Actions can speak more clearly than words

Pulkit Grover

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2011-1
http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-1.html

January 5, 2011

Acknowledgement

Vodafone fellowship (2005-07), NSF grant CNS-0932410

Actions can speak more clearly than words

by

Pulkit Grover

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering — Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anant Sahai, Chair
Professor Pravin P Varaiya
Professor Andrew EB Lim

Fall 2010

Actions can speak more clearly than words

Abstract

Actions can speak more clearly than words

by

Pulkit Grover

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Anant Sahai, Chair

Shannon theory tells us how to communicate explicit sources across explicit channels. However, systems in nature and human society are rife with examples where neither the source nor channel is explicit, and actions, not words, appear to "speak." This phenomena of what we can call *implicit communication* is little understood in the theory of control, and little explored in theory of communication. Consequently, almost no engineering systems systematically exploit implicit communication. In this dissertation, using toy models, we first argue that dramatic improvements could be possible in control precision and control costs with proper use of actions that communicate.

Theoretically, implicit communication has proven to be a hard nut to crack. From a control stand-point, implicit communication makes problems hard because the same actions that are traditionally used exclusively for control can now communicate as well. From a communications view, there is often another conceptual difficulty: since the source is not specified explicitly, the message can be altered by control actions!

Consequently, even the minimalist toy problem that distills these two difficulties — the infamous Witsenhausen counterexample — has remained unsolved for the past four decades. Worse, its discrete counterpart is known to be NP-complete, ruling out the possibility of an algorithmic solution. Since the problem is hard as well as minimalist, it is a bottleneck in understanding implicit communication in particular and decentralized control in general.

The main contribution of this dissertation is two-fold. First, using a sequence of three simplifications of the counterexample, we release this bottleneck by providing the first provably approximately-optimal solutions to the Witsenhausen counterexample. Second, we generalize this sequence of simplifications and propose them as a program for addressing more complicated problems of decentralized control. As an indication of the potential success of this program, we provide approximately-optimal solutions to various problems where implicit communication is possible. Using our refined understanding of implicit communication, we also identify a few practical situations where the phenomena may prove useful.

To my first teachers: my mother, my father, Mohnish, and nana.

Contents

Acknowledgments

It has been wonderful being a student of Anant Sahai for the last five years. His excitement for learning anything new is infectious, and I hope to keep this spirit with me wherever I go next. Anant let me work on *two* interdisciplinary problems across *three* fields, and on the way I learned a lot from him about the art of problem formulation, the necessity of being a good scholar, and the need for asking bold, insightful questions. His thoughtful criticism and constant encouragement have helped me immensely in my research, technical writing, and presentation, and I hope they keep coming in the years ahead. I am also grateful to him for taking out time during his sabbatical for giving me valuable feedback on this dissertation.

Thanks to the collaborators who helped develop results in this dissertation: Anant, Aaron Wagner, Gireeja Ranade and Se Yong Park. This dissertation has also been helped immensely by discussions with Venkat Anantharam. I have also learned quite a lot about writing and presenting by watching David Tse. His conviction in simplicity and the power of intuition has influenced me and this dissertation in many ways, some more obvious than others. Pravin Varaiya's deep insights and knowledge of the field, as well as timely help and encouragement helped shape this dissertation. He also helped me connect my work with practice. Andrew Lim's references in finance and economics have broadened the scope of my research, and are the source of discussions on economics in this dissertation. I was also fortunate to have Kameshwar Poolla on my qualifying exam committee — digging for answers to questions raised by him proved to be an extremely valuable exercise. Interactions with Sanjay Lall, Aditya Mahajan, Nuno Martins, Mike Rotkowitz, and Serdar Yüksel helped understand the bigger picture. Towards the end of my grad school, I was fortunate to have interactions with Tamer Başar, John Doyle, Sanjoy Mitter and Chris Sims whose perspectives have opened my mind to many possibilities.

Thanks also to my collaborators in the work on green communication, which does not appear in this dissertation: Bora Nikolic, Hari Palaiyanur, Jan Rabaey, Matt Weiner, and Kris Woyach (and of course Anant), and to Elad Alon and Zhengya Zheng for introducing me to decoder and receiver implementations. Thanks also to the great group of students and faculty at BWRC, interactions with whom broadened my perspective significantly.

I am lucky to have the most wonderful of friends. Many of them were there for me at a time when not all was going well. In particular, I am extremely grateful for the encouragement from Vinod Prabhakaran, Amin Gohari and Bobak Nazer, and the support from Asako Toda, that I received at the time. I am also thankful to Amin for being a wonderful sounding board for technical ideas. I fondly remember seemingly endless discussions with him on life and universe and the theory of mind. Alex Dimakis's profound simplicity and Anand Sarwate's breadth of interests are much admired. Over the last couple of years, Se Yong Park's insights on decentralized control helped develop my own perspective. Gireeja Ranade's constructive criticism, friendship, and infectious enthusiasm continue to be invaluable. Collaboration with Hari Palaiyanur was fruitful and rewarding, and he has made my life here fun in many ways. Multiple hours spent with Kris Woyach in understanding im-

plicit communication have affected more than just this dissertation, and I look forward to continuing discussions on life and universe with her. Her war cry "you can do it!" in the last couple of months worked wonders on dull days.

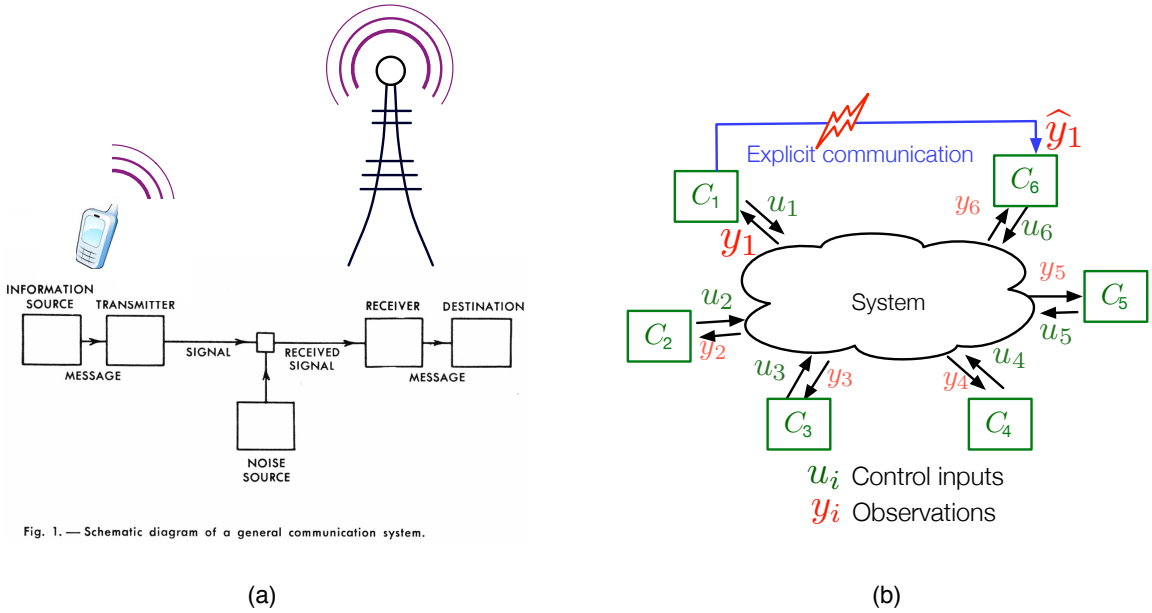
Wireless foundations is a great place to be, and it was made even better by Amy Lee and Kim Kail who have helped me through the departmental work. I particularly want to acknowledge the help of Cheng Chang, Sudeep Kamath, Mubarak Mishra, Hari Palaiyanur, Gireeja Ranade, Anand Sarwate, Rahul Tandra, and Kris Woyach for technical discussions and giving feedback on my papers. Thanks to Kris, Venkatesan Ekambaram, Lav Varshney, Hari, Sudeep and Gireeja for feedback on my dissertation, and to Ruth Gjerde and Dana Jantz for helping me with paperwork.

I am most grateful to my loving mother, father and brother: wonderful people who are leading inspirational lives in ways more than just professional. I will soon join the rest of my immediate family — Dr. Chander Kumar Grover, Dr. Neerja Grover, Dr. Mohnish Grover, and Dr. Shruti Grover — as a doctor, albeit of a different kind (one who does not cure diseases yet). Only Nehal Grover is not a doctor (another reason why she's so special!), and she's also the only one who is yet to celebrate her first birthday. I dedicate this thesis to my family as a token of gratitude for their encouragement, and to my nana (grandfather) Mr. P.C. Jain, for the boldness in his thoughts, words, and actions that will continue to inspire me for the rest of my life.

Chapter 1

Introduction

1.1 Communication for decentralized control



(a) (b)

Figure 1.1: (a) An example of explicit communication. The source (voice) and the channel (wireless) are explicitly specified. Shannon's model of point-to-point communication is a good abstraction for the problem (the block-diagram is taken from Shannon's original paper [1]). (b) A decentralized control system. Multiple agents act on a control system. Is Shannon's model still a good approximation to "communication" in this system?

Fig. 1.1(a) shows an example of what we call problems of *explicit communication*. These are the traditional problems of communication where the goal is to have the encoders commu-

nicate given messages across communication channels to the decoders. The modern theory of explicit communication started with Shannon's seminal work [1], where he says,

> *"Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem."*

In other words, Shannon's intent was to address problems where communication could be viewed as a goal in itself. As communication systems get better integrated in our daily lives and activities (see "cyber-physical systems" [2]), of increasing engineering interest are problems where communication is a *means* to a goal. For instance, consider the problem of facilitating coordination among agents in a decentralized control system (e.g. a set of robots assembling a car, sets of sensors/actuators keeping a building temperature comfortable and uniform, nanobots detecting or killing a tumor etc.). Does the theory of explicit communication allow us to facilitate this coordination? Fig. 1.1(b) illustrates one possible way: if the designer has the engineering freedom of attaching external channels between agents, one can hope to simulate a centralized system by disseminating the observations of various agents quickly and reliably over these external channels. The theory of explicit communication tells us how to engineer this communication so that this hope can be realized.

Practically, one cannot always engineer external channels of arbitrarily high capacity to connect these controllers. In the extreme case of nanobots, for instance, electromagnetic communication can be extremely expensive to engineer and run because the size of a nanobot (smaller than a micrometer) would require an extremely high frequency[1] thereby consuming more power. Not surprisingly, chemical communication techniques that use existing chemicals in the body have been proposed[2] for communicating implicitly between nanobots [3]. Even when an external channel is feasible, the use of these channels assumes a conceptual "separation" that is reflected best in Witsenhausen's words [4]:

> *"[the information transmission theory] deals with an essentially simple problem, because the transmission of information is considered independently of its use"*

By taking away the meaning of the message, explicit communication separates it from the act of communication, thereby potentially over-simplifying the problem[3]. Natural (and human) interactions suggest an alternative way to forge this coordination without necessarily resorting to external channels.

[1] To stick a dipole antenna on a nanobot would require a frequency of about 10^{14} Hz, which lies in the visible spectrum!

[2] This is discussed at greater length in Chapter 1.4.2.

[3] Shannon's intent was to separate the *semantic* content of the source (which can be subjective to the observer for whom the message is intended) from the engineering problem of communication. In case when the meaning is measured by per-letter distortion, Shannon's source-channel separation theorem [1] in information theory shows that this separation of semantic content and communication does not have any performance penalty and thus can be viewed as an optimal strategy. More precisely, lossy-compression of the source (which is done, for instance, in JPEG images) followed by reliable communication across the channel can attain the same asymptotic end-to-end distortion as would any other optimal scheme.

1.2 When actions speak: *implicit* communication

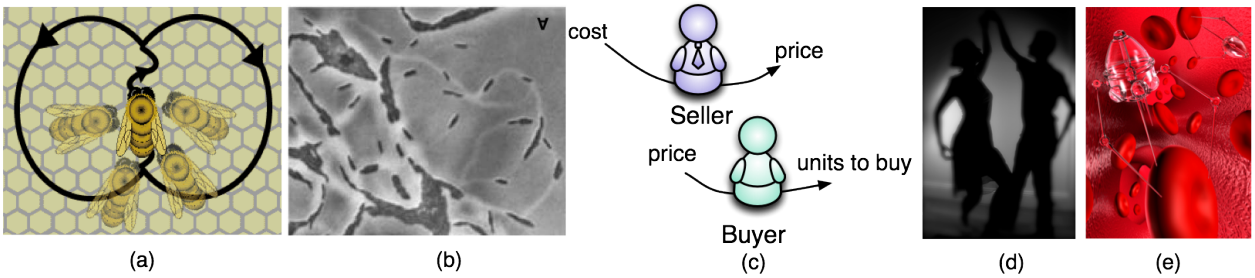


Figure 1.2: Examples of interactions in nature, economics, and human society that do not fit the mold of explicit communication. (a) The waggle dance of bees that indicates location of food and quantity, (b) The slime-trails of myxobacteria that help other bacteria glide, (c) An example from economics where the seller communicates cost through the price to the consumer, and (d) Two dancers communicate implicitly using body contact and motion. (e) An engineered system where nanobots are flowing in the bloodstream. Literature on nanobots proposes chemical communication between them [3].

While explicit communication is used commonly in engineered systems, natural systems often appear to develop coordination without explicit communication. Consider the rather fanciful example of ballroom dancing (Fig. 1.2(d)). Even though the dancers do not use the verbal channel, they are coordinated while dancing. Evidently, the dancers are communicating to each other in some fashion. Looking at this communication closely, it is apparent that the leader in the dance communicates to the follower using the 'channel' of body contact and motion, and the follower responds with movements while simultaneously signaling back through the contact. But what are these agents communicating? The 'message' itself can evolve as the dance proceeds with the moods of the dancers. It is therefore possible that the communication 'message' can be affected by the control actions themselves. Clearly, communication in dancing cannot be cast in the mold of explicit communication: the message source and the communication channel are specified only implicitly. Similar implicit specification of sources and/or channels can be observed in many examples of natural and human interactions (see Fig. 1.2; these are discussed in greater detail in the next section).

Taking inspiration from these examples, we informally define *implicit communication*, or communicating using actions, by contrasting it with explicit communication. Problems of implicit communication are those problems that possess any one of the following two features (a) *implicit sources/messages*: where control agents use actions to generate messages endogenously, and/or (b) *implicit channels*: where agents use control actions to communicate through the plant (*i.e.* the implicit channel) while simultaneously using these actions to control the same plant.

1.2.1 Implicit communication in natural systems

Although our definition of implicit communication is at the moment mathematically imprecise, it helps classify and distinguish the nature of implicit communication in examples of natural and human interactions. For instance, when ants crawl, they leave trails of pheromone along their path. These trails are strengthened by other ants following the pheromone trail [5]. The chosen path is an implicit *message* because it is generated by an agent (i.e. the ant).

Similarly, honey bees (see Fig. 1.2(a)) are known to perform wiggly motions[4] with their abdomen, and walk in semi-circles, to communicate[5] the location of the food to the other bees at the hive [7]. Even though it appears that the interaction of bees in this waggle dance is a form of implicit communication, the communication message, namely the food location, is not implicit because it corresponds to the actual location of food which is specified exogenously, and cannot be modified by the bees. Even so, the "channel inputs" are determined by the control actions of the bees: the act of moving in semi-circles and dancing with their abdomen. Since the same "control plant," namely the feet (and to a lesser extent, the abdomen), are used for locomotion in general, the channel can be thought of as implicit[6].

Is there any example of a natural system which exhibits both of these notions of implicit signaling? Sure enough! Bacteria that live in cultivated soils, called myxobacteria, provide a ready example [8]. Much like pheromones for ants, slime secreted by a myxobacterium signals its path to other bacteria. The mode of signaling works differently — the slime that these bacteria secrete aids the motion of the other bacteria by allowing them to glide over it. Just as for ants, these bacteria communicate an endogenous, implicit message. Since slime, which is meant to aid other bacteria for gliding, is also serving the purpose of signaling to them the chosen path, it acts as an implicit channel.

1.2.2 Implicit communication in human society

Natural systems are not the only ones exhibiting implicit communication. Our day-to-day life is rife with examples of implicit communication. Games of cards often involve implicit communication between partners, where the implicit channel is the cards being played and the act of viewing these cards. In the game of contract bridge, the act of bidding can also be viewed as implicit communication between the players. Even though the bids

[4] See [6] for a beautiful video!

[5] Karl von Frisch was one of the first to translate the meaning of the waggle dance, and he received a Nobel prize for this work in 1973.

[6] This example also brings out the fact that while identifying implicit messages is straightforward, identifying implicit channels can sometimes be a matter of interpretation. Conceptually, however, the identification of channels as implicit is important. The fact that the same actions serve a dual purpose: that of control and determining the input to the implicit communication channel, is one of the features that is widely believed to make decentralized control hard. The issue is discussed at length in Chapter 3.

are made verbally, they help determine the cost (winning or losing) while simultaneously communicating messages about the bidder's cards to other players.

We all know that the textbook way of signaling while driving is signaling explicitly. An indicator indicates a lane-change or a turn, brake-lights indicate slow-downs, explicit hand signals can be used to indicate intent, etc. Even so, real-world traffic uses implicit communication extensively: a gentle movement to your right indicates a desire to change lanes, tailgating urges the driver in front to be faster, tapping brakes suggests a traffic slow-down (through decreased velocity, or through flashing of brake-lights), perhaps even an accident[7]. These are valuable pieces of information that are available that perhaps semi-automated systems, or even completely automated ones, may make use of. Similar possibilities for implicit communication exist in all decentralized systems where the agents have partial observations of the state. For instance, submarines used in coordinated search missions [9], robots moving articles in a warehouse [10], ship-maneuvering [11], etc., all have the potential for implicit communication.

Mathematical modeling of implicit communication in human interactions can be difficult because the goal of the interacting agents may not be readily quantifiable. However, simplistic situations in economics offer us platforms where the modeling of implicit communication may be easier. Not surprisingly, implicit communication has received significant attention in the economics literature, most notably in the Nobel-prize-winning [12] work of Spence [13], where it is referred to as 'signaling.' The problem addressed by Spence is that of job-market signaling, where the candidate signals his or her ability using, for example, the level of education that the candidate has received. The implicit channel is the level of education. One can think of the ability as an implicit message, because it can be *enhanced* (and hence modified) by the action of getting education[8]. One such problem of implicit communication inspired by signaling in economics is addressed in Chapter 5.5.

In the control-theoretic literature, the term 'signaling' was first used[9] by Ho, Kastner and Wong [16,17]. Ho and Kastner [17] also connect the control-theoretic notion to Spence's signaling model in a game-theoretic formulation, where they consider a toy *stochastic* version of Spence's job-market signaling problem.

[7]The language and the extent of such implicit communication depends on the place and the traffic-culture!

[8]This idea of 'enhancement' in job-market signaling came out of discussions with Prof. Varaiya and is a simplification of Spence's original model. Spence's model instead takes an approach that we can call behaviorist. It also allows for enhancement of 'signals', that is, the alterable quantities such as education-level (as opposed to 'indices', e.g. race, gender, etc. — the unalterable ones). But the correlation between signals (and indices) and 'productivity' is based on the experience of the employer. The tussle between the signaling job-candidates and the observing employer becomes a dynamic game where beliefs of the employer change with their consistency with the actual productivity of the hired candidate.

[9]Signaling as a role of control actions seems to have first appeared in works of Witsenhausen [14, 15]. In [15], Witsenhausen thinks of the entire system as a "communication channel" with control inputs and the state as the inputs to the channel, and observations as the output.

1.3 Exploring implicit communication through toy problems

What should be our starting point for exploring implicit communication? We take inspiration from the history of the modern theory of explicit communication which started with a simple toy problem — that of communicating a message from one point to another [1]. Emulating the beginnings of explicit-communication theory, in this dissertation we focus on toy problems that will help us study one or more aspects of implicit communication in isolation. A fortuitous advantage of looking at toy problems is the following: in simple problems such as these, it is possible to compare the costs for implicit communication with those attained using an explicit communication-based architecture. These problems can thus be used as experiments that provide hints about when implicit communication might offer a useful engineering alternative in practical situations. This comparison is done in Chapter 2.

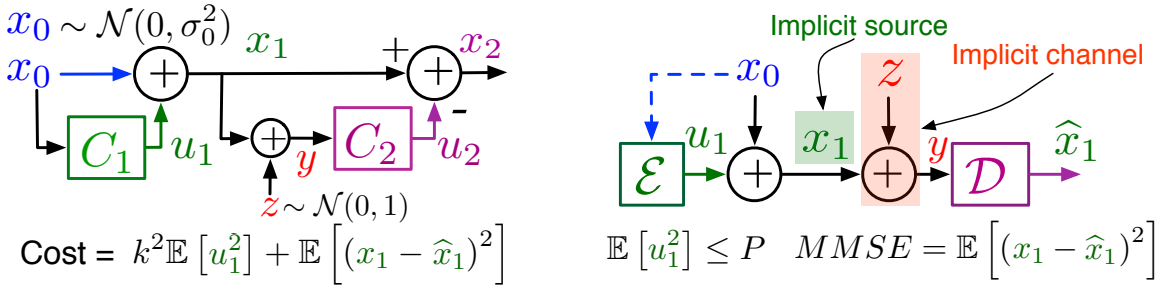


Figure 1.3: The Witsenhausen counterexample and an equivalent implicit communication interpretation.

Our first toy problem is the minimalist problem that exhibits both an implicit source and an implicit channel: the Witsenhausen counterexample [14]. An implicit communication interpretation of the counterexample is shown in Fig. 1.3, which brings out the implicit source and the implicit channel in the counterexample.

In Chapter 2, for an estimate of the performance of implicit communication, we use a strategy where the control input is used to quantize the initial state at the first controller (this strategy was developed by Witsenhausen [14] and extended by Mitter and Sahai [18]). Using this strategy, in Chapter 2, we compare the engineering alternatives of implicit and explicit communication. It turns out that when high precision is required in estimation, this quantization-based implicit communication strategy can significantly outperform the *optimal* explicit communication strategy. Naturally, one would want to know an optimal implicit communication strategy for the counterexample and for problems in its neighborhood. This desire motivates a deeper investigation of the phenomena of implicit communication, which forms the core of the dissertation.

Unfortunately, despite its minimalist simplicity, finding an optimal strategy for the Witsenhausen counterexample is an infamously hard problem [19]. At the same time, its mini-

malist nature demands that any satisfactory theory of implicit communication must have a good understanding of the counterexample. Figuratively, the problem is located just outside the boundary of what is thought to be the set of "tractable" control problems. Significant research effort has been invested into understanding what makes the problem hard [20–24][10], and into obtaining brute-force solutions to the problem [25–27]. In fact, we argue in Chapter 3 that the hardness of the problem influenced the development of decentralized control[11] — problem formulations carefully avoided the possibility of implicit communication. This motivates our exploration of the Witsenhausen counterexample in Chapter 4, culminating in the first approximately-optimal solutions for the problem.

Building on this understanding, we explore quite a few other problems of implicit communication which are detailed in Chapter 1.5.

1.4 How can these toy problems give insights into practical system design?

Suppose a designer wants to design a decentralized control network. Why does exploring toy problems help? While toy problems may not be directly applicable to the real-world, they allow us to distill aspects of real-world problems and study them in isolation. The 'toyness' of the problem is really just a proposed separation of the "grain from the chaff," *i.e.* the essence of the problem from the details (that needs to be tested by taking the insights back into real-world). When faced with the problem of designing a large system, the designer breaks down the problem into sub-problems each of which is *inspired* by one or more toy problems. For instance, once the idealized point-to-point toy problem of explicit communication was well understood, it was natural to ask if one could make larger communication systems work. Interference is a consequence of having a larger system, and one needs to know how to deal with interference. An initial justification for treating this interference as merely "chaff" (*i.e.* detail) came from the observation that the worst-case interference distribution is the familiar Gaussian noise [28].

Subsequent refined understanding has shown that this strategy that ignores interference as a mere detail can lead to arbitrarily large gaps from the optimal attainable rate [29]. Nevertheless, the toy model of point-to-point communication found its utility in practice (e.g. in early CDMA systems), and laid the foundation for studying the more complex interference problem.

In order to integrate the point-to-point solution into a network, there are still many details that are unresolved. For instance, which transmitter is the message coming from, which receiver is it intended for, what is the packet size, etc. For simplicity of design, the

[10] A detailed historical survey of the problem and its hardness is provided in Chapter 3.

[11] This historical perspective is based on discussions with Prof. Anant Sahai on his own involvement with the development of the field.

network is split into various 'layers.' A layered structure helps because it abstracts away details of, for instance, addressing from the designers of information theoretic strategies[12].

In the same spirit, a layered architecture for decentralized control networks was proposed by Varaiya in [31]. The architecture abstracts previous approaches for highway traffic [32], traffic surveillance [33], etc., and tacitly uses explicit communication for coordination. Could implicit communication suggest an alternative architecture?

Consider the concrete problem of controlling traffic flow by designing an automated highway traffic systems (see, for example, Varaiya's proposal for smart cars [32]). Varaiya's proposed layered architecture is shown in Fig. 1.4.

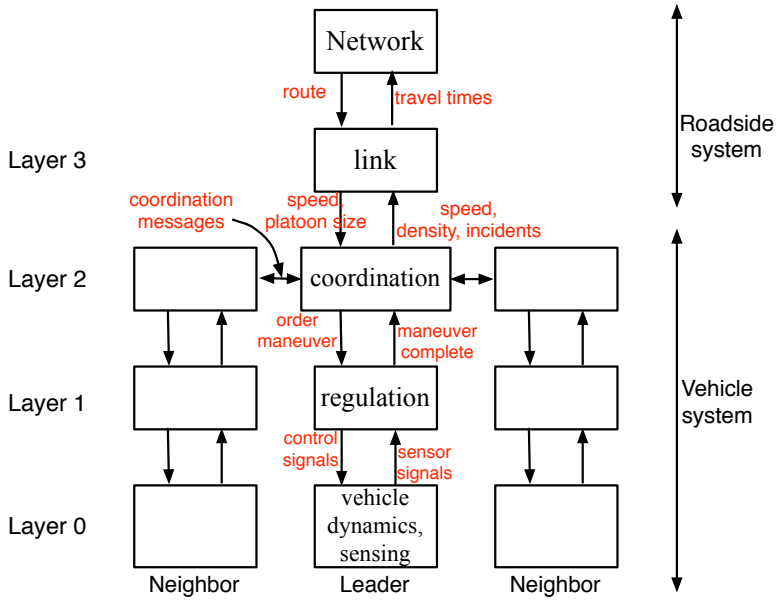


Figure 1.4: Varaiya's layered approach to decentralized control [31] exemplified in the architecture of automated traffic control using 'smart cars' on 'smart roads' [32].

What do these layers do? Of our interest are Layers 0, 1 and 2 that deal with the car and its neighbors. The lowermost Layer 0 is open loop: it receives control signals from Layer 1, the regulation layer, and implements the dynamics. It also 'senses' the environment, which is a broad term that could include sensing the relative position and the velocity of each car in the neighborhood. It also passes these observations to Layer 1. The regulation layer, Layer 1, is responsible for completing maneuvers successfully for which it uses feedback of sensor observations from Layer 0.

[12]This separation is a conceptual simplification, and theoretical justifications [30] are few and unsatisfactory. These abstractions are useful even when such a separation is suboptimal because they provide a yardstick to compare cross-layer strategies with.

Layer 1 receives its maneuver commands from Layer 2, the coordination layer. This coordination layer communicates with its peers in neighborhood to determine which maneuver (e.g. lane change, exit/entry into highway) to execute to fulfill its goal, which is reaching an exit.

1.4.1 Coordination using explicit communication

One way to build coordination in the coordination layer is to connect the cars using external wireless channels. Results from the theory of explicit communication, suitably adapted, can be used to exchange information (e.g. location, velocity, intent of lane-change etc.) at the coordination layer. For instance, consider the case of transmitting location of one car to another. How can we model the movement of a car? In the moving frame of reference of our car, other cars can be modeled as performing one dimensional random walks along the highway (possibly with a drift), with occasional perpendicular motion for lane changes. The lane-changes can be communicated easily using traditional techniques, and the small frequency also requires only low rates of communication. The random-walk in the direction of motion is harder to communicate, but communicating a random walk to within a bounded moment is precisely the problem addressed by Sahai [34].

As we noted, explicit communication inspires the architecture shown in Fig. 1.4. The coordination layer uses explicit communication to help coordinate with the neighboring cars. Based on messages from neighboring cars, it orders maneuvers to the regulation layer, tacitly separating communication from control.

1.4.2 Coordination using implicit communication: a modified layered architecture

Is the separation between communication and control assumed by explicit communication strategy necessary? We noted earlier that the examples from real-life traffic: sideways movement while changing lanes, tapping breaks for a slowdown, etc. are all arguably examples of implicit communication. Taking inspiration from these examples, let us first speculate if we can make the *channel* implicit. Since sensors could replace eyes in automated systems, a natural way of making the channel implicit is to use the sensors to communicate messages.

Can the cars communicate implicitly in the architecture shown in Fig. 1.4? As noted, the separation between control and communication aspects of explicit communication is reflected in the layered architecture: the sensors are used by the regulation layer for completing the maneuvers ordered by at the coordination layer. However, because the regulation layer sends only the one bit message: "maneuver complete," to the coordination layer, the coordination layer receives no message about the other cars from the regulation layer. Consequently, the cars cannot coordinate using the sensors-based implicit channel in the layered architecture of Fig. 1.4, even though our human experience from driving suggests that sensors can likely be used for implicit communication.

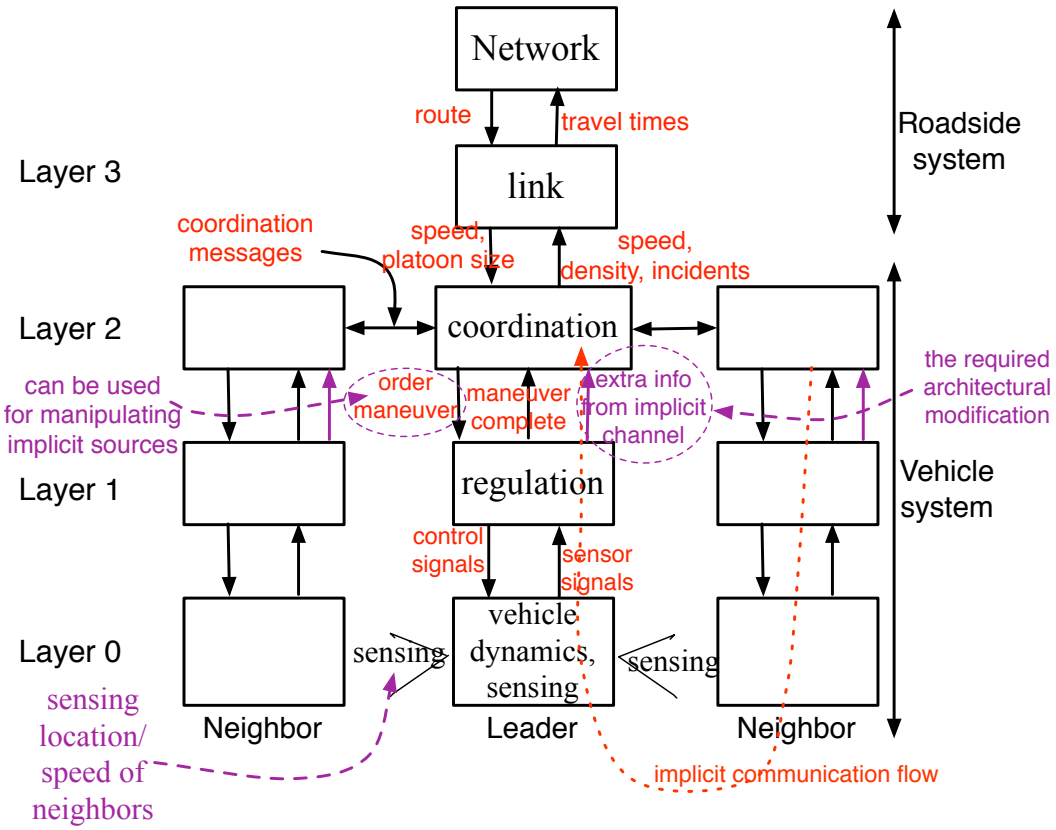


Figure 1.5: The required architectural modifications to the layered structure of Fig. 1.4 that allow for the actuator-sensor implicit communication. The higher the "SNR" on the implicit channel, the more the "extra" information (about an implicit or explicit source) that a car can communicate to its neighbor using the implicit channel. Even though the architecture of [31] tacitly assumed explicit communication for coordination, no architectural change is required for making the *source* (e.g. location of the car) implicit as long as it is available at the coordination layer.

A modification to the layered architecture of Fig. 1.4 that allows for the use of the implicit channel through these sensors is shown in Fig. 1.5. If the sensor observation noise is small, the cars can exchange more information to attain improved coordination. On the other hand, a large sensor noise (for instance, when the conditions are foggy) will reduce the influence of the implicit channel (which is also what happens in current non-automated traffic).

What can the cars communicate through this implicit channel? The coordination messages of lane-change or slowing down can be communicated just as in real-world driving today. Let us speculate if we can make the *sources* implicit as well. Can the controller, *i.e.* a car, affect the sources? Coming back to the example of communicating the location of the car, it is clear that the source (in this case the location itself) can be modified by the maneuvers at the coordination layer. When could such source-modification be useful? One possibility utility is "source-simplification," *i.e.* simplifying the source so that the error in source-estimation is smaller. Looking at real-world driving, lane-driving can be thought of as a form of source-simplification where the source is the location of the car, and it is simplified by forcing it to exist in "quantized" lanes. This simplified source can be estimated more easily by other drivers. A source-simplification such as this could also be performed in automated systems. Even if explicit communication channels are available, the source-simplification can help reduce the required rate across these channels.

Indeed, recent automated robotic systems for warehouse management (see Fig. 1.6) actually use source-simplification to communicate implicitly to the neighboring robots. Consequently, the explicit communication overhead[13] is quite small (about 50 bits-per-second [35]). At what point is there value to using implicit communication along with explicit communication? This question is explored in Chapter 2.

1.5 Main contributions

1.5.1 Substitutes for certainty-equivalence: semi-deterministic abstractions

The dominant conceptual framework for designing control strategies in the face of uncertainty is the theory of "certainty-equivalence." What is certainty-equivalence theory? At its core, this theory suggests separating estimation and control[14] by splitting each agent into an estimator followed by a controller. The controllers first arrive at a strategy by pretending that the observations is noiseless and the system state is known perfectly (*i.e.* and hence with "certainty"). The estimators use the observations to estimate the state and feed the estimates into the controllers. The controllers use these estimates as inputs to the strategy

[13] We believe that the goal is to reduce communication as well as computational overhead. Path-planning for robots could become algorithmically simpler to implement if the robots move on a grid rather than everywhere in the space.

[14] See [4] for an excellent survey on the separation of estimation and control.

Figure 1.6: A warehouse (of Kiva systems) where mobile robots move packages for delivery (used with the permission of Prof. D'Andrea of ETH Zurich). The cost of collision is huge, and therefore accurate estimation of the location of neighboring robots is a must. The chosen strategy, that of having the robots move on a grid in the space, can be thought of as making the source implicit. The robots are also equipped with sensors which together with movement of other robots can be thought of as creating implicit channels for this implicit source. Indeed, the movement of robots on the grid is so precise that they leave tell-tale tracks on the warehouse floor. We will see in Chapter 2 that implicit communication is specially useful when the required precision in estimation is high, thus substantiating the source-simplification used here.

obtained from the fictional noiseless version of the system.

This conceptually simpler design based on separation of estimation and control is optimal in quite a few interesting centralized cases [4, 36], including centralized LQ systems [4, Assertion 7]. What strategy does certainty-equivalence suggest for a decentralized system? If unobserved states are thought of as partial observations with extremely large observation noise, then a noiseless version of the system corresponds to all the controllers having complete knowledge of the state. This certainty-equivalence strategy will therefore be no different than what would be suggested if the system were a centralized one.

Suboptimality of certainty-equivalence for decentralized LQG problems

Agents in a decentralized system usually have different observations. There is therefore a strong temptation for the controllers to communicate among one another in order to simulate a centralized system. A certainty-equivalence approach suggests connecting these controllers using external channels: the controllers can now communicate over this channel and thereafter simulate a centralized system. However, real-world channels are imperfect, and simulating a centralized system may come at a very high communication cost. In order to understand the impact of imperfect external channels, we need to step back and understand the limiting case when external channels are absent.

Even though certainty-equivalence-based strategies are optimal for centralized LQG systems, the Witsenhausen counterexample shows that these strategies can be far from optimal for decentralized LQG systems[15]. There is a philosophical and pedagogical value to understanding *why* this suboptimality is present — the cause is intimately tied to implicit communication. Bar-Shalom and Tse [36] showed that certainty-equivalence-based strategies are suboptimal whenever control actions have a *dual role*: that of minimizing immediate costs, and reducing uncertainty in future estimation.

For instance, for linear systems, what difference can the inputs make in the posterior distribution of the state? If the system is centralized, the inputs can only affect the mean of the distribution, so the intuitive uncertainty in the state does not change. However, in decentralized systems, it is plausible that a controller with less noisy observations can reduce the uncertainty in the observations of the controllers that follow. Witsenhausen's counterexample demonstrates not only that this reduction in uncertainty is possible, but that it can really help. While certainty-equivalence suggests linear strategies for the problem[16], Mitter and Sahai [18] showed that nonlinear strategies that reduce uncertainty in state estimation can outperform linear strategies by an arbitrarily large factor.

A semi-deterministic model

While the appeal of the theory of certainty-equivalence is its simplicity, the Witsenhausen counterexample exposes the fact that it is not always applicable to decentralized control problems. There is essentially no theory to guide the design of decentralized control policies when 'signaling' or the dual role of control is a possibility[17]. Therefore, we propose a substitute for certainty-equivalence theory in Chapter 4 and Chapter 5. The substitute theory is one of semi-deterministic abstractions that are based on the recently proposed binary deterministic models for Gaussian network-communication problems [37–39]. Just as the deterministic model in information theory captures the flow of information in communication networks, our model might be able to capture the flow of information in networks of *implicit* communication as well.

Our abstractions have one notable modification: to capture the dual effect of control, we include influence of noise[18] which is why the abstractions are *semi*-deterministic. To demonstrate the applicability of these models, we show that they are useful in finding the first

[15]For the counterexample, quantization-based strategies can outperform certainty-equivalence-based strategies by an arbitrarily large factor (an observation that was first made by Mitter and Sahai [18])

[16]We shall see in Chapter 2 that certainty-equivalence does not even suggest the best *linear* strategy for the counterexample.

[17]An optimization perspective does not work for these problems: as we will see in Chapter 3, even the simplest of these problems, Witsenhausen's counterexample, is NP-complete (when discretized).

[18]In the original model of [37,39], the part of the signal below the noise level was ignored: in communication, these least-significant bits are indeed unimportant because they are mangled by noise. In control systems, however, these bits can be affected by controllers with better observations. Removing them from the model will bring us back to certainty-equivalence-based strategies.

provably-approximately-optimal solution to the Witsenhausen counterexample and many other problems of implicit communication.

1.5.2 Witsenhausen's counterexample: a provably-approximately-optimal solution

Based on our proposed semi-deterministic model, a fundamentally new approach to addressing Witsenhausen's counterexample forms the core of Chapter 4. Accepting that finding the optimal strategy is too hard, we instead ask for an approximate solution. However, an approximate solution is a meaningful solution only if it is known how far it could be from the optimal cost. Inspired by the approximation results obtained using the information-theoretic deterministic model (see [37]), we seek a similar approximation that is provably uniform over all problem parameters. Our approximate-optimality results thus have the following flavor: we characterize the control costs to within a constant factor that is uniform over all the choices of problem parameters[19]. The reason for considering a constant factor, instead of the other natural comparison using constant differences, is simple: the "costs" for most of these problems (as traditionally normalized) are bounded, and decrease to zero in certain limits.

Our approximate solution to Witsenhausen's counterexample is uniform over k and σ_0, the parameters of the counterexample, and the vector length m. The solution is obtained in a sequence of four steps:

1. The semi-deterministic abstraction of the problem is posed and addressed first. The optimal strategies for the semi-deterministic abstraction (which are based on quantization-based strategies complemented by linear strategies) are hypothesized to be good strategies for the LQG problem as well. The next three steps bring us to the original LQG problem.

2. As a first test experiment for our hypothesis, the strategies for the deterministic version are lifted to a variation on Witsenhausen's counterexample where the noise is uniform (instead of the Gaussian noise in the original LQG formulation). Quantization-based strategies (complemented by linear strategies) are shown to attain within a constant factor of the optimal cost for all problem parameters, thus completing the first experiment.

3. Our second experimental setup is an asymptotically infinite-length vector version of Witsenhausen's counterexample. The techniques developed for the uniform-noise counterexample extend naturally to this setup, proving approximate-optimality of natural extensions of the same strategies.

[19]The counterpart of this approximation in information theory is to obtain capacity within a constant number of bits (an additive approximation), which is equivalent to obtaining the required power within a constant factor at high SNR for most problems. Constant-factor approximations are also used for approximating solutions to NP-hard problems [40].

4. Arriving finally at the experimental setup of the original (scalar) counterexample, techniques from large-deviation theory are used to prove approximate optimality of these strategies for the scalar case and all finite-length vector extensions.

A few points of our approach and the approximately-optimal strategies themselves are notable:

- In contrast with the "observer-controller" formulation of tracking over an explicit communication channel [34, 41–43], the problem formulations here have two crucial differences. The observer is now not merely an observer, it can control too. The controller is not merely a controller either; it has direct (but noisy) observations of the channel itself. Not only can these enhancements lower costs (as shown in Chapter 2), it is difficult to work in the cost minimization framework without these enhancements: work in [34, 42] and ensuing work shows that one is forced to relax the goal merely attaining stability (see Chapter 3.4.3). Our enhanced observers and controllers allow us to stay within the cost framework.

- Quantization-based strategies (complemented by linear strategies) are shown to be approximately-optimal for the counterexample. This quantifies and proves the intuition of Witsenhausen [14] and Mitter and Sahai [18] on the goodness of quantization strategies.

- Many heuristic search-based techniques have yielded strategies that appear to be like quantization, only they have some slope in the flat parts of the quantization curve. These strategies are believed to attain the optimal cost (albeit without proof) because of the feeling of exhaustiveness in the search procedure. In Chapter 4.3.3, we show that these strategies can be arrived at using the procedure of *dirty-paper coding* in information theory. Further, at least in the limit of infinite-lengths, these strategies use the optimum required power for attaining zero distortion costs. Our results thus provide the first theoretical evidence for the believed optimality of these strategies.

- Nonlinear strategies can in general be extremely complicated functions. Prior to our work, there was no guarantee that a class of good nonlinear strategies for the counterexample would have any nice structure. The surprising simplicity of quantization-based strategies, or even dirty-paper coding based strategies, suggests that good strategies for decentralized control problems may not look extremely complicated. This is further substantiated by similarly simple structures of approximately-optimal solutions to a few other problems of implicit communication that we discuss below.

Our structured approach to understanding the counterexample extends to other problems in decentralized control as well. In Chapter 5.4, we address a (finite-horizon) dynamic version of the counterexample where the two controllers repeatedly take control actions on the system. In Chapter 5, we choose three other problems of decentralized control. Each

of these problems brings out phenomena of importance in decentralized control that the counterexample itself does not. These examples can also be thought of as the first few building blocks for a theory of implicit communication.

1.5.3 A problem of implicit and explicit channels

Is communicating implicitly at all useful when controllers are connected using external channels? We saw earlier that if the external channels are assumed to be perfect and instantaneous, then the system is effectively centralized and certainty-equivalence theory is applicable in many cases. But not only are the real-world channels imperfect, even for single controller (and hence seemingly centralized) systems, certainty-equivalence may not be applicable! Which single-controller systems are these? To understand this, let us consider the case of a single memoryless controller. Because the controller is memoryless, the situation is equivalent to one where the controller is replaced by its perfect copy at the next time-step. Coming back to non-memoryless controllers, realistically, any controller has only finite memory, and so it can be thought of as a decentralized system with *rate-limited channels* connecting it to its future self[20]! Is there any advantage, then, for the controller to communicate implicitly to its future self? In general, if a decentralized system has imperfect external channels connecting the controllers, is implicit communication between agents still useful? What strategies are good for these problems?

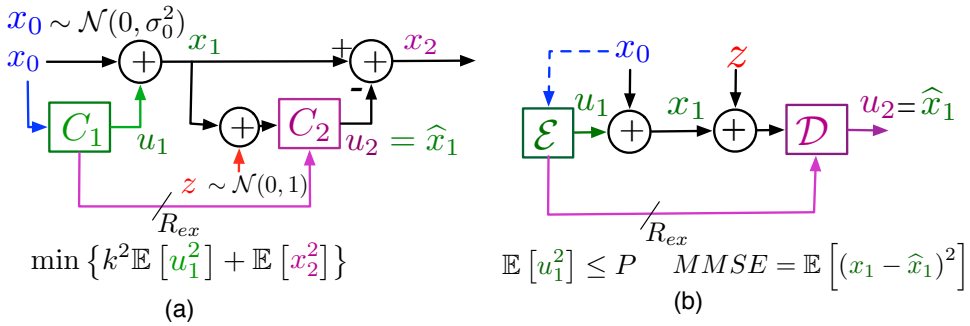


Figure 1.7: A problem of implicit and explicit channels. An external channel connects the two controllers. Should a linear scheme be used on the external channel? The answer is no: a linear scheme is good at communicating the most significant bits of the state. But these bits are already known at the decoder through the implicit channel. We propose a binning-based strategy that transmits finer information on the external channel. This strategy attains within a constant factor of the optimal cost.

To investigate these questions we construct the following toy problem: we consider an extension of Witsenhausen's counterexample where a finite capacity external channel connects

[20]The same happens in movie 'Memento' [44] where the protagonist, suffering from short-term memory loss, uses notes and tattoos to communicate with his future self.

the two controllers (see Fig. 1.7). What strategies would the theory of certainty-equivalence suggest? These strategies turn out to be those of inaction: the first controller does not use any control input on the external channel or the implicit channel. A more interesting strategy based on the certainty-equivalence philosophy is where the first controller communicates the state as well as possible on the external channel, and uses a linear strategy on the implicit channel.

From an implicit communication perspective, certainty-equivalence-inspired strategies lose performance because of redundancy: the implicit and explicit channel are essentially being used to send the same information. In Chapter 5.2, we use a deterministic abstraction of the problem to guide the strategy design for the LQG problem. In our strategy, the information of the state is split "orthogonally" on the two channels: the implicit channel is relied upon to communicate coarse information about the state, and finer information is communicated over the external channel. These strategies outperform certainty-equivalence-inspired strategies by a factor that can diverge to infinity. A proof of the asymptotic-approximately-optimality of these strategies is also provided.

If humans are thought of as finite-memory agents, then our results suggest how the visual system might organize a stimulus into object and its details or parts. Gestalt psychology [45] proposes that the visual stimulus conveys the object or the scene as a whole. The details or parts are determined by the intrinsic nature of the whole, and not the other way around. In our interpretation, the whole is coarse information about the object, *i.e.* the most significant bits of the system state. There is no need to remember these bits because they are available in the environment, and hence can be observed. The more refined information — the details in the stimulus — is what we remember (and hence store in the external channel of memory). This refined information can only be understood once the stimulus presents the coarse information (analogous to the "orthogonality" of information on the implicit and explicit channels in our strategy).

1.5.4 A problem exhibiting the triple nature of control laws

Varaiya calls the possibility of a single control action having three roles to play — control, improving the estimability of the state, and signaling — as the 'triple aspect' of control laws[21], or 'triple control' in decentralized control systems [47]. This triple aspect does not show up in Witsenhausen's counterexample: the first controller wants to communicate the state itself to the second controller. For the counterexample, therefore, the goals of improving state estimability and signaling collapse into one.

We need a toy problem where the three roles are not aligned. What will force the controllers to signal to other controllers beyond merely improving state estimability? We are

[21] In adaptive control, control actions have a fourth role to play — that of enabling the learning of system parameters [46]. This was explored first by Feldbaum in a series of papers starting with [46]. Similar to issues arise there: certainty-equivalence-based strategies are also suboptimal for problems where control actions have to learn as well as control [46].

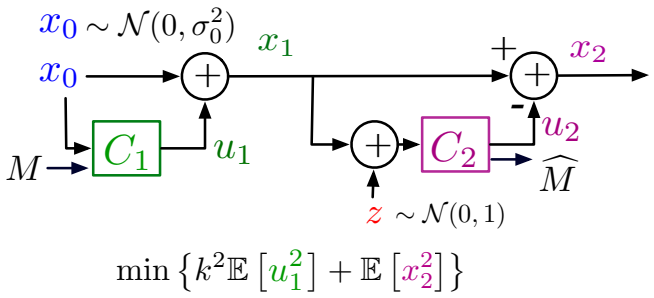


Figure 1.8: A problem that brings out the triple role of control actions in decentralized control. The control actions are used to reduce the immediate control costs, communicate a message, and improve state estimability at the second controller.

looking for a situation where the controller *embeds* information into the state for other other controllers to observe. Can one controller have information that the other needs? This can happen if the latter controller does not observe a part of the state which the former does.

Based on this observation, in Chapter 5.3 we formulate a new toy problem (shown in Fig. 1.8) by extending Witsenhausen's counterexample. In this problem, the initial state is denoted by the two-dimensional vector $[x_0, M]^T$. The first controller observes the state noiselessly, and the second controller only observes the state x_1 through noise. The goal is to have the second controller reconstruct x_1 and M. Clearly, not only is improving state estimability the goal, the first controller also wants to communicate the "message" M to the second controller. Again, a semi-deterministic abstraction provides guidance for obtaining approximately-optimal strategies for the problem. These approximately-optimal strategies show that there is an overhead cost associated with signaling beyond the cost required for mere state-estimability, thereby demonstrating that the goals of signaling and improving state-estimability do not collapse into one for this problem.

1.5.5 An economics-inspired problem of rational inattention

In any social system, economic modeling often assumes that the participating agents are maximizing either individual or joint utility. It is commonly observed (for instance, in prisoner's dilemma [48]) that the game-theoretic conclusions are not followed by players in practice [49]. In the last two decades, a number of formulations have shown that at least in part, this might be a consequence of *bounded rationality* of the participating agents. There is no unanimity on what model of bounded rationality suits all problems. For instance, for a game of repeated Prisoner's dilemma, Papadimitriou and Yannakakis [48] model the participating agents as finite-state automata. They show that the Nash equilibrium is for the prisoners to use a tit-for-tat strategy (which entails returning favors as well), rather than relentlessly (and unrealistically) back-stab each other. Models of noisy observation have also

been considered (e.g. [50]).

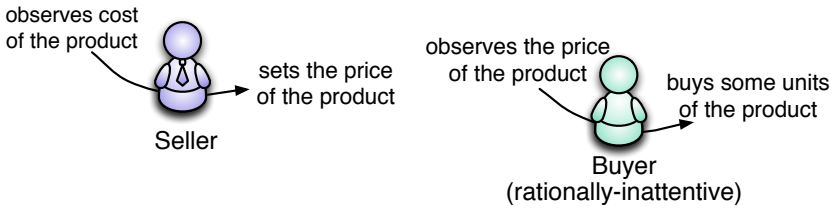


Figure 1.9: A toy model for a tiny market: a single seller sets the price that a rationally-inattentive consumer observes. The consumer selects the number of units to buy based on this observation.

Sims observed that sometimes it is not the noise in our observations, but it is the amount of attention we *choose* to allocate: if we choose to, we can focus obsessively on a particular problem and optimize it, but we will likely not pay attention on the others. To model this element of choice in how we allocate our attention, Sims [51] proposes what he calls the "rational-inattention" model. The rationally-inattentive agents can have arbitrary functions to map their observations to their decisions (control inputs), except for an information-processing constraint. In the hope of analytical tractability (inspired from the success of information theory), Sims assumes a mutual information between the observation and the control input is bounded by a constant I.

These ideas are closely connected to implicit communication. For instance, the concept of 'price signaling,' *i.e.* using price of an item to signal an aspect (e.g. quality) of the product to the consumer is quite useful in explaining pricing strategies [52]. What would be good pricing strategies to signal to a rationally-inattentive consumer?

Unfortunately, it turns out that these models are hard to analyze analytically. Even a simple two agent problem of seller and consumer (see Fig. 1.9), where each agent operates just once — the seller fixes the price, and the consumer buys some units of the object — is hard. What are the "partial observations" of the consumer? In the rational-inattention model, the observations are *a priori* "noisy," but under a mutual-information constraint, they can be chosen by of the second agent.

