The source coding game with a cheating switcher

Hari Palaiyanur, *Student Member*, Cheng Chang, Anant Sahai, *Member*

Abstract

Motivated by the lossy compression of an active-vision video stream, we consider the problem of finding the rate-distortion function of an arbitrarily varying source (AVS) composed of a finite number of subsources with known distributions. Berger's paper 'The Source Coding Game', *IEEE Trans. Inform. Theory*, 1971, solves this problem under the condition that the adversary is allowed only strictly causal access to the subsource realizations. We consider the case when the adversary has access to the subsource realizations non-causally. Using the type-covering lemma, this new rate-distortion function is determined to be the maximum of the IID rate-distortion function over a set of source distributions attainable by the adversary. We then extend the results to allow for partial or noisy observations of subsource realizations. We further explore the model by attempting to find the rate-distortion function when the adversary is actually helpful.

Finally, a bound is developed on the uniform continuity of the IID rate-distortion function for finite-alphabet sources. The bound is used to give a sufficient number of distributions that need to be sampled to compute the rate-distortion function of an AVS to within a certain accuracy. The bound is also used to give a rate of convergence for the estimate of the rate-distortion function for an unknown IID finite-alphabet source .

Index Terms

Rate-distortion, arbitrarily varying source, uniform continuity of rate-distortion function, switcher, lossy compression, source coding game, estimation of rate-distortion function

I. Introduction

A. Motivation

Active vision/sensing/perception [2] is an approach to computer vision, the main principle of which is that sensors should choose to explore their environment actively *based on what they currently sense or have previously sensed*. As Bajcsy states it in [2], "We do not just see, we look." The contrast to passive sensors can be seen by comparing a fixed security camera (non-active) to a person holding a camera (active). Even if the person is otherwise stationary, they may zoom the camera into any part of their visual field to obtain a better view (e.g. if they see a trespasser). There is also the possibility that the sensor has noncausal information about the environment. For example, a cameraman at a sporting event generally has only causal knowledge of the environment. A cameraman on a movie set, however, has noncausal information about the environment through the script. The noncausal information can be advantageous to the cameraman in (actively) capturing the important features of a scene.

There is a subtle distinction between causal and strictly causal information and this distinction is related to the time-scales on which the environment changes. A causal active sensor knows both the present and the past, but a strictly causal one knows only the past. If the environment changes at a pace much slower than the sensor can actively look, there is essentially no difference between knowing the immediate past and knowing the present. However, if the environment changes at a pace faster then the sensor can actively look (and process information), there is intuitively a substantial difference between knowing only the past and knowing the present.

As motivation for this paper, we are interested in the fixed-rate lossy compression of an active-vision source. In reality, there are many interesting questions that need to be answered to truly understand the problem, including:

- What is the relevant distortion measure for active-video?
- Is there a distinction between the compression of an active-video source for use by the closed-loop control system that points the camera as compared to compression for later off-line use?
- How to model the entire plenoptic function that the active-video source will be dynamically sampling? [3]

This work was supported by an NSF Graduate Research Fellowship. An earlier version [1] of the material in this paper was presented in part at the 2007 International Symposium on Information Theory, Nice, France.

The authors are with Wireless Foundations, Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley. (email: hpalaiya@eecs.berkeley.edu; cchang@eecs.berkeley.edu; sahai@eecs.berkeley.edu)

It is also clear that the core issues here extend well beyond vision. They also arise in a series of sensor-measurements that were dynamically sampled by a distributed sensor network as well as the case of measurements taken by an autonomously moving sensor that chooses where to go in part based on what it is observing. More provocatively, similar issues of active-sources arise when the successive source symbols are brought by customers, each of which has free will and can choose among competing codecs for compression.[1]

We concentrate entirely on the simplest aspect of the problem: what is the impact on the rate-distortion function of having the source being actively sampled by an entity that knows something about the realizations of the environment as it does the sampling. Thus, we assume an overly simplified traditional rate-distortion setting with known finite alphabets and bounded distortion measures. The goal is the traditional block-coding one: meet an average distortion constraint with high probability using as little rate as possible.

The modeling question is whether or not it is worth building a detailed model for how the active-source is going to be doing its dynamic sampling of the source. Three basic ways to model the goals of the camera are worst case (adversarial), random (agnostic), and helpful (joint optimization of camera and coding system). Admittedly, the most interesting problems involve the compression of sources with memory, but following tradition we focus on memoryless sources to understand the basic differences between active and non-active sources for lossy compression.

In the context of active-vision, a strictly causal adversary pointing a camera is intuitively no more threatening than a robot randomly pointing the camera when the scene being captured is memoryless. This intuition was formally proved correct in [5] by Berger as he determined the rate-distortion function for memoryless sources and a strictly causal adversarial model. This paper determines the rate-distortion function for the additional cases of causal and non-causal adversaries. The model is then extended to allow only noisy observations by the adversary doing the sampling of the scene. To see the impact of the details of the dynamic sampling on the rate-distortion function, the paper also considers how the rate-distortion function changes when the ‘adversary’ is actually a helpful party.

B. Causality in information theory

The issue of causality arises naturally in several major problems of information theory where noncausal knowledge of the realizations of randomness in the problem can be advantageous. Shannon [6] studied the problem of transmitting information over a noisy channel with memoryless state parameter revealed to the encoder causally. Gelfand and Pinsker [7] studied the same problem with the state parameter available to the encoder noncausally. In general, the capacity is larger when the channel state is available noncausally to the encoder. When the channel state corresponds to Gaussian interference known noncausally, Costa [8] showed that the capacity is the same as when the interference is not present at all. Willems ([9], [10]) gave achievable strategies when the Gaussian interference is known only causally. Lattice strategies for both causal and non-causal knowledge of the interference are discussed in [11], but the advantage of finitely anticipatory knowledge of interference is not yet explicitly understood even in the case of Gaussian interference.

Agarwal et.al. [12] find the capacity for an arbitrarily varying channel whose input is constrained to look like an IID source with known distribution. The adversary is constrained to distort over a block to at most some (additive) distortion, but is not constrained to act causally. [12] shows that the rate-distortion function turns out to be the capacity for this channel. Because the codewords are constrained to look IID, simulating the action of a causal memoryless channel turns out to be sufficient for the adversary to minimize the capacity.

Causality also has implications for the problem of lossy source coding, as studied by Neuhoff and Gilbert [13]. There, for an IID source, causal source codes generally require a higher rate to achieve distortion D than non-causal source codes. It is also shown that optimal causal source codes can be constructed by time-sharing between memoryless codes. Hence, there is a rate penalty for using causal coders (as opposed to noncausal coders), but no further penalty for using memoryless coders. Similar results have been derived by Weissman and Merhav [14] for lossy source coding with causal and noncausal side information. In [13], the channel was implicitly assumed to noiseless and binary. Tatikonda, et.al [15] show that even if the channel is matched properly to achieve the *sequential* rate-distortion function, there is a penalty for using causal coders when the sources have memory. For example, they show that proper matching for a Gauss-Markov source is a Gaussian channel with feedback, but the rate-distortion performance with this causal matching still does not meet the performance of noncausal coders.

[1] This is related to a particularly odd kind of moral hazard in private health insurance markets. Somewhat counterintuitively, private health insurers actually have a disincentive to provide good treatment of chronic conditions since they fear attracting patients that are intrinsically likely to get sick! [4]

C. Results and organization of paper

Section II sets up the notation, model and briefly reviews the literature on lossy compression of arbitrarily varying sources. Section III gives the rate-distortion function for an AVS when the adversary has noncausal access to realizations of a finite collection of memoryless subsources and can sample among them. As shown in Theorem 3.1, the rate-distortion function for this problem is the maximization of the IID rate-distortion function over the memoryless distributions the adversary can simulate. The adversary requires only causal information to impose this rate-distortion function. This establishes that when the subsources are memoryless, the rate-distortion function can strictly increase when the adversary has knowledge of the present subsource realizations, but no further increase occurs when the adversary is allowed knowledge of the future.

We then extend the AVS model to include noisy or partial observations of the subsource realizations and determine the rate-distortion function for this setting in Section IV. As shown in Theorem 4.1, the form of the solution is the same as for the adversary with clean observations, with the set of attainable distributions essentially being related to the original distributions through Bayes' rule.

Next, Section V explores the problem when the goal of the active sensor is to help the coding system achieve a low distortion. Theorem 5.1 gives a characterization of the rate-distortion functions if the helper is fully noncausal in terms of the rate-distortion function for an associated lossy compression problem. As a corollary, we also give bounds for the cases of causal observations and noisy observations.

Simple examples illustrating these results are given in Section VI. In Section VII, we discuss how to compute the rate-distortion function for arbitrarily varying sources to within a given accuracy using the uniform continuity of the IID rate-distortion function. The main tool there is an explicit bound on the uniform continuity of the IID rate-distortion function that is of potentially independent interest. Finally, we conclude in Section VIII.

All the problems in this paper are studied in the context of fixed-length block coding. Variable-length coding could perform better in a universal sense by using only as much rate as required when the active sensor is not adversarial. However, we are interested in determining upper and lower bounds for the rate that active sensors might end up needing and for this purpose, fixed-length block coding is appropriate.

II. Problem Setup

A. Notation

Let $\mathcal{X}$ and $\widehat{\mathcal{X}}$ be the finite source and reconstruction alphabets respectively. Let $\mathbf{x}^n = (x_1, \ldots, x_n)$ denote an arbitrary vector from $\mathcal{X}^n$ and $\widehat{\mathbf{x}}^n = (\widehat{x}_1, \ldots, \widehat{x}_n)$ an arbitrary vector from $\widehat{\mathcal{X}}^n$. When needed, $\mathbf{x}^k = (x_1, \ldots, x_k)$ will be used to denote the first k symbols in the vector $\mathbf{x}^n$.

Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion measure on the product set $\mathcal{X} \times \widehat{\mathcal{X}}$ with maximum distortion $d^* < \infty$. Let

$$\widetilde{d} = \min_{(x,\widehat{x}):\ d(x,\widehat{x})>0} d(x, \widehat{x}) \tag{1}$$

be the minimum nonzero distortion. Define $d_n : \mathcal{X}^n \times \widehat{\mathcal{X}}^n \to [0, d^*]$ for $n \geq 1$ to be

$$d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) = \frac{1}{n} \sum_{k=1}^{n} d(x_k, \widehat{x}_k). \tag{2}$$

Let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$, let $\mathcal{P}_n(\mathcal{X})$ be the set of types of length n strings from $\mathcal{X}$, and let $\mathcal{W}$ be the set of probability transition matrices from $\mathcal{X}$ to $\widehat{\mathcal{X}}$. Let $p_{\mathbf{x}^n} \in \mathcal{P}_n(\mathcal{X})$ be the empirical type of a vector $\mathbf{x}^n$. For a $p \in \mathcal{P}(\mathcal{X})$, let

$$D_{\min}(p) = \sum_{x \in \mathcal{X}} p(x) \min_{\widehat{x} \in \widehat{\mathcal{X}}} d(x, \widehat{x}) \tag{3}$$

be the minimum average distortion achievable for the source distribution p. The rate-distortion function of $p \in \mathcal{P}(\mathcal{X})$ at distortion $D > D_{\min}(p)$ with respect to distortion measure d is defined to be

$$R(p, D) = \min_{W \in \mathcal{W}(p,D)} I(p, W), \tag{4}$$

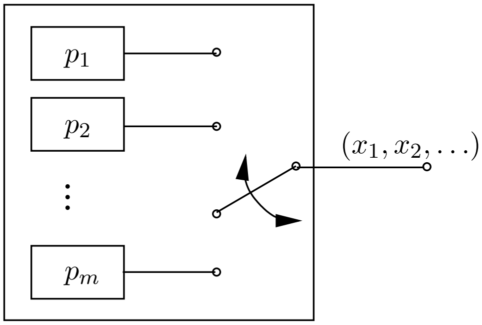

Fig. 1. A class of models for an AVS. The switcher can set the switch position according to the rules of the model.

where

$$\mathcal{W}(p, D) = \left\{ W \in \mathcal{W} : \sum_{x \in \mathcal{X}} \sum_{\widehat{x} \in \widehat{\mathcal{X}}} p(x) W(\widehat{x}|x) d(x, \widehat{x}) \leq D \right\} \tag{5}$$

and $I(p, W)$ is the mutual information[2]

$$I(p, W) = \sum_{x \in \mathcal{X}} \sum_{\widehat{x} \in \widehat{\mathcal{X}}} p(x) W(\widehat{x}|x) \ln \left[\frac{W(\widehat{x}|x)}{\sum_{x' \in \mathcal{X}} p(x') W(\widehat{x}|x')} \right]. \tag{6}$$

Let $\mathcal{B} = \{\widehat{\mathbf{x}}^n(1), \ldots, \widehat{\mathbf{x}}^n(K)\}$ be a codebook with K length-n vectors from $\widehat{\mathcal{X}}^n$. Define

$$d_n(\mathbf{x}^n; \mathcal{B}) = \min_{\widehat{\mathbf{x}}^n \in \mathcal{B}} d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n). \tag{7}$$

If $\mathcal{B}$ is used to represent an IID source with distribution p, then the average distortion of $\mathcal{B}$ is defined to be

$$d(\mathcal{B}) = \sum_{\mathbf{x}^n \in \mathcal{X}^n} P(\mathbf{x}^n) d_n(\mathbf{x}^n; \mathcal{B}) = \mathbb{E}[d_n(\mathbf{x}^n; \mathcal{B})], \tag{8}$$

where

$$P(\mathbf{x}^n) = \prod_{k=1}^{n} p(x_k). \tag{9}$$

For $n \geq 1$, $D > D_{\min}(p)$, let $K(n, D)$ be the minimum number of codewords needed in a codebook $\mathcal{B} \subset \widehat{\mathcal{X}}^n$ so that $d(\mathcal{B}) \leq D$. By convention, if no such codebook exists, $K(n, D) = \infty$. Let the rate-distortion function[3] of an IID source be $R(D) = \limsup_n \frac{1}{n} \ln K(n, D)$. Shannon's rate-distortion theorem ([16], [17]) states that for all n, $\frac{1}{n} \ln K(n, D) \geq R(p, D)$ and

$$\liminf_{n \to \infty} \frac{1}{n} \ln K(n, D) = R(D) = R(p, D). \tag{10}$$

B. *Arbitrarily varying sources*

The source coding game is a two-player game introduced in [5] by Berger as a model for an AVS. The two players are called the 'switcher' and 'coder'. In a coding context, the coder corresponds to the designer of a lossy source code and the switcher corresponds to a potentially malicious adversary pointing the camera.

