A vector version of Witsenhausen's counterexample: Towards convergence of control, communication and computation

Pulkit Grover and Anant Sahai
Wireless Foundations, Department of EECS
University of California at Berkeley, CA-94720, USA
{pulkit, sahai}@eecs.berkeley.edu

***Abstract*— At its core, the renowned Witsenhausen's counterexample contains an implicit communication problem. Consequently, we argue that the counterexample provides a useful conceptual bridge between distributed control and communcation problems. Inspired by the success in studying long block lengths in information theory, we consider a vector version of the Witsenhausen counterexample. For this example, information-theoretic arguments relating to lossy compression, channel coding, and dirty-paper-coding are used to show the existence of nonlinear encoding-decoding control strategies that outperform optimal linear laws and have the ratio of costs go to infinity asymptotically in the vector-space dimension over a much broader range of cost parameters than the previous scalar examples.**

The vector example is then in turn viewed as a collection of scalar random variables with a four-phase distributed control strategy. First a set of agents make observations and communicate with each other to coordinate a first-stage control strategy, then they individually act on their state. A second set of agents now make noisy observations and communicate to coordinate a control strategy, and finally they act on the state again. The vector case can be considered one in which the first and third phase are free. It is thus natural to impose a cost on the length of the first and third phases and this can in turn be viewed as inducing a natural cost function on the information pattern itself.

Inspired by this, we close by considering the simplest possible information-theoretic analog of the problem — lossless compression of a binary state vector. It turns out that the information-pattern can be used as a natural proxy for computational complexity and this gives a new result on the fundamental complexity of lossless compression in terms of the tradeoff between rate, effort, and the probability of error.

I. Introduction

For LQG systems with perfectly classical information patterns, it was well known that control laws affine in the observation are optimal. In [1], Witsenhausen gave an explicit "counterexample" that demonstrated the importance of information patterns in control problems. The counterexample was a chosen distributed control system (and hence a system with a non-classical information pattern) that was otherwise quadratic and Gaussian. For this system, Witsenhausen provided a nonlinear control law that outperformed the optimal linear control law and also demonstrated that a measurable optimal control law should exist.

The counterexample has inspired a large volume of research along three related themes. The first body of work is devoted to finding the elusive optimal control law for the problem. For the simplicity with which the problem is stated, it is interesting to note that the optimal control law is still unknown. In [2], a discrete version of the problem is introduced. This allows for a convex formulation over a set of complicated constraints. However, in [3], the discrete version was shown to be NP complete. In search of an optimal law, a sequence of results were obtained in (amongst other works) [4]–[6] using tools from information theory, neural networks and stochastic optimization respectively. Since the problem is nonconvex, this work has also inspired numerical methods for solving nonconvex problems.

The second theme is in refining the classification of distributed LQG systems into those for which affine laws are optimal, and those for which affine laws are not optimal. In [4], the authors consider a parametrized family of two-stage stochastic control problems. The family includes the Witsenhausen counterexample. The authors show that whenever the cost function does not contain a product of two decision variables, affine control laws are optimal. The authors use results from information theory to arrive at the optimality result. In [7], the author shows that affine controls are still optimal for a deterministic variant of the Witsenhausen counterexample if the cost function is the induced two-norm instead of the expected two-norm in the stochastic variant.

The third theme has been in viewing the counterexample as a bridge between control and communication [8]. In [9], the authors observe that the original Witsenhausen problem is in essence a communication problem between the two controllers. They back up this observation by proposing control strategies that are explicitly based on quantization of the initial state. The strategy is conceptually related to Tomlinson-Harashima precoding (see e.g. [10, Pg.454]) for what is called dirty-paper coding in information theory. The authors then generate a sequence of problem parameters for which nonlinear strategies based on quantization outperform the optimal linear strategies by a factor that tends to infinity. This work inspired a larger body of work that considered explicit (rather than implicit) communication channels connecting the two controllers and took asymptotics in time [11]–[16] and even the idea of implicit communication plays a vital role in [17], [18]. The counterexample itself was revisited yet again in [19] where the author adapts the standard information-theoretic concern with *side-information* into a modified Witsenhausan

problem. The side-information of the initial state is passed through a noisy AWGN channel before being received by the second controller, and is itself subject to an SNR constraint. The author shows that nonlinear schemes still outperform linear ones. In fact, at low SNR, nonlinear schemes that do not make use of the side-information outperform all linear ones, including those that make use of the side information.

It can be argued that the root of all these connections between information theory and control can be traced back to Witsenhausen's counterexample. It might seem that the exploration of connections between information theory and control is mature and no longer needs to consider the counterexample as a bridge. In this work, we challenge that view by returning to the Witsenhausen counterexample. We investigate how tools in information theory, specifically the use of asymptotically long block lengths, can contribute towards improving our understanding of the counterexample. In Section II, we state the vector version of Witsenhausen problem. Assuming the vector length is asymptotically large, in Section III we propose a pair of nonlinear schemes building on the scalar quantization ideas introduced in [9]. The first scheme is based on the information theoretic concepts of lossy compression (vector quantization) and joint source-channel coding. The second is inspired by dirty-paper coding [20]. We show that the proposed schemes outperform all affine schemes as well as the scalar scheme of [9].

The new control schemes as proposed treat the entire vector all at once. While usually accepted without question in information-theoretic circles, this seems aphysical in the context of distributed control. The vector example can be viewed as a distributed collection of scalar random variables with a four-phase distributed control strategy. First a set of agents make observations and communicate with each other to coordinate a first-stage control strategy, then they act on the state, a second set of agents now make noisy observations and communicate to coordinate a second-stage control strategy, and finally act on the state again. It is thus natural to impose a cost on the length of the first and third phases and this can in turn be viewed as inducing a natural cost function on the information pattern itself. In Section IV, we observe that the system is a collection of scalar Witsenhausen problems, with an additional freedom that controllers can send messages (iteratively) to each other in order to perform the encoding at time 1, and decoding at time 2. The Witsenhausen counterexample thus leads naturally to a new information-theoretic problem of understanding the complexity of distributed lossy compression. Building on our work in [21], we formulate a toy lossless source-coding problem to explore the tradeoff between various costs for operating such a distributed system.