Computer calculations of Matejka [53, 54] provide evidence that for a toy model of a seller and a rationally-inattentive consumer (and more generally, in problems of tracking with rationally inattentive agents), the numerically-optimal pricing strategies are discrete. This discreteness is consistent with what we observe in practice[22], and is reasoned in [53, 54] as follows. Because the consumer has only a limited attention to allocate to observing the prices, the seller makes it easy for her by making the prices discrete. A discretization of prices

[22]The prices are often pegged at round figures, e.g. 10.00, or for psychological reasons at figures such as 9.99. This is the tacit code that the sellers and consumers understand, but is harder to capture mathematically. The more general phenomena of discretization can, however, be captured.

makes it easy for the consumer to decide quickly on the price-changes, thereby stimulating her to consume more.

This interpretation is very similar to our "source-simplification" interpretation of the counterexample (Chapter 1.4.2): in both cases, good solutions to a continuous state-space problem are discrete. In Chapter 5.5, we show that this is not a mere coincidence. We consider a quadratic version of the Matejka's problem of tracking using rationally-inattentive agents, and show that quantization-based strategies (complemented by linear strategies) are approximately optimal for this problem as well.

Publications in which some of this work has appeared

Some results in this dissertation have appeared in various journals and conferences, and a few others were developed in the course of finalizing this dissertation. The following articles helped develop the perspective and the results in Chapter 4: in [55,56], we proposed the vector version of Witsenhausen's counterexample and provided approximately optimal solutions to the asymptotically infinite length problem. An improved bound on the infinite-length problem appeared in [57] which characterizes the optimal power for zero $MMSE$ for the asymptotic problem. In [58, 59], we provided approximately optimal solutions to finite-length Witsenhausen counterexample, including the scalar version of the problem. The perspective that we adopt in this dissertation evolved over time. An early perspective appeared in [60].

The extensions of the counterexample (some of which appear in Chapter 5) have appeared in the following papers. In [61] and [62], we obtained approximate-optimality results for an extension of the counterexample with costs on all states and inputs, and noise in all state evolutions, inputs and observations, respectively. In [63], we show that the proofs simplify considerably for a version of the counterexample where the noise is non-Gaussian and bounded, and even considered an adversarial robust-control formulation. Using a semi-deterministic abstraction of an extension of the problem, we obtain asymptotically-approximately-optimal strategies for an extension of the problem with an external channel in [64]. The other extensions in this dissertation have not yet appeared in print.

Chapter 2

Why communicate implicitly: Actions can speak more clearly than words

Actions can speak: they can be used to communicate implicitly (see Chapter 1). But when should we use actions to speak? Clearly, when attaching external channels that connect various agents is infeasible (e.g. some economic systems and human interactions), actions are the only possible way to speak.

But is it still useful to communicate using actions when one *can* communicate using words, *i.e.* an external channel can be attached? This chapter investigates this question using simple toy models. Our main conclusion is that even when an external channel can be attached, communicating implicitly can significantly outperform explicit communication. While this does not conclusively imply that implicit communication will be useful in practice, it identifies the nature of problems where there might be a substantial reason to explore it as an alternative to explicit communication.

2.1 A toy problem for comparing implicit and explicit communication

How can we compare implicit and explicit communication? We need a problem where the designer can use any of these two options. Let us construct a simple setting: a two controller system where two controllers want to operate sequentially in order to force a state to be small. The first controller observes the state perfectly but has limited power, so it wants to communicate the state to the second controller. To emphasize the communication aspect of the system, the input of the second controller is assumed to be free and is allowed to be unboundedly large. Thus the second controller only needs to have a good state estimate in order to force the final state to be close to zero.

We impose quadratic costs on the input of the first controller, and quadratic costs on the state after the action of the second controller. A weighted sum of these costs yields

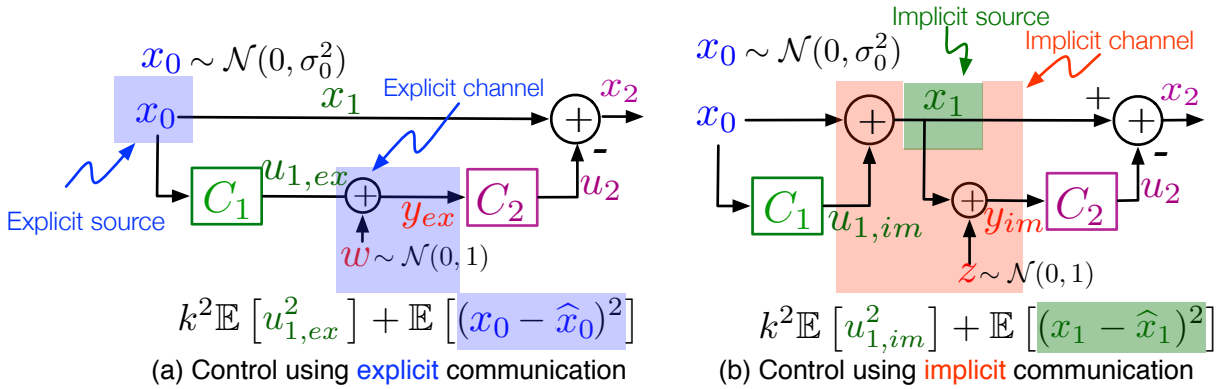


Figure 2.1: (a) A problem of "words" where the message as well as the communication channel are explicit. (b) A problem of "actions" where the message as well as the channel are implicit. The cost function for both of these problems is a weighted sum of power and $MMSE$ costs, $k^2P + MMSE$, where power P is the power of the (channel or control) input. Fig. 2.2 shows that actions, used wisely, can "beat" words handsomely.

a total average cost of $k^2P + \mathbb{E}[x_2^2]$, where P is the power of the input ($u_{1,ex}$ or $u_{1,im}$, depending on whether the communication is explicit or implicit) and $\mathbb{E}[x_2^2]$ is the mean-squared reconstruction error in estimating x_1.

If the designer chooses explicit communication, then the resulting block-diagram is shown in Fig. 2.1(a). On the other hand, a choice of implicit communication yields the block-diagram in Fig. 2.1(b). Which option — (a) or (b) — should the designer choose?

Naturally, the two options have different architectural costs[1]. For simplicity, we only compare their running costs.

Alternative architectures are also possible. Also, it is also possible (and indeed likely) that the weight on the input cost and the "bandwidth" (*i.e.* the number of control inputs for each observation) can be different for the particular implicit and explicit communication setups. These differences are important, but we will see in the next section that these two architectures capture the essence of the difference between implicit and explicit communication

2.1.1 Costs using explicit communication: an optimal strategy

The explicit communication option (Fig. 2.1(a)) is a problem of communicating a Gaussian "source" across a Gaussian channel. There is exactly one source symbol, and exactly

[1]The explicit-communication option requires the controllers to be equipped with an external link connecting the first controller to the second[2]. The second controller does not observe the state directly, and only estimates the state from the channel output. In the implicit-communication option, the first controller is equipped with an actuator as well, using which it can change the system state. The second controller has a sensor to sense the state. But no external link connects the two controllers.

one channel use (in information-theoretic lexicon, the source and channel are "bandwidth-matched"). The optimal solution for this problem was first found by Goblick [65], and is well known to be linear [65, 66], *i.e.* $u_1 = \alpha x_0$ for $\alpha = \frac{\sqrt{P}}{\sigma_0}$.

The resulting $MMSE$ is $\frac{\sigma_0^2}{P+1}$. The total cost can be calculated easily

$$\overline{\mathcal{J}}_{x_0,comm} = k^2 P + \frac{\sigma_0^2}{P+1}, \tag{2.1}$$

2.1.2 Costs using implicit communication: a quantization-based strategy

As we shall see later, the problem resulting from adopting the implicit communication option (Fig. 2.1(b)) is the Witsenhausen counterexample. The optimal strategy for the problem is unknown. We therefore use a strategy which has been observed in the literature[3] to be reasonably good: a quantization-based strategy. In fact, this is also the strategy that the semi-deterministic model suggests, and will be proved to be approximately-optimal in Chapter 4. The strategy is described next.

The controllers agree on uniformly spaced quantization points with bin size B. The first controller uses its input to force x_0 to the quantization point nearest to x_0. The second controller now performs a maximum-likelihood estimation for x_1 based on its observation y_{im}. That is, it decodes to the quantization point closest (in Euclidean distance) to the received $y_{im} = x_1 + z$. We numerically optimize over the choice of bin-sizes to obtain the minimum total cost using quantization. The resulting cost is plotted in Fig. 2.2, which is what we discuss next.

2.2 The tipping point: when should one use actions to speak?

2.2.1 Comparison of explicit and implicit communication of 2.1.1 and 2.1.2

A comparison of costs attained attained using the optimal strategy of Chapter 2.1.1 for explicit communication and using quantization-based strategy of Chapter 2.1.2 for implicit communication is shown in Fig. 2.2. The figure shows that in *all* cases, implicit communication outperforms explicit communication. Although surprising, in part this is because the weight k^2 on the costs of the inputs for the two options is assumed to be the same. At large

[3] In [14], Witsenhausen proposed a two-point quantization strategy. Bansal and Başar optimized Witsenhausen's strategy in [23]. In [18], Mitter and Sahai used a multipoint quantization strategy which is the strategy we use here.

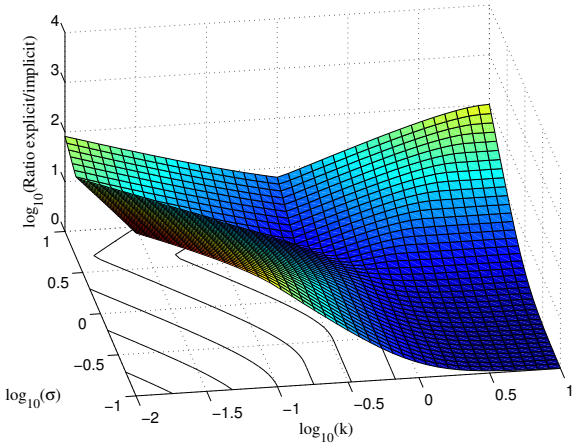


Figure 2.2: The log of ratio of costs attained for problems (a) and (b) of Fig. 2.1. The costs for problems (a) is the optimal costs. For problem (b), the costs are those attained using a quantization strategy (the optimal costs for this problem, which is equivalent to Witsenhausen's counterexample, are still unknown). The figure shows the importance of having actions speak: not only are the attained costs always better with implicit communication (the log of the ratio is always greater than 0, thus the ratio is always larger than 1), the attained costs are better by a factor that diverges to infinity in the limits $k \to 0$ and $(k, \sigma_0) \to (\infty, \infty)$.

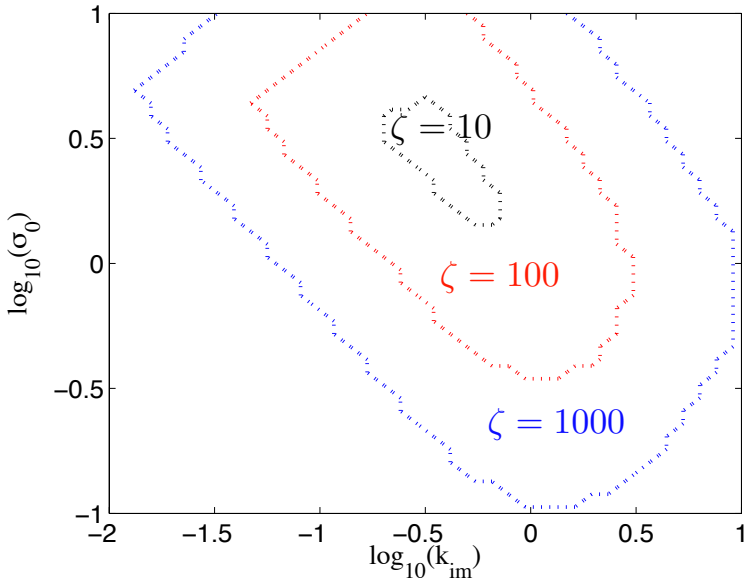


Figure 2.3: Regions in which explicit communication performs better than implicit communication for $\zeta = \frac{k_{ex}^2}{k_{im}^2} = 10, 100$ and 1000. The region is always bounded, thereby showing that implicit communication outperforms explicit communication in "most" of the parameter-space (for fixed ζ).

values of k, this happens in part because the implicit communication is helped somewhat unfairly by the implicit channel: even if the first controller uses zero input, *i.e.* $u_{1,im} \equiv 0$, the second controller can perform an estimation on the observation, and the $MMSE$ is $\frac{\sigma_0^2}{\sigma_0^2+1}$. For the explicit-communication problem, if the first controller uses $u_{1,ex} \equiv 0$, the second controller receives no information about the state, and the corresponding $MMSE$ is σ_0^2.

The next section shows that these reasons do not provide a sufficient explanation for why implicit communication can outperform explicit communication. The fact that the source is implicit is very important as well.

2.2.2 When control and communication inputs have different costs

What does the weight k^2 signify? This weight measures the relative importance of the input power and the reconstruction error. In situations where high precision in state estimation (at the second controller) is not required, the input is relatively more important and k^2 is large. On the other hand, when high precision is required, k^2 is small.

Realistically, the cost of communication input can be different from the cost of control input, so we should assign different weights to the input costs for implicit and explicit communication. How do we choose the different weights? In this section, we assume a weight of k_{ex}^2 on input power on the explicit channel, and k_{im}^2 for power on the implicit channel. In order to emphasize on the effect of relative importance of power and $MMSE$, we fix the ratio $\zeta = \frac{k_{ex}^2}{k_{im}^2}$, and let one of them vary. The resulting plots are shown in Fig. 2.3.

2.2.3 Comparisons with other architectural options

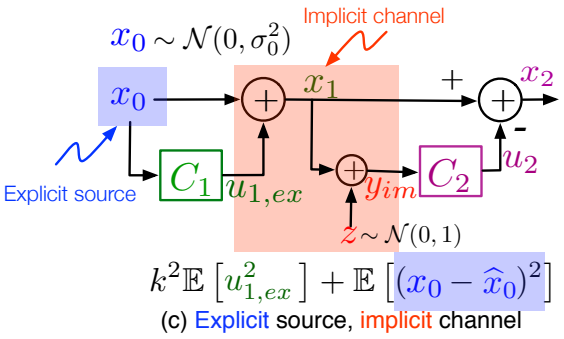


Figure 2.4: A problem with an explicit source (x_0) that needs to be communicated across the implicit channel $X_1 - Y_{im}$.

Comparing problems in Fig. 2.1, there are two major differences in the toy problems of implicit and explicit communication. The first difference is that we noted as one of implicit sources: the first controller can modify the state x_1 that is to be communicated. There

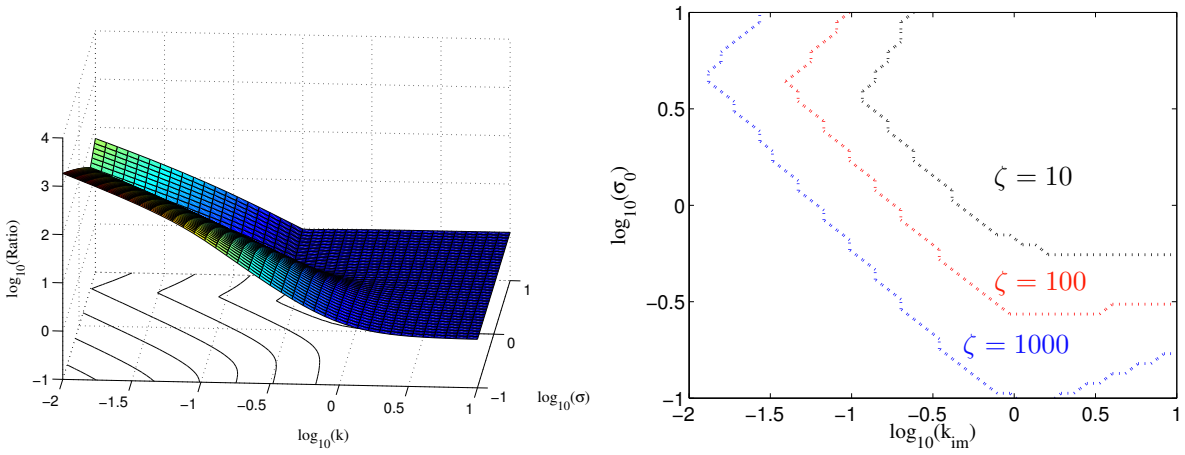


Figure 2.5: The (log of) ratio of the costs required by the problem of explicit source, implicit channel shown in Fig. 2.4 with the implicit communication problem of Fig. 2.1(b). The ratio is observed to diverge to infinity whenever $k \to 0$.

is a second, and a more subtle, difference too: the power on communication channel in implicit communication problem (Fig. 2.1(b)) can be much larger than that on the explicit communication problem. This is because the power in the initial state x_0 can be used to boost the power on the implicit channel (because the channel input is a sum of x_0 and u_1), but the power on the explicit channel is determined solely by the input u_1. Where is the advantage of implicit communication coming from?

To investigate this, we consider the problem shown in Fig. 2.4. In this problem, an explicit source x_0 is to be communicated across a channel whose power is boosted by the source itself in the same way as that in Fig 2.1(b). As a sanity check, with $u_{1,ex} \equiv 0$, the cost for this problem is the same as that for $u_{1,im} \equiv 0$ the problem in Fig. 2.1(b), because in that case, $x_1 \equiv x_0$. Fig 2.5 shows that even in comparison to this problem, the costs for the implicit communication problem (b) are better by a factor that diverges to infinity in the limit $k \to 0$.

Because the distinguishing feature between the two problems considered here is the implicit nature of the source, it has to be the case that the implicit source, and not the implicit channel, brings about the gains in implicit communication. What is so special about an implicit source?

Conclusions: what aspect of implicit communication makes it better?

From results in Chapter 2.2.1, even if explicit communication inputs cost much less than the same real-number inputs for implicit communication, the total costs using implicit commu-

nication can still be significantly smaller than that for explicit communication. But implicit communication has two aspects: an implicit source (a source that can be "simplified") and an implicit channel. The literature in 'signaling' in control emphasizes on the aspect of communicating through the plant (e.g. [4]), *i.e.* the implicit-channel aspect of implicit communication. Chapter 2.2.3 attempts to isolate the effect that makes implicit communication powerful by allowing for the same power on the explicit and implicit channels. The fact that implicit communication is still a superior strategy in some cases shows that the "source-simplification" (e.g. price-signaling in economics) *i.e.* the implicit-source aspect of implicit communication might even be more important.

So when can implicit communication be useful? It is clearly useful when it is the only alternative, *i.e.* when the engineering freedom of attaching external channels does not exist (e.g. some economic and human interactions). It also could be useful when the cost of communicating over an external channel is comparable to the control costs, and/or when extremely high precision control is required.

In this dissertation, we use this potential advantage of implicit communication in toy problems to motivate a deeper understanding of Witsenhausen's counterexample, which is the same as problem (d) in Fig. 2.1. There are likely other situations where implicit communication can be advantageous. The next chapter discusses historical reasons why understanding the counterexample and understanding implicit communication using the counterexample has been of immense interest.

Chapter 3

The historical importance of the minimalist implicit communication problem: Witsenhausen's counterexample

In Chapter 2, in a toy setting, we compared the costs of implicit and explicit communication, and noted that implicit communication can be a useful alternative in some cases. We observed that the implicit communication problem we compared with is the Witsenhausen counterexample, which is also the minimalist problem that exhibits aspects of both implicit sources and implicit channels. Studying the counterexample is therefore important in order to understand implicit communication.

Historically, the counterexample has been studied for many other (though not unrelated) intellectual reasons. This chapter discusses these reasons, and looks at how the counterexample has influenced the development of the theory of decentralized control. We also talk about the literature related to signaling in the counterexample, and observe how in the signaling context, the counterexample sits naturally within a set of related information theory problems. To set up the notation for this discussion, we begin with a formal statement of a vector version of the counterexample. Witsenhausen's original counterexample, which is scalar, is just the one-dimensional case of this problem. Why look at a vector version of the counterexample? As we will see in Chapter 4, much like in traditional information-theoretic problems, the vector version provides conceptual simplification: large vector lengths allow us to use laws of large numbers and side-step the complications associated with the geometry of finite-dimensional spaces.

3.1 Notation and a formal statement of the vector Witsenhausen counterexample

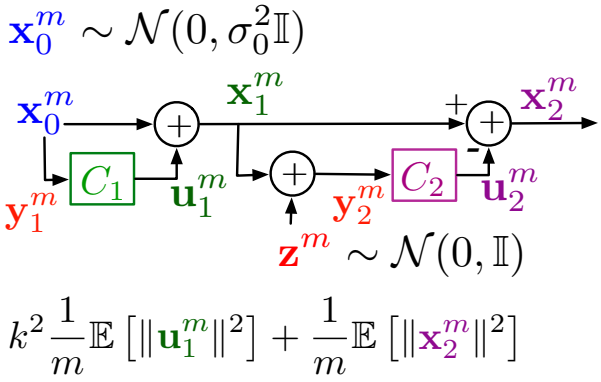


Figure 3.1: Block-diagram for Witsenhausen's counterexample of dimension m.

Vectors are denoted in bold. Upper case tends to be used for random variables, while lower case symbols represent their realizations. $W(m, k^2, \sigma_0^2)$ denotes the dimension-m vector version of Witsenhausen's problem, which is a time-horizon-2 problem of stochastic control described as follows.

- The initial state $\mathbf{X}_0^m$ is Gaussian, distributed $\mathcal{N}(0, \sigma_0^2 \mathbb{I}_m)$ (*i.e.* the elements of the state are iid), where $\mathbb{I}_m$ is the identity matrix of size $m \times m$.

- The states are denoted using vectors $\mathbf{X}_t^m$, $t = 0, 1, 2$. The first controller C_1 acts at $t = 1$ and uses control input $\mathbf{u}_1^m$. Similarly, the second controller acts at $t = 2$, and uses input $\mathbf{u}_2^m$. The state transition functions describe the state evolution with time. The state transitions are linear:

$$\begin{aligned} \mathbf{X}_1^m &= \mathbf{X}_0^m + \mathbf{U}_1^m, \quad \text{and} \\ \mathbf{X}_2^m &= \mathbf{X}_1^m - \mathbf{U}_2^m. \end{aligned}$$

- The outputs $\mathbf{Y}_t^m$ are observed by the controllers:

$$\begin{aligned} \mathbf{Y}_1^m &= \mathbf{X}_0^m, \quad \text{and} \\ \mathbf{Y}_2^m &= \mathbf{X}_1^m + \mathbf{Z}^m, \end{aligned} \tag{3.1}$$

  where $\mathbf{Z}^m \sim \mathcal{N}(0, \sigma_Z^2 \mathbb{I}_m)$ is Gaussian observation noise. Without loss of generality, we assume that $\sigma_Z^2 = 1$.

- The objective is to choose a control strategy that minimizes the expected cost, averaged over the random realizations of $\mathbf{X}_0^m$ and $\mathbf{Z}^m$. The total cost is a quadratic function of the states and the inputs given by the sum of two terms:

$$\begin{aligned} J_1(\mathbf{x}_1^m, \mathbf{u}_1^m) &= \frac{1}{m} k^2 \|\mathbf{u}_1^m\|^2, \text{ and} \\ J_2(\mathbf{x}_2^m, \mathbf{u}_2^m) &= \frac{1}{m} \|\mathbf{x}_2^m\|^2 \end{aligned}$$

  where $\|\cdot\|$ denotes the usual Euclidean 2-norm. The cost expressions are normalized by the dimension m to allow natural comparisons between different dimensions. A control strategy is denoted by $\gamma = (\gamma_1, \gamma_2)$, where γ_i is the function that maps the observation $\mathbf{y}_i^m$ at C_i to the control input $\mathbf{u}_i^m$. For a fixed γ, the time-1 state $\mathbf{x}_1^m = \mathbf{x}_0^m + \gamma_1(\mathbf{x}_0^m)$ is a function of $\mathbf{x}_0^m$. Thus the first stage cost can instead be written as a function $J_1^{(\gamma)}(\mathbf{x}_0^m) = J_1(\mathbf{x}_0^m + \gamma_1(\mathbf{x}_0^m), \gamma_1(\mathbf{x}_0^m))$ and the second stage cost can be written as $J_2^{(\gamma)}(\mathbf{x}_0^m, \mathbf{z}^m) = J_2(\mathbf{x}_0^m + \gamma_1(\mathbf{x}_0^m) - \gamma_2(\mathbf{x}_0^m + \gamma_1(\mathbf{x}_0^m) + \mathbf{z}^m), \gamma_2(\mathbf{x}_0^m + \gamma_1(\mathbf{x}_0^m) + \mathbf{z}^m))$.

  For given γ, the expected costs (averaged over $\mathbf{x}_0^m$ and $\mathbf{z}^m$) are denoted by $\overline{\mathcal{J}}^{(\gamma)}(m, k^2, \sigma_0^2)$ and $\overline{\mathcal{J}}_i^{(\gamma)}(m, k^2, \sigma_0^2)$ for $i = 1, 2$. We define $\overline{\mathcal{J}}_{\min}^{(\gamma)}(m, k^2, \sigma_0^2)$ as follows

$$\overline{\mathcal{J}}_{\min}(m, k^2, \sigma_0^2) := \inf_{\gamma} \overline{\mathcal{J}}^{(\gamma)}(m, k^2, \sigma_0^2). \tag{3.2}$$

  Because of the diagonal dynamics and diagonal covariance matrices, the optimal linear strategies act on a component-by-component basis. Therefore, even if $m > 1$, the relevant linear strategies are still essentially scalar in nature.

  What happens if the initial state is not distributed iid across its elements? The problem does not change much because the correlation across various elements can be zeroed out by simply rotating the axes. This rotation of axes does not affect the noise because it is white. What if the noise is also not white? This case, though potentially interesting, is not discussed in this dissertation.

3.2 What conjecture is refuted by the counterexample?

We saw in Chapter 1.5.1 that strategies based on the theory of certainty-equivalence are optimal for linear-quadratic (LQ) problems with classical information patterns [4, Assertion 7]. The optimal solution can be obtained by splitting each agent into an estimator followed by a controller. What happens when the system is not just LQ, but Linear-Quadratic-Gaussian (LQG)? Using dynamic-programming, one can show that at each step of "backtracking" in the program, the optimization problem is convex. The resulting optimal strategy is that predicted by certainty-equivalence, and turns out to be linear! Further, it can be found

recursively using Riccatti equations [67]. Because of this simplicity and the generality of the formulation, LQG control strategies have found applicability in diverse practical problems (many such examples are talked about in [68]).

A natural conjecture is that the optimality of certainty-equivalence strategies extends to *decentralized* LQ problems. In particular, for decentralized LQG systems, this would imply that linear strategies are optimal. Indeed, this belief was widespread at the time Witsenhausen came up with his counterexample (e.g. see abstract of [14]). Witsenhausen's counterexample explicitly demonstrated that linear laws can indeed be suboptimal for decentralized LQG problems. To show this, Witsenhausen constructed a two-point quantization strategy which can outperform all linear strategies for the counterexample. Even today, many formulations in decentralized LQG control restrict their attention to linear strategies (e.g. [69]) without a proof of their optimality in the larger set of all possible strategies. How much can this approach hurt? The answer, surprisingly, is that it can hurt a lot: by constructing multi-point quantization strategies, and by choosing an appropriate sequence of problem parameters in Witsenhausen's formulation, Mitter and Sahai [18] showed that nonlinear strategies can outperform linear strategies by a factor that diverges to infinity. A designer of decentralized control systems ignores nonlinear strategies at her own peril.

What makes linear strategies suboptimal for the counterexample? We saw in Chapter 1.5.1 that certainty-equivalence based strategies are suboptimal for centralized systems when control actions can perform a dual role[1]: that of control and signaling (explored more deeply in Chapter 3.4.2).

3.3 The counterexample as an optimization problem

3.3.1 Nonconvexity of the counterexample

In the last section, we saw that the theory of certainty-equivalence yields strategies that are optimal for centralized LQG problems, and these strategies can be found efficiently. We also saw that the Witsenhausen counterexample demonstrates that the theory does not extend to decentralized control. Staying within the framework of solving control problems by minimizing costs, can numerical optimization techniques help solve the counterexample? If so, there is some hope that good strategies for larger problems can be obtained through optimization as well.

Convex optimization is a commonly-used framework that provides efficient algorithms for finding numerical solutions to many optimization problems. Finding whether a problem is convex (and can therefore be solved using convex optimization) is therefore the often first approach to take when addressing an optimization problem. When is an optimization

[1] Close parallels exist in the literature of adaptive control. There, the control actions play a dual role as well, that of control and learning the system. It turns out that certainty-equivalence-based strategies are suboptimal there as well.

problem said to be convex? It is convex when it can be cast into the following form:

$$\begin{aligned} \min \quad & f_0(\gamma) \\ \text{subject to} \quad & f_i(\gamma) \leq b_i, \quad i = 1, \ldots, n, \end{aligned}$$

where the functions $f_0, f_1, \ldots, f_n$ are convex-$\cup$. Equivalently, if the objective function being minimized is convex-$\cup$ and the set of feasible solutions is also convex[2], the problem is said to be convex [70]. Convex problems are considered easy because algorithms such as the interior-point method, gradient-descent, etc. efficiently solve these optimization problems [70].

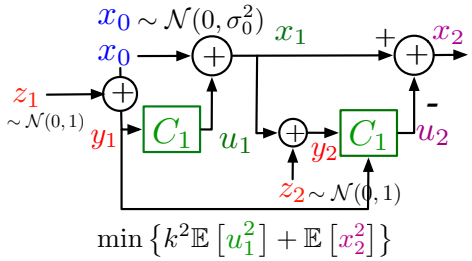


Figure 3.2: A centralized LQG problem. If the controller is memoryless, and noise z_1 is removed, the problem becomes the Witsenhausen counterexample.

As an example, let us consider a centralized problem whose illustration (Fig 3.2) resembles that of Witsenhausen's counterexample. Let us verify if this problem is convex. Given γ_1, the choice of γ_2 is clear: the MMSE estimate of X_1. Is the problem convex in γ_1? It is well known that for any centralized LQG problem, calculation of cost-to-go at any step in dynamic programming is a convex optimization problem (of minimizing a quadratic function). In order to later show the nonconvexity of Witsenhausen's counterexample, let us design a test using the centralized problem. We choose two strategies $\gamma_1^{(a)} = 0$ (the zero-input strategy), and $\gamma_1^{(b)} = -x_1$ (the zero-forcing strategy). The MMSE estimation of X_1 based on observations[3] Y_1 and Y_2 yields a cost of $\overline{\mathcal{J}}(a) = \frac{\sigma_0^2}{2\sigma_0^2+1}$ for $\gamma_1^{(a)}$ and $\overline{\mathcal{J}}(b) = k^2\sigma_0^2$ for $\gamma_2^{(b)}$.

Now let us consider a strategy $\gamma_1^{(c)} = 0.5\gamma_1^{(a)} + 0.5\gamma_1^{(b)} = -\frac{x_1}{2}$. The cost for this strategy (with MMSE estimation at time 2) turns out to be $\overline{\mathcal{J}}(c) = \frac{k^2\sigma_0^2}{4} + \frac{5\sigma_0^2}{5\sigma_0^2+4}$. This is also the cost-to-go at time 1 because we have chosen the optimal γ_2 based on our choice of γ_1. If the problem is convex, the cost $\overline{\mathcal{J}}(c)$ should be smaller than $0.5\overline{\mathcal{J}}(a) + 0.5\overline{\mathcal{J}}(b)$, where $\overline{\mathcal{J}}(a)$ and $\overline{\mathcal{J}}(b)$ are the costs attained using strategies a and b respectively. What is this sum? $0.5\overline{\mathcal{J}}(a) + 0.5\overline{\mathcal{J}}(b) = \frac{k^2\sigma_0^2}{2} + \frac{\sigma_0^2}{4\sigma_0^2+2}$. Each of the two terms is larger than the corresponding

[2]A set $A \subset \mathbb{R}^m$ is said to be convex if for any $\gamma^{(a)}, \gamma^{(b)} \in A$, $\alpha\gamma^{(a)} + (1-\alpha)\gamma^{(b)} \in A$ for any $\alpha \in [0,1]$.

[3]This MMSE estimation depends on the strategy. For $\gamma_1^{(a)}$, the MMSE choice is $\gamma_2^{(a)} = \frac{\sigma_0^2}{2\sigma_0^2+1}(Y_1 + Y_2)$, and for $\gamma_1^{(b)}$, the choice is $\gamma_2^{(b)} = 0$.

term in $\overline{\mathcal{J}}(c)$, the cost with $\gamma_1^{(c)}$. This centralized LQG problem therefore passes our simple convexity test.

Now let us run the same test for the Witsenhausen counterexample. Again, we use the same two strategies, $\gamma_1^{(a)}$ and $\gamma_1^{(b)}$. The strategies at second stage are again MMSE strategies based now on observing only $Y = X_1 + Z$. What are the attained costs? With $\gamma_1^{(a)}$, the zero-input strategy, the cost is $\overline{\mathcal{J}}(a) = \frac{\sigma_0^2}{\sigma_0^2+1}$. With $\gamma_1^{(b)}$, the cost is $\overline{\mathcal{J}}(b) = k^2\sigma_0^2$. What is the cost using $\gamma_1^{(c)}$? It is $\frac{k^2\sigma_0^2}{4} + \frac{\frac{\sigma_0^2}{4}}{\frac{\sigma_0^2}{4}+1} = \frac{k^2\sigma_0^2}{4} + \frac{\sigma_0^2}{\sigma_0^2+4}$. Again, if the counterexample is convex, then $\overline{\mathcal{J}}(c)$ must be smaller than the average of $\overline{\mathcal{J}}(a)$ and $\overline{\mathcal{J}}(b)$. This average is $0.5\overline{\mathcal{J}}(a) + 0.5\overline{\mathcal{J}}(b) = \frac{k^2\sigma_0^2}{2} + \frac{\sigma_0^2}{2(\sigma_0^2+1)} \approx 0.505$ for $k^2 = 0.01$, $\sigma_0^2 = 10$. In comparison, for the same parameter choice, $\overline{\mathcal{J}}(c) \approx 0.739$, which is larger than $0.5\overline{\mathcal{J}}(a) + 0.5\overline{\mathcal{J}}(b)$! The Witsenhausen counterexample, therefore, is not convex in γ_1.

An alternative proof of nonconvexity of the problem was provided by Witsenhausen himself in [14], where he notes that the cost function can be written down as:

$$\overline{\mathcal{J}} = k^2\mathbb{E}\left[(x_0 - f(x_0))^2\right] + 1 - I(D_f), \tag{3.3}$$

where $f(x_0) = x_0 + \gamma(x_0)$, and $I(D_f)$ is the Fisher information of the observation $Y = f(X_0)$. The nonconvexity of the problem results from the negative sign in front of the term $I(D_f)$: the function $I(D_f)$ itself is a convex-$\cup$ function of f, and therefore $-I(D_f)$ is concave-$\cap$ in f.

Is it possible that the problem is convex jointly in (γ_1, γ_2)? We show in Appendix A.1 that even this convexity does not hold. As far as we are aware, this result, while simple, is not discussed in the existing literature[4].

3.3.2 Hardness of the discrete counterpart of the counterexample

A convex-optimization approach may not work, but can we just quantize the problem and hope to use some other computational approach to solve it? A discretization can be performed as follows. First discretize the Gaussian distributions of x_0 and z. Next, constrain the domains of γ_1 and γ_2 to be finite. An exhaustive search in this discrete space for optimal γ_1 and γ_2 will yield the optimal solution for this discretized problem, which will hopefully be "close" to the optimal solution of the original problem. However, notice that the total number of possible γ_1 and γ_2 mappings is exponential in the size of their domain-spaces. Nevertheless, this approach was explored by Ho and Chang [20]. They provided a discussion on why such approaches can fail. However, the discussion was unsatisfactory[5].

[4]As an interesting aside, it may seem surprising that even the *centralized* LQG problem is nonconvex in (γ_1, γ_2). The trick to show this is to choose strategies that pretend as if the problem is decentralized even when it is not.

[5]Ho and Chang attributed the failure partly to the lack of "partial-nestedness:" a concept we discuss in the next section. They further claimed that the lack of partial-nestedness means that the problem cannot be

Papdimitriou and Tsitsiklis [19] provide a concrete reason for this failure. Consider the algorithm that takes the discretized distributions of x_0 and z as inputs. The sizes of the supports of these discretized distributions are the sizes of the inputs to this algorithm. The algorithm is meant to check if there exists a strategy γ for which the cost $\overline{\mathcal{J}}(\gamma) < \beta$ for a given β. If this problem is solved, Witsenhausen's counterexample can be solved to any approximation-fidelity using a simple binary search. However, Papadimitriou and Tsitsiklis show a reduction of a problem of three-dimensional matching[6] to this discrete version of the counterexample by moving away from the Gaussian distribution, and allowing for arbitrary discrete distributions (that obey the cardinality constraints). Since their problem of three-dimensional matching is NP-complete [19], so also is this discrete version of the counterexample. The implicit philosophical argument is that since there is *per-se* no special structure of the discretized Gaussian distributions (as compared to any other distribution), the Gaussian problem is likely as hard as any other. Therefore, any algorithmic approach for finding the optimal cost of this discrete version of the counterexample will be computationally complex.

A version of three-dimensional matching was also used in information theory to show that the problem of decoding general linear codes is NP-complete (Berlekamp, McEliece and Tilborg [72]). Nevertheless, we know of suboptimal solutions for appropriately chosen structured codes that are "good enough" [73]. Even in the theoretical computer science literature, approximation algorithms are known for NP-complete problems such as the traveling salesman problem [40], and even the problem of 3-D mapping [74] (both to within a factor of 1.5). Even though the problem of 3-D mapping reduces to the Witsenhausen counterexample, approximate solutions to one may not yield approximate solution to another[7]. Finding approximation algorithms for the counterexample and other intractable problems in control [19] is therefore an open problem, and as we will see in Chapter 4, one that has had a philosophical influence on the results in this dissertation.

Two reasonable approaches — convex programming and discretization — do not work for the counterexample. What makes the counterexample so hard? We will see in the next section that there is a possibility of signaling in the counterexample which makes it hard.

---

reduced to a static one (where all the controllers act simultaneously and just once), which makes the problem hard. This claim was later proved to be wrong by Witsenhausen [15]: in a surprising result, Witsenhausen showed that many dynamic problems (including the counterexample and the problem of communicating a Gaussian source across a Gaussian average-power-constrained channel) can be reduced to static problems through a coupling introduced in the cost function.

[6]The 3-D matching problem of Papadimitriou and Tsitsiklis [19] is slightly different from the conventional one. The problem can be described as follows: Given a set S and a family F of subsets of S that have cardinality 3, can we subdivide F into three sub-families C_0, C_1 and C_2 such that subsets in each of the C_i are disjoint, and the union of subsets of C_0 equals S. The constraint of disjointness of C_i is not present in the usual problem of 3-D matching [71]. Nevertheless, this variation on 3-D matching is still NP-complete.

[7]A stronger notion of reduction, called L-reduction, introduced by Papadimitriou and Yannakakis [75], is required to preserve approximability with constants.

Will an optimization solution be sufficient?

But will a designer be satisfied with merely an algorithmic solution, approximate or not? To answer this question, we bring our attention back to the problem of system design. A purely algorithmic solution may not reveal the connection between good strategies and the structure of the problem — the solution may appear to be magical, rather than intuitive. For instance, the optimization solutions of Baglietto, Parisini and Zoppoli [25], Lee, Lau and Ho [26], Li, Marden and Shamma [27], Karlsson *et al.* [76] etc. offer little justification as to *why* the solutions suggested are optimal (we shall see in Chapter 4.3.3 that a justification can be arrived at using information-theoretic arguments[8]). The only reason that they are thought to be optimal is because different heuristic approaches all arrive at strategies that are very similar to each other. But how do we know that the goodness of the strategies provided by these approaches extends to other problems? Without guarantees on the gap from optimality, an algorithmic/optimization solution is insufficient.

3.4 How the difficulty of the counterexample shaped the understanding of decentralized control

Observing the difficulties in designing signaling strategies for Witsenhausen's counterexample (and larger problems of decentralized control), a survey paper in 1978 of Sandell *et al* [78] argues for a problem reformulation:

> *"Determination of these signaling strategies has been shown to be equivalent to an infinite-dimensional, nonconvex optimal control problem with neither analytical nor computational solution likely to be forthcoming in the foreseeable future. This fact of life forces one to re-evaluate the problem formulation."*

That is, at least as early as 1978, the community had started taking steps towards reformulation of problems in order to avoid the difficulties brought to their attention by Witsenhausen's counterexample. How would the development of the field of decentralized control have been different if the counterexample were understood much earlier? This question is too hard and open-ended to even speculate on. Instead, we take a path down the history of decentralized control examining how the counterexample influenced problem formulation and solution approaches.

[8] More precisely, we will see in Chapter 4 that the strategies proposed in [25–27] graphically resemble a strategy based on dirty-paper coding [77] in information theory. Further, the dirty-paper coding strategy is shown to be optimal at least in the limit of zero second stage cost for the asymptotic vector Witsenhausen counterexample.

3.4.1 Classifying problems as tractable and intractable

What aspect of the counterexample makes it intractable? The question has been of active interest for the last 40 years, and has motivated a sequence of problem formulations that can be identified as tractable. An early work of Ho and Chu [79] (1972) proposes the following sufficient condition for the problem to be easy: "if a decision-maker's action affects our information, then knowing what he knows [when he took the decision] will yield linear optimal solutions." [79, Pg. 21]. If this condition is satisfied, the problem is said to have a "partially-nested" information structure.[9] If the information-structure of the problem lacks partial-nestedness, the problem is said to have a "nonclassical" information structure. In Witsenhausen's counterexample, C_2 does not know x_0, which C_1 knows. Yet, the actions of C_1 affect x_1, which is observed noisily by C_2. The information structure of the problem is therefore nonclassical. From the perspective of Bar-Shalom and Tse [36] (1974), when the information structure is nonclassical, there can be an incentive to 'signal' to other controllers using the plant itself. This is precisely the dual role (as discussed in Chapter 1.5.1) of control actions — signaling and control — that blocks certainty-equivalence-theory and makes the counterexample hard.

Since it is this possibility of signaling that seems to be making problems hard, can we remove the incentive to signal? For instance, if the controllers can communicate perfectly and fast enough, they can send what they know to the subsequent controllers, resulting in a partially-nested information structure. In particular, an architectural change — that of connecting the controllers with an external channel — could possibly be used for this communication. Using this understanding, Rotkowitz and Lall arrive at an alternative characterization of the partial-nestedness condition [21]. They show that when *propagation delays* in system dynamics are slower than *transmission delays* on an external channel, there is no incentive to signal through the plant. Further, in such cases, the resulting problem can be formulated as a convex optimization problem. Their result is a special case of their own general criterion of quadratic-invariance: a condition which, if satisfied, ensures that the problem can be solved using convex optimization.

In the absence of a theory to complement certainty-equivalence for decentralized problems, a natural approach is to artificially restrict the search for the optimal strategy to the set of linear strategies. Although computational difficulties can exist even with this simplification [78], in some cases [69, 80, 81], the best linear strategy is efficient to compute. But is sticking with linear strategies reasonable? Again, the results of Mitter and Sahai [18] show that the loss associated with restricting attention to linear strategies can be arbitrarily large[10]. It is therefore imperative to study the use of nonlinear strategies for signaling.

[9] In some cases, even if this condition is satisfied only probabilistically, certainty-equivalence strategies are still sufficient (a condition known as "stochastic nestedness" discovered by Yüksel [24]).

[10] It is sometimes suggested that restricting to linear strategies is justified because they can be easy to implement. From a system perspective, the implementation complexity also results in some extra costs at installation and at run-time. A fair comparison would be to understand the costs associated with this

3.4.2 'Signaling' and the counterexample

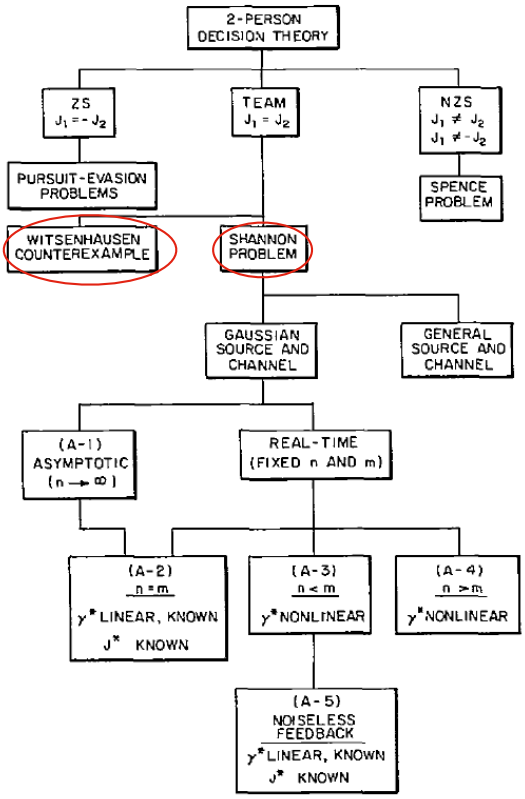


Fig. 1. Teams, signaling, and information theory.

Figure 3.3: Fig. 1 from the work of Ho, Kastner, and Wong [16]. They seem to be the first to identify the correspondence between Witsenhausen's counterexample and Shannon's point-to-point explicit communication problem.

The field of information theory is dedicated to understanding communication, which is a pure form of signaling. So why is the signaling in the counterexample so hard to understand? To explore this question, Ho, Kastner and Wong [16] view the Witsenhausen problem in relation to two other signaling problems: Shannon's problem of explicit communication [1] and Spence's problem of job-market signaling [12, 13] (see Fig. 3.3). They observe that the Gaussian version of Shannon's problem is one where the initial state is to be communicated. The problem turns out to be easy: Goblick [65] showed that linear strategies are optimal. They also formulate game-theoretic problems based on Spence's signaling problem (these

complexity (using models such as those in [82]) and then making a judicious decision on what strategy to use.

problems are explored in greater detail in [83]), and observe that when the goal of the first agent is to signal the initial state, the Nash equilibrium can be provided explicitly.

Bansal and Başar perform another exploration of the problem space that is complementary to that by Ho, Kastner and Wong [16]. They consider modifications of Witsenhausen's counterexample with parameterized cost functions that contain Witsenhausen's counterexample as a special case. Their main observation is that whenever the cost function does not contain a product of two decision variables, the problem can essentially be reduced to a variant of the problem of communicating a Gaussian source across a Gaussian channel, for which affine laws are optimal [65, 66].

These results showed that the signaling in the counterexample is somehow different from that in Shannon's problem, and therefore unaddressed and potentially hard (the connection with information theory problems is brought out explicitly in Chapter 3.5). The results cited in the previous section suggest that addressing the hardness introduced by signaling is easy in an engineering sense: just connect all the controllers using perfect channels. But realistic channels are never perfect, and even good channels can be costly (to install as well as run, *i.e.* will require high SNR). What happens when we take into account the fact that channels connecting the controllers are not perfect? Martins [22] points out that even in the presence of a non-perfect (but still pretty good) explicit communication link, nonlinear signaling strategies continue to outperform linear ones (we tackle this problem in Chapter 5.2). The incentive of signaling is present as long as the external communication links are imperfect! The impact of imperfect channels in control is interesting in greater generality, and a body of literature in the burgeoning field at the intersection of control and communication is intellectually motivated towards understanding control over imperfect channels.

3.4.3 Control under communication constraints

How can we understand control over imperfect communication channels when we cannot even understand signaling through the plant itself[11]? In order to rule out the option of signaling, many formulations (e.g. those in [34, 41, 42, 84]) do not allow the controller any direct observations of the plant. Instead, the system has an "observer" who can see the state, and has to communicate[12] the state to a "controller" who only observes the signal transmitted by the observer. Fig 3.4 takes a closer look at the observer-controller architecture. It illustrates how these formulations for control under communication constraints were inspired by the certainty-equivalence-based separation of estimation and control. The estimator (now the observer) observes the state, and communicates it (now through a noisy channel) to a controller.

[11] The historical perspective in this section is largely based on discussions with Prof. Anant Sahai as he witnessed the development of this field.

[12] As we will see in Chapter 3.5, ruling out the possibility of observer affecting the state also removes a difficulty associated with the counterexample: that of implicit sources.

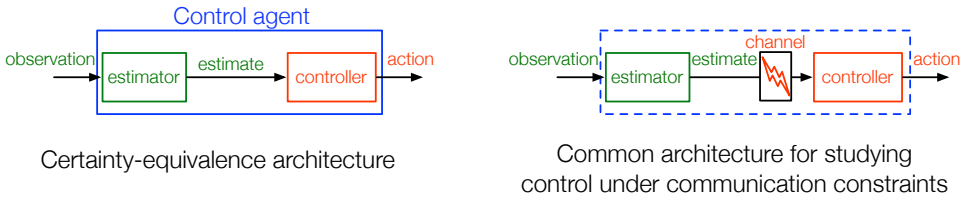


Figure 3.4: The problems of control under communication constraints have often been inspired from the certainty-equivalence architecture which may not be the optimal architecture for the problem under consideration.

But does this observer-controller architecture eliminate the possibility of signaling? It was observed by Sahai and Mitter in [85] that since the observer no longer knows the channel outputs seen by the controller, the controller can signal these outputs to the observer through the plant. The specter of signaling rose again! However, since the observer has noiseless observations of other control inputs, Sahai and Mitter notice that this signal can be embedded in the state. In order to do so, they embed the signal in the bits of the state that are just higher than those that would be affected by perturbation noise. The controller is thus forced to balance between signaling and control — the very issue that the observer-controller formulation was trying to avoid! Sahai and Mitter therefore step back from the goal of reducing costs, and instead find conditions for attaining stability, a coarser metric than minimizing costs. Their strategy of embedding information in the state is therefore reasonable: a stability formulation is not affected by gap from optimal costs.

What if a separate perfect feedback channel from controller to observer is available? Tatikonda [42] observed that the problem of minimizing costs is still hard because the causal geometry of the problem is not well understood within information theory. Thus the formulation was again forced to be driven away from an optimization perspective, and the goal was relaxed to that of merely attaining stability. For the perfect-feedback problem, Tatikonda showed that it is necessary and sufficient for the controller to track the state within a bounded-moment error in order to stabilize the system.

In [34], Sahai shows that the traditional information-theoretic formulation of capacity is not sufficient for stabilization. A new notion of *anytime capacity*, where constraints on delay emerge organically from the requirement of tracking, is needed in order to accomplish this. In order to bring out the difference between anytime capacity and Shannon capacity, he considered a version of the problem where the channel suffers from probabilistic erasures, but is otherwise noiseless and hence has infinite Shannon capacity (the "real-erasure channel"). Sahai shows that the anytime reliability of this problem is still finite.

In order to simplify the problem further and get rid of the possibility of signaling completely, Sinopoli *et al.* [86] disallow any encoding at the observer. They also require the controller to only estimate the state, stripping away any ability to modify the state by either agent. They focus on the real-erasure channel, thinking of the erasures as packet-drops on

networks intended for control. For a Gaussian version of this problem, they note that when a packet is not dropped, the optimal estimation strategy is the usual Kalman filtering. The problem is therefore called "intermittent Kalman-filtering[13]." The focus of [86] and the ensuing work is to quantify the erasure probabilities that allow the system to be stabilized in this setup.

New problem formulations

We saw that one of the core difficulties in understanding control under communication constraints was the lack of understanding of the aspect of signaling. This difficulty forced simplifications of problem formulations which limit the questions they can address. For instance, a question of central importance is that of partitioning the tasks: how should we allocate effort on the part of the agents? For two agents who are attempting to control a system, how much effort should we put in locally and how much should be put in by the agent farther away? The stability formulation considered commonly in the literature only allows the "controller" to invest any effort. Further, the goal of attaining stability is a coarse metric, and even when the goal is relaxed to merely attaining stability, the results are primarily negative: Sahai [34] shows that in order to have all moments of error bounded, the channel needs to have positive zero-error capacity[14]. Even when stabilization is possible, it is necessary to ensure that the total cost is also small for results to have practical applications.

With the understanding of signaling that we develop using the Witsenhausen counterexample (in Chapter 4), we believe that some problems of effort allocation can be begun to be addressed. Moreover, we can stay in the cost framework, thus potentially obtaining more relevant insights. Let us look at the problem of filtering to speculate how this could be done. An example of a filtering problem is shown in Fig. 3.5(a). The formulation avoids the possibility of signaling by disallowing the observer to encode or influence the state. What if the observer could put in a little effort, *i.e.* had a little power to change the state? The resulting problem is shown in Fig. 3.5(a). Does the problem allow for the possibility of signaling? Indeed, the controller C_1 can modify the state in order to signal to C_2. Is Witsenhausen's counterexample needed to understand signaling in this problem? Fig. 3.5(a) shows that a single time-step version of the problem (with Gaussian observation noise) is the information-theoretic interpretation (discussed in next section) of the Witsenhausen counterexample itself! It is clear that the while the formulation of filtering successfully avoided the difficulties introduced by signaling, the modified formulation that allows for signaling cannot be addressed (at least in the optimization framework of minimizing system costs) without addressing the counterexample.

[13]Elia's term for the problem of control over real-erasure channel was "indelible control," a name suggested by John Doyle [87]. He used the real-erasure channel to understand the idea of anytime reliability from a purely control-theoretic perspective.