Figure 1 shows a model of an AVS. There are m IID 'subsources' with common alphabet $\mathcal{X}$. In [5], the subsources are assumed to be independent, but that restriction turns out not to be required[4]. There can be multiple subsources governed by the same distribution. In that sense, the switcher has access to a *list* of m subsources, rather than a set

[2] We use natural log, denoted ln, and nats in most of the paper. In examples only, we use bits.

[3] We define $R(D_{\min}(p)) = \lim_{D \downarrow D_{\min}(p)} R(D)$. This is equivalent to saying that a sequence of codes represent a source to within distortion D if their average distortion is tending to D in the limit. The only distortion where this distinction is meaningful is $D_{\min}(p)$.

[4] In [5], the motivation was multiplexing data streams and independence is a reasonable assumption, but the proof does not require it. Active sources, however, would likely choose among correlated subsources in practice.

of m different distributions. The marginal distributions of the m subsources are known to be $\{p_l\}_{l=1}^m$ and we let $\mathcal{G} = \{p_1, \dots, p_m\}$. Let $P(x_{1,1}, \dots, x_{m,1})$ be the joint probability distribution for the IID source $\{(x_{1,k}, \dots, x_{m,k})\}_k$. Fix an $n \geq 1$ and consider a block of length n. We let $x_{l,k}$ denote the output of the l^{th} subsource at time k. We will use $\mathbf{x}_l^n$ to denote the vector $(x_{l,1}, \dots, x_{l,n})$. At each time k, the AVS outputs a letter x_k which is determined by the position of the switch inside the AVS. The switch positions are denoted $\mathbf{s}^n = (s_1, \dots, s_n)$ with $s_k \in \{1, 2, \dots, m\}$ for each $1 \leq k \leq n$. With this notation, $x_k = x_{s_k,k}$ for $1 \leq k \leq n$.

The switcher can set the switch position according to the model for the AVS. For example, in the compound source setting of Sakrison [18], the switcher chooses $s \in \{1, \dots, m\}$ and sets $s_k = s$ for $1 \leq k \leq n$. The main case analyzed in [5] allowed the switcher to change s_k arbitrarily, but the switcher only had knowledge at time k of $\mathbf{s}^{k-1}$ and $\mathbf{x}^{k-1}$. That is, the switcher only had knowledge of past switch positions and past AVS outputs before deciding the switch position at each time. One of the cases analyzed in this paper is termed full-lookahead, where the switcher makes a (possibly random) decision about the full $\mathbf{s}^n$ with knowledge of $\mathbf{x}_1^n, \mathbf{x}_2^n, \dots, \mathbf{x}_m^n$ beforehand. The other case is termed 1-step lookahead[5], where for each k, s_k is a (possibly random) function of $\mathbf{x}_1^k, \dots, \mathbf{x}_m^k$. The switcher may or may not have knowledge of the codebook, but this knowledge turns out to be inconsequential for the rate-distortion function.

The coder's goal is to design a codebook $\mathcal{B}$ of minimal size to represent $\mathbf{x}^n$ to within distortion D on average. The codebook must be able to do this for *every* allowable strategy for the switcher according to the model. Define

$$M(n, D) = \min \left\{ |\mathcal{B}| : \begin{array}{c} \mathcal{B} \subset \widehat{\mathcal{X}}^n, \ \mathbb{E}[d_n(\mathbf{x}^n; \mathcal{B})] \leq D \\ \text{for all allowable} \\ \text{switcher strategies} \end{array} \right\}. \tag{11}$$

Here, $\mathbb{E}[d_n(\mathbf{x}^n; \mathcal{B})]$ is defined to be $\sum_{\mathbf{x}^n} \left(\sum_{\mathbf{s}^n} P(\mathbf{s}^n, \mathbf{x}^n)\right) d_n(\mathbf{x}^n; \mathcal{B})$, where $P(\mathbf{s}^n, \mathbf{x}^n)$ is an appropriate probability mass function on $\{1, \dots, m\}^n \times \mathcal{X}^n$ that agrees with the model of the AVS. When the switcher has full lookahead, $P(\mathbf{s}^n, \mathbf{x}^n)$ must be composed of conditional distributions of the form

$$P(\mathbf{s}^n, \mathbf{x}^n | \mathbf{x}_1^n, \dots, \mathbf{x}_m^n) = P(\mathbf{s}^n | \mathbf{x}_1^n, \dots, \mathbf{x}_m^n) \cdot \prod_{k=1}^n 1(x_k = x_{s_k,k}). \tag{12}$$

Then, $P(\mathbf{s}^n, \mathbf{x}^n)$ is simply obtained by averaging over $(\mathbf{x}_1^n, \dots, \mathbf{x}_m^n)$.

$$P(\mathbf{s}^n, \mathbf{x}^n) = \sum_{(\mathbf{x}_1^n, \dots, \mathbf{x}_m^n)} \left(\prod_{k=1}^n P(x_{1,k}, \dots, x_{m,k}) \right) P(\mathbf{s}^n | \mathbf{x}_1^n, \dots, \mathbf{x}_m^n). \tag{13}$$

For a set of distributions $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$, let $D_{\min}(\mathcal{Q}) = \sup_{p \in \mathcal{Q}} D_{\min}(p)$. We are interested in the exponential rate of growth of $M(n, D)$ with n. Define the rate-distortion function of an AVS to be

$$R(D) \triangleq \limsup_{n \to \infty} \frac{1}{n} \ln M(n, D). \tag{14}$$

In every case considered, it will be also be clear that $R(D) = \liminf_{n \to \infty} \frac{1}{n} \ln M(n, D)$.

C. Literature Review

a) One IID source: Suppose $m = 1$. Then there is only one IID subsource $p_1 = p$ and the switch position is determined to be $s_k = 1$ for all time. This is exactly the classical rate-distortion problem considered by Shannon [16], and he showed

$$R(D) = R(p, D). \tag{15}$$

Computing $R(p, D)$ can be done with the Blahut-Arimoto algorithm [19], and also falls under the umbrella of convex programming.

[5] We use the term 1-step lookahead even though this term is meant to represent the causal (but not strictly) switcher. In most of the information theory literature, 'causal' knowledge includes knowledge of the present.

b) Compound source: Now suppose that $m > 1$, but the switcher is constrained to choose $s_k = s \in \{1, \ldots, m\}$ for all k. That is, the switch position is set once and remains constant afterwards. Sakrison [18] studied the rate-distortion function for this class of *compound* sources and showed that planning for the worst case subsource is both necessary and sufficient. Hence, for compound sources,

$$R(D) = \max_{p \in \mathcal{G}} R(p, D). \tag{16}$$

This result holds whether the switch position is chosen with or without knowledge of the realizations of the m subsources. Here, $R(D)$ can be computed easily since m is finite and each individual $R(p, D)$ can be computed.

c) Causal adversarial source: In Berger's setup [5], the switcher is allowed to choose $s_k \in \{1, \ldots, m\}$ arbitrarily at any time k , but must do so in a strictly causal manner without access to the current time step's subsource realizations. More specifically, the switch position s_k is chosen as a (possibly random) function of $(s_1, \ldots, s_{k-1})$ and $(x_1, \ldots, x_{k-1})$. The conclusion of [5] is that under these rules,

$$R(D) = \max_{p \in \textbf{conv}(\mathcal{G})} R(p, D), \tag{17}$$

where $\textbf{conv}(\mathcal{G})$ is the convex hull of $\mathcal{G}$. It should be noted that this same rate-distortion function applies in the following cases:

- The switcher chooses s_k at each time k without *any* observations at all.
- The switcher chooses s_k as a function of the first $k - 1$ outputs of *all* m subsources.

Note that in (17), evaluating $R(D)$ involves a maximization over an infinite set, so the computation of $R(D)$ is not trivial since $R(p, D)$ is not necessarily a concave $\cap$ function. A simple, provable, approximate (to any given accuracy) solution is discussed in Section VII.

III. $R(D)$ FOR THE CHEATING SWITCHER

In the conclusion of [5], Berger poses the question of what happens to the rate-distortion function when the rules are tilted in favor of the switcher. Suppose that the switcher were given access to the m subsource realizations before having to choose the switch positions; we call such a switcher a 'cheating switcher'. In this paper, we deal with two levels of noncausality and show they are essentially the same when the subsources are IID over time:

- The switcher chooses s_k based on the realizations of the m subsources at time k. We refer to this case as 1-step lookahead for the switcher.
- The switcher chooses $(s_1, \ldots, s_n)$ based on the entire length n realizations of the m subsources. We refer to this case as full lookahead for the switcher.

Theorem 3.1: Suppose the switcher has 1-step lookahead or full lookahead. In both cases, for $D > D_{\min}(\mathcal{C})$,

$$R(D) = \widetilde{R}(D) \triangleq \max_{p \in \mathcal{C}} R(p, D), \tag{18}$$

where

$$\mathcal{C} = \left\{ \begin{array}{ccc} & & \sum_{i \in \mathcal{V}} p(i) \geq P\big(x_l \in \mathcal{V}, 1 \leq l \leq m\big) \\ p \in \mathcal{P} & : & \forall\ \mathcal{V} \text{ such that} \\ & & \mathcal{V} \subseteq \mathcal{X} \end{array} \right\}. \tag{19}$$

For $D < D_{\min}(\mathcal{C})$, $R(D) = \infty$ by convention because the switcher can simulate a distribution for which the distortion D is infeasible for the coder.

Remarks:

- If there are at least two non-deterministic subsources and $\textbf{conv}(\mathcal{G}) \neq \mathcal{P}(\mathcal{X})$, then $\textbf{conv}(\mathcal{G})$ is a strict subset of $\mathcal{C}$, and thus $R(D)$ can strictly increase when the switcher is allowed to look at the present subsource realizations before choosing the switch position. Hence, extra rate must be provisioned for active sensors in general.
- As a consequence of the theorem, we see that when the subsources within an AVS are IID, knowledge of past subsource realizations is useless to the switcher, knowledge of the current step's subsource realizations is useful, and knowledge of future subsource realizations beyond the current step is useless if 1-step lookahead is already given.

- Note that computing $R(D)$ requires further discussion given in Section VII, just as it does for the strictly causal case of Berger.

Proof: We give a short outline of the proof here. See Appendix I for the complete proof. To show $R(D) \leq \widetilde{R}(D)$, we use the type-covering lemma from [5]. It says for a fixed type p in $\mathcal{P}_n(\mathcal{X})$ and $\epsilon > 0$, all sequences with type p can be covered within distortion D with at most $\exp(n(R(p, D) + \epsilon))$ codewords for large enough n. Since there are at most $(n+1)^{|\mathcal{X}|}$ distinct types, we can cover all n-length strings with types in $\mathcal{C}$ with at most $\exp(n(\widetilde{R}(D) + \frac{|\mathcal{X}|}{n} \ln(n+1) + \epsilon))$ codewords. Furthermore, we can show that types not in $\mathcal{C}$ occur exponentially rarely even if the switcher has full lookahead, meaning that their contribution to the average distortion can be bounded by d^* times an exponentially decaying term in n. Hence, the rate needed regardless of the switcher strategy is at most $\widetilde{R}(D) + \epsilon$ with $\epsilon > 0$ arbitrarily small.

Now, to show $R(D) \geq \widetilde{R}(D)$, we describe one potential strategy for the adversary. This strategy requires only 1-step lookahead and it forces the coder to use rate at least $\widetilde{R}(D)$. For each set $\mathcal{V} \subset \mathcal{X}$ with $\mathcal{V} \neq \emptyset$ and $|\mathcal{V}| \leq m$, the adversary has a random rule $f(\cdot|\mathcal{V})$, which is a probability mass function (PMF) on $\mathcal{V}$. At each time k, if the switcher observes a candidate set $\{x_{1,k}, \ldots, x_{m,k}\}$, the switcher chooses to output $x \in \{x_{1,k}, \ldots, x_{m,k}\}$ with probability $f(x|\{x_{1,k}, \ldots, x_{m,k}\})$. If $\beta(\mathcal{V}) = P(\{x_{1,k}, \ldots, x_{m,k}\} = \mathcal{V})$, let

$$\mathcal{D} \triangleq \left\{ p \in \mathcal{P} \quad : \quad \begin{array}{c} p(x) = \sum_{\mathcal{V} \subseteq \mathcal{X}, |\mathcal{V}| \leq m} \beta(\mathcal{V}) f(x|\mathcal{V}), x \in \mathcal{X} \\ f(\cdot|\mathcal{V}) \text{ is a PMF on } \mathcal{V}, \\ \forall\ \mathcal{V} \text{ s.t. } \mathcal{V} \subseteq \mathcal{X},\ |\mathcal{V}| \leq m \end{array} \right\}. \tag{20}$$

$\mathcal{D}$ is the set of IID distributions the AVS can 'simulate' using these memoryless rules requiring 1-step lookahead. It is clear by construction that $\mathcal{D} \subseteq \mathcal{C}$. Also, it is clear that both $\mathcal{C}$ and $\mathcal{D}$ are convex sets of distributions. Lemma 1.3 in Appendix I uses a separating hyperplane argument to show $\mathcal{D} = \mathcal{C}$. The adversary can therefore simulate any IID source with distribution in $\mathcal{C}$ and hence $R(D) \geq \widetilde{R}(D)$. ■

Qualitatively, allowing the switcher to 'cheat' gives access to distributions $p \in \mathcal{C}$ which may not be in **conv**$(\mathcal{G})$. Quantitatively, the conditions placed on the distributions in $\mathcal{C}$ are precisely those that restrict the switcher from producing symbols that do not occur often enough on average. For example, let $\mathcal{V} = \{1\}$ where $1 \in \mathcal{X}$, and suppose that the subsources are independent of each other. Then for every $p \in \mathcal{C}$,

$$p(1) \geq \prod_{l=1}^{m} p_l(1). \tag{21}$$

$\prod_{l=1}^{m} p_l(1)$ is the probability that all m subsources produce the letter 1 at a given time. In this case, the switcher has no option but to output the letter 1, hence any distribution the switcher mimics must have $p(1) \geq \prod_{l=1}^{m} p_l(1)$. The same logic can be applied to all subsets $\mathcal{V}$ of $\mathcal{X}$.

IV. Noisy observations of subsource realizations

A natural extension of the AVS model is to consider the case when the adversary has noisy access to subsource realizations through a discrete memoryless channel before pointing the camera. Since the subsource probability distributions are already known, this model is equivalent to one in which the switcher observes a state noiselessly. Conditioned on the state, the m subsources output symbols independent of the past according to a conditional distribution. This model is depicted in Figure 2.

The overall AVS is comprised now of a 'state generator' and a 'symbol generator' that outputs m symbols at a time. The state generator produces the state t_k at time k from a finite set $\mathcal{T}$. We assume the states are generated IID across time with distribution $\alpha(t)$. At time k, the symbol generator outputs $(x_{1,k}, \ldots, x_{m,k})$ according to $P(x_{1,k}, \ldots, x_{m,k}|t_k)$. This model allows for correlation among the subsources at a fixed time. Let $p_l(\cdot|t), l = 1, \ldots, m$, be the marginals of this joint distribution so that conditioned on t_k, $x_{l,k}$ has marginal distribution $p_l(\cdot|t_k)$. For an $t \in \mathcal{T}$, let $\overline{\mathcal{G}}(t) = \textbf{conv}(p_1(\cdot|t), \ldots, p_m(\cdot|t))$.

