in the obvious ways.

Finally, the preferences between outcomes are expressed consistently through an additive cost function on the state and the controls: $\sum_{t=1}^T h_t(x_t, u_t^1, \cdots, u_t^K)$. The goal of the designer is to pick a *design* γ specifying control *laws* γ_t^k that select the u_t^k to minimize the expected cost. Furthermore, once all the γ_t^k are selected, all the variables in the closed loop system become well defined random variables. More technically, given a complete design γ and a pair of sets of values for some arbitrary sets of the output and control variables, Y and U, we have a clearly defined σ-field $\mathcal{F}(Y, U; \gamma)$ in probability space and thus conditional

[2]All "uncertainty" in the system is modeled probabilistically

[3]For example, consider geographically distributed sensors

[4]These usually represent distributed controllers

distributions[5] for all the variables in the system[6].

2.2 Information Patterns

As stated above, the problem is still incompletely specified. We need to know the sets from which we are allowed to pick the functions γ_t^k. Stated informally, the key questions are "who knows what when" and "what are they allowed to do with that information?". To formalize the first of these questions, the notion of *information pattern* is defined. This assigns to every control variable u_t^k, two sets $Y_{t,k}$ and $U_{t,k}$ of pairs of indices specifying which observation variables y_τ^μ and control variables u_θ^κ the control law γ_t^k has access to[7]. Generally, no restriction is put on the functional form or range of γ_t^k, except the trivial one of saying that it should be measurable over the σ-field generated by its arguments. However, sometimes it is interesting to restrict attention to jointly affine γ_t^k.

For the idea of *information pattern* to be useful, we need a notion of equivalence over it. So, patterns $(Y_{t,k}, U_{t,k})$ and $(\tilde{Y}_{t,k}, \tilde{U}_{t,k})$ are *equivalent* if for any design γ feasible with the first, there is a design $\tilde{\gamma}$ feasible with the second such that every system variable agrees under the two designs almost surely.[8] Witsenhausen next defines some classifications of *information patterns.* A pattern is said to have *perfect recall* if $Y_{t,k} \subseteq Y_{t+1,k}$ and $U_{t,k} \subseteq U_{t+1,k}$. A pattern is said to be *classical* if it has *perfect recall* and moreover $Y_{t,k}$ and $U_{t,k}$ are independent of k.[9] We define two related terms that will also be useful. A pattern is said to be *perfectly classical* if every station has knowledge of all past outputs and controls. For the common case when the observation posts have a natural identification with the control stations[10], a pattern is said to be *locally classical* if every station can remember all of its past inputs and outputs.

Now, the point of these definitions is to begin to get at the notion that as long as we have information about the relevant past control *variables* and outputs, we might not need to know all the control *laws* in order to have well

[5]The underlying probability space and measure are determined by the primitive random variables.

[6]For example, the conditional probability $P(y_3^2 \in [-1,1] | y_2^4 = 7, y_3^3 = 5, u_4^2 = 0.5, \gamma)$ should be defined and make sense

[7]To be precise, γ_t^k takes as arguments all the y_τ^μ and u_θ^κ where $(\tau, \mu) \in Y_{t,k}$ and $(\theta, \kappa) \in U_{t,k}$

[8]With respect to the probability measure defined by the basic set of independent random variables.

[9]Independence of k means that all the control laws at any given time have access to the same information.

[10]In block diagrams for example, for each block there is a natural identification of the input arrows with the output ones.

defined random variables. Let L be a set of indices (θ, k). We use γ_L to refer to the restriction of design γ to just the laws γ_θ^k. Now, call a triple (Y, U, L) a *field basis* if for any two designs γ, $\hat{\gamma}$, $\gamma_L = \hat{\gamma}_L$ implies $\mathcal{F}(Y, U; \gamma) = \mathcal{F}(Y, U; \hat{\gamma})$. So, knowledge of the values of these particular Y and U together with knowing the laws γ_L is sufficient to understand the underlying probability space.[11]

2.3 Results

With these definitions in hand, Witsenhausen proceeds to state 11 distinct "Assertions" in the paper. Rather than going through all of them, we restate 4 of them that seem most important.

