The Source Coding Game With a Cheating Switcher

Hari Palaiyanur, Cheng Chang, *Member, IEEE*, and Anant Sahai, *Member, IEEE*

***Abstract*—The problem of finding the rate-distortion function of an arbitrarily varying source (AVS) composed of a finite number of memoryless subsources is revisited. Berger's 1971 paper "The Source Coding Game" solves this problem when the adversary is allowed only strictly causal access to the subsource realizations. The case when the adversary has access to the subsource realizations non-causally is considered. This new rate-distortion function is determined to be the maximum of the rate-distortion function over a set of independent and identically distributed (IID) random variables that can be simulated by the adversary. The results are extended to allow for partial or noisy observations of subsource realizations. The model is further explored 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 desired accuracy. The bound is also used to give a rate of convergence for the estimate of the rate-distortion function for an unknown IID source.**

***Index Terms*—Adversarial source coding, arbitrarily varying source, finite alphabet, rate-distortion, source coding game, type covering, uniform continuity of rate-distortion functions.**

I. Introduction

A. Motivation

THE arbitrarily varying source (AVS) was introduced by Berger [4] as a source that samples other "subsources" under the control of an agent called a switcher. The AVS was used in the model of an information-theoretic 'source coding game' between two players, the afore-mentioned switcher and a coder. The goal of the coder was to encode the output of the AVS to within a specified distortion, and the goal of the switcher was to make the coder use as large a rate to attain the specified distortion as possible. Berger studied the adversarial rate-distortion function (the rate the coder needs to achieve a target distortion regardless of switcher strategy) under certain rules for the switcher. Primarily, [4] gives the rate-distortion function when the switcher is not allowed to observe present or future subsource realizations.

The purpose of this paper is to deepen understanding of the source coding game by looking into variations where the capabilities of the switcher are enhanced. In [4], Berger himself asks what happens to the rate-distortion function when the switcher is allowed to "cheat" and observe present or future realizations of the subsources. In addition to tackling this question, we further study scenarios where the switcher receives noisy observations of the subsources or the switcher is not adversarial, but helpful.

As a motivation for studying the source coding game, Berger mentions that the results might have application to situations where multiple data streams are multiplexed into a single data stream. Another potential application is in the field of active sensing or active vision [5], a subfield of computer vision in which sensors actively explore their environment using information they have previously sensed. The idea of using an AVS with specialized models for the switcher as a model for an active source is explored in [3].

B. Results and Organization of Paper

Section II sets up the notation and model, and briefly reviews the literature on lossy compression of arbitrarily varying sources. Intuitively, a strictly causal adversary switching amongst memoryless subsources is no more threatening than a switcher that randomly switches. This intuition was proved correct in [4] by Berger as he determined the rate-distortion function for memoryless subsources and a strictly causal adversarial model. 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 rate-distortion function over the IID sources 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 realizations.

In order to provide more ways to restrict the knowledge of the switcher, 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.

Manuscript received December 18, 2007; revised November 03, 2010; accepted November 17, 2010. Date of current version June 22, 2011. This work was supported by a National Science Foundation (NSF) Graduate Research Fellowship and by NSF Grants ANI-326503, CNS-403427, and CCF-729122. An earlier version of the material in this paper was presented in part at the 2007 International Symposium on Information Theory, Nice, France. Some portions have also been presented at the 2008 International Symposium on Information Theory, Toronto, ON, Canada.

H. Palaiyanur and A. Sahai are with Wireless Foundations, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA (e-mails: hpalaiya@eecs.berkeley.edu; sahai@eecs.berkeley.edu).

C. Chang was with Wireless Foundations, Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley, Berkeley, CA 94720 USA. He is now with D.E. Shaw Group, New York, NY 10036 USA (e-mail: cchang@eecs.berkeley.edu).

Communicated by E.-H. Yang, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2145730

0018-9448/$26.00 © 2011 IEEE

Next, Section V changes the perspective from the traditional adversarial setting to a cooperative setting. It explores the problem when the goal of the switcher is to help the coder achieve a low distortion. Theorem 5.1 gives a characterization of the rate-distortion function if the helper has access to future realizations 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. However, for most helpful switcher settings, a tight characterization of the rate-distortion function is lacking.

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. This task needs some discussion because of the fact that the IID rate-distortion function is generally nonconcave as a function of the distribution [6]. The main tool there is an explicit bound on the uniform continuity of the IID rate-distortion function that is of potentially independent interest, as we use it to quickly analyze the behavior (in probability) of a simple rate-distortion estimator for IID sources. Finally, we conclude in Section VIII.

II. Problem Setup

A. Notation

Let $\mathcal{X}$ and $\hat{\mathcal{X}}$ be the finite source and reconstruction alphabets respectively. Let $\mathbf{x}^n = (x_1, \ldots, x_n)$ denote a vector from $\mathcal{X}^n$ and $\hat{\mathbf{x}}^n = (\hat{x}_1, \ldots, \hat{x}_n)$ a vector from $\hat{\mathcal{X}}^n$. 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 $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, d^*]$ be a distortion measure on the product set $\mathcal{X} \times \hat{\mathcal{X}}$ with maximum distortion $d^* < \infty$. Let

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

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

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

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 (see [7], [8]) of length-n strings from $\mathcal{X}$, and let $\mathcal{W}$ be the set of probability transition matrices from $\mathcal{X}$ to $\hat{\mathcal{X}}$. 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

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

be the minimum average distortion achievable for the source distribution p. The (functional) IID 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

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

where $\mathcal{W}(p, D)$ is a set of admissable probability transition matrices

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

and $I(p, W)$ is the mutual information[1]

$$I(p, W) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p(x) W(\hat{x} \,|\, x) \ln \left[\frac{W(\hat{x} \,|\, x)}{(pW)(\hat{x})} \right]$$

with $(pW)(\hat{x}) = \sum_{x' \in \mathcal{X}} p(x') W(\hat{x}|x')$. Let $\mathcal{B} = \{\hat{\mathbf{x}}^n(1), \ldots, \hat{\mathbf{x}}^n(K)\}$ be a codebook with K length-n vectors from $\hat{\mathcal{X}}^n$. Define

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

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

$$d(p; \mathcal{B}) = \sum_{\mathbf{x}^n \in \mathcal{X}^n} d_n(\mathbf{x}^n; \mathcal{B}) \prod_{k=1}^{n} p(x_k) = \mathbb{E}_p[d_n(\mathbf{x}^n; \mathcal{B})].$$

For $n \geq 1$, $D > D_{\min}(p)$, let $K(n, D)$ be the minimum number of codewords needed in a codebook $\mathcal{B} \subset \hat{\mathcal{X}}^n$ so that $d(p; \mathcal{B}) \leq D$. By convention, if no such codebook exists, $K(n, D) = \infty$. Let the (operational) rate-distortion function[2] of an IID source be $R(D) = \limsup_n \frac{1}{n} \ln K(n, D)$. Shannon's rate-distortion theorem ([9], [10]) states that for all n, $\frac{1}{n} \ln K(n, D) \geq R(p, D)$ and

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

B. Arbitrarily Varying Sources

As mentioned earlier, the AVS is a model of a source in the 'source coding game' introduced by Berger in [4]. 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 selecting the sequence of symbols to be encoded.

