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— We argue that Witsenhausen's counterexample provides a useful conceptual bridge between distributed control, communication and computation. Inspired by the utility of studying long block lengths in information theory, we formulate a vector version of the counterexample. Information-theoretic arguments are then used to derive upper and lower bounds on the minimum cost for the vector problem. Restricted to the scalar case, the lower bounds are a strict improvement over Witsenhausen's lower bound for some parameter values. The upper bounds are based on two strategies that asymptotically outperform optimal linear and nonlinear scalar strategies. To investigate the computational aspects of the problem, we then consider a simpler problem of lossless source coding. From a distributed control perspective, the computations required for encoding and decoding can be viewed as internal communication between virtually distributed agents. We derive new lower bounds that establish a tradeoff between the computation, communication and distortion costs for lossless source coding.

I. INTRODUCTION

For LQG systems with perfectly classical information patterns, it was well known that control laws affine in the observation are optimal. To show why distributed control is hard, Witsenhausen gave an explicit "counterexample" in [1] that is a two-step distributed control system that is otherwise quadratic and Gaussian. For this system, Witsenhausen provided a nonlinear control law that outperformed the optimal linear law, and gave a non-constructive proof of the existence of a measurable optimal law. He also gave a lower bound to the minimum cost. However, even for this seemingly simple distributed system, the optimal strategy and the optimal cost are still unknown[1].

In [3], the authors observe that it helps to interpret Witsenhausen's counterexample as a communication problem with an implicit noisy channel. The cost at time 1 can be interpreted as the power that is input to this channel. The cost at time 2 is the distortion in estimating the state at time 1. Using this interpretation, the authors in [3] propose control strategies[2] based on quantization of the initial state. By generating a sequence of problems and providing the appropriate quantization levels, they show that nonlinear strategies can outperform the linear strategies by an arbitrary factor. 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 [5]–[12]. The idea of implicit communication plays only a small (if vital) role in [11], [12].

In this paper, we revisit the interpretation of Witsenhausen's counterexample as an *implicit* communication problem, viewing the counterexample as a bridge between control and communication. The success of information theory in understanding communication problems raises hopes that information-theoretic tools may help in improving our understanding of the counterexample. Information-theoretic formulations often use long block-lengths to allow for the use of the law of large numbers and to avoid the complications associated with the geometry of finite-dimensional spaces. Following this intuition, we formulate a vector version of the counterexample in Section II that is really a collection of scalar Witsenhausen problems with enhanced information patterns. Because it is still a simple two-step distributed control problem where linear strategies are not optimal, the vector extension simplifies the counterexample and yet retains its essence.

This conceptual simplification allows us to obtain a new information-theoretic lower bound on the optimal cost that holds for all vector lengths[3]. This bound is presented in Section III and it is a strict improvement over Witsenhausen's lower bound [1] for some parameter values even when restricted to the scalar case. Section IV then provides information-theoretic control strategies that outperform all linear and nonlinear strategies for some parameter values. The lower bound, together with the schemes, provides a new dimension to the applicability of information theory to distributed control.

Long block-lengths are easily interpreted in standard communication problems as the introduction of an additional delay. Thus information-theoretic results for these problems can be viewed as taking asymptotics *in time*. Introducing additional delay generally worsens the performance in control systems. Instead, it is more natural to take asymptotics *in space* to obtain long vectors. The agents that make the observations are

---

[1]Authors in [2] propose an achievable strategy obtained through ordinal optimization techniques. It is conjectured by some to be very close to optimal.

[2]The strategies are conceptually related to Tomlinson-Harashima precoding [4, Pg.454] for what is called dirty-paper coding in information theory.

[3]Because of the strict space limitations on the final version of this conference paper, proofs of many results appear in [13].

then spatially separated. Thus the computation of the controls is performed in a distributed manner, with agents communicating with each other to enhance their information patterns. This communication between the agents can be interpreted as computation required to perform a collective control operation. It is of interest to understand how this internal computation cost trades off with the external costs of the actuation power and the distortion.

To investigate this tradeoff, we simplify the problem by first making the channel explicit and noiseless. We further simplify the problem by assuming that the source is binary instead of Gaussian. Thus we arrive at a simpler communication problem — lossless source coding, with encoding and decoding performed in a distributed manner. In Section V, we derive a fundamental tradeoff between the computation costs and the performance of a lossless coding system.