This first assertion is perhaps the most fundamental, and is the basis for many of the separation results for linear systems.

Assertion 1 *If, for every (t, k), $(Y_{t,k}, U_{t,k}, \emptyset)$ is a* field basis*, then the given feedback control problem is equivalent to a feedforward control problem.*

A feedforward control problem is defined as one in which the observation functions depend only on the primitive random variables, and not on the actual control variables applied. Let $(\check{x}, \check{y}, \check{u}, \check{f}, \check{g})$ be the suitably constructed feedforward control problem depending on the *same primitive random variables* as the original problem. The systems are equivalent if $\forall\gamma\exists\check{\gamma}$ such that $P(u = \check{u}) = 1$ and similarly $\forall\check{\gamma}\exists\gamma$ such that $P(u = \check{u}) = 1$.

Assertion 2 *Consider a problem with perfectly classical information pattern. Let F_t be the conditional distribution for x_{t-1} given all the past outputs and applied controls. Then, there is no loss if we restrict our control laws γ_t to be of the form: $\gamma_t = \phi_t(F_t)$ where ϕ_t is a function defined over the (possibly infinite dimensional) space of distributions for x_{t-1}.*

This second assertion states that the conditional law for the state is a sufficient statistic for the purpose of control. Thus, for a perfectly classical information pattern, a clear separation exists between filtering (estimating F_t) and control. Although Witsenhausen does not point this out, it is important to note

[11]This is not enough to know all the conditional distributions for all the random variables in the system. To understand this, consider the following example. For a simple single-input single-output scalar system, suppose $Y = (1, 1)$, $U = (1, 1)$, $L = \emptyset$. Now this is a field basis because knowledge of the control law does not tell us anything more about the underlying probability space than what we already know by seeing y_1, u_1. However, unless we have a control law in hand, we can not talk about the conditional distribution of u_2.

that this assertion rests on the assumption that the primitive random variables are all independent. Without that, we must first explicitly augment the state to capture the dependence before this most basic separation can hold.

It is also important to notice that any nontrivial distributed system will not have a perfectly classical information pattern. This will be brought out sharply in the discussion of the "counterexample" in the next section.

Assertion 3 *For a perfectly classical linear Gaussian system, the conditional distribution of x_{t-1} has a Gaussian version with covariance independent of the data and mean affine in the data.*

The above assertion tells us that in the case of linear Gaussian systems, the filtering problem can be solved (since Gaussian random variables have their distributions parameterized by the mean and covariance) even if we restrict ourselves to time-varying affine functions to do the filtering. However, notice that no assertion is made about the form of the control law ϕ_t. For that, we need some extra assumption on the cost function.

Assertion 4 *For a perfectly classical linear system with quadratic cost criteria*[12] *consider the same system, except with perfect state observation and setting all the primitive random variables v_t to their mean values.*[13] *Let $\phi_t^*(x_{t-1})$ be the (obviously affine) optimal control law for this simpler system except thinking of it as starting at time t with the initial distribution for the state x_{t-1} being a point mass at x_{t-1}. Then, $\gamma_t = \phi_t^*(\bar{F}_t)$ is an optimal control law for the original system where $\bar{F}_t$ is the mean of the conditional distribution for x_{t-1}.*

This assertion represents a phenomenon often called "certainty equivalence". Here, the mean of the conditional distribution is sufficient to determine the optimal control action. The variance just contributes to the expected cost. Notice that here, only the ϕ^* part is affine. But for LQG problems with perfectly classical information patterns, we can combine this assertion with the previous one, and so both the ϕ_t^* and the $\bar{F}_t$ are affine. Thus, so are the optimal γ_t. This is the separation result that we are all most familiar with.

[12]By quadratic cost we mean that the incremental cost functions h_t should be quadratic in state x_t and in the individual controls u_t^i.

[13]Witsenhausen states this assertion subtly incorrectly in his paper. He says to use the same system except "*fixing all the primitive random variables at their mean values.*" This is too much of a restriction. To see this, suppose that all the primitive random variables, which includes x_0, had zero-mean. Then, identically zero control laws ϕ^* would be optimal for this system since everything would be zero. Clearly, this need not be optimal for the original problem!