The switcher can observe states either with full lookahead or 1-step lookahead, but these two cases will once again have the same rate-distortion function when the switcher is an adversary. So assume that at time k, the switcher chooses the switch position s_k with knowledge of $\mathbf{t}^n, \mathbf{x}_1^{k-1}, \ldots, \mathbf{x}_m^{k-1}$. The non-cheating and cheating switcher can be recovered as special cases of this model. If the conditional distributions $p_l(x|t)$ do not depend

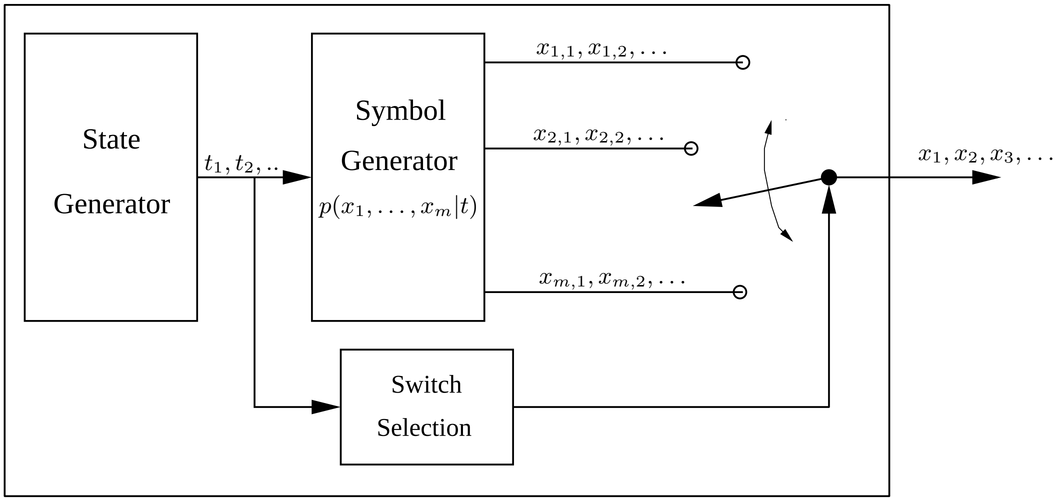

Fig. 2. A model of an AVS encompassing both cheating and non-cheating switchers. Additionally, this model allows for noisy observations of subsource realizations by the switcher.

on t, the non-cheating switcher is recovered. The cheating switcher is recovered by setting $\mathcal{T} = \mathcal{X}^m$ and letting $p_l(x|t) = 1(x = t(l))$ where the state t is an m dimensional vector consisting of the outputs of each subsource.

With this setup, we have the following extension of Theorem 3.1.

Theorem 4.1: For the AVS problem of Figure 2, where the adversary has access to the states either with 1-step lookahead or full lookahead,

$$R(D) = \max_{p \in \mathcal{D}_{states}} R(p, D), \tag{22}$$

where

$$\mathcal{D}_{states} = \left\{ p \in \mathcal{P}(\mathcal{X}) : \begin{array}{c} p(\cdot) = \sum_{t \in \mathcal{T}} \alpha(t) f(\cdot|t) \\ f(\cdot|t) \in \overline{\mathcal{G}}(t), \forall\ t \in \mathcal{T} \end{array} \right\}. \tag{23}$$

Proof: See Appendix II. ■

One can see that in the case of the cheating switcher of the previous section, the set $\mathcal{D}$ of equation (20) equates directly with $\mathcal{D}_{states}$ of equation (23). In that sense, from the switcher's point of view, $\mathcal{D}$ is a more natural description of the set of distributions that can be simulated than $\mathcal{C}$. Again, computing $R(D)$ in (22) falls into the discussion of Section VII.

V. The Helpful Switcher

In general, the active-source may be acting in such a way that optimizes its own objectives. When its objective is to output a source sequence that is not well represented by the codebook, we arrive at the traditional adversarial setting considered above. The objective of the switcher, however, may vary from adversarial to agnostic to helpful. In this section, we consider the *helpful* cheating switcher. The model is as follows:

- The coder chooses a codebook that is made known to the switcher.
- The switcher chooses a strategy to help the coder achieve distortion D on average with the minimum number of codewords. We consider the cases where the switcher has full lookahead or 1-step lookahead.

As opposed to the adversarial setting, a rate R is now achievable at distortion D if *there exist* switcher strategies and codebooks for each n with expected distortion at most D and the rates of the codebooks tend to R. The following theorem establishes $R(D)$ if the cheating switcher has full lookahead.

Theorem 5.1: Let $\mathcal{X}^* = \{\mathcal{V} \subseteq \mathcal{X} : \mathcal{V} \neq \emptyset, |\mathcal{V}| \leq m\}$. Let $\rho : \mathcal{X}^* \times \widehat{\mathcal{X}} \to [0, d^*]$ be defined by

$$\rho(\mathcal{V}, \widehat{x}) = \min_{x \in \mathcal{V}} d(x, \widehat{x}). \tag{24}$$

Let $\mathcal{V}_k = \{x_{1,k}, \ldots, x_{m,k}\}$ for all k. Note that $\mathcal{V}_i, i = 1, 2, \ldots$ is a sequence of IID random variables with distribution $\beta(\mathcal{V}) = P(\{x_{1,1}, \ldots, x_{m,1}\} = \mathcal{V})$. Let $R^*(\beta, D)$ be the rate-distortion function for the IID source with distribution

β at distortion D with respect to the distortion measure $\rho(\cdot,\cdot)$. For the helpful cheating switcher with full lookahead,

$$R(D) = R^*(\beta, D). \tag{25}$$

Proof: Rate-distortion problems are essentially covering problems, so we equate the rate-distortion problem for the helpful switcher with the classical covering problem for the observed sets $\mathcal{V}_i$. If the switcher is helpful, has full lookahead, and knowledge of the codebook, the problem of designing the codebook is equivalent to designing the switcher strategy and codebook jointly. At each time k, the switcher observes a candidate set $\mathcal{V}_k$ and must select an element from $\mathcal{V}_k$. For any particular reconstruction codeword $\widehat{\mathbf{x}}^n$, and a string of candidate sets $(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n)$, the switcher can at best output a sequence $\mathbf{x}^n$ such that

$$d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) = \frac{1}{n}\sum_{k=1}^{n} \rho(\mathcal{V}_k, \widehat{x}_k) \tag{26}$$

Hence, for a codebook $\mathcal{B}$, the helpful switcher with full lookahead can select switch positions to output $\mathbf{x}^n$ such that

$$d_n(\mathbf{x}^n; \mathcal{B}) = \min_{\widehat{\mathbf{x}}^n \in \mathcal{B}} \frac{1}{n}\sum_{k=1}^{n} \rho(\mathcal{V}_k, \widehat{x}_k). \tag{27}$$

Therefore, for the helpful switcher, the problem of covering the $\mathcal{X}$ space with respect to the distortion measure $d(\cdot,\cdot)$ now becomes one of covering the $\mathcal{X}^*$ space with respect to the distortion measure $\rho(\cdot,\cdot)$. ■

Remarks:

- Computing $R(D)$ in (25) can be done by the Blahut-Arimoto algorithm[20].
- In the above proof, full lookahead was required in order for the switcher to align the entire output word of the source with the minimum distortion reconstruction codeword as a whole. This process cannot be done with 1-step lookahead and so the $R(D)$ function for a helpful switcher with 1-step lookahead remains an open question, but we have the following corollary of Theorems 3.1 and 5.1.

Corollary 5.1: For the helpful switcher with 1-step lookahead,

$$R^*(\beta, D) \leq R(D) \leq \min_{p \in \mathcal{C}} R(p, D) \tag{28}$$

Proof: If the switcher has at least 1-step lookahead, it immediately follows from the proof of Theorem 3.1 that $R(D) \leq \min_{p\in\mathcal{C}} R(p, D)$. The question is whether or not any lower rate is achievable. We can make the helpful switcher with 1-step lookahead more powerful by giving it n-step lookahead, which yields the lower bound $R^*(\beta, D)$. ■

An example in Section VI-B shows that in general, we have the strict inequality $R^*(\beta, D) < \min_{p\in\mathcal{C}} R(p, D)$.

One can also investigate the helpful switcher problem when the switcher has access to noisy or partial observations as in Section IV. This problem has the added flavor of remote source coding because the switcher can be thought of as an extension of the coder and observes data correlated with the source to be encoded. However, the switcher has the additional capability of choosing the subsource that must be encoded. For now, this problem is open and we can only say that $R(D) \leq \min_{p\in\mathcal{D}_{states}} R(p, D)$.

VI. Examples

We illustrate the results with several simple examples using binary alphabets and Hamming distortion, i.e. $\mathcal{X} = \widehat{\mathcal{X}} = \{0, 1\}$ and $d(x, \widehat{x}) = 1(x \neq \widehat{x})$. Recall that the rate-distortion function of an IID binary source with distribution $(p, 1-p)$, $p \in [0, \frac{1}{2}]$ is

$$R((1-p, p), D) = \begin{cases} h_b(p) - h_b(D) & D \in [0, p] \\ 0 & D > p \end{cases}, \tag{29}$$

where $h_b(p)$ is the binary entropy function (in bits for this section).

A. Bernoulli $1/4$ *and* $1/3$ *sources*

Let $m = 2$ so the switcher has access to two IID Bernoulli subsources. Subsource 1 outputs 1 with probability $1/4$ and subsource 2 outputs 1 with probability $1/3$, so $p_1 = (3/4, 1/4)$ and $p_2 = (2/3, 1/3)$. First, we consider the switcher as an adversary. Figure 3 shows this example in the traditional strictly causal setting of [5], where the switcher gets only outputs of the source after the switch position has been decided. Figure 4 shows the AVS in the noncausal setting, where the switcher has the subsource realizations before choosing the switch position.

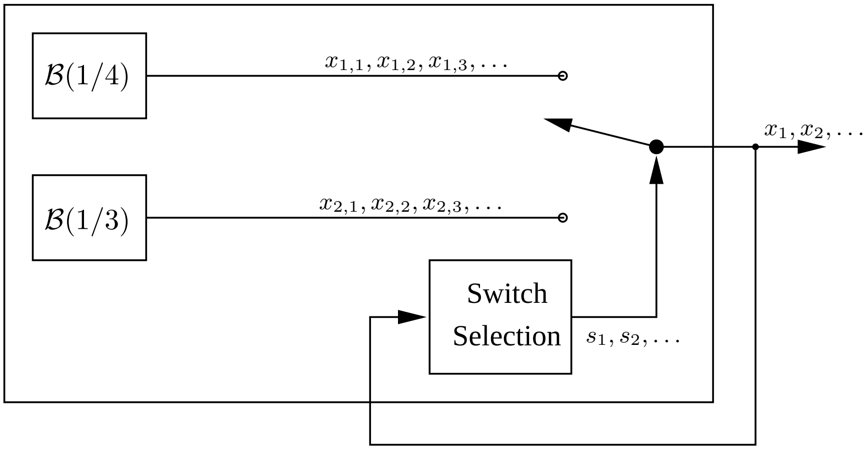

Fig. 3. The adversary chooses the switch position with knowledge only of the past AVS outputs. For Hamming distortion, the rate-distortion function is $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$.

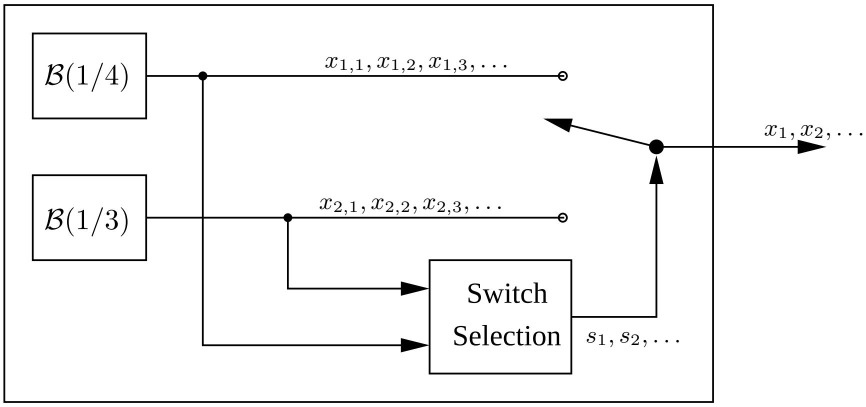

Fig. 4. The adversary chooses the switch position with knowledge of both subsource realizations. For Hamming distortion, the rate-distortion function is $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$.

For any time k,

$$P(x_{1,k} = x_{2,k} = 0) = \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2} \tag{30}$$

$$P(x_{1,k} = x_{2,k} = 1) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12} \tag{31}$$

$$P(\{x_{1,k}, x_{2,k}\} = \{0, 1\}) = 1 - \frac{1}{2} - \frac{1}{12} = \frac{5}{12}. \tag{32}$$

If the switcher is allowed 1-step lookahead and has the option of choosing either 0 or 1, suppose the switcher chooses 1 with probability f_1. The coder then sees an IID binary source with a probability of a 1 occurring being equal to:

$$p(1) = \frac{1}{12} + \frac{5}{12} f_1. \tag{33}$$

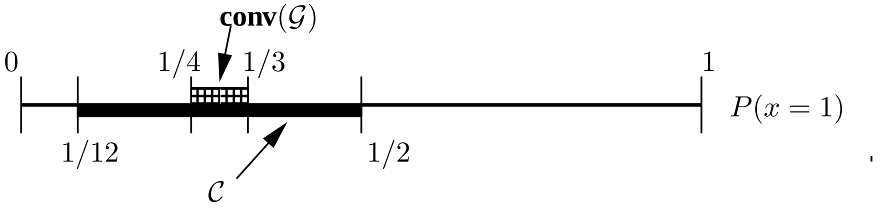

Fig. 5. The binary distributions the switcher can mimic. **conv**$(\mathcal{G})$ is the set of distributions the switcher can mimic with causal access to subsource realizations, and $\mathcal{C}$ is the set attainable with noncausal access.