This paper is an attempt to demonstrate that to understand general distributed control problems, it is insufficient to understand control, communication and computation in isolation. Instead, there is a need to bring together ideas from all three fields and understand their interplay. Witsenhausen's counterexample provides a good setting to explore these connections, without the clutter associated with more realistic problems.
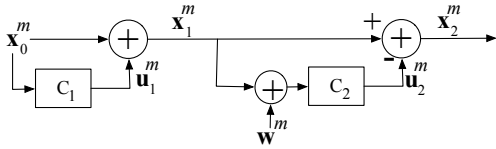
II. VECTOR WITSENHAUSEN PROBLEM

Fig. 1.   The block-diagram of the vector Witsenhausen problem.

We generalize the scalar Witsenhausen problem to a vector case. The states and the inputs are now vectors of length m. The system is shown in Fig. 1. A vector is represented in bold font, with the superscript used to denote a vector length (e.g. $\mathbf{x}^m$). As is conventional, x denotes the 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:

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

- The output equations:

$$\mathbf{y}_1^m = g_1(\mathbf{x}_0^m) = \mathbf{x}_0^m, \quad \text{and} \tag{1}$$
$$\mathbf{y}_2^m = g_2(\mathbf{x}_1^m) = \mathbf{x}_1^m + \mathbf{w}^m, \tag{2}$$

where the observation noise $\mathbf{w}^m \sim \mathcal{N}(0, \sigma_w^2 \mathbb{I})$.
- The cost expressions:

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

The cost expressions are normalized by the vector-length so that they do not grow with the problem size.
- The information patterns (following the notation of [14]):

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

Observe that in (2) there is an implicit channel defined by $\mathbf{x}_1^m = \mathbf{u}_1^m + \mathbf{x}_0^m$ and $\mathbf{y}_2^m$. We denote the average input power at time 1 by $P = \frac{1}{m} E[\|\mathbf{u}_1^m\|^2]$.

III. LOWER BOUNDS ON THE REQUIRED COSTS

In [1, Pg. 145], Witsenhausen provides a lower bound to the optimal cost for the scalar counterexample. However, his argument does not extend to the vector problem of Section II. The following theorem provides a lower bound for the vector problem for any vector length m.

Theorem 1 (Lower bound to the vector problem): The total average cost for the vector Witsenhausen problem in Section II is lower bounded for all $m \geq 1$ by

$$E[h_1 + h_2] \geq \sup_{P \geq 0} k^2 P + \eta(P), \tag{3}$$

where

$$\eta(P) \geq \begin{cases} \left(\sqrt{\kappa(P)} - \sqrt{P} \right)^2 & \text{if } P < \kappa(P) \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where

$$\kappa(P) = \frac{\sigma_0^2 \sigma_w^2}{\sigma_0^2 + P + 2\sqrt{P}\sigma_0 + \sigma_w^2}. \tag{5}$$

Proof: See [13]. We only outline the proof here.
Using the triangle inequality, we show in [13] that

$$\sqrt{E[\|\mathbf{x}_0^m - \mathbf{u}_2^m\|^2]} \leq \sqrt{E[\|\mathbf{x}_0^m - \mathbf{x}_1^m\|^2]} + \sqrt{E[\|\mathbf{x}_1^m - \mathbf{u}_2^m\|^2]}.$$

We wish to lower bound $E[\|\mathbf{x}_1^m - \mathbf{u}_2^m\|^2]$. The first term on the RHS is $\sqrt{P}$, it therefore suffices now to bound to lower bound the term on the LHS. To that end, we interpret $\mathbf{u}_2^m$ as an estimate for $\mathbf{x}_0^m$. The average power that is input to the implicit channel defined by $\mathbf{y}_2^m$ and $\mathbf{x}_1^m$ can be at most $P_{ch} = P + \sigma_0^2 + 2\sqrt{P\sigma_0^2}$. The channel capacity can be upper bounded by $\bar{C}$, the maximum mutual information at power P_{ch}. Then $E[\|\mathbf{x}_0^m - \mathbf{u}_2^m\|^2]$ can be lower bounded by the distortion-rate function $D(R)$ (for the Gaussian source that generates $\mathbf{x}_0^m$) evaluated at rate equal to $\bar{C}$. Denoting this lower bound by $\kappa(P)$, we get the desired expression. ∎

IV. UPPER BOUNDS ON REQUIRED COSTS

In this section we provide two nonlinear strategies for the vector Witsenhausen problem in the limit $m \to \infty$. We only sketch the proofs[4] here and convey the intuition with the help of Fig. 2. The details can be found in [13].