Fig. 1 shows a model of an AVS. There are m IID "subsources" with common alphabet $\mathcal{X}$. In [4], the subsources are assumed to be independent, but that restriction turns out not to be required[3]. There can also 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 of m different distributions. The marginal distributions of the m subsources are known to be $\{p_l\}_{l=1}^m$ and we let $\mathcal{G} = \{p_1, \ldots, p_m\}$. Let $P(x_{1,1}, \ldots, x_{m,1})$ be the joint probability distribution for

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

[2] 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 $D_{\min}(p)$.

[3] In [4], the motivation was multiplexing data streams and independence is a reasonable assumption, but the proofs did not require it.

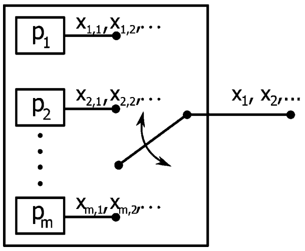


Fig. 1. Model for an AVS. The switcher can set the switch position according to the rules of the model.

the IID source $\{(x_{1,k},\dots,x_{m,k})\}_{k\geq 1}$. Fix an $n \geq 1$ and consider a block of length n. We let $x_{l,k}$ denote the output of the lth 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 rules for the AVS. In the next few sections, we will discuss different rules for the switcher, particularly different levels of causality in knowledge of the subsource realizations. The switcher may or may not have knowledge of the codebook, but this knowledge turns out to be inconsequential for the worst-case 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

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

Here, $\mathbb{E}[d_n(\mathbf{x}^n;\mathcal{B})]$ is defined to be $\sum_{\mathbf{x}^n}(\sum_{\mathbf{s}^n} P(\mathbf{s}^n,\mathbf{x}^n))d_n(\mathbf{x}^n;\mathcal{B})$, 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.

We are interested in the exponential rate of growth of $M(n,D)$ with n, and so we define the rate-distortion function of an adversarial AVS to be

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

In every case considered, it will also be clear that $R(D) = \liminf_{n\to\infty} \frac{1}{n} \ln M(n,D)$. For notational convenience, we only refer to the rate-distortion function as $R(D)$, removing its dependence on the subsource distributions as well as all the different cases of switcher power.

C. Literature Review

a) One IID Source: Suppose $m = 1$. Then there is only one IID subsource $p_1 = p$ and the switch position must be $s_k = 1$ for all time. This is exactly the classical rate-distortion problem considered by Shannon [9], and he showed

$$R(D) = R(p,D).$$

Computing $R(p,D)$ can be done with the Blahut-Arimoto algorithm [8], and also falls under the umbrella of convex programming.

b) Compound Source: Now suppose that $m > 1$, but the switcher is constrained to choose $s_k = s \in \{1,\dots,m\}$ for all k. That is, the switch position is set once and remains constant afterwards. Sakrison [11] studied the rate-distortion function for this class of *compound* sources and showed that planning for the worst subsource is both necessary and sufficient. Hence, for compound sources

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

Recall that $\mathcal{G} = \{p_1,\dots,p_m\}$ is the set of marginal distributions of the m subsources. 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) Strictly Causal Adversarial Source: In Berger's setup [4], the switcher is allowed to choose $s_k \in \{1,\dots,m\}$ 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 $(s_1,\dots,s_{k-1})$ and $(x_1,\dots,x_{k-1})$. The conclusion of [4] is that under these rules

$$R(D) = \max_{p\in\overline{\mathcal{G}}} R(p,D) \tag{2}$$

where $\bar{\mathcal{G}}$ is the convex hull of $\mathcal{G}$. It should be noted that this same rate-distortion function applies in the following cases [4].

- 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 (2), 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 [4], Berger poses the question of what happens to the rate-distortion function when the rules are tilted in favor of the switcher. Paraphrasing Berger:

> As another example, suppose the switcher is permitted to observe the candidates for x_k generated by each [subsource] before (randomly) selecting one of them. Then it can be shown that [$R(D)$ (except in certain special cases) strictly increases]. The determination of $R(D)$ under these rules appears to be a challenging task.

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: Define the set of distributions

$$\mathcal{C} = \left\{ p \quad : \quad \begin{array}{c} \sum_{x \in \mathcal{V}} p(x) \geq P(\forall l, x_l \in \mathcal{V}) \\ \forall\, \mathcal{V} \text{ such that} \\ \mathcal{V} \subseteq \mathcal{X} \end{array} \right\} \tag{3}$$

where the event $\{\forall\, l, x_l \in \mathcal{V}\}$ is shorthand for $\{(x_1, \ldots, x_m) : x_l \in \mathcal{V}, l = 1, \ldots, m\}$. Also, define

$$\tilde{R}(D) \triangleq \max_{p \in \mathcal{C}} R(p, D).$$

For a general set of distributions $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$, let $D_{\min}(\mathcal{Q}) \triangleq \sup_{p \in \mathcal{Q}} D_{\min}(p)$. Suppose the switcher has either 1-step lookahead or full lookahead. In both cases, for $D > D_{\min}(\mathcal{C})$

$$R(D) = \tilde{R}(D)$$

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:

- In non-degenerate cases, $\bar{\mathcal{G}}$ 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.
- 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 $\tilde{R}(D)$ requires the 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 A for the complete proof. To show $R(D) \leq \tilde{R}(D)$, we use the type-covering lemma from [4]. 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 $\exp(n(\tilde{R}(D) + \frac{|\mathcal{X}|}{n} \ln(n + 1) + \epsilon))$ 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 $\tilde{R}(D) + \epsilon$ with $\epsilon > 0$ arbitrarily small.

Now, to show $R(D) \geq \tilde{R}(D)$, 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 $\tilde{R}(D)$. For each subset $\mathcal{V} \subseteq \mathcal{X}$ 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

$$\mathcal{D} \triangleq \left\{ p \in \mathcal{P} \quad : \quad \begin{array}{c} p(\,\cdot\,) = \sum_{\mathcal{V}} \beta(\mathcal{V}) f(\cdot|\mathcal{V}), \\ f(\cdot|\mathcal{V}) \text{ is a PMF on } \mathcal{V}, \\ \forall\, \mathcal{V} \text{ s.t. } \mathcal{V} \subseteq \mathcal{X}, |\mathcal{V}| \leq m \end{array} \right\}. \tag{4}$$

$\mathcal{D}$ 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 A.3 in Appendix A 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 $R(D) \geq \tilde{R}(D)$. ■

Qualitatively, allowing the switcher to "cheat" gives access to distributions $p \in \mathcal{C}$ which may not be in $\bar{\mathcal{G}}$. 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 $p \in \mathcal{C}$