This paper does not represent the end of a story, but rather an attempt to demonstrate that the Witsenhausen counterexample still has plenty of life left in it even after 40 years of providing inspiration to control researchers.

II. The problem: distributed control

We generalize the scalar Witsenhausen problem to a vector case. The system is still a two-step control system. The states and the inputs are now vectors of length m. A vector is represented in bold font, with the superscript used to denote a vector length (e.g. $\mathbf{x}^m$). As in conventional notation, x is used to denote states, u the input, and y the observation.

- The state $\mathbf{x}_0^m$ is distributed $\mathcal{N}(0, \sigma_0^2 \mathbb{I})$.
- The state transition functions :

$$\begin{aligned} \mathbf{x}_1^m &= f_1(\mathbf{x}_0^m, \mathbf{u}_1^m) = \mathbf{x}_0^m + \mathbf{u}_1^m, \quad \text{and} \\ \mathbf{x}_2^m &= f_2(\mathbf{x}_1^m, \mathbf{u}_2^m) = \mathbf{x}_1^m - \mathbf{u}_2^m. \end{aligned}$$

- The output equations:

$$\begin{aligned} \mathbf{y}_1^m &= g_1(\mathbf{x}_0^m) = \mathbf{x}_0^m, \quad \text{and} \\ \mathbf{y}_2^m &= g_2(\mathbf{x}_1^m) = \mathbf{x}_1^m + \mathbf{w}^m, \end{aligned}$$

  where $w \sim \mathcal{N}(0, \sigma_w^2 \mathbb{I})$. We assume that $\sigma_w^2 < \sigma_0^2$.
- The cost expressions:

$$\begin{aligned} h_1(\mathbf{x}_1^m, \mathbf{u}_1^m) &= \frac{1}{m} k^2 ||\mathbf{u}_1^m||^2, \text{ and} \\ h_2(\mathbf{x}_2^m, \mathbf{u}_2^m) &= \frac{1}{m} ||\mathbf{x}_2^m||^2. \end{aligned}$$

  The cost expressions are normalized by the vector-length, so that they do not grow with the problem size.
- The information patterns :

$$\begin{aligned} Y_1 &= \{\mathbf{y}_1^m\};\ U_1 = \varnothing, \\ Y_2 &= \{\mathbf{y}_2^m\};\ U_2 = \varnothing. \end{aligned}$$

  Observe that the first controller as assumed to have complete knowledge of $\mathbf{y}_1^m$, and similarly the second controller has complete knowledge of $\mathbf{y}_2^m$. Therefore the system is not completely distributed. Section IV shows that there are computational costs associated with making the system completely distributed.

In the next section we provide a pair of nonlinear schemes that outperform the optimal linear scheme.

III. The schemes: control and communication

In this section we provide a nonlinear coding scheme that is based on the concept of joint source-channel coding in information theory. To enable understanding of the scheme, we review some fundamental results and definitions from information theory in Appendix I. These are taken from [22], and the reader is referred to [22] for further details.

A. The first joint-source channel scheme

We now briefly describe the scheme, before giving a detailed description and analyzing its performance.

As in [9], the idea is to quantize the space of realizations of $\mathbf{x}_0^m$ to arrive at $\mathbf{x}_1^m$. These points are chosen carefully so that with high probability, the second controller can recover $\mathbf{x}_1^m$ from the noisy observation $\mathbf{y}_1^m$. By making $\mathbf{u}_2^m = \mathbf{x}_1^m$, the second controller can now force $\mathbf{x}_2^m$, and hence the second cost, to zero. In the vector case, for a careful choice of points, the probability of error in recovering $\mathbf{x}_1^m$ converges to zero exponentially in m [23]. Therefore, for large enough m, the average cost at time 2 can be made as small as desired.

At time 1, the state of the system is $\mathbf{x}_1^m = \mathbf{x}_0^m + \mathbf{u}_1^m$. We use the following construction to find $\mathbf{u}_1^m$ for each $\mathbf{x}_1^m$. First, we design a rate-R source code for distortion D where $\sigma_0^2 > D > \sigma_w^2$. A random codebook is constructed, with each codeword drawn randomly from distribution $\mathcal{N}(0, \sigma_D^2 \mathbb{I})$ for $\sigma_D^2 = \sigma_0^2 - D$. If the code rate R satisfies

$$R \geq R(D) = \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{D}\right), \tag{1}$$

the average distortion is no greater than D (in the limit). The choice of $\mathbf{u}_1^m$ is the distortion $\hat{\mathbf{x}}_0^m - \mathbf{x}_0^m$, and the resulting $\mathbf{x}_1^m = \hat{\mathbf{x}}_0^m$.

Thus $\mathbf{x}_1^m$, which is the quantized $\mathbf{x}^m$, is itself transmitted across the channel. Since a random Gaussian codebook achieves the channel capacity [22] for an average power constraint equal to the average power of the codebook, the points in the codebook form a good channel code as well. Since these codewords are generated $\mathcal{N}(0, \sigma_D^2 \mathbb{I})$, the average power of the codebook is $\sigma_D^2 = \sigma_0^2 - D$. Therefore, $\mathbf{x}_1^m$ can be recovered reliably at the second controller for rates $R < C$ where

$$C = \frac{1}{2}\log_2\left(1 + \frac{\sigma_D^2}{\sigma_w^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2 - D}{\sigma_w^2}\right). \tag{2}$$

Simplifying the capacity expression,

$$\begin{aligned} C &= \frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2 - D}{\sigma_w^2}\right) \\ &= \frac{1}{2}\log_2\left(\frac{\sigma_w^2 + \sigma_0^2 - D}{\sigma_w^2}\right) \\ &= \frac{1}{2}\log_2\left(\frac{\sigma_0^2 - (D - \sigma_w^2)}{D - (D - \sigma_w^2)}\right) \\ &\geq \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{D}\right) = R(D) \end{aligned}$$

where the last inequality uses the fact[1] that $D > \sigma_w^2$. Therefore, reliable communication is possible at rate $R(D) < R < C$.