---

[4]The proofs use a random coding argument. A random coding argument is based on generating a strategy randomly according to some distribution. It is then shown that the performance averaged over these strategies is good. Therefore, there exists at least one strategy that has good performance. There exist practical schemes that approach the predicted performance of these strategies. See, for example, [15] for a scheme that approaches the performance of dirty-paper coding.
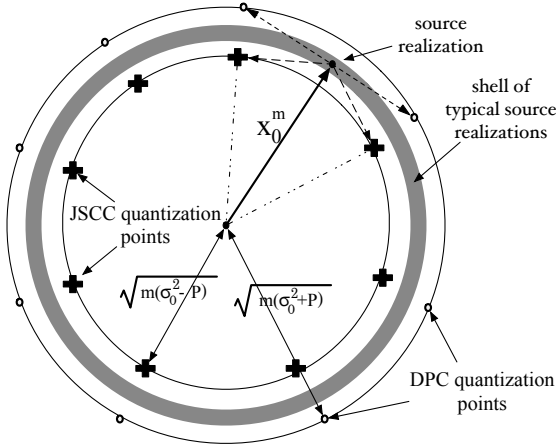
Fig. 2. The figure is a geometric representation of the joint source-channel coding scheme and the dirty-paper coding scheme of Section IV. The grey shell contains the typical source realizations. The JSCC scheme quantizes to points inside this shell. The DPC scheme, on the other hand, quantizes to points outside this shell. For the same input power, the distance between the quantization points of DPC scheme is larger than that for the JSCC scheme, making it robust to larger observation noise variance.

A. *The joint source-channel coding (JSCC) scheme*

This strategy is conceptually similar to vector quantization, and is thus a vector generalization of the scheme in [3]. About $(\sigma_0^2/P)^{m/2}$ quantization points are chosen iid randomly with distribution $\mathcal{N}(0, \sigma_0^2 - P)$. Given a particular $\mathbf{x}_0^m$, the first controller finds the quantization point closest to $\mathbf{x}_0^m$. The input $\mathbf{u}_1^m$ then drives $\mathbf{x}_0^m$ to this quantization point. These points are chosen carefully so that with high probability, the second controller can recover $\mathbf{x}_1^m$ from the noisy observation $\mathbf{y}_1^m$. The scheme is pictorially represented in Fig. 2, with '+' denoting the JSCC quantization points.

The average cost at time 1 is $k^2 P$. We show in [13] that the second controller can recover $\mathbf{x}_1^m$ perfectly as long as $P \geq \min\{\sigma_0^2, \sigma_w^2\}$. Therefore, the asymptotic total cost is $k^2 \min\{\sigma_0^2, \sigma_w^2\}$.

B. *A Dirty-Paper Coding (DPC) scheme*

The vector version of Witsenhausen's counterexample is similar to the communication problem of multiaccess channels with states known to some encoders [16]. The strategy we propose in this section is also similar to that in [16]. In [13] we discuss the difference between the problem addressed in [16] and the vector Witsenhausen problem.

As in the last section, dirty-paper coding (DPC) techniques [17] can also be thought of as performing a quantization. The quantization points[5] are chosen randomly in the space of realizations of $\mathbf{x}_1^m$ according to the distribution $\mathcal{N}(0, \sigma_0^2 + P)$. This is shown in Fig. 2, with 'o' denoting the DPC quantization points. Unlike the JSCC scheme, DPC quantization points increase the power in the state. This suggests that DPC technique can tolerate a higher observation noise variance for the same P.

[5]The number of these quantization points is determined by an optimization [13].