[14]Zero-error capacity is the maximum rate that can be achieved with exactly zero error probability [88]. Zero-error capacity of many practical channels is zero.

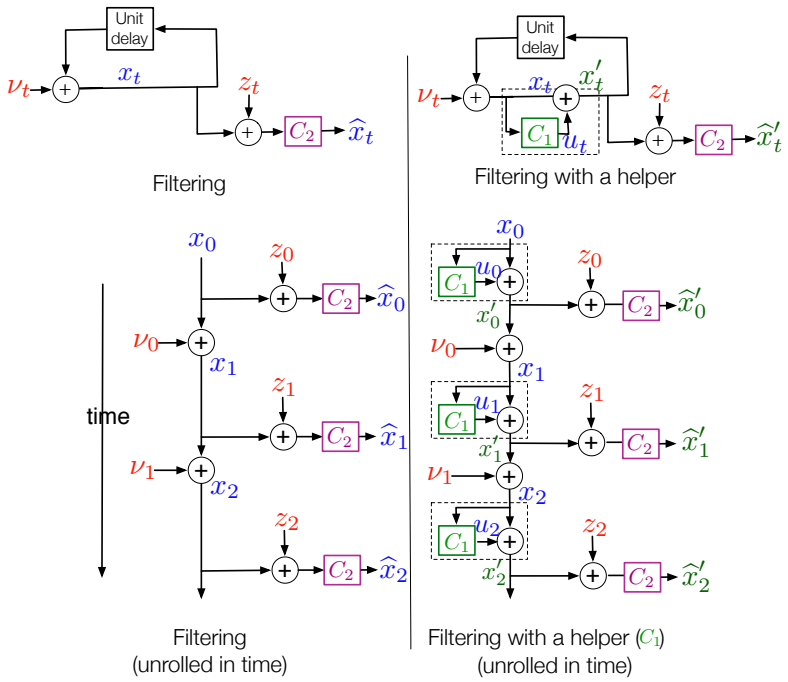


Figure 3.5: (a) The problem of filtering with a helper can be thought of as an extension of Witsenhausen's counterexample to multiple time-steps. (b) Communicating an evolving source, the problem of [34], can now be understood in an "active" context where a helper now assists the encoder by modifying the source. This can be thought of as an extension of the problem with implicit source and explicit channel, considered in Chapter 5.1.

We also notice that with this modification, we observe that the system is always stabilizable: the empowered "observer" can simply force the state all the way to zero. Of course, this will incur high costs, which is why it is important to find low-cost strategies for this problem. In Chapter 5.4, using our understanding for the counterexample (in Chapter 4), we provide strategies that attain within a constant factor of the optimal cost for a version of this filtering problem.

Let us turn our attention now to the more general problem of control under communication constraints. For the formulations that we discussed in the last section, can we make similar modifications? Indeed, Fig. 3.5(b) shows the resulting problem. The possibility of observer zero-forcing the state again makes the problem stabilizable[15]. The important question is how low a cost can be attained. In Chapter 5.1, we address a single time-step version of this problem.

Understanding the counterexample therefore opens up the possibility of addressing such "active" versions of the problems of control under communication constraints[16]. Further, dramatic reductions total cost may now be possible (as suggested by results in Chapter 2), and thus studying this alternative architecture is practically important.

3.5 Related problems in information theory

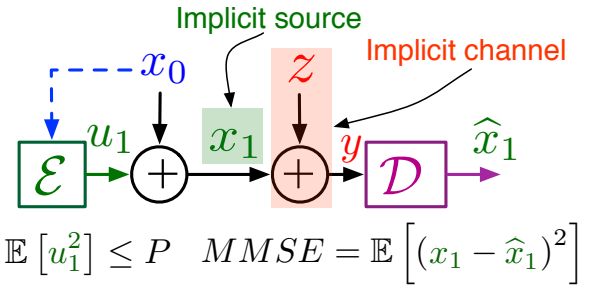


Figure 3.6: The Witsenhausen counterexample can be interpreted as a problem of communicating $\mathbf{x}_1^m$ to the second controller. The first controller is interpreted as an encoder $\mathcal{E}$, and the second controller as a decoder $\mathcal{D}$. An equivalent problem is that of having the decoder estimate $\mathbf{x}_1^m$ which is the result of a state-modification by an encoder operating under a power constraint.

[15]Not all formulations in the observer-controller architecture focus on stability. The work of Bansal and Başar [89], for instance, shows that strategies linear in the innovation can be optimal in the cost framework for first-order Gaussian ARMA model.

[16]The usage of the term "active" is inspired from the active-vision literature [90], where a camera is fitted with a controller and that chooses what part of the entire scene the camera should point to. This perspective introduces new problems in information theory even without the control objective: the source statistics into the encoder now change with the movement of the camera. What is the optimal compression that can be attained [91]?

The implicit-communication interpretation of the counterexample yields the block-diagram, shown in Fig. 3.6, that resembles block diagrams for problems in communications. The first controller is interpreted as an encoder, with an average power constraint P. The second controller can be interpreted as a decoder who wants to estimate $\mathbf{x}_1^m$ to within the smallest MMSE error[17]. The problem of obtaining the optimal tradeoff curve between P and $MMSE$ is equivalent[18] to characterizing the optimal cost $k^2P + MMSE$ for all k and σ_0.

We now examine the intellectual motivations for information-theoretic formulations that look similar to the Witsenhausen counterexample. This helps us view the counterexample as one among many related information theory problems while at the same time it helps us isolate the main difficulty.

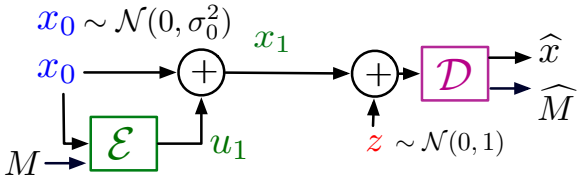


Figure 3.7: A block-diagram that can represent the Witsenhausen counterexample (where there is no message), our formulation for a triple role of control in Chapter 5.3 (where we have a message), dirty-paper coding [77] (where the state does not need to be communicated), state amplification [92] (where the goal is to communicate x_0 and message M) and state-masking [93] (where the goal is to hide x_0 and communicate message M).

Figure 3.7 shows the block-diagram of three related problems in information theory. These problems are inspired from the formulation of Gel'fand-Pinsker [94]. In their formulation, the encoder modifies the noiselessly observed state in order to communicate an independent message to the decoder. They characterize the achievable rates (in the asymptotic limit of zero error probability) in the form of an optimization problem, which can be solved using brute-force search in a finite-dimensional state-space. The "LQG version" of the problem was addressed by Costa in [77]. The block-diagram of Costa's problem, which he called "Dirty-Paper Coding[19]," is shown in Fig. 3.7. The objective here is the maximization of the communication rate of message M under a constraint on the input power $\mathbb{E}[u_1^2] \leq P$.

[17]The second controller chooses $\mathbf{u}_2^m$ in order to minimize $\mathbb{E}\left[\|\mathbf{X}_2^m\|^2\right] = \mathbb{E}\left[\|\mathbf{X}_1^m - \mathbf{U}_2^m\|^2\right]$. Equivalently, the controller chooses $\mathbf{u}_2^m$ as the MMSE estimate of $\mathbf{X}_1^m$ given $\mathbf{Y}_2^m$.

[18]A rigorous proof of this statement appears in [56]. Despite the equivalence of finding optimal solutions to these two problems, we will see in Chapter 4 that an *approximately-optimal* solution to one does not yield an approximately-optimal solution to another. This is analogous to approximations in computational-complexity theory: approximate solutions to an NP-complete problem may not yield an approximate solution to another [75].

[19]The initial state is thought of "dirt" on a paper. The goal is to write on this dirty-paper (thus changing the state) in order to communicate the message. One way to write on a dirty-paper is to erase the dirt (force the state to zero) and then write the message (appropriately coded) on it. Costa's result suggests a power-efficient strategy where existing dirt is modified to communicate the message. The result has turned out to be very important and powerful in addressing problems of watermarking [95], broadcast [96],

Costa's formulation does not care about communicating anything pertaining to the initial state, x_0. However, it ends up conveying a part of this initial state[20]. This raises issues of technical and intellectual interest: what if conveying information about the initial state is a part of the objective? This leads to the formulation of Kim, Sutivong, and Cover, called "State Amplification" [92], where the authors characterize the tradeoff between the deliverable information about x_0 and the achievable rate for communicating the message under a power constraint on the control input. A counterpart of this problem is the "State Masking" problem considered by Merhav and Shamai [93]. They consider a problem of *hiding* x_0 (*i.e.* revealing as little information about x_0 as possible) while maximizing the rate of communicating the message.

One interesting aspect shows up in this comparison: unlike in the rest of the problems, the goal in the counterexample to communicate a *modified source*, or what we called an implicit source in Chapter 1. In DPC, state-amplification, and state-masking, the source — the message M and/or the state x_0 — is explicit in that it cannot be modified by the controllers. As we saw in Chapter 1, for the counterexample, the source x_1 depends on the choice of control strategy.

Is there any instance in information theory where the goal is to communicate a modified source? Lossy transmission of a source across a channel suggests one such instance. As suggested by the source-channel separation theorem [1], after lossy compression, the goal is to transmit reliably a distorted (*i.e.* modified) version of the source. It may therefore seem that that the source can be thought of as implicit. However, the performance-measure is the error in reconstructing the *original* source. This is unlike the counterexample, where the performance is measured using the error in reconstructing the *modified* source. What if we artificially force a separation architecture to the problem? That is, the source is first compressed using lossy-source coding, and then transmitted reliably across the channel. In that case, communicating the lossy-source-codewords is indeed a problem of a communicating modified source.

This artificial separation constraint can therefore be abstracted as having the encoder know in advance the eventual reconstruction (which need not be the actual source) at the decoder. In a surprising result, Steinberg [97] noticed that this added constraint made the distributed source-coding problem solvable even though in its traditional form, the problem famously remains open except for the two-user Gaussian case [98]. Following this lead, in [99], Sumszyk and Steinberg consider the setup of dirty-paper coding (in a discrete state-space), and impose the constraint of the encoder knowing perfectly the reconstruction of the *modified* state (x_1) at the decoder. They are able to solve this problem as well (in an asymptotic setting). However, a lossy version of this problem (of which Witsenhausen's

etc. As we will see in Chapter 4, it also provides an understanding of the best-known strategies for the Witsenhausen counterexample, and provides asymptotically-optimal strategies in the limit of zero distortion for the counterexample.

[20]What the decoder is able to decode is not the initial state x_0, but a linear combination of input and the initial state, $u_1 + \alpha x_0$.

counterexample is a special case), where x_1 only needs to be reconstructed within some distortion, remains unsolved[21].

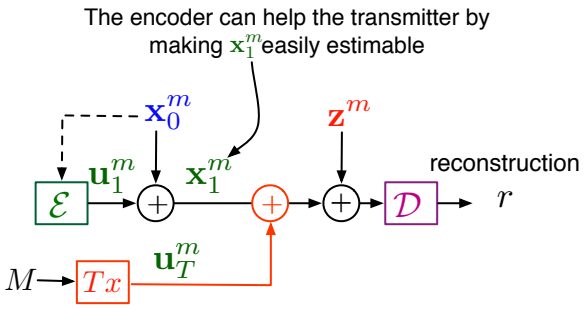


Figure 3.8: The distributed dirty-paper coding problem of Kotagiri and Laneman [100]. The encoder helps the transmitter Tx send its message to the decoder by polishing the "dirt" (*i.e.* the state).

Although this constraint of having the encoder know the reconstruction at the decoder seems artificial, it may arise naturally out of an explicit communication problem. The problem addressed by Kotagiri and Laneman [100] brings out this aspect. Motivated towards understanding a distributed implementation of dirty-paper coding, they investigate the problem of communicating across a multiple access channel with partial state information at some encoders. A special case of this problem is the distributed dirty-paper-coding, where one encoder knows the state, and the other (called the transmitter) knows the message to be communicated (see Fig. 3.8). In order to help the transmitter send its message reliably to the receiver, the encoder can simplify the estimation of $\mathbf{x}_1^m$ at the decoder. The problem therefore incentivizes signaling with an implicit "source:" $\mathbf{x}_1^m$. Kotagiri and Laneman only provide upper and lower bounds to the rate region. A complete solution to the problem is still elusive.

What aspect of these problems makes them solvable (or not)? The problems of dirty-paper coding, state-amplification, and state-masking are completely solved at least in the asymptotic case. These are also the problems where the sources/messages M and $\mathbf{x}_0^m$ are explicitly specified and do not depend on the choice of the control policy. On the other hand, for the problems that are not solved, *i.e.* the Witsenhausen counterexample, Sumszyk and Steinberg's problem in a lossy-reconstruction setting, and distributed dirty-paper coding, the controllers have the ability to modify what is being estimated. It therefore seems to us that the implicit-source aspect of the problem (combined with lossy reconstruction of the implicit source) is what makes these problems hard from an information-theoretic perspective.

[21] A Gaussian version is addressed in Chapter 5.3.

Chapter 4

An approximately optimal solution to Witsenhausen's counterexample

In Chapter 3, we saw that the lack of understanding of implicit communication (*i.e.* signaling) is one of the core reasons why problem formulations in decentralized control were forced to back-off from minimizing system costs to just attaining stability. How do we even begin to understand implicit communication? We noted in Chapter 1 that Witsenhausen's counterexample is the minimalist problem of implicit communication, and therefore is the right place to start.

We also noted why the counterexample is a hard problem from complexity-theoretic (because the problem is nonconvex, and its discrete version is NP-complete; Chapter 3.3), control-theoretic (because of the dual role of control actions: control and signaling; Chapter 3.4.2), and information-theoretic (because it is a problem of implicit source reconstruction in a distortion setting; Chapter 3.5) perspectives.

In this chapter, we develop an understanding of the signaling inherent in the Witsenhausen counterexample in a sequence of four steps. In the first step, we formulate a semi-deterministic abstraction of the problem that is much easier to analyze. It lets us understand the flow of information within signal interactions in the problem and provides us with optimal strategies that are intuitive and interpretable. The interpreted strategies are hypothesized to also be good for the original problem. However, the abstraction is an oversimplification of the Witsenhausen counterexample, and proving this hypothesis requires constructing models that bring us closer to the counterexample. These strategies may not be optimal for the original problem (and in fact, in most cases they are not optimal), but they capture the essence — the conceptual "most-significant bits" — of the information-flow and signal interactions. To show that these strategies indeed capture the essence, we need to provide guarantees on their proximity to optimality. To that end, in Step 2, we consider an LQ version of the Witsenhausen counterexample where the noise-distribution has a bounded support. We obtain a lower bound on the total costs for this problem. Using this lower bound, we show that the strategies intuited from the deterministic model attain within a *uniform constant factor of*

the optimal cost for all problem parameters.

It remains to see whether this non-Gaussian nature of the noise fundamentally alters the essence of the problem. In Step 3, we investigate the asymptotic LQG vector Witsenhausen problem and observe that the natural counterparts of the strategies that were approximately optimal for the bounded-noise counterexample continue to be approximately-optimal for this asymptotic LQG problem. This is not surprising: at asymptotically infinite vector-lengths, the Gaussian concentrates like every well-behaved distribution (including bounded-noise distributions). In particular, Gaussian distribution also asymptotically concentrates onto a bounded compact set: the same shell onto which all random variables of same variance concentrate. What happens in the scalar case, when the Gaussian distribution has a finite probability of falling far outside the typical sphere? By deriving lower bounds on the minimum possible costs, we show that the fact that Gaussian noise distribution can push the noise outside a comfort-zone is something even an optimal controller cannot deal with! This is done in Step 4 where we consider the finite-length vector Witsenhausen problem and prove that the same strategies attain within a constant factor of the optimal cost for any finite-length. In particular, this yields the first provably approximately-optimal solution to the original Witsenhausen counterexample. This solution characterizes the optimal costs of the counterexample to within a (numerically evaluated) constant factor of 8 for all problem parameters.

We propose this four-step process as a program for obtaining approximately-optimal solutions to more general decentralized LQG problems with or without external channels connecting the controllers. In order to demonstrate the applicability of this process, we will consider several example problems in Chapter 5.

4.1 Step 1: A semi-deterministic abstraction of Witsenhausen's counterexample

Deterministic models of network information theory problems were recently proposed by Avestimehr, Diggavi and Tse [37–39]. These models abstract the structural layout of a wireless communication network and the available SNR at each agent in order to gain insights into the flow of information within the signal interactions of these networks.

Just as the deterministic models capture the flow of information in explicit communication networks, is it possible that these models, after suitable modifications, might be able to capture the flow of information in LQG control networks of *implicit* communication? If this approach succeeds (as we will see, it does), then the strategies obtained from understanding the flow of information may help us design control strategies for the Witsenhausen counterexample.

In this section we provide semi-deterministic models inspired from the information-

theoretic deterministic models for decentralized scalar[1] LQG networks. What aspect of the original Linear-Quadratic-Gaussian models do these semi-deterministic models abstract? These models of course retain the physical structure (*i.e.* the connectivity) and the temporal ordering of actions of the controllers. They preserve a simple yet crucial aspect of linear models: that the effect of small signals on interactions with larger signals is small, and hence limited to only the least-significant bits. This aspect preserves the flow of information in the original problem. Because of their binary alphabet, the semi-deterministic models do not retain the quadratic costs or the Gaussian priors.

The semi-deterministic model for LQG decentralized control problems is introduced below using Witsenhausen's counterexample as an example:

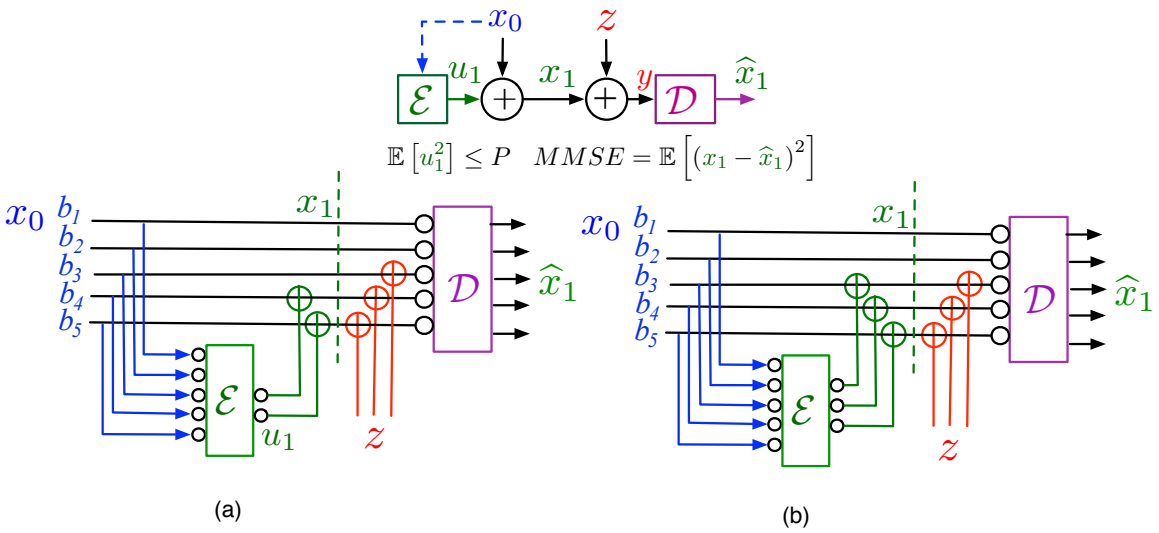


Figure 4.1: A semi-deterministic model for Witsenhausen's counterexample. The expansion $b_1 b_2 b_3 . b_4 b_5 \ldots$ runs until infinity. Figure (a) assumes that the power of the input u_1 is chosen such that the encoder can affect only the least-significant bits $b_4, b_5, \ldots$. Figure (b) assumes that the encoder can affect $b_3, b_4, b_5, \ldots$. If the power is chosen as in (b), then the encoder can force the bit b_3 and the ensuing bits in the expansion of x_1 to zero. The decoder then has a perfect estimate of x_1. Even though the model has an infinite-bit expansion of the state (unlike its information-theoretic counterpart in [37–39]), it can be truncated in visual representation once it is clear that further expansion does not aid intuition.

- Each real-valued system variable is represented in binary. For instance, in Fig. 4.1,

[1]Vector control problems where the initial state or noise is correlated across the vector elements (*i.e.* has non-diagonal covariance matrix) turn out to be the counterparts of Gaussian MIMO networks. The proposed deterministic models for MIMO networks, for instance by Anand and Kumar [101], do not appear to be as intuitive as those for SISO (*i.e.* scalar) networks. In the special case where one of the covariances is identity ("white"), the other axes can be rotated to attain a diagonal covariance matrix for both.

the state is represented by $b_1 b_2 b_3 . b_4 b_5 \ldots$, where b_1 is the most significant bit, and the expansion can run for infinitely many bits after the decimal point.[2]

- The location of the decimal point is determined by the signal-to-noise ratio (SNR), where signal refers to the state or input to which noise is added. It is given by $\lfloor \log_2 (SNR) \rfloor - 1$. Noise can only affect the bit just before the decimal point (*i.e.* bit b_3), and the bits following it (bits $b_4, b_5, \ldots$).

- The power[3] of a random variable A, denoted by $max(A)$ is defined as the most significant bit that is 1 among all the possible (binary-represented) values that A can take[4]. For instance, if $A \in \{0.01, 0.11, 0.1, 0.001\}$, then A has the power $max(A) = 0.1$.

- Additions/subtractions in the original control model are replaced by XORs. Noise is assumed to be Ber(0.5), *i.e.* it takes value 1 with probability 0.5 and probability 0 with probability 0.5.

- In addition to being LQG, if the original control model has an external channel connecting the controllers, then an external channel connects the controllers in its semi-deterministic version as well. The capacity of the external channel in the semi-deterministic version is the integer part (floor) of capacity of the actual external channel.

In the information-theoretic deterministic model [37], the binary expansions are limited to those above the noise-level. The bits below the noise-level are corrupted by noise, and hence are insignificant for communication at high SNR. These bits can therefore safely be ignored while communicating. Can we ignore these bits in control as well? Let us take the example of the Witsenhausen counterexample to see this. If the semi-deterministic version of the counterexample (shown in Fig. 4.1) is made deterministic, then the bits below the noise-level at each controller are removed. Therefore, the decoder does not observe or estimate the bits below the noise level, and there is no reason why the encoder should spend any power for modifying them. However, in this control problem there is interest in estimating these bits at the decoder because the goal is to reduce costs. Thus the binary expansions in our models are valuable even after the decimal point (below the noise level), and in fact we assume that these binary expansions are not truncated and can run until infinity. Because random noise is also modeled, these models are also not completely deterministic.

Alternatively, we can keep the models deterministic, but introduce erasures on links. This model is shown in Fig. 4.2. These models are equivalent to our semi-deterministic

[2]Though intuition can be gained from just a finite bit-expansion, but not by simply truncating the expansion below the observation noise level, as we will see soon.

[3]This power is really the log of the power of the original random variable, and is only a constant factor away from decibels (dBs) used to measure power in communications.

[4]We note that our definition of $max(A)$ is for clarity and convenience, and is far from unique amongst the good choices.

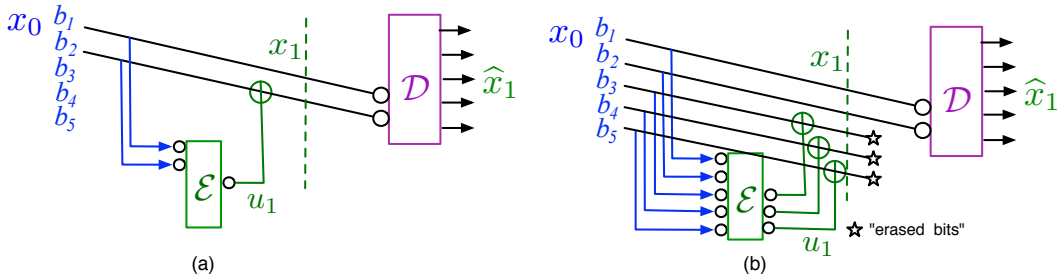


Figure 4.2: (a) A deterministic model of the counterexample based on the modeling in [39]. The model does not suffice because it does not allow the encoder to modify the least-significant bits. This is the reason we move to semi-deterministic models in Fig. 4.1. (b) An alternative to semi-deterministic models: an erasure-based deterministic model. These erasure-based models yield the same strategies as the semi-deterministic models. We prefer the semi-deterministic model with XORs with Ber(0.5) representing noise purely because we feel they are better at illustrating the possibility of improving state-estimability.

models, and therefore will yield the same strategies. We do not use it merely because the semi-deterministic model brings out the aspect of control — that it can be used to improve state-estimability — more explicitly and intuitively[5].

4.1.1 Optimal strategies for the semi-deterministic abstraction

The semi-deterministic abstraction introduced in last section is easy to solve. In this section, we provide a solution by characterizing the optimal tradeoff between the input power $max(u_1)$ and the power in the reconstruction error $max(x_2)$. The minimum total cost problem is a convex dual of this problem, and can be obtained easily. Let the power of x_0, $max(x_0)$ be σ_0^2. The noise power is assumed to be 1.

Case 1: $\sigma_0^2 > 1$. This problem is shown in Fig. 4.1.

Achievable strategies: If $max(u_1) < 1$, we use the zero-input strategy, *i.e.* use $u_1 = 0$. Because we still recover bits b_1 and b_2, we only have a reconstruction error of power 1.

On the other hand, for $max(u_1) \geq 1$, the encoder can affect the last three bits or more. But the decoder already knows bits b_1 and b_2 because these are not affected by noise. A good strategy is therefore to force the last three bits, b_3, b_4 and b_5, to zero, and the required power is just $max(u_1) = 1$.[6] The decoder now has a perfect estimate of $\widehat{x}_1$: the two most significant bits are received noiselessly, and the bits of lower significance are known to be zeros.

[5]The model also helps in visualizing atypical behavior of noise, when it creeps up to corrupt more bits. This interpretation is useful in deriving lower bounds in Step 4.

[6]More bits can be forced to zero if one wants to use $max(u_1) > 1$, but it cannot lower the reconstruction error since the error is already zero.

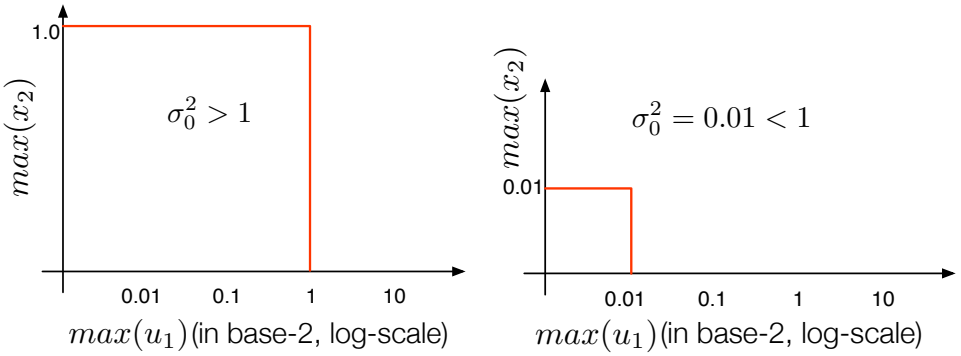


Figure 4.3: Optimal tradeoff between $max(u_1)$ and the $max(x_2)$, the reconstruction error, for the semi-deterministic version of Witsenhausen's counterexample. The figure on left is for $\sigma_0^2 > 1$, that is, the noise power is smaller than that of the initial state. The one on right is for $\sigma_0^2 = 0.01 < 1$.

Outer bound on the achievable region: For $\sigma_0^2 > 1$, can one do any better? For power $max(u_1) \geq 1$, we already have a zero-reconstruction error, and hence cannot do any better! If power $max(u_1) < 1$, as shown in Fig. 4.1(a), the bit b_3 of x_0 cannot be reconstructed at the decoder: the encoder has no ability to affect it, and the decoder only receives a completely noisy observation of it. The encoder cannot even hope to communicate b_3 because the bits that it can affect in order to signal b_3 are already mangled by noise. The reconstruction error is thus dominated by b_3, and has a power 1, the same if the encoder chose to affect no bits at all.

The matching of the achievable region and the outer bounds yields the optimal tradeoff curve shown in Fig. 4.3(a).

Case 2: $\sigma_0^2 < 1$.

Achievable strategy: If $max(u_1) < \sigma_0^2$, we use the zero input strategy, incurring an error-power-cost of σ_0^2. If $max(u_1) \geq \sigma_0^2$, we use the zero-forcing strategy where the encoder forces the state to zero. The decoder estimates the state to be zero, and the reconstruction error is also zero.

Outer bound on the achievable region: If $max(u_1) \geq \sigma_0^2$, we cannot hope to improve on error because it is already zero. If $max(u_1) < \sigma_0^2$, we cannot signal anything about the state to the decoder, so it is best to use no power at all.

Again, the achievable region and the outer bounds match, and the resulting tradeoff curves are shown in Fig. 4.3(b).

Interpretation of the strategy: How can we interpret the strategy suggested by this semi-deterministic abstraction? Clearly, the strategy depends on how large the observation noise is relative to the initial state. If initial state has power σ_0^2 smaller than noise power (*i.e.* $\sigma_0^2 < 1$), then the encoder should either force the entire state to zero (to force reconstruction error to zero), or use no input at all (because the power in reconstruction error does not change with input power smaller than noise power).

If the $\sigma_0^2 > 1$, and the input power $P < 1$, then the encoder should use no input at all, *i.e.* $u_1 = 0$. But if $P > 1$, then the encoder should force the bits that are affected by noise to zero. As shown in Figure 4.4, on a real line, this truncation operation is really just

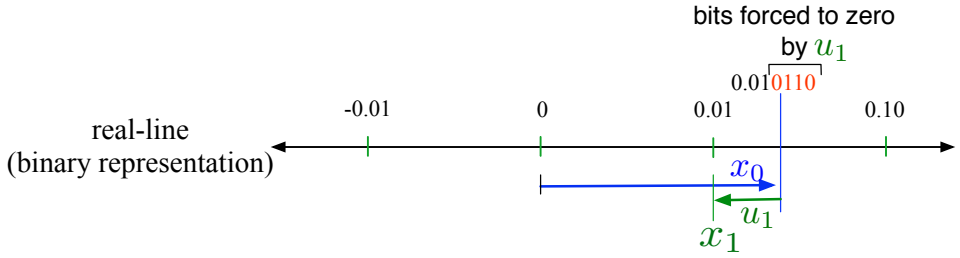


Figure 4.4: Scalar quantization can be thought of as truncation of bits — exactly the strategy suggested by the deterministic model.

quantization. This suggests a natural strategy for the counterexample — quantize the initial state x_0 onto a set of regularly-spaced quantization points. This is the strategy adopted by Mitter and Sahai [18], which is an extension of Witsenhausen's strategy [14] that uses just two quantization points.

Our hypothesis from the semi-deterministic model is therefore that the strategies of zero-input, zero-forcing, and quantization (depending on problem parameters) are good strategies for the counterexample. The three remaining steps prove this hypothesis.

4.2 Step 2: The uniform-noise counterexample

The benefit of the semi-deterministic model lies in its simplicity: the binary alphabet lays the problem structure bare, and helps see the possible information flows. However, the binary alphabet also contributes to a feel of extreme toyness: how can we be sure that the simplification offered by this binary alphabet is not an over-simplification?

In this section, we test the hypothesis of goodness of the strategies obtained from the semi-deterministic model on a model that has a continuous state-space. Notice that the noise is assumed to affect only the last few bits of the state in the semi-deterministic model. In order to keep the problem reasonably close to the semi-deterministic version, we retain this property by assuming that the noise takes values in a bounded support $(-a, a)$. At this point, we could continue with a power model that measures power of a variable by the maximum value it can take. But this feels too conservative because it allows for an adversarial choice of the noise[7]. In order to get away from this conservative model, we impose quadratic costs on power and error. As we will see, quadratic costs are easier to analyze. Fortuitously, they also bring us closer to the LQG formulation. The resulting model is shown in Fig. 4.5.

[7]The flavor of our results does not change even if the noise is adversarial and bounded. The proof appears in [63].

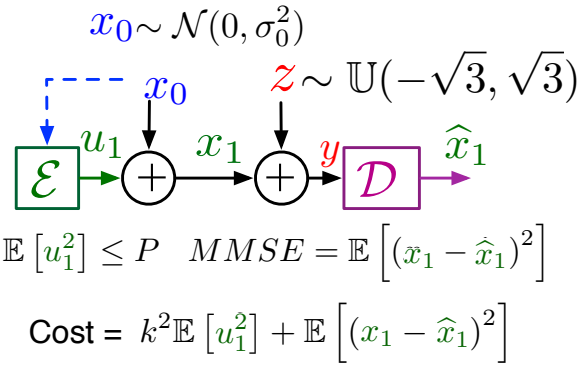


Figure 4.5: A version of Witsenhausen's counterexample with uniform noise distribution $Z \sim \mathbb{U}(-\sqrt{3}, \sqrt{3})$, which has variance 1.

In this section, we will focus on the case when the noise Z is distributed uniformly in the interval $(-\sqrt{3}, \sqrt{3})$ (so that the variance of Z is 1). For clarity of exposition as well as generality, our theorems only assume that $Z \in (-a, a)$ for some a, has variance 1, and the distribution of Z has a finite differential entropy [28] $h(Z)$.

4.2.1 Upper bounds on the costs based on the strategies obtained from the semi-deterministic model

We hypothesized in the last section that the strategies of zero-input, zero-forcing, and quantization (depending on problem parameters) are good strategies for the counterexample. The following theorem tests this for this uniform-noise counterexample.

Theorem 1. *An upper bound on the costs for Witsenhausen's formulation with bounded noise $Z \in (-a, a)$ with $Var(Z) = 1$, is given by*

$$\overline{\mathcal{J}}_{opt} \leq \min\left\{k^2a^2, \frac{\sigma_0^2}{\sigma_0^2+1}, k^2\sigma_0^2\right\}. \tag{4.1}$$

Proof. We consider the following three strategies 1) an essentially scalar quantization strategy that quantizes the entire real line with bins of sizes $2a$ in each dimension, 2) the zero-input strategy, followed by LLSE estimation at the second controller, and 3) the zero-forcing strategy. For a given (k, σ)-pair, the strategy with minimum cost is chosen.

For the quantization strategy, the input forces the state to the nearest quantization point. The magnitude of the input is therefore bounded by a. Since the bins are disjoint, there are never any errors at the second controller (because the noise is smaller than a). The cost is therefore upper bounded by k^2a^2. For the zero-input strategy with LLSE estimation, the cost is given by $\sigma_0^2 - \frac{\sigma_0^2\sigma_0^2}{\sigma_0^2+1} = \frac{\sigma_0^2}{\sigma_0^2+1}$. For zero-forcing, the input is forced to zero, and thus the cost is $k^2\sigma_0^2$. This completes the proof. □

4.2.2 A signaling-based lower bound on the costs

How well can any strategy do? For the semi-deterministic model, the bound on the performance of any strategy was rather easy to obtain: there were only finitely many possibilities! For the uniform-noise counterexample, however, we are forced to find limits on how well we can signal through the implicit channel. The derivation of the following theorem obtains these limits and exploits them to obtain a lower bound to the bounded noise problem.

Theorem 2. *A lower bound on the costs for Witsenhausen's formulation with noise Z distributed such that the differential entropy of Z, given by $h(Z)$, is finite, is given by*

$$\overline{\mathcal{J}}_{opt} \geq \inf_{P \geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P} \right)^+ \right)^2, \tag{4.2}$$

where $(x)^+ = \max\{x, 0\}$,

$$\kappa(P) = \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e \left(\left(\sigma_0 + \sqrt{P} \right)^2 + 1 \right)}. \tag{4.3}$$

Proof. For a fixed $P := \frac{1}{m}\mathbb{E}\left[\|\mathbf{U}_1^m\|^2\right]$, we will obtain a lower bound on the $MMSE$. First, we need the following lemma (which is a straightforward consequence of the triangle inequality):

Lemma 1. *For any three vector random variables A, B and C,*

$$\sqrt{\mathbb{E}\left[\|B - C\|^2\right]} \geq \left| \sqrt{\mathbb{E}\left[\|A - C\|^2\right]} - \sqrt{\mathbb{E}\left[\|A - B\|^2\right]} \right|. \tag{4.4}$$

Proof. See Appendix A.2. □

Substituting $\mathbf{X}_0^m$ for A, $\mathbf{X}_1^m$ for B, and $\mathbf{U}_2^m$ for C in Lemma 1, we get

$$\sqrt{\mathbb{E}\left[\|\mathbf{X}_1^m - \mathbf{U}_2^m\|^2\right]} \geq \sqrt{\mathbb{E}\left[\|\mathbf{X}_0^m - \mathbf{U}_2^m\|^2\right]} - \sqrt{\mathbb{E}\left[\|\mathbf{X}_0^m - \mathbf{X}_1^m\|^2\right]}. \tag{4.5}$$

We wish to lower bound $\mathbb{E}\left[\|\mathbf{X}_1^m - \mathbf{U}_2^m\|^2\right]$. The second term in the RHS is smaller than $\sqrt{mP}$. Therefore, it suffices to lower bound the first term on the RHS of (4.5). To that end, we will interpret $\mathbf{U}_2^m$ as an estimate for $\mathbf{X}_0^m$.

How can we lower bound this distortion term? The total power input to the implicit channel $X_1 - Y_2$ is bounded. Using information theory, we can find how many bits of information can be signaled through this channel. This is given by the channel capacity, which is the maximum possible mutual information $I(\mathbf{X}_1^m; \mathbf{Y}_2^m)$ across the channel. This

mutual information can be bounded as follows

$$\begin{aligned}
I(\mathbf{X}_1^m;\mathbf{Y}_2^m) &= h(\mathbf{Y}_2^m) - h(\mathbf{Y}_2^m|\mathbf{X}_1^m) \\
&\leq \sum_i h(Y_{2,i}) - h(\mathbf{Y}_2^m|\mathbf{X}_1^m) \\
&= \sum_i \left(h(Y_{2,i}) - h(Y_{2,i}|X_{1,i})\right) \\
&= \sum_i I(X_{1,i};Y_{2,i}) \\
&\stackrel{(a)}{=} mI(X_1;Y_2|Q) \\
&= m\left(h(Y_2|Q) - h(Y_2|X_1,Q)\right) \\
&= m\left(h(Y_2|Q) - h(Y_2|X_1)\right) \\
&\leq m\left(h(Y_2) - h(Y_2|X_1)\right) \\
&\leq mI(X_1;Y_2).
\end{aligned}$$

In (a), the random variables X_1, Y_2 and Q are defined as follows: $X_1 = X_{1,i}$ if $Q = i$ (and Y_2 is defined similarly), and Q is distributed uniformly on the discrete set $\{1, 2, \ldots, m\}$. Now,

$$\begin{aligned}
Y_2 &= X_1 + Z \\
&= X_0 + U_1 + Z.
\end{aligned}$$

The variance of Y_2 is maximized when X_0 and U_1 are aligned, and it equals $\left(\sigma_0 + \sqrt{P}\right)^2 + 1$. Thus,

$$\begin{aligned}
I(X_1;Y_2) &= h(Y_2) - h(Y_2|X_1) \\
&= h(Y_2) - h(Z) \\
&\stackrel{(a)}{\leq} \frac{1}{2}\log_2\left(2\pi e\left(\left(\sigma_0 + \sqrt{P}\right)^2 + 1\right)\right) - h(Z) \\
&= \frac{1}{2}\log_2\left(\frac{2\pi e\left(\left(\sigma_0 + \sqrt{P}\right)^2 + 1\right)}{2^{2h(Z)}}\right),
\end{aligned} \tag{4.6}$$

where (a) follows from the observation that for given second moment of the random variable, the distribution that maximizes the differential entropy is Gaussian.

Pretending we wish to communicate $\mathbf{X}_0^m$ across the $X_1 - Y_2$ channel (instead of $\mathbf{X}_1^m$), we can obtain a lower bound on the distortion in reconstructing $\mathbf{X}_0^m$ as follows: $\mathbf{X}_0^m$ is a Gaussian source that needs to be communicated across a channel of mutual information (and hence also the capacity) upper bounded by the expression in (4.6). The distortion in

reconstructing $\mathbf{X}_0^m$ is therefore lower bounded by $D_{\sigma_0^2}(C_{X_1-Y_2})$ where $D_{\sigma_0^2}(R) := \sigma_0^2 2^{-2R}$ is the distortion-rate function [28, Ch. 13] of a Gaussian source, and $C_{X_1-Y_2}$ is the capacity across the $X_1 - Y_2$ channel.

Thus, the mean-squared distortion in reconstructing $\mathbf{X}_0^m$ is lower bounded by

$$\begin{aligned} \frac{1}{m}\mathbb{E}\left[\|\mathbf{X}_0^m - \mathbf{U}_2^m\|^2\right] &\geq D_{\sigma_0^2}(C_{X_1-Y_2}) \\ &\geq \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e\left(\left(\sigma_0 + \sqrt{P}\right)^2 + 1\right)}. \end{aligned} \tag{4.7}$$

A lower bound on the $MMSE$ follows from (4.5) and (4.7). The theorem follows from the minimizing the sum of k^2P and $MMSE$ over non-negative values of P. □

We note that the theorem only makes use of finite differential entropy of Z, and not the bounded nature of its distribution. Thus the theorem is also valid for Gaussian distributions of noise.

4.2.3 Quantization-based strategies are approximately optimal for the uniform-noise counterexample

For the semi-deterministic version in Chapter 4.1, we could prove the optimality of the proposed strategies essentially by exhaustive search. We then hypothesized that these strategies will be good for more realistic problems as well. Can we hope that these strategies are exactly optimal? Even in information theory, the deterministic models of Avestimehr, Diggavi and Tse [37–39] only provide approximately optimal strategies: the strategies attain within a constant gap of the optimal rates where the gap is uniform for all problem parameters. So the deterministic model there does not capture all the modeling details.

Can we hope for similar approximation results here? What will be the right way to approximate the costs? In information theory, the capacity is approximated to within a finite number of bits. At high SNR, capacity is usually logarithmic in power, so these approximations can be thought of as multiplicative approximations to the optimal power. Since we are dealing with power (and error) here, could the right approximation be multiplicative here as well? Indeed, because the costs themselves converge to zero as k or σ_0^2 converge to zero, an additive approximation is not useful. Further, the problem itself is normalized by assuming the noise variance $\sigma_z^2 = 1$. What if $\sigma_z^2 \neq 1$? If the strategies are also scaled by σ_z, the total costs are also multiplied by σ_z^2. Thus this assumption of $\sigma_z^2 = 1$ retains validity with a multiplicative approximation: even if $\sigma_z^2 \neq 1$, the multiplicative factor of the approximation will remain the same.

The following theorem shows that the strategies hypothesized using the semi-deterministic model (in Chapter 4.1), namely the quantization-based strategies complemented by linear

strategies of zero-forcing and zero-input, are approximately-optimal in this constant-factor sense for the uniform-noise counterexample.

Theorem 3. *For Witsenhausen's formulation with non-Gaussian bounded noise $Z \in (-a, a)$,*

$$\inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2 \leq \overline{\mathcal{J}}_{opt}$$
$$\leq \mu \left(\inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2\right),$$

where $\mu \leq \frac{200a^2}{2^{2h(Z)}}$, and the upper bound is achieved by quantization-based strategies, complemented by linear strategies. For example, for $Z \sim \mathbb{U}(-\sqrt{3}, \sqrt{3})$, the uniform distribution of variance 1, $\mu \leq 50$.

Proof. See Appendix A.3. □

Note that the result is not asymptotic: the constant factor is uniform over all dimensions[8].

Remark: The constant factor of $\frac{200a^2}{2^{2h(Z)}}$ is not really uniform over all problem parameters, since it is a function of $h(Z)$ and a. However, scaling the distribution by a factor of β would increase both the numerator and the denominator by a factor of β^2, keeping the ratio constant. Thus, fixing the shape of noise distribution and the initial state distribution (allowing them to be scaled), scaling either of them is not going to alter the constant factor. We also note that tighter bounds on the constant factor, that depend only on the variance of the noise (and not on a), can be derived in the limit of large dimensions using laws of large numbers. A demonstration of this derivation is the Gaussian case, which is discussed next.

4.3 Step 3: The Gaussian counterexample: asymptotically infinite-length case

How do we conceptually move from a uniform distribution to a Gaussian one? One important aspect of uniform distribution is its bounded support. How can we generate a bounded support for a Gaussian distribution? One option is that we can truncate it, but this direct truncation yields very loose bounds [102]. An alternative is to consider a vector of iid Gaussian variables. If the Gaussian and the uniform distribution have the same variance σ^2, the laws of large-numbers ensure that the Gaussian vector falls very likely in a bounded shell of radius close to $m\sigma^2$.

[8]We note that the ratio improves as the number of dimensions increases to infinity because both upper and lower bounds improve due to concentration.

This observation inspires us to consider the vector Witsenhausen counterexample (introduced in Chapter 3.1) in the asymptotic limit of infinite vector length. In this limit, we characterize the asymptotically optimal costs for the problem to within a constant factor for all values of problem parameters k and σ_0^2. In the next section, we will use the understanding gained from this analysis to obtain approximate-optimality for the original (scalar) counterexample as well.

The following theorem provides the asymptotic characterization.

Theorem 4. *For the vector version of Witsenhausen's counterexample, in the limit of dimension $m \to \infty$, the optimal expected cost $\overline{\mathcal{J}}_{\min}(k^2)$ satisfies*

$$\frac{1}{\mu_1} \min\left\{k^2,\ k^2\sigma_0^2,\ \frac{\sigma_0^2}{\sigma_0^2+1}\right\} \le \overline{\mathcal{J}}_{\min}(k^2) \le \min\left\{k^2,\ k^2\sigma_0^2,\ \frac{\sigma_0^2}{\sigma_0^2+1}\right\}. \tag{4.8}$$

Alternatively, $\overline{\mathcal{J}}_{\min}(k^2)$ satisfies

$$\inf_{P\ge 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2 \le \overline{\mathcal{J}}_{\min}(k^2) \le \mu_2 \inf_{P\ge 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2, \tag{4.9}$$

where $(\cdot)^+$ is shorthand for $\max(\cdot, 0)$ and

$$\kappa(P) = \frac{\sigma_0^2}{\sigma_0^2 + 2\sigma_0\sqrt{P} + P + 1}. \tag{4.10}$$

The factors μ_1 and μ_2 are no more than 11 (numerical evaluation shows that $\mu_1 < 4.45$, and $\mu_2 < 2$).

Proof. As with the uniform-noise counterexample, the proof here proceeds in three steps. Chapter 4.3.1 provides a lower bound on the expected cost that is valid for all dimensions. This provides the expressions on the two sides of (4.9). An upper bound is then derived in Chapter 4.3.2 by providing three schemes, and taking the best performance among the three. This provides the expressions in (4.8).

Fig. 4.6 partitions the (k^2, σ_0^2) parameter space into three different regions, showing which of the three upper bounds is the tightest for various values of k^2 and σ_0^2. It is interesting to note that the nonlinear VQ scheme is required only in the small-k large-σ_0^2 regime. A similar figure in [25, Fig. 1] for the scalar problem shows that the same regime is interesting there as well.

A 3-D plot of the ratio between the upper and lower bounds for varying k^2 and σ_0^2 is shown in Fig. 4.7. The figure shows that the ratio is bounded by a constant μ_1, numerically evaluated to be 4.45, and attained at $k^2 = 0.5$ and $\sigma_0^2 = 1$. The figure also shows that for most of the (k^2, σ_0^2) parameter space, the ratio is in fact close to 1 so the upper and lower bounds are almost equal there.

This asymptotic characterization is tightened by improving the upper bound in Chapter 4.3.3. The new strategy uses a balanced combination of the information-theoretic strategy of Dirty-Paper Coding (DPC) and linear control described. Numerical evaluation of this ratio leads us to conclude that $\mu_2 < 2$, as is illustrated in Fig. 4.12. This yields (4.9).

Finally, Appendix A.5 complements the plots by giving an explicit proof that the ratio of the upper and lower bounds is always smaller than 11. □

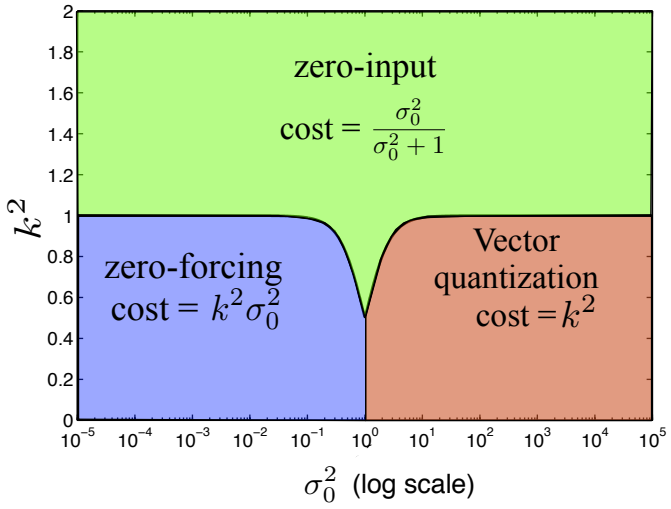


Figure 4.6: The plot maps the regions where each of the three schemes (VQ, zero-forcing $\mathbf{x}_0^m$, and zero input) perform better than the other two. For large k, zero input performs best. For small k and small σ_0^2, the cost of zero-forcing the state is small, and hence the zero-forcing scheme performs better than the other two. For small k but large σ_0^2, the nonlinear VQ cost is the smallest amongst the three.

4.3.1 A lower bound on the expected cost

Witsenhausen [14, Chapter 6] derived a lower bound on the costs for the counterexample. We first state his lower bound, and then provide our lower bound for the Gaussian case.

Witsenhausen's existing lower bound

Witsenhausen [14, Chapter 6] derived the following lower bound on the optimal costs for the scalar problem.

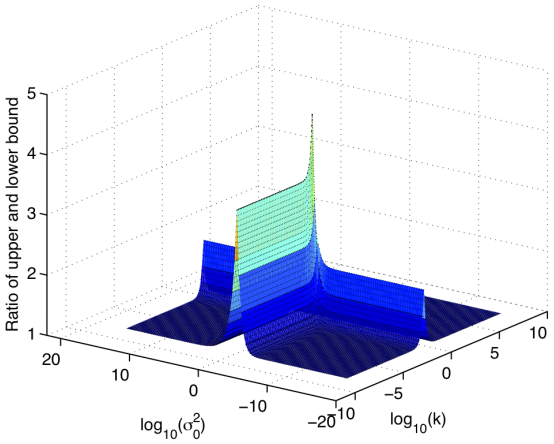


Figure 4.7: The plot shows the ratio of the upper bound in (4.8) to the lower bound in (4.9) for varying σ_0 and k. The ratio is upper bounded by 4.45. This shows that the proposed schemes achieve performance within a constant factor of optimal for the vector Witsenhausen problem in the limit of large number of dimensions. Notice the ridges along the parameter values where we switch from one control strategy to another in Fig. 4.6.

Theorem 5 (Witsenhausen's lower bound)**.** *The optimal cost for the scalar Witsenhausen counterexample is lower bounded by*

$$\overline{\mathcal{J}}_{\min}^{scalar}(k^2) \geq \frac{1}{\sigma_0}\int_{-\infty}^{+\infty} \phi\left(\frac{\xi}{\sigma_0}\right) V_k(\xi) d\xi, \tag{4.11}$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})$ *is the standard Gaussian density and*

$$V_k(\xi) := \min_a [k^2(a-\xi)^2 + h(a)], \tag{4.12}$$

where

$$h(a) := \sqrt{2\pi}a^2\phi(a)\int_{-\infty}^{+\infty} \frac{\phi(y)}{cosh(ay)} dy. \tag{4.13}$$

However, Witsenhausen's scalar-specific proof of this lower bound does not generalize to the vector case. The following theorem provides a newer, better, and simpler (to work with) lower bound that is valid for all vector lengths.

Our lower bound

Corollary 1 (Lower bound to the vector Witsenhausen counterexample)**.** *For all* $m \geq 1$*, and all strategies* S*, given an average power* P *of* $\mathbf{u}_1^m$*, the second stage cost,* $\overline{\mathcal{J}}_2(S)$ *is lower*

bounded by

$$\overline{\mathcal{J}}_2(S) \geq \overline{\mathcal{J}}_{2,\min}(P) \geq \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2, \tag{4.14}$$

where $\kappa(P)$ is the function of P given by (4.10)*. Equivalently, the optimal total cost is lower bounded by*

$$\overline{\mathcal{J}}_{\min}(k^2) \geq \inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2. \tag{4.15}$$

Proof. Follows directly from Theorem 2 with substitution of $h(Z)$ by $\frac{1}{2}\log_2(2\pi e)$, the differential entropy of a $\mathcal{N}(0,1)$ random variable. □

Fig. 4.8 plots Witsenhausen's lower bound from [14] and compares it with the lower bound of Corollary 1. A particular sequence of $k = \frac{100}{n^2}$ and $\sigma_0^2 = 0.01n^2$ is chosen to visually demonstrate that for this sequence of problem parameters, in the limit of $n \to \infty$, the ratio of the bounds diverges to infinity. Thus, we conclude that prior to this work, it was not possible to provide a uniform (over problem parameters) characterization of the optimal cost to within a constant factor for the scalar problem. Such a characterization needs a tightening of the lower bound in Corollary 1 as well, and is provided in Chapter 4.4.