Since D is the mean-square distortion $\frac{1}{m}\mathbb{E}\left[||\mathbf{x}_1^m - \mathbf{x}_0^m||_2^2\right]$, it is also the mean-square input required to drive $\mathbf{x}_0^m$ to $\mathbf{x}_1^m$. Therefore, the cost at time 1 is $k^2 D$. Observe that D is only constrained by the inequality $D > \sigma_w^2$. Asymptotically, therefore, the first stage cost is $k^2\sigma_w^2$. Since the error probability converges to zero exponentially in m, for large enough m, the average cost at second stage can be made as close to zero as desired. Therefore, the asymptotic total cost is just $k^2\sigma_w^2$.

B. Another information theoretic scheme

Note that in the above scheme, the cost at the second stage is zero. Dirty-paper coding [20] suggests another scheme for which the second stage cost is not zero.

Observe that the lossy source code reduces the power that is fed into the "channel". This imposes a constraint of $D > \sigma_w^2$. Alternatively, dirty-paper coding techniques [20] in information theory can be thought of as performing a similar quantization, without reducing the power. This suggests that dirty-paper schemes might perform better than the joint source channel scheme. We refer the reader to Costa's original paper [20] for more details. We also observe that due to a different problem formulation, our notation is different from that in [20].

The scheme proceeds by choosing an auxiliary random variable $V \sim N(0, P + \alpha^2\sigma_0^2)$, for some α that will be an optimization parameter. $M = 2^{nT}$ iid sequences are drawn uniformly at random from the set of typical $\mathbf{v}^m$, where[2]

$$T = \frac{1}{2}\log_2\left(\frac{(P + \sigma_0^2 + \sigma_w^2)(P + \alpha^2\sigma_0^2)}{P\sigma_0^2(1-\alpha)^2 + \sigma_w^2(P + \sigma_0^2)}\right). \tag{3}$$

These sequences are then distributed uniformly over 2^{nR} bins. A particular bin is chosen[3]. The encoding is now performed as follows. Given a source sequence $\mathbf{x}_0^m$, a $\mathbf{v}^m$ jointly typical with $\mathbf{x}_0^m$ is first found in the chosen bin. Then the control $\mathbf{u}_1^m$ is chosen as $\mathbf{u}_1^m = \mathbf{v}^m - \alpha\mathbf{x}_0^m$. The received sequence $\mathbf{y}_2^m$ is, therefore

$$\mathbf{y}_2^m = \mathbf{u}_1^m + \mathbf{x}_0^m + \mathbf{w}^m. \tag{4}$$

It is shown in [20] that the decoder (in our case the second controller), can recover $\mathbf{v}^m$ from the received sequence as long as the rate R is smaller than

$$C(\alpha, P) = \frac{1}{2}\log_2\left(\frac{P(P + \sigma_0^2 + \sigma_w^2)}{P\sigma_0^2(1-\alpha)^2 + \sigma_w^2(P + \alpha^2\sigma_0^2)}\right). \tag{5}$$

We are not interested in getting a high rate. However, we want to keep P small, since $k^2 P$ is our cost at time 1. At time 2, the cost is the average mean-square error in estimating $\mathbf{u}_1^m + \mathbf{x}_0^m$. The decoder can recover $\mathbf{v}^m$ with arbitrarily high probability. Now, $\mathbf{v}^m = \mathbf{u}_1^m + \alpha\mathbf{x}_0^m$. By design, $\mathbf{u}_1^m$ and $\mathbf{x}_0^m$ act as if drawn independently. Therefore, we can find the error in estimating $\mathbf{u}_1^m + \mathbf{x}_0^m$ from $\mathbf{v}^m$ by MMSE estimation. This turns out to be

$$MSE = \frac{P\sigma_0^2(1-\alpha)^2}{P^2 + \alpha^2\sigma_0^2}. \tag{6}$$

The total cost is, therefore,

$$k^2 P + \frac{P\sigma_0^2(1-\alpha)^2}{P^2 + \alpha^2\sigma_0^2}. \tag{7}$$

This cost can be achieved only if $C(\alpha, P)$ in (5) is greater than 0. Thus, the optimal cost is obtained by minimizing (7) under the constraint that $C(\alpha, P) > 0$.

Consider $\alpha = 1$. In this case, the MSE cost is zero, so only the first cost is retained. Also,

$$C(1, P) = \frac{1}{2}\log_2\left(\frac{P(P + \sigma_0^2 + \sigma_w^2)}{\sigma_w^2(P + \sigma_0^2)}\right), \tag{8}$$

which is strictly positive at $P = \sigma_w^2$. Therefore, zero second cost is possible for some values of $P < \sigma_w^2$ for this scheme. Since the MSE cost is zero, the net cost is smaller than σ_w^2. Notice that this was not possible for the joint source-channel scheme, where the cost D is constrained to be greater than σ_w^2.

[1] If $a > b > 0$, then $\frac{a-x}{b-x} > \frac{a}{b}$ for all $0 < x < b$.

[2] M corresponds to the mutual information between V and Y_2 [20].

[3] Eventually we will let $R \to 0$, so there's no loss in choosing any particular bin.

C. Comparison with linear and scalar schemes

In this section, we compare the vector scheme with the optimal linear scheme, and the scalar nonlinear schemes in [9].

For simplicity, assume $\sigma_w^2 = 1$. For given value of σ_0^2, the cost for the optimal linear scheme is (from [9])

$$\inf_a k^2 a^2 \sigma_0^2 + \frac{(1+a)^2\sigma_0^2}{1+(1+a)^2\sigma_0^2}. \tag{9}$$

Since $\sigma_w^2 = 1$, the asymptotic cost for the vector joint source channel coding based scheme is $k^2\sigma_w^2 = k^2$. The ratio of the optimal linear cost to the cost for the joint source channel scheme is, therefore,

$$\inf_a \frac{k^2 a^2 \sigma_0^2 + \frac{(1+a)^2\sigma_0^2}{1+(1+a)^2\sigma_0^2}}{k^2}$$
$$= \inf_a a^2\sigma_0^2 + \frac{(1+a)^2\frac{1}{k^2}}{\frac{1}{\sigma_0^2}+(1+a)^2}$$

Now let $k \to 0$ and $\sigma_0^2 \to \infty$. If a is close to 0, the second term is unbounded. If a is close to -1, the first term gets unbounded. For any other value of a, both terms are unbounded.