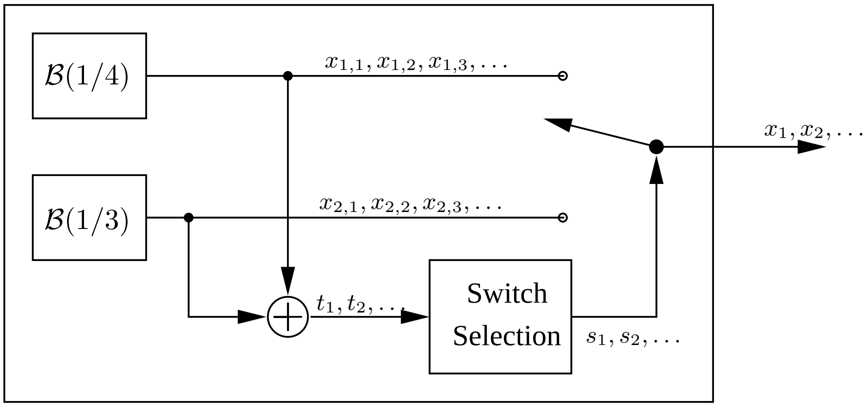

Fig. 6. The adversary observes the mod-2 sum of the two subsources, a Bernoulli $1/3$ subsource and a Bernoulli $1/4$ subsource. For Hamming distortion, the rate-distortion function is $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$.

By using f_1 as a parameter, the switcher can produce 1's with any probability between $1/12$ and $1/2$. The attainable distributions are shown in Figure 5. The switcher with lookahead can simulate a significantly larger set of distributions than the causal switcher, which is restricted to outputting 1's with probability in $[1/4, 1/3]$. Thus, for the strictly causal switcher, $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$ and for the switcher with 1-step or full lookahead, $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$.

We now look at several variations of this example to illustrate the utility of noisy or partial observations of the subsources for the switcher. In the first variation, shown in Figure 6, the switcher observes the mod-2 sum of the two subsources. Theorem 4.1 then implies that $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$. Hence, the mod-2 sum of these two subsources is useless to the switcher in deciding the switch position. This is intuitively clear from the symmetry of the mod-2 sum. If $t = 0$, either both subsources are 0 or both subsources are 1, so the switch position doesn't matter in this state. If $t = 1$, one of the subsources has output 1 and the other has output 0, but because of the symmetry of the mod-2 function, the switcher's prior as to which subsource output the 1 does not change and it remains that subsource 2 was more likely to have output the 1.

In the second variation, shown in Figure 7, the switcher observes the second subsource directly but not the first, so $t_k = x_{2,k}$ for all k. Using Theorem 4.1 again, it can be deduced that in this case $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$. This is also true if $t_k = x_{1,k}$ for all k, so observing just one of the subsources noncausally is as beneficial to the switcher as observing both subsources noncausally. This is clear in this example because the switcher is attempting to output as many 1's as possible. If $t = 1$, the switcher will set the switch position to 2 and if $t = 0$, the switcher will set the switch position to 1 as there is still a chance that the first subsource outputs a 1.

For this example, the helpful cheater with 1-step lookahead has a rate-distortion function that is upper bounded by $h_b(1/12) - h_b(D)$ for $D \in [0, 1/12]$. The rate-distortion function for the helpful cheater with full lookahead can be computed from Theorem 5.1. In Figure 8, the rate-distortion function is plotted for the situations discussed so far. In an active sensing situation, we see that there can be a large gap between the required rates for adversarially modelled active sensors and sensors which have been jointly optimized with the coding system.

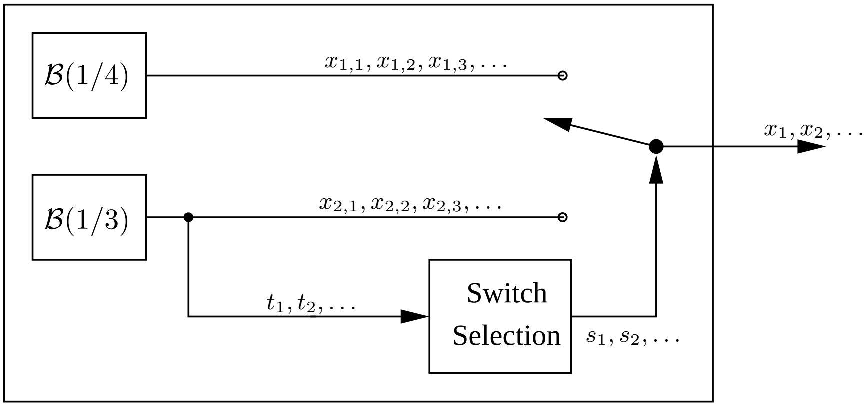

Fig. 7. The adversary observes the second subsource perfectly, but does not observe the first subsource. For Hamming distortion, the rate-distortion function is $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$.

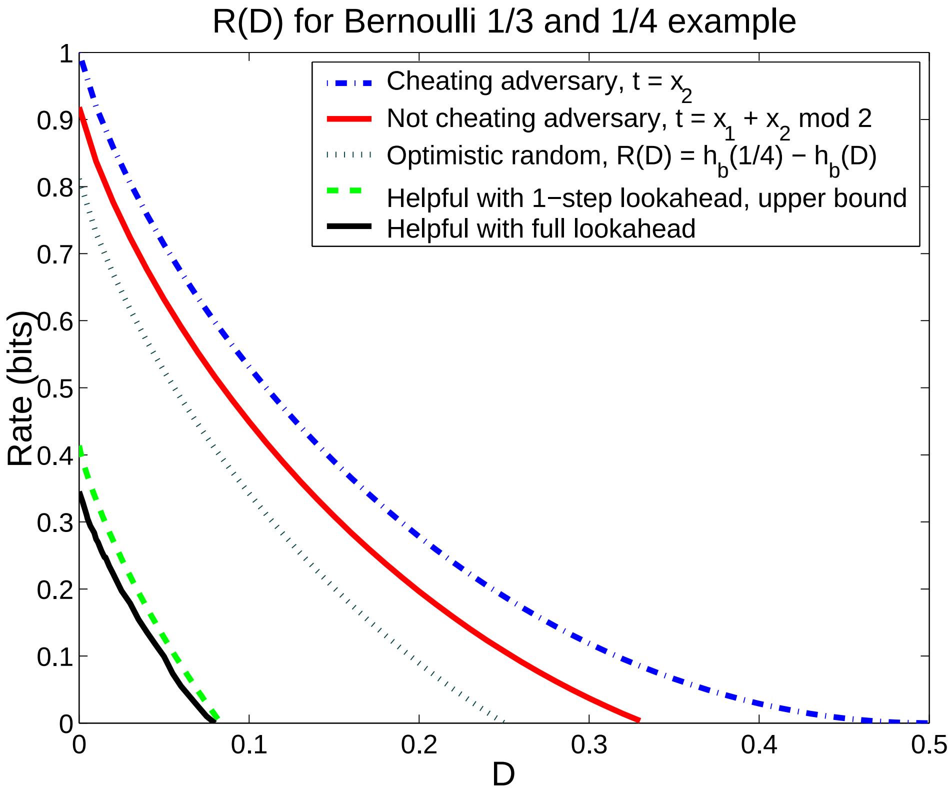

Fig. 8. $R(D)$ for the cheating switcher and the non-cheating switcher. Also, the rate-distortion function for the examples of Figures 6 and 7.

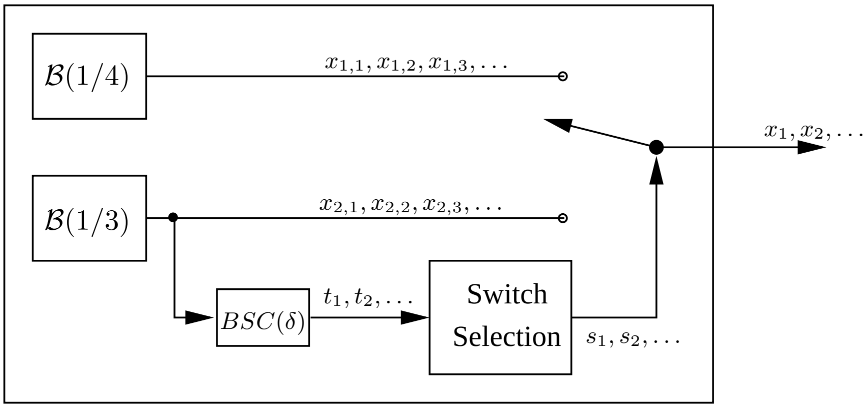

Fig. 9. The adversary observes the second subsource transmitted over a binary symmetric channel with crossover probability δ. For Hamming distortion, the rate-distortion function is $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$ if $\delta \in [2/5, 1/2]$. If $\delta \in [0, 2/5)$, $R(D) = h_b(1/2 - 5\delta/12) - h_b(D)$ for $D \in [0, 1/2 - 5\delta/12]$.

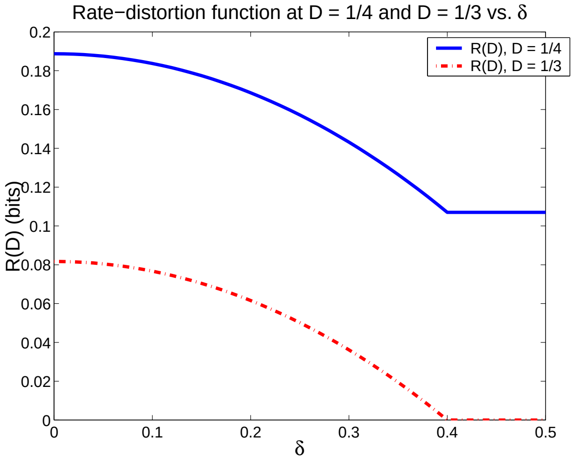

Fig. 10. $R(D)$ as a function of the noisy observation crossover probability δ for two different distortions for the example of Figure 9.

Finally, in Figure 9, an adversarial switcher observes the second subsource through a binary symmetric channel with crossover probability $\delta \in [0, 1/2]$. Applying Theorem 4.1 again, it can be shown that if $\delta \in [0, 2/5]$,

$$R(D) = h_b\left(\frac{1}{2} - \frac{5}{12}\delta\right) - h_b(D), \;\; D \in \left[0, \frac{1}{2} - \frac{5}{12}\delta\right] \tag{34}$$

and if $\delta \in [2/5, 1/2]$,

$$R(D) = h_b\left(\frac{1}{3}\right) - h_b(D), \;\; D \in \left[0, \frac{1}{3}\right]. \tag{35}$$

Here, increasing δ decreases the switcher's knowledge of the subsource realizations. Somewhat surprisingly, the utility of the observation is exhausted at $\delta = 2/5$, even before the state and observation are completely independent at $\delta = 1/2$. This can be explained through the switcher's *a posteriori* belief that second subsource output was a 1 given the state. If the switcher observes $t = 1$ and $\delta \leq 1/2$, $p(x_{2,k} = 1 | t_k = 1) \geq 1/3 > 1/4$ so the switch position will be set to 2. When the switcher observes $t = 0$, if $\delta \leq 2/5$, $p(x_{2,k} = 1 | t_k = 0) \leq 1/4$, so the switch will be set to position 1. However, if $\delta > 2/5$, $p(x_{2,k} = 1 | t_k = 0) > 1/4$, so the switch position will be set to 2 even if $t = 0$ because the switcher's *a posteriori* belief is that the second subsource is *still* more likely to have output a 1 than the first subsource. Figure 10 shows $R(D)$ for this example as a function of δ for two values of D.

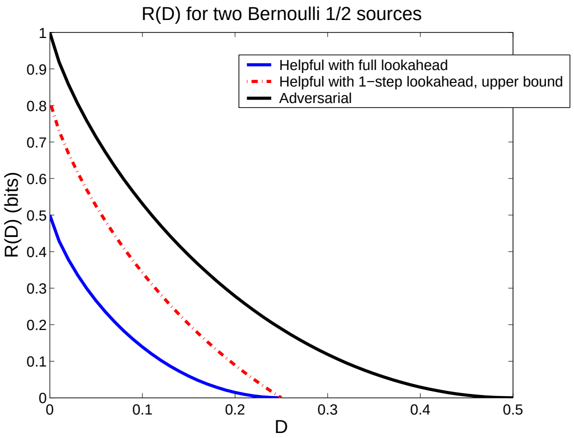

Fig. 11. The $R(D)$ function for a helpful switcher with full lookahead. For 1-step lookahead, the upper bound is shown.

B. *Two Bernoulli* $1/2$ *subsources*

Suppose $m = 2$, and both subsources are Bernoulli $1/2$ IID processes. For this example, the rate-distortion function is $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$ whether the adversarial switcher is strictly causal, causal or noncausal. When the helpful switcher has 1-step lookahead, $R(D) \leq R_U(D) = h_b(1/4) - h_b(D)$ for $D \in [0, 1/4]$. One can also think of this upper bound as being the rate-distortion function for the helpful switcher with 1-step lookahead that is restricted to using memoryless, time-invariant rules. Using Theorem 9.4.1 of [21], one can show that when the switcher has full lookahead,

$$R(D) = R^*(\beta, D) = \frac{1}{2}\left[1 - h_b(2D)\right], \; D \in [0, 1/4]. \tag{36}$$

The plot of these functions in Figure 11 shows that the rate-distortion function can be significantly reduced if the helpful switcher is allowed to observe the entire block of subsource realizations. It is also interesting to note *how* the switcher with full lookahead helps the coder achieve a rate of $R^*(\beta, D)$. In this example $\mathcal{X}^* = \{\{0\}, \{1\}, \{0, 1\}\}$, $\rho(\{0\}, \widehat{x}) = 1(0 \neq \widehat{x})$, $\rho(\{1\}, \widehat{x}) = 1(1 \neq \widehat{x})$, $\rho(\{0, 1\}, \widehat{x}) = 0$ and $\beta = (1/4, 1/4, 1/2)$. The $R^*(\beta, D)$ achieving distribution on $\widehat{\mathcal{X}}$ is $(1/2, 1/2)$, but $R^*(\beta, D) < 1 - h_b(D)$. The coder is attempting to cover strings with types near $(1/2, 1/2)$ but with far fewer codewords than are needed to do so. This problem is circumvented through the aid provided by the switcher in pushing the output of the source inside the Hamming D-ball of a codeword. This is in contrast to the strategy that achieves $R_U(D)$, where the switcher makes the output an IID sequence with as few 1's as possible and the coder is expected to cover *all* strings with types near $(3/4, 1/4)$.

VII. Computing $R(D)$ for an AVS

The $R(D)$ function for an AVS with either causal or noncausal access to the subsource realizations is of the form

$$R(D) = \max_{p \in \mathcal{Q}} R(p, D), \tag{37}$$

where $\mathcal{Q}$ is a set of distributions in $\mathcal{P}(\mathcal{X})$. In (17), (19), and (23) $\mathcal{Q}$ is defined by a finite number of linear inequalities and hence is a polytope. The number of constraints in the definition of $\mathcal{Q}$ is exponential in $|\mathcal{X}|$ or $|\mathcal{T}|$ when the adversary has something other than strictly causal knowledge. Unfortunately, the problem of finding $R(D)$ is not a convex program because $R(p, D)$ is not a concave $\cap$ function of p in general. In fact, $R(p, D)$ may not even be quasi-concave and may have multiple local maxima with values different from the global maximum as shown by Ahlswede [22].