3 Counterexample

The natural question that arises is whether Witsenhausen is being overly conservative in his separation assertions. For affine control laws to be optimum, do we really need all four of the properties: linear systems, Gaussian primitive variables, quadratic cost, and perfectly classical information patterns? That the LQG part is critical seems clear, but one may have a doubt when it comes to the perfectly classical information patterns. To see this, we consider Witsenhausen's famous "counterexample" [3].

3.1 Problem

The problem is deceptively simple. Stated using the notation above, let us consider the problem (k, σ) as:

- $T = 2$
- x is a scalar, with x_0 Gaussian - zero mean, variance σ^2
- The state transition functions:[14] $x_1 = f_1(x_0, u_1) = x_0 + u_1$ and $x_2 = f_2(x_1, u_2) = x_1 - u_2$
- The output equations: $y_1 = g_1(x_0) = x_0$ and $y_2 = g_2(x_1) = x_1 + w$ where w is a zero mean, unit variance Gaussian random variable.
- The cost expressions: $h_1(x, u) = k^2u^2$, $h_2(x, u) = x^2$
- The information patterns: memoryless[15]: $Y_1 = \{y_1\}; U_1 = \emptyset\ Y_2 = \{y_2\}; U_2 = \emptyset$

Before we proceed to analyze the problem as given, consider what would happen if we had a perfectly classical information pattern. In that case, we could take advantage of the given cost function and achieve zero cost with the following affine control laws: $\gamma_1(y_1) = 0$ and $\gamma_2(y_1, y_2, u_1) = y_1$.

3.2 Affine Controls

We want to now find the best possible affine control laws under the specified information pattern. By inspection, it is clear that since everything has zero-mean, they will be linear.

[14] We follow Witsenhausen's notation here.

[15] Recall that the perfectly classical information patterns for this system would have been: $Y_1 = \{y_1\}; U_1 = \emptyset$ and $Y_2 = \{y_1, y_2\}; U_2 = \{u_1\}$

Let $\gamma_1(y_1) = ay_1 = ax_0$ and $\gamma_2(y_2) = by_2$. Clearly, x_1 will be Gaussian, with zero-mean and variance $(1+a)^2\sigma^2$. So, since h_2 is just x_2^2, it is clear that the optimal $\gamma_2 = \hat{x_1} = E(x_1|y_2)$. So, using the familiar properties of sums of Gaussian random variables, $b = \frac{(1+a)^2\sigma^2}{1+(1+a)^2\sigma^2}$. We can also compute $E(h_2) = E(x_2^2) = E((x_1 - \hat{x_1})^2) = \frac{(1+a)^2\sigma^2}{1+(1+a)^2\sigma^2}$ Now, we have an expression for the expected total cost:

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

To find the minimum of this expression with respect to a, we take its derivative and set it equal to zero. After some simplification, we get the equation:

$$2k^2\sigma^2 a(1+\sigma^2(1+a)^2)^2 + 2\sigma^2(1+a) = 0 \tag{2}$$

We divide through by $2k^2\sigma^2$ and following Witsenhausen, we let $t = \sigma(1+a)$ to get:

$$(t-\sigma)(1+t^2)^2 + \frac{t}{k^2} = 0 \tag{3}$$

Which we can rewrite as

$$\frac{t}{(1+t^2)^2} = k^2(\sigma - t) \tag{4}$$

Now, let us compute them for the case $k = 0.1$, $\sigma = 10$. We can see graphically where the solutions will be in Figures 1 and 2. Numerically, we find that the optimal value for t is 9.899 which results in $a = -0.0101$ and total cost = 0.99.

3.3 Nonlinear Controls

As an alternative, Witsenhausen suggests that we try the nonlinear controllers:[16]

$$\gamma_1(y_1) = -y_1 + \sigma \mathrm{sgn}(y_1) \tag{5}$$

So, at the end of the first stage, x_1 is a two-point distribution at $\pm\sigma$ depending on the sign of x_0.

$$\gamma_2(y_2) = \sigma \tanh(\sigma y_2). \tag{6}$$

[16]Witsenhausen motivates these controllers by showing that this form (with σ replaced with an adjustable parameter a) is optimal if x_0 had been chosen as being $\pm\sigma$ with probability $\frac{1}{2}$ each. In this case, the first control pushes out the state, and the second control is the optimum response to the resulting two-point distribution.