Thus any choice of sequence (k, σ_0) such that $k \to 0$ and $\sigma_0 \to \infty$, the ratio diverges to infinity. Observe that there is more flexibility in choice of (k, σ_0) as compared to that in [9], where a careful choice has been made. The three schemes, viz. the optimal linear scheme and the two vector nonlinear schemes proposed here are compared in Fig. 1 and Fig. 2.

In Appendix II, we show that the proposed scheme outperforms the scalar nonlinear scheme in [9] by a factor of infinity as well. This is also evident from Fig. 2.

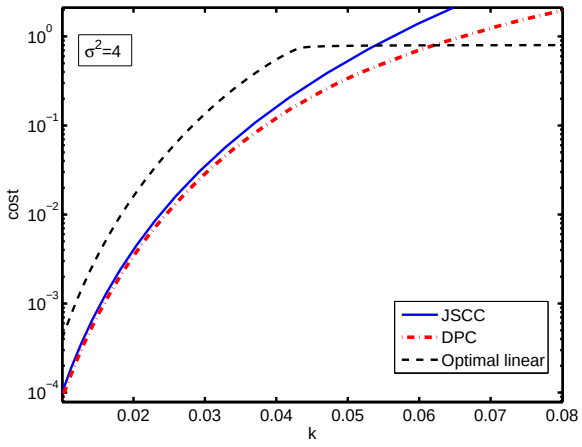

Fig. 1. The figure shows the variation of the total cost with k for $\sigma^2 = 4$. The Joint Source-Channel (JSCC) Scheme and the Dirty-Paper Coding(DPC) scheme perform better at low values of k. At large k, however, the cost at time 1 is larger, therefore the costs for DPC and JSCC schemes increase. However, the cost for linear schemes is still bounded by 1 by choice of $a = 0$.

IV. Redistributing the vector case: from control back to communication and computation

The schemes in Section III raise a natural question: what does it cost to implement such a scheme in a distributed control

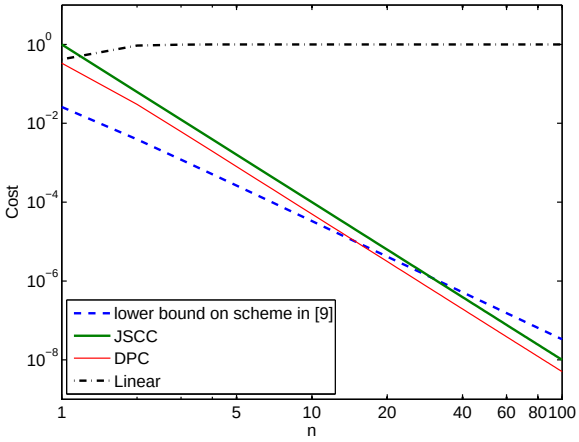

Fig. 2. This figure shows the variation of cost (on a log-log scale) with n, where n is the parameter that characterizes the family of control problems in [9]. Thus, $k_n = \frac{1}{n^2}$, $\sigma_{0,n} = n^2$, and for the scheme in [9], the size of bin $B_n = n$. A lower bound on cost for this scheme is derived in Appendix II. Since slopes for DPC and JSCC costs are better than that for a lower bound on scheme in [9], the ratio of costs for the scheme in [9] and these schemes converges to infinity.

context? While the limit of large block-lengths is justified in information theory by considering it as introducing longer and longer end-to-end delay in an inherently centralized communication problem [24], this is problematic in a distributed control setting where a longer vector seems to suggest a distributed controller acting over a larger geographical area.

It is natural to consider the vector as made up of m scalar Witsenhausen problems. Therefore, a centralized system might be required to perform the encoding, and another for decoding, which is contrary to the spirit of the counterexample. In order to address this, we allow for iterative message-passing before the actions of both sets of distributed controllers. Message-passing algorithms can be performed in a distributed manner, and have complexity that scales linearly with $l \times m$, where l is the number of iterations performed, and m is the block-length. Therefore, the normalized computational cost is only linear in l, and does not scale with block-length m. In addition, the success of sparse-graph codes in coding-theoretic literature, and success of channel coding and source coding techniques based on sparse-graphs gives hope that these codes may exist [25].

Next we describe the encoding and decoding model of this message-passing algorithm. An investigation into costs for the full joint lossy source-channel coding problem posed here is hard, and we are still working on it as it is a new problem in information theory. Based on results in [21], one expects to see some fundamental performance-cost tradeoffs. The schemes proposed in Section III are implemented in a couple of steps. In the first step, a quantization of $\mathbb{R}^m$ is performed. The resulting quantization, can be thought of as a lossy source code, is transmitted across the channel. This is followed by a channel decoding, that fails with an error probability that converges to zero exponentially fast in m. The performance-cost tradeoffs for channel coding can be understood from

the ideas in [21]. However, the problem of performance-cost tradeoffs in lossy source coding is entirely new.

To begin to understand what such tradeoffs could be for source coding, instead of a Gaussian source we analyze a binary source, which has the advantage of discrete alphabet. In information theory problems, lossless source coding is generally easier to understand than lossy source coding. Therefore, we restrict our attention here to lossless source coding. Since a binary source that produces 0 or 1 with equal probability is incompressible losslessly, we consider asymmetric binary sources that produce a 1 with probability $p < 0.5$.

A. The encoding/decoding model

We now describe a message passing model of the encoder and the decoder. The model is inspired by distributed Witsenhausen counterexample. We focus on the distributed nature of the encoding and the decoding. We assume that the encoder is physically made of computational nodes that have communication links with other nodes in the encoder. A subset of nodes are designated 'source nodes' in that each is responsible for storing the value of a particular source symbol in the initial state $\mathbf{x}_0^m$. Another subset of nodes, called the 'coded nodes' has members that are would eventually store the encoded symbols $\mathbf{u}_1^m$. There may be additional computational nodes that are just there to help encode. To arrive at $\mathbf{u}_1^m$, the encoding is performed in an iterative, distributed manner. At the start, each of the source nodes is first initialized with one element of the vector $\mathbf{x}_0^m$. In each subsequent iteration, all the nodes send messages to the nodes that they are connected to. At the end of l_e encoder iterations, the values stored in the coded nodes constitute the encoded symbols $\mathbf{u}_1^m$.