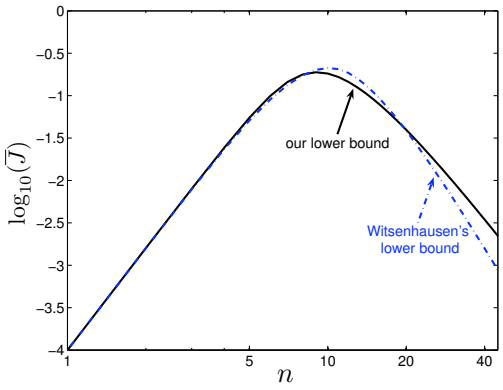


Figure 4.8: Plot of the two lower bounds on the optimal cost as a function of n, with $k_n = \frac{100}{n^2}$, $\sigma_{0,n} = 0.01n^2$ on a log-log scale for comparing the two lower bounds. The figure shows that the vector lower bound derived here is tighter than Witsenhausen's scalar lower bound in certain cases.

4.3.2 A vector-quantization upper bound on the asymptotic expected cost

In Theorem 4, the upper bound is a minimum of three terms. This section describes a nonlinear strategy that asymptotically (in the number of dimensions) attains the cost of

k^2 given by the first term of (4.8). We call the strategy the Vector Quantization (VQ) scheme. The proof uses a randomized code that exploits common randomness. For clarity of exposition, we only outline the proof here. For a rigorous proof, we refer the reader to [56]. Alternatively, an upper bound on the asymptotic cost for a quantization strategy can also be obtained by taking limits of our upper bound in Chapter 4.4 for the finite-dimensional problem. This alternative route yields a bound that is looser, but suffices to obtain constant-factor results.

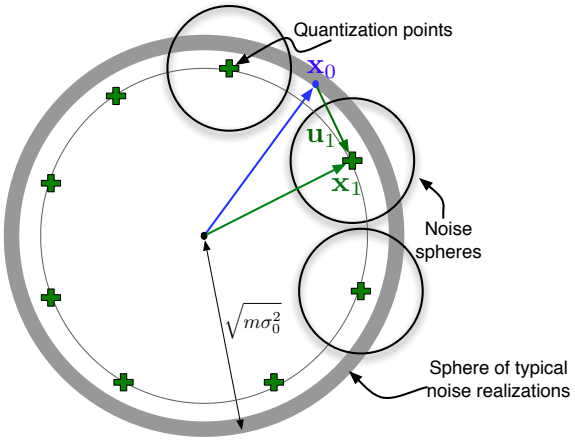


Figure 4.9: An illustration of the vector quantization scheme. The decoding is asymptotically error free as long as the noise-spheres do not intersect. This condition requires that the power P of the first controller exceed the noise variance $\sigma_z^2 = 1$.

This is a quantization-based control strategy and is illustrated in Fig. 4.9, where '+'s denote the VQ quantization points. The quantization points are generated randomly according to the distribution $\mathcal{N}(0, (\sigma_0^2 - P)\mathbb{I})$. This set of quantization points is referred to as the *codebook*, denoted by $\mathbb{Q}$. Given a particular realization of the initial state $\mathbf{x}_0^m$, the first controller finds the point $\mathbf{x}_1^m$ in the codebook closest to $\mathbf{x}_0^m$. The input $\mathbf{u}_1^m = \mathbf{x}_1^m - \mathbf{x}_0^m$ then drives the state to this point. The number of quantization points is chosen carefully — there are sufficiently many of them to ensure that the required average power of $\mathbf{u}_1^m$ is close to P, but not so many that there could be confusion at the second controller.

More precisely, a codebook $\mathbb{Q}$ of 2^{mR} quantization points $\{\mathbf{x}_q^m(1), \ldots, \mathbf{x}_q^m(2^{mR})\}$ is chosen by drawing the quantization points iid in $\mathbb{R}^m$ randomly from the distribution $\mathcal{N}(0, (\sigma_0^2 - P)\mathbb{I})$, where the operating "rate" R and the power P satisfy the pair of equalities

$$R = \mathcal{R}(P) + \frac{\delta}{2} = \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{P}\right) + \frac{\delta}{2} \tag{4.16}$$

$$C(P) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2 - P}{\sigma_z^2}\right) = R + \frac{\delta}{2}, \tag{4.17}$$

for small $\delta > 0$ where $\mathcal{R}(\cdot)$ is the rate-distortion function for a Gaussian source of variance σ_0^2 [28, Pg. 345], and $C(\cdot)$ is the capacity of an AWGN channel with input power constraint $\sigma_0^2 - P$.

With this careful choice, the state $\mathbf{x}_1^m$ can be recovered perfectly in the limit $m \to \infty$ because the capacity $C(P)$. Intuitively, we require the power to be large enough so that the noise-spheres in Fig. 4.9 do not intersect. We show in Appendix A.4 that the two conditions (4.16) and (4.17) are satisfied, and hence the noise-spheres do not intersect, when average the input power $P > \sigma_z^2 = 1$. Thus, asymptotically, $\overline{\mathcal{J}}_2 = 0$, $\overline{\mathcal{J}}_1 = P$ and the total cost approaches k^2.

4.3.3 An improved upper bound using dirty-paper coding, and a conjecture on the optimal strategy

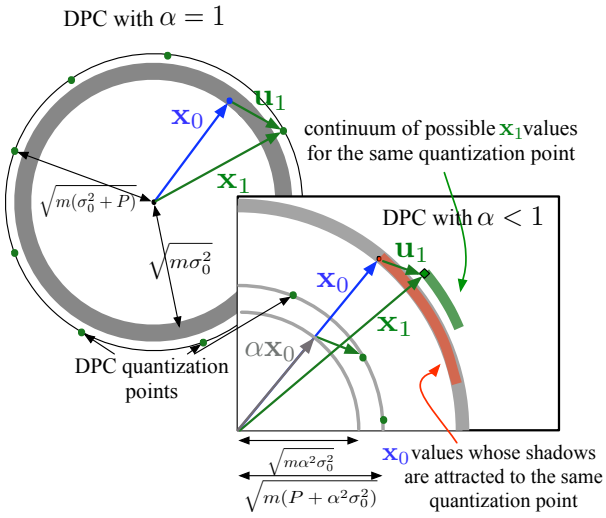


Figure 4.10: A geometric representation of the dirty-paper coding scheme (on left with the DPC-parameter $\alpha = 1$, and on right for $\alpha < 1$) of Chapter 4.3.3. The grey shell contains the typical $\mathbf{x}_0^m$ realizations. The VQ scheme (see Fig. 4.9) quantizes to points inside this shell. The DPC scheme quantizes the state to points outside this shell. For the same power in the input $\mathbf{u}_1^m$, the distances between the quantization points of the DPC scheme is larger than those for the VQ scheme, making it robust to larger observation noise variances.

We saw in Chapter 3.5 that the vector Witsenhausen counterexample is deeply connected to Costa's problem of dirty-paper coding (DPC) [77]. Dirty-paper coding techniques [77] can also be thought of as performing a (possibly soft) quantization. The quantization points are chosen randomly in the space of realizations of $\mathbf{x}_1^m$ according to the distribution $\mathcal{N}(0, (P +$

$\alpha^2\sigma_0^2)\mathbb{I}$). For $\alpha = 1$ the quantization is hard and a pictorial representation is given in Fig. 4.10, with '$\circ$' denoting the DPC quantization points. Given the vector $\mathbf{x}_0^m$, the first controller finds the quantization point $\mathbf{x}_1^m$ closest to $\mathbf{x}_0^m$ and again uses $\mathbf{u}_1^m = \mathbf{x}_1^m - \mathbf{x}_0^m$ to drive the state to the closest point. For $\sigma_0^2 > \sigma_z^2 = 1$, we show[9] in Appendix A.6 that asymptotically, $\overline{\mathcal{J}}_2 = 0$, and that this scheme performs better than VQ.

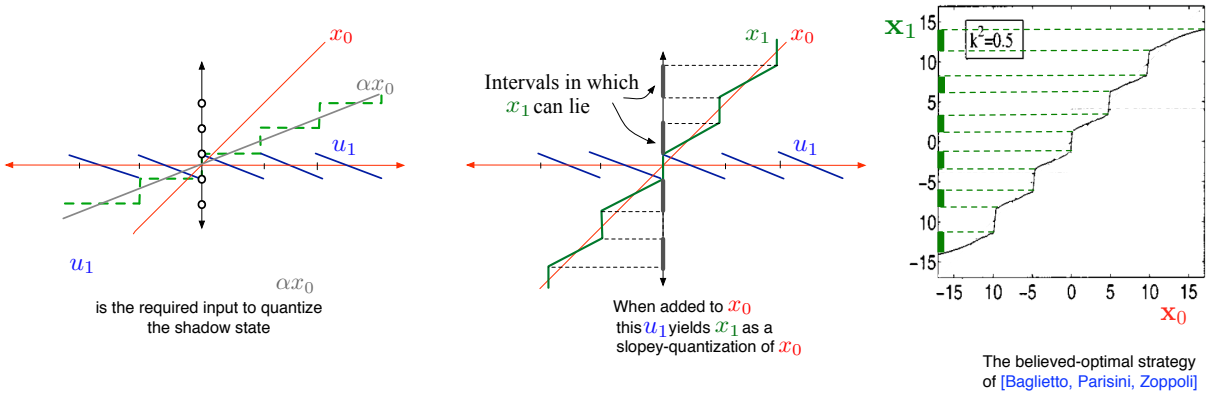


Figure 4.11: If the sequence of operations in the dirty-paper-coding strategy shown in Fig. 4.10 are followed for the scalar case, the resulting strategies look exactly like the slopey-quantization strategies of [25–27, 76].

For $\alpha \neq 1$, the transmitter does not drive the state all the way to a quantization point. Instead, the state $\mathbf{x}_1^m = \mathbf{x}_0^m + \mathbf{u}_1^m$ is merely correlated with the quantization point, given by $\mathbf{v}^m = \mathbf{x}_0^m + \alpha\mathbf{u}_1^m$. With high probability, the second controller can decode the underlying quantization point, and using the two observations $\mathbf{y}^m = \mathbf{x}_0^m + \mathbf{u}_1^m + \mathbf{z}^m$ and $\mathbf{v}^m = \mathbf{x}_0^m + \alpha\mathbf{u}_1^m$, it can estimate $\mathbf{x}_1^m = \mathbf{x}_0^m + \mathbf{u}_1^m$. This scheme has $\overline{\mathcal{J}}_2 \neq 0$, but when k is moderate, the total cost can be lower than that for DPC with $\alpha = 1$. Appendix A.6 describes this strategy and analyzes its performance in detail. Fig. 4.11 shows that for $\alpha \neq 1$, the DPC scheme is conceptually similar to the "neural schemes" numerically explored in [25] in that they are "soft quantization" schemes that tolerate some residual cost at stage 2 in order to reduce the cost at stage 1. Minor further improvements can be obtained by using a combination scheme that divides its power into two parts: a linear part and a part dedicated to dirty-paper coding. The linear component is used first to reduce the variance in $\mathbf{x}_0^m$ by scaling it down in a manner reminiscent of state-masking [93]. The remaining power is used to dirty-paper code against the resulting reduced interference. Appendix A.6 provides the details of this combination strategy. As shown in Fig. 4.12, using the combination scheme, the value of μ_2 is 2.

This combination strategy is shown to be optimal in the limit of asymptotically zero-reconstruction error using an improved lower bound in the next section.

[9]Only an outline of the proof is included in this dissertation. The full proof appears in [56].

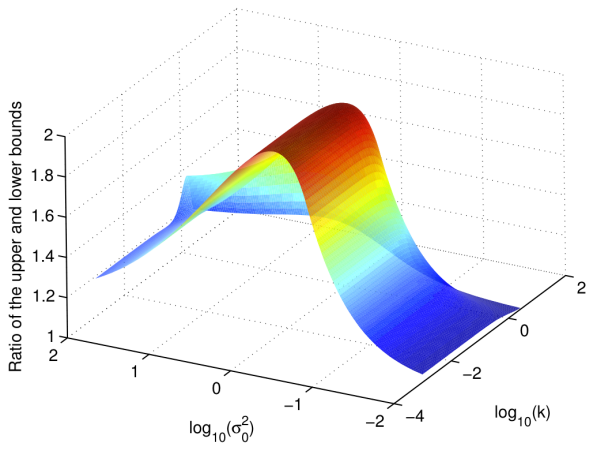


Figure 4.12: The plot shows the ratio of the performance of the combined DPC/linear scheme of Chapter 4.3.3 (analyzed in Appendix A.6) to the lower bound of (4.9) as σ_0 and k vary. Relative to Fig. 4.7, this new scheme has a maximum ratio of 2 attained on the ridge of $\sigma_0^2 = \frac{\sqrt{5}-1}{2}$ and small k. Also, the ridge along $k = 1$ is reduced as compared to Fig. (4.7). It is eliminated for small σ_0^2, while its asymptotic peak value of about 1.29 is attained at $k \approx 1.68$ and large σ_0^2.

4.3.4 Improved lower bounds, and improved ratios

The lower bound in Theorem 2 (and that in Corollary 1) allows for alignment of the input with the initial state when calculating the power input into the implicit channel. While this alignment maximizes the potential capacity of the channel, in reality this will also make the implicit source $\mathbf{X}_1^m$ Gaussian[10], the hardest source to estimate (in a rate-distortion sense [28]) with the worst possible (*i.e.* largest) variance. What this lower bound is ignoring is the fact that any correlation between $\mathbf{X}_0^m$ and $\mathbf{U}_1^m$ induces a different distribution (in particular, a different variance) on $\mathbf{X}_1^m$. We exploit this fact to obtain a tighter bound in this section. This improved bound shows that (Appendix A.8 in Appendix A.6) the our upper and lower bounds are tight in the limit of asymptotically zero-reconstruction error: the combination-strategy of previous section requires only the absolute minimum power required to attain perfect reconstruction.

Theorem 6. *For the vector Witsenhausen problem with* $\mathbb{E}\left[\|\mathbf{U}_1^m\|^2\right] \leq mP$*, the following is*

[10] A complete alignment corresponds to a scalar strategy where the input amplifies the initial state. This obviously retains the Gaussianity of the state as well.

a lower bound on the $MMSE$ in the estimation of $\mathbf{X}_1^m$.

$$MMSE \geq \inf_{\sigma_{X_0,U_1}} \sup_{\gamma>0} \frac{1}{\gamma^2} \left(\left(\sqrt{\frac{\sigma_0^2}{1+\sigma_0^2+P+2\sigma_{X_0,U_1}}} - \sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2 P-2\gamma(1-\gamma)\sigma_{X_0,U_1}} \right)^+ \right)^2 .$$

where $\sigma_{X_0,U_1} \in [-\sigma_0\sqrt{P}, \sigma_0\sqrt{P}]$. Further, the required power predicted by this lower bound turns out to be achievable in the limit of asymptotically zero reconstruction error.

Proof. See Appendix A.7. The tightness in the limit of asymptotically zero-error is shown in Appendix A.8. □

It is insightful to see how the lower bound in Theorem 6 is an improvement over that in Corollary 1. The lower bound in Corollary 1 is

$$MMSE \geq \left(\left(\sqrt{\frac{\sigma_0^2}{\sigma_0^2+P+2\sigma_0\sqrt{P}+1}} - \sqrt{P} \right)^+ \right)^2 , \tag{4.18}$$

which again holds for all m. Because any γ provides a valid lower bound in Theorem 6, choosing $\gamma = 1$ in Theorem 6 provides the following (loosened) bound,

$$MMSE \geq \inf_{|\sigma_{X_0,U_1}| \leq \sigma_0\sqrt{P}} \left(\left(\sqrt{\frac{\sigma_0^2}{\sigma_0^2+P+2\sigma_{X_0,U_1}+1}} - \sqrt{P} \right)^+ \right)^2 , \tag{4.19}$$

which is minimized for $\sigma_{X_0,U_1} = \sigma_0\sqrt{P}$. This immediately yields the lower bound (4.18) of Corollary 1.

Improved ratios, and a discussion of approximate optimality

Fig. 4.13 shows that asymptotically, the ratio of upper and new lower bounds (from Theorem 6) on the total weighted cost is bounded by 1.3, an improvement over the ratio of 2 obtained with the lower bound of Corollary 1. Comparing Fig. 4.7 and Fig. 4.13, the ridge of ratio 2 along $\sigma_0^2 = \frac{\sqrt{5}-1}{2}$ present in Fig. 4.7 does not exist anymore with the new lower bound. This is because the small-k regime corresponds to target $MMSE$s close to zero – where the new lower bound is tight. This point is further elucidated in Fig. 4.14(a). Also shown in Fig. 4.14(b) is the lack of tightness in the bounds at small P. The figure explains how this looseness results in the ridge along $k \approx 1.67$ still surviving in the new ratio plot.

Fig. 4.15 shows the ratio of upper and lower bounds on $MMSE$ versus P and σ_0. This figure brings out an important aspect of approximate-optimality results: while approximate-optimality may hold in one formulation of the problem (namely, minimizing weighted sum-cost), it may not hold in another equivalent formulation (namely, characterizing the optimal

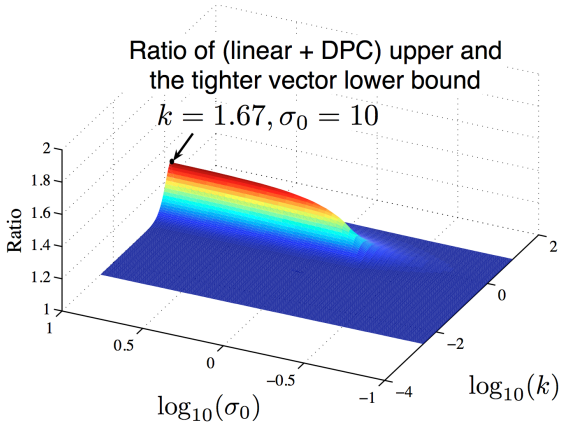


Figure 4.13: The ratio of upper and lower bounds on the total asymptotic cost for the vector Witsenhausen counterexample with the improved lower bound of Theorem 6. As compared to the maximum ratio of 2 using the lower bound of Corollary 1 (in Fig. 4.12), the ratio here is smaller than 1.3. Further, an infinitely long ridge along $\sigma_0^2 = \frac{\sqrt{5}-1}{2}$ and small k that is present in Fig. 4.12 is no longer present here. This is a consequence of the tightness lower bound at $MMSE = 0$, and hence for small k. A ridge remains along $k \approx 1.67$ ($\log_{10}(k) \approx 0.22$) and large σ_0, and this can be understood by observing Fig. 4.14 for $\sigma_0 = 10$.

tradeoff between P and $MMSE$)[11]. Thus, while Fig. 4.13 shows that the optimal total cost can be characterized to within a constant factor, Fig. 4.15(a) shows that the ratio of upper and lower bounds on $MMSE$ versus P and σ_0 diverges to infinity. Our improved lower bound in this section rectifies the problem: with the new bound of Theorem 6, the ratio is bounded by a factor of 1.5 (Fig. 4.15, right). This is again a reflection of the tightness of the bound at small $MMSE$.

However, a flipped perspective shown in Fig. 4.16 shows that the tradeoff curve is not yet completely understood. In this figure, we compute the ratio of upper and lower bounds on the required *power* to attain a specified $MMSE$. The ratio diverges to infinity along the path $MMSE = \frac{\sigma_0^2}{\sigma_0^2+1}$. This path is precisely the path corresponding to zero-input-power. Thus the question we do not completely understand is: how low a power is required when $MMSE$ is so bad that it is close to its maximum?

[11] As noted earlier, this phenomena is similar to approximation-algorithms in complexity theory where having an approximation algorithm for an NP-complete problem does not necessarily lead to an approximation algorithm for another, even though all NP-complete problems are computationally equivalent when solving exactly.

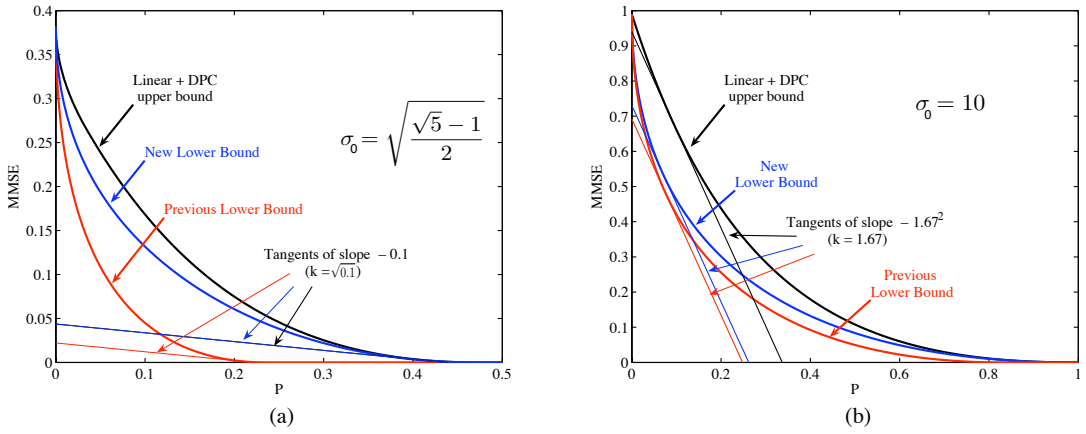


Figure 4.14: Upper and lower bounds on asymptotic $MMSE$ vs P for $\sigma_0 = \sqrt{\frac{\sqrt{5}-1}{2}}$ (square-root of the Golden ratio; Fig. (a)) and $\sigma_0 = 10$ (b) for zero-rate (the vector Witsenhausen counterexample). Tangents are drawn to evaluate the total cost for $k = \sqrt{0.1}$ for $\sigma_0 = \sqrt{\frac{\sqrt{5}-1}{2}}$, and for $k = 1.67$ for $\sigma_0 = 10$ (slope $= -k^2$). The intercept on the $MMSE$ axis of the tangent provides the respective bound on the total cost. The tangents to the upper bound and the new lower bound almost coincide for small values of k. At $k \approx 1.67$ and $\sigma_0 = 10$, however, this bound is not significantly better than that in Corollary 1 and hence the ridge along $k \approx 1.67$ remains in the new ratio plot in Fig. 4.13.

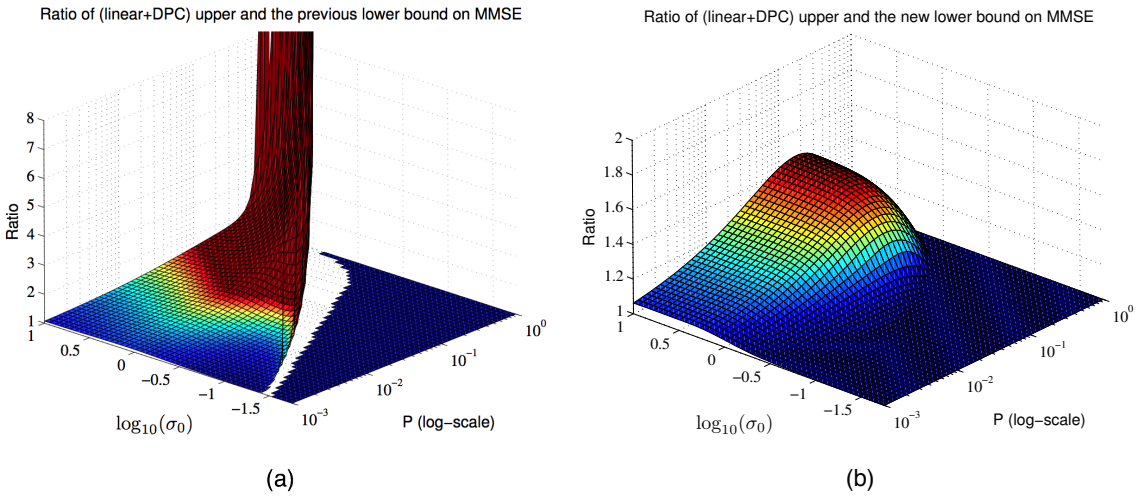


Figure 4.15: Ratio of upper and lower bounds on $MMSE$ vs P and σ_0. Whereas the ratio diverges to infinity in (a) with the lower bound of Corollary 1, it is bounded in (b) by 1.5 for the new bound. This is a consequence of the improved tightness of the new bound at small $MMSE$.

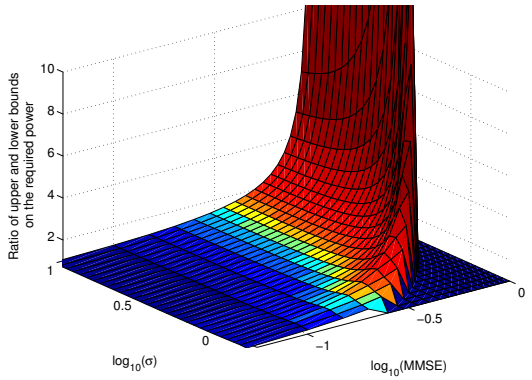


Figure 4.16: Ratio of upper and lower bounds on P vs $MMSE$ and σ_0. Interestingly, the ratio diverges to infinity as $\sigma_0 \to \infty$ along the path where P is close to zero (corresponding to $MMSE = \frac{\sigma_0^2}{\sigma_0^2+1}$).

4.4 Step 4: The Gaussian counterexample: finite number of dimensions (including the original scalar counterexample)

Now that we have approximate-optimality results for the asymptotically infinite-length version of the counterexample, can we use these to understand the original scalar counterexample? Using an asymptotic analysis to obtain results for finite-lengths is often a standard procedure in the theory of large-deviations [103]. Even in information theory, Shannon first addressed an asymptotic formulation of capacity, before dealing with error probability at finite-lengths[12] [104]. Although Shannon's bounds in [104] were derived for the power-constrained AWGN channel, the approach has been generalized and refined. Most of these bounds characterize the exponential rate of decay of error-probability with block-length. Recently, Polyanskiy, Poor, and Verdu [105, 106] use a central limit theorem-based approach to find bounds on the gap from capacity as a function of error probability and block-length based on a "dispersion" term. This yields fairly tight bounds on the error probability for what are traditionally considered small block-lengths (on the order of a hundred).

The challenge we face is two fold. The first challenge is obvious: we require results for the tiniest of block-lengths: the scalar case. Second, the bounds we require are bounds on a symbol-by-symbol distortion metric (the $MMSE$), and not a block-metric such as the block error probability. Most of the literature in information theory focuses on block-error

[12]Indeed, our step of addressing the asymptotic limit of the counterexample was very much inspired from Shannon's.

probability. Our results in [82, 107] for understanding the tradeoff of the size of the decoding neighborhood with the *bit*-error probability and the gap from capacity helps us in developing this understanding.

This section develops the theory that addresses these challenges. We first needs some definitions in order to provide the quantization strategy at finite dimensions.

Notation and definitions

Vectors are denoted in bold font, random variables in upper case, and their realizations in lower case. We use $A \perp\!\!\!\perp B$ to imply that the random variables A and B are independent. $\mathcal{B}$ is used to denote the unit ball in L_2-norm in $\mathbb{R}^m$.

Definition 1 (Packing and packing radius)**.** *Given an m-dimensional lattice Λ and a radius r, the set $\Lambda + r\mathcal{B} = \{\mathbf{x}^m + r\mathbf{y}^m : x \in \Lambda, \mathbf{y}^m \in \mathcal{B}\}$ is a packing of Euclidean m-space if for all points $\mathbf{x}^m, \mathbf{y}^m \in \Lambda$, $(\mathbf{x}^m + r\mathcal{B}) \bigcap (\mathbf{y}^m + r\mathcal{B}) = \emptyset$. The packing radius r_p is defined as $r_p := \sup\{r : \Lambda + r\mathcal{B}$ is a packing$\}$.*

Definition 2 (Covering and covering radius)**.** *Given an m-dimensional lattice Λ and a radius r, the set $\Lambda + r\mathcal{B}$ is a covering of Euclidean m-space if $\mathbb{R}^m \subset \Lambda + r\mathcal{B}$. The covering radius r_c is defined as $r_c := \inf\{r : \Lambda + r\mathcal{B}$ is a covering$\}$.*

Definition 3 (Packing-covering ratio)**.** *The packing-covering ratio (denoted by ξ) of a lattice Λ is the ratio of its covering radius to its packing radius, $\xi = \frac{r_c}{r_p}$.*

For this section, we denote the pdf of the elements of noise $\mathbf{Z}^m$ by $f_Z(\cdot)$. In our proof techniques, we also consider a hypothetical observation noise $\mathbf{Z}_G^m \sim \mathcal{N}(0, \sigma_G^2)$ with variance $\sigma_G^2 \geq 1$. The pdf of this test noise is denoted by $f_G(\cdot)$. We use $\psi(m, r)$ to denote $\Pr(\|\mathbf{Z}^m\| \geq r)$ for $\mathbf{Z}^m \sim \mathcal{N}(0, \mathbb{I})$. Subscripts in expectation expressions denote the random variable being averaged over (e.g. $\mathbb{E}_{\mathbf{X}_0^m, \mathbf{Z}_G^m}[\cdot]$ denotes averaging over the initial state $\mathbf{X}_0^m$ and the test noise $\mathbf{Z}_G^m$).

4.4.1 Upper and lower bounds on costs

An upper bound on costs

What will be a good strategy for a vector extension, say of dimension 2? One can break the problem down into two scalar problems, and operate separately on the two elements of the vector. But we know from information theory that strategies that perform vector operations commonly outperform strategies that treat vectors merely as a collection of scalars. Is there a possible improvement over a simple scalar quantization strategy?

The use of scalar quantization strategy in a problem of dimension 2 amounts to quantizing to a *grid lattice* shown in Fig. 4.17. The "error probability" of decoding to a wrong

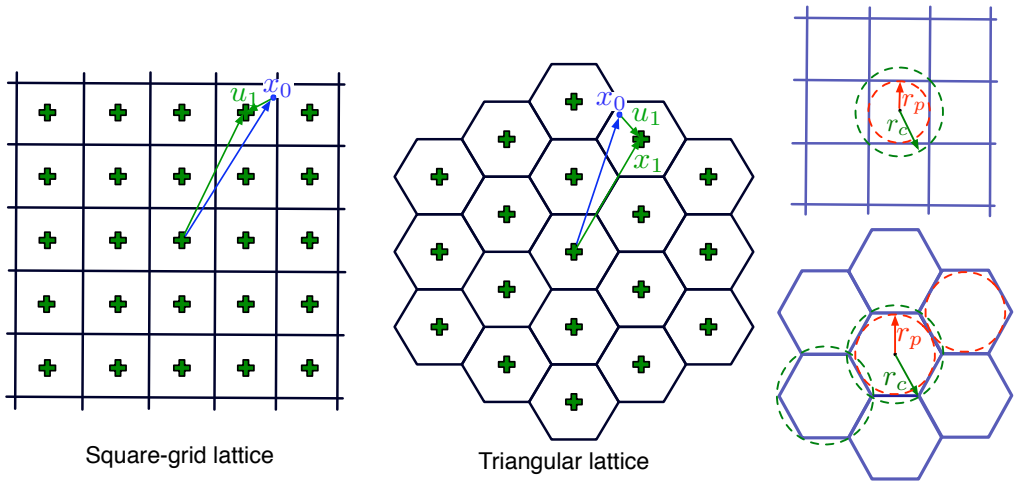


Figure 4.17: The most obvious generalization of the scalar-quantization strategy to the vector case is a square-grid, that quantizes each dimension of the vector problem separately. A triangular lattice attains a smaller error probability for the same average power costs because of an improved packing-covering ratio. On the right side is an illustration of packing radius r_p and covering radius r_c.

quantization point in either dimension is governed by the nearest lattice point, in Euclidean sense, that the noise can push the x_1 quantization point to. Euclidean distance between quantization points thus emerges as a proxy for the error probability. One can reduce the cost at the first stage by using an improved lattice, for example, a triangular lattice[13] (shown in Figure 4.17) while keeping the minimum distance between the lattice points the same. Since the error probability is dominated by this minimum distance, the second stage cost is also dominated by the term that corresponds to the error of decoding to the nearest neighbors. Thus one needs a lattice that performs a good *packing*, keeping the nearest lattice points far enough for small second stage costs, as well as a good *covering* so that no point in the space is too far – yielding small first stage-costs. These lattices thus correspond to ones that have a *good packing-covering ratio* — the ratio of covering radii to the packing radius of the lattice.

This lattice-quantization strategy yields the following upper bound on the cost for $W(m, k^2, \sigma_0^2)$, the dimension-m vector Witsenhausen problem.

Theorem 7. *Using a lattice-based strategy (as described above) for $W(m, k^2, \sigma_0^2)$ with r_c and r_p the covering and the packing radius for the lattice, the total average cost is upper bounded*

[13]Often also called 'hexagonal' lattice for its hexagonal Voronoi regions.

by

$$\overline{\mathcal{J}}^{(\gamma)}(m,k^2,\sigma_0^2) \leq \inf_{P\geq 0} k^2 P + \left(\sqrt{\psi(m+2,r_p)} + \sqrt{\frac{P}{\xi^2}}\sqrt{\psi(m,r_p)}\right)^2,$$

where $\xi = \frac{r_c}{r_p}$ *is the packing-covering ratio for the lattice, and* $\psi(m,r) = \Pr(\|\mathbf{Z}^m\| \geq r)$*. The following looser bound also holds*

$$\overline{\mathcal{J}}^{(\gamma)}(m,k^2,\sigma_0^2) \leq \inf_{P>\xi^2} k^2 P + \left(1+\sqrt{\frac{P}{\xi^2}}\right)^2 e^{-\frac{mP}{2\xi^2}+\frac{m+2}{2}\left(1+\ln\left(\frac{P}{\xi^2}\right)\right)}.$$

Remark: The latter loose bound is useful for analytical manipulations when proving explicit bounds on the ratio of the upper and lower bounds in Chapter 4.4.2.

Proof. Note that because Λ has a covering radius of r_c, $\|\mathbf{x}_1^m - \mathbf{x}_0^m\|^2 \leq r_c^2$. Thus the first stage cost is bounded above by $\frac{1}{m}k^2 r_c^2$. A tighter bound can be provided for a specific lattice and finite m (for example, for $m=1$, the first stage cost is approximately $k^2\frac{r_c^2}{3}$ if $r_c^2 \ll \sigma_0^2$ because the distribution of $\mathbf{x}_0^m$ conditioned on it lying in any of the quantization bins is approximately uniform at least for the most likely bins). For the second stage, observe that

$$\mathbb{E}_{\mathbf{X}_1^m,\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\right] = \mathbb{E}_{\mathbf{X}_1^m}\left[\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2|\mathbf{X}_1^m\right]\right]. \tag{4.20}$$

Denote by $\mathcal{E}_m$ the event $\{\|\mathbf{Z}^m\|^2 \geq r_p^2\}$. Observe that under the event $\mathcal{E}_m^c$, $\widehat{\mathbf{X}}_1^m = \mathbf{X}_1^m$, resulting in a zero second-stage cost. Thus,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2|\mathbf{X}_1^m\right] &= \mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\mathbf{1}_{\{\mathcal{E}_m\}}|\mathbf{X}_1^m\right] + \mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\mathbf{1}_{\{\mathcal{E}_m^c\}}|\mathbf{X}_1^m\right] \\ &= \mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\mathbf{1}_{\{\mathcal{E}_m\}}|\mathbf{X}_1^m\right].\end{aligned}$$

We now bound the squared-error under the error event $\mathcal{E}_m$, when either $\mathbf{x}_1^m$ is decoded erroneously, or there is a decoding failure. If $\mathbf{x}_1^m$ is decoded erroneously to a lattice point $\widetilde{\mathbf{x}}_1^m \neq \mathbf{x}_1^m$, the squared-error can be bounded as follows

$$\|\mathbf{x}_1^m - \widetilde{\mathbf{x}}_1^m\|^2 = \|\mathbf{x}_1^m - \mathbf{y}_2^m + \mathbf{y}_2^m - \widetilde{\mathbf{x}}_1^m\|^2 \leq \left(\|\mathbf{x}_1^m - \mathbf{y}_2^m\| + \|\mathbf{y}_2^m - \widetilde{\mathbf{x}}_1^m\|\right)^2 \leq \left(\|\mathbf{z}^m\| + r_p\right)^2.$$

If $\mathbf{x}_1^m$ is decoded as $\mathbf{y}_2^m$, the squared-error is simply $\|\mathbf{z}^m\|^2$, which we also upper bound by $\left(\|\mathbf{z}^m\| + r_p\right)^2$. Thus, under event $\mathcal{E}_m$, the squared error $\|\mathbf{x}_1^m - \widehat{\mathbf{x}}_1^m\|^2$ is bounded above by $\left(\|\mathbf{z}^m\| + r_p\right)^2$, and hence

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2|\mathbf{X}_1^m\right] &\leq \mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2\mathbf{1}_{\{\mathcal{E}_m\}}|\mathbf{X}_1^m\right] \\ &\overset{(a)}{=} \mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2\mathbf{1}_{\{\mathcal{E}_m\}}\right],\end{aligned} \tag{4.21}$$

where (a) uses the fact that the pair $(\mathbf{Z}^m, \mathbf{1}_{\{\mathcal{E}_m\}})$ is independent of $\mathbf{X}_1^m$. Now, let $P = \frac{r_c^2}{m}$, so that the first stage cost is at most $k^2 P$. The following lemma helps us derive the upper bound.

Lemma 2. *For a given lattice with $r_p^2 = \frac{r_c^2}{\xi^2} = \frac{mP}{\xi^2}$, the following bound holds*

$$\frac{1}{m}\mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] \leq \left(\sqrt{\psi(m+2, r_p)} + \sqrt{\frac{P}{\xi^2}}\sqrt{\psi(m, r_p)}\right)^2.$$

The following (looser) bound also holds as long as $P > \xi^2$,

$$\frac{1}{m}\mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] \leq \left(1 + \sqrt{\frac{P}{\xi^2}}\right)^2 e^{-\frac{mP}{2\xi^2} + \frac{m+2}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right)}.$$

Proof. See Appendix A.9. □

The theorem now follows from (4.20), (4.21) and Lemma 2. □

Lower bound on costs

Observe that the lower bound expression of Corollary 1 is the same for all vector lengths. In the following, large-deviation arguments [108, 109] (called sphere-packing style arguments in information theory for historical reasons[14]) are extended following [107, 110, 111] to a joint source-channel setting where the distortion measure is unbounded.

The main technical difficulty is posed by the unbounded support of the Gaussian distribution. Because the lower bounds discussed so far are valid asymptotically, they implicitly assume that the noise behavior is within a bounded sphere. In the scalar case, the noise can be extremely large, even though there is a small probability associated with it. How can we account for this? We use the technique of change-of-measure from large-deviation theory [103]. Heuristically, the idea is this: an atypically large behavior of (Gaussian) noise is typical for another (Gaussian) distribution (of larger variance). Using this, there can be a probability associated with an atypical behavior. Conditioned on this atypical behavior, a lower bound on the distortion is known from Theorem 2 (by bringing the new noise variance out explicitly). Multiplying this lower bound with the associated probability will bring us to an actual lower bound on the distortion. The resulting bounds are tighter than those in Corollary 1 and depend explicitly on the vector length m.

[14] The first bounds in this style were derived by Shannon in [104] by finding the number of spheres of a given size that can be packed in a given volume assuming a maximum allowed intersection between the spheres. This argument was used in Park, Grover and Sahai [102] to obtain the first constant-factor optimality result on the counterexample, albeit the constant factor there was quite large.

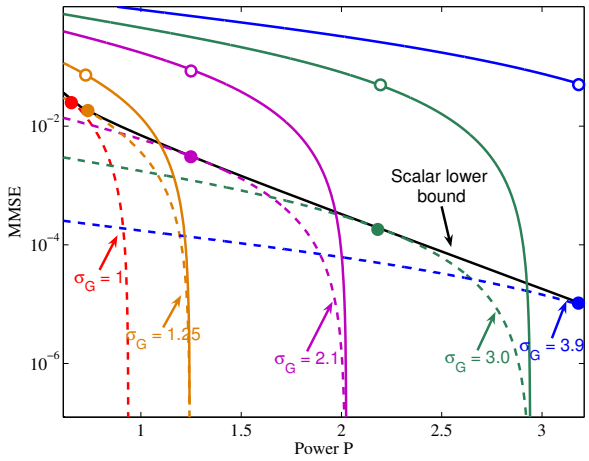


Figure 4.18: A pictorial representation of the proof for the lower bound assuming $\sigma_0^2 = 30$. The solid curves show the vector lower bound of Corollary 1 for various values of observation noise variances, denoted by σ_G^2. Conceptually, multiplying these curves by the probability of that channel behavior yields the shadow curves for the particular σ_G^2, shown by dashed curves. The scalar lower bound is then obtained by taking the maximum of these shadow curves. The circles at points along the scalar bound curve indicate the optimizing value of σ_G for obtaining that point on the bound.

Theorem 8. *For $W(m,k^2,\sigma_0^2)$, if for a strategy $\gamma(\cdot)$ the average power $\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m}[\|\mathbf{U}_1^m\|^2] = P$, the following lower bound holds on the second stage cost for any choice of $\sigma_G^2 \geq 1$ and $L > 0$*

$$\overline{\mathcal{J}}_2^{(\gamma)}(m,k^2,\sigma_0^2) \geq \eta(P,\sigma_0^2,\sigma_G^2,L).$$

where

$$\eta(P,\sigma_0^2,\sigma_G^2,L) = \frac{\sigma_G^m}{c_m(L)}\exp\left(-\frac{mL^2(\sigma_G^2-1)}{2}\right)\left(\left(\sqrt{\kappa_2(P,\sigma_0^2,\sigma_G^2,L)}-\sqrt{P}\right)^+\right)^2,$$

where $\kappa_2(P,\sigma_0^2,\sigma_G^2,L) :=$

$$\frac{\sigma_0^2\sigma_G^2}{c_m^{\frac{2}{m}}(L)e^{1-d_m(L)}\left(\left(\sigma_0+\sqrt{P}\right)^2+d_m(L)\sigma_G^2\right)},$$

$c_m(L) := \frac{1}{\Pr(\|\mathbf{Z}^m\|^2 \leq mL^2)} = (1-\psi(m,L\sqrt{m}))^{-1}$, $d_m(L) := \frac{\Pr(\|\mathbf{Z}^{m+2}\|^2 \leq mL^2)}{\Pr(\|\mathbf{Z}^m\|^2 \leq mL^2)} = \frac{1-\psi(m+2,L\sqrt{m})}{1-\psi(m,L\sqrt{m})}$, $0 < d_m(L) < 1$, *and* $\psi(m,r) = \Pr(\|\mathbf{Z}^m\| \geq r)$. *Thus the following lower bound holds on the total cost*

$$\overline{\mathcal{J}}_{\min}(m,k^2,\sigma_0^2) \geq \inf_{P\geq 0} k^2P + \eta(P,\sigma_0^2,\sigma_G^2,L), \tag{4.22}$$

for any choice of $\sigma_G^2 \geq 1$ and $L > 0$ (the choice can depend on P). Further, these bounds are at least as tight as those of Corollary 1 for all values of k and σ_0^2.

Proof. From Corollary 1, for a given P, a lower bound on the average second stage cost is $\left(\left(\sqrt{\kappa}-\sqrt{P}\right)^+\right)^2$. We derive another lower bound that is equal to the expression for $\eta(P,\sigma_0^2,\sigma_G^2,L)$. The high-level intuition behind this lower bound is presented in Fig. 4.18. Define $\mathcal{S}_L^G := \{\mathbf{z}^m : \|\mathbf{z}^m\|^2 \leq mL^2\sigma_G^2\}$ and use subscripts to denote which probability model is being used for the second stage observation noise. Z denotes white Gaussian of variance 1 while G denotes white Gaussian of variance $\sigma_G^2 \geq 1$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}^m)\right] &= \int_{\mathbf{z}^m}\int_{\mathbf{x}_0^m} J_2^{(\gamma)}(\mathbf{x}_0^m,\mathbf{z}^m)f_0(\mathbf{x}_0^m)f_Z(\mathbf{z}^m)d\mathbf{x}_0^m d\mathbf{z}^m \\
&\geq \int_{\mathbf{z}^m\in\mathcal{S}_L^G}\left(\int_{\mathbf{x}_0^m} J_2^{(\gamma)}(\mathbf{x}_0^m,\mathbf{z}^m)f_0(\mathbf{x}_0^m)d\mathbf{x}_0^m\right)f_Z(\mathbf{z}^m)d\mathbf{z}^m \\
&= \int_{\mathbf{z}^m\in\mathcal{S}_L^G}\left(\int_{\mathbf{x}_0^m} J_2^{(\gamma)}(\mathbf{x}_0^m,\mathbf{z}^m)f_0(\mathbf{x}_0^m)d\mathbf{x}_0^m\right)\frac{f_Z(\mathbf{z}^m)}{f_G(\mathbf{z}^m)}f_G(\mathbf{z}^m)d\mathbf{z}^m
\end{aligned} \tag{4.23}$$

The ratio of the two probability density functions is given by

$$\frac{f_Z(\mathbf{z}^m)}{f_G(\mathbf{z}^m)} = \frac{e^{-\frac{\|\mathbf{z}^m\|^2}{2}}}{\left(\sqrt{2\pi}\right)^m}\frac{\left(\sqrt{2\pi\sigma_G^2}\right)^m}{e^{-\frac{\|\mathbf{z}^m\|^2}{2\sigma_G^2}}} = \sigma_G^m e^{-\frac{\|\mathbf{z}^m\|^2}{2}\left(1-\frac{1}{\sigma_G^2}\right)}.$$

Observe that $\mathbf{z}^m \in \mathcal{S}_L^G$, $\|\mathbf{z}^m\|^2 \leq mL^2\sigma_G^2$. Using $\sigma_G^2 \geq 1$, we obtain

$$\frac{f_Z(\mathbf{z}^m)}{f_G(\mathbf{z}^m)} \geq \sigma_G^m e^{-\frac{mL^2\sigma_G^2}{2}\left(1-\frac{1}{\sigma_G^2}\right)} = \sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}}. \tag{4.24}$$

Using (4.23) and (4.24),

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}^m)\right] &\geq \sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}} \int_{\mathbf{z}^m\in\mathcal{S}_L^G}\left(\int_{\mathbf{x}_0^m} J_2^{(\gamma)}(\mathbf{x}_0^m,\mathbf{z}^m) f_0(\mathbf{x}_0^m)d\mathbf{x}_0^m\right) f_G(\mathbf{z}^m)d\mathbf{z}^m \\
&= \sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}} \mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)\mathbf{1}_{\left\{\mathbf{Z}_G^m\in\mathcal{S}_L^G\right\}}\right] \\
&= \sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}} \mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]\Pr(\mathbf{Z}_G^m\in\mathcal{S}_L^G) \qquad (4.25)
\end{aligned}$$

Analyzing the probability term in (4.25),

$$\begin{aligned}
\Pr(\mathbf{Z}_G^m\in\mathcal{S}_L^G) &= \Pr\left(\|\mathbf{Z}_G^m\|^2 \leq mL^2\sigma_G^2\right) = \Pr\left(\left(\frac{\|\mathbf{Z}_G^m\|}{\sigma_G}\right)^2 \leq mL^2\right) \\
&= 1-\Pr\left(\left(\frac{\|\mathbf{Z}_G^m\|}{\sigma_G}\right)^2 > mL^2\right) = 1-\psi(m,L\sqrt{m}) = \frac{1}{c_m(L)}, \qquad (4.26)
\end{aligned}$$

because $\frac{\mathbf{Z}_G^m}{\sigma_G}\sim\mathcal{N}(0,\mathbb{I}_m)$. From (4.25) and (4.26),

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}^m)\right] &\geq \sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}} \mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right](1-\psi(m,L\sqrt{m})) \\
&= \frac{\sigma_G^m e^{-\frac{mL^2(\sigma_G^2-1)}{2}}}{c_m(L)} \mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]. \qquad (4.27)
\end{aligned}$$

We now need the following lemma, which connects the new finite-dimensional lower bound to the infinite-dimensional lower bound of Corollary 1.

Lemma 3.

$$\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right] \geq \left(\left(\sqrt{\kappa_2(P,\sigma_0^2,\sigma_G^2,L)}-\sqrt{P}\right)^+\right)^2,$$

for any $L>0$.

Proof. See Appendix A.10. □

The lower bound on the total average cost now follows from (4.27) and Lemma 3. We now verify that $d_m(L) \in (0,1)$. That $d_m(L) > 0$ is clear from definition. $d_m(L) < 1$ because $\{\mathbf{z}^{m+2} : \|\mathbf{z}^{m+2}\|^2 \leq mL^2\sigma_G^2\} \subset \{\mathbf{z}^{m+2} : \|\mathbf{z}^m\|^2 \leq mL^2\sigma_G^2\}$, *i.e.*, a sphere sits inside a cylinder.

Finally we verify that this new lower bound is at least as tight as the one in Corollary 1. Choosing $\sigma_G^2 = 1$ in the expression for $\eta(P, \sigma_0^2, \sigma_G^2, L)$,

$$\eta(P, \sigma_0^2, \sigma_G^2, L) \geq \sup_{L>0} \frac{1}{c_m(L)} \left(\left(\sqrt{\kappa_2(P, \sigma_0^2, 1, L)} - \sqrt{P} \right)^+ \right)^2 .$$

Now notice that $c_m(L)$ and $d_m(L)$ converge to 1 as $L \to \infty$. Thus $\kappa_2(P, \sigma_0^2, 1, L) \overset{L\to\infty}{\longrightarrow} \kappa(P, \sigma_0^2)$ and therefore, $\eta(P, \sigma_0^2, \sigma_G^2, L)$ is lower bounded by $\left(\left(\sqrt{\kappa} - \sqrt{P} \right)^+ \right)^2$, the lower bound in Corollary 1. □

4.4.2 Combination of linear and lattice-based strategies attain within a constant factor of the optimal cost

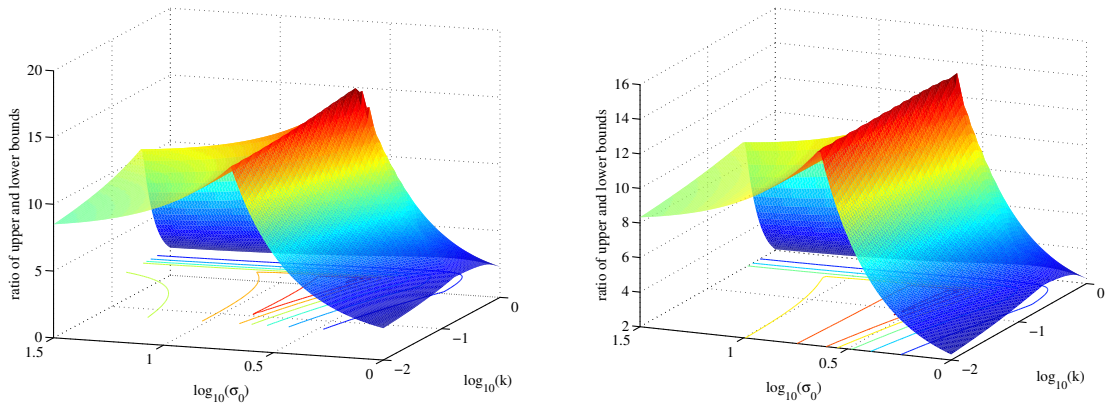


Figure 4.19: The ratio of the upper and the lower bounds for the scalar Witsenhausen problem (left), and the 2-D Witsenhausen problem (right, using triangular lattice of $\xi = \frac{2}{\sqrt{3}}$) for a range of values of k and σ_0. The ratio is bounded above by 17 for the scalar problem, and by 14.75 for the 2-D problem.

Theorem 9 (Constant-factor optimality)**.** *The costs for* $W(m, k^2, \sigma_0^2)$ *are bounded as follows*

$$\inf_{P\geq 0} \sup_{\sigma_G^2 \geq 1, L>0} k^2 P + \eta(P, \sigma_0^2, \sigma_G^2, L) \leq \overline{\mathcal{J}}_{min}(m, k^2, \sigma_0^2) \leq \mu \left(\inf_{P\geq 0} \sup_{\sigma_G^2 \geq 1, L>0} k^2 P + \eta(P, \sigma_0^2, \sigma_G^2, L) \right),$$

where $\mu = 100\xi^2$, ξ is the packing-covering ratio of any lattice in $\mathbb{R}^m$, and $\eta(\cdot)$ is as defined in Theorem 8. For any m, $\mu < 1600$. Further, depending on the (m, k^2, σ_0^2) values, the upper bound can be attained by lattice-based quantization strategies or linear strategies. For $m = 1$, a numerical calculation (MATLAB code available at [112]) shows that $\mu < 8$ (see Fig. 4.20).