More generally, the DPC strategy recovers an "auxiliary" quantization point $\mathbf{v}^m = \mathbf{u}_1^m + \alpha\mathbf{x}_0^m$, where $\mathbf{u}_1^m$ is distributed $\mathcal{N}(0, P)$ and is statistically independent of $\mathbf{x}_0^m$. The second controller also observes the output of the implicit channel $\mathbf{y}_2^m$. Therefore, the second controller can perform linear MMSE estimation on $\mathbf{v}^m$ and $\mathbf{y}_2^m$ to estimate $\mathbf{x}_1^m = \mathbf{u}_1^m + \mathbf{x}_0^m$. The resulting average total cost is $k^2 P + MMSE$, which can now be optimized over P and α to obtain the minimum total cost.

C. *Comparison with linear and scalar schemes*

The three schemes (the optimal linear scheme and the two vector nonlinear schemes proposed here) are compared in Fig. 3. Also shown is a lower bound derived in [13] for the performance of the scheme in [3]. In Fig. 4, we compare the costs attained by the DPC scheme with the information-theoretic bound of Section III. We also plot the lower bound derived by Witsenhausen [1] for the scalar problem. For many parameter values, as shown in Fig. 4, our bound is tighter.
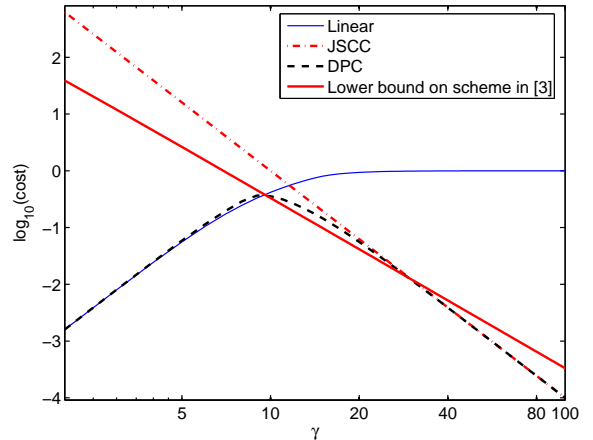

Fig. 3. This figure shows the variation of cost (on a log-log scale) with γ, where γ is the parameter that characterizes the family of control problems in [3]. Thus, $k_\gamma = \frac{100}{\gamma^2}$, $\sigma_{0,\gamma} = 0.01\gamma^2$, and for the scheme in [3], the size of bin $B_\gamma = \gamma$. A lower bound on cost for this scheme is derived in Appendix II of [13]. Since the slopes for DPC and JSCC costs are steeper than that for the lower bound on scheme in [3], the ratio of costs for the scheme in [3] and these schemes diverges to infinity.

V. LOSSLESS SOURCE CODING VIEWED AS A DISTRIBUTED CONTROL PROBLEM

The lower bound of Section III and the two schemes of Section IV give insight into the tradeoff between communication and distortion costs for the vector Witsenhausen problem. However, the schemes assume that a centralized system implements the control operations. From a distributed control perspective, this is unrealistic, especially with asymptotically long vector lengths. Because the delay constraint cannot be tampered with in a control setting, it is more natural to assume that the agents that observe the individual states are spatially distributed. Clearly, these agents need to communicate with each other to compute the control laws. Thus arises the need to understand the tradeoffs between this internal communication
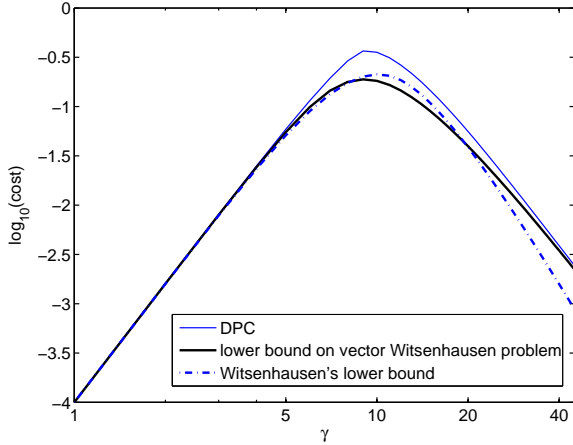
Fig. 4. The figure illustrates the tightness of the DPC scheme in the asymptotic regime of $\gamma \to \infty$ with $k_\gamma = \frac{100}{\gamma^2}$ and $\sigma_{0,\gamma} = 0.01\gamma^2$. Also shown is the lower bound derived by Witsenhausen for the scalar problem. The vector lower bound is tighter than Witsenhausen's bound for large γ. It is shown in [13] that the vector *upper bound* falls below Witsenhausen's lower bound in some cases, thereby showing that the vector strategies can outperform all scalar strategies.