Since standard convex optimization tools are unavailable for this problem, we consider the question of how to approximate $R(D)$ to within some (provable) precision. That is, for any $\epsilon > 0$, we will consider how to provide an approximation $R_a(D)$ such that $|R_a(D) - R(D)| \leq \epsilon$. Note that for fixed p, $R(p, D)$ can be computed efficiently by the Blahut-Arimoto algorithm to any given precision, say much less than ϵ. Therefore, we assume that $R(p, D)$

can be computed for a fixed p and D. We also assume $D \geq D_{\min}(\mathcal{Q})$ since otherwise $R(D) = \infty$. Checking this condition is a linear program since $\mathcal{Q}$ is a polytope and $D_{\min}(p)$ is linear in p.

We will take a 'brute-force' approach to computing $R(D)$. That is, we wish to compute $R(p, D)$ for (finitely) many p and then maximize over the computed values to yield $R_a(D)$. Since $R(p, D)$ is uniformly continuous in (p, D) and hence in p, it is possible to do this and have $|R_a(D) - R(D)| \leq \epsilon$ provided enough distributions p are 'sampled'. Undoubtedly, there are other algorithms to compute $R(D)$ that likely have better problem-size dependence. In this section, we are only interested in showing that $R(D)$ can provably be computed to within any required precision with a finite number of computations.

A. *Uniform continuity of* $R(p, D)$

The main tool used to show that the rate-distortion function can be approximated is an explicit bound on the uniform continuity of $R(p, D)$ in terms of $\|p - q\|_1 = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$ for distortion measures that allow for 0-distortion to be achieved regardless of the source. In [20], a bound on the continuity of the entropy of a distribution is developed in terms of $\|p - q\|_1$.

Lemma 7.1 ($\mathcal{L}_1$ bound on continuity of entropy [20]): Let p and q be two probability distributions on $\mathcal{X}$ such that $\|p - q\|_1 \leq 1/2$, then

$$|H(p) - H(q)| \leq \|p - q\|_1 \ln \frac{|\mathcal{X}|}{\|p - q\|_1}. \tag{38}$$

In the following lemma, a similar uniform continuity is stated for $R(p, D)$. The proof makes use of Lemma 7.1.

Lemma 7.2 (Uniform continuity of $R(p, D)$): Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion function. $\widetilde{d}$ is the minimum nonzero distortion from (1). Also, assume that for each $x \in \mathcal{X}$, there is an $\hat{x}_0(x) \in \widehat{\mathcal{X}}$ such that $d(x, \hat{x}_0(x)) = 0$. Then, for $p, q \in \mathcal{P}(\mathcal{X})$ with $\|p - q\|_1 \leq \frac{\widetilde{d}}{4d^*}$, for any $D \geq 0$,

$$|R(p, D) - R(q, D)| \leq \frac{7d^*}{\widetilde{d}} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \tag{39}$$

Proof: See Appendix III. ■

The restriction that $d(x, \cdot)$ has at least one zero for every x can be relaxed if we are careful about recognizing when $R(p, D)$ is infinite. For an arbitrary distortion measure $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$, define

$$d_0(x, \widehat{x}) = d(x, \widehat{x}) - \min_{\widetilde{x} \in \widehat{\mathcal{X}}} d(x, \widetilde{x}). \tag{40}$$

Now let $d_0^* = \max_{x, \widehat{x}} d_0(x, \widehat{x})$ and $\widetilde{d}_0 = \min_{(x, \widehat{x}) : d_0(x, \widehat{x}) > 0} d_0(x, \widehat{x})$. We have defined $d_0(x, \widehat{x})$ so that Lemma 7.2 applies, so we can prove the following lemma.

Lemma 7.3: Let $p, q \in \mathcal{P}(\mathcal{X})$ and let $D \geq \max(D_{\min}(p), D_{\min}(q))$. If $\|p - q\|_1 \leq \widetilde{d}_0 / 4d^*$,

$$|R(p, D) - R(q, D)| \leq \frac{11d^*}{\widetilde{d}_0} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \tag{41}$$

Proof: See Appendix IV. ■

As $\|p - q\|_1$ goes to 0, $-\ln \|p - q\|_1$ goes to infinity slowly and it can be shown that for any $\delta \in (0, 1)$ and $\gamma \in [0, 1/2]$,

$$\gamma \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\gamma} \leq \frac{(|\mathcal{X}||\widehat{\mathcal{X}}|)^\delta}{e\delta} \gamma^{1-\delta}. \tag{42}$$

In the sequel, we let $f(\gamma) = \gamma \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\gamma}$ for $\gamma \in [0, 1/2]$ with $f(0) = 0$ by continuity. It can be checked that f is strictly monotonically increasing and continuous on $[0, 1/2]$ and hence has an inverse function $g : f([0, 1/2]) \to [0, 1/2]$, i.e. $g(f(\gamma)) = \gamma$ for all $\gamma \in [0, 1/2]$. Note that g is not expressible in a simple 'closed-form', but can be computed numerically.

B. A bound on the number of distributions to sample

Returning to the problem of computing $R(D)$ in equation (37), consider the following simple algorithm. Without loss of generality, assume $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. Let $\gamma \in (0,1)$ and let $\gamma\mathbb{Z}^{|\mathcal{X}|-1}$ be the $|\mathcal{X}|-1$ dimensional integer lattice scaled by γ. Let $\widetilde{\mathcal{O}} = [0,1]^{|\mathcal{X}|-1} \bigcap \gamma\mathbb{Z}^{|\mathcal{X}|-1}$. Now, define

$$\mathcal{O} = \left\{ q \in \mathcal{P}(\mathcal{X}) : \begin{array}{c} \exists\, \widetilde{q} \in \widetilde{\mathcal{O}}, \\ q(i) = \widetilde{q}(i), i = 1, \ldots, |\mathcal{X}|-1, \\ q(|\mathcal{X}|) = 1 - \sum_{i=1}^{|\mathcal{X}|-1} \widetilde{q}(i) \geq 0 \end{array} \right\}. \tag{43}$$

In words, sample the $|\mathcal{X}|-1$ dimensional unit cube, $[0,1]^{|\mathcal{X}|-1}$, uniformly with points from a scaled integer lattice. Embed these points in $\mathbb{R}^{|\mathcal{X}|}$ by assigning the last value of the new vector to be 1 minus the sum of the values in the original point. If this last value is non-negative, the new point is a distribution in $\mathcal{P}(\mathcal{X})$. The algorithm to compute $R_a(D)$ is then one where we compute $R(p,D)$ for distributions $q \in \mathcal{O}$ that are also in or close enough to $\mathcal{Q}$.

1) Fix a $q \in \mathcal{O}$. If $\min_{p\in\mathcal{Q}} \|p-q\|_1 \leq 2|\mathcal{X}|\gamma$, compute $R(q,D)$, otherwise do not compute $R(q,D)$. Repeat for all $q \in \mathcal{O}$.
2) Let $R_a(D)$ be the maximum of the computed values of $R(q,D)$, i.e.

$$R_a(D) = \max\left\{ R(q,D) : q \in \mathcal{O}, \min_{p\in\mathcal{Q}} \|p-q\|_1 \leq 2|\mathcal{X}|\gamma \right\}. \tag{44}$$

Checking the condition $\min_{p\in\mathcal{Q}} \|p-q\|_1 \leq \gamma 2|\mathcal{X}|$ is essentially a linear program, so it can be efficiently solved. By setting γ according to the accuracy $\epsilon > 0$ we want, we get the following result.

Theorem 7.1: The preceding algorithm computes an approximation $R_a(D)$ such that $|R_a(D) - R(D)| \leq \epsilon$ if

$$\gamma \leq \frac{1}{2|\mathcal{X}|} g\left(\frac{\epsilon \widetilde{d}_0}{11 d^*}\right). \tag{45}$$

The number of distributions for which $R(q,D)$ is computed to determine $R(D)$ to within accuracy ϵ is at most[6]

$$N(\epsilon) \leq \left(\frac{2|\mathcal{X}|}{g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right)} + 2 \right)^{|\mathcal{X}|-1}. \tag{46}$$

Proof: The bound on $N(\epsilon)$ is clear because the number of points in $\widetilde{\mathcal{O}}$ is at most $(\lceil 1/\gamma \rceil + 1)^{|\mathcal{X}|-1}$ and every distribution in $\mathcal{O}$ is associated with one in $\widetilde{\mathcal{O}}$, so $|\mathcal{O}| \leq |\widetilde{\mathcal{O}}|$.

Now, we prove $|R_a(D) - R(D)| \leq \epsilon$. For this discussion, we let $\gamma = \frac{1}{2|\mathcal{X}|} g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right)$. First, for all $p \in \mathcal{Q}$, there is a $q \in \mathcal{O}$ with $\|p-q\|_1 \leq g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right) = 2|\mathcal{X}|\gamma$. To see this, let $\widetilde{q}(i) = \lfloor \frac{p(i)}{\gamma} \rfloor \gamma$ for $i = 1, \ldots, |\mathcal{X}|-1$. Then $\widetilde{q} \in \widetilde{\mathcal{O}}$, and we let $q(i) = \widetilde{q}(i)$ for $i = 1, \ldots, |\mathcal{X}|-1$. Note that

$$q(|\mathcal{X}|) = 1 - \sum_{i=1}^{|\mathcal{X}|-1} q(i) = 1 - \sum_{i=1}^{|\mathcal{X}|-1} \left\lfloor \frac{p(i)}{\gamma} \right\rfloor \gamma \geq 1 - \sum_{i=1}^{|\mathcal{X}|-1} p(i) = p(|\mathcal{X}|) \geq 0. \tag{47}$$

Therefore $q \in \mathcal{O}$ and furthermore,

$$\begin{aligned} \|p-q\|_1 &\leq \left(1 - \sum_{i=1}^{|\mathcal{X}|-1} (p(i) - \gamma) - p(|\mathcal{X}|)\right) + \sum_{i=1}^{|\mathcal{X}|-1} \left(p(i) - \left\lfloor \frac{p(i)}{\gamma} \right\rfloor \gamma\right) &\quad (48) \\ &\leq 2(|\mathcal{X}|-1)\gamma &\quad (49) \\ &\leq 2|\mathcal{X}|\gamma &\quad (50) \\ &\leq g\left(\frac{\epsilon\widetilde{d}_0}{11d^*}\right). &\quad (51) \end{aligned}$$

[6] This is clearly not the best bound as many of the points in the unit cube on do not yield distributions on $\mathcal{P}(\mathcal{X})$. The factor by which we are overbounding is roughly $|\mathcal{X}|!$, but this factor does not affect the dependence on ϵ.

By Lemma 7.3, $R(q, D) \geq R(p, D) - \epsilon$. This distribution q (or possibly one closer to p) will always be included in the maximization yielding $R_a(D)$, so we have $R_a(D) \geq \max_{p \in \mathcal{Q}} R(p, D) - \epsilon = R(D) - \epsilon$.

Conversely, for a $q \in \mathcal{O}$, if $\min_{p \in \mathcal{Q}} \|p - q\|_1 \leq 2|\mathcal{X}|\gamma$, Lemma 7.3 again gives

$$R(q, D) \leq \max_{p \in \mathcal{Q}} R(p, D) + \epsilon = R(D) + \epsilon \tag{52}$$

Therefore, $|R_a(D) - R(D)| \leq \epsilon$. ■

C. Estimation of the rate-distortion function of an unknown IID source

An explicit bound on the continuity of the rate-distortion function has other applications. Recently, Harrison and Kontoyiannis [23] have studied the problem of estimating the rate-distortion function of the marginal distribution of an unknown source. Let $p_{\mathbf{x}^n}$ be the (marginal) empirical distribution of a vector $\mathbf{x}^n \in \mathcal{X}^n$. They show that the 'plug-in' estimator $R(p_{\mathbf{x}^n}, D)$, the rate-distortion function of the empirical marginal distribution of a sequence, is a consistent estimator for a large class of sources beyond just IID sources with known alphabets. However, if the source is known to be IID with alphabet size $|\mathcal{X}|$, estimates of the convergence rate (in probability) of the estimator can be provided using the uniform continuity of the rate-distortion function.

Suppose the true source is IID with distribution $p \in \mathcal{P}(\mathcal{X})$ and fix a probability $\tau \in (0, 1)$ and an $\epsilon \in (0, \ln |\mathcal{X}|)$. We wish to answer the question: How many samples n need to be taken so that $|R(p_{\mathbf{x}^n}, D) - R(p, D)| \leq \epsilon$ with probability at least $1 - \tau$? The following lemma gives a sufficient number of samples n.

Theorem 7.2: Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion measure for which Lemma 7.2 holds. For any $p \in \mathcal{P}(\mathcal{X})$, $\tau \in (0, 1)$, and $\epsilon \in (0, \ln |\mathcal{X}|)$, then

$$P(|R(p_{\mathbf{x}^n}, D) - R(p, D)| \geq \epsilon) \leq \tau \tag{53}$$

if

$$n > \frac{2}{g\left(\frac{\epsilon \widetilde{d}}{7d^*}\right)^2} \left(\ln \frac{1}{\tau} + |\mathcal{X}| \ln 2 \right). \tag{54}$$

Proof: From Lemma 7.2, we have

$$\begin{aligned} P(|R(p_{\mathbf{x}^n}, D) - R(p, D)| \geq \epsilon) \quad &\leq \quad P\left(\|p_{\mathbf{x}^n} - p\|_1 \geq g\left(\frac{\epsilon \widetilde{d}}{7d^*} \right) \right) && (55) \\ &\leq \quad 2^{|\mathcal{X}|} \exp\left(-\frac{n}{2} g\left(\frac{\epsilon \widetilde{d}}{7d^*} \right)^2 \right) && (56) \end{aligned}$$

The last line follows from Theorem 2.1 of [24]. This bound is similar to, but a slight improvement over, the method-of-types bound of Sanov's Theorem. Rather than an $(n+1)^{|\mathcal{X}|}$ term, we just have a $2^{|\mathcal{X}|}$ term multiplying the exponential. Taking $\ln$ of both sides gives the desired result. ■

We emphasize that this number n is a sufficient number of samples regardless of what the true distribution $p \in \mathcal{P}(\mathcal{X})$ is. The bound of (54) depends only on the distortion measure d, alphabet sizes $|\mathcal{X}|$ and $|\widehat{\mathcal{X}}|$, desired accuracy ϵ and 'estimation error' probability τ.