Proof. See Appendix A.11. □

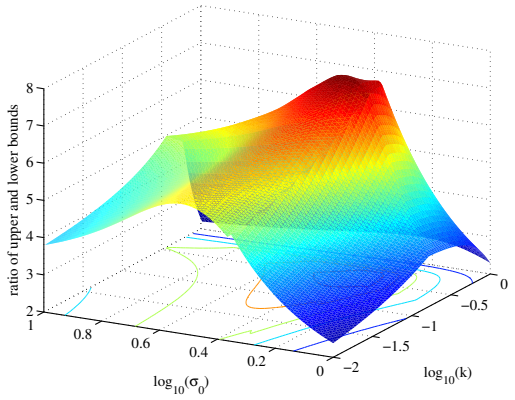


Figure 4.20: An exact calculation of the first and second stage costs yields an improved maximum ratio smaller than 8 for the scalar Witsenhausen problem.

Although the proof in Appendix A.11 succeeds in showing that the ratio is uniformly bounded by a constant, it is not very insightful and the constant is large. Of importance here is that such a constant exists. The value of the constant can now be evaluated numerically. Such a numerical evaluation (using Theorem 7 and 8 for upper and lower bounds respectively) shows that the ratio is smaller than 17 for $m = 1$ (see Fig. 4.19). A precise calculation of the cost of the quantization strategy improves the upper bound (by calculating the cost of the quantization strategy to a greater precision for $m = 1$) to yield a maximum ratio smaller than 8 (see Fig. 4.20). A simple grid lattice has a packing-covering ratio $\xi = \sqrt{m}$. Therefore, while the grid lattice has the best possible packing-covering ratio of 1 in the scalar case, it has a rather large packing covering ratio of $\sqrt{2}$ (≈ 1.41) for $m = 2$. On the other hand, a triangular lattice (for $m = 2$) has an improved packing-covering ratio of $\frac{2}{\sqrt{3}} \approx 1.15$. In contrast with $m = 1$, where the ratio of upper and lower bounds of Theorem 7 and 8 is approximately 17, a triangular lattice yields a ratio smaller than 14.75, despite having a larger packing-covering ratio. This is a consequence of the tightening of the sphere-packing lower bound (Theorem 8) as m gets large[15].

[15]Indeed, in the limit $m \to \infty$, the ratio of the asymptotic average costs attained by a vector-quantization strategy and the vector lower bound of Corollary 1 is bounded by 4.45.

Chapter 5

Beyond the counterexample: towards a theory of implicit communication

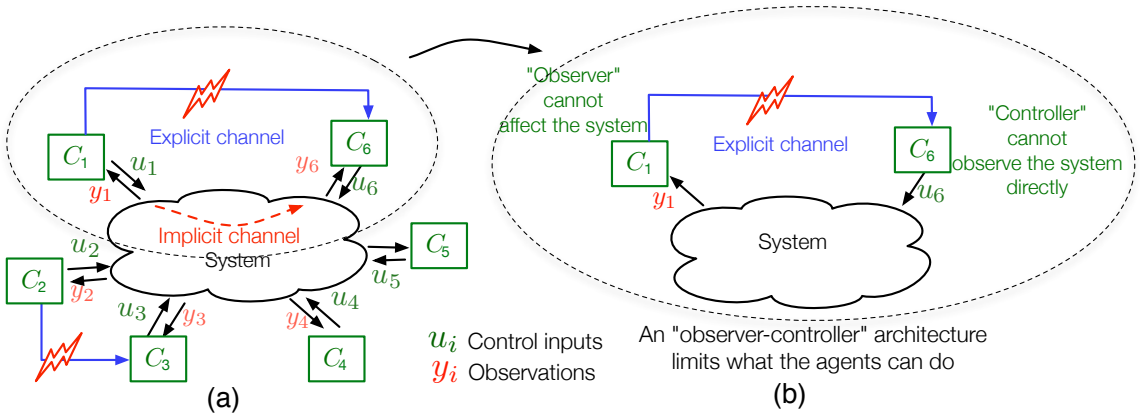


Figure 5.1: (a) A set of agents operating dynamically in order to attain a control goal. The controllers can communicate implicitly through the plant, and may also communicate to each other using external channels. We are interested in designing cost-efficient strategies for such a general system. (b) Many current formulations, which we call "observer-controller" formulations, limit the ability of the agents to either observing the system state, or taking actions to alter the state (using messages received across a channel), but not both.

One of our grand goals is to address the system shown in Fig. 5.1: how should we design cost-efficient strategies for such a general decentralized control system? Notice that the controllers can communicate through the implicit channel of the system as well as external channels that may connect them. This chapter aims at building and addressing toy problems that can help us design efficient strategies for such larger systems.

Most of the results in the existing literature (e.g. [34,42,43,84,86,89,113–118]) focus on an "observer-controller" architecture where the observer is connected to the controller through a

communication channel. These results and formulations are unsatisfactory for three reasons. First, they are designed to disallow implicit communication. This is done in two steps: the "observer" is stripped of its ability to modify the state (making the "source" unmodifiable), and the "controller" cannot observe the state directly, and only estimates it from the channel output. Second, the difficulty of the optimal cost framework forces the formulation away from that of minimizing costs to a coarse measure of attaining stability (see Chapter 3.4.3). Third, the overall flavor of results is negative: they suggest that even stabilizing a system across a channel can be really hard to accomplish! Our goal in this chapter is to show that an understanding of the counterexample that we developed in Chapter 4 can help make some progress in addressing these issues.

In Chapter 4, we provided an approximately-optimal solution to the Witsenhausen counterexample, and proposed a program of using semi-deterministic abstractions for gaining insights into provably-good strategy-design for more general problems of decentralized control. While the solutions obtained are not exactly optimal, they bring us closer to the problem of minimizing costs because they perform within a uniform constant factor of the minimum *possible* cost for all problem parameters. While the jump from a stability formulation to an approximate-optimality formulation is qualitative, arriving at the optimal cost from an approximately-optimal cost would be more of a quantitative improvement. These results thus also allow us to operate in a cost-framework, which is of much greater practical interest than the stability framework.

The question of interest in this chapter is: can we now investigate problems of control under communication constraints from a cost perspective? It was partly the lack of understanding of signaling[1] that forced us into stability formulations. If Witsenhausen's counterexample indeed distills some important aspects of signaling in decentralized control, and if our semi-deterministic abstractions indeed capture the essence of signaling within the counterexample, we should now be able to extend our proposed program to toy versions of the problem shown in Fig. 5.1. In this chapter, we use simplistic toy versions to show that this extension might indeed by possible. For each of these toy problems, we use the semi-deterministic model to gain insight into strategy design. To show that the model captures the most significant aspects (the "most significant bits") of the problem, we prove asymptotic-approximate-optimality (the counterpart of Step 3 in Chapter 4) of the strategy obtained from the semi-deterministic model for each of these problems.

We begin with a problem in which an external channel connects two agents who are jointly trying to force the system state close to zero (Chapter 5.1; see Fig. 5.2). As in other formulations of control under communication constraints, the first agent, the "observer," has perfect observations of the state. The observer wants to communicate the state to the second agent, the "controller," through an external channel (the controller has no direct observations of the state). The controller is allowed to use large control inputs to force the

[1]The other technical difficulty arises from the difficulty in understanding causality in information theory.

state close to zero. In a departure from most other models, we use an "enhanced" observer who can modify the state so that the possibility of implicit communication exists. Unlike for the counterexample, we assume that the controller still has no direct observations of the state. The role of control actions in the counterexample is often thought to be that of signaling (see, for example, [16]). If actions could be used to simply signal the state (very much akin to explicit communication), then in this problem the observer should not take any control action at all: there is no possibility of signaling through the plant! Instead, by this hypothesis, it should merely communicate its observations over the external channel (as well as possible). Our analysis shows that the controller *should* take an action: it should modify the state in order to simplify it, so that it can be estimated better across the channel. In particular, quantization-based strategies can be shown to be asymptotically approximately-optimal for this problem.

What if the controller in the problem of Chapter 5.1 is also enhanced by allowing it to see the state directly? This problem is addressed in Chapter 5.2. Using a semi-deterministic abstraction of the problem, we arrive at a strategy that essentially uses the plant along with the external channel for communication, treating it as a problem of parallel channels. The control input is used to simplify the source in order to communicate it over these parallel channels. These strategies are connected to the notion of "binning" in information theory, and they outperform the best known strategies for this problem (obtained in [22]) by a factor that can diverge to infinity.

In Chapter 5.3 we formulate a complementary problem to that in Chapter 5.2. Again, the first agent has complete observations of the (vector) state. However, there is no external channel connecting the agents. Further, the second agent does not observe some state dimensions at all (and only has partial observations of other state dimensions). The first agent is thus forced to signal the state in the hidden dimensions using the dimensions that are observed at the second agent. This problem also highlights the triple role of control actions, namely control, communication, and improving state estimability. For the Witsenhausen counterexample, as we noted in Chapter 1.5.4, the roles of communication and improving state estimability are aligned. But here control actions are forced to balance between all three roles.

So far, all the problems addressed in this dissertation have time horizon two. However, almost all realistic control problems have a larger time horizon, where controllers repeatedly act on the system as it evolves. Does our understanding of signaling extend to such problems? In Chapter 5.4 we address a problem of decentralized filtering where the time-horizon can be larger than 2. The first agent has perfect observations of an evolving system state, but it "actively" participates in helping the second agent estimate the state: it implicitly communicates the state through the plant. The problem turns out to be an extension of Witsenhausen's counterexample to multiple time-steps, and surprisingly, can be analyzed for a restricted parameter space using the results from the counterexample in a straightforward manner.

Does our understanding of signaling extend to problems beyond the LQG setup? At

one level, our problems in of uniform noise in Chapter 4.2 and problems with rate-limited external channels in Chapter 5.2 are not LQG. But a more stark example comes from agents in an economic system. In these systems, observations are often not noise limited, but instead they are limited by the bounded processing ability, or "bounded rationality" (as it is often called in game-theoretic literature [48]) of agents. The key difference of these problems is that even though the observations are partial, the agent has some freedom in choosing what the observations are. For instance, a common model of such agents assumes that they are finite-state machines [48]. Another recent model of Sims [51], called the "rational-inattention mode," assumes that there is a mutual-information constraint between what a controller observes and the actions that it takes. In Chapter 5.5, we consider a toy version of the problem of pricing by a seller in order to capture the attention of a rationally-inattentive consumer. Numerical studies (for a slightly different formulation) show that the chosen prices should occupy only a finite set of discrete points, rather than the entire real-line [53, 54]. We prove that for our closely-related problem, a discrete pricing strategy is indeed asymptotically-approximately optimal.

Finally, in Chapter 5.6, we consider a problem where the observations of all the controllers are noisy. The goal is to attain a deeper understanding of the goodness of approximately-optimal solutions. Many problems in the field of control under communication constraints, and also in this dissertation, assume that the first controller has perfect observations of the state. This leads to a convenient interpretation of the controllers as *encoders* and *decoders*. Are the suggested solutions robust to observation noise at the first controller? We use the semi-deterministic model to suggest that modifications of existing strategies should suffice for these new formulations as well. The claim is substantiated by showing that these modified strategies are asymptotically approximately-optimal. The problem also helps raise the question of when approximate-optimality captures the essence of the problem, which we discuss in Chapter 6.

For simplicity, we only analyze the asymptotic infinite-dimensional versions of these problems. As is standard in this dissertation, for each problem, we will provide an approximately optimal solution that attains within a constant factor of the optimal cost for all problem parameters (except for the problem in Chapter 5.4 where we provide approximately-optimal strategies for a large subset of the parameter space). It is not a given, however, that the solutions will be approximately-optimal at finite lengths as well. We make observations in this regard for each problem.

5.1 A problem of an implicit source with an explicit channel

In Witsenhausen's counterexample, the first controller injects power into the system to modify x_1 in order to communicate it to the second controller. We noted that the counterexample

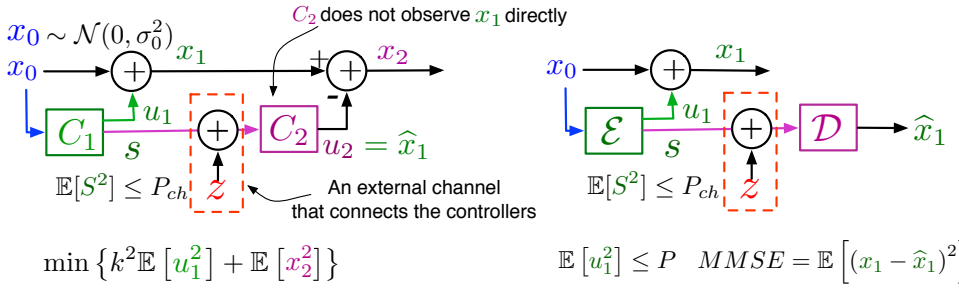


Figure 5.2: A problem of an implicit source, and an explicit channel: actions are used to modify the source and communicate the modified source x_1 across an explicit channel. On the right, an information-theoretic interpretation of the problem is shown.

differs from problems of explicit communication in two ways: it has an implicit (modifiable) source, and an implicit channel, *i.e.* the plant itself is used as a channel. An additional difficulty introduced by the implicit channel is that the message and the messenger coincide: both are x_1. In particular, the capacity of the channel changes with the choice of distribution of x_1. What if we isolated the aspects of implicit source and implicit channel? Could we arrive at a problem that is conceptually simpler than the counterexample? Fig 5.2 shows one such formulation where the source is implicit but the channel is explicit. From this perspective, the counterexample may not be simplest problem of implicit communication: the problem in Fig. 5.2 may even be simpler!

In this problem, the controller C_1 uses a control input u_1 to modify the state x_0. The controller has an AWGN channel of power constraint P_{ch} and noise $Z \sim \mathcal{N}(0, 1)$ connecting it to the controller C_2. The resulting state x_1 needs to be estimated by C_2, who only observes the channel output $y = s + z$, where s is the channel input chosen by C_1. The goal is again to minimize the average cost $k^2\mathbb{E}\left[U_1^2\right] + \mathbb{E}\left[X_2^2\right]$.

This problem also brings out one of the oversimplifications in the "observer-controller" architecture for problems of control under communication constraints: the observer there cannot modify the state, making the state an explicit source. Our problem here enhances the observer by allowing it to modify the source, and considers a problem of just one-shot communication. The problem with a dynamically evolving state can be addressed in a way similar to the problem in Chapter 5.4.

Can this problem be understood using the program we propose? Let us first formulate a semi-deterministic abstraction of the problem.

A semi-deterministic abstraction

A semi-deterministic abstraction of the problem is shown in Fig. 5.3. A minor technical difficulty is that the state is not observed at the decoder, and hence it is unclear where the decimal point in the binary expansion of the state should be placed. For sake of convenience,

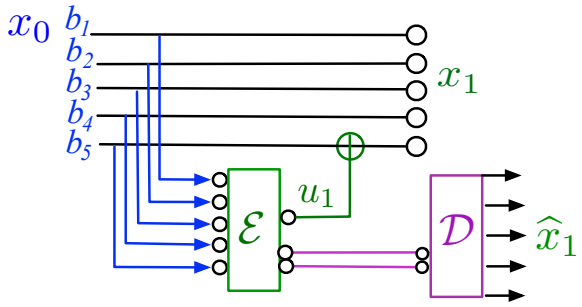


Figure 5.3: A semi-deterministic abstraction of the problem shown in Fig. 5.2.

let us assume that the decimal point is before bit b_1. The SNR on the external channel limits the capacity of this channel. How much power should the encoder use? In the example shown, only two bits can be communicated over the external channel. The encoder should clearly communicate the most significant two bits, b_1 and b_2, on the external channel. The strategy for control input u_1 is also obvious: if enough power is available, the input u_1 should be used to force all least-significant bits to zero (*i.e.* bits b_3 and bits of lower significance). If such power is not available, a zero-input strategy should be used.

What do these strategies look like on the real-line? If the input u_1 has enough power, it forces the state to a quantization point. The resulting quantization-point is sent over the (external) channel, and is easily estimated at the decoder.

Asymptotic-approximate-optimality

The following theorem proves that the strategy obtained from the semi-deterministic abstraction is indeed approximately-optimal.

Theorem 10. *For the problem of implicit messages, but explicit communication, the following lower bound holds on costs.*

$$\inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa_{simpler}} - \sqrt{P}\right)^+\right)^2 \leq \overline{\mathcal{J}}_{opt} \leq \mu \inf_{P\geq 0} \left(\left(\sqrt{\kappa_{simpler}} - \sqrt{P}\right)^+\right)^2,$$

where $\mu \leq 4$, $\kappa_{simpler} = \frac{\sigma_0^2}{P_{ch}+1}$, *and the upper bound is achieved by quantization-based strategies, complemented by linear strategies. Further, quantization-based strategies require the optimal power for forcing MMSE to zero.*

Proof. **A lower bound on the minimum achievable costs by any strategy**

Following the triangle-inequality argument used in proof of Theorem 2, a lower bound on distortion in reproducing $\mathbf{X}_1^m$ is given by

$$\sqrt{\mathbb{E}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\right]} \geq \sqrt{\mathbb{E}\left[\|\mathbf{X}_0^m - \widehat{\mathbf{X}}_1^m\|^2\right]} - \sqrt{\mathbb{E}\left[\|\mathbf{X}_0^m - \mathbf{X}_1^m\|^2\right]}. \tag{5.1}$$

We wish to lower bound $\mathbb{E}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\right]$. The second term in the RHS is smaller than $\sqrt{mP}$. Therefore, it suffices to lower bound the first term on the RHS of (5.1). To that end, we will interpret $\widehat{\mathbf{X}}_1^m$ as an estimate for $\mathbf{X}_0^m$.

The input power constraint P_{ch} limits the channel capacity to $C_{ch} = \frac{1}{2}\log_2(1+P_{ch})$. This in turn determines the amount of power required for source-simplification. A lower bound on the mean-square reconstruction error of $\mathbf{x}_0^m$ is given by

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_0^m - \widehat{\mathbf{X}}_1^m\|^2\right] &\geq D(C_{ch}) \\ &\geq \frac{\sigma_0^2}{P_{ch}+1}.\end{aligned}$$

Thus,

$$MMSE(P) \geq \left(\left(\sqrt{\tfrac{\sigma_0^2}{P_{ch}+1}} - \sqrt{P}\right)^+\right)^2,$$

and a lower bound on the average cost for the problem is

$$\overline{\mathcal{J}}_{\min} \geq \inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\tfrac{\sigma_0^2}{P_{ch}+1}} - \sqrt{P}\right)^+\right)^2.$$

Upper bounds (achieved by quantization and linear strategies)
The upper bounds we use are those of quantization and zero-input. If the quantization strategy uses a power P, then the resulting modified state $\mathbf{X}_1^m$ can be communicated across the channel reliably (error probability converging to zero as $m \to \infty$) as long as the rate-distortion function of the Gaussian source evaluated at the 'distortion' P is smaller than the channel capacity. That is,

$$\frac{1}{2}\log_2\left(\frac{\sigma_0^2}{P}\right) < \frac{1}{2}\log_2(1+P_{ch}).$$

Thus asymptotically, the required power P to have $MMSE = 0$ with this quantization-based strategy is

$$P = \frac{\sigma_0^2}{1+P_{ch}}. \tag{5.2}$$

Thus the asymptotic average achievable cost is upper bounded by

$$\overline{\mathcal{J}}_{VQ} = k^2\frac{\sigma_0^2}{1+P_{ch}}. \tag{5.3}$$

For zero-input strategy, the cost is upper bounded by how well the encoder can represent $\mathbf{X}_0^m$ across a channel of capacity C_{ch}. This is asymptotically the distortion-rate function of a $\mathcal{N}(0, \sigma_0^2)$ source evaluated at C_{ch}, which is

$$\overline{\mathcal{J}}_{ZI} = 0 + D(C_{ch}) = \frac{\sigma_0^2}{P_{ch}+1}.$$

Thus, we obtain the following upper bound

$$\overline{\mathcal{J}}_{\min} \leq \min\left\{k^2, 1\right\} \frac{\sigma_0^2}{P_{ch}+1}.$$

Bounded ratios
In the lower bound, if the optimizing $P^* < \frac{\sigma_0^2}{4(P_{ch}+1)}$, then the $MMSE$, and hence the cost itself, is lower bounded by

$$\overline{\mathcal{J}}_{min} \geq \left(\left(\sqrt{\frac{\sigma_0^2}{P_{ch}+1}} - \sqrt{P}\right)^+\right)^2 > \left(\left(\sqrt{\frac{\sigma_0^2}{P_{ch}+1}} - \sqrt{\frac{\sigma_0^2}{4(P_{ch}+1)}}\right)^+\right)^2 = \frac{\sigma_0^2}{4(P_{ch}+1)}.$$

Thus the ratio of upper and lower bounds is smaller than 4.

If $P^* \geq \frac{\sigma_0^2}{4(P_{ch}+1)}$, the lower bound is larger than $k^2 P^* \geq \frac{\sigma_0^2}{4(P_{ch}+1)}$. Using the quantization-upper-bound of $k^2 \frac{\sigma_0^2}{1+P_{ch}}$ from (5.3), the ratio is again no larger than than 4. □

We observe that the quantization strategy used in our upper bound is also the optimal strategy used for a different problem: that of lossy-reconstruction of the source $\mathbf{x}_0^m$ across a channel. This is a well-known consequence of Shannon's result on the optimality of separating source and channel coding for point-to-point communication. The source is first quantized to a "source-codeword" $\mathbf{x}_1^m$. This codeword is then communicated reliably across the channel. Since in our upper bound, the quantization strategy recovers $\mathbf{x}_1^m$ exactly, it is mathematically equivalent to the separation strategy for source-channel coding. What is different is the goal. Our goal is to minimize the distortion $\mathbb{E}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\right]$, whereas the goal in point-to-point communication is to minimize the distortion $\mathbb{E}\left[\|\mathbf{X}_0^m - \widehat{\mathbf{X}}_0^m\|^2\right]$.

For the Witsenhausen counterexample, in the asymptotic limit of zero-reconstruction error, the question is: what is the power that needs to be injected into the system so that the state can be reconstructed (asymptotically) perfectly across an implicit channel? A complication is that the power injected into the system can not only affect the "information content" of the state, but it can also affect the channel capacity by changing the average power input to the channel, as well as the input distribution. Our problem here gets rid of this added complication by making removing the dependance between the channel input and the system state that was forced by the problem structure. A "verification" of this

simplification is in the results: while simple quantization is optimal for the problem here, one needs to use dirty-paper coding (which historically came [77] much after quantization [1]) for the Witsenhausen counterexample (as shown in Chapter 4.3.3.

This suggests a natural discrete-alphabet problem: a source $\mathbf{X}_0^m$ can be modified $\mathbf{X}_1^m$ under a distortion constraint $\mathbb{E}\left[d(\mathbf{X}_0^m, \mathbf{X}_1^m)\right] \leq P$, for a given P. What is the minimum distortion with which $\mathbf{X}_1^m$ can be communicated across a channel? While the limiting case of zero-distortion (in reconstructing $\mathbf{X}_1^m$) can be solved easily using the separation theorem, the non-zero distortion case is open.

5.2 Witsenhausen with an external channel: control across implicit and explicit channels

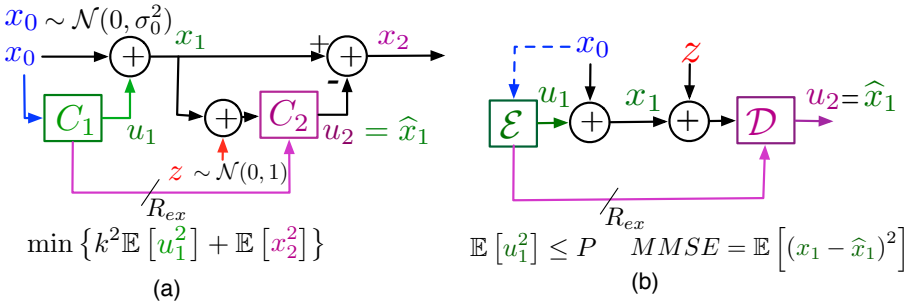


Figure 5.4: A problem of implicit and explicit channels. An external channel connects the two controllers. In a manner similar to the Witsenhausen counterexample, the agents in this problem also lend themselves to an encoder-decoder interpretation.

Witsenhausen's counterexample contains an implicit (modifiable) source and an implicit channel. Our problem in last section contains the aspects of an implicit source and an explicit (external) channel. In this section, we put the two together: we consider a problem where the the source is implicit, and both implicit and explicit channels connect the agents.

The block-diagram for the problem is shown in Fig. 5.4. The formulation is the same as that for Witsenhausen's counterexample except that an explicit channel of finite rate R_{ex} connects the two controllers[2]. The goal is again to minimize the average cost $k^2\mathbb{E}\left[U_1^2\right] + \mathbb{E}\left[X_2^2\right]$.

A similar formulation — one where the external channel has Gaussian noise — was considered by Shoarinejad *et al.* [119] and Martins in [22]. Martins used nonlinear quantization-based strategies that outperform linear strategies even without using an external channel. Here, we use a semi-deterministic abstraction of the problem to obtain improved strategies

[2]A shared finite memory between the controllers can be thought of as a rate-limited channel connecting the two.

based on the concept of binning in information theory. These strategies outperform Martins's strategies by a factor that can diverge to infinity and are shown to be asymptotically-approximately optimal. We first use a semi-deterministic abstraction to obtain intuition into strategy design.

A semi-deterministic abstraction

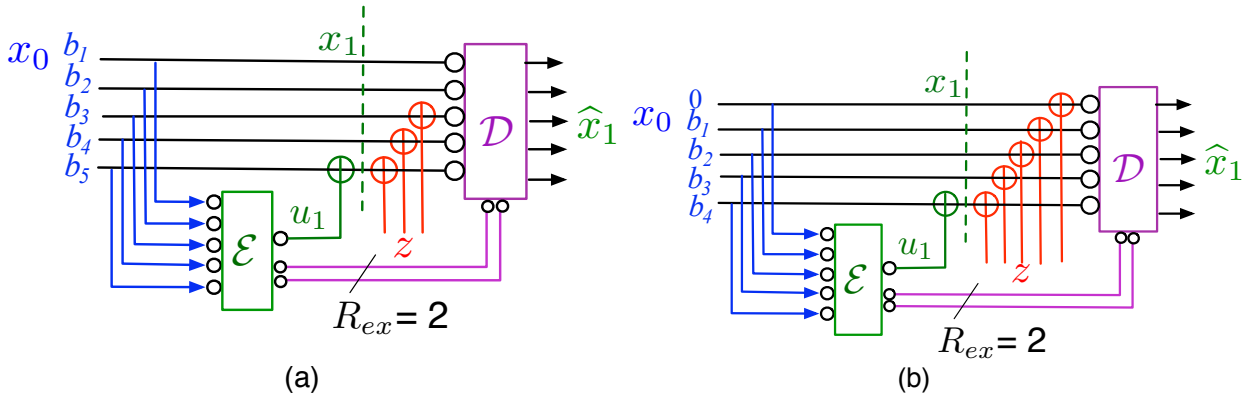


Figure 5.5: A semi-deterministic model for the toy problem of implicit and explicit channels. An external channel (for this example, of capacity $R_{ex} = 2$ bits) connects the two controllers. (a) and (b) show different levels of noise (as compared to initial state variance), and therefore require different strategies.

We characterize the optimal tradeoff between the input power $max(u_1)$ and the power in the reconstruction error $max(x_2)$. Let the power of x_0, $max(x_0)$ be σ_0^2. The noise power is assumed to be 1.

Case 1: $\sigma_0^2 > 1$.

This case is shown in Fig. 5.5(a). The bits b_1, b_2 are communicated noiselessly to the decoder, so the encoder does not need to communicate them explicitly. The external channel has a capacity of two bits, so it can be used to communicate two of the bits b_3, b_4 and b_5. Clearly, we should communicate the more significant bits among those corrupted by noise, *i.e.*, bits b_3 and b_4. If the power $max(u_1)$ of the control input u_1 is large enough, u_1 should be used to modify the least-significant bits (bit b_5 in Fig. 5.5). Else it is best not to spend any power on u_1 and use a zero-input strategy. In the example shown, if $max(u_1) < 0.01$, $MMSE = 0.01$, else $MMSE = 0$.

Case 2: $\sigma_0^2 \leq 1$.

In this case (shown in Fig. 5.5(b)), the signal power is smaller than noise power. All the bits are therefore corrupted by noise, and nothing can be communicated across the implicit channel. In order for the decoder to be able to decode any bit in the representation of x_1, it must either a) know the bit in advance (for instance, encoder can force the bit to 0), or b)

be communicated the bit on the external channel. Since the encoder should use minimum power, it is clear that the most significant bits of the state (bits b_1, b_2 in Fig. 5.5(b)) should be communicated on the external channel. The encoder, if it has sufficient power, can then force the least-significant bits (b_3, b_4 in Fig. 5.5(b)) of x_1 to zero. In the example shown in Fig. 5.5(b), if $max(u_1) < 0.001$, then $max(x_2) = 0.001$, else $max(x_2) = 0$.

What scheme does the semi-deterministic model suggest over the reals?

The inefficiency of linear strategies becomes clear once we look at the semi-deterministic model in Fig. 5.5(a). A linear strategy would communicate the most significant bits of the state on the implicit as well as external channels[3], thereby communicating the same information on two parallel channels. A similar problem, where both the channels are explicit, was considered by Ho, Kastner, and Wong [16], where they also show that nonlinear strategies outperform linear strategies.

For our problem, the scheme obtained from the semi-deterministic abstraction (Case 1) also suggests using a nonlinear strategy that communicates different bits on different channels. The implicit channel is used to communicate the most significant bits. The external channel is used to communicate bits in the middle — bits b_3 and b_4 in Fig. 5.5(a) — which are the most significant bits remaining once the bits above the noise level are taken out. The least significant bits are zeroed out by control input (or cause a reconstruction error, depending on the available power).

How do we port this scheme to the reals? Fig. 5.6 illustrates this. The encoder forces least-significant bits of the state to zero, thereby truncating the binary expansion, or effectively quantizing the state into bins. Unlike for the counterexample, however, the implicit channel by itself does not help us distinguish in which bin the state lies: the channel noise is too large.

The more significant bits among those that are corrupted by noise (b_3, b_4 in Fig. 5.5(a)) are communicated via the external channel. These bits can be thought of as representing the color, *i.e.* the bin index, of quantization bins, where set of $2^{R_{ex}}$ consecutive quantization-bins are labelled with $2^{R_{ex}}$ colors with a fixed ordering (with zero, for instance, colored blue). The bin-index associated with the color of the bin is sent across the external channel. The decoder finds the quantization point nearest to y_2 that has the same bin-index as that received across the external channel.

The scheme is very similar to the binning scheme used for Wyner-Ziv coding of a Gaussian source with side information [120], which is not surprising because of the similarity of our problem with the Wyner-Ziv formulation. The implicit channel provides the "side-information" to the decoder. The external channel is the coding problem. The main difference from the Wyner-Ziv formulation is that the source here is modifiable. The problem is

[3]A linear strategy simply scales the observations. The least-significant bits of the signal are therefore the ones mangled by noise.

also related to that of "successive refinement" in information theory [121], where the communication data is first approximated using a few bits of information, and is then successively refined as more information is supplied. Here, the additional information on the external channel is used to refine the estimate of x_1 obtained from the implicit channel.

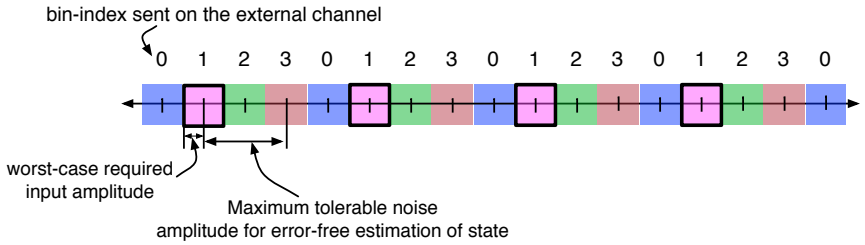


Figure 5.6: The strategy intuited from the semi-deterministic model naturally yields a binning-based strategy for reals that leads to a synergistic use of implicit and explicit channels. The external channel get the decoder the bin-index (in this example, the index is 1). The more significant bits (coarse bin) are received on the implicit channel.

Gaussian external channel

In order to compare the performance of our strategy with that of Martins [22], we consider the Gaussian external channel model used in his work. There, the channel is a power constrained additive Gaussian noise channel. Without loss of generality, we assume that the noise in the external channel is also of variance 1 (the same as the variance of observation noise Z).

Martins's strategy suggests using the control action to quantize on the implicit channel, and communicate the resulting x_1 linearly over the external channel. With strategically chosen problem parameters, our binning-strategy can outperform Martins's strategy in [22]. The key is to choose the set of problems where the initial state variance σ_0^2 and the power on the external channel, denoted by P_{ex}, are almost equal. In this case, Martins's strategy is extremely inefficient since it uses both implicit and explicit channels to communicate the state when the fidelity across both the channels is almost the same. Fig. 5.7 shows that fixing the relation $P_{ex} = \sigma_0^2$, as $\sigma_0^2 \to \infty$, the ratio of costs attained by the binning strategy to that attained by Martins's strategy diverges to infinity.

Asymptotic version of the problem

We now show that the binning strategy of Chapter 5.2 is approximately-optimal in the limit of infinitely many dimensions.

Theorem 11. *For the extension of Witsenhausen's counterexample with an external channel*

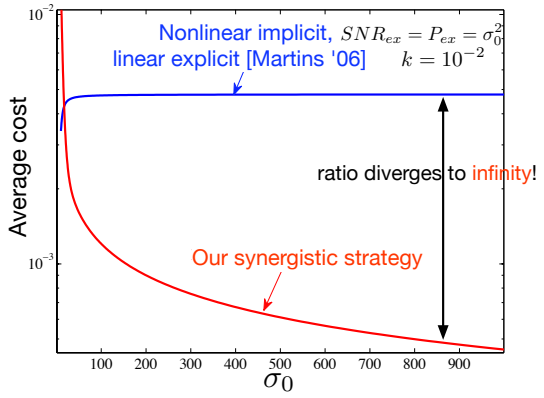


Figure 5.7: If the SNR on the external channel is made to scale with SNR of the initial state, then our binning-based strategy outperforms strategy in [22] by a factor that diverges to infinity.

connecting the two controllers,

$$\begin{aligned} &\inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa_{new}} - \sqrt{P}\right)^+\right)^2 \\ \leq\ & \overline{\mathcal{J}}_{opt} \leq \mu \inf_{P\geq 0} \left(\left(\sqrt{\kappa_{new}} - \sqrt{P}\right)^+\right)^2, \end{aligned}$$

where $\mu \leq 64$, $\kappa_{new} = \frac{\sigma_0^2 2^{-2R_{ex}}}{\overline{P}+1}$, *where* $\overline{P} = \left(\sigma_0 + \sqrt{P}\right)^2$ *and the upper bound is achieved by binning-based quantization strategies. Numerical evaluation shows that* $\mu < 8$.

Proof. **Lower bound**

As before, we wish to lower bound $\mathbb{E}\left[\|\mathbf{X}_1^m - \mathbf{U}_2^m\|\right]$. The second term on the RHS is smaller than $\sqrt{mP}$. Therefore, it suffices to lower bound the first term on the RHS of (4.5).

With what distortion can $\mathbf{x}_0^m$ be communicated to the decoder? The capacity of the parallel channel is the sum of the two capacities $C_{sum} = R_{ex} + C_{implicit}$. The capacity $C_{implicit}$ is upper bounded by $\frac{1}{2}\log_2\left(1+\overline{P}\right)$ where $\overline{P} := \left(\sigma_0 + \sqrt{P}\right)^2$. Using Lemma 1, the distortion in reconstructing $\mathbf{x}_0^m$ is lower bounded by

$$D(C_{sum}) = \sigma_0^2 2^{-2C_{sum}} = \sigma_0^2 2^{-2R_{ex}-2C_{implicit}} \geq \frac{\sigma_0^2 2^{-2R_{ex}}}{\overline{P}+1} = \kappa_{new}.$$

Thus the distortion in reconstructing $\mathbf{x}_1^m$ is lower bounded by

$$\left(\left(\sqrt{\kappa_{new}} - \sqrt{P}\right)^+\right)^2.$$

This proves the lower bound in Theorem 11.

Upper bound

Quantization: This strategy is used for $\sigma_0^2 > 1$. Quantize $\mathbf{x}_0^m$ at rate $C_{sum} = R_{ex} + C_{implicit}$. Bin the codewords randomly into $2^{nR_{ex}}$ bins, and send the bin index on the external channel. On the implicit channel, send the codeword closest to the vector $\mathbf{x}_0^m$.

The decoder looks at the bin-index on the external channel, and keeps only the codewords that correspond to the bin index. This subset of the codebook, which now corresponds to the set of valid codewords, has rate $C_{implicit}$. The required power P (which is the same as the distortion introduced in the source $\mathbf{x}_0^m$) is thus given by

$$\frac{1}{2}\log_2\left(\frac{\sigma_0^2}{P}\right) \leq R_{ex} + \frac{1}{2}\log_2\left(1+\sigma_0^2-P\right),$$

which yields the solution $P = \frac{(1+\sigma_0^2)-\sqrt{(1+\sigma_0^2)^2-4\sigma_0^2 2^{-2R_{ex}}}}{2}$ which is smaller than 1. Thus,

$$P = \frac{(1+\sigma_0^2)-\sqrt{(1+\sigma_0^2)^2-4\sigma_0^2 2^{-2R_{ex}}}}{2} = \frac{1}{2}(1+\sigma_0^2)\left(1-\sqrt{1-4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}}}\right).$$

Now note that $\frac{\sigma_0^2}{(1+\sigma_0^2)^2}$ is a decreasing function of σ_0^2 for $\sigma_0^2 > 1$. Thus, $\frac{\sigma_0^2}{(1+\sigma_0^2)^2} < \frac{1}{4}$ for $\sigma_0^2 > 1$, and $1 - 4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}} > 0$. Because $0 < 1 - 4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}} < 1$,

$$\sqrt{1-4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}}} \geq 1-4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}},$$

and therefore

$$\begin{aligned} P &\leq \frac{1}{2}(1+\sigma_0^2)\left(1-\left(1-4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}}\right)\right) = \frac{1}{2}(1+\sigma_0^2)\left(4\frac{\sigma_0^2}{(1+\sigma_0^2)^2}2^{-2R_{ex}}\right) \\ &= \frac{2\sigma_0^2}{1+\sigma_0^2}2^{-2R_{ex}} \leq 2\times 2^{-2R_{ex}}. \end{aligned}$$

The other strategies that complement this binning strategy are the analogs of zero-forcing and zero-input.

Analog of the zero-forcing strategy

The state $\mathbf{x}_0^m$ is quantized using a rate-distortion codebook of $2^{mR_{ex}}$ points. The encoder sends the bin-index of the nearest quantization-point on the external channel. Instead of forcing the state all the way to zero, the input is used to force the state to the nearest quantization point. The required power is given by the distortion $\sigma_0^2 2^{-2R_{ex}}$. The decoder knows exactly which quantization point was used, so the second stage cost is zero. The total cost is therefore $k^2\sigma_0^2 2^{-2R_{ex}}$.

Analog of the zero-input strategy

Case 1: $\sigma_0^2 \leq 4$.

Quantize the space of initial state realizations using a random codebook of rate R_{ex}, with the codeword elements chosen i.i.d $\mathcal{N}(0, \sigma_0^2(1-2^{-2R_{ex}}))$. Send the index of the nearest codeword on the external channel, and ignore the implicit channel. The asymptotic achieved distortion is given by the distortion-rate function of the Gaussian source $\sigma_0^2 2^{-2R_{ex}}$.

Case 2: $R_{ex} \leq 2$. Do not use the external channel. Perform an MMSE operation at the decoder on the state $\mathbf{x}_0^m$. The resulting error is $\frac{\sigma_0^2}{\sigma_0^2+1}$.

Case 3: $\sigma_0^2 > 4, R_{ex} > 2$.

Our proofs in this part follow those in [122]. Let $R_{code} = R_{ex} + \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{3}\right) - \epsilon$. A codebook of rate R_{code} is designed as follows. Each codeword is chosen randomly and uniformly inside a sphere centered at the origin and of radius $m\sqrt{\sigma_0^2 - D}$, where $D = \sigma_0^2 2^{-2R_{code}} = 3 \times 2^{-2(R_{ex}-\epsilon)}$. This is the attained asymptotic distortion when the codebook is used to represent[4] $\mathbf{x}_0^m$.

Distribute the $2^{mR_{code}}$ points randomly into $2^{mR_{ex}}$ bins that are indexed $\{1, 2, \ldots, 2^{mR_{ex}}\}$. The encoder chooses the codeword $\mathbf{x}_{code}^m$ that is closest to the initial state. It sends the bin-index (say i) of the codeword across the external channel.

Let $\mathbf{z}_{code}^m = \mathbf{x}_0^m - \mathbf{x}_{code}^m$. The received signal $\mathbf{y}_2^m = \mathbf{x}_0^m + \mathbf{z}^m = \mathbf{x}_{code}^m + \mathbf{z}_{code}^m + \mathbf{z}^m$, which can be thought of as receiving a noisy version of codeword $\mathbf{x}_{code}^m$ with a total noise of variance $D+1$, since $\mathbf{z}_{code}^m \perp\!\!\!\perp \mathbf{z}^m$.

The decoder receives the bin-index i on the external channel. Its goal is to find $\mathbf{x}_{code}^m$. It looks for a codeword from bin-index i in a sphere of radius $D+1+\epsilon$ around $\mathbf{y}_2^m$. We now show that it can find $\mathbf{x}_{code}^m$ with probability converging to 1 as $m \to \infty$. A rigorous proof that $MMSE$ also converges to zero can be obtained along the lines of proof in [56].

To prove that the error probability converges to zero, consider the total number of codewords that lie in the decoding sphere. This, on average, is bounded by

$$\frac{2^{mR_{code}}}{Vol\left(\mathcal{S}^m\left(m\sqrt{(\sigma_0^2 - D + \epsilon)}\right)\right)} Vol\left(\mathcal{S}^m\left(m\sqrt{D+1+\epsilon}\right)\right) = \frac{2^{m\left(R_{ex}-\epsilon+\frac{1}{2}\log_2\left(\frac{\sigma_0^2}{3}\right)\right)}}{Vol\left(\mathcal{S}^m\left(m\sqrt{(\sigma_0^2 - D + \epsilon)}\right)\right)} Vol\left(\mathcal{S}^m\left(m\sqrt{D+1+\epsilon}\right)\right)$$

$$= \frac{2^{m\left(R_{ex}-\epsilon+\frac{1}{2}\log_2\left(\frac{\sigma_0^2}{3}\right)\right)}}{\left(m\sqrt{\sigma_0^2 - D + \epsilon}\right)^m} \left(m\sqrt{D+1+\epsilon}\right)^m = 2^{m(R_{ex}-\epsilon)} 2^{\left(\frac{m}{2}\log_2\left(\frac{\sigma_0^2(D+1+\epsilon)}{3(\sigma_0^2 - D + \epsilon)}\right)\right)}.$$

Let us pick another codeword in the decoding sphere. Probability that this codeword has index i is $2^{-mR_{ex}}$. Using union bound, the probability that there exists another codeword in the decoding sphere of index i is bounded by

$$2^{-mR_{ex}} 2^{m(R_{ex}-\epsilon)} 2^{\left(\frac{m}{2}\log_2\left(\frac{\sigma_0^2(D+1+\epsilon)}{3(\sigma_0^2 - D + \epsilon)}\right)\right)} = 2^{-m\epsilon} 2^{\left(\frac{m}{2}\log_2\left(\frac{\sigma_0^2(D+1+\epsilon)}{3(\sigma_0^2 - D + \epsilon)}\right)\right)}.$$

It now suffices to show that the second term converges to zero as $m \to \infty$. Since $D = 3 \times 2^{-2(R_{ex}-\epsilon)}$. Since $R_{ex} > 2$, $D < \frac{3}{4} \times 2^{\epsilon} < \frac{5}{6} - \epsilon$ for small enough ϵ. Since $\sigma_0^2 > 4$,

[4] In the limit of infinite block-lengths, average distortion attained by a uniform-distributed random-codebook and a Gaussian random-codebook of the same variance is the same [122].

$D < \frac{5}{6}\frac{\sigma_0^2}{4} < \frac{\sigma_0^2}{4} + \epsilon$,

$$\frac{\sigma_0^2(D+1+\epsilon)}{3(\sigma_0^2 - D + \epsilon)} < \frac{\sigma_0^2 \times \left(\frac{5}{6}+1\right)}{3\frac{3\sigma_0^2}{4}} = \frac{\frac{11}{6}}{\frac{9}{4}} = \frac{22}{27} < 1.$$

Thus the cost here is bounded by $3 \times 2^{-2(R_{ex}-\epsilon)}$ which is bounded by $4 \times 2^{-2R_{ex}}$ for small enough ϵ.

Bounded ratios for the asymptotic problem

The upper bound is the best of the vector-quantization bound, $2k^2 2^{-2R_{ex}}$, zero-forcing $k^2\sigma_0^2 2^{-2R_{ex}}$, and zero-input bounds of $\sigma_0^2 2^{-2R_{ex}}$ and $4 \times 2^{-2R_{ex}}$.

Case 1: $P^* > \frac{2^{-2R_{ex}}}{16}$.

In this case, the lower bound is larger than $k^2 \frac{2^{-2R_{ex}}}{16}$. Using the upper bound of $4 \times 2^{-2R_{ex}}$, the ratio is smaller than 64.

Case 2: $P^* \leq \frac{2^{-2R_{ex}}}{16}, \sigma_0^2 \geq 1$.

Since $R_{ex} \geq 0$, $P^* \leq \frac{1}{16}$. Thus,

$$\kappa_{new} = \frac{\sigma_0^2 2^{-2R_{ex}}}{(\sigma_0 + \sqrt{P^*})^2 + 1} > \frac{1}{\left(1+\frac{1}{4}\right)^2 + 1} = \frac{16}{41} 2^{-2R_{ex}}.$$

Thus, the lower bound is greater than the $MMSE$ which is larger than

$$\left(\sqrt{\frac{16}{41}} - \sqrt{\frac{1}{16}}\right)^2 2^{-2R_{ex}} \approx 0.14 \times 2^{-2R_{ex}}. \tag{5.4}$$

Using the upper bound of $4 \times 2^{-2R_{ex}}$, the ratio is smaller than 29.

Case 3: $P^* \leq \frac{2^{-2R_{ex}}}{16}, \sigma_0^2 < 1$.

If $P^* > \frac{\sigma_0^2 2^{-2R_{ex}}}{25}$, using the upper bound of $\sigma_0^2 2^{-2R_{ex}}$, the ratio is smaller than 25.

If $P^* \leq \frac{\sigma_0^2 2^{-2R_{ex}}}{25} < \frac{1}{25}$,

$$\kappa_{new} = \frac{\sigma_0^2 2^{-2R_{ex}}}{(\sigma_0 + \sqrt{P^*})^2 + 1} \geq \frac{\sigma_0^2 2^{-2R_{ex}}}{\left(1+\frac{1}{5}\right)^2 + 1} \sigma_0^2 2^{-2R_{ex}} = \frac{25}{61}\sigma_0^2 2^{-2R_{ex}}.$$

Thus, a lower bound on $MMSE$, and hence also on the total costs, is

$$\left(\sqrt{\frac{25}{61}} - \sqrt{\frac{1}{25}}\right)^2 \sigma_0^2 2^{-2R_{ex}} \approx 0.19\sigma_0^2 2^{-2R_{ex}}.$$

Using the upper bound of $\sigma_0^2 2^{-2R_{ex}}$, the ratio is smaller than $\frac{1}{0.19} < 6$. □

Finite-vector length problem

Are the proposed binning strategies approximately-optimal for finite vector lengths? Following the lead from Chapter 4 for the Witsenhausen counterexample, we can consider lattice-based strategies. In [64] we investigate the problem for the scalar case. Our lower bounds for the original counterexample extend naturally to this problem. While the ratio of upper and lower bounds is bounded uniformly for each R_{ex}, it diverges to infinity as $R_{ex} \to \infty$. We believe that a tightening of the upper bound (i.e. a better achievable strategy) in the regime of large-k, large-σ_0 is required to attain within a constant factor of the optimal cost, and to not have the constant depend on R_{ex}.

5.3 A problem exhibiting the triple role of control actions

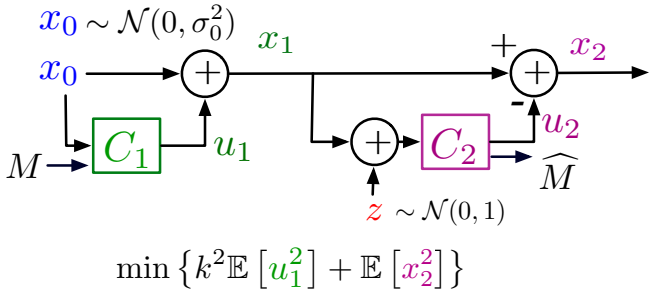


Figure 5.8: A problem that brings out the triple role of control actions in decentralized control. The control actions are used to reduce the immediate control costs, communicate a message, and improve state estimability at the second controller.

As discussed in Chapter 1.5.4, control actions in decentralized systems can play a triple role: control, communication and improving estimability[5]. In Witsenhausen's counterexample, the two roles of communication and improving state estimability are aligned: the state x_1 is both the message and the messenger, so improving estimability of x_1 also communicates the message, which is also x_1. In more general problems, such an alignment need not be present. Are such problems much harder than the counterexample? After all, the dual role in the counterexample made the problem much harder than the problems where all three roles are aligned.

[5]As noted earlier, in adaptive control, control actions have a fourth role to play — that of enabling the learning of system parameters [46]. This was explored first by Feldbaum in a series of papers starting with [46]. Similar issues arise there: certainty-equivalence-based strategies are also suboptimal for problems where control actions have to learn as well as control [46].

Our toy problem to test this question is a simple extension on the vector Witsenhausen counterexample. Not only does the first controller want to improve the state estimability at the second controller (thus keeping the weighted sum of power and $MMSE$ costs low), it also wants to communicate an independent message at rate R.

A semi-deterministic model

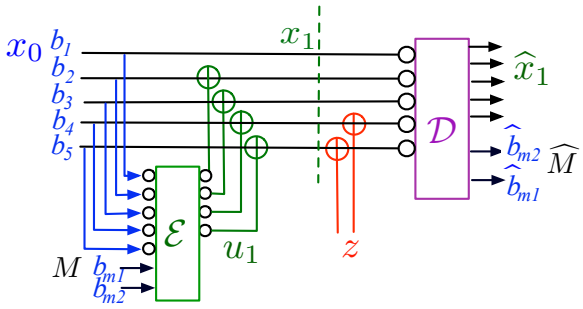


Figure 5.9: A semi-deterministic model for the problem shown in Fig. 5.8. The encoder wants to communicate a two-bit message b_{m1}, b_{m2} to the decoder, as well as minimize the system costs.

Based on the semi-deterministic model for the counterexample, the optimal strategy for the semi-deterministic model shown in Fig. 5.9 is obvious. In order to communicate one bit across the channel, the encoder must encode this information in bits that are not affected by noise. In the particular example of Fig. 5.9, a two-bit message is encoded in bit b_2, b_3. At the same time, because the most significant bit to be modified is already determined by the number of message bits, the least-significant bits b_4 and b_5 can be forced to zero for free (This is an artifact of our choice of cost function in Chapter 4.1. The chosen power function depends only on the most-significant bit that is modified.).

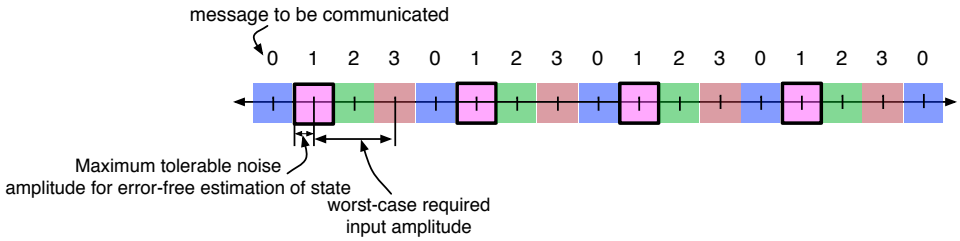


Figure 5.10: Strategies for the problem of triple role of control shown in Fig. 5.8. Comparison with Fig. 5.6 reveals an interesting duality of this problem with that of Chapter 5.2.

What strategies does this semi-deterministic version suggest for the actual problem? These strategies, shown in Fig. 5.10, are conceptually dual (see Fig. 5.6 for comparison) to the strategies for the problem in last section of signaling the implicit source x_1 across an

implicit and explicit channels. This is not surprising: the problem in Chapter 5.2 is one where two channels are used to communicate one (implicit) source. In this case, a single channel is being to communicate two sources.