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

$\prod_{l=1}^{m} p_l(1)$ is the probability that all m subsources simultaneously 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 of Fig. 1 is to consider the case when the adversary has noisy access to subsource realizations through a discrete memoryless channel. Suppose we let the switcher observe y_k at time k, which is probabilistically related to the subsource realizations through a discrete memoryless multiple access channel W by

$$W(y_k \,|\, x_{1,k}, x_{2,k}, \ldots, x_{m,k}).$$

Since the subsource probability distributions are already known, through an application of Bayes' rule, this model is equivalent to one in which the switcher observes a state, $t_k = y_k$, noiselessly. Namely

$$Pr(x_{1,k} = x_1, \ldots, x_{m,k} = x_m | t_k = t) = \frac{P(x_1, \ldots, x_m) W(t|x_1, \ldots, x_m)}{\sum_{x'_1, \ldots, x'_m} W(t|x'_1, \ldots, x'_m) P(x'_1, \ldots, x'_m)}.$$

Conditioned on the state, the m subsources emit symbols independent of the past according to a conditional distribution. This model is depicted in Fig. 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

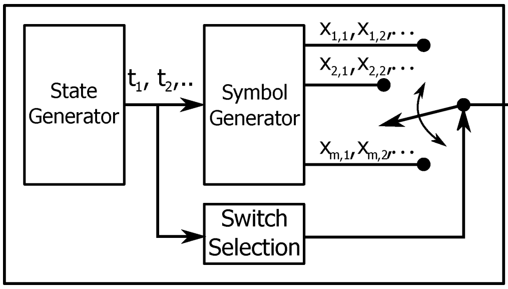


Fig. 2. Model of an AVS encompassing both cheating and non-cheating switchers. Additionally, this model allows for noisy observations of subsource realizations by the switcher.

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 a fixed $t \in \mathcal{T}$, let $\bar{\mathcal{G}}(t) = \mathbf{conv}(p_1(\cdot|t), \ldots, p_m(\cdot|t))$, where **conv** denotes the convex hull of a set of distributions.

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 strictly causal and 1-step lookahead switchers with noiseless subsource observations can be recovered as special cases of this model. If the conditional distributions $p_l(x|t)$ do not depend on t, the strictly causal switcher is recovered. The full lookahead switcher with noiseless subsource observations 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 Fig. 2, where the adversary has access to the states either with 1-step lookahead or full lookahead,

$$R(D) = \max_{p \in \mathcal{D}_{\text{states}}} R(p, D) \tag{5}$$

where

$$\mathcal{D}_{\text{states}} = \left\{ p : \begin{array}{c} p(\cdot) = \sum_{t \in \mathcal{T}} \alpha(t) f(\cdot|t) \\ f(\cdot|t) \in \bar{\mathcal{G}}(t), \forall\, t \in \mathcal{T} \end{array} \right\}. \tag{6}$$

Proof: See Appendix B. ■

One can see that in the case of the cheating switcher of the previous section, the set $\mathcal{D}$ of (4) equates directly with $\mathcal{D}_{\text{states}}$ of (6). 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, actually computing $R(D)$ in (5) falls into the discussion of Section VII.

V. The Helpful Switcher

Arbitrarily varying sources and channels have generally been associated with adversarial source and channel coding, but in this section, we consider the *helpful* cheating switcher to more thoroughly explore the information-theoretic game established in [4]. The goal of the helpful switcher is to help the coding system achieve low distortion. 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 $\rho : \mathcal{X}^* \times \hat{\mathcal{X}} \to [0, d^*]$ be defined by

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

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 this new 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

$$R(D) = R^*(\beta, D). \tag{7}$$

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 $\hat{\mathbf{x}}^n$, 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

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

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

$$\begin{aligned} d_n(\mathbf{x}^n; \mathcal{B}) &= \min_{\hat{\mathbf{x}}^n \in \mathcal{B}} \frac{1}{n} \sum_{k=1}^{n} \min_{x \in \mathcal{V}_k} d(x, \hat{x}_k) \\ &= \min_{\hat{\mathbf{x}}^n \in \mathcal{B}} \frac{1}{n} \sum_{k=1}^{n} \rho(\mathcal{V}_k, \hat{x}_k). \end{aligned}$$

Therefore, for the helpful switcher with full lookahead, 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 (7) can be done by the Blahut-Arimoto algorithm [7].
- 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,

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

Proof: If the switcher has at least 1-step lookahead, it immediately follows from the proof of Theorem 3.1 that $R(D) \le \min_{p\in\mathcal{C}} R(p, D)$. 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 $R(D) \le \min_{p\in\mathcal{D}_{\text{states}}} R(p, D)$.

VI. Examples

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

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

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

A. Bernoulli 1/4 and 1/3 Sources

Consider the example shown in Fig. 3, where the switcher has access to two independent 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 $p_2 = (2/3, 1/3)$. At time k, the switcher is given access to an observation $t_k = T(x_{1,k}, x_{2,k}, z_k)$ where T is a function and z_k is independent noise (that is, the switcher observes a potentially noisy version of the subsource realizations).

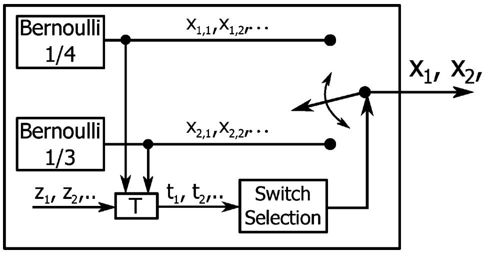


Fig. 3. Two independent Bernoulli subsources, which produce 1's with probabilities 1/4 and 1/3.

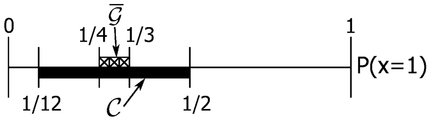


Fig. 4. Binary distributions the switcher can mimic. $\bar{\mathcal{G}}$ is the set of distributions the switcher can mimic with causal access to subsource realizations, and $\mathcal{C}$ is the set attainable with noncausal access.

First, we consider the switcher as an adversary in the traditional strictly causal setting of [4] and the 1-step lookahead setting, where the switcher has the subsource realizations $t_k = (x_{1,k}, x_{2,k})$ before choosing the switch position. For any time k,

$$\begin{aligned} P(x_{1,k} = x_{2,k} = 0) &= \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2} \\ P(x_{1,k} = x_{2,k} = 1) &= \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12} \\ P(\{x_{1,k}, x_{2,k}\} = \{0, 1\}) &= 1 - \frac{1}{2} - \frac{1}{12} = \frac{5}{12}. \end{aligned}$$

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

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

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 Fig. 4. The switcher with lookahead can simulate a significantly larger set of distributions than the strictly causal switcher, which is restricted to outputting 1's with a 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, the switcher observes the mod-2 sum of the two subsources $t_k = x_{1,k} \oplus x_{2,k}$. 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 $t_k = 0$, either both subsources are 0 or both subsources are 1, so the switch position doesn't matter in this state. If $t_k = 1$, 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