The implementation technology is assumed to dictate that each computational node is connected to at most $\alpha + 1 > 2$ other nodes. No other restriction is assumed on the topology of the decoder. No restriction is placed on the size or content of the messages except for the fact that they must depend on the information that has reached the computational node in previous iterations. If a node wants to communicate with a more distant node, it has to have its message relayed through other nodes.

The neighborhood size of each node at the encoder after l_e iterations, which is the number of nodes it has communicated with, is denoted by $n_e \leq \alpha^{l_e+1}$. The per-node cost associated with the number of iterations is some function $\phi(l)$, that is increasing with l.

The decoding model is analogous, with 'reconstruction nodes' responsible for storing the reconstructed symbols, and another subset of nodes that store the encoding symbols $\mathbf{u}_1^m$.

B. Derivation of lower bound on complexity for lossless source coding

The source generates m symbols that are encoded losslessly into k symbols at rate $R = \frac{k}{m} > h_b(p)$. The encoding and decoding are performed iteratively using a message passing algorithm. Encoding is performed in l_e encoder iterations, and the decoding is performed in l_d decoder iterations. Reconstruction of each bit is performed by using messages from at most α^{l_d} output bits. Each of these output bits depends on at most α^{l_e} source bits. Therefore, each reconstruction is based on a "neighborhood" of $\alpha^{l_e+l_d}$ source symbols (See Fig. 3). We refer to this as the source neighborhood of the particular symbol. Intuitively, an atypical source realization for this local neighborhood of the reconstruction bit should cause errors in the reconstruction.

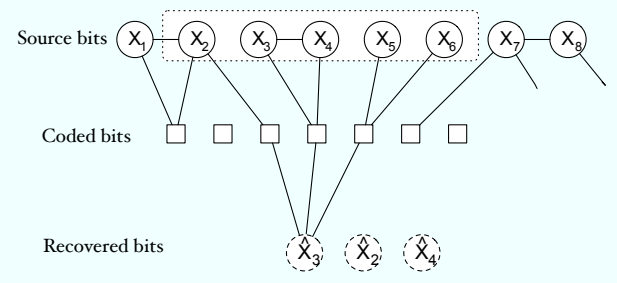

Fig. 3. The dashed box in the figure shows the source neighborhood on one iteration of encoding and decoding for reconstruction bit $\hat{x}_3$. Whether the reconstruction is in error depends only on the source realization in the neighborhood.

The following theorem gives a lower bound on error probability for given size of local neighborhood n. Turned around, these bounds give lower bounds on n, and hence the total number of iterations at the encoder and the decoder $l_e + l_d \geq \log_\alpha(n)$ for given error probability.

Theorem 1: Consider a binary source P that generates iid Bernoulli(p) symbols, $p < 0.5$. Let n be the maximum size of source neighborhood for each reconstructed bit. Then the following lower bound holds on the average probability of bit error

$$\langle P_e \rangle_P \geq \sup_{h_b^{-1}(R) < g \leq \frac{1}{2}} \frac{p h_b^{-1}(\delta(G))}{2} 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-p)} \right)^{\epsilon\sqrt{n}}, \tag{10}$$

where $h_b(\cdot)$ is the binary entropy function, $D(g||p) = g \log_2 \left(\frac{g}{p}\right) + (1-g) \log_2 \left(\frac{1-g}{1-p}\right)$,

$$\delta(G) = h_b(g) - R, \tag{11}$$

$$\epsilon = \sqrt{\frac{1}{K(g)} \log_2 \left(\frac{2}{p h_b^{-1}(\delta(G))} \right)}. \tag{12}$$

and

$$K(g) = \frac{1}{1-2g} \log_2 \left(\frac{1-g}{g} \right). \tag{13}$$

Proof: See Appendix III. ∎

We note that the lower bound in Theorem 1 results look much like that in [21]. Conceptually, the two problems differ only in their source of randomness and the neighborhood.

Observe that the neighborhood here is determined by the number of encoding and decoding operations. This suggests that the encoding costs can be reduced by making the decoding costs larger. We believe this is an artifact of our bounding technique, and is not fundamental to the problem at hand.

C. Tradeoff between control, communication, and computation costs

In Section III, we determined the communication and control costs for the system. The decentralized encoding and decoding framework above allows us to calculate the computation costs.

Let $gap = R - h_b(p)$ to denote the gap from optimality for the lossless source coding problem above. For extremely low error probabilities, analogous to results in [21], we get the following approximate lower bound on the neighborhood size as a function of the error probability and the gap.

$$n \gtrsim K_2 \frac{\log_2\left(\frac{1}{\langle P_e \rangle}\right)}{gap^2}, \tag{14}$$

for some constant K_2 that does not depend on gap and $\langle P_e \rangle$. This lower bound implies that for low computational complexity, the rate R should be at a finite gap from $h_b(p)$. This suggests that for the joint source-channel scheme proposed in Section III, a similar result could hold. That is, to reduce the computational costs, the rate should be bounded away from $R(D)$ for distortion D. Observe that a similar result holds for gap from the channel capacity [21]. Therefore, for optimal costs, the system should be operated at rate $C > R > R(D)$, where R is at a finite gap from both C and $R(D)$. Such an operating point requires that for chosen rate R, the distortion D be strictly larger than $D(R) > \sigma_w^2$, thus leading to higher costs at time 1 than those estimated in Section III.

APPENDIX I
SOME USEFUL INFORMATION THEORETIC CONCEPTS

A. Lossy source coding

Assume that we have a source that produces sequence $\mathbf{x}^m \in \mathcal{X}^m$. The encoder describes the source sequence $\mathbf{x}^m$ by an index $f_m(\mathbf{x}^m) \in \{1, 2, \ldots, 2^{nR}\}$. The decoder represents $\mathbf{x}^m$ by an estimate $\hat{\mathbf{x}}^m \in \hat{\mathcal{X}}^m$.

Definition 1: A distortion function or distortion measure is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+ \tag{15}$$

from the set of source alphabet-reproduction alphabet pairs into the set of non-negative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$.