Indeed, the following theorem shows that the attained strategies are approximately optimal for all rates.

Theorem 12. *For the problem exhibiting triple role of control,*

$$\inf_{P\geq 2^{2R}-1} k^2 P + \left(\left(\sqrt{\kappa_{triple}} - \sqrt{P}\right)^+\right)^2 \leq \overline{\mathcal{J}}_{opt}$$
$$\leq \mu \inf_{P\geq 2^{2R}-1} k^2 P + \left(\left(\sqrt{\kappa_{triple}} - \sqrt{P}\right)^+\right)^2,$$

where $\mu \leq 21$*, and* $\kappa_{triple} = \frac{\sigma_0^2 2^{2R}}{\sigma_0^2+P+2\sigma_0\sqrt{P}+1}$*. The upper bound is achieved by quantization-based strategies, complemented by linear strategies.*

Upper bound

DPC with $\alpha = 1$.

In [77], the DPC parameter is chosen to be $\alpha = \alpha_{mmse} = \frac{P}{P+1}$ to achieve the maximum rate of $\frac{1}{2}\log_2(1+P)$. Since the cost here is not merely that of input power (an additional $MMSE$ term is also present), same α may no longer be the optimizing one. For general α, as shown in [77], the achievable rate is

$$R(\alpha) = \frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+1)}{P\sigma_0^2(1-\alpha)^2+P+\alpha^2\sigma_0^2}\right). \tag{5.5}$$

With $\alpha = 1$ the decoder decodes $\mathbf{X}_1^m$ perfectly (in the limit $m \to \infty$), thereby attaining asymptotically zero $MMSE$. The attained rate is

$$R(1) = \frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+1)}{P+\sigma_0^2}\right) = \frac{1}{2}\log_2\left(P\left(1+\frac{1}{P+\sigma_0^2}\right)\right) \geq \frac{1}{2}\log_2(P). \tag{5.6}$$

Thus, to attain a rate R, the cost is upper bounded by the cost attained by $DPC(1)$ strategy, yielding

$$\overline{\mathcal{J}}_{\min} \leq k^2 2^{2R} + 0 = k^2 2^{2R}. \tag{5.7}$$

DPC with $\alpha = \alpha_{Costa} = \frac{P}{P+1}$.

For this choice of α, the achievable rate is well known to equal the channel capacity for interference-free version of the channel [77]

$$R(\alpha_{Costa}) = \frac{1}{2}\log_2(1+P), \tag{5.8}$$

and the required power is, therefore, $P = 2^{2R} - 1$. The expression for MMSE-error in estimation of $\mathbf{X}_1^m = \mathbf{X}_0^m + \mathbf{U}_1^m$ can be unwieldy because it is estimated using $\mathbf{V}^m = \mathbf{U}_1^m + \alpha \mathbf{X}_0^m$ as well as $\mathbf{Y}_2^m = \mathbf{X}_1^m + \mathbf{Z}^m$. For analytical simplicity, we use two upper bounds. Instead of estimating $\mathbf{X}_1^m$ using both $\mathbf{Y}_2^m$ and $\mathbf{V}^m$, we use just $\mathbf{Y}_2^m$, or just $\mathbf{V}^m$. In the first case, using just $\mathbf{Y}_2^m$, the MMSE error is $\frac{\sigma_0^2+P}{\sigma_0^2+P+1}$. In the second case, assuming asymptotically perfect decoding, the MMSE error is

$$\begin{aligned} MMSE &= \mathbb{E}\left[X_1^2\right] - \frac{\left(\mathbb{E}\left[X_1 V\right]\right)^2}{\mathbb{E}\left[V^2\right]} = P + \sigma_0^2 - \frac{(P+\alpha\sigma_0^2)^2}{P+\alpha^2\sigma_0^2} \\ &= \frac{P^2 + \alpha^2\sigma_0^4 + P\sigma_0^2(1+\alpha^2) - P^2 - \alpha^2\sigma_0^4 - 2\alpha P\sigma_0^2}{P+\alpha^2\sigma_0^2} \\ &= \frac{P\sigma_0^2(1-\alpha)^2}{P+\alpha^2\sigma_0^2} \overset{(\alpha=\frac{P}{P+1})}{=} \frac{P\sigma_0^2\left(\frac{1}{P+1}\right)^2}{P+\frac{P^2}{(P+1)^2}\sigma_0^2} = \frac{\sigma_0^2}{(P+1)^2+P\sigma_0^2} \\ &\overset{(P=2^{2R}-1)}{=} \frac{\sigma_0^2}{2^{4R}+(2^{2R}-1)\sigma_0^2}. \end{aligned} \tag{5.9}$$

Straight coding

In this strategy, we first force the initial state to zero, and then add a codeword to communicate across the channel. Since the message (and hence the codeword) is independent of the initial state $\mathbf{X}_0^m$, the total power required is the sum of the powers of the codeword and the initial state. Using Costa's result [77], the required codebook power is $2^{2R}-1$. Thus the required total power is $P = \sigma_0^2 + 2^{2R} - 1$, and the required cost is

$$\overline{\mathcal{J}}_{\min} \leq k^2(\sigma_0^2 + 2^{2R} - 1). \tag{5.10}$$

Lower bounds on $MMSE(P,R)$

Theorem 13. *For the problem stated above, for communicating reliably at rate R with input power P, the asymptotic average mean-square error in recovering $\mathbf{X}_1^m$ is lower bounded as follows. For $P \geq 2^{2R}-1$,*

$$MMSE(P,R) \geq$$
$$\inf_{\sigma_{X_0,U_1}} \sup_{\gamma>0} \frac{1}{\gamma^2}\left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{1+\sigma_0^2+P+2\sigma_{X_0,U_1}}} - \sqrt{(1-\gamma)^2\sigma_0^2 + \gamma^2 P - 2\gamma(1-\gamma)\sigma_{X_0,U_1}}\right)^+\right)^2,$$

where $\max\left\{-\sigma_0\sqrt{P}, \frac{2^{2R}-1-P-\sigma_0^2}{2}\right\} \leq \sigma_{X_0,U_1} \leq \sigma_0\sqrt{P}$. *For $P < 2^{2R}-1$, reliable communication at rate R is not possible. Further, in the asymptotic limit of zero $MMSE$, the strategy that attains the optimal tradeoff between power P and rate R is a dirty-paper coding-based strategy.*

Proof. See Appendix A.7. □

For analytical ease in proving approximate optimality, we simplify the above bound. Choosing $\gamma = 1$ in (5.11) we can obtain the following (loosened) bound.

$$MMSE(P,R) \geq \inf_{|\sigma_{X_0,U_1}| \leq \sigma_0\sqrt{P}} \left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{\sigma_0^2 + P + 2\sigma_{X_0,U_1} + 1}} - \sqrt{P} \right)^+ \right)^2, \tag{5.11}$$

which is minimized for $\sigma_{X_0,U_1} = \sigma_0\sqrt{P}$, yielding

$$MMSE(P,R) \geq \left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{\sigma_0^2 + P + 2\sigma_0\sqrt{P} + 1}} - \sqrt{P} \right)^+ \right)^2, \tag{5.12}$$

Thus a lower bound on the total cost is given by

$$\overline{\mathcal{J}}_{\min} \geq \inf_{P \geq 2^{2R}-1} k^2 P + \left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{\sigma_0^2 + P + 2\sigma_0\sqrt{P} + 1}} - \sqrt{P} \right)^+ \right)^2. \tag{5.13}$$

Proof that the ratio of upper and lower bounds is bounded

Case 1: $R \geq \frac{1}{4}$.
We use the DPC($\alpha = 1$) upper bound from (5.7) of $2^{2R}k^2$. The lower bound is clearly larger than $k^2(2^{2R} - 1)$, since $P \geq 2^{2R} - 1$ in (5.13). The ratio of upper and lower bounds is therefore smaller than

$$\frac{k^2 2^{2R}}{k^2(2^{2R}-1)} \overset{R \leq \frac{1}{4}}{\leq} \frac{\sqrt{2}}{\sqrt{2}-1} \approx 3.4 < 4.$$

Case 2: $P^* \geq \frac{2^{2R}}{8\sqrt{2}}$.
Again, we use the DPC($\alpha = 1$) upper bound of $2^{2R}k^2$. The lower bound is larger than $k^2 P^* \geq k^2 \frac{2^{2R}}{8\sqrt{2}}$. Thus, the ratio is smaller than $8\sqrt{2} \approx 11.3 < 12$.

Case 3: $R < \frac{1}{4}$, $P^* < \frac{2^{2R}}{8\sqrt{2}}, \sigma_0^2 > 1$.
For the lower bound, note that

$$\begin{aligned} \kappa_{triple} &= \frac{\sigma_0^2 2^{2R}}{(\sigma_0 + \sqrt{P^*})^2 + 1} \overset{\sigma_0^2 \geq 1}{\geq} \frac{2^{2R}}{(1 + \sqrt{P^*})^2 + 1} \\ &\overset{P^* < \frac{2^2 R}{8\sqrt{2}}}{\geq} \frac{2^{2R}}{(1 + \frac{2^R}{2})^2 + 1} \overset{R < \frac{1}{4}}{\geq} \frac{\sqrt{2}}{\left(1 + \sqrt{\frac{1}{8}}\right)^2 + 1} \geq 0.49. \end{aligned}$$

Thus,

$$\begin{aligned}\overline{\mathcal{J}}_{\min} \geq k^2(2^{2R}-1)+MMSE &\geq k^2(2^{2R}-1)+\left(\left(\sqrt{\kappa_{triple}}-\sqrt{P^*}\right)^+\right)^2 \\ &\geq k^2(2^{2R}-1)+\left(0.7-\sqrt{\frac{1}{8}}\right)^2 \geq k^2(2^{2R}-1)+0.12.\end{aligned}$$

Upper bound of DPC($\alpha=\alpha_{Costa}$) is smaller than $k^2(2^{2R}-1)+1$. Thus the ratio is smaller than $\frac{1}{0.12}<9$.

Case 4: $R<\frac{1}{4}$, $P^*<\frac{2^{2R}}{8\sqrt{2}}, \sigma_0^2\leq 1$.

Case 4a: If $\sigma_0^2<20(2^{2R}-1)$, using the straight coding upper bound, the cost is smaller than

$$\overline{\mathcal{J}}_{\min}\leq k^2(\sigma_0^2+2^{2R}-1)\leq 21k^2(2^{2R}-1).$$

Since the lower bound is larger than $k^2(2^{2R}-1)$, the ratio is smaller than 21.

Case 4b: If $20(2^{2R}-1)<\sigma_0\leq 1$, then the straight coding upper bound yields

$$\overline{\mathcal{J}}_{\min}\leq k^2(\sigma_0^2+2^{2R}-1)\leq \frac{21}{20}k^2\sigma_0^2. \tag{5.14}$$

For the lower bound, if $P^*>\frac{\sigma_0^2}{20}$, the ratio is smaller than 21.

If $P^*\leq\frac{\sigma_0^2}{20}$, $P^*\leq\frac{1}{20}$. Thus,

$$\begin{aligned}\kappa_{triple} &= \frac{\sigma_0^2 2^{2R}}{\left(\sigma_0+\sqrt{P}\right)^2+1} \\ &\overset{R\geq 0}{\geq} \frac{\sigma_0^2}{\left(\sigma_0+\sqrt{P}\right)^2+1} \\ &\overset{\sigma_0\leq 1, P^*\leq\frac{1}{20}}{\geq} \frac{\sigma_0^2}{1.05^2+1}\geq\frac{\sigma_0^2}{3}.\end{aligned}$$

Thus the lower bound is larger than

$$\overline{\mathcal{J}}_{\min}\geq k^2(2^{2R}-1)+\left(\left(\frac{\sigma_0^2}{3}-\frac{\sigma_0^2}{20}\right)^+\right)^2\geq k^2(2^{2R}-1)+0.1251\sigma_0^2.$$

Upper bound is based on DPC($\alpha=\alpha_{Costa}$). Using (5.9), the upper bound is smaller than

$$\begin{aligned}\overline{\mathcal{J}}_{\min} &\leq k^2(2^{2R}-1)+\frac{\sigma_0^2}{2^{4R}+(2^{2R}-1)\sigma_0^2} \\ &\overset{R\geq 0}{\leq} k^2(2^{2R}-1)+\sigma_0^2.\end{aligned}$$

The ratio is smaller[6] than $\frac{1}{0.1251} < 8$.

5.4 Introducing feedback: dynamic version of the Witsenhausen counterexample

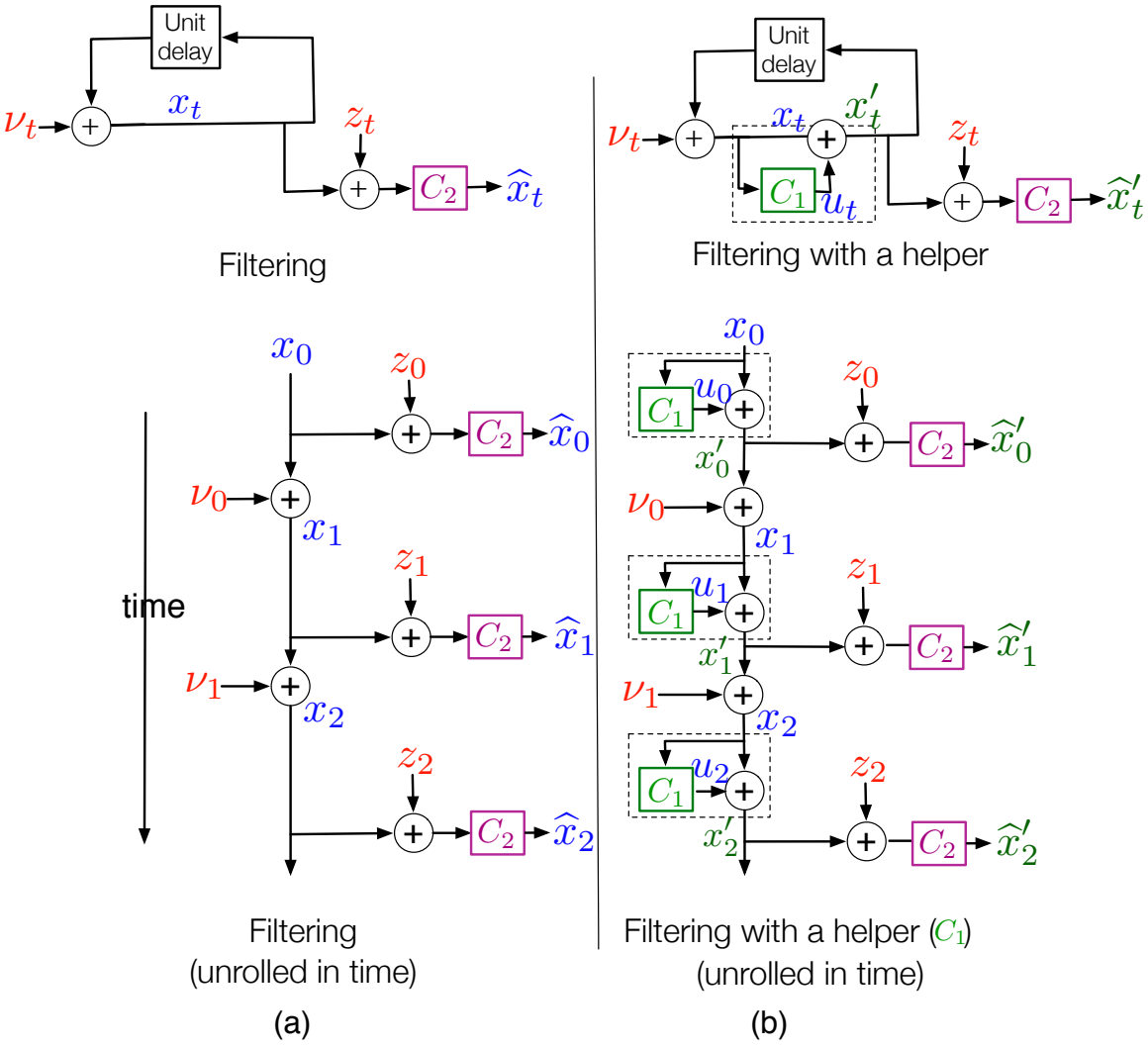


Figure 5.11: (a) A simplistic filtering problem. (b) The problem of filtering with a helper, which is a dynamic version of Witsenhausen's counterexample (unrolled to multiple time-steps).

In this section, we extend the results for the counterexample to a dynamic setting. The setting is as follows. Fig. 5.11(a) shows a system that is perturbed at each time instant by

[6]Figures illustrating this bounded ratio can be found in [57, 123].

an independent perturbation ν_t^m. The resulting state evolution is

$$\mathbf{x}_{t+1}^m = \mathbf{x}_t^m + \nu_t^m. \tag{5.15}$$

The state is observed noisily by a controller C_2 (suggestively named because we will soon modify the problem to have another controller C_1). The problem is one of simple filtering: the controller C_2 wants to minimize the mean-square error in estimation of x_t. The optimal strategy is well known to be linear, and can be obtained using simple Kalman filtering [67].

A modified "active" version of the problem is shown in Fig. 5.11(b). Here, a controller C_1 has the ability to modify the state x_t before it is (partially) observed by C_2. The state evolution is, therefore,

$$\mathbf{x}_{t+1}^m = \mathbf{x}'^m_t + \nu_t^m, \;\; \mathbf{x}'^m_t = \mathbf{x}_t^m + \mathbf{u}_t^m. \tag{5.16}$$

As in the counterexample, the observations of C_1 are assumed to be perfect, and the the observations of C_2 are given by $\mathbf{y}_t^m = \mathbf{x}'^m_t + \mathbf{z}_t^m$. The time-horizon is assumed to be n. Extending the cost function of the counterexample to larger time-horizon, the cost here is a weighted sum of the average power of control input $\mathbf{u}_t^m$ and the mean-square estimation error over the entire time, *i.e.*

$$\overline{\mathcal{J}} = \sum_{t=1}^{n} \left(k^2 \frac{1}{m} \mathbb{E}\left[\|\mathbf{U}_t^m\|^2 \right] + \frac{1}{m} \mathbb{E}\left[\|\mathbf{X}'^m_t - \widehat{\mathbf{X}}'^m_t\|^2 \right] \right). \tag{5.17}$$

The LQG version that we address is a special case where $\mathbf{X}_0^m$ and ν_t^m are iid distributed $\mathcal{N}(0, \sigma_0^2 \mathbb{I})$, and the observation noise $\mathbf{z}_t^m \sim \mathcal{N}(0, \mathbb{I})$.

A strategy and achievable costs: For $k^2 < 1$, $\sigma_0^2 > 1$, we use the Vector Quantization strategy developed for the counterexample. At time 0, the first controller quantizes the state $\mathbf{x}_0^m$, and the second controller estimates the state to be the nearest quantization point. The error probability at this time is close to zero as long as the input power at the first controller, $P > 1$. For any $t > 0$, the controller shifts the origin to $\mathbf{x}'^m_{t-1}$, and uses the same quantization-codebook in these shifted coordinates to quantize $\mathbf{x}'^m_t$. As long as C_2 estimates the state perfectly, the asymptotic total cost using this strategy is nk^2 (at each time-step, the cost is $k^2 P + 0$, where P can be made as close to 1 as desired). Thus $\overline{\mathcal{J}}_{opt} \leq nk^2$.

A lower bound on the minimum possible cost: The lower bound simulates the shifting of axes in the upper bound by giving C_2 the side-information of $\mathbf{x}'^m_{t-1}$ at time t. This new problem with side-information effectively decouples the state-evolution across different time-steps. Therefore, a lower bound to this problem is simply the lower bound for the counterexample (Theorem 2) multiplied by the time-horizon n:

$$\overline{\mathcal{J}}_{opt} \geq n \left(\inf_{P \geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P} \right)^+ \right)^2 \right). \tag{5.18}$$

Bounded ratios: We focus on the region $k^2 \leq 1$, $\sigma_0^2 > 1$. In this region, both the upper bound and the lower bound for the problem are simply the time-horizon n multiplied

with the corresponding bounds for the Witsenhausen counterexample in Theorem 2. The ratio is again bounded by a factor of 4.45 for this quadrant in the (k, σ) parameter space for all time-horizons n. However, we do not have a result for approximate optimality over the entire space because the ratio of the costs attained by linear strategies and the lower bound in (5.18) diverges to infinity as $n \to \infty$ (even though it is bounded for each n).

5.5 A problem of rational inattention

The problem is motivated by signaling in economics literature. Here, control actions are often the only way of speaking because external channels are either unavailable, or signals sent across these channels are not trustworthy[7].

A model proposed by Sims [51] allows for an arbitrary function to map observations of various economic agents to inputs. In order to bound this function with an information-processing constraint, this *rational-inattention model* assumes that the mutual information between the observation and the control input is bounded by a constant I. The justification is that the information-processing ability of each agent is limited, even though the agent has a choice in how to allocate that ability.

Computer calculations of Matejka [53, 54] provide evidence that for a toy model of a seller and a rationally-inattentive consumer, the numerically-optimal pricing strategies that "catch the consumer's attention" are discrete. A discretization of prices makes it easy for the consumer (who has a limited attention) to decide quickly on the price-changes, thereby stimulating her to consume more. The discreteness here arises out of reasons that are quite similar to our understanding of using actions for source-simplification: a simplified source is more easily estimated. Can we obtain theoretical guarantees on the goodness of Matejka's numerically-optimal strategies?

In order to obtain such guarantees, we simplify focus on Matejka's formulation in [54], where he addresses a problem of information-constrained tracking. The goal is to understand how well a pricing strategy can help the consumer track the prices. Here we address a version of the problem with quadratic cost function[8].

The block-diagram (shown in Fig. 5.12) is the same as that for Witsenhausen's counterexample except that the second controller no longer has noise in its observations. Instead, it is limited by the following mutual-information constraint

$$I(X_1; U_2) \leq I. \tag{5.19}$$

[7] A case when a jammer acts on the signals on an external channels has been addressed recently in [124]. It will be interesting to see if an implicit channel can be used to counter the jammer.

[8] The translation from utility function of [54] to cost function here is often simple: the negative of the utility can be thought of as the cost incurred.

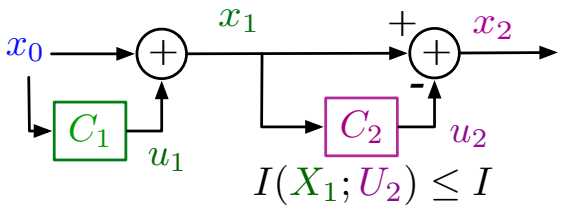


Figure 5.12: A problem of rational inattention. The second controller has perfect observations of the state X_1, but is limited by an information-processing constraint $I(X_1;U_2) \le I$.

Theorem 14. *For the problem of rational-inattention,*

$$\inf_{P\ge 0} k^2 P + \left(\left(\sqrt{\kappa_{RI}} - \sqrt{P}\right)^+\right)^2 \le \overline{\mathcal{J}}_{opt} \le \mu \inf_{P\ge 0} \left(\left(\sqrt{\kappa_{RI}} - \sqrt{P}\right)^+\right)^2,$$

where $\mu \le 4$, $\kappa_{RI} = \sigma_0^2 2^{-2I}$, and the upper bound is achieved by quantization-based strategies, complemented by linear strategies.

Proof. **A lower bound**
Following the lines of proof of Theorem 2,

$$\mathbb{E}\left[\left(X_0 - \widehat{X}_1\right)^2\right] \ge D(I) = \sigma_0^2 2^{-2I}.$$

Using the triangle-inequality argument (Lemma 1), this immediately yields the following lower bound on the total cost

$$\overline{\mathcal{J}} \ge \inf_P k^2 P + \left(\left(\sqrt{\sigma_0^2 2^{-2I}} - \sqrt{P}\right)^+\right)^2. \tag{5.20}$$

We now provide two upper bounds. Remember that under the mutual-information constraint of (5.19), we are free to choose the mapping from $X_1 \to Y_2$ as we like. In either case, we will choose $Y_2 = X_1 + Z$ for *additive Gaussian noise* $Z \sim \mathcal{N}(0,N)$ of some variance N and independent of X_1.

Zero-input upper bound
For zero-input, $X_1 = X_0$. The mutual information $I(X_1;Y_2)$ is therefore given by

$$I(X_1;Y_2) = I(X_0;X_0+Z) = \frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2}{N}\right) \le I. \tag{5.21}$$

Thus we choose $N = \frac{\sigma_0^2}{2^{2I}-1}$. Correspondingly, the $MMSE$ is given by

$$MMSE = \frac{\sigma_0^2 \frac{\sigma_0^2}{2^{2I}-1}}{\sigma_0^2 + \frac{\sigma_0^2}{2^{2I}-1}} = \frac{\sigma_0^2}{2^{2I}}. \tag{5.22}$$

Thus we get the following upper bound on the costs

$$\overline{\mathcal{J}}_{RI} \leq \frac{\sigma_0^2}{2^{2I}}. \tag{5.23}$$

Quantization upper bound
Using quantization strategy with power P, and choosing a noise variance of N, the mutual information condition becomes

$$\frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2 - P}{N}\right) \leq I. \tag{5.24}$$

We know from vector-quantization results for the vector Witsenhausen counterexample that any choice of $P > N$ suffices in the asymptotic limit of large dimensions. Thus, in the limit $m \to \infty$, the required condition is

$$\begin{aligned} & \frac{1}{2}\log_2\left(1 + \frac{\sigma_0^2 - N}{N}\right) < I \\ \text{i.e.} \quad & \frac{\sigma_0^2}{N} < 2^{2I} \\ \Rightarrow \quad & N > \sigma_0^2 2^{-2I}. \end{aligned}$$

With this condition satisfied, the second stage costs can be made to converge to zero as $m \to \infty$. Thus, an achievable cost, in the limit $m \to \infty$ is

$$\overline{\mathcal{J}}_{RI} \leq k^2 \sigma_0^2 2^{-2I}. \tag{5.25}$$

Proof of approximate-optimality
Case 1: If $P^* \geq \frac{\sigma_0^2 2^{-2I}}{4}$.
In this case, the lower bound is larger than $k^2 \frac{\sigma_0^2 2^{-2I}}{4}$. Using the quantization upper bound, the upper bound is smaller than $k^2 \sigma_0^2 2^{-2I}$. The ratio of upper and lower bounds is therefore smaller than 4.
Case 2: If $P^* < \frac{\sigma_0^2 2^{-2I}}{4}$.
In this case, in the lower bound, $MMSE > \frac{\sigma_0^2 2^{-2I}}{4}$. The zero-input upper bound is smaller than $\sigma_0^2 2^{-2I}$. Thus the ratio is again smaller than 4. □

5.6 A noisy version of Witsenhausen's counterexample, and viewing the counterexample as a corner case

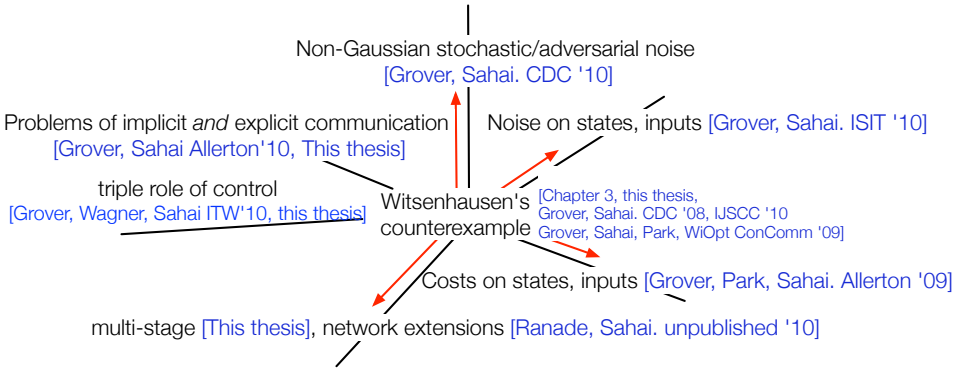


Figure 5.13: Witsenhausen's counterexample as a corner case in an investigation of implicit communication. Is the problem too idealistic for it to be useful?

In Chapter 3.5, we argued that the counterexample is an information-theoretic problem where the controllers can be interpreted as encoders and decoders. Thus, in hindsight, it is not surprising that information-theoretic techniques provide insights, and even approximately optimal solutions to the problem. This leads to a natural question: is the counterexample too idealistic and therefore impractical? After all, the counterexample is a corner case (see Fig. 5.13) where there is no noise in observations at the first controller, much like an encoder, and no cost on the input of the second, much like a decoder. The resemblance is too obvious for comfort: *a priori*, it is unclear whether this understanding extends to more complicated problems when these controllers are less caricatured.

In this section, we shall demonstrate that the understanding and the techniques built for Witsenhausen's counterexample also extend to problems where the encoder/decoder interpretation of the controllers is not strictly valid. We will consider a noisy version of the counterexample where there are noises not only in the observation of the first controller, but also in the state evolution and inputs of the controller. We shall see that approximately optimal solutions can be derived for this problem as well, even though the first controller is no longer quite like an encoder. Later we will see that the approximate-solution to this problem also addresses a version of the counterexample with noises in all observations, state-evolutions, and inputs.

A complementary problem is one in which costs are imposed on the input of the second controller, as well as all the states. This costlier counterexample does not naturally allow for the second controller to be interpreted as a decoder. Approximately optimal solutions

to the problem are not included in this dissertation to keep it relatively focused, and can be found in [61].

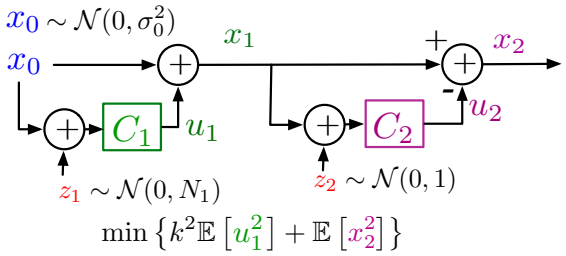


Figure 5.14: A noisy version of Witsenhausen's counterexample where there is noise in the observation of the first controller as well.

Equivalence of the two problems

In this section we show that the problem of Fig. 5.14 is equivalent to a problem with noise in evolution of state $\mathbf{X}_1^m$, but noiseless observation at the encoder, shown in Fig. 5.15(c).

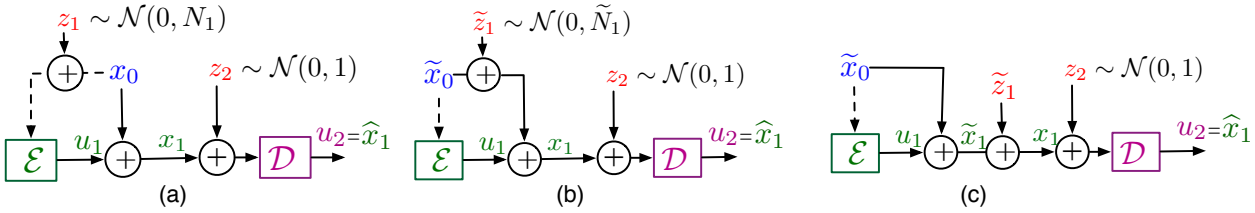


Figure 5.15: These figures show how the signal cancelation problem shown in Fig. 5.14 is equivalent to a problem with noise in the evolution of state $\mathbf{X}_1^m$, instead of noise in the observation at the encoder. From (c), it is clear that the encoder cannot help much in the reconstruction of $\widetilde{\mathbf{Z}}_1^m$ since its observations are independent of $\widetilde{\mathbf{Z}}_1^m$.

In Fig. 5.14, the encoder takes an action based on its observation of $\mathbf{X}_0^m + \mathbf{Z}_1^m$. Define $\widetilde{\mathbf{X}}_0^m := \alpha(\mathbf{X}_0^m + \mathbf{Z}_1^m)$, the MMSE estimate of $\mathbf{X}_0^m$ given $\mathbf{X}_0^m + \mathbf{Z}_1^m$, where $\alpha = \frac{\sigma_0^2}{\sigma_0^2 + N_1}$. Since $\widetilde{\mathbf{X}}_0^m$ can be obtained from $\mathbf{X}_0^m + \mathbf{Z}_1^m$ with an invertible mapping, we can equivalently assume that the encoder observes $\widetilde{\mathbf{X}}_0^m$. The initial state can be written as $\mathbf{X}_0^m = \widetilde{\mathbf{X}}_0^m + \widetilde{\mathbf{Z}}_1^m$, where $\widetilde{\mathbf{X}}_0^m \perp\!\!\!\perp \widetilde{\mathbf{Z}}_1^m$ (orthogonality principle), and $\widetilde{\mathbf{Z}}_1^m \sim \mathcal{N}\left(0, \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1}\right)$. The resulting block diagram (which represents an equivalent problem) is shown in Fig. 5.15(b). By commutativity of addition, we get the equivalent problem with noise $\widetilde{\mathbf{Z}}_1^m$ in state evolution, as shown in Fig. 5.15(c). An intermediate state $\widetilde{\mathbf{X}}_1^m = \widetilde{\mathbf{X}}_0^m + \mathbf{U}_1^m$ is also introduced.

In summary, the equivalent noisy-state evolution problem is the following: the initial state $\widetilde{\mathbf{X}}_0^m \sim \mathcal{N}(0, \widetilde{\sigma}_0^2 \mathbb{I})$ is observed noiselessly by the encoder $\mathcal{E}$, where $\widetilde{\sigma}_0^2 = \frac{\sigma_0^4}{\sigma_0^2+N_1}$. The encoder modifies the state using an input $\mathbf{U}_1^m$, resulting in the system state $\widetilde{\mathbf{X}}_1^m$. State evolution noise $\widetilde{\mathbf{Z}}_1^m \sim \mathcal{N}(0, \widetilde{N}_1 \mathbb{I})$ is added to the state $\widetilde{\mathbf{X}}_1^m$ resulting in state $\mathbf{X}_1^m$. Here, $\widetilde{N}_1 = \frac{\sigma_0^2 N_1}{\sigma_0^2+N_1}$. The objective, as before, is to minimize

$$\overline{\mathcal{J}} = \frac{1}{m} k^2 \mathbb{E}\left[\|\mathbf{U}_1^m\|^2\right] + \frac{1}{m}\mathbb{E}\left[\|\mathbf{X}_1^m - \widehat{\mathbf{X}}_1^m\|^2\right], \tag{5.26}$$

where $\widehat{\mathbf{X}}_1^m$ is the estimate of $\mathbf{X}_1^m$ at the decoder based on noisy observations of $\mathbf{X}_1^m$.

A lower bound on the average costs

A coarse lower bound on the average cost is given in the following.

Theorem 15. *For the noisy version of Witsenhausen's counterexample,*

$$\overline{\mathcal{J}}_{opt} \geq \max\left\{\frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}, \inf_{P\geq 0} k^2 P + \left(\left(\sqrt{\widetilde{\kappa}(P)} - \sqrt{P}\right)^+\right)^2\right\},$$

where $\widetilde{\kappa}(P) = \frac{\widetilde{\sigma}_0^2}{(\widetilde{\sigma}_0+\sqrt{P})^2+1}$, *and* $\widetilde{\sigma}_0^2 = \frac{\sigma_0^4}{\sigma_0^2+N_1}$.

Proof. Consider the equivalent problem of noise in state evolution of Chapter 5.6. A lower bound can be derived as follows.

If the decoder is given side information $\widetilde{\mathbf{X}}_0^m$, it can simulate the encoder, reconstructing $\mathbf{U}_1^m$ perfectly. Thus the decoder only has to estimate $\widetilde{\mathbf{Z}}_1^m$, which is independent of $\widetilde{\mathbf{X}}_0^m$. The resulting $MMSE$ is therefore given by $\frac{\widetilde{N}_1}{\widetilde{N}_1+1} = \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1+\sigma_0^2+N_1}$, yielding the first term in the lower bound.

Alternatively, if side-information $\widetilde{\mathbf{Z}}_1^m$ is given to the decoder, the problem reduces to the vector Witsenhausen counterexample, where the encoder observes the source $\widetilde{\mathbf{X}}_0^m$ noiselessly and there is no noise $\widetilde{\mathbf{Z}}_1^m$ in state evolution. A lower bound can now be obtained from [56, Theorem 1] (using $\widetilde{\sigma}_0$ in place of σ_0), yielding the second term in the lower bound. □

An upper bound on the average costs

Theorem 16. *For the noisy extension of Witsenhausen's counterexample an upper bound on the optimal costs is*

$$\overline{\mathcal{J}}_{opt} \leq \min\left\{\overline{\mathcal{J}}_{\widetilde{ZI}}, \overline{\mathcal{J}}_{\widetilde{ZF}}, \overline{\mathcal{J}}_{\widetilde{VQ}}\right\},$$

where $\overline{\mathcal{J}}_{\widetilde{ZI}} = \frac{\sigma^2}{\sigma^2+1}$, $\overline{\mathcal{J}}_{\widetilde{ZF}} = k^2\frac{\sigma^4}{\sigma^2+N_1} + \frac{\sigma^2 N_1}{\sigma^2+N_1+\sigma^2 N_1}$, *and* $\overline{\mathcal{J}}_{\widetilde{VQ}} \leq k^2(\widetilde{N}_1+1) + \widetilde{N}_1$.

Proof. As usual, we provide three strategies. The strategies are defined on the equivalent problem of noise in the state evolution (of Chapter 5.6).

The first strategy is the Zero-Input ($\widetilde{ZI}$) strategy, where the input $\mathbf{U}_1^m = 0$. The decoder merely estimates $\widetilde{\mathbf{X}}_0^m + \widetilde{\mathbf{Z}}_1^m = \mathbf{X}_0^m$ from the noisy observation $\mathbf{X}_0^m + \mathbf{Z}_2^m$. Since $\mathbf{Z}_2^m \sim \mathcal{N}(0, \mathbb{I})$, the LLSE error is given by

$$MMSE = \frac{\sigma_0^2}{\sigma_0^2 + 1}, \tag{5.27}$$

which is also the attained cost since $P = 0$.

Our second strategy is a Zero-Forcing ($\widetilde{ZF}$) strategy, applied to the equivalent noisy state-evolution problem. The first input forces the state $\widetilde{\mathbf{X}}_0^m$ to zero, requiring an average power of $P = \widetilde{\sigma_0}^2 = \frac{\sigma_0^4}{\sigma_0^2 + N_1}$. The decoder merely performs an LLSE estimation for $\widetilde{\mathbf{Z}}_1^m \sim \mathcal{N}(0, \widetilde{N}_1)$. The $MMSE$ error is therefore given by

$$MMSE_{\widetilde{ZF}} = \frac{\widetilde{N}_1}{\widetilde{N}_1 + 1} = \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1 + \sigma_0^2 N_1}. \tag{5.28}$$

The cost for $\widetilde{ZF}$ is, therefore, $\overline{\mathcal{J}}_{\widetilde{ZF}} = k^2 \frac{\sigma_0^4}{\sigma_0^2 + N_1} + \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1 + \sigma_0^2 N_1}$.

The third strategy is the Vector Quantization ($\widetilde{VQ}$) strategy, but with a difference. The encoder quantizes assuming the two noises ($\widetilde{\mathbf{Z}}_1^m$ and $\mathbf{Z}_2^m$) add up in the observation at the decoder. The decoder thus has an asymptotically error-free estimate of $\widetilde{\mathbf{X}}_1^m$ as long as $P > \widetilde{N}_1 + 1$.

the decoder's input. The resulting MMSE error is the variance of noise $\widetilde{\mathbf{Z}}_1^m$, which is given by $\widetilde{N}_1 = \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1}$. The total cost for this strategy is therefore given by $\overline{\mathcal{J}}_{\widetilde{VQ}} = k^2(\widetilde{N}_1 + 1) + \widetilde{N}_1$.

The upper bound can now be obtained by using the best of $\widetilde{ZI}$, $\widetilde{ZF}$, and $\widetilde{VQ}$ strategies depending on the values of k and σ. □

Proof of approximate asymptotic optimality

Theorem 17 (Approximate asymptotic optimality)**.** *For the noisy version of vector Witsenhausen counterexample (with noise in the observations of the two controllers), in the limit of $m \to \infty$,*

$$\max\left\{\frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}, \inf_{P \geq 0} k^2 P + \left(\left(\sqrt{\widetilde{\kappa}(P)} - \sqrt{P}\right)^+\right)^2\right\}$$
$$\leq \overline{\mathcal{J}}_{opt} \leq \gamma \max\left\{\frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}, \inf_{P \geq 0} k^2 P + \left(\left(\sqrt{\widetilde{\kappa}(P)} - \sqrt{P}\right)^+\right)^2\right\},$$

where $\gamma \leq 41$.

Proof. See Appendix B. □

Chapter 6

Discussions and concluding remarks

Constant-factor approximate-optimality: what is it good for?

In this dissertation, we investigate an intersection of control and communication from an optimization perspective. Our goal is to obtain provable guarantees on the gap from optimality of approximately-optimal strategies. Such provable guarantees have been explored previously in each of the fields that this dissertation has connections with. Theoretical computer science has such guarantees on approximation algorithms [40] for many NP-complete problems. Information theory has explored the concept of degrees of freedom of a wireless channel [122] as a form of asymptotic approximate optimality, and more recently, the deterministic approach [37, 39] has helped solve many problems to within a constant number of bits [29, 38, 39]. At high SNR, a constant additive gap in capacity is equivalent to a multiplicative gap in required power to achieve a specified rate. Thus the constant difference approximation results in information theory can also be interpreted as constant factor results in this high SNR regime. Even in decentralized control, Cogill and Lall [125, 126] provide provably approximately-optimal solutions that also use a constant-factor optimality criterion.

What good is approximate-optimality? The coarsest answer to the question is that in the absence of an optimal solution, it is the next best alternative. While correct, this answer does not help us understand which approximations are good, and which are not. For instance, why do we need provable guarantees on approximate-optimality of solutions? The results of Baglietto, Parisini and Zoppoli [25], Lee, Lau and Ho [26], Lee, Marden and Shamma [27], and Karlsson *et al.* [76] provide us with solutions that are believed to be extremely close to optimal for the Witsenhausen counterexample. In this particular case, provable guarantees provide us the satisfaction that we have not missed any significantly better strategies.

But there is another more powerful motivation to obtain such guarantees: approaches based on approximate optimality often capture the most significant aspects of the problem.

The second-order details may often be left out. In practice, as long as the approximations are not too loose, the second-order details may be of little or no significance. Mathematical models themselves are inaccurate, and one needs to question if the second-order details indeed capture an aspect of the core problem, or merely a detail of the model. For instance, capturing the implicit communication in the counterexample helps design strategies that work for the original Gaussian problem as well as a bounded noise version (in Chapter 4.2), and even an adversarial version where the the bounded noise has no distribution on it (see [63]). It is the same quantization strategies that attain within a constant factor, a stronger justification for the goodness of these strategies even in the presence of modeling errors.

Does constant-factor optimality capture the most significant aspects of the problem when the solutions are not uniform over the entire parameter space, but only a subset of it? For instance, consider the noisy version of Witsenhausen's counterexample (Chapter 5.6), where the first controller has noise of variance N_1 in its observations. This variance is an additional problem parameter. Restricting our space to $N_1 > \epsilon$ (for any $\epsilon > 0$), it can be shown that even linear strategies attain within a constant factor of the optimal (uniform over all k, σ_0), with a factor that depends on ϵ. The Mitter and Sahai's result [18] for Witsenhausen counterexample tells us that this factor must diverge to infinity as $\epsilon \to 0$. But approximate-optimality seems to suggest that linear strategies are good for $N_1 > \epsilon$! It is clear therefore that such restrictions of parameter-space can yield misleading results. Does this mean that we must have a solution that is uniformly approximately-optimal over the entire space? We go back to our problem in Chapter 5.6: notice that for $N_1 > 1$, our results in Appendix B show that linear strategies attain within a reasonably small factor of 2 of the optimal. Indeed, for $N_1 > 1$, this result captures the most significant aspect of the problem: when the noise variance of the first controller is larger than that of the second, there is little incentive to signal. The key to obtaining insightful results within such restrictions is therefore to ensure that the constant factor is reasonably small.

Modeling implicit communication in spontaneous synchronization

Spontaneous synchronization of oscillators has been observed in nature as well as in social systems. For instance, fireflies have been observed to flash synchronously in Malaysian jungles. Initially they flash incoherently, but after a short period of time the whole swarm flashes in unison. Pacemakers in our hearts also fire in a coordinated fashion. Many other examples exist in natural and social systems, and we refer the reader to Strogatz's book [127] for more examples.

Explanation of these phenomena using mathematical models is in general hard because of the intractability of most such formulations [128]. In 1975, Yoshiki Kuramoto developed a

model for identical oscillators with weak coupling [129]. In his model, Kuramoto introduced a coupling term in the differential equations that describe the behavior of the oscillators. In the limit of large number of oscillators, with sufficient strength in the coupling, he was able to show spontaneous synchronization of the operating frequencies and the phases of the coupled oscillators. Most other mathematical models are, however, thought to be hard [128].

The transmission of the frequency and phase information through the substrate can be thought of as communication between the participating agents. Is this communication implicit, or explicit? The communication message is not specified *a priori*, and can thus be affected by actions of an agent. Further, the message is communicated through the system itself. Therefore it appears that the source and channel are both implicit. A natural question therefore is: can an understanding of implicit communication help address intractable models in spontaneous synchronization?

The bigger decentralized-control picture: what other problems need our attention

The intersection of control and communication is an area fertile in intellectually stimulating and practically relevant problems. In this dissertation, we explored the possibility of communication using control actions and provided a program that can address quite a few problems of control of a system under communication constraints. The potential success of the program is suggested by obtaining approximately-optimal solutions to some toy problems in decentralized control, including the celebrated Witsenhausen counterexample. The goal of addressing these toy problems is to develop an understanding of the multiple roles of control actions: control, signaling, source-simplification, and improving state estimability in various static and dynamic settings.

A comprehensive theory of decentralized control will need to have a good understanding of many other issues, some of which we outline here. It is well known [46] that in adaptive control, control actions can play another role: that of helping us learn the system. We therefore need toy problems to help understand role of learning in conjunction these other roles. While a finite-memory controller can be thought of as different controllers connected using rate-limited channels, modeling finite-computational ability of controllers is probably harder, but needs to be understood, possibly in restricted settings.

Some other issues are being considered concurrently in the literature. In this dissertation, one of our main interests is to understand the following question: how do we use communication to facilitate coordination among decentralized agents? Here, communication is not an end in itself, but a means to an end of creating coordination. Our focus is on the possibility of implicit communication: where the source can be simplified before transmission, and the plant itself can be used as a channel. Work of Cuff [130] investigates the same question from a different perspective that measures coordination by the *dependance* that can be created

at different agents. He characterizes the joint distributions that can be achieved given the rate-limitations on the external channels connecting the control agents. This dependance can be used, for instance, to generate mixed strategies in cooperative games.

In this dissertation, we assume that the sensor noise at each controller is fixed (except for the formulation in Chapter 5.5). What happens when sensing itself is expensive, and improved sensing comes at a higher cost? Weissman *et al.* [131–133] consider the cost of sensing and its tradeoffs with rate of communication. Improved sensing can increase the channel capacity, but comes at an increased cost. We need to understand this issue in a control setting so that one can understand how to divide resources among sensing and communication in order to minimize control costs.

Just as sensing can be expensive, so also can be the computation of the control input. For instance, for agents that operate at short distances from each other, the cost of communication can be comparable to the cost of computation [82]. It is unclear what problems can help us understand control, communication and computation together. We suggest possible formulations in [55], but the question needs a deeper investigation.

Our proposed strategies assume that each agent knows the strategies of other agents. For instance, if the approximately-optimal strategies are known to be based on quantization, we assume that the quantization bin-size for each controller is known at every other controller. How can this information be communicated? In particular, what if there is no established protocol for the controllers to talk to each other? Recent work of Juba and Sudan [134] develops some understanding of this extremely difficult problem. The hope is that in restricted settings, computationally efficient methods of arriving at agreement on strategies will be possible.

Appendix A

Proofs for Witsenhausen's counterexample

A.1 Nonconvexity of the counterexample in (γ_1, γ_2).

Consider two strategies, $\gamma^{(a)}$ and $\gamma^{(b)}$. The first strategy is $\gamma_1^{(a)} = |x_0|$, and $\gamma_2^{(a)} = \frac{4\sigma_0^2}{4\sigma_0^2+1} y$. For the second strategy, we use $\gamma_1^{(b)} = -|x_0|$, and $\gamma_2^{(b)} = \frac{4\sigma_0^2}{4\sigma_0^2+1} y$. To check for convexity, we will consider a convex combination $\gamma^{(c)} = 0.5\gamma^{(a)} + 0.5\gamma^{(b)}$ of these strategies and check if the resulting strategy has lower costs.

By the symmetry of the counterexample about zero, the attained total cost using $\gamma^{(a)}$ and $\gamma^{(b)}$ is the same. Focusing on $\gamma^{(a)}$, the first-stage cost is $k^2\mathbb{E}\left[|x_0|^2\right] = k^2\sigma_0^2$. The second stage cost needs to be understood in two (equally-likely) cases: conditioned on $X_0 < 0$, the cost is $\mathbb{E}\left[\left(Z - \frac{4\sigma_0^2}{4\sigma_0^2+1} Z\right)^2\right] = \mathbb{E}\left[\left(\frac{Z}{4\sigma_0^2+1}\right)^2\right] = \frac{1}{(4\sigma_0^2+1)^2}$ because $\mathbb{E}\left[Z^2\right] = 1$. Conditioned on $X_0 > 0$, the second-stage cost is

$$\begin{aligned}
\mathbb{E}\left[\left(2X_0 - \frac{4\sigma_0^2}{4\sigma_0^2+1}(2X_0 + Z)\right)^2 \middle| X_0 > 0\right] &= \mathbb{E}\left[\left(\frac{2X_0}{4\sigma_0^2+1} + \frac{4\sigma_0^2 Z}{4\sigma_0^2+1}\right)^2\right] \\
&= \frac{4\sigma_0^2}{(4\sigma_0^2+1)^2} + \left(\frac{4\sigma_0^2}{4\sigma_0^2+1}\right)^2 \\
&= \frac{4\sigma_0^2(4\sigma_0^2+1)}{(4\sigma_0^2+1)^2} = \frac{4\sigma_0^2}{4\sigma_0^2+1}.
\end{aligned}$$

The total cost for $\gamma^{(a)}$ (and by symmetry so also for $\gamma^{(b)}$) is, therefore,

$$\begin{aligned}\overline{\mathcal{J}}(\gamma^{(i)}) &= k^2\sigma_0^2 + \frac{1}{2}\frac{4\sigma_0^2}{4\sigma_0^2+1} + \frac{1}{2}\frac{1}{(4\sigma_0^2+1)^2} \\ &= k^2\sigma_0^2 + \frac{16\sigma_0^4+4\sigma_0^2+1}{2(4\sigma_0^2+1)^2}. \end{aligned} \tag{A.1}$$

Now consider the third strategy, $\gamma^{(c)} = 0.5\gamma^{(a)} + 0.5\gamma^{(b)}$, a convex combination of the first two strategies. If the counterexample were convex, then $\overline{\mathcal{J}}(\gamma^{(c)})$ would be no larger than $0.5\overline{\mathcal{J}}(\gamma^{(a)}) + 0.5\overline{\mathcal{J}}(\gamma^{(b)}) = \overline{\mathcal{J}}(\gamma^{(a)})$.

Now, $\gamma_1^{(c)} = 0$, and $\gamma^{(c)}(2) = \frac{4\sigma_0^2}{4\sigma_0^2+1}Y$. The total cost for this strategy is

$$\begin{aligned}\overline{\mathcal{J}}(\gamma^{(c)}) &= k^2 \times 0 + \mathbb{E}\left[X_0 - \frac{4\sigma_0^2}{4\sigma_0^2+1}(X_0+Z)\right] \\ &= \sigma_0^2\frac{1}{(4\sigma_0^2+1)^2} + \frac{(4\sigma_0^2)^2}{(4\sigma_0^2+1)^2} \\ &= \frac{16\sigma_0^4+\sigma_0^2}{(4\sigma_0^2+1)^2}. \end{aligned} \tag{A.2}$$

Now let us compare costs for $\gamma^{(a)}$ and $\gamma^{(b)}$ (see (A.1)) with the cost for $\gamma^{(c)}$ (see (A.2)). Choosing $k^2 = 0.01$, and $\sigma_0^2 = 10$, the cost $\overline{\mathcal{J}}(\gamma^{(a)}) = \overline{\mathcal{J}}(\gamma^{(b)}) \approx 0.59$, whereas the cost $\overline{\mathcal{J}}(\gamma^{(c)}) = \overline{\mathcal{J}}\left(0.5\gamma^{(a)} + 0.5\gamma^{(b)}\right) \approx 0.95$. That is, $\overline{\mathcal{J}}\left(0.5\gamma^{(a)} + 0.5\gamma^{(b)}\right) \geq 0.5\overline{\mathcal{J}}\left(\gamma^{(a)}\right) + 0.5\overline{\mathcal{J}}\left(\gamma^{(b)}\right)$. Clearly for the counterexample, the objective function (*i.e.* the total cost) is not convex in the choice of strategy $\gamma = (\gamma_1, \gamma_2)$.