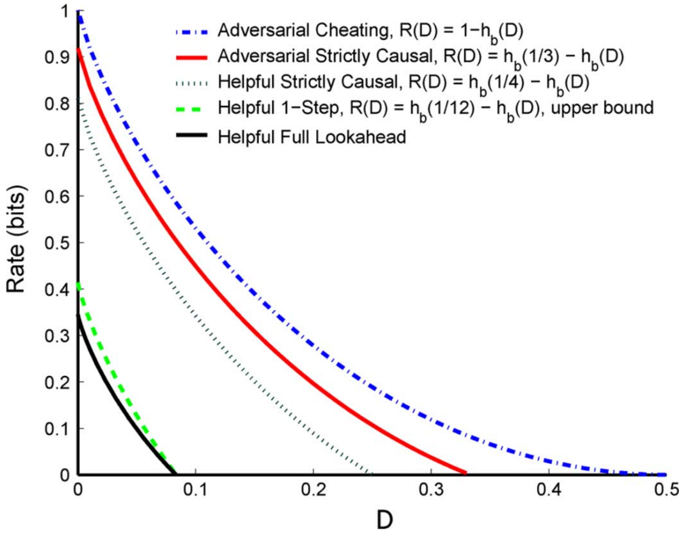


Fig. 5. $R(D)$ for the cheating switcher and the non-cheating switcher with Bernoulli 1/4 and 1/3 subsources. Also, the rate-distortion function for the examples of Fig. 3 where $t_k = x_{1,k} \oplus x_{2,k}$ and $t_k = x_{2,k}$.

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, 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 Fig. 5, the rate-distortion function is plotted for the situations discussed so far.

Finally, consider an example where an adversarial switcher observes only the second subsource through a binary symmetric channel with crossover probability $\delta \in [0, 1/2]$, i.e., $t_k = x_{2,k} \oplus z_k$ where z_k is a Bernoulli sequence that produces 1's with probability δ. Applying Theorem 4.1 again, it can be shown that if $\delta \in [0, 2/5]$

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

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

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

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 the second subsource output was a 1 given the state. If the switcher observes $t_k = 1$ 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 $t_k = 0$, 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. Fig. 6 shows $R(D)$ for this example as a function of δ for two values of D.

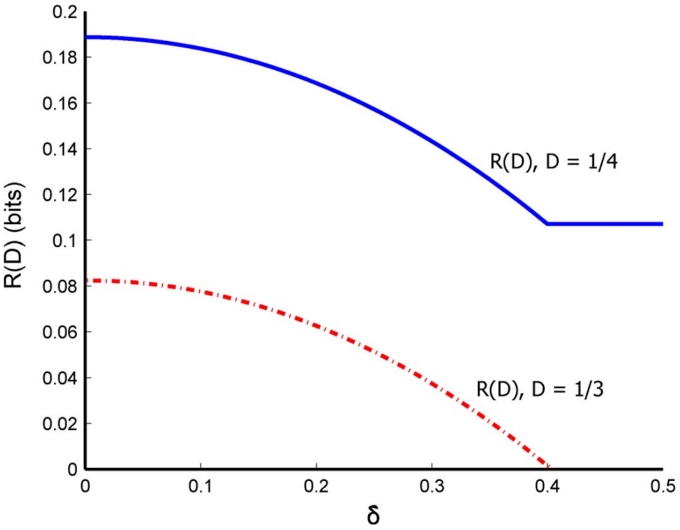


Fig. 6. $R(D)$ as a function of the noisy observation crossover probability δ for $D = 1/3$ and $D = 1/4$ for the example of Fig. 3 with $t_k = x_{2,k} \oplus z_k$ and $z_k \sim \mathcal{B}(\delta)$.

B. Two Bernoulli 1/2 Subsources

Suppose $m = 2$, and the subsources are independent 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 [12] and Theorem 5.1, one can show that when the helpful switcher has full lookahead with $t_k = (x_{1,k}, x_{2,k})$

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

The plot of these functions in Fig. 7 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 $\mathcal{X}^* = \{\{0\}, \{1\}, \{0, 1\}\}$, $\rho(\{0\}, \hat{x}) = 1(0 \neq \hat{x})$, $\rho(\{1\}, \hat{x}) = 1(1 \neq \hat{x})$, $\rho(\{0, 1\}, \hat{x}) = 0$ and $\beta = (1/4, 1/4, 1/2)$. The $R^*(\beta, D)$ achieving distribution on $\hat{\mathcal{X}}$ 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 actually cover all such strings. This problem is circumvented through the aid

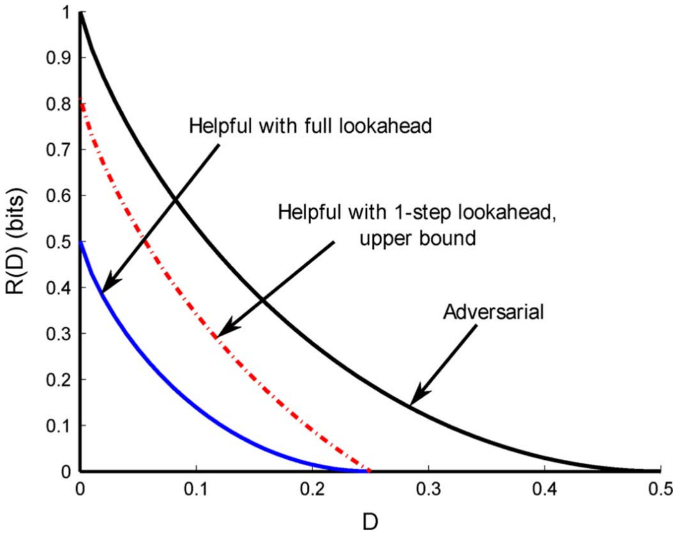


Fig. 7. $R(D)$ function for an AVS with two Bernoulli 1/2 sources when the switcher is helpful with full lookahead. For 1-step lookahead, the upper bound is shown.

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 adversarial AVS with either causal or noncausal access to subsource realizations is of the form

$$R(D) = \max_{p \in \mathcal{Q}} R(p, D) \tag{8}$$

where $\mathcal{Q}$ is a set of distributions in $\mathcal{P}(\mathcal{X})$. In (2), (3), and (6) $\mathcal{Q}$ 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 [6].

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, 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 [7], a bound on the continuity of the entropy of a distribution is developed in terms of $\|p - q\|_1$.