(that can be interpreted as a computation for a collective control operation) and the external power and distortion costs.

Investigating this tradeoff turns out to be hard for the vector Witsenhausen problem. Therefore, we simplify the problem by making the channel both explicit and noiseless, as well as considering a binary source instead of a Gaussian source. We thus arrive at a simpler communication problem — lossless source coding — in the hope that the tradeoffs for this problem may offer insight into the tradeoffs for the vector Witsenhausen problem.

The lossless source coding problem is as follows. A source generates k iid binary symbols $\mathbf{s}^k$ according to a Bernoulli distribution where the probability of each symbol being 1 is p (we denote this distribution Ber(p)). At the encoder, there is a set of agents, each of which observes a source symbol. These agents then communicate with each other to encode $\mathbf{s}^k$, computing the codeword $\mathbf{c}^m$. At the decoder, the codeword $\mathbf{c}^m$ is observed by a set of agents. These agents then communicate with each other to estimate the source symbols.

Next we describe a model for performing distributed encoding/decoding. We then provide a lower bound on the tradeoff between the computational complexity, the error probability, and the gap from the optimal rate. The result suggests that there is a fundamental tradeoff between the three quantities for the vector Witsenhausen problem as well.

A. The encoding/decoding model

We assume that the set of agents communicate by an iterative message-passing algorithm. For consistency with the message-passing literature [18], the agents can be thought of as nodes in a graph. The encoder is physically made of computational nodes that are connected to each other using local communication links. A subset of nodes are designated 'source nodes'. Each source node is responsible for storing

an element of the source vector $\mathbf{s}^k$. Another subset of nodes, called the 'coded nodes' has members that will eventually store the encoded symbols $\mathbf{c}^m$. There may be additional computational nodes that are there just to help encode. To arrive at $\mathbf{c}^m$, 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 $\mathbf{s}^k$. 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 $\mathbf{c}^m$.

The decoding model is analogous, with a subset of nodes called 'coded nodes' storing the received codeword $\mathbf{c}^m$, and the 'reconstruction nodes' responsible for storing the reconstructed symbols $\hat{\mathbf{s}}^k$.

The implementation technology or the underlying physical topology is assumed to dictate that each computational node is connected to at most $\alpha + 1 > 2$ neighboring nodes. Because we are interested in deriving lower bounds on the required comptuation, no other restriction is assumed on the topology of the encoder/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, its message must be relayed through other nodes. After $l_e \geq 1$ encoder iterations, the "neighborhood" size of each node at the encoder, which is the number of nodes it has communicated with, is bounded above by $(\alpha + 1)\alpha^{l_e}$.

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

Encoding is performed for $l_e \geq 1$ encoder iterations, and the decoding is performed for $l_d \geq 1$ decoder iterations. Reconstruction of each source symbol is performed by using messages from at most $(\alpha + 1)\alpha^{l_d}$ coded symbols. Each of these coded symbols depends on at most $(\alpha + 1)\alpha^{l_e}$ source symbols. Therefore, each reconstruction is based on a neighborhood of $(\alpha + 1)^2\alpha^{l_e + l_d}$ source symbols (see Fig. 5). We refer to this as the source neighborhood of the particular reconstructed symbol. Intuitively, errors in reconstruction occur when the source realization in this neighborhood is atypical.

We assume that the source neighborhood of the i−th source symbol includes the i−th source node. If this were not the case, the error probability would be p, which is larger than the lower bound we use in the derivation. Denote the maximum
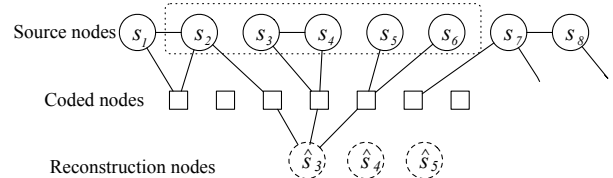

Fig. 5. The dashed box in the figure shows the source neighborhood after one iteration each of encoding and decoding for reconstruction symbol $\hat{s}_3$. Whether the reconstruction is in error depends only on the source realization in the neighborhood.

of size of the source neighborhood by $n + 1$. The following theorem gives a lower bound on the average error probability $\langle P_e \rangle$ for given $n + 1$. Turned around, these bounds give lower bounds on the maximum neighborhood size $n + 1$, and hence a lower bound on the total number of iterations $l_e + l_d$ as a function of $\langle P_e \rangle$ and R.