A.2 Derivation of Lemma 1

Proof. Using the triangle inequality on Euclidian distance,

$$\sqrt{d(B,C)} \geq \sqrt{d(A,C)} - \sqrt{d(A,B)}. \tag{A.3}$$

Similarly,

$$\sqrt{d(B,C)} \geq \sqrt{d(A,B)} - \sqrt{d(A,C)}. \tag{A.4}$$

Thus,

$$\sqrt{d(B,C)} \geq |\sqrt{d(A,C)} - \sqrt{d(A,B)}|, \tag{A.5}$$

Squaring both sides,

$$d(B,C) \geq d(A,C) + d(A,B) - 2\sqrt{d(A,C)}\sqrt{d(A,B)}. \tag{A.6}$$

Taking the expectation on both sides,

$$\mathbb{E}\left[d(B,C)\right] \;\geq\; \mathbb{E}\left[d(A,C)\right] + \mathbb{E}\left[d(A,B)\right] - 2\mathbb{E}\left[\sqrt{d(A,C)}\sqrt{d(A,B)}\right].$$

Now, using the Cauchy-Schwartz inequality [135, Pg. 13],

$$\left(\mathbb{E}\left[\sqrt{d(A,C)}\sqrt{d(A,B)}\right]\right)^2 \leq \mathbb{E}\left[d(A,C)\right]\mathbb{E}\left[d(A,B)\right]. \tag{A.7}$$

Using (A.7) and (A.7),

$$\begin{aligned}\mathbb{E}\left[d(B,C)\right] &\geq \mathbb{E}\left[d(A,C)\right] + \mathbb{E}\left[d(A,B)\right] - 2\sqrt{\mathbb{E}\left[d(A,C)\right]}\sqrt{\mathbb{E}\left[d(A,B)\right]} \\ &= \left(\sqrt{\mathbb{E}\left[d(A,C)\right]} - \sqrt{\mathbb{E}\left[d(A,B)\right]}\right)^2.\end{aligned}$$

Taking square-roots on both the sides completes the proof. □

A.3 Proof of Theorem 3: bounded ratios for the uniform-noise counterexample

We consider two cases:

Case 1: $\sigma_0^2 < 1$.

If $P > \frac{\sigma_0^2 2^{2h(Z)}}{200}$, using the zero-forcing upper bound of $k^2\sigma_0^2$, the ratio is smaller than $\frac{200}{2^{2h(Z)}}$.

If $P \leq \frac{\sigma_0^2 2^{2h(Z)}}{200}$,

$$\begin{aligned}\kappa(P) &= \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e\left(\left(\sigma_0 + \sqrt{P}\right)^2 + 1\right)} \\ &\overset{\sigma_0^2\leq 1, P\leq\frac{\sigma_0^2 2^{2h(Z)}}{200}}{\geq} \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e\left((1+\sqrt{\frac{2^{2h(Z)}}{200}})^2 + 1\right)} \\ &\overset{(a)}{\geq} \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e\left((1+\sqrt{\frac{\pi e}{100}})^2 + 1\right)} \\ &\geq \frac{\sigma_0^2 2^{2h(Z)}}{46},\end{aligned}$$

where (a) follows from the fact that $h(Z) \leq \frac{1}{2}\log_2(2\pi e)$, the differential entropy for the

$\mathcal{N}(0,1)$ random variable. Thus,

$$\begin{aligned}\left(\left(\kappa-\sqrt{P}\right)^{+}\right)^{2} &\geq \sigma_0^2 2^{2h(Z)}\left(\frac{1}{\sqrt{46}}-\frac{1}{\sqrt{200}}\right)^{2}\\ &\geq \frac{\sigma_0^2 2^{2h(Z)}}{173} > \frac{\sigma_0^2 2^{2h(Z)}}{200}.\end{aligned}$$

Using the zero-input upper bound of $\frac{\sigma_0^2}{\sigma_0^2+1} < 1$, the ratio in this case is bounded by $\frac{200}{2^{2h(Z)}}$.

Case 2: $\sigma_0^2 \geq 1$.

If $P > \frac{2^{2h(Z)}}{200}$, using the upper bound of k^2a^2, the ratio of upper and lower bounds is smaller than $\frac{k^2a^2}{k^2\frac{2^{2h(Z)}}{200}} = \frac{200a^2}{2^{2h(Z)}}$.

If $P \leq \frac{2^{2h(Z)}}{200} \leq \frac{2\pi e}{200}$ (again, because Gaussian distribution maximizes the differential entropy for given variance),

$$\begin{aligned}\kappa(P) &= \frac{\sigma_0^2 2^{2h(Z)}}{2\pi e\left(\left(\sigma_0+\sqrt{P}\right)^2+1\right)}\\ &\geq \frac{2^{2h(Z)}}{2\pi e\left((1+\sqrt{P})^2+1\right)}\\ &\geq \frac{2^{2h(Z)}}{2\pi e\left((1+\frac{\pi e}{100})^2+1\right)} \geq \frac{2^{2h(Z)}}{46}.\end{aligned}$$

Thus, the following lower bound holds for the $MMSE$

$$MMSE \geq 2^{2h(Z)}\left(\frac{1}{46}-\frac{1}{200}\right)^2 \geq 2^{2h(Z)}0.0058.$$

Using the zero-input upper bound, the ratio is smaller than $\frac{1}{2^{2h(Z)}0.0058} < \frac{173}{2^{2h(Z)}}$

Using the fact that $a > 1$, we get the theorem.

A.4 Required P for error probability converging to zero using the vector quantization scheme

We now derive the required power P that satisfies (4.16) and (4.17). Let ξ satisfy $\frac{1}{2}\log_2(1+\xi) = \delta$. Then (4.16) and (4.17) are satisfied whenever

$$\begin{aligned}
&\frac{1}{2}\log_2\left(\frac{\sigma_z^2+\sigma_0^2-P}{\sigma_z^2}\right) = \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{P}\right) + \frac{1}{2}\log_2(1+\xi),\\
&\text{i.e. } \frac{\sigma_z^2+\sigma_0^2-P}{\sigma_z^2} = \frac{\sigma_0^2}{P}(1+\xi)\\
&\text{i.e. } P^2 - P(\sigma_0^2+\sigma_z^2) + \sigma_z^2\sigma_0^2(1+\xi) = 0. \qquad \text{(A.8)}
\end{aligned}$$

Now, some algebra reveals that (A.8) is satisfied if

$$\begin{aligned}
P &= \frac{\sigma_0^2+\sigma_z^2-\sqrt{(\sigma_0^2-\sigma_z^2)^2 - 4\sigma_0^2\sigma_z^2\xi^2}}{2}\\
&= \sigma_0^2\left(\frac{1-\sqrt{1-\frac{4\sigma_0^2\sigma_z^2\xi^2}{(\sigma_0^2-\sigma_z^2)^2}}}{2}\right) + \sigma_z^2\left(\frac{1+\sqrt{1-\frac{4\sigma_0^2\sigma_z^2\xi^2}{(\sigma_0^2-\sigma_z^2)^2}}}{2}\right),
\end{aligned}$$

which is along the line segment joining σ_z^2 and σ_0^2, and is hence smaller than σ_0^2. For this P to exist, $\xi < \frac{\sigma_0^2-\sigma_z^2}{2\sigma_0\sigma_z}$, and therefore $\delta < \frac{1}{2}\log_2\left(1+\frac{\sigma_0^2-\sigma_z^2}{2\sigma_0\sigma_z}\right)$. Also, in the limit $\xi \to 0$ (or equivalently, $\delta \to 0$), P converges to $\sigma_z^2 = 1$.

A.5 Proof of bounded ratios for the asymptotic vector Witsenhausen counterexample

The performance of the scheme that zero-forces $\mathbf{x}_0^m$ and the JSCC scheme is identical for $\sigma_0^2 = 1$, as is evident from Fig. 4.6. Therefore, we consider two different cases: $\sigma_0^2 \leq 1$ and $\sigma_0^2 \geq 1$. In either case, we show that the ratio is bounded by 11. The result can be tightened by a more detailed analysis by dividing the (k, σ_0^2) space into finer partitions. However, we do not present the detailed analysis here for ease of exposition.

Region 1: $\sigma_0^2 \leq 1$.

We consider the upper bound as the minimum of $k^2\sigma_0^2$ and $\frac{\sigma_0^2}{\sigma_0^2+1}$. Consider the lower bound

$$\overline{\mathcal{J}} \geq \min_{P\geq 0} k^2 P + \left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2. \qquad \text{(A.9)}$$

Now if the optimizing power P is greater than $\sigma_0^2/11$, then the first term of the lower bound is greater than $k^2\sigma_0^2/11$. Thus the ratio of the upper bound $k^2\sigma_0^2$ and the lower bound is smaller than 11.

If the optimizing $P \leq \frac{\sigma_0^2}{11}$,

$$\begin{aligned}
\kappa(P) &= \frac{\sigma_0^2}{\left(\sigma_0 + \sqrt{P}\right)^2 + 1} \\
&\geq \frac{\sigma_0^2}{\left(\sigma_0 + \frac{\sigma_0}{\sqrt{11}}\right)^2 + 1} \\
&\overset{(\sigma_0^2 \leq 1)}{\geq} \frac{\sigma_0^2}{\left(1 + \frac{1}{\sqrt{11}}\right)^2 + 1} \geq 0.37\sigma_0^2
\end{aligned}$$

which is greater than $\sigma_0^2/11 > P$. Thus,

$$\begin{aligned}
\left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2 &\geq \left(\sqrt{0.37\sigma_0^2} - \sqrt{\frac{\sigma_0^2}{11}}\right)^2 \\
&> 0.094\sigma_0^2 > \frac{\sigma_0^2}{11}.
\end{aligned}$$

The lower bound is no smaller than $\left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2$. Thus, even for $P \leq \frac{\sigma_0^2}{11}$ the ratio of the upper bound $\frac{\sigma_0^2}{\sigma_0^2+1}$ and the lower bound is smaller than 11.

Region 2: $\sigma_0^2 \geq 1$.

The upper bound relevant here is the minimum of k^2 and $\frac{\sigma_0^2}{\sigma_0^2+1}$. Again, looking at (A.9), if $P > \frac{1}{11}$, the ratio of the upper bound k^2 to the lower bound is no more than 11.

Now, if $P \leq \frac{1}{11}$,

$$\kappa(P) \geq \frac{\sigma_0^2}{(\sigma_0 + 1/\sqrt{11})^2 + 1}.$$

Therefore,

$$\left(\left(\sqrt{\kappa(P)} - \sqrt{P}\right)^+\right)^2 = \left(\left(\sqrt{\frac{\sigma_0^2}{\left(\sigma_0 + \frac{1}{\sqrt{11}}\right)^2 + 1}} - \frac{1}{\sqrt{11}}\right)^+\right)^2.$$

For $\sigma_0^2 \geq 1$, the first term on the RHS attains its minima at $\sigma_0^2 = 1$. Evaluated at this point,

the term is larger than $\frac{1}{\sqrt{11}}$. Therefore, a bound on the ratio for $P < \frac{1}{11}$ is

$$\begin{aligned}\frac{\sigma_0^2/(\sigma_0^2+1)}{\left(\sqrt{\frac{\sigma_0^2}{\left(\sigma_0+\frac{1}{\sqrt{11}}\right)^2+1}}-\frac{1}{\sqrt{11}}\right)^2} &\leq \frac{1}{\left(\sqrt{\frac{\sigma_0^2}{\left(\sigma_0+\frac{1}{\sqrt{11}}\right)^2+1}}-\frac{1}{\sqrt{11}}\right)^2} \\ &\leq \frac{1}{\left(\sqrt{\frac{1}{\left(1+\frac{1}{\sqrt{11}}\right)^2+1}}-\frac{1}{\sqrt{11}}\right)^2} \\ &\approx 10.56 < 11.\end{aligned}$$

Thus, for $\sigma_0^2 \geq 1$, the ratio is bounded by 11 as well. Therefore, γ_1 and γ_2 are both smaller than 11.

A.6 Dirty-paper coding and tightness at $MMSE = 0$

We summarize the strategy briefly, and refer the interested reader to [56] for a detailed description and analysis of the achievability. The encoder divides its input into two parts $\mathbf{U}_{lin}^m$ and $\mathbf{U}_{dpc}^m$ of powers P_{lin} and P_{dpc} respectively, such that $P = P_{lin} + P_{dpc}$ (by construction, $\mathbf{U}_{lin}^m$ and $\mathbf{U}_{dpc}^m$ turn out to be orthogonal in the limit). We refer to P_{lin} as the *linear* part of the power, and P_{dpc} the *dirty-paper coding* part of the power. The linear part is used to scale the host signal down by a factor β (using $\mathbf{U}_{lin}^m = -\beta\mathbf{X}_0^m$) so that the scaled down host signal has variance $\widetilde{\sigma}_0^2 = \sigma_0^2(1-\beta)^2$, where $\beta^2\sigma_0^2 = P_{lin}$. Using the remaining P_{dpc} power, the transmitter dirty-paper codes against the scaled-down host signal $(1-\beta)\mathbf{X}_0^m$ with the DPC parameter α [77] allowed to be arbitrary (unlike in [77], where it is eventually chosen to be the MMSE parameter). A plain DPC strategy achieves the following rate [77, Eq. (6)]

$$R = \frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+1)}{P\sigma_0^2(1-\alpha)^2 + P + \alpha^2\sigma_0^2}\right), \tag{A.10}$$

The strategy recovers $\mathbf{U}^m + \alpha\mathbf{X}_0^m$ at the decoder with high probability. Because we also have a linear part here, the achieved rate is

$$R = \frac{1}{2}\log_2\left(\frac{P_{dpc}(P_{dpc}+\widetilde{\sigma}_0^2+1)}{P_{dpc}\widetilde{\sigma}_0^2(1-\alpha)^2 + P_{dpc} + \alpha^2\widetilde{\sigma}_0^2}\right). \tag{A.11}$$

The decoder now decodes the codeword $\mathbf{U}_{dpc}^m + \alpha(1-\beta)\mathbf{X}_0^m$. It then performs an MMSE estimation for estimating $\mathbf{X}^m = \mathbf{X}_0^m + \mathbf{U}^m = (1-\beta)\mathbf{X}_0^m + \mathbf{U}_{dpc}^m$ using the channel output $\mathbf{Y}_2^m = (1-\beta)\mathbf{X}_0^m + \mathbf{U}_{dpc}^m + \mathbf{Z}^m$ and the decoded codeword $\alpha(1-\beta)\mathbf{X}_0^m + \mathbf{U}_{dpc}^m$. The obtained $MMSE$ can now be minimized over the choice of α and β under the constraint (A.11).

Corollary 2. *For a given power P, a combination of linear and DPC-based strategies* ***achieve the maximum possible rate*** $C(P)$ *in the perfect recovery limit $MMSE(P,R) = 0$, where $C(P)$ is given by*

$$C(P) = \sup_{\sigma_{X_0,U_1} \in [-\sigma_0\sqrt{P},0]} \frac{1}{2}\log_2\left(\frac{(P\sigma_0^2 - \sigma_{X_0,U_1}^2)(1+\sigma_0^2+P+2\sigma_{X_0,U_1})}{\sigma_0^2(\sigma_0^2+P+2\sigma_{X_0,U_1})}\right). \tag{A.12}$$

Proof. See Appendix A.8. The special case of $R = 0$ shows that DPC-based strategies are asymptotically-optimal for perfect-recovery limit for Witsenhausen's counterexample. □

DPC strategy performs better than vector-quantization

For $\alpha = 1$, P needs to satisfy $C(1,P) > \epsilon$, where

$$C(1,P) = \frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+\sigma_z^2)}{(P+\sigma_0^2)\sigma_z^2}\right). \tag{A.13}$$

Let ξ be such that $\epsilon = \frac{1}{2}\log_2(1+\xi)$. Then,

$$\frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+\sigma_z^2)}{(P+\sigma_0^2)\sigma_z^2}\right) = \frac{1}{2}\log_2(1+\xi)$$
$$\text{i.e. } P^2 + (\sigma_0^2 - \xi\sigma_z^2)P - (1+\xi)\sigma_0^2\sigma_z^2 = 0$$

Taking the positive root of the quadratic equation,

$$P = (\sigma_0^2 - \xi\sigma_z^2)\frac{\sqrt{1+\frac{4(1+\xi)^2\sigma_z^2\sigma_0^2}{(\sigma_0^2-\xi\sigma_z^2)^2}}-1}{2}. \tag{A.14}$$

Now letting ϵ go to zero (and thus $\xi \to 0$) by increasing m to infinity, the required P approaches $\sigma_0^2\frac{\sqrt{1+\frac{4\sigma_z^2}{\sigma_0^2}}-1}{2}$. The asymptotic expected cost for the scheme is, therefore, $k^2\sigma_0^2\frac{\sqrt{1+\frac{4\sigma_z^2}{\sigma_0^2}}-1}{2}$. This expression turns out to be an increasing function in σ_0^2 which is bounded above by $k^2\sigma_z^2$, the cost for the JSCC scheme. Thus even in the special case of $\alpha = 1$, the DPC scheme asymptotically outperforms the VQ scheme.

Costs for DPC with $\alpha \neq 1$

The total asymptotic costs (assuming no errors in decoding the auxiliary codeword) are given by

$$k^2(P + (1+|\alpha|)^2) + MMSE(\alpha, P), \tag{A.15}$$

where P satisfies

$$C(\alpha, P) = I(v; y_2) - I(v; x_0) = \frac{1}{2}\log_2\left(\frac{P(P+\sigma_0^2+\sigma_z^2)}{P\sigma_0^2(1-\alpha)^2+\sigma_z^2(P+\alpha^2\sigma_0^2)}\right) = \epsilon. \quad (A.16)$$

Concentrating on the case of interest of $\epsilon \to 0$ by letting $m \to \infty$, the condition (A.16) is equivalent to

$$P(P+\sigma_0^2+\sigma_z^2) = P\sigma_0^2(1-\alpha)^2+\sigma_z^2(P+\alpha^2\sigma_0^2).$$

Taking the positive root,

$$P = \frac{\sqrt{\sigma_0^2\alpha(2-\alpha)}}{2}\left(\sqrt{1+\frac{4\sigma_z^2}{\sigma_0^2(2-\alpha)^2}}-1\right) \quad (A.17)$$

By letting $m \to \infty$, we can have $\epsilon_1 \to 0$ and also $\epsilon \to 0$. Optimizing the total cost over α, the asymptotic total cost achieved is

$$\min_{\alpha} k^2 P + MMSE(\alpha, P), \quad (A.18)$$

where P is given by (A.17).

A.7 Tighter outer bound for the vector Witsenhausen problem: proof of Theorems 6 and 13

Achievability: a combination of linear and DPC-based strategies

The combination of linear and DPC-based strategies of Chapter 4.3.3 recovers $\mathbf{U}_{dpc}^m + \alpha(1-\beta)\mathbf{X}_0^m$ at the decoder with high probability. In order to perfectly recover $\mathbf{X}_1^m = (1-\beta)\mathbf{X}_0^m + \mathbf{U}_{dpc}^m$, we can use $\alpha = 1$, and hence the strategy achieves a rate of (from (A.11))

$$R_{ach} = \sup_{P_{lin}, P_{dpc}: P = P_{lin}+P_{dpc}} \frac{1}{2}\log_2\left(\frac{P_{dpc}(P_{dpc}+\widetilde{\sigma}_0^2+1)}{P_{dpc}+\widetilde{\sigma}_0^2}\right), \quad (A.19)$$

where we take a supremum over P_{lin}, P_{dpc} such that they sum up to P. Let $\sigma_{X_0,U_1} = -\sigma_0\sqrt{P_{lin}}$ (note that as P_{lin} varies from 0 to P, σ_{X_0,U_1} varies from 0 to $-\sigma_0\sqrt{P}$). Then, $P_{dpc} = P - \frac{\sigma_{X_0,U_1}^2}{\sigma_0^2}$, and $P_{dpc} + \widetilde{\sigma}_0^2 = P_{dpc} + \sigma_0^2 + P_{lin} - 2\sigma_0\sqrt{P_{lin}} = P + \sigma_0^2 + 2\sigma_{X_0,U_1}$. Thus,

$$R_{ach} = \sup_{\sigma_{X_0,U_1}\in[-\sigma_0\sqrt{P},0]} \frac{1}{2}\log_2\left(\frac{\left(P-\frac{\sigma_{X_0,U_1}^2}{\sigma_0^2}\right)(P+\sigma_0^2+2\sigma_{X_0,U_1}+1)}{P+\sigma_0^2+2\sigma_{X_0,U_1}}\right). \quad (A.20)$$

Simple algebra shows that this expression matches that in Corollary 2.

Proof. [Of Theorem 6]

For any chosen pair of encoding map $\mathcal{E}_m$ and decoding map $\mathcal{D}_m$, there is a Markov chain $\mathbf{X}_0^m \to \mathbf{X}_1^m \to \mathbf{Y}_2^m \to \widehat{\mathbf{X}}_1^m$. Using the data-processing inequality

$$I(\mathbf{X}_0^m; \widehat{\mathbf{X}}_1^m) \leq I(\mathbf{X}_1^m; \mathbf{Y}_2^m). \tag{A.21}$$

The terms in the inequality can be bounded by single letter expressions as follows. Define Q as a random variable uniformly distributed over $\{1, 2, \ldots, m\}$. Define $X_0 = X_{0,Q}$, $U = U_Q$, $X_1 = X_{1,Q}$, $Z = Z_Q$, $Y = Y_Q$ and $\widehat{X}_1 = \widehat{X}_{1,Q}$. Then,

$$\begin{aligned}
I(\mathbf{X}_1^m; \mathbf{Y}_2^m) &= h(\mathbf{Y}_2^m) - h(\mathbf{Y}_2^m|\mathbf{X}_1^m) \\
&\overset{(a)}{\leq} \sum_i h(Y_{2,i}) - h(\mathbf{Y}_2^m|\mathbf{X}_1^m) \\
&= \sum_i h(Y_{2,i}) - h(Y_{2,i}|X_{1,i}) \\
&= \sum_i I(X_{1,i}; Y_{2,i}) \\
&= mI(X_1; Y_2|Q) \\
&= m\left(h(Y_2|Q) - h(Y_2|X_1, Q)\right) \\
&\leq m\left(h(Y_2) - h(Y_2|X_1, Q)\right) \\
&\overset{(b)}{=} m\left(h(Y_2) - h(Y_2|X_1)\right) = mI(X_1; Y_2),
\end{aligned} \tag{A.22}$$

where (a) follows from an application of the chain-rule for entropy followed by using the fact that conditioning reduces entropy, and (b) follows from the observation that the additive noise Z_i is iid across time, and independent of the input $X_{1,i}$ (thus $Y \perp\!\!\!\perp Q|X$). Also,

$$\begin{aligned}
I(\mathbf{X}_0^m; \widehat{\mathbf{X}}_1^m) &= h(\mathbf{X}_0^m) - h(\mathbf{X}_0^m|\widehat{\mathbf{X}}_1^m) \\
&= \sum_i h(X_{0,i}) - h(\mathbf{X}_0^m|\widehat{\mathbf{X}}_1^m) \\
&\overset{(a)}{\geq} \sum_i \Big(h(X_{0,i}) - h(X_{0,i}|\widehat{X}_{1,i})\Big) \\
&= \sum_i I(X_{0,i}; \widehat{X}_{1,i}) = mI(X_0; \widehat{X}_1|Q) \\
&= m\Big(h(X_0|Q) - h(X_0|\widehat{X}_1, Q)\Big) \\
&\overset{(b)}{\geq} m\Big(h(X_0) - h(X_0|\widehat{X}_1)\Big) = mI(X_0; \widehat{X}_1),
\end{aligned} \tag{A.23}$$

where (a) and (b) again follow from the fact that conditioning reduces entropy, and (b) also uses the observation that since $X_{0,i}$ are iid, X_0, $X_{0,i}$, and $X_0|Q = q$ are distributed identically.

Now, using (A.21), (A.22) and (A.23),

$$mI(X_0;\widehat{X}) \leq I(\mathbf{X}_0^m;\widehat{\mathbf{X}}_1^m) \leq I(\mathbf{X}_1^m;\mathbf{Y}^m) \leq mI(X_1;Y). \tag{A.24}$$

Also observe that from the definitions of X_0, X_1, $\widehat{X}_1$ and Y, $\mathbb{E}\left[d(\mathbf{X}_0^m,\mathbf{X}_1^m)\right] = \mathbb{E}\left[d(X_0,X_1)\right]$, and $\mathbb{E}\left[d(\mathbf{X}_1^m,\widehat{\mathbf{X}}_1^m)\right] = \mathbb{E}\left[d(X_1,\widehat{X}_1)\right]$. Using the Cauchy-Schwartz inequality, the correlation $\sigma_{X_0,U_1} = \mathbb{E}\left[X_0U_1\right]$ must satisfy the following constraint,

$$|\sigma_{X_0,U_1}| = |\mathbb{E}\left[X_0U_1\right]| \leq \sqrt{\mathbb{E}\left[X_0^2\right]}\sqrt{\mathbb{E}\left[U_1^2\right]} \leq \sigma_0\sqrt{P}. \tag{A.25}$$

Also,

$$\mathbb{E}\left[X_1^2\right] = \mathbb{E}\left[(X_0+U_1)^2\right] = \sigma_0^2 + P + 2\sigma_{X_0,U_1}. \tag{A.26}$$

Since $Z = Y - X_1 \perp\!\!\!\perp X_1$, and a Gaussian input distribution maximizes the mutual information across an average-power-constrained AWGN channel,

$$I(X_1;Y) \leq \frac{1}{2}\log_2\left(1 + \frac{P+\sigma_0^2+2\sigma_{X_0,U_1}}{1}\right). \tag{A.27}$$

$$\begin{aligned} I(X_0;\widehat{X}_1) &= h(X_0) - h(X_0|\widehat{X}_1) \\ &= h(X_0) - h(X_0-\gamma\widehat{X}|\widehat{X}_1)\ \forall\gamma \\ &\overset{(a)}{\geq} h(X_0) - h(X_0-\gamma\widehat{X}_1) \\ &= \frac{1}{2}\log_2\left(2\pi e\sigma_0^2\right) - h(X_0-\gamma\widehat{X}_1), \end{aligned} \tag{A.28}$$

where (a) follows from the fact that conditioning reduces entropy. Also note here that the result holds for any $\gamma > 0$, and in particular, γ can depend on σ_{X_0,U_1}. Now,

$$\begin{aligned} h(X_0-\gamma\widehat{X}_1) &= h(X_0-\gamma(\widehat{X}_1-X_1)-\gamma X_1) \\ &= h\left(X_0-\gamma(\widehat{X}_1-X_1)-\gamma X_0-\gamma U\right) \\ &= h\left((1-\gamma)X_0-\gamma U_1-\gamma(\widehat{X}_1-X_1)\right). \end{aligned} \tag{A.29}$$

The second moment of a sum of two random variables A and B can be bounded as follows

$$\begin{aligned} \mathbb{E}\left[(A+B)^2\right] &= \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + 2\mathbb{E}\left[AB\right] \\ &\overset{\text{Cauchy-Schwartz ineq.}}{\leq} \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + 2\sqrt{\mathbb{E}\left[A^2\right]}\sqrt{\mathbb{E}\left[B^2\right]} \\ &= \left(\sqrt{\mathbb{E}\left[A^2\right]} + \sqrt{\mathbb{E}\left[B^2\right]}\right)^2, \end{aligned} \tag{A.30}$$

with equality when A and B are aligned, i.e. $A = \lambda B$ for some $\lambda \in \mathbb{R}$. For the random variable under consideration in (A.29), choosing $A = (1-\gamma)X_0 - \gamma U_1$, and $B = -\gamma(\widehat{X}_1 - X_1)$ in (A.30)

$$\begin{aligned} &\mathbb{E}\left[\Big((1-\gamma)X_0 - \gamma U_1 - \gamma(\widehat{X}_1 - X_1)\Big)^2\right] \\ &\leq \left(\sqrt{(1-\gamma)^2\sigma_0^2 + \gamma^2 P - 2\gamma(1-\gamma)\sigma_{X_0,U_1}} + \gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1 - X_1)^2\right]}\right)^2. \end{aligned} \tag{A.31}$$

Equality is obtained by aligning[1] $X_1 - \widehat{X}_1$ with $(1-\gamma)X_0 - \gamma U_1$. Thus,

$$\begin{aligned} & I(X_0; \widehat{X}_1) \\ \geq\ & \frac{1}{2}\log_2\left(2\pi e \sigma_0^2\right) - h(X_0 - \gamma\widehat{X}_1) \\ \geq\ & \frac{1}{2}\log_2\left(\frac{\sigma_0^2}{\left(\sqrt{(1-\gamma)^2\sigma_0^2 + \gamma^2 P - 2\gamma(1-\gamma)\sigma_{X_0,U_1}} + \gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1 - X_1)^2\right]}\right)^2}\right) \end{aligned} \tag{A.32}$$

Using (A.24), $I(X_0; \widehat{X}_1) \leq I(X_1; Y)$. Using the lower bound on $I(X_0; \widehat{X}_1)$ from (A.32) and the upper bound on $I(X_1; Y)$ from (A.27), we get

$$\begin{aligned} &\frac{1}{2}\log_2\left(\frac{\sigma_0^2}{\left(\sqrt{(1-\gamma)^2\sigma_0^2 + \gamma^2 P - 2\gamma(1-\gamma)\sigma_{X_0,U_1}} + \gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1 - X_1)^2\right]}\right)^2}\right) \\ &\leq \frac{1}{2}\log_2\left(1 + \frac{P + \sigma_0^2 + 2\sigma_{X_0,U_1}}{1}\right), \end{aligned}$$

[1] In general, since $\widehat{\mathbf{X}}_1^m$ is a function of $\mathbf{Y}_2^m$, this alignment is not actually possible when the recovery of $\mathbf{X}_1^m$ is not exact. The derived bound is therefore loose.

for the choice of $\mathcal{E}_m$ and $\mathcal{D}_m$. Since $\log_2(\cdot)$ is a monotonically increasing function,

$$\begin{aligned}
&\frac{\sigma_0^2}{\left(\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}+\gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]}\right)^2}\\
&\leq 1+P+\sigma_0^2+2\sigma_{X_0,U_1}\\
&\text{i.e.}\quad \left(\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}+\gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]}\right)^2\\
&\geq \frac{\sigma_0^2}{1+P+\sigma_0^2+2\sigma_{X_0,U_1}},
\end{aligned}$$

Since $\gamma>0$, $\gamma\sqrt{\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]}\geq\sqrt{\frac{\sigma_0^2}{1+P+\sigma_0^2+2\sigma_{X_0,U_1}}}-\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}$.

Because the RHS may not be positive, we take the maximum of zero and the RHS and obtain the following lower bound for $\mathcal{E}_m$ and $\mathcal{D}_m$.

$$\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]\geq\frac{1}{\gamma^2}\left(\left(\sqrt{\frac{\sigma_0^2}{1+P+\sigma_0^2+2\sigma_{X_0,U_1}}}-\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}\right)^+\right)^2. \tag{A.33}$$

Because the bound holds for every $\gamma>0$,

$$\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]\geq\sup_{\gamma>0}\frac{1}{\gamma^2}\left(\left(\sqrt{\frac{\sigma_0^2}{1+P+\sigma_0^2+2\sigma_{X_0,U_1}}}-\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}\right)^+\right)^2, \tag{A.34}$$

for the chosen $\mathcal{E}_m$ and $\mathcal{D}_m$. Now, from (A.25), σ_{X_0,U_1} can take values in $[-\sigma_0\sqrt{P},\sigma_0\sqrt{P}]$. Because the lower bound depends on $\mathcal{E}_m$ and $\mathcal{D}_m$ only through σ_{X_0,U_1}, we obtain the following lower bound for all $\mathcal{E}_m$ and $\mathcal{D}_m$,

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{X}_1-X_1)^2\right]\\
&\geq \inf_{|\sigma_{X_0,U_1}|\leq\sigma_0\sqrt{P}}\sup_{\gamma>0}\frac{1}{\gamma^2}\left(\left(\sqrt{\frac{\sigma_0^2}{1+P+\sigma_0^2+2\sigma_{X_0,U_1}}}-\sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2P-2\gamma(1-\gamma)\sigma_{X_0,U_1}}\right)^+\right)^2,
\end{aligned}$$

which proves Theorem 6. Notice that we did not take limits in m anywhere, and hence the lower bound holds for all values of m. □

A.8 Proof of Corollary 2: optimality of DPC-based strategy for asymptotically perfect reconstruction.

The case of nonzero rate

Proof. To prove Theorem 13, consider now the problem when the encoder wants to also communicate a message M reliably to the decoder at rate R.

Using Fano's inequality, since $\Pr(M \neq \widehat{M}) = \epsilon_m \to 0$ as $m \to \infty$, $H(M|\widehat{M}) \leq m\delta_m$ where $\delta_m \to 0$. Thus,

$$\begin{aligned} I(M;\widehat{M}) &= H(M) - H(M|\widehat{M}) \\ &= mR - H(M|\widehat{M}) \\ &\geq mR - m\delta_m = m(R-\delta_m). \end{aligned} \tag{A.35}$$

As before, we consider a mutual information inequality that follows directly from the Markov chain $(M, \mathbf{X}_0^m) \to \mathbf{X}_1^m \to \mathbf{Y}^m \to (\widehat{\mathbf{X}}^m, \widehat{M})$:

$$I(M, \mathbf{X}_0^m; \widehat{M}, \widehat{\mathbf{X}}^m) \leq I(\mathbf{X}_1^m; \mathbf{Y}^m). \tag{A.36}$$

The RHS can be bounded above as in (A.22). For the LHS,

$$\begin{aligned} I(M, \mathbf{X}_0^m; \widehat{M}, \widehat{\mathbf{X}}_1^m) &= I(M; \widehat{M}, \widehat{\mathbf{X}}_1^m) + I(\mathbf{X}_0^m; \widehat{M}, \widehat{\mathbf{X}}_1^m | M) \\ &\geq I(M; \widehat{M}) + I(\mathbf{X}_0^m; \widehat{M}, \widehat{\mathbf{X}}_1^m | M) \\ &= I(M; \widehat{M}) + h(\mathbf{X}_0^m|M) - h(\mathbf{X}_0^m|\widehat{M}, \widehat{\mathbf{X}}_1^m, M) \\ &\overset{\mathbf{X}_0^m \perp\!\!\!\perp M}{=} I(M; \widehat{M}) + h(\mathbf{X}_0^m) - h(\mathbf{X}_0^m|\widehat{M}, \widehat{\mathbf{X}}_1^m, M) \\ &\geq I(M; \widehat{M}) + h(\mathbf{X}_0^m) - h(\mathbf{X}_0^m|\widehat{\mathbf{X}}_1^m) \\ &\geq I(M; \widehat{M}) + I(\mathbf{X}_0^m; \widehat{\mathbf{X}}_1^m) \\ &\overset{\text{using (A.23)}}{\geq} I(M; \widehat{M}) + mI(X_0; \widehat{X}). \end{aligned} \tag{A.37}$$

From (A.35), (A.36) and (A.37), we obtain

$$\begin{aligned} m(R-\delta_m) + mI(X_0;\widehat{X}) &\overset{\text{using (A.35)}}{\leq} I(M;\widehat{M}) + mI(X_0;\widehat{X}) \\ &\overset{\text{using (A.37)}}{\leq} I(M, \mathbf{X}_0^m; \widehat{M}, \widehat{\mathbf{X}}_1^m) \\ &\overset{\text{using (A.36)}}{\leq} I(\mathbf{X}_1^m; \mathbf{Y}_2^m) \overset{\text{using (A.22)}}{\leq} mI(X_1;Y_2). \end{aligned} \tag{A.38}$$

$I(X_1;Y_2)$ and $I(X_0;\widehat{X}_1)$ can be bounded as before in (A.27) and (A.32). Observing that as

$m \to \infty$, $\delta_m \to 0$, we get the following lower bound on the $MMSE$ for nonzero rate,

$$\begin{aligned} & MMSE(P,R) \\ \geq & \inf_{\sigma_{X_0,U_1}} \sup_{\gamma>0} \frac{1}{\gamma^2} \left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{1+\sigma_0^2+P+2\sigma_{X_0,U_1}}} - \sqrt{(1-\gamma)^2\sigma_0^2+\gamma^2 P - 2\gamma(1-\gamma)\sigma_{X_0,U_1}} \right)^+ \right)^2 \end{aligned}$$

In the limit $\delta_m \to 0$, we require from (A.38) that $I(X_1;Y_2) \geq R$. This gives the following constraint on σ_{X_0,U_1},

$$\begin{aligned} \frac{1}{2}\log_2\left(1+P+\sigma_0^2+2\sigma_{X_0,U_1}\right) &\geq R \\ \text{i.e. } \sigma_{X_0,U_1} &\geq \frac{2^{2R}-1-P-\sigma_0^2}{2}, \end{aligned} \tag{A.39}$$

yielding (in conjunction with (A.25)) the constraint on σ_{X_0,U_1} in Theorem 13. The constraint on P in the Theorem follows from Costa's result [77], because the rate R must be smaller than the capacity over a power constrained AWGN channel with known interference, $\frac{1}{2}\log_2(1+P)$. □

Since we are free to choose γ, let $\gamma = \gamma^* = \frac{\sigma_0^2+\sigma_{X_0,U_1}}{\sigma_0^2+P+2\sigma_{X_0,U_1}}$. Then, $1-\gamma^* = \frac{P+\sigma_{X_0,U_1}}{\sigma_0^2+P+2\sigma_{X_0,U_1}}$. Thus, we get

$$0 \geq \inf_{\sigma_{X_0,U_1}} \frac{1}{\gamma^{*^2}} \left(\left(\sqrt{\frac{\sigma_0^2 2^{2R}}{1+\sigma_0^2+P+2\sigma_{X_0,U_1}}} - \sqrt{(1-\gamma^*)^2\sigma_0^2+\gamma^{*^2} P - 2\gamma^*(1-\gamma^*)\sigma_{X_0,U_1}} \right)^+ \right)^2 . \tag{A.40}$$

It has to be the case that the term inside $(\cdot)^+$ is non-positive for some value of σ_{X_0,U_1}. This immediately yields

$$\begin{aligned} 2^{2R} &\leq \sup_{\sigma_{X_0,U_1}} \frac{1}{\sigma_0^2}\left((1-\gamma^*)^2\sigma_0^2+\gamma^{*^2}P-2\gamma^*(1-\gamma^*)\sigma_{X_0,U_1}\right)(1+\sigma_0^2+P+2\sigma_{X_0,U_1}) \\ &= \sup_{\sigma_{X_0,U_1}} \frac{1}{\sigma_0^2}\frac{((P+\sigma_{X_0,U_1})^2\sigma_0^2+(\sigma_0^2+\sigma_{X_0,U_1})^2P-2(P+\sigma_{X_0,U_1})(\sigma_0^2+\sigma_{X_0,U_1})\sigma_{X_0,U_1})}{(\sigma_0^2+P+2\sigma_{X_0,U_1})^2} \\ &\quad \times(1+\sigma_0^2+P+2\sigma_{X_0,U_1}) \\ &= \sup_{\sigma_{X_0,U_1}} \frac{1}{\sigma_0^2}\frac{\left(P^2\sigma_0^2-\sigma_{X_0,U_1}^2\sigma_0^2+2P\sigma_{X_0,U_1}\sigma_0^2+P\sigma_0^4-P\sigma_{X_0,U_1}^2-2\sigma_{X_0,U_1}^3\right)}{(\sigma_0^2+P+2\sigma_{X_0,U_1})^2} \\ &\quad \times(1+\sigma_0^2+P+2\sigma_{X_0,U_1}) \\ &= \sup_{\sigma_{X_0,U_1}} \frac{1}{\sigma_0^2}\frac{\left((P\sigma_0^2-\sigma_{X_0,U_1}^2)(P+\sigma_0^2+2\sigma_{X_0,U_1})\right)}{(\sigma_0^2+P+2\sigma_{X_0,U_1})^2}(1+\sigma_0^2+P+2\sigma_{X_0,U_1}) \\ &= \sup_{\sigma_{X_0,U_1}} \frac{(P\sigma_0^2-\sigma_{X_0,U_1}^2)(1+\sigma_0^2+P+2\sigma_{X_0,U_1})}{\sigma_0^2(\sigma_0^2+P+2\sigma_{X_0,U_1})} \end{aligned}$$

Thus, we get the following upper bound on $C(P)$,

$$C(P) \leq \sup_{\sigma_{X_0,U_1} \in [-\sigma_0\sqrt{P}, \sigma_0\sqrt{P}]} \frac{1}{2}\log_2\left(\frac{(P\sigma_0^2 - \sigma_{X_0,U_1}^2)(1+\sigma_0^2+P+2\sigma_{X_0,U_1})}{\sigma_0^2(\sigma_0^2+P+2\sigma_{X_0,U_1})}\right). \tag{A.41}$$

The term $(P\sigma_0^2 - \sigma_{X_0,U_1}^2)$ is oblivious to the sign of σ_{X_0,U_1}. However, the term

$$\frac{1+\sigma_0^2+P+2\sigma_{X_0,U_1}}{\sigma_0^2+P+2\sigma_{X_0,U_1}} = 1 + \frac{1}{\sigma_0^2+P+2\sigma_{X_0,U_1}} \tag{A.42}$$

is clearly larger for $\sigma_{X_0,U_1} < 0$ if we fix $|\sigma_{X_0,U_1}|$. Thus the supremum in (A.41) is attained at some $\sigma_{X_0,U_1} < 0$, and we get

$$C(P) \leq \sup_{\sigma_{X_0,U_1} \in [-\sigma_0\sqrt{P}, 0]} \frac{1}{2}\log_2\left(\frac{(P\sigma_0^2 - \sigma_{X_0,U_1}^2)(1+\sigma_0^2+P+2\sigma_{X_0,U_1})}{\sigma_0^2(\sigma_0^2+P+2\sigma_{X_0,U_1})}\right), \tag{A.43}$$

which matches the expression in Corollary 2. Thus for perfect reconstruction ($MMSE = 0$), the combination of linear and DPC strategy proposed in Chapter 4.3.3 is optimal.

A.9 Proof of Lemma 2

$$\begin{aligned}
& \mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] \\
= \;& \mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] + r_p^2 \Pr(\mathcal{E}_m) + 2r_p \mathbb{E}_{\mathbf{Z}^m}\left[\left(\mathbf{1}_{\{\mathcal{E}_m\}}\right)\left(\|\mathbf{Z}^m\|\mathbf{1}_{\{\mathcal{E}_m\}}\right)\right] \\
\overset{(a)}{\leq} \;& \mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] + r_p^2 \Pr(\mathcal{E}_m) + 2r_p \sqrt{\mathbb{E}_{\mathbf{Z}^m}\left[\mathbf{1}_{\{\mathcal{E}_m\}}\right]}\sqrt{\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right]} \\
= \;& \left(\sqrt{\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right]} + r_p\sqrt{\Pr(\mathcal{E}_m)}\right)^2,
\end{aligned} \tag{A.44}$$

where (a) uses the Cauchy-Schwartz inequality [135, Pg. 13].

We wish to express $\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right]$ in terms of

$$\psi(m, r_p) := \Pr(\|\mathbf{Z}^m\| \geq r_p) = \int_{\|\mathbf{z}^m\| \geq r_p} \frac{e^{-\frac{\|\mathbf{z}^m\|^2}{2}}}{\left(\sqrt{2\pi}\right)^m} d\mathbf{z}^m. \tag{A.45}$$

Denote by $\mathcal{A}_m(r) := \frac{2\pi^{\frac{m}{2}} r^{m-1}}{\Gamma\left(\frac{m}{2}\right)}$ the surface area of a sphere of radius r in $\mathbb{R}^m$ [136, Pg. 458], where $\Gamma(\cdot)$ is the Gamma-function satisfying $\Gamma(m) = (m-1)\Gamma(m-1)$, $\Gamma(1) = 1$, and

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Dividing the space $\mathbb{R}^m$ into shells of thickness dr and radii r,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] &= \int_{\|\mathbf{z}^m\|\geq r_p} \|\mathbf{z}^m\|^2 \frac{e^{-\frac{\|\mathbf{z}^m\|^2}{2}}}{\left(\sqrt{2\pi}\right)^m} d\mathbf{z}^m = \int_{r\geq r_p} r^2 \frac{e^{-\frac{r^2}{2}}}{\left(\sqrt{2\pi}\right)^m} \mathcal{A}_m(r) dr \\
&= \int_{r\geq r_p} r^2 \frac{e^{-\frac{r^2}{2}}}{\left(\sqrt{2\pi}\right)^m} \frac{2\pi^{\frac{m}{2}} r^{m-1}}{\Gamma\left(\frac{m}{2}\right)} dr \\
&= \int_{r\geq r_p} \frac{e^{-\frac{r^2}{2}} 2\pi}{\left(\sqrt{2\pi}\right)^{m+2}} \frac{2\pi^{\frac{m+2}{2}} r^{m+1}}{\pi^{\frac{2}{m}} \Gamma\left(\frac{m+2}{2}\right)} dr = m\psi(m+2, r_p). \qquad \text{(A.46)}
\end{aligned}$$

Using (A.44), (A.46), and $r_p = \sqrt{\frac{mP}{\xi^2}}$

$$\mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\| + r_p\right)^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] \leq m\left(\sqrt{\psi(m+2, r_p)} + \sqrt{\frac{P}{\xi^2}}\sqrt{\psi(m, r_p)}\right)^2,$$

which yields the first part of Lemma 2. To obtain a closed-form upper bound we consider $P > \xi^2$. It suffices to bound $\psi(\cdot,\cdot)$.

$$\begin{aligned}
\psi(m, r_p) &= \Pr(\|\mathbf{Z}^m\|^2 \geq r_p^2) = \Pr(\exp(\rho \sum_{i=1}^m Z_i^2) \geq \exp(\rho r_p^2)) \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathbf{Z}^m}\left[\exp(\rho \sum_{i=1}^m Z_i^2)\right] e^{-\rho r_p^2} = \mathbb{E}_{Z_1}\left[\exp(\rho Z_1^2)\right]^m e^{-\rho r_p^2} \overset{\text{(for } 0<\rho<0.5\text{)}}{=} \frac{1}{(1-2\rho)^{\frac{m}{2}}} e^{-\rho r_p^2},
\end{aligned}$$

where (a) follows from the Markov inequality, and the last inequality follows from the fact that the moment generating function of a standard χ_2^2 random variable is $\frac{1}{(1-2\rho)^{\frac{1}{2}}}$ for $\rho \in (0, 0.5)$ [137, Pg. 375]. Since this bound holds for any $\rho \in (0, 0.5)$, we choose the minimizing $\rho^* = \frac{1}{2}\left(1 - \frac{m}{r_p^2}\right)$. Since $r_p^2 = \frac{mP}{\xi^2}$, ρ^* is indeed in $(0, 0.5)$ as long as $P > \xi^2$. Thus,

$$\psi(m, r_p) \leq \frac{1}{(1-2\rho^*)^{\frac{m}{2}}} e^{-\rho^* r_p^2} = \left(\frac{r_p^2}{m}\right)^{\frac{m}{2}} e^{-\frac{1}{2}\left(1-\frac{m}{r_p^2}\right) r_p^2} = e^{-\frac{r_p^2}{2} + \frac{m}{2} + \frac{m}{2}\ln\left(\frac{r_p^2}{m}\right)}.$$

Using the substitutions $r_c^2 = mP$, $\xi = \frac{r_c}{r_p}$ and $r_p^2 = \frac{mP}{\xi^2}$,

$$\Pr(\mathcal{E}_m) = \psi(m, r_p) = \psi\left(m, \sqrt{\frac{mP}{\xi^2}}\right) \leq e^{-\frac{mP}{2\xi^2} + \frac{m}{2} + \frac{m}{2}\ln\left(\frac{P}{\xi^2}\right)}, \text{ and} \qquad \text{(A.47)}$$

$$\mathbb{E}_{\mathbf{Z}^m}\left[\|\mathbf{Z}^m\|^2 \mathbf{1}_{\{\mathcal{E}_m\}}\right] \leq m\psi\left(m+2, \sqrt{\frac{mP}{\xi^2}}\right) \leq m e^{-\frac{mP}{2\xi^2} + \frac{m+2}{2} + \frac{m+2}{2}\ln\left(\frac{mP}{(m+2)\xi^2}\right)}. \qquad \text{(A.48)}$$

From (A.44), (A.47) and (A.48),

$$\begin{aligned}
& \mathbb{E}_{\mathbf{Z}^m}\left[\left(\|\mathbf{Z}^m\|+r_p\right)^2\mathbf{1}_{\{\mathcal{E}_m\}}\right] \\
\leq\ & \left(\sqrt{m}e^{-\frac{mP}{4\xi^2}+\frac{m+2}{4}+\frac{m+2}{4}\ln\left(\frac{mP}{(m+2)\xi^2}\right)}\sqrt{\frac{mP}{\xi^2}}e^{-\frac{mP}{4\xi^2}+\frac{m}{4}+\frac{m}{4}\ln\left(\frac{P}{\xi^2}\right)}\right)^2 \\
\overset{(\text{since } P>\xi^2)}{<}\ & \left(\sqrt{m}\left(1+\sqrt{\frac{P}{\xi^2}}\right)e^{-\frac{mP}{4\xi^2}+\frac{m+2}{4}+\frac{m+2}{4}\ln\left(\frac{P}{\xi^2}\right)}\right)^2 \\
=\ & m\left(1+\sqrt{\frac{P}{\xi^2}}\right)^2 e^{-\frac{mP}{2\xi^2}+\frac{m+2}{2}+\frac{m+2}{2}\ln\left(\frac{P}{\xi^2}\right)}.
\end{aligned}$$

A.10 Proof of Lemma 3

Choosing $A=\mathbf{X}_0^m$, $B=\mathbf{X}_1^m$ and $C=\widehat{\mathbf{X}}_1^m$ in Lemma 1,

$$\begin{aligned}
& \mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}_G^m)|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]=\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_1^m-\widehat{\mathbf{X}}_1^m\|^2|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right] \\
\geq\ & \left(\left(\sqrt{\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_1^m\|^2|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]}-\sqrt{\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\mathbf{X}_1^m\|^2|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]}\right)^+\right)^2 \\
=\ & \left(\left(\sqrt{\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_1^m\|^2|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]}-\sqrt{P}\right)^+\right)^2, \qquad \text{(A.49)}
\end{aligned}$$

since $\mathbf{X}_0^m-\mathbf{X}_1^m=\mathbf{U}_1^m$ is independent of $\mathbf{Z}_G^m$ and $\mathbb{E}\left[\|\mathbf{U}_1^m\|^2\right]=mP$. Define $\mathbf{Y}_L^m:=\mathbf{X}_1^m+\mathbf{Z}_L^m$ to be the output when the observation noise $\mathbf{Z}_L^m$ is distributed as a truncated Gaussian distribution:

$$f_{Z_L}(\mathbf{z}_L^m)=\begin{cases} c_m(L)\dfrac{e^{-\frac{\|\mathbf{z}_L^m\|^2}{2\sigma_G^2}}}{\left(\sqrt{2\pi\sigma_G^2}\right)^m} & \mathbf{z}_L^m\in\mathcal{S}_L^G \\ 0 & \text{otherwise.}\end{cases} \qquad \text{(A.50)}$$

Let the estimate at the second controller on observing $\mathbf{y}_L^m$ be denoted by $\widehat{\mathbf{X}}_L^m$. Then, by the definition of conditional expectations,

$$\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_1^m\|^2|\mathbf{Z}_G^m\in\mathcal{S}_L^G\right]=\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_L^m\|^2\right]. \qquad \text{(A.51)}$$

To get a lower bound, we now allow the controllers to optimize themselves with the additional knowledge that the observation noise $\mathbf{z}^m$ must fall in $\mathcal{S}_L^G$. In order to prevent the first controller from "cheating" and allocating different powers to the two events (*i.e.* $\mathbf{z}^m$ falling or not falling in $\mathcal{S}_L^G$), we enforce the constraint that the power P must not change with this additional knowledge. Since the controller's observation $\mathbf{X}_0^m$ is independent of $\mathbf{Z}^m$, this

constraint is satisfied by the original controller (without the additional knowledge) as well, and hence the cost for the system with the additional knowledge is still a valid lower bound to that of the original system. The rest of the proof uses ideas from channel coding and the rate-distortion theorem [28, Ch. 13] from information theory. We view the problem as a problem of implicit communication from the first controller to the second. Notice that for a given $\gamma(\cdot)$, $\mathbf{X}_1^m$ is a function of $\mathbf{X}_0^m$, $\mathbf{Y}_L^m = \mathbf{X}_1^m + \mathbf{Z}_L^m$ is conditionally independent of $\mathbf{X}_0^m$ given $\mathbf{X}_1^m$ (since the noise $\mathbf{Z}_L^m$ is additive and independent of $\mathbf{X}_1^m$ and $\mathbf{X}_0^m$). Further, $\widehat{\mathbf{X}}_L^m$ is a function of $\mathbf{Y}_L^m$. Thus $\mathbf{X}_0^m - \mathbf{X}_1^m - \mathbf{Y}_L^m - \widehat{\mathbf{X}}_L^m$ form a Markov chain. Using the data-processing inequality [28, Pg. 33],

$$I(\mathbf{X}_0^m; \widehat{\mathbf{X}}_L^m) \leq I(\mathbf{X}_1^m; \mathbf{Y}_L^m), \tag{A.52}$$

where $I(A, B)$ is the expression for mutual information expression between two random variables A and B (see, for example, [28, Pg. 18, Pg. 231]). To estimate the distortion to which $\mathbf{X}_0^m$ can be communicated across this truncated Gaussian channel (which, in turn, helps us lower bound the $MMSE$ in estimating $\mathbf{X}_1^m$), we need to upper bound the term on the RHS of (A.52).