VIII. Concluding Remarks

As mentioned in the introduction, the active-source problem is truly interesting when the sources have memory. Dobrushin [25] has analyzed the case of the non-anticipatory AVS composed of independent sources with memory with different distributions when the switcher is passive and blindly chooses the switch position. In the case of sources with memory, additional knowledge will no doubt increase the adversary's power to increase the rate-distortion function. If we let $R^{(k)}(D)$ be the rate-distortion function for an AVS composed of sources with memory and an adversary with k step lookahead, one could imagine that in general,

$$R^{(0)}(D) < R^{(1)}(D) < R^{(2)}(D) < \cdots < R^{(\infty)}(D). \tag{57}$$

Another interesting problem, at least mathematically, is the arbitrarily varying channel formulation analogous to the problems of Sections III and IV. Similar techniques to those developed here might prove useful in considering a cheating 'jammer' for an arbitrarily varying channel. While the problem is well defined, it seems unphysical in the usual context of jamming or channel noise. The idea may make more sense in the context of watermarking, where the adversary can try many different attacks on different letters of the input before deciding to choose one for each.

For the original motivation of compressing active-vision sources, the results here suggest that treating it as an adversarial black box might be overly conservative. There is a large gap between the adversarial and helpful rate-distortion functions. This suggests that an interesting question to study is one of mismatched objectives where the switcher is trying to be helpful for some particular distortion metric but the source is actually being encoded with a different metric in mind. Finally, if the active-sensor and coding system are part of a tightly delay-constrained control loop, we would want to study these issues from the causal source code perspective of [13]. It seems likely that the adversarial results of Theorems 3.1 and 4.1 would follow straightforwardly with the same sets of distributions $\mathcal{C}$ and $\mathcal{D}$, with the IID rate-distortion function for noncausal source codes replaced by the the IID rate-distortion functions for causal source codes.

Appendix I
Proof of Theorem 3.1

A. Achievability for the coder

The main tool of the proof is:

Lemma 1.1 (Type Covering): Let $S_D(\widehat{\mathbf{x}}^n) \triangleq \{\mathbf{x}^n \in \mathcal{X}^n : d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) \leq D\}$ be the set of $\mathcal{X}^n$ strings that are within distortion D of a $\widehat{\mathcal{X}}^n$ string $\widehat{\mathbf{x}}^n$. Fix a $p \in \mathcal{P}_n(\mathcal{X})$ and an $\epsilon > 0$. Then for all n large enough, there exist codebooks $\mathcal{B} = \{\widehat{\mathbf{x}}^n(1), \widehat{\mathbf{x}}^n(2), \ldots, \widehat{\mathbf{x}}^n(M)\}$ where $M < \exp(n(R(p,D) + \epsilon))$ and

$$T_p^n \subseteq \bigcup_{\widehat{\mathbf{x}}^n \in \mathcal{B}} S_D(\widehat{\mathbf{x}}^n), \tag{58}$$

where T_p^n is the set of $\mathcal{X}^n$ strings with type p.

Proof: See [5], Lemma 1. ■

We now show how the coder can get arbitrarily close to $\widetilde{R}(D)$ for large enough n. For $\delta > 0$, define $\mathcal{C}_\delta$ as

$$\mathcal{C}_\delta \triangleq \left\{ \begin{array}{ccc} & & \sum_{x \in \mathcal{V}} p(x) \geq P(x_l \in \mathcal{V}, 1 \leq l \leq m) - \delta \\ p \in \mathcal{P}(\mathcal{X}) & : & \forall\ \mathcal{V} \text{ such that} \\ & & \mathcal{V} \subseteq \mathcal{X} \end{array} \right\}. \tag{59}$$

Lemma 1.2 (Converse for switcher): Let $\epsilon > 0$. For all n sufficiently large

$$\frac{1}{n} \ln M(n,D) \leq \widetilde{R}(D) + \epsilon. \tag{60}$$

Proof: We know $R(p,D)$ is a continuous function of p ([19]). It follows then that because $\mathcal{C}_\delta$ is monotonically decreasing (as a set) with δ that for all $\epsilon > 0$, there is a $\delta > 0$ so that

$$\max_{p \in \mathcal{C}_\delta} R(p,D) \leq \max_{p \in \mathcal{C}} R(p,D) + \epsilon/2. \tag{61}$$

We will have the coder use a codebook such that all $\mathcal{X}^n$ strings with types in $\mathcal{C}_\delta$ are covered within distortion D. The coder can do this for large n with at most M codewords in the codebook $\mathcal{B}$, where

$$\begin{aligned} M &< (n+1)^{|\mathcal{X}|} \exp(n \max_{p \in \mathcal{C}_\delta} R(p,D)) \qquad (62) \\ &\leq \exp(n(\max_{p \in \mathcal{C}} R(p,D) + \epsilon)). \qquad (63) \end{aligned}$$

Explicitly, this is done by taking a union of the codebooks provided by the type-covering lemma and noting that the number of types in $\mathcal{P}_n(\mathcal{X})$ is less than $(n+1)^{|\mathcal{X}|}$. Next, we will show that the probability of the switcher being able to produce a string with a type not in $\mathcal{C}_\delta$ goes to 0 exponentially with n.

Consider a type $p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)$. By definition, there is some $\mathcal{V} \subseteq \mathcal{X}$ such that $\sum_{x \in \mathcal{V}} p(x) < P(x_l \in \mathcal{V}, 1 \leq l \leq m) - \delta$. Let $\zeta_k(\mathcal{V})$ be the indicator function

$$\zeta_k(\mathcal{V}) = \prod_{l=1}^{m} \mathbf{1}(x_{l,k} \in \mathcal{V}). \tag{64}$$

ζ_k indicates the event that the switcher cannot output a symbol outside of $\mathcal{V}$ at time k. Then $\zeta_k(\mathcal{V})$ is a Bernoulli random variable with a probability of being 1 equal to $\kappa(\mathcal{V}) \triangleq P(x_l \in \mathcal{V}, 1 \leq l \leq m)$. Since the subsources are IID over time, $\zeta_k(\mathcal{V})$ is a sequence of IID binary random variables with distribution $q' \triangleq (1-\kappa(\mathcal{V}), \kappa(\mathcal{V}))$.

Now for the type $p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)$, we have that for all strings $\mathbf{x}^n$ in the type class T_p, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i \in \mathcal{V}) < \kappa(\mathcal{V}) - \delta$. Let p' be the binary distribution $(1-\kappa(\mathcal{V})+\delta, \kappa(\mathcal{V})-\delta)$. Therefore $||p'-q'||_1 = 2\delta$, and hence we can bound the binary divergence $D(p'||q') \geq 2\delta^2$ by Pinsker's inequality. Using standard types properties [20] gives

$$P\left(\frac{1}{n}\sum_{k=1}^n \zeta_k(\mathcal{V}) < \kappa(\mathcal{V}) - \delta\right) \leq (n+1)\exp(-nD(p'||q')) \quad (65)$$

$$\leq (n+1)\exp(-2n\delta^2). \quad (66)$$

This bound holds for all $\mathcal{V} \subset \mathcal{X}, \mathcal{V} \neq \emptyset$, so we sum over types not in $\mathcal{C}_\delta$ to get

$$P(p_{\mathbf{x}^n} \notin \mathcal{C}_\delta) \leq \sum_{p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)} (n+1)\exp(-2n\delta^2) \quad (67)$$

$$\leq (n+1)^{|\mathcal{X}|}\exp(-2n\delta^2) \quad (68)$$

$$= \exp\left(-n\left(2\delta^2 - |\mathcal{X}|\frac{\ln(n+1)}{n}\right)\right). \quad (69)$$

Then, regardless of the switcher strategy,

$$\mathbb{E}[d(\mathbf{x}^n; \mathcal{B})] \leq D + d^* \cdot \exp\left(-n\left(2\delta^2 - |\mathcal{X}|\frac{\ln(n+1)}{n}\right)\right). \quad (70)$$

So for large n we can get arbitrarily close to distortion D while the rate is at most $\widetilde{R}(D) + \epsilon$. Using the fact that the IID rate-distortion function is continuous in D gives us that the coder can achieve at most distortion D on average while the asymptotic rate is at most $\widetilde{R}(D) + \epsilon$. Since ϵ is arbitrary, $R(D) \leq \widetilde{R}(D)$. ■

B. Achievability for the switcher

This section shows that $R(D) \geq \widetilde{R}(D)$ when the switcher has 1-step lookahead. We will show that the switcher can target any distribution $p \in \mathcal{C}$ and produce a sequence of IID symbols with distribution p. In particular, the switcher can target the distribution that yields $\max_{p\in\mathcal{C}} R(p, D)$, so $R(D) \geq \widetilde{R}(D)$.

The switcher will use a memoryless randomized strategy. Let $\mathcal{V} \subseteq \mathcal{X}$ and suppose that at some time k the set of symbols available to choose from for the switcher is exactly $\mathcal{V}$, i.e. $\{x_{1,k}, \ldots, x_{m,k}\} = \mathcal{V}$. Recall $\beta(\mathcal{V}) \triangleq P(\{x_{1,1}, \ldots, x_{m,1}\} = \mathcal{V})$ is the probability that at any time the switcher must choose among elements of $\mathcal{V}$ and no other symbols. Then let $f(x|\mathcal{V})$ be a probability distribution on $\mathcal{X}$ with support $\mathcal{V}$, i.e. $f(x|\mathcal{V}) \geq 0,\ \forall\ x \in \mathcal{X}$, $f(x|\mathcal{V}) = 0$ if $x \notin \mathcal{V}$, and $\sum_{x\in\mathcal{V}} f(x|\mathcal{V}) = 1$. The switcher will have such a randomized rule for every nonempty subset $\mathcal{V}$ of $\mathcal{X}$ such that $|\mathcal{V}| \leq m$. Let $\mathcal{D}$ be the set of distributions on $\mathcal{X}$ that can be achieved with these kinds of rules,

$$\mathcal{D} = \left\{ p \in \mathcal{P}(\mathcal{X}) \ : \ \begin{array}{c} p(\cdot) = \sum_{\mathcal{V}\subseteq\mathcal{X}, |\mathcal{V}|\leq m} \beta(\mathcal{V}) f(\cdot|\mathcal{V}), \\ \forall\ \mathcal{V} \text{ s.t. } \mathcal{V} \subseteq \mathcal{X},\ |\mathcal{V}| \leq m, \\ f(\cdot|\mathcal{V}) \text{ is a PMF on } \mathcal{V} \end{array} \right\}. \quad (71)$$

It is clear by construction that $\mathcal{D} \subseteq \mathcal{C}$ because the conditions in $\mathcal{C}$ are those that only prevent the switcher from producing symbols that do not occur enough on average, but put no further restrictions on the switcher. So we need only show that $\mathcal{C} \subseteq \mathcal{D}$. The following gives such a proof by contradiction.

Lemma 1.3 (Achievability for switcher): The set relation $\mathcal{C} \subseteq \mathcal{D}$ is true.

Proof: Without loss of generality, let $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$. Suppose $p \in \mathcal{C}$ but $p \notin \mathcal{D}$. It is clear that $\mathcal{D}$ is a convex set. Let us view the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$. Since $\mathcal{D}$ is a convex set, there is a hyperplane through p that does not intersect $\mathcal{D}$. Hence, there is a vector $(a_1, \ldots, a_{|\mathcal{X}|})$ such that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) = t$ for some real t but $t < \min_{q\in\mathcal{D}} \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$. Without loss of generality, assume $a_1 \geq a_2 \geq \ldots \geq a_{|\mathcal{X}|}$ (otherwise permute symbols).

Now, we will construct $f(\cdot|\mathcal{V})$ so that the resulting q has $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) \geq \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$, which contradicts the initial assumption. Let

$$f(i|\mathcal{V}) \triangleq \begin{cases} 1 & \text{if } i = \max(\mathcal{V}) \\ 0 & \text{else} \end{cases}, \tag{72}$$

so for example, if $\mathcal{V} = \{1, 5, 6, 9\}$, then $f(9|\mathcal{V}) = 1$ and $f(i|\mathcal{V}) = 0$ if $i \neq 9$. Call q the distribution on $\mathcal{X}$ induced by this choice of $f(\cdot|\mathcal{V})$. Recall that $\kappa(\mathcal{V}) = P(x_l \in \mathcal{V}, 1 \leq l \leq m)$. Then, we have

$$\begin{aligned} \sum_{i=1}^{|\mathcal{X}|} a_i q(i) \;&=\; a_1 \kappa(\{1\}) + a_2[\kappa(\{1,2\}) - \kappa(\{1\})] + \\ &\quad \cdots + a_{|\mathcal{X}|}[\kappa(\{1, \ldots, |\mathcal{X}|\}) - \kappa(\{1, \ldots, |\mathcal{X}| - 1\})] \end{aligned} \tag{73}$$

By the constraints in the definition (19) of $\mathcal{C}$, we have the following inequalities for p:

$$p(1) \;\geq\; \kappa(\{1\}) = q(1) \tag{74}$$

$$p(1) + p(2) \;\geq\; \kappa(\{1,2\}) = q(1) + q(2) \tag{75}$$

$$\vdots$$

$$\sum_{i=1}^{|\mathcal{X}|-1} p(i) \;\geq\; \kappa(\{1, \ldots, |\mathcal{X}| - 1\}) = \sum_{i=1}^{|\mathcal{X}|-1} q(i). \tag{76}$$

Therefore, the difference of the objective is

$$\begin{aligned} \sum_{i=1}^{|\mathcal{X}|} a_i(p(i) - q(i)) \;&=\; a_{|\mathcal{X}|}\left[\sum_{i=1}^{|\mathcal{X}|} p(i) - q(i)\right] + \\ &\quad (a_{|\mathcal{X}|-1} - a_{|\mathcal{X}|})\left[\sum_{i=1}^{|\mathcal{X}|-1} p(i) - q(i)\right] + \\ &\quad \cdots + (a_1 - a_2)\Big[p(1) - q(1)\Big] \end{aligned} \tag{77}$$

$$= \sum_{i=1}^{|\mathcal{X}|-1} (a_i - a_{i+1})\left[\sum_{j=1}^{i} p(j) - \sum_{j=1}^{i} q(j)\right] \tag{78}$$

$$\geq 0. \tag{79}$$

The last step is true because of the monotonicity in the a_i and the inequalities we derived earlier. Therefore, we see that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) \geq \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$ for the p we had chosen at the beginning of the proof. This contradicts the assumption that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) < \min_{q \in \mathcal{D}} \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$, therefore it must be that $\mathcal{C} \subseteq \mathcal{D}$. ■

APPENDIX II
PROOF OF THEOREM 4.1

It is clear that $R(D) \geq \max_{p \in \mathcal{D}_{states}} R(p, D)$ because the switcher can select distributions $f(\cdot|t) \in \overline{\mathcal{G}}(t)$ for all $t \in \mathcal{T}$ and upon observing a state t, the switcher can randomly select the switch position according to the convex combination that yields $f(\cdot|t)$. With this strategy, the AVS is simply an IID source with distribution $p(\cdot) = \sum_t \alpha(t) f(\cdot|t)$. Hence, $R(D) \geq \max_{p \in \mathcal{D}_{states}} R(p, D)$.