Definition 2: The distortion between sequences $\mathbf{x}^m$ and $\hat{\mathbf{x}}^m$ is defined by

$$d(\mathbf{x}^m, \hat{\mathbf{x}}^m) = \frac{1}{n} \sum_{i=1}^{m} d(x_i, \hat{x}_i) \tag{16}$$

Definition 3: A $(2^{mR}, m)$ rate distortion code consists of an encoding function,

$$f_m : \mathcal{X}^m \to \{1, 2, \ldots, 2^{mR}\} \tag{17}$$

and a decoding (reproduction) function,

$$g_m : \{1, 2, ..., 2^{mR}\} \to \hat{\mathcal{X}}^m. \tag{18}$$

The distortion associated with the $(2^{mR}, m)$ code is defined as

$$D = \mathbb{E}\left[d(\mathbf{x}^m, g_m(f_m(\mathbf{x}^m)))\right] \tag{19}$$

where the expectation is with respect to the probability distribution on x.

Definition 4: A rate distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{mR}, m)$ rate distortion codes (f_m, g_m) with $\lim_{m\to\infty} \mathbb{E}\left[d(\mathbf{x}^m, g_m(f_m(\mathbf{x}^m)))\right] \leq D$. The *rate-distortion* function $R(D)$ is the infimum of rates R such that (R, D) is achievable for a given distortion D. The *distortion-rate* function $D(R)$ is the inflmum of all distortions D such that (R, D) achievable for a given rate R.

Theorem 2 ($R(D)$ for Gaussian source): The rate-distortion function for Gaussian source $\mathcal{N}(0, \sigma_0^2)$ with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{D}\right), & 0 \leq D \leq \sigma_0^2 \\ 0, & D > \sigma_0^2. \end{cases} \tag{20}$$

The proof of this theorem tells us that this codebook can be constructed by choosing 2^{nR} points independently from $\mathcal{N}(0, (\sigma_0^2 - D)\mathbb{I})$ distribution.

B. Channel coding

Definition 5: An *Additive White Gaussian Noise (AWGN) channel* with an average power constraint consists of a channel input $X \in \mathbb{R}$ and a channel output $Y = X + Z$, where $Z \sim \mathcal{N}(0, \sigma_w^2)$. The input X has an average power constraint P, that is, over m channel uses, $\frac{1}{m}\sum_{i=1}^{m} ||X_i||^2 \leq P$

Definition 6: An (M, m) code for the AWGN channel consists of the following:

1) An index set $\{1, 2, \ldots, M\}$.
2) An encoding function $\mathbf{X}^m : \{1, 2, \ldots, M\} \to \mathbb{R}^m$, yielding codewords $\mathbf{X}^m(1), \mathbf{X}^m(2), \ldots, \mathbf{X}^m(M)$. The set of codewords is called the codebook.
3) A decoding function $g : \mathbb{R}^m \to \{1, 2, \ldots, M\}$, which is a deterministic rule which assigns a guess to each possible received vector.

Definition 7 (Probability of error): Let

$$\lambda_i = \Pr(g(\mathbf{Y}^m) \neq i | \mathbf{X}^m = \mathbf{X}^m(i)) \tag{21}$$

be the conditional probability of error given that i was sent. The average probability of error is defined as

$$P_e^m = \frac{1}{M} \sum_{i=1}^{M} \lambda_i, \tag{22}$$

and the maximal probability of error is defined as

$$\lambda^{(m)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i \tag{23}$$

Definition 8: A rate R is said to be *achievable* if there exists a sequence of $(2^{mR}, m)$ codes such that the maximal probability of error $\lambda^{(m)} \to 0$ as $n \to \infty$.

Definition 9: The *capacity* of a memoryless channel is the supremum of all achievable rates.

Theorem 3 (Channel coding theorem): The capacity for an additive white Gaussian noise channel of noise variance σ_w^2 with an average power constraint P is

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma_w^2}\right) \tag{24}$$

In addition, the error probability converges to zero *exponentially* in m [23], and the capacity can be achieved by a choosing a codebook of 2^{mR} points independently from $\mathcal{N}(0, P\mathbb{I})$ distribution.

APPENDIX II
PERFORMANCE COMPARISON WITH SCALAR SCHEME IN [9]

For the family of problems and the quantization scheme in [9], we find lower bounds on the cost at time 1. We follow the notation of [9] in this section. B^0 is used to denote the 0-th bin (bin that includes the origin), and B is the bin-size.

$$\begin{aligned}\frac{cost1}{k^2} &= \mathbb{E}\left[(\gamma_1^B(x_0))^2\right] \\ &\geq \mathbb{E}\left[x_0^2 \mathbb{1}_{\{B^0\}}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-B/2}^{B/2} x^2 e^{-x^2/2\sigma^2}dx.\end{aligned}$$

For the particular sequence of problem parameters n in [9], the size of n^{th} bin is $B_n = n$, $\sigma_n^2 = n^2$ and $k_n = \frac{1}{n^2}$. Therefore,

$$\begin{aligned}\frac{cost1_n}{k_n^2} &\geq \frac{1}{\sqrt{2\pi n^4}}\int_{-n/2}^{n/2} x^2 e^{-x^2/2n^4}dx \\ &\geq \frac{1}{\sqrt{2\pi n^4}}\int_{-n/2}^{n/2} x^2 e^{-n^2/8n^4}dx \\ &= \frac{1}{\sqrt{2\pi n^4}} 2\int_0^{n/2} x^2 dx \ \times e^{-1/8n^2} \\ &= \frac{n}{12\sqrt{2\pi}} e^{-1/8n^2},\end{aligned}$$

that increases to infinity as $n \to \infty$. In comparison, the joint source-channel scheme proposed in Section III has cost of k_n^2. Thus the ratio $\frac{cost1_n}{k_n^2} = 1$ for the joint source-channel scheme. Hence, the ratio of the costs for the scalar scheme in [9] and the vector scheme proposed here diverges to infinity.