Lemma 7.1 (Uniform Continuity of Entropy, [7]): Let p and q be two probability distributions on $\mathcal{X}$ such that $\|p - q\|_1 \leq 1/2$, then

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

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

$$|R(p, D) - R(q, D)| \leq \frac{7d^*}{\tilde{d}} \|p - q\|_1 \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1}. \tag{9}$$

Proof: See Appendix C. ■

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 $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, d^*]$, define another distortion measure $d_0 : \mathcal{X} \times \hat{\mathcal{X}} \to [0, d^*]$ by

$$d_0(x, \hat{x}) = d(x, \hat{x}) - \min_{\tilde{x} \in \hat{\mathcal{X}}} d(x, \tilde{x}).$$

Now let $d_0^* = \max_{x, \hat{x}} d_0(x, \hat{x})$ and $\tilde{d}_0 = \min_{(x, \hat{x}): d_0(x, \hat{x}) > 0} d_0(x, \hat{x})$. We have defined $d_0(x, \hat{x})$ 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 $\|p - q\|_1 \leq \tilde{d}_0/4d^*$

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

Proof: See Appendix D. ■

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 $\gamma \in [0, 1/2]$

$$\gamma \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\gamma} \leq \frac{(|\mathcal{X}||\hat{\mathcal{X}}|)^\delta}{e\delta} \gamma^{1-\delta}. \tag{10}$$

In the sequel, we let $f(\gamma) = \gamma \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\gamma}$ 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. Also, by inverting (10), we have a lower bound on $g(r)$ for any $r \in [0, f(1/2)]$ and $\delta \in (0, 1)$

$$g(r) \geq \left(\frac{e\delta}{(|\mathcal{X}||\hat{\mathcal{X}}|)^{\delta}} r \right)^{1/(1-\delta)}. \tag{11}$$

B. A Bound on the Number of Distributions to Sample

Returning to the problem of computing $R(D)$ in (8), 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 $\tilde{\mathcal{O}} = [0, 1]^{|\mathcal{X}|-1} \bigcap \gamma\mathbb{Z}^{|\mathcal{X}|-1}$. Now, define

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

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 coordinate 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 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.,

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

Checking the condition $\min_{p \in \mathcal{Q}} \|p - q\|_1 \leq \gamma 2|\mathcal{X}|$ is a linear program, so it can be done efficiently. 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

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

The number of distributions for which $R(q, D)$ is computed to determine $R(D)$ to within accuracy ϵ is at most[4]

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

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

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

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

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

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

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

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

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

To get a sense of how $N(\epsilon)$ scales as ϵ goes to 0, we can use the bound of (11) with an arbitrary value of $\delta \in (0, 1)$. For example, with $\delta = 1/2$, the scaling becomes

$$\begin{aligned} N(\epsilon) &\leq \left(\frac{2|\mathcal{X}|}{\left(\frac{e \tilde{d}_0}{22 d^* \sqrt{|\mathcal{X}||\hat{\mathcal{X}}|}} \right)^2} \cdot \frac{1}{\epsilon^2} + 2 \right)^{|\mathcal{X}|-1} \\ &= O\left((1/\epsilon)^{2(|\mathcal{X}|-1)} \right). \end{aligned}$$

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 [13] 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

[4] This is clearly not the best bound as many of the points in the unit cube 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.

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.

Lemma 7.4: Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, d^*]$ be a distortion measure for which Lemma 7.2 holds. For any $p \in \mathcal{P}(\mathcal{X})$, $\tau \in (0,1)$, and $\epsilon \in (0, \ln |\mathcal{X}|)$

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

if

$$n > \frac{2}{g\left(\frac{\epsilon \tilde{d}}{7d^*}\right)^2} \left(\ln \frac{1}{\tau} + |\mathcal{X}| \ln 2 \right). \qquad (12)$$

Proof: From Lemma 7.2, we have

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

The last line follows from [14, Theorem 2.1]. 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 (12) depends only on the distortion measure d, alphabet sizes $|\mathcal{X}|$ and $|\hat{\mathcal{X}}|$, desired accuracy ϵ and "estimation error" probability τ.

VIII. Concluding Remarks

In this paper, we have seen how the rate-distortion function for an AVS is affected by various constraints on the switcher's knowledge involving causality and noise in observations (see Table I). Several other natural constraints come to mind. First, there might be a constraint on how much information the switcher has when making its decisions on subsampling. This could be handled by performing an optimization in Theorem 4.1 over all channels from the subsources to the state observations that satisfy a mutual information constraint. Secondly, one might be interested in studying the rate-distortion function if the switching speed is fixed or constrained in some way. Another interesting area to study might be "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. Here, some understanding of how the rate-distortion function behaves with continuity of the metric might prove useful.

TABLE I
Summary of Results

Switcher model	$R(D)$
Time-invariant (adversarial) [11]	$\max_{p \in \mathcal{G}} R(p, D)$
Strictly causal (adversarial) [4]	$\max_{p \in \overline{\mathcal{G}}} R(p, D)$
Causal or noncausal (adversarial)	$\max_{p \in \mathcal{C}} R(p, D)$
Casual or noncausal noisy observations (adversarial)	$\max_{p \in \mathcal{D}_{states}} R(p, D)$
Noncausal (helpful)	$R^*(\beta, D)$

Aside from thinking of different settings for switcher knowledge, analyzing subsources with memory makes the study of the source coding game even more interesting and potentially relevant to compression. Dobrushin [15] has analyzed the case of the non-anticipatory AVS composed of independent subsources with memory with different distributions when the switcher is passive and blindly chooses the switch position. In the case of subsources 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 subsources with memory and an adversary with k step lookahead, one could imagine that in general

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

Appendix A
Proof of Theorem 3.1

1) Achievability for the Coder: The main tool of the proof is:

Lemma A.1 (Type Covering): Let $S_D(\hat{\mathbf{x}}^n) \triangleq \{\mathbf{x}^n \in \mathcal{X}^n : d_n(\mathbf{x}^n, \hat{\mathbf{x}}^n) \leq D\}$ be the set of $\mathcal{X}^n$ strings that are within distortion D of a given $\hat{\mathcal{X}}^n$ string $\hat{\mathbf{x}}^n$. Fix an $\epsilon > 0$. Then for all $n \geq n_0(d, \epsilon)$, for any $p \in \mathcal{P}_n(\mathcal{X})$, there exists a codebook $\mathcal{B} = \{\hat{\mathbf{x}}^n(1), \hat{\mathbf{x}}^n(2), \ldots, \hat{\mathbf{x}}^n(M)\}$ where $M \leq \exp(n(R(p, D) + \epsilon))$ and

$$T_p^n \subseteq \bigcup_{\hat{\mathbf{x}}^n \in \mathcal{B}} S_D(\hat{\mathbf{x}}^n)$$

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

Proof: See [8, Lemma 2.4.1]. Note that $n_0(d, \epsilon)$ is independent of both p and D. ■

We now show how the coder can get arbitrarily close to $\tilde{R}(D)$ for large enough n. For a $\delta > 0$,

$$\mathcal{C}_\delta \triangleq \left\{ p \in \mathcal{P} \quad : \quad \begin{array}{c} \sum_{x \in \mathcal{V}} p(x) \geq P(\forall\, l, x_l \in \mathcal{V}) - \delta \\ \forall\, \mathcal{V} \text{ such that} \\ \mathcal{V} \subseteq \mathcal{X} \end{array} \right\}.$$

Lemma A.2 (Converse for Switcher): Let $\epsilon > 0$. For all n sufficiently large

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

Proof: Fix a $\lambda > 0$ and $\lambda \leq \lambda(\epsilon) < D - D_{\min}(\mathcal{C})$ to be defined later. We know $R(p, D - \lambda)$ is a continuous function of p ([8]). 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

$$\max_{p \in \mathcal{C}_\delta} R(p, D - \lambda) \leq \max_{p \in \mathcal{C}} R(p, D - \lambda) + \epsilon/3.$$