We will now show that $R(D) \leq \max_{p \in \mathcal{D}_{states}} R(p, D)$. This can be done in the same way as in Appendix I. We can use the type covering lemma to cover sequences with types in or very near $\mathcal{D}_{states}$ and then we need only show that the probability of $\mathbf{x}^n$ having a type ϵ far from $\mathcal{D}_{states}$ goes to 0 with block length n.

Lemma 2.1: Let $p_{\mathbf{x}^n}$ be the type of $\mathbf{x}^n$ and for $\epsilon > 0$ let $\mathcal{D}_{states,\epsilon}$ be the set of $p \in \mathcal{P}(\mathcal{X})$ with $\mathcal{L}_1$ distance at most ϵ from a distribution in $\mathcal{D}_{states}$. Then, for $\epsilon > 0$,

$$P(p_{\mathbf{x}^n} \notin \mathcal{D}_{states,\epsilon}) \leq 4|\mathcal{T}||\mathcal{X}| \exp(-n\xi(\epsilon)), \tag{80}$$

where $\xi(\epsilon) > 0$ for all $\epsilon > 0$. So for large n, $p_{\mathbf{x}^n}$ is in $\mathcal{D}_{states,\epsilon}$ with high probability.

Proof: Let $\mathbf{t}^n$ be the n-length vector of the observed states. We assume that the switcher has advance knowledge of all these states before choosing the switch positions. First, we show that with high probability, the states that are observed are strongly typical. Let $N(t|\mathbf{t}^n)$ be the count of occurrence of $t \in \mathcal{T}$ in the vector $\mathbf{t}^n$. Fix a $\delta > 0$ and for $t \in \mathcal{T}$, define the event

$$A_\delta^t = \left\{ \left| \frac{N(t|\mathbf{t}^n)}{n} - \alpha(t) \right| > \delta \right\}. \tag{81}$$

Since $N(t|\mathbf{t}^n) = \sum_{i=1}^n \mathbf{1}(t_i = t)$ and each term in the sum is an IID Bernoulli variable with probability of 1 equal to $\alpha(t)$, we have by Hoeffding's tail inequality [26],

$$P(A_\delta^t) \leq 2\exp(-2n\delta^2). \tag{82}$$

Next, we need to show that the substrings output by the AVS at the times when the state is t have a type in or very near $\overline{\mathcal{G}}(t)$. This will be done by a martingale argument similar to that given in Lemma 3 of [5]. Let $\mathbf{t}^\infty$ denote the infinite state sequence $(t_1, t_2, \ldots)$ and let $\mathcal{F}_0 = \sigma(\mathbf{t}^\infty)$ be the sigma field generated by the states $\mathbf{t}^\infty$. For $i = 1, 2, \ldots$, let $\mathcal{F}_i = \sigma(\mathbf{t}^\infty, \mathbf{s}^i, \mathbf{x}_1^i, \ldots, \mathbf{x}_m^i)$. Note that $\{\mathcal{F}_i\}_{i=0}^\infty$ is a filtration and for each i, the x_i is included in $\mathcal{F}_i$ trivially because $x_i = x_{s_i,i}$.

Let C_i be the $|\mathcal{X}|$-dimensional unit vector with a 1 in the position of x_i. That is, $C_i(x) = \mathbf{1}(x_i = x)$ for each $x \in \mathcal{X}$. Define T_i to be

$$T_i = C_i - \mathbb{E}[C_i|\mathcal{F}_{i-1}] \tag{83}$$

and let $S_0 = 0$. For $k \geq 1$,

$$S_k = \sum_{i=1}^k T_i. \tag{84}$$

We claim that $S_k, k \geq 1$ is a martingale[7] with respect to the filtration $\{\mathcal{F}_i\}$ defined previously. To see this, note that $\mathbb{E}[|S_k|] < \infty$ for all k since S_k is bounded (not uniformly). Also, $S_k \in \mathcal{F}_k$ because $T_i \in \mathcal{F}_i$ for each i. Finally,

$$\begin{aligned}
\mathbb{E}[S_{k+1}|\mathcal{F}_k] &= \mathbb{E}[T_{k+1} + S_k|\mathcal{F}_k] \\
&= \mathbb{E}[T_{k+1}|\mathcal{F}_k] + S_k \\
&= \mathbb{E}[C_{k+1} - \mathbb{E}[C_{k+1}|\mathcal{F}_k]|\mathcal{F}_k] + S_k \\
&= \mathbb{E}[C_{k+1}|\mathcal{F}_k] - \mathbb{E}[C_{k+1}|\mathcal{F}_k] + S_k \\
&= S_k.
\end{aligned}$$

Now, define for each $t \in \mathcal{T}$,

$$T_i^t = T_i \cdot \mathbf{1}(t_i = t) \tag{85}$$

and analogously,

$$S_k^t = \sum_{i=1}^k T_i^t. \tag{86}$$

It can be easily verified that S_k^t is a martingale with respect to $\mathcal{F}_i$ for each $t \in \mathcal{T}$. Expanding, we also see that

$$\begin{aligned}
\frac{1}{N(t|\mathbf{t}^n)} S_n^t &= \frac{1}{N(t|\mathbf{t}^n)} \sum_{i=1}^n T_i \mathbf{1}(t_i = t) & (87)\\
&= \frac{1}{N(t|\mathbf{t}^n)} \sum_{i:\ t_i = t} C_i - \frac{1}{N(t|\mathbf{t}^n)} \sum_{i:\ t_i = t} \mathbb{E}[C_i|\mathcal{F}_{i-1}]. & (88)
\end{aligned}$$

The first term in the difference above is the type of the output of the AVS during times when the state is t. For any i such that $t_i = t$,

$$\mathbb{E}[C_i|\mathcal{F}_{i-1}] = \sum_{l=1}^m P(l|\mathcal{F}_{i-1}) p_l(\cdot|t) \in \overline{\mathcal{G}}(t). \tag{89}$$

[7] S_k is a vector, so we show that each component of the vector is martingale. For ease of notation, we drop the dependence on the component of the vector until it is explicitly needed.

In the above, $P(l|\mathcal{F}_{i-1})$ represents the switcher's possibly random strategy because the switcher chooses the switch position at time i with knowledge of events in $\mathcal{F}_{i-1}$. The source generator's outputs, conditioned on the state at the time are independent of all other random variables, so $\sum_{l=1}^{m} P(l|\mathcal{F}_{i-1})p_l(\cdot|t)$ is the probability distribution of the output at time i conditioned on $\mathcal{F}_{i-1}$.

Thus, the second term in the difference of equation (88) is in $\overline{\mathcal{G}}(t)$ because it is the average of $N(t|\mathbf{t}^n)$ terms in $\overline{\mathcal{G}}(t)$ and $\overline{\mathcal{G}}(t)$ is a convex set. Therefore, $S_n^t/N(t|\mathbf{t}^n)$ measures the difference between the type of symbols output at times when the state is t and some distribution guaranteed to be in $\overline{\mathcal{G}}(t)$.

Let $p_{\mathbf{x}^n}$ be the empirical type of the string $\mathbf{x}^n$, and let $p_{\mathbf{x}^n}^t$ be the empirical type of the sub-string of $\mathbf{x}^n$ corresponding to the times i when $t_i = t$. Then,

$$p_{\mathbf{x}^n} \;=\; \sum_{t\in\mathcal{T}} \frac{N(t|\mathbf{t}^n)}{n} p_{\mathbf{x}^n}^t. \tag{90}$$

Let $\overline{\mathcal{G}}(t)_\epsilon$ be the set of distributions at most ϵ in $\mathcal{L}_1$ distance from a distribution in $\overline{\mathcal{G}}(t)$. Recall that for $|\mathcal{X}|$ dimensional vectors, $\|p-q\|_\infty < \epsilon/|\mathcal{X}|$ implies $\|p-q\|_1 < \epsilon$. Hence, we have

$$P\left(\bigcup_{t\in\mathcal{T}} \{p_{\mathbf{x}^n}^t \notin \overline{\mathcal{G}}(t)_\epsilon\}\right) \;\le\; \sum_{t\in\mathcal{T}} P\left(\bigcup_{x\in\mathcal{X}} \left\{\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right\}\right) \tag{91}$$

$$\le\; \sum_t \sum_x P\left(\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right). \tag{92}$$

Let $(A_\delta^t)^c$ denote the complement of the event A_δ^t. So, for every (t,x) we have

$$P\left(\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right) \;\le\; P(A_\delta^t) + P\left(\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right) \tag{93}$$

$$\le\; 2\exp(-2n\delta^2) + P\left(\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right). \tag{94}$$

In the event of $(A_\delta^t)^c$, we have $N(t|\mathbf{t}^n) \ge n(\alpha(t)-\delta)$, so

$$P\left(\left|\frac{1}{N(t|\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right) \;\le\; P\left(|S_n^t(x)| > n(\alpha(t)-\delta)\frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right) \tag{95}$$

$$\le\; P\left(|S_n^t(x)| > n(\alpha(t)-\delta)\frac{\epsilon}{|\mathcal{X}|}\right). \tag{96}$$

$S_k^t(x)$ is a martingale with bounded differences since $|S_{k+1}^t(x) - S_k^t(x)| = |T_{k+1}^t(x)| \le 1$. Hence, we can apply Azuma's inequality [27] to get

$$P\left(|S_n^t(x)| > n(\alpha(t)-\delta)\frac{\epsilon}{|\mathcal{X}|}\right) \;\le\; 2\exp\left(-n\frac{(\alpha(t)-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right). \tag{97}$$

Plugging this back into equation (92),

$$P\left(\bigcup_{t\in\mathcal{T}} \{p_{\mathbf{x}^n}^t \notin \overline{\mathcal{G}}(t)_\epsilon\}\right) \;\le\; 2|\mathcal{T}||\mathcal{X}|\left(\exp(-2n\delta^2) + \exp\left(-n\frac{(\alpha_*-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right)\right) \tag{98}$$

$$\le\; 4|\mathcal{X}||\mathcal{T}|\exp(-n\xi(\epsilon,\delta)) \tag{99}$$

where

$$\xi(\epsilon,\delta) \;=\; \min\left\{2\delta^2, \frac{(\alpha_*-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right\} \tag{100}$$

$$\alpha_* \;\triangleq\; \min_{t\in\mathcal{T}} \alpha(t). \tag{101}$$

We assume without loss of generality that $\alpha_* > 0$ since $\mathcal{T}$ is finite. We will soon need that $\delta \le \epsilon/|\mathcal{T}|$, so let

$$\widetilde{\xi}(\epsilon) = \max_{0<\delta<\min\{\epsilon/|\mathcal{T}|,\alpha_*\}} \xi(\epsilon,\delta) \tag{102}$$

and note that it is always positive provided $\epsilon > 0$, since $\xi(\epsilon, \delta) > 0$ whenever $\delta \in (0, \alpha_*)$. Hence,

$$P\left(\bigcup_{t\in\mathcal{T}} \{p^t_{\mathbf{x}^n} \notin \overline{\mathcal{G}}(t)_\epsilon\}\right) \leq 4|\mathcal{X}||\mathcal{T}|\exp(-n\widetilde{\xi}(\epsilon)) \xrightarrow{n} 0. \tag{103}$$

We have shown that with probability at least $1 - 4|\mathcal{X}||\mathcal{T}|\exp(-n\widetilde{\xi}(\epsilon))$, for each $t \in \mathcal{T}$ there is some $p^t \in \overline{\mathcal{G}}(t)$ such that $\|p^t_{\mathbf{x}^n} - p^t\|_1 \leq \epsilon$ and $(A^t_{\epsilon/|\mathcal{T}|})^c$ occurs. Let

$$p = \sum_{t\in\mathcal{T}} \alpha(t)p^t. \tag{104}$$

By construction, $p \in \mathcal{D}_{states}$. To finish, we show that $\|p_{\mathbf{x}^n} - p\|_1 \leq 2\epsilon$.

$$\begin{aligned}
\|p_{\mathbf{x}^n} - p\|_1 &= \sum_{x\in\mathcal{X}} |p_{\mathbf{x}^n}(x) - p(x)| && (105)\\
&= \sum_x \left|\sum_{t\in\mathcal{T}} \frac{N(t|\mathbf{t}^n)}{n} p^t_{\mathbf{x}^n}(x) - \alpha(t)p^t(x)\right| && (106)\\
&\leq \sum_t \sum_x \left|\frac{N(t|\mathbf{t}^n)}{n} p^t_{\mathbf{x}^n}(x) - \alpha(t)p^t(x)\right| && (107)\\
&= \sum_t \alpha(t) \sum_x \left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} p^t_{\mathbf{x}^n}(x) - p^t(x)\right| && (108)\\
&\leq \sum_t \alpha(t) \sum_x |p^t_{\mathbf{x}^n}(x) - p^t(x)| + \left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} - 1\right| p^t_{\mathbf{x}^n}(x). && (109)
\end{aligned}$$

From (81), we are assumed to be in the event that

$$\left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} - 1\right| \leq \frac{\delta}{\alpha(t)} \tag{110}$$

Hence,

$$\begin{aligned}
\|p_{\mathbf{x}^n} - p\|_1 &\leq \sum_t \alpha(t)\left(\epsilon + \frac{\delta}{\alpha(t)}\right) && (111)\\
&= \epsilon + |\mathcal{T}|\delta \leq 2\epsilon. && (112)
\end{aligned}$$

We have proved $P(p_{\mathbf{x}^n} \notin \mathcal{D}_{states,2\epsilon}) \leq 4|\mathcal{X}||\mathcal{T}|\exp(-n\widetilde{\xi}(\epsilon))$, so we arrive at the conclusion of the lemma by letting $\xi(\epsilon) = \widetilde{\xi}(\epsilon/2)$.