Theorem 2: Consider a binary source P that generates iid Ber(p) symbols, $p < 0.5$. Let $n + 1$ be the maximum size of the source neighborhoods for the reconstructed symbols. Then the average probability of bit error

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

where $h_b(\cdot)$ is the base-2 binary entropy function, $D(g||p) = g \log_2 \left(\frac{g}{p} \right) + (1-g) \log_2 \left(\frac{1-g}{1-p} \right)$, and $\delta(G) = h_b(g) - R$,

$$\epsilon = \sqrt{\frac{1}{K(g)} \log_2 \left(\frac{2}{ph_b^{-1}(\delta(G))} \right)}, \qquad (6)$$

and

$$K(g) = \frac{1}{1 - 2g} \log_2 \left(\frac{1-g}{g} \right). \qquad (7)$$

Proof: See Appendix I. ∎

To understand the result better, let $gap = R - h_b(p)$ denote the gap from the theoretically optimal rate. For extremely low error probabilities we get the following approximate lower bound on the neighborhood size as a function of the error probability and the gap.

$$n \gtrsim K_2 \frac{\log_2 \left(\frac{1}{\langle P_e \rangle} \right)}{gap^2}, \qquad (8)$$

for some constant K_2 that does not depend on gap and $\langle P_e \rangle$. Thus the bound implies that for low computational complexity, the rate should not be too close to $h_b(p)$.

We note that the lower bound in Theorem 2 and the approximation in (8) are similar to those in [19, Thm. 4] for channel coding. Unlike in [19], where the neighborhood size is determined solely by the number of decoding operations, here it is determined by the number of decoding *and encoding* operations. This seems to suggest 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.

VI. DISCUSSION AND CONCLUSIONS

Information theory provides fundamental limits to the system performance that can be attained asymptotically. We argue that taking complexity into account, one must, in fact, operate at a certain gap from these fundamental limits. This is shown for the lossless source coding in Section V-B, and for channel coding in [19].

For the vector Witsenhausen counterexample, consider the JSSC scheme. The tradeoff for lossless source coding suggests that for low computational costs, the operating rate R should be far from $R(D)$, at a distortion D that is not too small. Similarly, the gap $C - R$ should be substantially large.

Thus, low computational cost would require high input power (to increase C) and high distortion (to decrease $R(D)$) in order to have substantial gaps $R - R(D)$ and $C - R$. This suggests a tradeoff between the computation, distortion, and power costs should exist for the vector Witsenhausen problem.

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

The proof is similar to that for the performance-complexity tradeoff for channel coding over a BSC [19]. In the following, we use P to denote the underlying source that generates symbols distributed Ber(p). G denotes a test source generating symbols Ber(g). We use $\Pr_P(\mathbf{s}^n)$ to denote the probability of a sequence of length n when the underlying source is distributed according to P. $\langle P_{e,i} \rangle_{P,j}$ denotes the error probability of the $i-$th source symbol conditioned on it being $j \in \{0, 1\}$. The average error probability for the $i-$th symbol is $\langle P_{e,i} \rangle_P = p \langle P_{e,i} \rangle_{P,1} + (1-p) \langle P_{e,i} \rangle_{P,0}$, and

$$\langle P_e \rangle_P = \frac{1}{k} \sum_{i=1}^{k} \langle P_{e,i} \rangle_P, \qquad (9)$$

is the average error probability over the source symbols. $\langle P_e \rangle_{P,0}$ and $\langle P_e \rangle_{P,1}$ are defined analogously.

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 the average probability of bit-error

$$\langle P_e \rangle_G \geq h_b^{-1}(h_b(g) - R) =: D_G(R). \qquad (10)$$

Proof: Follows from the distortion-rate function $D_G(R)$ for a Ber(g) source [20, Pg. 343]. ∎

Given the source sequence $\mathbf{s}^k$, the encoding and decoding of $i-$th source symbol s_i are based on a particular neighborhood of source symbols of size $n + 1$. We denote the neighborhood of s_i *excluding the source symbol* by $\mathbf{s}_{\mathrm{nbd},i}^n$. We assume that the encoding is error-free if $\mathbf{s}_{\mathrm{nbd},i}^n$ lies in the region $\mathcal{D}_{i,0}$ when the $i-$th symbol is 0, and in $\mathcal{D}_{i,1}$ when the $i-$th symbol is 1.