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 $D - \lambda$. The coder can do this for large n with at most M codewords in the codebook $\mathcal{B}$, where

$$\begin{aligned} M &\leq (n+1)^{|\mathcal{X}|} \exp\left(n\left(\max_{p \in \mathcal{C}_\delta} R(p, D-\lambda) + \epsilon/3\right)\right) \\ &\leq \exp\left(n\left(\max_{p \in \mathcal{C}} R(p, D-\lambda) + \epsilon\right)\right). \end{aligned}$$

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

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

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 there exists a $\mathcal{V}$ such 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 properties of types [7] gives

$$\begin{aligned} \tau &\triangleq P\left(\frac{1}{n}\sum_{k=1}^{n} \zeta_k(\mathcal{V}) < \kappa(\mathcal{V}) - \delta\right) \\ &\leq (n+1)\exp(-nD(p'||q')) \\ &\leq (n+1)\exp(-2n\delta^2). \end{aligned}$$

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

$$\begin{aligned} P(p_{\mathbf{x}^n} \notin \mathcal{C}_\delta) &\leq \sum_{p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)} (n+1)\exp(-2n\delta^2) \\ &\leq (n+1)^{|\mathcal{X}|}\exp(-2n\delta^2) \\ &= \exp\left(-n\left(2\delta^2 - |\mathcal{X}|\frac{\ln(n+1)}{n}\right)\right). \end{aligned}$$

Then, regardless of the switcher strategy

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

So for large n we can get arbitrarily close to distortion $D - \lambda$ while the rate is at most $\max_{p \in \mathcal{C}} R(p, D - \lambda) + \epsilon$. Using the fact that the IID rate-distortion function is continuous in D (uniformly over p such that $D_{\min}(p) < D$, see (20)) gives us that the coder can achieve at most distortion D on average while the asymptotic rate is at most $\tilde{R}(D) + 2\epsilon$ (provided $\lambda \leq \lambda(\epsilon)$ is small enough). Since ϵ is arbitrary, $R(D) \leq \tilde{R}(D)$. ■

2) Achievability for the Switcher: This section shows that $R(D) \geq \tilde{R}(D)$ when the switcher has 1-step lookahead. We 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 $R(D) \geq \tilde{R}(D)$.

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

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

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 A.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

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

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 $\kappa(\mathcal{V}) = P(x_l \in \mathcal{V}, 1 \le l \le m)$. Then, we have

$$\sum_{i=1}^{|\mathcal{X}|} a_i q(i) = a_1\kappa(\{1\}) + a_2[\kappa(\{1,2\}) - \kappa(\{1\})] + \cdots + a_{|\mathcal{X}|}[\kappa(\{1,\ldots,|\mathcal{X}|\}) - \kappa(\{1,\ldots,|\mathcal{X}|-1\})].$$

By the constraints in the definition (3) of $\mathcal{C}$, we have the following inequalities for p:

$$\begin{aligned} p(1) &\ge \kappa(\{1\}) = q(1) \\ p(1) + p(2) &\ge \kappa(\{1,2\}) = q(1) + q(2) \\ &\vdots \\ \sum_{i=1}^{|\mathcal{X}|-1} p(i) &\ge \kappa(\{1,\ldots,|\mathcal{X}|-1\}) = \sum_{i=1}^{|\mathcal{X}|-1} q(i). \end{aligned}$$

Therefore, the difference of the objective is

$$\begin{aligned} \lambda &\triangleq \sum_{i=1}^{|\mathcal{X}|} a_i(p(i) - q(i)) \\ &= a_{|\mathcal{X}|}\left[\sum_{i=1}^{|\mathcal{X}|} p(i) - q(i)\right] + \\ &\quad (a_{|\mathcal{X}|-1} - a_{|\mathcal{X}|})\left[\sum_{i=1}^{|\mathcal{X}|-1} p(i) - q(i)\right] + \\ &\quad \cdots + (a_1 - a_2)[p(1) - q(1)] \\ &= \sum_{i=1}^{|\mathcal{X}|-1} (a_i - a_{i+1})\left[\sum_{j=1}^{i} p(j) - \sum_{j=1}^{i} q(j)\right] \\ &\ge 0. \end{aligned}$$

The last step is true because of the monotonicity in the a_i and the inequalities we derived earlier. Therefore, we see that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) \ge \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$ 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 B
Proof of Theorem 4.1

It is clear that $R(D) \ge \max_{p\in\mathcal{D}_{\text{states}}} R(p,D)$ because the switcher can select distributions $f(\cdot|t) \in \bar{\mathcal{G}}(t)$ 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, $R(D) \ge \max_{p\in\mathcal{D}_{\text{states}}} R(p,D)$.

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

Lemma B.1: Let $p_{\mathbf{x}^n}$ be the type of $\mathbf{x}^n$ and for $\epsilon > 0$ let $\mathcal{D}_{\text{states},\epsilon}$ be the set of $p \in \mathcal{P}(\mathcal{X})$ with $\mathcal{L}_1$ distance at most ϵ from a distribution in $\mathcal{D}_{\text{states}}$. Then, for $\epsilon > 0$

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

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

$$A_\delta^t = \left\{\left|\frac{N(t|\mathbf{t}^n)}{n} - \alpha(t)\right| > \delta\right\}. \tag{13}$$

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 [16]

$$P(A_\delta^t) \le 2\exp(-2n\delta^2).$$

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 $\bar{\mathcal{G}}(t)$. This will be done by a martingale argument similar to that given in [4, Lemma 3]. 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, 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

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

and let $S_0 = 0$. For $k \ge 1$

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

We claim that $S_k, k \ge 1$ is a martingale[5] 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 $t \in \mathcal{T}$

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

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

and analogously

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

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

$$\begin{aligned}\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t &= \frac{1}{N(t\,|\,\mathbf{t}^n)} \sum_{i=1}^{n} T_i \mathbf{1}(t_i = t)\\ &= \frac{1}{N(t\,|\,\mathbf{t}^n)} \sum_{i:t_i=t} C_i -\\ &\quad \frac{1}{N(t\,|\,\mathbf{t}^n)} \sum_{i:t_i=t} \mathbb{E}[C_i|\mathcal{F}_{i-1}]. \qquad (14)\end{aligned}$$

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 $t_i = t$,

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

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 symbol 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 (14) is in $\bar{\mathcal{G}}(t)$ because it is the average of $N(t\,|\,\mathbf{t}^n)$ terms in $\bar{\mathcal{G}}(t)$ and $\bar{\mathcal{G}}(t)$ 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 $\bar{\mathcal{G}}(t)$.