APPENDIX III
PROOF OF LOWER BOUND ON COMPLEXITY FOR LOSSLESS SOURCE CODING

The proof is similar to that for the rate-complexity tradeoff for channel coding over a BSC [21]. In the following, we use P to denote the underlying source that generates symbols distributed Bernoulli(p). G denotes a test source generating symbols Bernoulli(g). We use $\Pr_P(\mathbf{x}^n)$ to denote the probability of a sequence of length n under source behavior P. $\langle P_{e,i}\rangle_{P,1}$ denote the error probability of $i-$th source bit conditioned on it being 1. Similar notation is used for channel G, conditioning on bit 1. $\langle P_{e,i}\rangle_P$ denotes the average error probability

$$\langle P_e\rangle_P = \frac{1}{m}\sum_{i=1}^{m}\langle P_{e,i}\rangle_P \tag{25}$$

We proceed by a sequence of Lemmas.

Lemma 1 (Lower bound on $\langle P_e\rangle$ under test source G): Consider a test source G that generates iid binary symbols distributed Ber(g). If a rate R code is used for lossless coding of G with $R < h_b(g)$, then average probability of bit-error is lower bounded by

$$\langle P_e\rangle_G \geq h_b^{-1}(h_b(g) - R) \tag{26}$$

Proof: For $R < h_b(g)$, consider hamming distortion between the source symbols and the reconstruction. The average hamming distortion is also the average error probability. Using the distortion-rate function for a binary source [22, Pg. 343], the distortion $D(R)$ is bounded below by

$$D(R) \geq h_b^{-1}(h_b(g) - R) \tag{27}$$

■

Let $\mathbf{x}^m$ denote the source sequence. Consider the $i-$th message bit x_i. Its encoding and decoding are based on a particular neighborhood of source symbols $\mathbf{x}^n_{\text{nbd,i}}$ of size n. The encoding is error-free if these neighborhood source symbols $\mathbf{x}^n_{\text{nbd,i}}$ lie in the region $\mathcal{D}_{i,0}$ if the $i-$th bit is 0, and in $\mathcal{D}_{i,1}$ if the $i-$th bit is 1.

Lemma 2: Let $\mathcal{A}$ be a set of source sequences $\mathbf{x}^n$ such that $\Pr_G(\mathcal{A}) = \delta$. Then,

$$\Pr_P(\mathcal{A}) \geq f(\delta) \tag{28}$$

where

$$f(y) = \frac{y}{2}2^{-nD(g||p)}\left(\frac{p(1-g)}{g(1-p)}\right)^{\epsilon(y)\sqrt{n}} \tag{29}$$

is a convex-$\cup$ increasing function of y, and where

$$\epsilon(y) = \sqrt{\frac{1}{K(g)}\log_2\left(\frac{2}{y}\right)}, \tag{30}$$

Proof: Define typical set $\mathcal{T}_{\epsilon,G}$ as follows

$$\mathcal{T}_{\epsilon,G} = \{\mathbf{x}^n \ s.t. \ \sum_{i=1}^{n} s_i - ng \leq \epsilon\sqrt{n}\} \tag{31}$$

Then, as shown in [21, Lemma 9], for

$$\epsilon = \sqrt{\frac{1}{K(g)}\log_2\left(\frac{2}{\Pr_G(\mathcal{A})}\right)}, \tag{32}$$

$$\Pr_G(\mathcal{T}^c_{\epsilon,G}) \leq \frac{\Pr_G(\mathcal{A})}{2}. \tag{33}$$

That $K(g)$ is as in (13) is derived in [26, Prop. 4.2]. Now, under test source G,

$$\begin{aligned}\Pr_G(\mathbf{x}^n \in A) &= \sum_{\mathbf{x}^n\in\mathcal{A}^c}\Pr_G(\mathbf{x}^n) \\ &= \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}}\Pr_G(\mathbf{x}^n) + \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}^c_{\epsilon,G}}\Pr_G(\mathbf{x}^n) \\ &\leq \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}}\Pr_G(\mathbf{x}^n) + \sum_{\mathbf{x}^n\in\mathcal{T}^c_{\epsilon,G}}\Pr_G(\mathbf{x}^n)\end{aligned}$$

Choosing ϵ as in (32), it follows that

$$\sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}}\Pr_G(\mathbf{x}^n) \geq \frac{\Pr_G(\mathcal{A})}{2}. \tag{34}$$

Let $n_{\mathbf{x}^n}$ be the number of ones in $\mathbf{x}^n$. Then,

$$\begin{aligned}
\Pr_P(\mathcal{A}) &= \sum_{\mathbf{x}^n\in\mathcal{A}^c} \Pr_P(\mathbf{x}^n) \\
&\geq \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \Pr_P(\mathbf{x}^n) \\
&= \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \frac{\Pr_P(\mathbf{x}^n)}{\Pr_G(\mathbf{x}^n)} \Pr_P(\mathbf{x}^n) \\
&= \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \frac{p^{n_{\mathbf{x}^n}}(1-p)^{n-n_{\mathbf{x}^n}}}{g^{n_{\mathbf{x}^n}}(1-g)^{n-n_{\mathbf{x}^n}}} \Pr_G(\mathbf{x}^n) \\
&= \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \frac{p^{n_{\mathbf{x}^n}}(1-p)^{n-n_{\mathbf{x}^n}}}{g^{n_{\mathbf{x}^n}}(1-g)^{n-n_{\mathbf{x}^n}}} \Pr_G(\mathbf{x}^n) \\
&= \frac{(1-p)^n}{(1-g)^n} \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \left(\frac{p(1-g)}{g(1-g)}\right)^{n_{\mathbf{x}^n}} \Pr_G(\mathbf{x}^n) \\
&\geq \frac{(1-p)^n}{(1-g)^n} \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \left(\frac{p(1-g)}{g(1-g)}\right)^{ng+\epsilon\sqrt{n}} \Pr_G(\mathbf{x}^n) \\
&= \frac{(1-p)^n}{(1-g)^n} \left(\frac{p(1-g)}{g(1-g)}\right)^{ng+\epsilon\sqrt{n}} \sum_{\mathbf{x}^n\in\mathcal{A}^c\cap\mathcal{T}_{\epsilon,G}} \Pr_G(\mathbf{x}^n) \\
&\geq 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-g)}\right)^{\epsilon\sqrt{n}} \frac{\Pr_G(\mathcal{A})}{2}.
\end{aligned}$$