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

$$\Pr_P(\mathcal{A}) \geq f(\delta) \qquad (11)$$

where

$$f(y) = \frac{y}{2} 2^{-nD(g||p)} \left(\frac{p(1-g)}{g(1-p)} \right)^{\eta(y)\sqrt{n}} \qquad (12)$$

is a convex-∪ increasing function of y, and where

$$\eta(y) = \sqrt{\frac{1}{K(g)} \log_2 \left(\frac{2}{y} \right)}, \qquad (13)$$

with $K(g)$ as in (7).

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

$$\mathcal{T}_{\epsilon,G} = \{\mathbf{s}^n \ s.t. \ \sum_{i=1}^{n} s_i - ng \leq \epsilon\sqrt{n}\}. \tag{14}$$

Then, as shown in [19, Lemma 9], for

$$\epsilon = \eta\left(\Pr_G(\mathcal{A})\right), \tag{15}$$

$$\Pr_G(\mathcal{T}_{\epsilon,G}^c) \leq \frac{\Pr_G(\mathcal{A})}{2}. \tag{16}$$

Now, under test source G,

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

Choosing ϵ as in (15) and using (16), it follows that

$$\sum_{\mathbf{s}^n \in \mathcal{A} \cap \mathcal{T}_{\epsilon,G}} \Pr_G(\mathbf{s}^n) \geq \frac{\Pr_G(\mathcal{A})}{2}. \tag{17}$$

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

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

The function $f(\cdot)$ obtained is the same as that in [19, 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 [19, Lemma 8]. The lemma then follows from monotonicity. ∎

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

$$\langle P_{e,i} \rangle_{P,1} \geq f(\langle P_{e,i} \rangle_{G,1}). \tag{18}$$

A similar result holds for conditioning on $s_i = 0$. Averaging over the source bits, and using the convexity of $f(\cdot)$,

$$\begin{aligned} \langle P_e \rangle_P &= p\langle P_e \rangle_{P,1} + (1-p)\langle P_e \rangle_{P,0} \\ &\geq pf(\langle P_e \rangle_{G,1}) + (1-p)f(\langle P_e \rangle_{G,0}) \end{aligned}$$

$$\begin{aligned} &\geq f\left(p\langle P_e \rangle_{G,1} + (1-p)\langle P_e \rangle_{G,0}\right) \\ &\geq f\left(p\langle P_e \rangle_{G,1} + p\langle P_e \rangle_{G,0}\right) \\ &\geq f\left(p \max\{\langle P_e \rangle_{G,1}, \langle P_e \rangle_{G,0}\}\right), \end{aligned}$$

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,

$$\max\{\langle P_e \rangle_{G,0}, \langle P_e \rangle_{G,1}\} \geq D_G(R). \tag{19}$$

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] J. T. Lee, E. Lau, and Y.-C. L. Ho, "The Witsenhausen counterexample: A hierarchical search approach for nonconvex optimization problems," *IEEE Trans. Automat. Contr.*, vol. 46, no. 3, 2001.

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

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

[12] 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.

[13] P. Grover and A. Sahai, "Vector Witsenhausen problem as assisted interference cancelation," *Submitted to International Journal on Systems, Control and Communications (IJSCC)*, 2008. [Online]. Available: http://www.eecs.berkeley.edu/~sahai/

[14] H. S. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proceedings of the IEEE*, vol. 59, no. 11, pp. 1557–1566, Nov. 1971.

[15] S. ten Brink and U. Erez, "A close-to-capacity dirty paper coding scheme," in *Proceedings of the 2004 IEEE Symposium on Information Theory*, Chicago, USA, Jun. 2004, p. 536.

[16] S. Kotagiri and J. Laneman, "Multiaccess channels with state known to some encoders and independent messages," *EURASIP Journal on Wireless Communications and Networking*, no. 450680, 2008.

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

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

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

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