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

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

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

$$\begin{aligned}\varsigma &\triangleq P\left(\bigcup_{t \in \mathcal{T}} \{p_{\mathbf{x}^n}^t \notin \bar{\mathcal{G}}(t)_\epsilon\}\right)\\ &\le \sum_{t \in \mathcal{T}} P\left(\bigcup_{x \in \mathcal{X}} \left\{\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right\}\right)\\ &\le \sum_{t}\sum_{x} P\left(\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right). \qquad (15)\end{aligned}$$

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

$$\begin{aligned}\mu &\triangleq P\left(\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right)\\ &\le P\left(A_\delta^t\right) + P\left(\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, \left(A_\delta^t\right)^c\right)\\ &\le 2\exp(-2n\delta^2) +\\ &\quad P\left(\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right).\end{aligned}$$

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

$$\begin{aligned}\nu &\triangleq P\left(\left|\frac{1}{N(t\,|\,\mathbf{t}^n)} S_n^t(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right)\\ &\le P\left(\left|S_n^t(x)\right| > n(\alpha(t) - \delta)\frac{\epsilon}{|\mathcal{X}|}, (A_\delta^t)^c\right)\\ &\le P\left(\left|S_n^t(x)\right| > n(\alpha(t) - \delta)\frac{\epsilon}{|\mathcal{X}|}\right).\end{aligned}$$

$S_k^t(x)$ 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 [17] to get

$$\nu \le 2\exp\left(-n\frac{(\alpha(t) - \delta)^2 \epsilon^2}{2|\mathcal{X}|^2}\right). \qquad (16)$$

Plugging this back into (15)

$$\begin{aligned}\varsigma &= P\left(\bigcup_{t \in \mathcal{T}} \{p_{\mathbf{x}^n}^t \notin \bar{\mathcal{G}}(t)_\epsilon\}\right)\\ &\le 2|\mathcal{T}||\mathcal{X}| \Bigg(\exp(-2n\delta^2) +\\ &\qquad \exp\left(-n\frac{(\alpha_* - \delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right)\Bigg)\\ &\le 4|\mathcal{X}||\mathcal{T}|\exp(-n\xi(\epsilon, \delta))\end{aligned}$$

where

$$\begin{aligned}\xi(\epsilon, \delta) &= \min\left\{2\delta^2, \frac{(\alpha_* - \delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right\}\\ \alpha_* &\triangleq \min_{t \in \mathcal{T}} \alpha(t).\end{aligned}$$

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

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

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

$$P\left(\bigcup_{t \in \mathcal{T}} \{p_{\mathbf{x}^n}^t \notin \bar{\mathcal{G}}(t)_\epsilon\}\right) \le 4|\mathcal{X}||\mathcal{T}|\exp(-n\tilde{\xi}(\epsilon)).$$

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

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

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

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

From (13), we are assumed to be in the event that

$$\left| \frac{N(t\,|\,\mathbf{t}^n)}{n\alpha(t)} - 1 \right| \le \frac{\delta}{\alpha(t)}.$$

Hence

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

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

Appendix C
Proof of Lemma 7.2

Let $W^*_{p,D} \in \operatorname{argmin}_{W \in \mathcal{W}(p,D)} I(p, W)$. Then

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

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

$$\begin{aligned}d(p, W^*_{q,D}) &\le d(q, W^*_{q,D}) + |d(p, W^*_{q,D}) - d(q, W^*_{q,D})| \\ &\le D + \|p - q\|_1 d^*.\end{aligned}$$

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

$$\begin{aligned}R(p, d(p, W^*_{q,D})) &\le I(p, W^*_{q,D}) \\ &\le |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad I(q, W^*_{q,D}) \\ &= |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad + R(q, D).\end{aligned} \tag{17}$$

Expanding mutual informations yields

$$\begin{aligned}\omega &\triangleq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \\ &\le |H(p) - H(q)| + |H(pW^*_{q,D}) - H(qW^*_{q,D})| + \\ &\quad |H(p, W^*_{q,D}) - H(q, W^*_{q,D})|.\end{aligned}$$

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

$$\begin{aligned}\omega &= |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \\ &\le 3\|p - q\|_1 \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1}.\end{aligned}$$

Now, we have seen $d(p, W^*_{q,D}) \le D + d^*\|p - q\|_1$. 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 (17), namely

$$\begin{aligned}\chi &\triangleq R(p,D) - R(q,D) \\ &\le |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad R(p, D) - R(p, d(p, W^*_{q,D})) \\ &\le |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\quad R(p, D) - R(p, D + d^*\|p - q\|_1)\end{aligned} \tag{18}$$

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-∪ and decreasing and so has steepest descent at $D = 0$. Therefore, for any $0 \le D_1, D_2 \le d^*$

$$|R(p, D_1) - R(p, D_2)| \le |R(p, 0) - R(p, |D_2 - D_1|)|.$$

Hence, we can restrict our attention to continuity of $R(p, D)$ around $D = 0$. By assumption, $\mathcal{W}(p, 0) \ne \emptyset\ \forall p \in \mathcal{P}(\mathcal{X})$. 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

$$\begin{aligned}D &\ge \sum_x p(x) \sum_{\hat{x}} W(\hat{x}\,|\,x) d(x, \hat{x}) \\ &= \sum_x p(x) \sum_{\hat{x}: d(x,\hat{x}) > 0} W(\hat{x}\,|\,x) d(x, \hat{x}) \\ &\ge \tilde{d} \sum_x p(x) \sum_{\hat{x}: d(x,\hat{x}) > 0} W(\hat{x}\,|\,x).\end{aligned} \tag{19}$$

Now, we will construct a channel in $\mathcal{W}(p, 0)$, denoted W_0. First, for each $x, \hat{x}$ such that $d(x, \hat{x}) = 0$, let $V(\hat{x}\,|\,x) = W(\hat{x}\,|\,x)$. For all other $(x, \hat{x})$, set $V(\hat{x}\,|\,x) = 0$. 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 $(x, \hat{x})$ with $d(x, \hat{x}) = 0$. Namely, for $(x, \hat{x})$ with $d(x, \hat{x}) = 0$, we define

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

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

$$\begin{aligned}\varrho(x) &\triangleq \sum_{\hat{x}} |W(\hat{x}\,|\,x) - W_0(\hat{x}\,|\,x)| \\ &= \sum_{\hat{x}:d(x,\hat{x})>0} W(\hat{x}\,|\,x) + \\ &\quad \sum_{\hat{x}:d(x,\hat{x})=0} |W(\hat{x}\,|\,x) - W_0(\hat{x}\,|\,x)| \\ &= \sum_{\hat{x}:d(x,\hat{x})>0} W(\hat{x}\,|\,x) + \\ &\quad \sum_{\hat{x}:d(x,\hat{x})=0} \left| \frac{\sum_{\hat{x}':d(x,\hat{x}')>0} W(\hat{x}'|x)}{|\{\hat{x}' : d(x,\hat{x}') = 0\}|} \right| \\ &= 2 \sum_{\hat{x}:d(x,\hat{x})>0} W(\hat{x}\,|\,x).\end{aligned}$$

Therefore, using (19)

$$\sum_{x} p(x) \sum_{\hat{x}} |W(\hat{x}\,|\,x) - W_0(\hat{x}\,|\,x)| \le \frac{2D}{\tilde{d}}.$$

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 $2D/\tilde{d}$. Going back to the bound on $|R(p, 0) - R(p, D)|$

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

It can be easily verified that $\|pW_0 - pW^*_{p,D}\|_1$ is at most $2D/\tilde{d}$. Similarly, $\|(p, W_0) - (p, W^*_{p,D})\|_1 \le 2D/\tilde{d}$.