Lemma 4.

$$\frac{1}{m} I(\mathbf{X}_1^m; \mathbf{Y}_L^m) \leq \frac{1}{2} \log_2 \left(\frac{e^{1-d_m(L)} (\overline{P} + d_m(L)\sigma_G^2) c_m^{\frac{2}{m}}(L)}{\sigma_G^2} \right).$$

Proof. We first obtain an upper bound to the power of $\mathbf{X}_1^m$ (this bound is the same as that used in Corollary 1):

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{X}_1^m\|^2 \right] &= \mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{X}_0^m + \mathbf{U}_1^m\|^2 \right] = \mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{X}_0^m\|^2 \right] + \mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{U}_1^m\|^2 \right] + 2\mathbb{E}_{\mathbf{X}_0^m} \left[\mathbf{X}_0^{mT} \mathbf{U}_1^m \right] \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{X}_0^m\|^2 \right] + \mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{U}_1^m\|^2 \right] + 2\sqrt{\mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{X}_0^m\|^2 \right]} \sqrt{\mathbb{E}_{\mathbf{X}_0^m} \left[\|\mathbf{U}_1^m\|^2 \right]} \\
&\leq m \left(\sigma_0 + \sqrt{P} \right)^2,
\end{aligned}$$

where (a) follows from the Cauchy-Schwartz inequality. We use the following definition of *differential entropy* $h(A)$ of a continuous random variable A [28, Pg. 224]:

$$h(A) = -\int_{\mathbb{S}} f_A(a) \log_2 \left(f_A(a) \right) da, \tag{A.53}$$

where $f_A(a)$ is the pdf of A, and $\mathbb{S}$ is the support set of A. Conditional differential entropy is defined similarly [28, Pg. 229]. Let $\overline{P} := \left(\sigma_0 + \sqrt{P} \right)^2$. Now, $\mathbb{E}\left[Y_{L,i}^2 \right] = \mathbb{E}\left[X_{1,i}^2 \right] + \mathbb{E}\left[Z_{L,i}^2 \right]$ (since $X_{1,i}$ is independent of $Z_{L,i}$ and by symmetry, $Z_{L,i}$ are zero mean random variables). Denote $\overline{P}_i = \mathbb{E}\left[X_{1,i}^2 \right]$ and $\sigma_{G,i}^2 = \mathbb{E}\left[Z_{L,i}^2 \right]$. In the following, we derive an upper bound $C_{G,L}^{(m)}$

on $\frac{1}{m}I(\mathbf{X}_1^m;\mathbf{Y}_L^m)$.

$$
\begin{aligned}
C_{G,L}^{(m)} &:= \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} I(\mathbf{X}_1^m;\mathbf{Y}_L^m) \\
&\overset{(a)}{=} \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} h(\mathbf{Y}_L^m) - \frac{1}{m} h(\mathbf{Y}_L^m|\mathbf{X}_1^m) \\
&= \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} h(\mathbf{Y}_L^m) - \frac{1}{m} h(\mathbf{X}_1^m + \mathbf{Z}_L^m|\mathbf{X}_1^m) \\
&\overset{(b)}{=} \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} h(\mathbf{Y}_L^m) - \frac{1}{m} h(\mathbf{Z}_L^m|\mathbf{X}_1^m) \\
&\overset{(c)}{=} \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} h(\mathbf{Y}_L^m) - \frac{1}{m} h(\mathbf{Z}_L^m) \\
&\overset{(d)}{\le} \sup_{p(\mathbf{X}_1^m):\mathbb{E}\left[\|\mathbf{X}_1^m\|^2\right]\le m\overline{P}} \frac{1}{m} \sum_{i=1}^{m} h(Y_{L,i}) - \frac{1}{m} h(\mathbf{Z}_L^m) \\
&\overset{(e)}{\le} \sup_{\overline{P}_i:\sum_{i=1}^m \overline{P}_i \le m\overline{P}} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}\log_2\left(2\pi e(\overline{P}_i + \sigma_{G,i}^2)\right) - \frac{1}{m} h(\mathbf{Z}_L^m) \\
&\overset{(f)}{\le} \frac{1}{2}\log_2\left(2\pi e(\overline{P} + d_m(L)\sigma_G^2)\right) - \frac{1}{m} h(\mathbf{Z}_L^m). \qquad \text{(A.54)}
\end{aligned}
$$

Here, (a) follows from the definition of mutual information [28, Pg. 231], (b) follows from the fact that translation does not change the differential entropy [28, Pg. 233], (c) uses independence of $\mathbf{Z}_L^m$ and $\mathbf{X}_1^m$, and (d) uses the chain rule for differential entropy [28, Pg. 232] and the fact that conditioning reduces entropy [28, Pg. 232]. In (e), we used the fact that Gaussian random variables maximize differential entropy. The inequality (f) follows from the concavity-$\cap$ of the $\log(\cdot)$ function and an application of Jensen's inequality [28, Pg.

25]. We also use the fact that $\frac{1}{m}\sum_{i=1}^{m}\sigma_{G,i}^2 = d_m(L)\sigma_G^2$, which can be proven as follows

$$\begin{aligned}
\frac{1}{m}\mathbb{E}\left[\sum_{i=1}^{m} Z_{L,i}^2\right] &\overset{\text{(using (A.50))}}{=} \frac{\sigma_G^2}{m}\int_{\mathbf{z}^m\in\mathcal{S}_L^G}\frac{\|\mathbf{z}^m\|^2}{\sigma_G^2}c_m(L)\frac{\exp\left(-\frac{\|\mathbf{z}_G^m\|^2}{2\sigma_G^2}\right)}{\left(\sqrt{2\pi\sigma_G^2}\right)^m}d\mathbf{z}_G^m \\
&= \frac{c_m(L)\sigma_G^2}{m}\mathbb{E}\left[\|\mathbf{Z}_G^m\|^2\mathbf{1}_{\left\{\|\mathbf{Z}_G^m\|\le\sqrt{mL^2\sigma_G^2}\right\}}\right] \\
&\overset{(\widetilde{\mathbf{Z}}^m:=\frac{\mathbf{Z}_G^m}{\sigma_G})}{=} \frac{c_m(L)\sigma_G^2}{m}\mathbb{E}\left[\|\widetilde{\mathbf{Z}}^m\|^2\mathbf{1}_{\left\{\|\widetilde{\mathbf{Z}}^m\|\le\sqrt{mL^2}\right\}}\right] \\
&= \frac{c_m(L)\sigma_G^2}{m}\left(\mathbb{E}\left[\|\widetilde{\mathbf{Z}}^m\|^2\right] - \mathbb{E}\left[\|\widetilde{\mathbf{Z}}^m\|^2\mathbf{1}_{\left\{\|\widetilde{\mathbf{Z}}^m\|>\sqrt{mL^2}\right\}}\right]\right) \\
&\overset{\text{(using (A.46))}}{=} \frac{c_m(L)\sigma_G^2}{m}\left(m - m\psi(m+2,\sqrt{mL^2})\right) \\
&= c_m(L)\left(1-\psi(m+2,L\sqrt{m})\right)\sigma_G^2 = d_m(L)\sigma_G^2. \qquad \text{(A.55)}
\end{aligned}$$

We now compute $h(\mathbf{Z}_L^m)$

$$\begin{aligned}
h(\mathbf{Z}_L^m) &= \int_{\mathbf{z}^m\in\mathcal{S}_L^G} f_{Z_L}(\mathbf{z}^m)\log_2\left(\frac{1}{f_{Z_L}(\mathbf{z}^m)}\right)d\mathbf{z}^m = \int_{\mathbf{z}^m\in\mathcal{S}_L^G} f_{Z_L}(\mathbf{z}^m)\log_2\left(\frac{\left(\sqrt{2\pi\sigma_G^2}\right)^m}{c_m(L)e^{-\frac{\|\mathbf{z}^m\|^2}{2\sigma_G^2}}}\right)d\mathbf{z}^m \\
&= -\log_2\left(c_m(L)\right) + \frac{m}{2}\log_2\left(2\pi\sigma_G^2\right) + \int_{\mathbf{z}^m\in\mathcal{S}_L^G} c_m(L)f_G(\mathbf{z}^m)\frac{\|\mathbf{z}^m\|^2}{2\sigma_G^2}\log_2(e)\,d\mathbf{z}^m. \qquad \text{(A.56)}
\end{aligned}$$

Analyzing the last term of (A.56),

$$\begin{aligned}
&\int_{\mathbf{z}^m\in\mathcal{S}_L^G} c_m(L)f_G(\mathbf{z}^m)\frac{\|\mathbf{z}^m\|^2}{2\sigma_G^2}\log_2(e)\,d\mathbf{z}^m \\
&= \frac{\log_2(e)}{2\sigma_G^2}\int_{\mathbf{z}^m\in\mathcal{S}_L^G} c_m(L)\frac{e^{-\frac{\|\mathbf{z}^m\|^2}{2\sigma_G^2}}}{\left(\sqrt{2\pi\sigma_G^2}\right)^m}\|\mathbf{z}^m\|^2 d\mathbf{z}^m = \frac{\log_2(e)}{2\sigma_G^2}\int_{\mathbf{z}^m} f_{Z_L}(\mathbf{z}^m)\|\mathbf{z}^m\|^2 d\mathbf{z}^m \\
&\overset{\text{(using (A.50))}}{=} \frac{\log_2(e)}{2\sigma_G^2}\mathbb{E}_G\left[\|\mathbf{Z}_L^m\|^2\right] = \frac{\log_2(e)}{2\sigma_G^2}\mathbb{E}_G\left[\sum_{i=1}^{m} Z_{L,i}^2\right] \\
&\overset{\text{(using (A.55))}}{=} \frac{\log_2(e)}{2\sigma_G^2}md_m(L)\sigma_G^2 = \frac{m\log_2\left(e^{d_m(L)}\right)}{2}. \qquad \text{(A.57)}
\end{aligned}$$

The expression $C_{G,L}^{(m)}$ can now be upper bounded using (A.54), (A.56) and (A.57) as follows.

$$\begin{aligned} C_{G,L}^{(m)} &\leq \frac{1}{2}\log_2\left(2\pi e(\overline{P}+d_m(L)\sigma_G^2)\right)+\frac{1}{m}\log_2\left(c_m(L)\right)-\frac{1}{2}\log_2\left(2\pi\sigma_G^2\right)-\frac{1}{2}\log_2\left(e^{d_m(L)}\right) \\ &= \frac{1}{2}\log_2\left(2\pi e(\overline{P}+d_m(L)\sigma_G^2)\right)+\frac{1}{2}\log_2\left(c_m^{\frac{2}{m}}(L)\right)-\frac{1}{2}\log_2\left(2\pi\sigma_G^2\right)-\frac{1}{2}\log_2\left(e^{d_m(L)}\right) \\ &= \frac{1}{2}\log_2\left(\frac{2\pi e(\overline{P}+d_m(L)\sigma_G^2)c_m^{\frac{2}{m}}(L)}{2\pi\sigma_G^2 e^{d_m(L)}}\right) \\ &= \frac{1}{2}\log_2\left(\frac{e^{1-d_m(L)}(\overline{P}+d_m(L)\sigma_G^2)c_m^{\frac{2}{m}}(L)}{\sigma_G^2}\right). \end{aligned} \tag{A.58}$$

□

Now, recall that the rate-distortion function $D_m(R)$ for squared error distortion for source $\mathbf{X}_0^m$ and reconstruction $\widehat{\mathbf{X}}_L^m$ is,

$$D_m(R) := \inf_{\substack{p(\widehat{\mathbf{X}}_L^m|\mathbf{X}_0^m) \\ \frac{1}{m}I(\mathbf{X}_0^m;\widehat{\mathbf{X}}_L^m)\leq R}} \frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_L^m\|^2\right], \tag{A.59}$$

which is the dual of the rate-distortion function [28, Pg. 341]. Since $I(\mathbf{X}_0^m;\widehat{\mathbf{X}}_L^m)\leq mC_{G,L}^{(m)}$, using the converse to the rate distortion theorem [28, Pg. 349] and the upper bound on the mutual information represented by $C_{G,L}^{(m)}$,

$$\frac{1}{m}\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[\|\mathbf{X}_0^m-\widehat{\mathbf{X}}_L^m\|^2\right]\geq D_m(C_{G,L}^{(m)}). \tag{A.60}$$

Since the Gaussian source is iid, $D_m(R)=D(R)$, where $D(R)=\sigma_0^2 2^{-2R}$ is the distortion-rate function for a Gaussian source of variance σ_0^2 [28, Pg. 346]. Thus, using (A.49), (A.51) and (A.60),

$$\mathbb{E}_{\mathbf{X}_0^m,\mathbf{Z}_G^m}\left[J_2^{(\gamma)}(\mathbf{X}_0^m,\mathbf{Z}^m)|\mathbf{Z}^m\in\mathcal{S}_L^G\right]\geq\left(\left(\sqrt{D(C_{G,L}^{(m)})}-\sqrt{P}\right)^+\right)^2.$$

Substituting the bound on $C_{G,L}^{(m)}$ from (A.58),

$$D(C_{G,L}^{(m)})=\sigma_0^2 2^{-2C_{G,L}^{(m)}}=\frac{\sigma_0^2\sigma_G^2}{c_m^{\frac{2}{m}}(L)e^{1-d_m(L)}(\overline{P}+d_m(L)\sigma_G^2)}.$$

Using (A.49), this completes the proof of the lemma. Notice that $c_m(L)\to 1$ and $d_m(L)\to 1$ for fixed m as $L\to\infty$, as well as for fixed $L>1$ as $m\to\infty$. So the lower bound on $D(C_{G,L}^{(m)})$ approaches κ of Corollary 1 in both of these two limits.

A.11 Proof for bounded ratios for the finite-dimensional Witsenhausen counterexample

Let P^* denote the power P in the lower bound in Theorem 8. We show here that for any choice of P^*, the ratio of the upper and the lower bound is bounded. Consider the two simple linear strategies of zero-forcing ($\mathbf{u}_1^m = -\mathbf{x}_0^m$) and zero-input ($\mathbf{u}_1^m = 0$) followed by LLSE estimation at C_2. The average cost attained using these two strategies is $k^2\sigma_0^2$ and $\frac{\sigma_0^2}{\sigma_0^2+1} < 1$ respectively. An upper bound is obtained using the best amongst the two linear strategies and the lattice-based quantization strategy.

Case 1: $P^* \geq \frac{\sigma_0^2}{100}$.

The first stage cost is larger than $k^2\frac{\sigma_0^2}{100}$. Consider the upper bound of $k^2\sigma_0^2$ obtained by zero-forcing. The ratio of the upper bound and the lower bound is no larger than 100. *Case 2*: $P^* < \frac{\sigma_0^2}{100}$ and $\sigma_0^2 < 16$.

Using the bound from Corollary 1 (which is a special case of the bound in Theorem 8),

$$\begin{aligned}\kappa &= \frac{\sigma_0^2}{(\sigma_0+\sqrt{P^*})^2+1} \overset{\left(P^*<\frac{\sigma_0^2}{100}\right)}{\geq} \frac{\sigma_0^2}{\sigma_0^2\left(1+\frac{1}{\sqrt{100}}\right)^2+1} \\ &\overset{(\sigma_0^2<16)}{\geq} \frac{\sigma_0^2}{16\left(1+\frac{1}{\sqrt{100}}\right)^2+1} = \frac{\sigma_0^2}{20.36} \geq \frac{\sigma_0^2}{21}.\end{aligned}$$

Thus, for $\sigma_0^2 < 16$ and $P^* \leq \frac{\sigma_0^2}{100}$,

$$\overline{\mathcal{J}}_{min} \geq \left(\left(\sqrt{\kappa}-\sqrt{P^*}\right)^+\right)^2 \geq \sigma_0^2\left(\frac{1}{\sqrt{21}}-\frac{1}{\sqrt{100}}\right)^2 \approx 0.014\sigma_0^2 \geq \frac{\sigma_0^2}{72}.$$

Using the zero-input upper bound of $\frac{\sigma_0^2}{\sigma_0^2+1}$, the ratio of the upper and lower bounds is at most $\frac{72}{\sigma_0^2+1} \leq 72$.

Case 3: $P^* \leq \frac{\sigma_0^2}{100}, \sigma_0^2 \geq 16, P^* \leq \frac{1}{2}$.

In this case,

$$\begin{aligned}\kappa &= \frac{\sigma_0^2}{(\sigma_0+\sqrt{P^*})^2+1} \overset{(P^*\leq\frac{1}{2})}{\geq} \frac{\sigma_0^2}{(\sigma_0+\sqrt{0.5})^2+1} \\ &\overset{(a)}{\geq} \frac{16}{(\sqrt{16}+\sqrt{0.5})^2+1} \approx 0.6909 \geq 0.69,\end{aligned}$$

where (a) uses $\sigma_0^2 \geq 16$ and the observation that $\frac{x^2}{(x+b)^2+1} = \frac{1}{\left(1+\frac{b}{x}\right)^2+\frac{1}{x^2}}$ is an increasing

function of x for $x, b > 0$. Thus,

$$\left(\left(\sqrt{\kappa} - \sqrt{P}\right)^+\right)^2 \geq ((\sqrt{0.69} - \sqrt{0.5})^+)^2 \approx 0.0153 \geq 0.015.$$

Using the upper bound of $\frac{\sigma_0^2}{\sigma_0^2+1} < 1$, the ratio of the upper and the lower bounds is smaller than $\frac{1}{0.015} < 67$.

Case 4: $\sigma_0^2 > 16$, $\frac{1}{2} < P^* \leq \frac{\sigma_0^2}{100}$ Using $L = 2$ in the lower bound,

$$\begin{aligned} c_m(L) &= \frac{1}{\Pr(\|\mathbf{Z}^m\|^2 \leq mL^2)} = \frac{1}{1 - \Pr(\|\mathbf{Z}^m\|^2 > mL^2)} \\ &\overset{\text{(Markov's ineq.)}}{\leq} \frac{1}{1 - \frac{m}{mL^2}} \overset{(L=2)}{=} \frac{4}{3}, \end{aligned}$$

Similarly,

$$\begin{aligned} d_m(2) &= \frac{\Pr(\|\mathbf{Z}^{m+2}\|^2 \leq mL^2)}{\Pr(\|\mathbf{Z}^m\|^2 \leq mL^2)} \\ &\geq \Pr(\|\mathbf{Z}^{m+2}\|^2 \leq mL^2) = 1 - \Pr(\|\mathbf{Z}^{m+2}\|^2 > mL^2) \\ &\overset{\text{(Markov's ineq.)}}{\geq} 1 - \frac{m+2}{mL^2} = 1 - \frac{1 + \frac{2}{m}}{4} \overset{(m\geq 1)}{\geq} 1 - \frac{3}{4} = \frac{1}{4}. \end{aligned}$$

In the bound, we are free to use any $\sigma_G^2 \geq 1$. Using $\sigma_G^2 = 6P^* > 1$,

$$\begin{aligned} \kappa_2 &= \frac{\sigma_G^2 \sigma_0^2}{\left((\sigma_0 + \sqrt{P^*})^2 + d_m(2)\sigma_G^2\right) c_m^{\frac{2}{m}}(2) e^{1 - d_m(2)}} \\ &\overset{(a)}{\geq} \frac{6P^*\sigma_0^2}{\left((\sigma_0 + \frac{\sigma_0}{10})^2 + \frac{6\sigma_0^2}{100}\right)\left(\frac{4}{3}\right)^{\frac{2}{m}} e^{\frac{3}{4}}} \overset{(m\geq 1)}{\geq} 1.255P^*. \end{aligned}$$

where (a) uses $\sigma_G^2 = 6P^*$, $P^* < \frac{\sigma_0^2}{100}$, $c_m(2) \leq \frac{4}{3}$ and $1 > d_m(2) \geq \frac{1}{4}$. Thus,

$$\left(\left(\sqrt{\kappa_2} - \sqrt{P^*}\right)^+\right)^2 \geq P^*(\sqrt{1.255} - 1)^2 \geq \frac{P^*}{70}. \tag{A.61}$$

Now, using the lower bound on the total cost from Theorem 8, and substituting $L = 2$,

$$\begin{aligned}
\overline{\mathcal{J}}_{min}(m, k^2, \sigma_0^2) &\geq k^2 P^* + \frac{\sigma_G^m}{c_m(2)} \exp\left(-\frac{mL^2(\sigma_G^2 - 1)}{2}\right) \left(\left(\sqrt{\kappa_2} - \sqrt{P^*}\right)^+\right)^2 \\
&\overset{(\sigma_G^2 = 6P^*)}{\geq} k^2 P^* + \frac{(6P^*)^m}{c_m(2)} \exp\left(-\frac{4m(6P^* - 1)}{2}\right) \frac{P^*}{70} \\
&\overset{(a)}{\geq} k^2 P^* + \frac{3^m}{\frac{4}{3}} e^{2m} e^{-12P^* m} \frac{1}{70 \times 2} \\
&\overset{(m \geq 1)}{\geq} k^2 P^* + \frac{3 \times 3 \times e^2}{4 \times 70 \times 2} e^{-12mP^*} \\
&> k^2 P^* + \frac{1}{9} e^{-12mP^*}, \qquad \text{(A.62)}
\end{aligned}$$

where (a) uses $c_m(2) \leq \frac{4}{3}$ and $P^* \geq \frac{1}{2}$. We loosen the lattice-based upper bound from Theorem 7 and bring it into a form similar to (A.62). Here, P is a part of the optimization:

$$\begin{aligned}
& \overline{\mathcal{J}}_{min}(m, k^2, \sigma_0^2) \\
\leq\ & \inf_{P > \xi^2} k^2 P + \left(1 + \sqrt{\frac{P}{\xi^2}}\right)^2 e^{-\frac{mP}{2\xi^2} + \frac{m+2}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right)} \\
\leq\ & \inf_{P > \xi^2} k^2 P + \frac{1}{9} e^{-\frac{0.5mP}{\xi^2} + \frac{m+2}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right) + 2\ln\left(1 + \sqrt{\frac{P}{\xi^2}}\right) + \ln(9)} \\
\leq\ & \inf_{P > \xi^2} k^2 P + \frac{1}{9} e^{-m\left(\frac{0.5P}{\xi^2} - \frac{m+2}{2m}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right) - \frac{2}{m}\ln\left(1 + \sqrt{\frac{P}{\xi^2}}\right) - \frac{\ln(9)}{m}\right)} \\
=\ & \inf_{P > \xi^2} k^2 P + \frac{1}{9} e^{-\frac{0.12mP}{\xi^2}} \times e^{-m\left(\frac{0.38P}{\xi^2} - \frac{1 + \frac{2}{m}}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right) - \frac{2}{m}\ln\left(1 + \sqrt{\frac{P}{\xi^2}}\right) - \frac{\ln(9)}{m}\right)} \\
\overset{(m \geq 1)}{\leq}\ & \inf_{P > \xi^2} k^2 P + \frac{1}{9} e^{-\frac{0.12mP}{\xi^2}} e^{-m\left(\frac{0.38P}{\xi^2} - \frac{3}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right) - 2\ln\left(1 + \sqrt{\frac{P}{\xi^2}}\right) - \ln(9)\right)} \\
\leq\ & \inf_{P \geq 34\xi^2} k^2 P + \frac{1}{9} e^{-\frac{0.12mP}{\xi^2}}, \qquad \text{(A.63)}
\end{aligned}$$

where the last inequality follows from the fact that $\frac{0.38P}{\xi^2} > \frac{3}{2}\left(1 + \ln\left(\frac{P}{\xi^2}\right)\right) + 2\ln\left(1 + \sqrt{\frac{P}{\xi^2}}\right) + \ln(9)$ for $\frac{P}{\xi^2} > 34$. This can be checked easily by plotting it.[2] Using $P = 100\xi^2 P^* \geq 50\xi^2 >$

[2]It can also be verified symbolically by examining the expression $g(b) = 0.38b^2 - \frac{3}{2}(1 + \ln b^2) - 2\ln(1 + b) - \ln(9)$, taking its derivative $g'(b) = 0.76b - \frac{3}{b} - \frac{2}{1+b}$, and second derivative $g''(b) = 0.76 + \frac{3}{b^2} + \frac{2}{(1+b)^2} > 0$. Thus $g(\cdot)$ is convex-$\cup$. Further, $g'(\sqrt{34}) \approx 3.62 > 0$, and $g(\sqrt{34}) \approx 0.09$ and so $g(b) > 0$ whenever $b \geq \sqrt{34}$.

$34\xi^2$ (since $P^* \geq \frac{1}{2}$) in (A.63),

$$\begin{aligned} \overline{\mathcal{J}}_{min}(m, k^2, \sigma_0^2) &\leq k^2 100\xi^2 P^* + \frac{1}{9} e^{-m\frac{0.12\times 100\xi^2 P^*}{\xi^2}} \\ &= k^2 100\xi^2 P^* + \frac{1}{9} e^{-12mP^*}. \end{aligned} \tag{A.64}$$

Using (A.62) and (A.64), the ratio of the upper and the lower bounds is bounded for all m since

$$\mu \leq \frac{k^2 100\xi^2 P^* + \frac{1}{9} e^{-12mP^*}}{k^2 P^* + \frac{1}{9} e^{-12mP^*}} \leq \frac{k^2 100\xi^2 P^*}{k^2 P^*} = 100\xi^2. \tag{A.65}$$

For $m = 1$, $\xi = 1$, and thus in the proof the ratio $\mu \leq 100$. For m large, $\xi \approx 2$ [138, Chapter VIII], and $\mu \lesssim 400$. For arbitrary m, using the recursive construction in [139, Theorem 8.18], $\xi \leq 4$, and thus $\mu \leq 1600$ regardless of m.

Appendix B

Approximate-optimality for a noisy version of Witsenhausen's counterexample

The proof involves showing that the ratio of the upper bound of Theorem 16 and the lower bound of Theorem 15 is no larger than 41. This is done by dividing the (k, σ, N_1) space into different regions, which are dealt with separately.

An optimal value of P that attains the minimum in the second expression in the lower bound of Theorem 15 is denoted by P^*.

Case 1: $N_1 \geq 1$.

A lower bound is

$$\overline{\mathcal{J}}_{opt} \geq \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} \overset{(N_1 \geq 1)}{\geq} \frac{\sigma_0^2}{\sigma_0^2 + \sigma_0^2 + 1} = \frac{\sigma_0^2}{2\sigma_0^2 + 1}.$$

The zero-input upper bound $\overline{\mathcal{J}}_{\widetilde{ZI}} = \frac{\sigma_0^2}{\sigma_0^2+1}$. The ratio of the upper and lower bounds is therefore smaller than

$$\frac{2\sigma_0^2 + 1}{\sigma_0^2 + 1} < 2. \tag{B.1}$$

Case 2: $\sigma_0^2 < N_1 < 1$.

If $N_1 > \sigma_0^2$, using the first term in the lower bound of Theorem 15,

$$\begin{aligned} \overline{\mathcal{J}}_{opt} \quad &\geq \quad \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} \\ &\overset{(N_1 > \sigma_0^2)}{>} \quad \frac{\sigma_0^2 \sigma_0^2}{\sigma_0^2 \sigma_0^2 + \sigma_0^2 + \sigma_0^2} = \frac{\sigma_0^4}{\sigma_0^4 + 2\sigma_0^2} \overset{(\sigma_0^2 < 1)}{>} \frac{\sigma_0^4}{\sigma_0^2 + 2\sigma_0^2} = \frac{\sigma_0^2}{3}. \end{aligned}$$

The $\widetilde{ZI}$ upper bound $\overline{\mathcal{J}}_{\widetilde{ZI}} = \frac{\sigma_0^2}{\sigma_0^2+1} < \sigma_0^2$. Thus the ratio of upper and lower bounds is smaller than 3.

Case 3: $N_1 < \sigma_0^2 < 1$.
Case 3a: $P^* \geq \frac{\sigma_0^2}{16}$.
Since the lower bound is the larger of the two terms in Theorem 15, it is larger than any convex combination of the two terms as well. That is,

$$\begin{aligned}\overline{\mathcal{J}}_{opt} &\geq \frac{1}{2}\left(k^2P^* + \left(\left(\sqrt{\widetilde{\kappa}} - \sqrt{P^*}\right)^+\right)^2\right) + \frac{1}{2}\frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} \\ &\overset{\left(P^* \geq \frac{\sigma_0^2}{16}\right)}{\geq} \frac{k^2\sigma_0^2}{32} + \frac{\sigma_0^2 N_1}{2(\sigma_0^2 N_1 + \sigma_0^2 + N_1)}.\end{aligned}$$

Now for the upper bound, we use the zero-forcing strategy

$$\begin{aligned}\overline{\mathcal{J}}_{\widetilde{ZF}} &= \frac{k^2\sigma_0^4}{\sigma_0^2 + N_1} + \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} \\ &\leq \frac{k^2\sigma_0^4}{\sigma_0^2} + \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} = k^2\sigma_0^2 + \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}.\end{aligned}$$

The ratio of upper and lower bound is therefore smaller than $\max\{32, 2\} = 32$.
Case 3b: $P^* < \frac{\sigma_0^2}{16}$.
Since $N_1 < \sigma_0^2$,

$$\widetilde{\sigma}_0^2 = \frac{\sigma_0^4}{\sigma_0^2 + N_1} \overset{(N_1 < \sigma_0^2)}{\geq} \frac{\sigma_0^4}{\sigma_0^2 + \sigma_0^2} = \frac{\sigma_0^2}{2}.$$

Thus,

$$\begin{aligned}\widetilde{\kappa} &= \frac{\widetilde{\sigma}_0^2}{(\widetilde{\sigma}_0 + \sqrt{P^*})^2 + 1} \geq \frac{\sigma_0^2/2}{\left(\frac{\sigma}{\sqrt{2}} + \frac{\sigma}{4}\right)^2 + 1} \\ &\overset{(\sigma_0^2 \leq 1)}{\geq} \frac{\sigma_0^2}{2\left(\frac{1}{\sqrt{2}} + \frac{1}{4}\right)^2 + 1} \geq \frac{\sigma_0^2}{3}.\end{aligned}$$

Thus,

$$(\sqrt{\widetilde{\kappa}} - \sqrt{P^*})^2 \geq \sigma_0^2\left(\frac{1}{\sqrt{3}} - \frac{1}{4}\right)^2 > 0.1\sigma_0^2.$$

Using $\overline{\mathcal{J}}_{\widetilde{ZI}} = \frac{\sigma_0^2}{\sigma_0^2+1} < \sigma_0^2$, the ratio of the upper and lower bounds is smaller than 10.
Case 4: $N_1 \leq 1 < \sigma_0^2$.

Case 4a: $P^* \leq \frac{1}{9}$.
In this case,

$$\widetilde{\sigma}_0^2 = \frac{\sigma_0^4}{\sigma_0^2 + N_1} \stackrel{(N_1 \leq 1 \leq \sigma_0^2)}{\geq} \frac{\sigma_0^4}{\sigma_0^2 + \sigma_0^2} = \frac{\sigma_0^2}{2}$$

Therefore,

$$\widetilde{\kappa} = \frac{\widetilde{\sigma}_0^2}{(\widetilde{\sigma}_0 + \sqrt{P^*})^2 + 1} \geq \frac{\sigma_0^2/2}{\left(\frac{\sigma}{\sqrt{2}} + \frac{1}{3}\right)^2 + 1} \geq 0.24.$$

Thus, $\left(\left(\sqrt{\widetilde{\kappa}} - \sqrt{P^*}\right)^+\right)^2 \geq 0.024$. The zero-input upper bound is smaller than 1. Thus the ratio is smaller than $\frac{1}{0.024} < 41$.

Case 4b: $P^* > \frac{1}{9}$
A lower bound is

$$\begin{aligned} \overline{\mathcal{J}}_{opt} &\geq \max\left\{\frac{k^2}{9}, \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}\right\} \\ &\geq \frac{k^2}{9} \times \frac{9}{10} + \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1} \times \frac{1}{10} = \frac{k^2}{10} + \frac{\sigma_0^2 N_1}{10(\sigma_0^2 N_1 + \sigma_0^2 + N_1)}. \end{aligned}$$

Now, we use the asymptotic vector quantization upper bound of

$$\lim_{m \to \infty} \overline{\mathcal{J}}_{\widetilde{VQ}} \leq k^2 \left(\frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1} + 1\right) + \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1}. \tag{B.2}$$

Since $N_1 < 1$, this upper bound is smaller than $2k^2 + \frac{\sigma_0^2 N_1}{\sigma_0^2 N_1 + \sigma_0^2 + N_1}$. The ratio of the first terms in the upper bound and the lower bound of (B.2) is at most 20. The ratio of the second terms is

$$\begin{aligned} \frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1} \times \frac{10(\sigma_0^2 N_1 + \sigma_0^2 + N_1)}{\sigma_0^2 N_1} &= 10\frac{\sigma_0^2 N_1}{\sigma_0^2 + N_1} + 10 \\ &\leq 10 + 10 = 20. \end{aligned}$$

Thus the ratio of the upper and lower bounds is no larger than 41 in all cases.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul./Oct. 1948.

[2] E. Lee, "Cyber physical systems: Design challenges," in *11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing (ISORC)*, 2008, pp. 363–369.

[3] A. Cavalcanti, T. Hogg, B. Shirinzadeh, and H. Liaw, "Nanorobot communication techniques: A comprehensive tutorial," in *9th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2006, pp. 1–6.

[4] 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.

[5] L. Guptha, M. Prasaanth, and S. Srikrishna, "Swarm Intelligence-Ant Colony Foraging Behavior," *International Journal of Computer Applications*, vol. 1, pp. 109–112, 2010.

[6] Z. Huang, "Waggle dance video." [Online]. Available: http://www.cyberbee.net/biology/ch6/dance2.html

[7] J. Riley, U. Greggers, A. Smith, D. Reynolds, and R. Menzel, "The flight paths of honeybees recruited by the waggle dance," *Nature*, vol. 435, no. 7039, pp. 205–207, 2005.

[8] E. M. F. Mauriello, T. Mignot, Z. Yang, and D. R. Zusman, "Gliding Motility Revisited: How Do the Myxobacteria Move without Flagella?" *Microbiol. Mol. Biol. Rev.*, vol. 74, no. 2, pp. 229–249, 2010. [Online]. Available: http://mmbr.asm.org/cgi/content/abstract/74/2/229

[9] J. de Sousa, K. Johansson, A. Speranzon, and J. Silva, "A control architecture for multiple submarines in coordinated search missions," in *Proceedings of the 16th IFAC World Congress*, 2005.

[10] "Kiva systems." [Online]. Available: http://www.kivasystems.com

[11] I. Ihle, R. Skjetne, and T. Fossen, "Nonlinear formation control of marine craft with experimental results," *Atlantis*, 2004.

[12] "Signaling in retrospect and the informational structure of markets," nobel Prize Lecture, December 8, 2001. [Online]. Available: http://nobelprize.org/nobel_prizes/economics/laureates/2001/spence-lecture.pdf

[13] A. M. Spence, "Job market signaling," *Quarterly journal of economics*, vol. 87, pp. 355–374, 1973.

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

[15] ——, "Equivalent stochastic control problems," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 1, pp. 3–11, 1988, 10.1007/BF02551232. [Online]. Available: http://dx.doi.org/10.1007/BF02551232

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

[17] Y.-C. Ho and M. P. Kastner, "Market signaling: an example of a two-person decision problem with dynamic information structure," *IEEE Trans. Autom. Control*, vol. 23, no. 2, pp. 350 – 361, Apr. 1978.

[18] 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.

[19] 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.

[20] Y.-C. Ho and T. Chang, "Another look at the nonclassical information structure problem," *IEEE Trans. Autom. Control*, vol. 25, no. 3, pp. 537–540, 1980.

[21] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Trans. Autom. Control*, vol. 51, no. 2, pp. 1984–1996, Feb. 2006.

[22] 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.

[23] R. Bansal and T. Başar, "Stochastic teams with nonclassical information revisited: When is an affine control optimal?" *IEEE Trans. Autom. Control*, vol. 32, pp. 554–559, Jun. 1987.

[24] S. Yüksel, "Stochastic nestedness and the belief sharing information pattern," *IEEE Trans. Autom. Control*, pp. 2773–2786, 2009.

[25] 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.

[26] J. T. Lee, E. Lau, and Y.-C. L. Ho, "The Witsenhausen counterexample: A hierarchical search approach for nonconvex optimization problems," *IEEE Trans. Autom. Control*, vol. 46, no. 3, pp. 382–397, 2001.

[27] N. Li, J. R. Marden, and J. S. Shamma, "Learning approaches to the Witsenhausen counterexample from a view of potential games," *Proceedings of the 48th IEEE Conference on Decision and Control (CDC)*, 2009.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1st ed. New York: Wiley, 1991.

[29] R. Etkin, D. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, Dec. 2008.

[30] R. Gallager, "Basic limits on protocol information in data communication networks," *IEEE Trans. Inf. Theory*, vol. 22, no. 4, pp. 385 – 398, Jul. 1976.

[31] P. Varaiya, "Towards a layered view of control," *Proceedings of the 36th IEEE Conference on Decision and Control (CDC)*, 1997.

[32] ——, "Smart cars on smart roads: Problems of control," *IEEE Trans. Autom. Control*, vol. 38, no. 2, pp. 195 – 207, Feb. 1993.

[33] J. Malik and S. Russel, "Traffic surveillance and detection technology development: new traffic sensor technology final report," University of California, Berkeley, Tech. Rep. UCB-ITS-PRR-976, 1997.

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

[35] R. D'Andrea, "personal communication," Sep. 2010.

[36] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Trans. Autom. Control*, vol. 19, no. 5, pp. 494 – 500, Oct. 1974.

[37] A. S. Avestimehr, "Wireless network information flow: A deterministic approach," Ph.D. dissertation, UC Berkeley, Berkeley, CA, 2008.

[38] A. S. Avestimehr, S. Diggavi, and D. N. C. Tse, "A deterministic approach to wireless relay networks," in *Proc. of the Allerton Conference on Communications, Control and Computing*, October 2007.

[39] ——, "Wireless network information flow: a deterministic approach," *Submitted to IEEE Transactions on Information Theory*, Jul. 2009.

[40] S. Arora, "Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems," *Journal of the ACM (JACM)*, vol. 45, no. 5, p. 782, 1998.

[41] V. Borkar and S. K. Mitter, "LQG control with communication constraints," in *Communications, Computation, Control, and Signal Processing: a Tribute to Thomas Kailath*. Norwell, MA: Kluwer Academic Publishers, 1997, pp. 365–373.

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

[43] S. Yüksel and T. Başar, "Achievable rates for stability of LTI systems over noisy forward and feedback channels," in *Proceedings of the 2005 Conference on Information Sciences and Systems*, Baltimore, MD, Mar. 2005, paper 12.

[44] C. Nolan, "Memento, the movie," Oct. 2000.

[45] M. Wertheimer, *Gestalt theory*. Hayes Barton Press, 1944.

[46] A. Fe'ldbaum, "Dual-control theory I," *Autom. Remote Control*, vol. 21, pp. 874–880, 1961.

[47] A. Mahajan, "Sequential decomposition of sequential teams: applications to real-time communication and networked control systems," Ph.D. dissertation, University of Michigan, Ann Arbor, Sep. 2008.

[48] C. Papadimitriou and M. Yannakakis, "On complexity as bounded rationality (extended abstract)," in *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*. ACM, 1994, pp. 726–733.

[49] "Prisoner's dilemma and the Axelrod experiment." [Online]. Available: http://www.slideshare.net/amitsinha1964/prisoners-dilemma-and-axelrod-experiment

[50] C. Ioannou, "Algorithmic bounded rationality, optimality and noise," Ph.D. dissertation, University of Minnesota, 2009.

[51] C. Sims, "Implications of rational inattention," *Journal of Monetary Economics*, vol. 50, no. 3, pp. 665–690, 2003.

[52] P. Milgrom and J. Roberts, "Limit pricing and entry under incomplete information: An equilibrium analysis," *Econometrica: Journal of the Econometric Society*, vol. 50, no. 2, pp. 443–459, 1982.

[53] F. Matejka, "Rigid pricing and rationally-inattentive consumer," Princeton University, Working paper, Feb. 2010, http://www.pacm.princeton.edu/publications/FMatejka_RI_consumer.pdf.

[54] F. Matejka and C. Sims, "Discrete Actions in Information-Constrained Tracking Problems," *Working Paper*, 2010.

[55] P. Grover and A. Sahai, "A vector version of Witsenhausen's counterexample: Towards convergence of control, communication and computation," in *Proceedings of the 47th IEEE Conference on Decision and Control (CDC)*, 2008.

[56] ——, "Vector Witsenhausen counterexample as assisted interference suppression," *Special issue on Information Processing and Decision Making in Distributed Control Systems of the International Journal on Systems, Control and Communications (IJSCC)*, vol. 2, pp. 197–237, 2010.

[57] P. Grover, A. B. Wagner, and A. Sahai, "Information embedding meets distributed control," in *IEEE Information Theory Workshop (ITW)*, 2010, pp. 1–5.

[58] P. Grover, A. Sahai, and S. Y. Park, "The finite-dimensional Witsenhausen counterexample," in *Proceedings of the 7th IEEE International conference on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, 2009, pp. 604–613.

[59] P. Grover, S. Y. Park, and A. Sahai, "The finite-dimensional Witsenhausen counterexample," *Submitted to IEEE Transactions on Automatic Control, Arxiv preprint arXiv:1003.0514*, 2010.

[60] A. Sahai and P. Grover, "Demystifying the Witsenhausen Counterexample," *"Ask the Experts," IEEE Control Systems Magazine,*, vol. 30, no. 6, pp. 20 –24, Dec. 2010.

[61] P. Grover, S. Y. Park, and A. Sahai, "On the generalized Witsenhausen counterexample," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2009.

[62] P. Grover and A. Sahai, "Distributed signal cancelation inspired by Witsenhausen's counterexample," in *IEEE International Symposium on Information Theory (ISIT)*, 2010, pp. 151–155.

[63] ——, "Is Witsenhausen's counterexample a relevant toy?" *Proceedings of the 49th IEEE Conference on Decision and Control (CDC), Altanta, Georgia, USA*, Dec. 2010.

[64] ——, "Implicit and explicit communication in decentralized control," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2010. [Online]. Available: http://arxiv.org/abs/1010.4854

[65] T. J. Goblick, "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inf. Theory*, vol. 11, no. 4, Oct. 1965.

[66] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, 2003.

[67] D. Bertsekas, *Dynamic Programming*. Belmont, MA: Athena Scientific, 1995.

[68] A. Johnson, "LQG applications in the process industries," *Chemical Engineering Science*, vol. 48, no. 16, pp. 2829 – 2838, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/B6TFK-444P65G-DW/2/11297f05c55ec1ecc203290854705f25

[69] A. Gattami, "Optimal Decisions with Limited Information," Ph.D. dissertation, Lund University, 2007.

[70] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[71] M. Garey and D. Johnson, *Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences*. WH Freeman and Company, San Francisco, Calif, 1979.

[72] E. Berlekamp, R. McEliece, and H. van Tilborg, "On the inherent intractability of certain coding problems (corresp.)," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 384 – 386, may. 1978.

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

[74] E. Hazan, S. Safra, and O. Schwartz, "On the hardness of approximating k-dimensional matching," in *Electronic Colloquium on Computational Complexity, TR03-020*. Citeseer, 2003.

[75] C. Papadimitriou and M. Yannakakis, "Optimization, approximation, and complexity classes," in *STOC '88: Proceedings of the twentieth annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 1988, pp. 229–234.

[76] J. Karlsson, A. Gattami, T. Oechtering, and M. Skoglund, "Iterative source-channel coding approach to Witsenhausen's counterexample," in *Submitted to American Control Conference (ACC)*, 2010.

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

[78] N. Sandell, P. Varaiya, M. Athans, and M. Safonov, "Survey of decentralized control methods for large scale systems," *Automatic Control, IEEE Transactions on*, vol. 23, no. 2, pp. 108 – 128, Apr. 1978.

[79] Y. Ho and K. Chu, "Team Decision Theory and Information Structures in Optimal Control Problems- Part I," *IEEE Trans. Autom. Control*, vol. 17, no. 1, pp. 15–22, Feb. 1972.

[80] D. A. Castanon, "Decentralized Estimation of Linear Gaussian Systems," LIDS, MIT, Tech. Rep., 1981.

[81] S.-H. Wang and E. Davison, "On the stabilization of decentralized control systems," *IEEE Trans. Autom. Control*, vol. 18, no. 5, pp. 473 – 478, Oct. 1973.

[82] P. Grover, K. Woyach, and A. Sahai, "Towards a communication-theoretic understanding of system-level power consumption," *Arxiv preprint arXiv:1010.4855, submitted to IEEE Journal on Selected Areas in Communication*, 2010.

[83] Y.-C. Ho and M. Kastner, "Market signaling: An example of a two-person decision problem with dynamic information structure," *IEEE Trans. Autom. Control*, vol. 23, no. 2, pp. 350 – 361, Apr. 1978.

[84] W. Wong and R. Brockett, "Systems with finite communication bandwidth constraints II: stabilization with limited information feedback. ," *IEEE Trans. Autom. Contr.*, pp. 1049–53, 1999.

[85] 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. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[86] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, and S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1453–1464, Sep. 2004.

[87] N. Elia, "Indelible control," in *Multidisciplinary Research in Control*, ser. Lecture Notes in Control and Information Sciences, L. Giarr and B. Bamieh, Eds. Springer Berlin / Heidelberg, 2003, vol. 289, pp. 33–46, 10.1007/3-540-36589-3_3. [Online]. Available: http://dx.doi.org/10.1007/3-540-36589-3_3

[88] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[89] R. Bansal and T. Başar, "Simultaneous design of measurement and control strategies for stochastic systems with feedback* 1," *Automatica*, vol. 25, no. 5, pp. 679–694, 1989.

[90] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 2002.

[91] H. Palaiyanur, C. Chang, and A. Sahai, "Lossy compression of active sources," in *IEEE International Symposium on Information Theory (ISIT)*, 2008, pp. 1977–1981.

[92] Y.-H. Kim, A. Sutivong, and T. M. Cover, "State amplification," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.

[93] N. Merhav and S. Shamai, "Information rates subject to state masking," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2254–2261, Jun. 2007.

[94] S. I. Gel'Fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[95] A. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1639–1667, Jun. 2002.

[96] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[97] Y. Steinberg, "Simultaneous transmission of data and state with common knowledge," in *IEEE International Symposium on Information Theory (ISIT)*, 2008, pp. 935–939.

[98] A. B. Wagner, S. Tavildar, and P. Viswanath, "The rate region of the quadratic Gaussian two-terminal source-coding problem," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1938–1961, 2008.

[99] O. Sumszyk and Y. Steinberg, "Information embedding with reversible stegotext," in *IEEE International Symposium on Information Theory (ISIT)*, Jul. 2009, pp. 2728 –2732.

[100] 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.

[101] M. Anand and P. Kumar, "A digital interface for Gaussian relay and interference networks: Lifting codes from the discrete superposition model," *Arxiv preprint arXiv:1005.0167*, 2010.

[102] S. Y. Park, P. Grover, and A. Sahai, "A constant-factor approximately optimal solution to the Witsenhausen counterexample," *Proceedings of the 48th IEEE Conference on Decision and Control (CDC)*, Dec. 2009.

[103] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.

[104] C. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J*, vol. 38, no. 3, pp. 611–656, 1959.

[105] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Dispersion of Gaussian channels," in *IEEE International Symposium on Information Theory*, Seoul, Korea, 2009.

[106] ——, "New channel coding achievability bounds," in *IEEE International Symposium on Information Theory*, Toronto, Canada, 2008.

[107] A. Sahai and P. Grover, "The price of certainty : "waterslide curves" and the gap to capacity," Dec. 2007. [Online]. Available: http://arXiv.org/abs/0801.0352v1

[108] R. Blahut, "A hypothesis testing approach to information theory," Ph.D. dissertation, Cornell University, Ithaca, NY, 1972.

[109] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[110] M. S. Pinsker, "Bounds on the probability and of the number of correctable errors for nonblock codes," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 44–55, Oct./Dec. 1967.

[111] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Trans. Inf. Theory*, no. 5, pp. 1860–1886, May 2008.

[112] "Code for performance of lattice-based strategies for Witsenhausen's counterexample." [Online]. Available: http://www.eecs.berkeley.edu/∼pulkit/FiniteWitsenhausenCode.htm

[113] S. Yüksel and T. Başar, "Optimal signaling policies for decentralized multicontroller stabilizability over communication channels," *IEEE Trans. Autom. Control*, vol. 52, no. 10, pp. 1969 –1974, oct. 2007.

[114] ——, "Communication constraints for decentralized stabilizability with time-invariant policies," *IEEE Trans. Autom. Control*, pp. 1060–1066, Jun. 2007.

[115] N. C. Martins and M. A. Dahleh, "Feedback control in the presence of noisy channels: "Bode-like" fundamental limitations of performance," *IEEE Trans. Autom. Control*, vol. 53, no. 7, pp. 56–66, Aug. 2008.

[116] N. Martins, M. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 438–447, 2006.

[117] N. Martins, M. Dahleh, and J. Doyle, "Fundamental limitations of disturbance attenuation in the presence of side information," *IEEE Trans. Autom. Control*, vol. 52, no. 1, pp. 56–66, Jan. 2007.

[118] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, Jul. 2004.

[119] K. Shoarinejad, J. L. Speyer, and I. Kanellakopoulos, "A stochastic decentralized control problem with noisy communication," *SIAM Journal on Control and optimization*, vol. 41, no. 3, pp. 975–990, 2002.

[120] A. D. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, p. 1, 1976.

[121] W. H. R. Equitz and T. M. Cover, "Successive Refinement of Information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, 1991.

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

[123] P. Grover, A. B. Wagner, and A. Sahai, "Information embedding meets distributed control," *Submitted to IEEE Transactions on Information Theory*, 2010.

[124] A. Gupta, C. Langbort, and T. Başar, "Optimal control in the presence of an intelligent jammer with limited actions," *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, 2010.

[125] R. Cogill and S. Lall, "Suboptimality bounds in stochastic control: A queueing example," in *American Control Conference (ACC)*, Jun. 2006, pp. 1642–1647.

[126] R. Cogill, S. Lall, and J. P. Hespanha, "A constant factor approximation algorithm for event-based sampling," in *American Control Conference (ACC)*, Jul. 2007, pp. 305–311.

[127] S. Strogatz, *Sync: The emerging science of spontaneous order*. Hyperion, 2003.

[128] J. Acebrón, L. Bonilla, C. Pérez Vicente, F. Ritort, and R. Spigler, "The Kuramoto model: A simple paradigm for synchronization phenomena," *Reviews of modern physics*, vol. 77, no. 1, pp. 137–185, 2005.

[129] Y. Kuramoto, *Chemical oscillations, waves, and turbulence*. Dover Publications, 2003.

[130] P. W. Cuff, "Communication in networks for coordinating behavior," Ph.D. dissertation, Stanford University, 2009.

[131] H. Asnani, H. Permuter, and T. Weissman, "Probing capacity," *Arxiv preprint arXiv:1010.1309*, 2010.

[132] T. Weissman, "Capacity of channels with action-dependent states," in *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory*, ser. ISIT'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1794–1798. [Online]. Available: http://portal.acm.org/citation.cfm?id=1701116.1701188

[133] H. Permuter and T. Weissman, "Source coding with a side information'vending machine'at the decoder," in *IEEE International Symposium on Information Theory (ISIT)*, 2009, pp. 1030–1034.

[134] B. Juba and M. Sudan, "Universal semantic communication I," in *Proceedings of the 40th annual ACM symposium on Theory of computing*. ACM, 2008, pp. 123–132.

[135] R. Durrett, *Probability: Theory and Examples*, 1st ed. Belmont, CA: Brooks/Cole, 2005.

[136] R. Courant, F. John, A. A. Blank, and A. Solomon, *Introduction to Calculus and Analysis*. Springer, 2000.

[137] S. M. Ross, *A first course in probability*, 6th ed. Prentice Hall, 2001.

[138] U. Erez, S. Litsyn, and R. Zamir, "Lattices which are good for (almost) everything," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3401–3416, Oct. 2005.

[139] D. Micciancio and S. Goldwasser, *Complexity of Lattice Problems: A Cryptographic Perspective*. Springer, 2002.