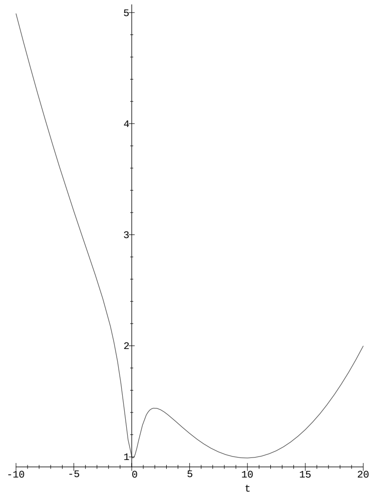

Figure 1: Expected cost vs t parameter

We analyze the resulting expected costs, term by term. $E(h_1) = k^2 E((x_0 - \sigma \mathrm{sgn}(x_0))^2)$. Simplifying this, we get $2k^2\sigma^2(1 - E(|\frac{x_0}{\sigma}|))$. But since $\frac{x_0}{\sigma}$ is just a unit-variance Gaussian, $E(h_1) = 2k^2\sigma^2(1 - \sqrt{\frac{2}{\pi}})$. The second term, $E(h_2) = E(x_2^2)$, can not be evaluated symbolically. But, after some simplifications:[17]

$$E(h_2) = \sigma^2 \int_{-\infty}^{+\infty} \frac{(1 - tanh(\sigma^2 + \sigma w))^2}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw \tag{7}$$

Setting $k = 0.1$ and $\sigma = 10$ as before, we compute numerically that the expected cost is: 0.404. Compare this with the best value possible with affine controllers, 0.99! The nonlinear controller is more than twice as good as the best affine control law.[18]

3.4 "Quantizing" Controllers

We would like to point out that Witsenhausen's example non-linear controllers are unnecessarily confusing — the integrals and hyperbolic functions obfuscate

[17] Witsenhausen simplifies this further, but since we were going to integrate it numerically anyway, there was no point in getting bogged down in additional unnecessary manipulations.

[18] No claim is made for the optimality of this nonlinear controller. In fact Witsenhausen says that we can numerically construct even better nonlinear controllers for this problem.

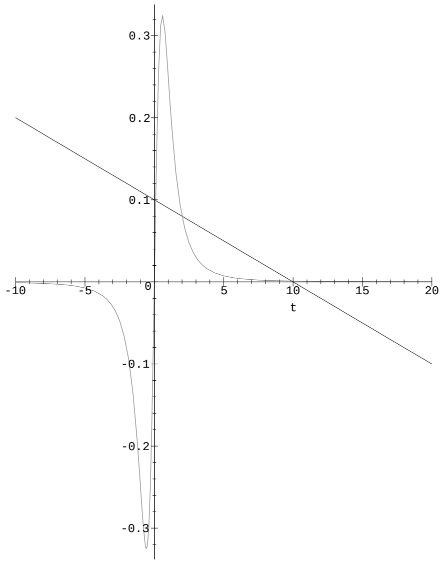

Figure 2: Graphical setting of the first derivative to zero

the essential simplicity of what is going on. Consider the following controller pair that is much clearer and still close to Witsenhausen's pair:

$$\gamma_1'(y_1) = -y_1 + \sigma \mathrm{sgn}(y_1) \tag{8}$$

$$\gamma_2'(y_2) = \sigma \mathrm{sgn}(y_2) \tag{9}$$

We can think of this as a 1-bit quantizer, followed by simple ML 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 $e^{-\frac{\sigma^2}{2}}$ 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 1 simple nonlinear 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" nonlinearities to take advantage of.