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

$$\begin{aligned}|R(p, 0) - R(p, D)| &\le \frac{2D}{\tilde{d}} \ln \frac{\tilde{d}|\mathcal{X}|}{2D} + \frac{2D}{\tilde{d}} \ln \frac{\tilde{d}|\mathcal{X}||\hat{\mathcal{X}}|}{2D} \\ &\le \frac{4D}{\tilde{d}} \ln \frac{\tilde{d}|\mathcal{X}||\hat{\mathcal{X}}|}{2D}. \qquad (20)\end{aligned}$$

Going back to (18), we see that if $\|p - q\|_1 \le \frac{\tilde{d}}{4d^*}$,

$$\begin{aligned}\psi &\triangleq |R(p, D + d^*\|p - q\|_1) - R(p, D)| \\ &\le \frac{4d^*\|p - q\|_1}{\tilde{d}} \ln \frac{\tilde{d}|\mathcal{X}||\hat{\mathcal{X}}|}{2d^*\|p - q\|_1} \\ &\le \frac{4d^*\|p - q\|_1}{\tilde{d}} \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1}.\end{aligned}$$

The last step follows because $\tilde{d}/d^* \le 1$. Substituting into (18) gives

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

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

$$R(q, D) - R(p, D) \le \frac{7d^*}{\tilde{d}}\|p - q\|_1 \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1}.$$

Appendix D
Proof of Lemma 7.3

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

$$d_0(x, \hat{x}) = d(x, \hat{x}) - \min_{\tilde{x} \in \hat{\mathcal{X}}} d(x, \tilde{x})$$

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 $d_0(x, \hat{x})$. By definition, $R(p, D)$ is the IID rate-distortion function for p with respect to distortion measure $d(x, \hat{x})$. From [7, Problem 13.4], for any $D \ge D_{\min}(p)$,

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

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

$$\begin{aligned}\zeta &\triangleq |R(p, D) - R(q, D)| \\ &= |R_0(p, D - D_{\min}(p)) - R_0(q, D - D_{\min}(q))| \\ &\le |R_0(p, D - D_{\min}(p)) - R_0(p, D - D_{\min}(q))| + \\ &\quad |R_0(p, D - D_{\min}(q)) - R_0(q, D - D_{\min}(q))|. \qquad (21)\end{aligned}$$

Now, we note that $|D_{\min}(p) - D_{\min}(q)| \le d^*\|p - q\|_1$. The first term of (21) can be bounded using (20) and the second term of (21) can be bounded using Lemma 7.2. The first term can be bounded if $\|p - q\|_1 \le \tilde{d}_0/4d^*$ and the second can be bounded if $\|p - q\|_1 \le \tilde{d}_0/4d_0^*$. Since $d_0^* \le d^*$, we only require $\|p - q\|_1 \le \tilde{d}_0/4d^*$

$$\begin{aligned}\zeta &\le \frac{4d^*}{\tilde{d}_0}\|p - q\|_1 \ln \frac{\tilde{d}_0|\mathcal{X}||\hat{\mathcal{X}}|}{2d^*\|p - q\|_1} + \\ &\quad \frac{7d_0^*}{\tilde{d}_0}\|p - q\|_1 \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1} \\ &\le \frac{11d^*}{\tilde{d}_0}\|p - q\|_1 \ln \frac{|\mathcal{X}||\hat{\mathcal{X}}|}{\|p - q\|_1}.\end{aligned}$$

Acknowledgment

We thank the Associate Editor and anonymous reviewers for their insightful comments that helped to clarify the presentation of these results.

References

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

[2] H. Palaiyanur and A. Sahai, "On the uniform continuity of the rate-distortion function," in *Proc. Int. Symp. Information Theory*, Toronto, ON, Canada, Jul. 2008.

[3] H. Palaiyanur, C. Chang, and A. Sahai, "Lossy compression of active sources," in *Proc. Int. Symp. Information Theory*, Toronto, ON, Canada, Jul. 2008.

[4] T. Berger, "The source coding game," *IEEE Trans. Inform. Theory*, vol. 17, pp. 71–76, Jan. 1971.

[5] R. Bajcsy, "Active perception," *Proc. IEEE*, vol. 76, no. 8, pp. 966–1005, Aug. 1988.

[6] R. Ahlswede, "Extremal properties of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 36, pp. 166–171, Jan. 1990.

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

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

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

[10] J. Wolfowitz, "Approximation with a fidelity criterion," in *Proc. 5th Berkeley Symp. Math. Stat. and Prob.*, Berkeley, CA, 1967, vol. 1, pp. 565–573.

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

[12] R. Gallager*, Information Theory and Reliable Communication*. New York: Wiley, 1971.

[13] M. Harrison and I. Kontoyiannis, "Estimation of the Rate-Distortion Function" 2007 [Online]. Available: http://arxiv.org/abs/cs/0702018v1

[14] 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

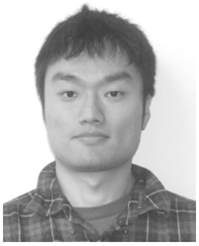
[15] R. Dobrushin, "Unified methods for the transmission of information: The general case," *Soviet Math.*, vol. 4, pp. 284–292, 1963.

[16] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

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

Hari Palaiyanur received the B.Sc. degree in electrical and computer engineering from Cornell University, Ithaca, NY, in August 2004 and the M.Sc. degree in electrical engineering and computer sciences in May 2006 from the University of California at Berkeley, where he is currently pursuing the Ph.D. degree.

In the summer of 2007, he was a research intern at Nokia Research Center, Palo Alto, CA. His research interests are in information theory and probability.



Cheng Chang (M'08) received the B.E. degree from Tsinghua University, Beijing, China, in 2000, and the Ph.D. degree from the University of California at Berkeley in 2007.

He is currently a Quantitative Analyst with the D.E. Shaw Group, New York. In 2008, he spent a year in the Information Theory Group at HP Labs as a Postdoctoral Researcher. His research interests include signal processing, control theory, information theory, and machine learning.

Anant Sahai (S'94–M'00) received the B.S. degree from the University of California at Berkeley in 1994 and the S.M. and Ph.D. degrees from Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2001, respectively.

He is a member of the Wireless Foundations Center and an Associate Professor in the Department of Electrical Engineering and Computer Sciences, both at the University of California at Berkeley, where he joined the faculty in 2002. In 2001, he spent a year at the wireless startup Enuvis, developing adaptive software-radio algorithms for extremely sensitive GPS receivers. Prior to that, he was a graduate student at the Laboratory for Information and Decision Systems, MIT. His research interests span wireless communication, decentralized control, and information theory. He is particularly interested in spectrum sharing, the nature of information in control systems, and power-consumption.