■

Appendix III
Proof of Lemma 7.2

Let $W^*_{p,D} = \text{arg min}_{W\in\mathcal{W}(p,D)} I(p, W)$. Then

$$|R(p, D) - R(q, D)| = |I(p, W^*_{p,D}) - I(q, W^*_{q,D})|. \tag{113}$$

Consider $d(p, W^*_{q,D})$, the distortion of source p across q's distortion D achieving channel.

$$\begin{aligned}
d(p, W^*_{q,D}) &\leq d(q, W^*_{q,D}) + |d(p, W^*_{q,D}) - d(q, W^*_{q,D})| && (114)\\
&= d(q, W^*_{q,D}) + \left|\sum_x \sum_{\hat{x}} (p(x) - q(x))W^*_{q,D}(\widehat{x}|x)d(x, \widehat{x})\right| && (115)\\
&\leq D + \sum_x |p(x) - q(x)| \sum_{\widehat{x}} W^*_{q,D}(\widehat{x}|x)d(x, \widehat{x}) && (116)\\
&\leq D + \|p - q\|_1 d^*. && (117)
\end{aligned}$$

By definition, $W^*_{q,D}$ is in $\mathcal{W}(p, d(p, W^*_{q,D}))$, so $R(p, d(p, W^*_{q,D})) \leq I(p, W^*_{q,D})$.

$$\begin{align} R(p, d(p, W^*_{q,D})) &\leq I(p, W^*_{q,D}) \tag{118}\\ &\leq I(q, W^*_{q,D}) + |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \tag{119}\\ &= R(q, D) + |I(p, W^*_{q,D}) - I(q, W^*_{q,D})|. \tag{120} \end{align}$$

Expanding mutual informations yields

$$\begin{align} |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| &= |H(p) + H(pW^*_{q,D}) - H(p, W^*_{q,D}) \cdots \tag{121}\\ &\qquad -H(q) - H(qW^*_{q,D}) + H(q, W^*_{q,D})| \nonumber\\ &\leq |H(p) - H(q)| + |H(pW^*_{q,D}) - H(qW^*_{q,D})| + \cdots \nonumber\\ &\qquad |H(p, W^*_{q,D}) - H(q, W^*_{q,D})|. \tag{122} \end{align}$$

Above, for a distribution p on $\mathcal{X}$ and channel W from $\mathcal{X}$ to $\widehat{\mathcal{X}}$, $H(pW)$ denotes the entropy of a distribution on $\widehat{\mathcal{X}}$ with probabilities $(pW)(\widehat{x}) = \sum_x p(x) W(\widehat{x}|x)$. $H(p, W)$ denotes the entropy of the joint source on $\mathcal{X} \times \widehat{\mathcal{X}}$ with probabilities $(p, W)(x, \widehat{x}) = p(x)W(\widehat{x}|x)$. It is straightforward to verify that $\|pW - qW\|_1 \leq \|p - q\|_1$ and $\|(p, W) - (q, W)\|_1 \leq \|p - q\|_1$. So using Lemma 7.1 three times, we have

$$\begin{align} |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| &\leq \|p - q\|_1 \ln \frac{|\mathcal{X}|}{\|p - q\|_1} + \|p - q\|_1 \ln \frac{|\widehat{\mathcal{X}}|}{\|p - q\|_1} + \nonumber\\ &\qquad \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1} \tag{123}\\ &\leq 3\|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \tag{124} \end{align}$$

Now, we have seen $d(p, W^*_{q,D}) \leq D + d^*\|p - q\|_1$. We will use the uniform continuity of $R(p, D)$ in D to bound $|R(p, D) - R(p, D + d^*\|p - q\|_1)|$. This will give an upper bound on $R(p, D) - R(q, D)$ as seen through equation (120), namely,

$$\begin{align} R(p, D) - R(q, D) &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + R(p, D) - R(p, d(p, W^*_{q,D})) \tag{125}\\ &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + R(p, D) - R(p, D + d^*\|p - q\|_1), \tag{126} \end{align}$$

where the last step follows because $R(p, D)$ is monotonically decreasing in D. For a fixed p, the rate-distortion function in D is convex $\cup$ and decreasing and so has steepest descent at $D = 0$. Therefore, for any $0 \leq D_1, D_2 \leq d^*$,

$$|R(p, D_1) - R(p, D_2)| \leq |R(p, 0) - R(p, |D_2 - D_1|)|. \tag{127}$$

Hence, we can restrict our attention to continuity of $R(p, D)$ around $D = 0$. By assumption, $\mathcal{W}(p, 0) \neq \emptyset \; \forall p \in \mathcal{P}(\mathcal{X})$. Now consider an arbitrary $D > 0$, and let $W \in \mathcal{W}(p, D)$. We will show that there is some $W_0 \in \mathcal{W}(p, 0)$ that is close to W in an $\mathcal{L}_1$-like sense (relative to the distribution p). Since $W \in \mathcal{W}(p, D)$, we have by definition

$$\begin{align} D &\geq \sum_x p(x) \sum_{\widehat{x}} W(\widehat{x}|x) d(x, \widehat{x}) \tag{128}\\ &= \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x) d(x, \widehat{x}) \tag{129}\\ &\geq \tilde{d} \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x). \tag{130} \end{align}$$

Now, we will construct a channel in $\mathcal{W}(p, 0)$, denoted W_0. First, for each $x, \widehat{x}$ such that $d(x, \widehat{x}) = 0$, let $V(\widehat{x}|x) = W(\widehat{x}|x)$. For all other $(x, \widehat{x})$, set $V(\widehat{x}|x) = 0$. Note that V is not a channel matrix if $W \notin \mathcal{W}(p, 0)$ since it is missing some probability mass. To create W_0, for each x, we redistribute the missing mass from $V(\cdot|x)$ to the pairs $(x, \widehat{x})$ with $d(x, \widehat{x}) = 0$. Namely, for $(x, \widehat{x})$ with $d(x, \widehat{x}) = 0$, we define

$$W_0(\widehat{x}|x) = V(\widehat{x}|x) + \frac{\sum_{\widehat{x}':\ d(x,\widehat{x}')>0} W(\widehat{x}'|x)}{|\{\hat{x}' :\ d(x, \hat{x}') = 0\}|}. \tag{131}$$

For all $(x,\widehat{x})$ with $d(x,\widehat{x})>0$, define $W_0(\widehat{x}|x)=0$. So, W_0 is a valid channel in $\mathcal{W}(p,0)$. Now for a fixed $x\in\mathcal{X}$,

$$\begin{aligned}
\sum_{\widehat{x}}|W(\widehat{x}|x)-W_0(\widehat{x}|x)| &= \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x)+\sum_{\widehat{x}:\ d(x,\widehat{x})=0}|W(\widehat{x}|x)-W_0(\widehat{x}|x)| && (132)\\
&= \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x)+\cdots && (133)\\
&\qquad \sum_{\widehat{x}:\ d(x,\widehat{x})=0}\left|W(\widehat{x}|x)-W(\widehat{x}|x)-\frac{\sum_{\hat{x}':\ d(x,\hat{x}')>0}W(\hat{x}'|x)}{|\{\hat{x}'\ :\ d(x,\hat{x}')=0\}|}\right| \\
&= 2\sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x). && (134)
\end{aligned}$$

Therefore, using (130)

$$\sum_x p(x)\sum_{\widehat{x}}|W(\widehat{x}|x)-W_0(\widehat{x}|x)| \;\le\; \frac{2D}{\widetilde{d}}. \tag{135}$$

So, for $W=W^*_{p,D}$, there is a $W_0\in\mathcal{W}(p,0)$ with the above 'modified $\mathcal{L}_1$ distance' with respect to p between W and W_0 being less than $2D/\widetilde{d}$. Going back to the bound on $|R(p,0)-R(p,D)|$,

$$\begin{aligned}
|R(p,0)-R(p,D)| &= \min_{W\in\mathcal{W}(p,0)} I(p,W)-I(p,W^*_{p,D}) && (136)\\
&\le I(p,W_0)-I(p,W^*_{p,D}) && (137)\\
&\le |H(pW_0)-H(pW^*_{p,D})|+|H(p,W_0)-H(p,W^*_{p,D})|. && (138)
\end{aligned}$$

Now, note that the $\mathcal{L}_1$ distance between pW_0 and $pW^*_{p,D}$ is

$$\begin{aligned}
\|pW_0-pW^*_{p,D}\|_1 &= \sum_{\widehat{x}}\left|\sum_x p(x)W_0(\widehat{x}|x)-p(x)W^*_{p,D}(\widehat{x}|x)\right| && (139)\\
&\le \sum_x p(x)\sum_{\widehat{x}}|W_0(\widehat{x}|x)-W^*_{p,D}(\widehat{x}|x)| && (140)\\
&\le \frac{2D}{\widetilde{d}}. && (141)
\end{aligned}$$

Similarly, $\|(p,W_0)-(p,W^*_{p,D})\|_1\le 2D/\widetilde{d}$.

Now, assuming $D\le\widetilde{d}/4$, we can again invoke Lemma 7.1 to get

$$\begin{aligned}
|R(p,0)-R(p,D)| &\le \frac{2D}{\widetilde{d}}\ln\frac{\widetilde{d}|\mathcal{X}|}{2D}+\frac{2D}{\widetilde{d}}\ln\frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D} && (142)\\
&\le \frac{4D}{\widetilde{d}}\ln\frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D}. && (143)
\end{aligned}$$

Going back to (126), we see that if $\|p-q\|_1\le\frac{\widetilde{d}}{4d^*}$,

$$\begin{aligned}
|R(p,d+d^*\|p-q\|_1))-R(p,D)| &\le \frac{4d^*\|p-q\|_1}{\widetilde{d}}\ln\frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} && (144)\\
&\le \frac{4d^*\|p-q\|_1}{\widetilde{d}}\ln\frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. && (145)
\end{aligned}$$

The last step follows because $\widetilde{d}/d^*\le 1$. Substituting into equation (126) gives

$$\begin{aligned}
R(p,D)-R(q,D) &\le 3\|p-q\|_1\ln\frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}+4\frac{d^*}{\widetilde{d}}\|p-q\|_1\ln\frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} && (146)\\
&\le \frac{7d^*}{\widetilde{d}}\|p-q\|_1\ln\frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. && (147)
\end{aligned}$$

Finally, this bound holds uniformly on p and q as long as the condition on $\|p-q\|_1$ is satisfied. Therefore, we can interchange p and q to get the other side of the inequality.

$$R(q,D) - R(p,D) \quad \leq \quad \frac{7d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \tag{148}$$

This concludes the proof.

Appendix IV
Proof of Lemma 7.3

We now assume $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ to be arbitrary. However, we let

$$d_0(x,\widehat{x}) = d(x,\widehat{x}) - \min_{\widetilde{x}\in\widehat{\mathcal{X}}} d(x,\widetilde{x}) \tag{149}$$

so that Lemma 7.2 applies to d_0. Let $R_0(p,D)$ be the IID rate-distortion function for $p \in \mathcal{P}(\mathcal{X})$ at distortion D with respect to distortion measure $d_0(x,\widehat{x})$. By definition, $R(p,D)$ is the IID rate-distortion function for p with respect to distortion measure $d(x,\widehat{x})$. From Problem 13.4 of [20], for any $D \geq D_{\min}(p)$,

$$R(p,D) = R_0(p, D - D_{\min}(p)). \tag{150}$$

Hence, for $p, q \in \mathcal{P}(\mathcal{X})$, $D \geq \max(D_{\min}(p), D_{\min}(q))$,

$$\begin{aligned} |R(p,D) - R(q,D)| &= |R_0(p, D - D_{\min}(p)) - R_0(q, D - D_{\min}(q)| && (151)\\ &\leq |R_0(p, D - D_{\min}(p)) - R_0(p, D - D_{\min}(q))| + \\ &\qquad |R_0(p, D - D_{\min}(q)) - R_0(q, D - D_{\min}(q))|. && (152) \end{aligned}$$

Now, we note that $|D_{\min}(p) - D_{\min}(q)| \leq d^*\|p-q\|_1$. The first term of equation (152) can be bounded using equation (143) and the second term of (152) can be bounded using Lemma 7.2. The first term can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$ and the second can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d_0^*$. Since $d_0^* \leq d^*$, we only require $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$.

$$\begin{aligned} |R(p,D) - R(q,D)| &\leq \frac{4d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{\widetilde{d}_0|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} + \frac{7d_0^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} && (153)\\ &\leq \frac{4d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} + \frac{7d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. && (154) \end{aligned}$$

References

[1] H. Palaiyanur, C. Chang, and A. Sahai, "The source coding game with a cheating switcher," in *Proc. Int. Symp. Inform. Theory*, Nice, France, June 2007.

[2] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, Aug. 1988.

[3] A. Chebira, P. Dragotti, L. Sbaiz, and M. Vetterli, "Sampling and interpolation of the plenoptic function," in *Proc.of IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[4] P. Longman, "The best care anywhere," *Washington Monthly*, Jan. 2005.

[5] T. Berger, "The source coding game," *IEEE Transactions on Information Theory*, vol. 17, pp. 71–76, Jan. 1971.

[6] C. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Devel.*, vol. 2, pp. 289–293, Oct. 1958.

[7] S. Gelfand and M. Pinsker, "Coding for channel with random parameters," *Probl. Pered. Inform. (Probl. Inf. Transm.)*, vol. 9, pp. 19–31, 1980.

[8] M. H. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983.

[9] F. Willems, "On Gaussian channels with side information at the transmitter," in *Proc. Int. Symp. Inform. Theory*, Benelux, Enschede, The Netherlands, May 1988, pp. 129–135.

[10] ——, "Signalling for the Gaussian channel with side information at the transmitter," in *Proc. Int. Symp. Inform. Theory*, Sorrento, Italy, June 2000.

[11] U. Erez, S. Shamai (Shitz), and R. Zamir, "Capacity and lattice strategies for canceling known interference," *IEEE Transactions on Information Theory*, vol. 51, Nov. 2005.

[12] M. Agarwal, A. Sahai, and S. Mitter, "Coding into a source: a direct inverse rate-distortion theorem," in *Forty-fourth Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sept. 2006. [Online]. Available: http://arxiv.org/abs/cs.IT/0610142

[13] D. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Transactions on Information Theory*, vol. 28, pp. 701–713, Sept. 1982.

[14] T. Weissman and N. Merhav, "On causal source codes with side information," *IEEE Transactions on Information Theory*, vol. 51, pp. 4003–4013, Nov. 2005.

[15] S. Tatikonda, A. Sahai, and S. Mitter, "Stochastic linear control over a communication channel," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1549–1561, Sept. 2004.

[16] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Natl. Conv. Rec.*, 1959, pp. 142–163.

[17] J. Wolfowitz, "Approximation with a fidelity criterion," in *5th Berkeley Symp. on Math. Stat. and Prob.*, vol. 1. Berkeley, California: University of California, Press, 1967, pp. 565–573.

[18] D. Sakrison, "The rate-distortion function for a class of sources," *Information and Control*, vol. 15, pp. 165–195, Mar. 1969.

[19] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. New York, NY: Academic Press, 1997.

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

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

[22] R. Ahlswede, "Extremal properties of rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 36, pp. 166–171, Jan. 1990.

[23] M. Harrison and I. Kontoyiannis, "Estimation of the rate-distortion function," 2007. [Online]. Available: http://arxiv.org/abs/cs/0702018v1

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

[25] R. Dobrushin, "Unified methods for the transmission of information: The general case," *Sov. Math.*, vol. 4, pp. 284–292, 1963.

[26] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, Mar 1963.

[27] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. Journal*, vol. 19, pp. 357 – 367, 1967.