Building on the intuition given above, consider the following family of "quantizing" controllers, parametrized by a single number B. [19]

$$\gamma_1^B(y_1) = -y_1 + B\lfloor \frac{y_1}{B} + \frac{1}{2} \rfloor \tag{10}$$

$$\gamma_2^B(y_2) = B\lfloor \frac{y_2}{B} + \frac{1}{2} \rfloor \tag{11}$$

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:

$$k_n = \frac{1}{n^2} \tag{12}$$

$$\sigma_n = n^2 \tag{13}$$

$$B_n = n \tag{14}$$

For our purposes, the analysis of the performance of these controllers is also simple. The first stage cost is $k^2 E((\gamma_1^B(x_0))^2)$ which by inspection can certainly be bounded by $\frac{k^2 B^2}{4}$ since the absolute value of the control is clearly bounded above by $\frac{B}{2}$. Since, $k_n^2 B_n^2 = \frac{1}{n^2}$, 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, that the second stage cost is zero, unless the noise w has magnitude greater than $\frac{B}{2} = \frac{n}{2}$. But since w is Gaussian, this tail event happens with a probability that tends to zero as $e^{-\frac{n^2}{8}}$. So, in the limit of large n, the second stage cost is zero as well. Thus:

$$\lim_{n\to\infty} E(J_n|\gamma^{B_n}) = 0 \tag{15}$$

But what happens to the affine cost? Examining Equation 1, and substituting, we have:

$$E(J_n|\gamma_{\text{affine}}) = a^2 + \frac{(1+a)^2}{\frac{1}{n^4} + (1+a)^2} \tag{16}$$

Clearly,

$$\lim_{n\to\infty} E(J_n|\gamma_{\text{affine}}) = a^2 + 1 \tag{17}$$

[19] This family has an important role to play in another situation as well. Consider the paramaterized pair $(\alpha * \gamma_1^B(y), \beta * (y - \gamma_1^B(y)))$. It can be shown [1] that based on appropriate choices of (α, β, B) this pair of joint source-channel encoders, together with suitable decoders, can achieve higher end-to-end distortion meeting a given power constraint for a 2 dimensional AWGN channel than is possible with the best linear encoding. In fact, as power tends to infinity, the non-linear encoder/decoder's distortion tends to zero faster than the best linear encoder/decoder's distortion.

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

$$\lim_{n\to\infty} E(J_n|\gamma_{\text{bestaffine}}) = 1 \tag{18}$$

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

3.5 Discussion

We have seen that in the case of this particular information pattern, a nonlinear 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[20] and the messenger[21] is what is causing this seemingly strange behavior.

Ideally, what we would like is for the message to be simple (*ie* low entropy = informative prior[22]) 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[23]). 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. If you look at the plot in Figure 1, you will see that the two minima correspond to exactly these two cases. In the one near $t = 0$, the entropy of x_1 is low. In the other one near $t = \sigma = 10$, the power in x_1 is high.

The nonlinear controllers have no such tension and they try to achieve the

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

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

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

[23] 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.

best of both. The resulting x_1 has differential entropy equal to zero[24], and still manages to have significant power — allowing the messenger to be decoded over the noise with a low probability of error. So, the cost can be driven all the way to zero.

4 Conclusion

Fundamentally, we can now say that even through the general stochastic control problem formulation gives us a single cost function for the control objective, there seem to be intrinsically three distinct things going on naturally in the closed-loop system.

1. The first and most obvious is the overt control-objective itself. We want to use information in order to keep the state and control small in some sense.
2. The second is estimation. The system needs to have good estimates of the true state to be able to act. This can be viewed as aggregating information.
3. The third is communication. Different parts of the system need to share information.

The importance of the first two is widely recognized (Dual control, etc.), but the Witsenhausen counterexample effectively shows how a problem with non-classical information pattern really has a strong communication aspect to it. It also showed by example that the class of affine functions may not have sufficient freedom to do a good job in balancing the various factors involved and hence will not lead to optimal solutions. We are currently looking at control problems that explicitly contain a communications channel[2].

References

[1] Anant Sahai. Sending 1 signal over 2 channels. Unpublished Work, 1998.

[2] Sekhar Tatikonda, Anant Sahai, and Sanjoy Mitter. LQG control under communication constraints. Submitted to the 1998 CDC.

[24] We realize that differential entropy is not the best thing to look at in this case, however no matter how you look at it, the signal x_1 constructed by the first stage is very simple — effectively a discrete random variable.

[3] H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal of Control*, 6(1):131–147, 1968.

[4] H. S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, 1971.