The function $f(\cdot)$ obtained is the same as that in [21, Lemma 8] for the case of rate-complexity tradeoffs for channel coding over a BSC. Therefore, the proof of convexity and monotonicity of $f(\cdot)$ are the same as that of [21, Lemma 8]. The Lemma then follows from the monotonicity. ∎

Now, to complete the proof of Theorem 1, note that $\langle P_e\rangle_P = p\langle P_e\rangle_{P,1} + (1-p)\langle P_e\rangle_{P,0}$. Conditioned on $x_i = 1$, choose $\mathcal{A} = \mathcal{D}_{i,0}$ in Lemma 2. Then,

$$\langle P_{e,i}\rangle_{P,1} \geq f(\langle P_{e,i}\rangle_{G,1}). \tag{35}$$

Summing from $i = 1, 2, \ldots, m$, diving by m, and using convexity of $f(\cdot)$,

$$\langle P_e\rangle_{P,1} = \frac{1}{m}\sum_{i=1}^{m} f(\langle P_{e,i}\rangle_{G,1}) \geq f(\langle P_e\rangle_{G,1}). \tag{36}$$

Similarly,

$$\langle P_e\rangle_{P,0} \geq f(\langle P_e\rangle_{G,0}).$$

Thus,

$$\begin{aligned}
\langle P_e\rangle_P &= p\langle P_e\rangle_{P,1} + (1-p)\langle P_e\rangle_{P,0} \\
&\geq pf(\langle P_e\rangle_{G,1}) + (1-p)f(\langle P_e\rangle_{G,0}) \\
&\geq f\left(p\langle P_e\rangle_{G,1} + (1-p)\langle P_e\rangle_{G,0})\right) \qquad (37)\\
&\geq f\left(p\langle P_e\rangle_{G,1} + p\langle P_e\rangle_{G,0}\right) \qquad (38)\\
&\geq f\left(p\max\{\langle P_e\rangle_{G,1}, \langle P_e\rangle_{G,0}\}\right), \qquad (39)
\end{aligned}$$

since $p < 1-p$. From Lemma 1, $g\langle P_e\rangle_{G,1} + (1-g)\langle P_e\rangle_{G,0} \geq D_G(R)$. Therefore,

$$\max\{\langle P_e\rangle_{G,1}, \langle P_e\rangle_{G,2}\} \geq D_G(R). \tag{40}$$

The Theorem follows.

REFERENCES

[1] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, Jan. 1968.

[2] Y.-C. Ho and T. Chang, "Another look at the nonclassical information structure problem," *IEEE Transactions on Automatic Control*, 1980.

[3] C. H. Papadimitriou and J. N. Tsitsiklis, "Intractable problems in control theory," *SIAM Journal on Control and Optimization*, vol. 24, no. 4, pp. 639–654, 1986.

[4] R. Bansal and T. Basar, "Stochastic teams with nonclassical information revisited: When is an affine control optimal?" *IEEE Transactions on Automatic Control*, 1987.

[5] M. Baglietto, T. Parisini, and R. Zoppoli, "Nonlinear approximations for the solution of team optimal control problems," *Proceedings of the IEEE Conference on Decision and Control CDC*, pp. 4592–4594, 1997.

[6] J. T. Lee, E. Lau, and Y.-C. L. Ho, "The witsenhausen counterexample: A hierarchical search approach for nonconvex optimization problems," *IEEE Transaction on Automatic Control*, vol. 46, no. 3, 2001.

[7] M. Rotkowitz, "Linear controllers are uniformly optimal for the witsenhausen counterexample," *Proceedings of the 45th IEEE Conference on Decision and Control CDC*, pp. 553–558, Dec. 2006.

[8] Y. C. Ho, M. P. Kastner, and E. Wong, "Teams, signaling, and information theory," *IEEE Trans. Automat. Contr.*, vol. 23, no. 2, pp. 305–312, Apr. 1978.

[9] S. K. Mitter and A. Sahai, "Information and control: Witsenhausen revisited," in *Learning, Control and Hybrid Systems: Lecture Notes in Control and Information Sciences 241*, Y. Yamamoto and S. Hara, Eds. New York, NY: Springer, 1999, pp. 281–293.

[10] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York: Cambridge University Press, 2005.

[11] S. Tatikonda, A. Sahai, and S. K. Mitter, "Control of LQG systems under communication constraints," in *Proceedings of the 37th IEEE Conference on Decision and Control*, Tampa, FL, Dec. 1998, pp. 1165–1170.

[12] A. Sahai, S. Tatikonda, and S. K. Mitter, "Control of LQG systems under communication constraints," in *Proceedings of the 1999 American Control Conference*, San Diego, CA, Jun. 1999, pp. 2778–2782.

[13] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2000.

[14] A. Sahai, "Any-time information theory," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[15] N. Elia, "When Bode meets Shannon: control-oriented feedback communication schemes," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1477–1488, Sep. 2004.

[16] S. Tatikonda, A. Sahai, and S. K. Mitter, "Stochastic linear control over a communication channel," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1549–1561, Sep. 2004.

[17] A. Sahai, "Evaluating channels for control: Capacity reconsidered," in *Proceedings of the 2000 American Control Conference*, Chicago, CA, Jun. 2000, pp. 2358–2362.

[18] A. Sahai and S. K. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. part I: scalar systems," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[19] N. C. Martins, "Witsenhausen's counter example holds in the presence of side information," *Proceedings of the 45th IEEE Conference on Decision and Control CDC*, pp. 1111–1116, 2006.

[20] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[21] A. Sahai and P. Grover, "The price of certainty : "waterslide curves" and the gap to capacity," *Submitted to IEEE Transactions on Information Theory, available online at http://arXiv.org/abs/0801.0352v1*, Dec. 2007.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[23] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley, 1971.

[24] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Trans. Inform. Theory*, may 2008. [Online]. Available: http://www.eecs.berkeley.edu/~sahai/Papers/FocusingBound.pdf

[25] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2007.

[26] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the l_1 deviation of the empirical distribution," Tech. Rep., Jun. 2003. [Online]. Available: http://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.html